axioms

MDPI

Article

Weighted Least Squares Regression with the Best Robustness and High Computability

Yijun Zuo [1,*] and Hanwen Zuo [2]

[1] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA
[2] Department of Computer Science, Michigan State University, East Lansing, MI 48824, USA; zuohanwe@msu.edu
* Correspondence: zuo@msu.edu

Abstract: A novel regression method is introduced and studied. The procedure weights squared residuals based on their magnitude. Unlike the classic least squares which treats every squared residual as equally important, the new procedure exponentially down-weights squared residuals that lie far away from the cloud of all residuals and assigns a constant weight (one) to squared residuals that lie close to the center of the squared-residual cloud. The new procedure can keep a good balance between robustness and efficiency; it possesses the highest breakdown point robustness for any regression equivariant procedure, being much more robust than the classic least squares, yet much more efficient than the benchmark robust method, the least trimmed squares (LTS) of Rousseeuw. With a smooth weight function, the new procedure could be computed very fast by the first-order (first-derivative) method and the second-order (second-derivative) method. Assertions and other theoretical findings are verified in simulated and real data examples.

Keywords: weighted least squares; robustness; efficiency; computability; finite sample breakdown point

MSC: 62G35; 62J05; 62G99; 62G05

Citation: Zuo, Y.; Zuo, H. Weighted Least Squares Regression with the Best Robustness and High Computability. *Axioms* **2024**, *13*, 295. https://doi.org/10.3390/axioms13050295

Academic Editor: Angel Ricardo Plastino

Received: 21 March 2024
Revised: 15 April 2024
Accepted: 21 April 2024
Published: 27 April 2024

1. Introduction

In the classical regression analysis, we assume that there is a relationship for a given data set $\{(\boldsymbol{x}_i^\top, y_i)^\top, i \in \{1, 2, \cdots, n\}\}$:

$$y_i = (1, \boldsymbol{x}_i^\top)\boldsymbol{\beta}_0 + e_i, \;\; i \in \{1, \cdots, n\} \tag{1}$$

where $y_i \in \mathbb{R}^1$, $\top$ stands for the transpose, $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})^\top$ (the true unknown parameter) in $\mathbb{R}^p$ and $\boldsymbol{x_i} = (x_{i1}, \cdots, x_{i(p-1)})^\top$ in $\mathbb{R}^{p-1}$ ($p \geq 2$), $e_i \in \mathbb{R}^1$ is called an error term (or random fluctuation/disturbances, which is usually assumed to have zero mean and variance σ^2 in classic regression theory). That is, β_{01} is the intercept term of the model. We write $\boldsymbol{w}_i = (1, \boldsymbol{x}_i')^\top$, then one has $y_i = \boldsymbol{w}_i^\top \boldsymbol{\beta}_0 + e_i$, which is used interchangeably with (1).

One wants to estimate the $\boldsymbol{\beta}_0$ based on a given sample $\mathbf{z}^{(n)} := \{(\boldsymbol{x}_i^\top, y_i)^\top, i \in \{1, \cdots, n\}\}$ from the model $y = (1, \boldsymbol{x}^\top)\boldsymbol{\beta}_0 + e$. Call the difference between y_i and $\boldsymbol{w}_i^\top \boldsymbol{\beta}$ the i-th residual $r_i(\boldsymbol{\beta})$ for a candidate coefficient vector $\boldsymbol{\beta}$ (which is often suppressed). That is,

$$r_i := r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta}. \tag{2}$$

To estimate $\boldsymbol{\beta}_0$, the classic *least squares* (LS) minimizes the sum of squares of residuals,

$$\widehat{\boldsymbol{\beta}}_{ls} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} r_i^2.$$

Alternatively, one can replace the square above by the absolute value to obtain the least absolute deviations estimator (i.e., L_1 estimator, in contrast to the L_2 (LS) estimator).

The LS estimator is very popular in practice across a broader spectrum of disciplines due to its great computability and optimal properties when the error e_is is i.i.d. and follows a normal $N(\mathbf{0}, \sigma^2)$ distribution. It, however, can behave badly when the error distribution is slightly departed from the normal distribution, particularly when the errors are heavy-tailed or contain outliers.

Robust alternatives to the $\widehat{\boldsymbol{\beta}}_{ls}$ have abounded in the literature for a long time. The most popular ones are M-estimators [1], the least median squares (LMS) and least trimmed squares (LTS) estimators [2], S-estimators [3], MM-estimators [4], τ-estimators [5], maximum depth estimators ([6,7]), and the recent least squares of trimmed residuals (LST) regression [8], among others. For more related discussions, see Sections 1.2 and 4.4 of [9], and Section 5.14 of [10].

Robust methods that have a high breakdown point are usually computationally intensive and with a non-differentiable objective function (e.g., LMS, LTS, and LST). In this article, we will introduce a smooth and differentiable objective function that greatly facilitates the computation of the underlying estimator. We introduce a new class of alternatives for robust regression, weighted least squares (WLS) estimators $\hat{\boldsymbol{\beta}}_{wls}$:

$$\widehat{\boldsymbol{\beta}}_{wls} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_i^n w_i r_i^2(\boldsymbol{\beta}), \tag{3}$$

where w_i is the weight associated with r_i with a fundamental feature: it assigns equal weight to all r_i^2 that are small (no greater than a cut-off value) and exponentially down-weights (penalizes) the large ones (when r_i^2s are greater than the cut-off value).

Weighted least squares estimation has been proposed and discussed in the literature, including the famous Huber's M-estimators which, however, can have the lowest breakdown point if the derivative of the weight (or loss) function is non-decreasing; see [9] (p. 13) or [10,11]. For more discussions, see Section 1.2 of [9] or Section 5.11 of [10]. Previous weight functions in the literature are either constants (e.g., LS with 1, or LMS and LTS with 0 and 1 weight), rank-based weight, do not down-weight large residuals sufficiently, or non-differentiable. Among these weight-induced regression estimators, there are few that possess a high breakdown point (50%), a high efficiency, and a high computability, simultaneously.

On the other hand, there is much room for smooth weight functions. Successful examples in location setting have already appeared in the literature, e.g., [12]. This motivates us to extend those smooth weight functions to regression setting and to achieve a high breakdown point and high efficiency and high computablity simultaneously. We propose using a differentiable $w(r)$, which would assign weight one to r_is that lies close to the center of the data (all r_is) cloud. The other points that lie on the outskirts of the data (all r_is) cloud could be viewed as outliers, so a lower positive weight (not zero) should be given. This would balance efficiency with robustness. The weighted procedure proposed in this article has never appeared before. The specially chosen w_i's in (3) will recover the famous LMS and the LTS in [2], and LST in [8]. More discussions on w and $\widehat{\boldsymbol{\beta}}_{wls}$ are carried out in Section 2, where an ad hoc choice of the weight function with the above property in mind will be introduced.

The rest of this article is organized as follows. Section 2 introduces a class of differentiable weight functions and a class of weighted least squares estimators. Section 3 establishes the existence of $\widehat{\boldsymbol{\beta}}_{wls}$ and studies its properties including its finite sample breakdown robustness. Section 4 discusses the computation of $\widehat{\boldsymbol{\beta}}_{wls}$. Section 5 presents some concrete examples, comparing the performance of $\widehat{\boldsymbol{\beta}}_{wls}$ with other leading estimators. Section 6 ends the article with some concluding remarks. Long proofs of the main results are deferred to in Appendix A.

2. A Class of Weighted Least Squares

2.1. A Class of Weight Functions

An ad hoc choice of the weight function with the property mentioned in Section 1 takes the form of

$$w(x) = \mathbb{1}(|x| \le c) + \mathbb{1}(|x| > c)\frac{e^{-k(1-c/|x|)^2} - e^{-k}}{1 - e^{-k}}, \quad \forall c, k > 0, \tag{4}$$

where the tuning parameter $k > 1$ is a positive number (say, between 1 and 10) controlling the steepness of the exponential decrease (see the left panel of Figure 1), where the larger the k, the steeper the curve (the key difference from the trimmed procedures where the weight becomes zero). Tuning parameter c is the point where the weight function will change from a constant one to being exponentially decreasing. c (>1) usually can be set to be a large positive number (say 10), or it can be residual dependent, say 50% or 75% percentile of the residuals, and a larger c is favorable for higher efficiency. c is assumed to be finite to exclude the LS case (i.e., $w(x)$ will not be a constant one).

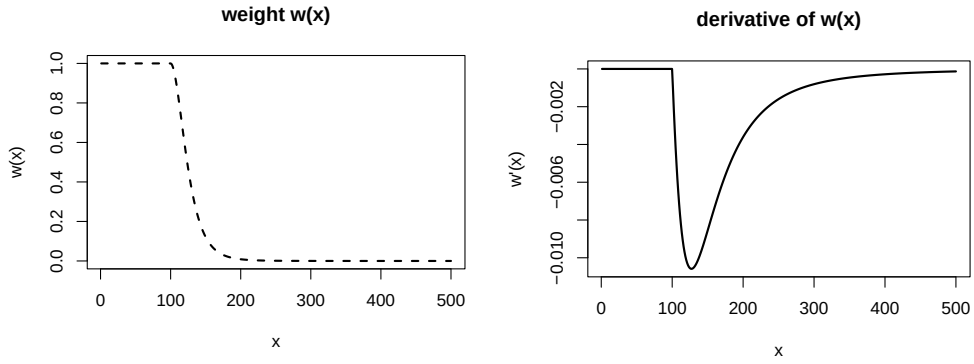

Figure 1. Weight function $w(x)$ when $k = 5$ and $c = 100$. **Left**: $w(x)$, **right**: $w'(x)$.

One of the examples of $w(x)$ is given in Figure 1, where $w(x)$ and its derivative are given and $k = 5$ and $c = 100$. For a general $w(x)$, it is straightforward to verify that

P1 $w(x)$ is twice differentiable and $0 < w(x) \le 1$. When $x \to \infty$, $w(x)$ is asymptotically equivalent to $\alpha(e^{\gamma x^{-1}} - 1)$ for some positive constants α and γ.

P2 If $r_i \to \infty$, then $w(r_i^2/c^*)r_i^2 \to 2ckc^*/(e^k - 1)$, where $c^* := \text{Med}_i\{y_i^2\}$, the median of $\{y_1^2, y_2^2, \cdots, y_n^2\}$.

2.2. Weighted Least Squares Estimators

With the weight function given above, we are ready to specify the weighted least squares estimator in (3) with more details:

$$\widehat{\boldsymbol{\beta}}_{wls} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_i^n w_i r_i^2(\boldsymbol{\beta}), \tag{5}$$

where weight $w_i := w(r_i^2/c^*)$, with $w(x)$ being a weight function in (4) that satisfies **P2** and c^* defined in **P2**.

The behavior of function $w(r^2/c^*)r^2$ when $r > c$ for different c^*s is illustrated in Figure 2 below. Inspecting the figure reveals that it is strictly convex.

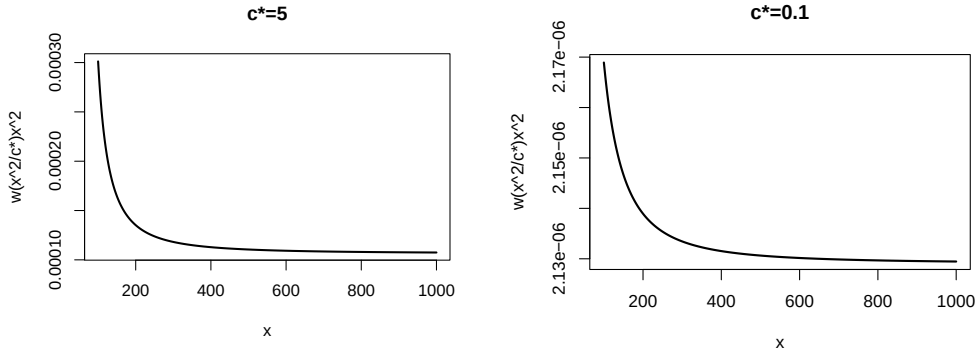

Figure 2. Behavior of function $w(x^2/c^*)x^2$ when $k = 5$ and $c = 100$, $x > c$.

3. Properties of the $\widehat{\boldsymbol{\beta}}_{wls}$

3.1. Existence

Does the minimizer of the objective function $O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)}) := \sum_i^n w_i r_i^2(\boldsymbol{\beta})$ on the right-hand side (RHS) of (5) exist? We now formally address this. We need the following assumption.

A1: For a given sample $\boldsymbol{z}^{(n)} := \{(\boldsymbol{z}_i)_{i=1}^n\} = \{(\boldsymbol{x}_i^\top, y_i)^\top, i \in \{1, 2, \cdots, n\}\}$ and any $\boldsymbol{\beta} \in \mathbb{R}^p$, all points $\{(\boldsymbol{x}_i^\top, y_i)^\top\}$ with r_is satisfying $r_i^2/c^* \leq c$ do not lie in a vertical hyperplane.

The assumption holds true with probability one if the sample comes from a distribution of $(\boldsymbol{x}^\top, y)^\top$ that has a density. Now, we have the following existence result.

Theorem 1. *If **A1** holds true, then the minimizer $\widehat{\boldsymbol{\beta}}_{wls}$ of $O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)})$ always exists.*

Proof. See the Appendix A. □

3.2. Equivariance

Desirable fundamental properties of regression estimators include regression, scale, and affine equivarince. For $\mathbf{x} \in \mathbb{R}^{n\times(p-1)}$ and $\mathbf{y} \in \mathbb{R}^n$, a regression estimator $\hat{\boldsymbol{\beta}} := \mathbf{t}(\mathbf{w}, \mathbf{y})$ with $\mathbf{w} = (1, \boldsymbol{x}^\top)^\top$ satisfying

$$\mathbf{t}(\mathbf{w}, \mathbf{y} + \mathbf{b}^\top \mathbf{w}) = \mathbf{t}(\mathbf{w}, \mathbf{y}) + \mathbf{b}, \ \forall \mathbf{b} \in \mathbb{R}^p; \tag{6}$$

$$\mathbf{t}(\mathbf{w}, s\mathbf{y}) = s\mathbf{t}(\mathbf{w}, \mathbf{y}), \ \forall s \in \mathbb{R}; \tag{7}$$

$$\mathbf{t}(\mathbf{A}^\top \mathbf{w}, \mathbf{y}) = \mathbf{A}^{(-1)}\mathbf{t}(\mathbf{w}, \mathbf{y}), \ \forall \text{ nonsingular } \mathbf{A} \in \mathbb{R}^{p\times p}. \tag{8}$$

is called *regression*, *scale*, and *affine* equivariant, respectively (see page 116 of [9]). All aforementioned regression estimators are regression, scale, and affine equivariant.

Theorem 2. $\widehat{\boldsymbol{\beta}}_{wls}$ *defined in (3) is* regression, scale, *and* affine *equivariant.*

Proof. Notice the identities $r_i = y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} = y_i + \boldsymbol{b}^\top \boldsymbol{w}_i - \boldsymbol{w}_i^\top(\boldsymbol{\beta} + \boldsymbol{b})$, $sr_i = sy_i - \boldsymbol{w}_i^\top(s\boldsymbol{\beta})$, and $r_i = y_i - (\boldsymbol{A}^\top \boldsymbol{w}_i)^\top \boldsymbol{A}^{-1}\boldsymbol{\beta}$. Meanwhile, r_i^2/c^* is regression, scale, and affine invariant. The desired result follows. □

3.3. Robustness

As an alternative to the least-squares $\widehat{\boldsymbol{\beta}}_{ls}$, is the $\widehat{\boldsymbol{\beta}}_{wls}$ more robust?

The most prevailing quantitative measure of the global robustness of any location or regression estimators in the finite sample practice is the *finite sample breakdown point* (FSBP), introduced in [13].

Roughly speaking, the FSBP is the minimum fraction of 'bad' (or contaminated) data that the estimator can be affected by to an arbitrarily large extent. For example, in the context of estimating the center of a data set, the sample mean has a breakdown point of $1/n$ (or 0%) because even one bad observation can change the mean by an arbitrary amount;

in contrast, the sample median has a breakdown point of $\lfloor (n+1)/2 \rfloor / n$ (or 50%), where $\lfloor \cdot \rfloor$ is the floor function.

Definition 1 ([13]). *The finite sample* replacement breakdown point *(RBP) of a regression estimator* $\mathbf{t}$ *at the given sample* $\mathbf{z}^{(n)} = \{z_1, \cdots, z_n\}$, *where* $z_i := (\boldsymbol{x}_i^\top, y_i)^\top$, *is defined as*

$$RBP(\mathbf{t}, \mathbf{z}^{(n)}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\mathbf{z}_m^{(n)}} \|\mathbf{t}(\mathbf{z}_m^{(n)}) - \mathbf{t}(\mathbf{z}^{(n)})\| = \infty \right\}, \tag{9}$$

where $\mathbf{z}_m^{(n)}$ *denotes an arbitrary contaminated sample by replacing* m *original sample points in* $\mathbf{z}^{(n)}$ *with arbitrary points in* $\mathbb{R}^p$. *Namely, the RBP of an estimator is the minimum replacement fraction that could drive the estimator beyond any bound. It turns out that both* L_1 *(least absolute deviations) and* L_2 *(least squares) estimators have RBP* $1/n$ *(or* 0%*), the lowest possible value, whereas* $\widehat{\boldsymbol{\beta}}_{wls}$ *can have* $(\lfloor (n-p)/2 \rfloor + 1)/n$ *(or* 50%*), the highest possible value for any regression equivariant estimators (see p. 125 of [9]).*

We shall say $\mathbf{z}^{(n)}$ is *in a general position* when any p of observations in $\mathbf{z}^{(n)}$ gives a unique determination of $\boldsymbol{\beta}$. In other words, any (p-1) dimensional subspace of the space $(\boldsymbol{x}^\top, y)^\top$ contains at most p observations of $\mathbf{z}^{(n)}$. When the observations come from continuous distributions, the event ($\mathbf{z}^{(n)}$ being in the general position) happens with a probability of one.

Theorem 3. *Assume that* ***A1*** *holds true,* $n > p$, *and* $\mathbf{z}^{(n)}$ *is in the general position. Then,*

$$RBP(\widehat{\boldsymbol{\beta}}_{wls}^n, \mathbf{z}^{(n)}) = \begin{cases} \lfloor (n+1)/2 \rfloor / n, & \text{if } p = 1, \\ (\lfloor (n-p)/2 \rfloor + 1)/n, & \text{if } p > 1. \end{cases} \tag{10}$$

Proof. see the Appendix A. □

We need the following important result for the proof of Theorem 3.

Lemma 1. *For any* $r_i^2 > r_j^2 > c^* c$, $w(r_i^2/c^*) r_i^2 < w(r_j^2/c^*) r_j^2$ *when* $r_j^2 \to \infty$.

Proof. See the Appendix A. □

Remark 1. *The RBP result in Theorem 3 is the highest possible breakdown point for any regression equivariant estimators in the literature (see p. 125 of [9]). There are very few regression estimators that possess the highest breakdown point robustness.*

4. Computation of the WLS

Now, we address the most important issue with a high breakdown point estimator, the computation of the estimator. The objective function in (5) is

$$O(\boldsymbol{\beta}) := O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)}) = \sum_{i=1}^{n} w(r_i^2/c^*) r_i^2, \tag{11}$$

which is differentiable with respect to $\boldsymbol{\beta}$ since the weight function $w(x^2/c^*)$ is twice differentiable with

$$w'(x) = \alpha^* e^{-k(1-c/|x|)^2} (1 - c/|x|) \mathrm{sgn}(x)/x^2 \mathbb{1}(|x| > c),$$

$$w''(x) = \alpha^* e^{-k(1-c/|x|)^2} \left(-2kc(1-c/|x|)^2/|x| - (2 - 3c/|x|)\right)/x^3 \mathbb{1}(|x| > c), \tag{12}$$

where $\alpha^* = -2kc/(1 - e^{-k})$. The problem in (3) belongs to an unconstrained minimization. This type of problem has been thoroughly discussed and studied in the literature. Common

approaches to find the solution include (i) methods utilizing first-order derivatives (gradient descent/steepest descent/conjugate gradient), (ii) methods using second-order derivatives (Hessian matrix) (Newton's method), and (iii) quasi-Newton method, see [14,15]. We will select the conjugate gradient for speed/efficiency and accuracy consideration.

Note that

$$\begin{aligned}\nabla O(\boldsymbol{\beta}) = \frac{\partial O(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} (w'(r_i^2/c^*)r_i^2 + c^* w(r_i^2/c^*)) \frac{\partial r_i^2/c^*}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} (w'(r_i^2/c^*)r_i^2 + c^* w(r_i^2/c^*)) 2r_i/c^*(-\boldsymbol{w}_i) \\ &= \sum_{i=1}^{n} -2r_i/c^*(w'(r_i^2/c^*)r_i^2 + c^* w(r_i^2/c^*))\boldsymbol{w}_i. \end{aligned} \tag{13}$$

$$\begin{aligned}\nabla^2 O(\boldsymbol{\beta}) = \frac{\partial^2 O(\boldsymbol{\beta})}{\partial^2 \boldsymbol{\beta}} &= \frac{-2}{c^*} \sum_{i=1}^{n} \frac{\partial \left(r_i(w'(r_i^2/c^*)r_i^2 + c^* w(r_i^2/c^*)\right)}{\partial \boldsymbol{\beta}} \boldsymbol{w}_i \\ &= \frac{-2}{c^*} \sum_{i=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{w}_i \left(5r_i^2 w'(\frac{r_i^2}{c^*}) + c^* w(\frac{r_i^2}{c^*}) + 2\frac{r_i^4}{c^*} w''(\frac{r_i^2}{c^*})\right) \\ &= \boldsymbol{X}_n^\top \boldsymbol{W} \boldsymbol{X}_n, \end{aligned} \tag{14}$$

where $\boldsymbol{X}_n^\top = (\boldsymbol{w}_1^\top, \cdots, \boldsymbol{w}_n^\top)$, $\boldsymbol{W}$ is a diagonal matrix with its i-th diagonal entry $-2\gamma_i/c^*$ and

$$\gamma_i = 5r_i^2 w'(\frac{r_i^2}{c^*}) + c^* w(\frac{r_i^2}{c^*}) + 2\frac{r_i^4}{c^*} w''(\frac{r_i^2}{c^*}).$$

Write γ_i/c^* as $g(t_i)$, then $g(t_i) = 5t_i w'(t_i) + 2t_i^2 w''(t_i) + w(t_i)$, where $t_i = r_i^2/c^* > c$ and $g(t) < 0$ for $t > c$ for different $c > 0$ as indicated below in Figure 3. Namely, $\boldsymbol{W}$ is positive definite when $t_i > c$.

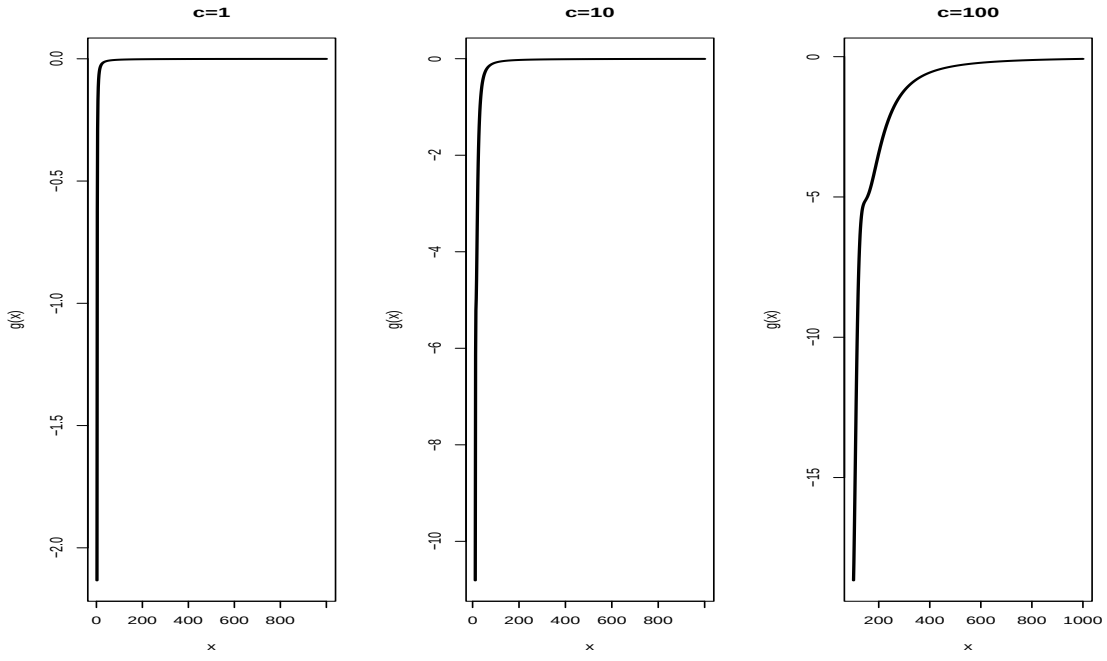

Figure 3. Behavior of function $\gamma_i(r_i)/c^*$ when $k = 5$ and $r_i > c$ with different values of c.

The algorithm for the conjugate gradient method (CGM) is as follows:

(i) **Step 1**. Pick a $\boldsymbol{\beta}^0$ (which can be an LS estimator, but for robustness, the LTS ([2]) or LST ([8]) is a better choice). Set $\boldsymbol{v}^0 = -\nabla O(\boldsymbol{\beta}^0)$. Set a tolerance ε. if $(\|v^0\| < \varepsilon)$ {return β^0}.

(ii) **Step 2**. For $k = 0, 1, \cdots, n-1$,

(a) Set $\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k + \alpha^k \boldsymbol{v}^k$, where α^k is the minimizer of $O(\boldsymbol{\beta}^k + \alpha \boldsymbol{v}^k)$ with respect to α (using backtracking line search, see page 464 of [14]), or set

$$\alpha^k = -\nabla^\top O(\boldsymbol{\beta}^k)\boldsymbol{v}^k / (\boldsymbol{v}^k)^\top H(\boldsymbol{\beta}^k)\boldsymbol{v}^k,$$

where $H(\boldsymbol{\beta}^k) = \nabla^2(O(\boldsymbol{\beta}^k))$.

(b) Compute $\nabla O(\boldsymbol{\beta}^{k+1})$, if $(\|\nabla O(\boldsymbol{\beta}^{k+1})\| < \varepsilon)$ {return $\boldsymbol{\beta}^{k+1}$}.

(c) If $(k = n-1)$ {break}; else set $v^{k+1} = -\nabla O(\boldsymbol{\beta}^{k+1}) + \alpha^k v^k$, where

$$\alpha^k = \nabla^\top O(\boldsymbol{\beta}^{k+1}) \nabla O(\boldsymbol{\beta}^{k+1}) / \nabla^\top O(\boldsymbol{\beta}^k) \nabla O(\boldsymbol{\beta}^k)$$

end for loop.

(iii) **Step 3**. Replace $\boldsymbol{\beta}^0$ by $\boldsymbol{\beta}^n$ and go to step 1.

Convergence of the gradient algorithm or gradient descent method to the global minimum has been thoroughly analyzed on pp. 466–467 of Boyd and Vandenberghe (2004) [14]. The global convergence of conjugate gradient methods specifically has been addressed in Gilbert and Nocedal (1992) [16].

5. Examples and Comparison

Now, we investigate the performance of our new procedure WLS and compare it with some leading competitors including the robust benchmark, the least trimmed squares LTS estimator, Rouseeuw [2] (known for its high robustness and fast computation), the MM estimator of Yohai [4] (known for its high robustness and high efficiency), and the classical least squares LS estimator (known for its high efficiency for i.i.d. normal errors) via some concrete examples.

5.1. Performance Criteria

Empirical mean squared error (EMSE) For a general regression estimator $\mathbf{t}$, we calculate EMSE $:= \sum_{i=1}^{R} \|\mathbf{t}_i - \boldsymbol{\beta}_0\|^2 / R$, the empirical mean squared error (EMSE) for $\mathbf{t}$. If $\mathbf{t}$ is regression equivariant, then we can assume (w.l.o.g.) that the true parameter $\boldsymbol{\beta}_0 = \mathbf{0} \in \mathbb{R}^p$ (see p.124 of [9]). Here, $\mathbf{t}_i$ is the realization of $\mathbf{t}$ obtained from the i-th sample with size n and dimension p, and replication number R is usually set to be 1000.

Total time consumed for all replications in the simulation (TT) This criterion measures the speed of a procedure, where the faster and more accurate, the better.One possible issue is the fairness of comparison of different procedures because different programming languages (e.g., C, Rcpp, Fortran, and R) are employed by different procedures.

Finite sample relative efficiency (FSRE) In the following, we investigate via simulation studies the finite-sample relative efficiency of the different robust alternatives of the LS with respect to the benchmark, the classical least squares line/hyperplane. The latter is optimal with normal models by the Gauss–Markov theorem. We generate $R = 1000$ samples from the linear regression model: $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + e_i, i \in \{1, \cdots, n\}$ with different sample size ns and dimension ps, where $e_i \sim N(0, \sigma^2)$. The finite sample RE of a procedure is the percentage of EMSE of the LS divided by the EMSE of the procedure.

All R code (downloadable via https://github.com/left-github-4-codes/WLS) accessed on 19 March 2024 for simulation, examples, and figures in this article were run on a desktop Intel(R)Core(TM) 21 i7-2600 CPU @ 3.40 GHz.

5.2. Examples

In the sequel, the cutoff value ε is set to be 10^{-4} for the procedure WLS. For simplicity, we set the tuning parameters $c = k = 6$ for the weight function.

Example 1 (Simple linear regression). *To take the advantage of the graphical illustration of data sets and plots, we start with $p = 2$, the simple linear regression case.*

We generated a data set with seven artificial highly correlated (with correlation 0.88 between x and y) bi-variate normal points. It is plotted in the left panel of Figure 4. Two reference regression lines ($y = 0$) and ($y = x$) are also provided.

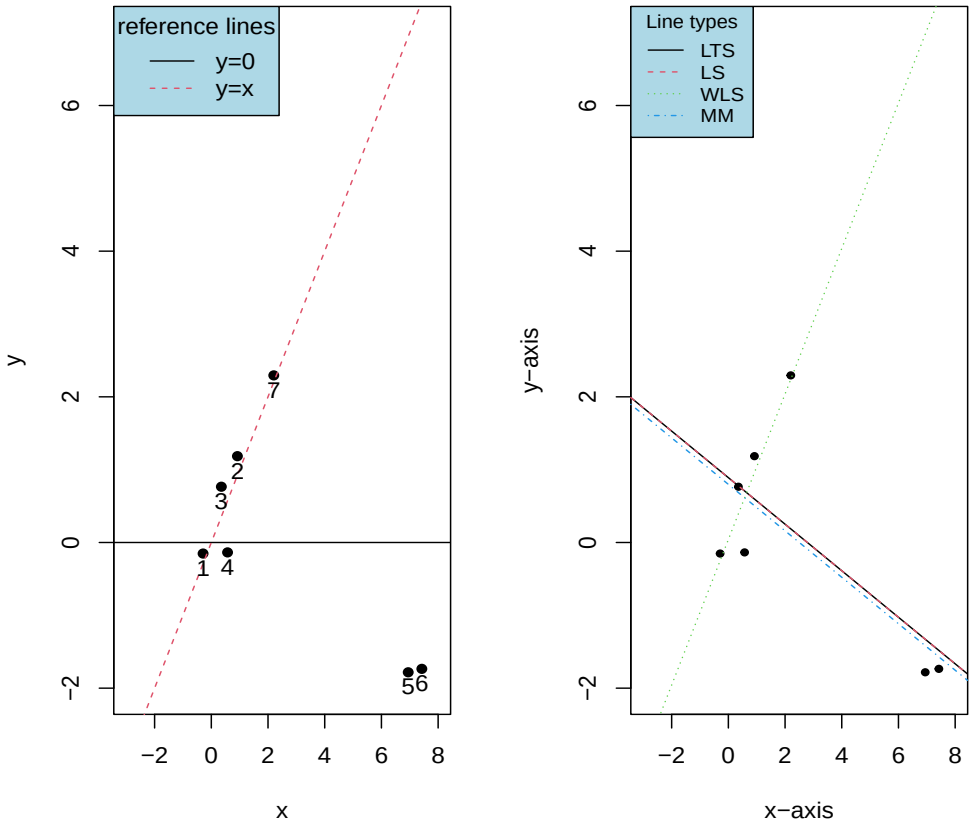

Figure 4. Left panel: plot of seven artificial points and two reference lines $y = 0$ and $y = x$. **Right panel**: the same seven points are fitted by LTS, WLS, MM, and the LS (benchmark). A solid black line is LTS given by ltsReg. Green dashed line is given by WLS. Red dotted line is given by the LS, which is identical to LTS line and is almost identical to the blue dot-dashed line given by MM in this case.

Inspecting the left panel of the figure immediately reveals that points 5 and 6 seem to be outliers and the overall pattern of the data set is linear $y = cx$ with $c > 0$. The right panel further reveals that the LS, the LTS, and the MM lines are very sensitive to the outlying points, whereas WLS still can catch the overall line pattern under the influence of two outliers.

One might immediately argue that the example above has at least two drawbacks: (i) the data set is too small, and (ii) it is purely artificial. In Figure 5, the sample size is increased to 80 highly correlated normal points with 30% of them contaminated by other normal points. Inspecting the figure reveals that the four procedures capture the linear pattern perfectly in the left panel of the figure for perfect bivariate normal points, while in the right panel, the LTS, MM, and LS lines are drastically changed due to the 24 contaminating points, while WLS well resists the influence of outliers, catching the original overall linear pattern.

In practice, there are more cases with more than one independent variable: in the following, we consider the case $p > 2$.

Example 2 (Multiple linear regression with contaminated normal points). *Now, we do not have the visual advantage like in the $p = 2$ case. To compare the performance of different procedures, we have to appeal the performance measures discussed in Section 5.1.*

We consider the contaminated highly correlated normal data points scheme. We generate 1000 samples $\{\mathbf{z}_i = (\boldsymbol{x}_i^\top, y_i)^\top, i \in \{1, \cdots, n\}\}$ with various ns from the normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a zero vector in $\mathbb{R}^p$, and $\boldsymbol{\Sigma}$ is a p by p matrix with diagonal entries being 1 and off-diagonal entries being 0.9. Then, ε% of them are contaminated by $m = \lceil n\varepsilon \rceil$ points, where $\lceil \cdot \rceil$ is the ceiling function. We randomly select m points of $\{\boldsymbol{z}_i, i \in \{1, \cdots, n\}\}$ and replace them by $(3, 3, \cdots, 3, -3)^\top$.

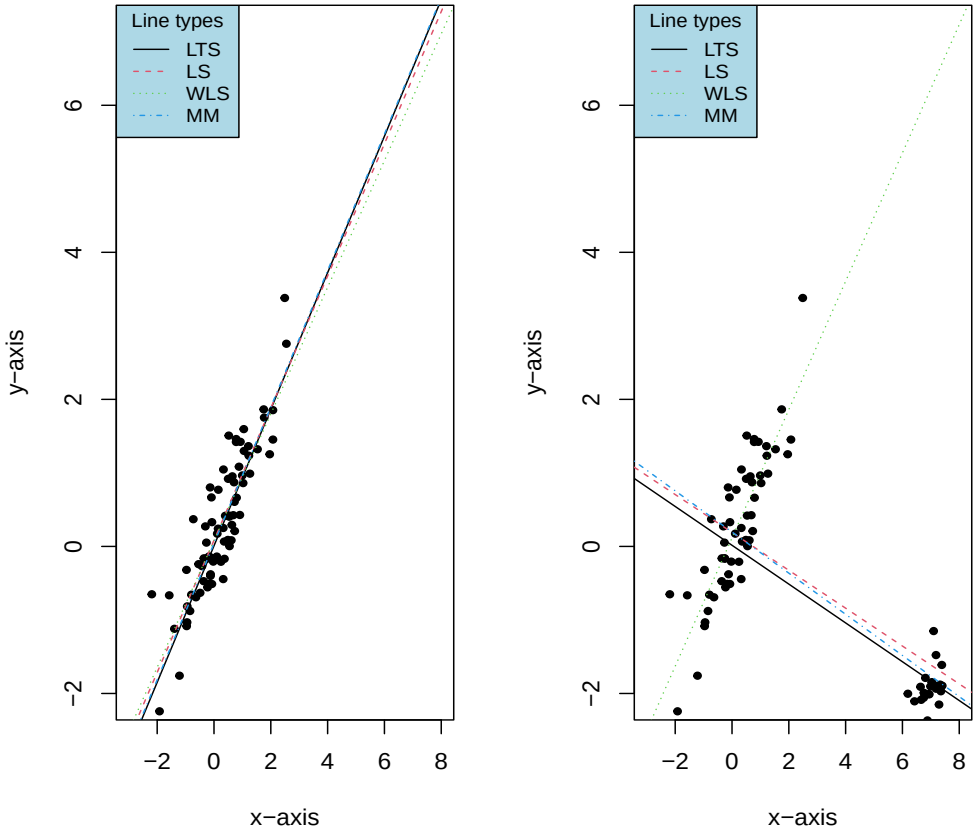

Figure 5. We show 80 highly correlated normal points with 30% of them contaminated by other normal points. Left: scatterplot of the uncontaminated perfect normal data set and four almost identical lines. Right: LTS, WLS, MM, and LS lines for the contaminated data set. The solid black is the LTS line, the dotted green is the WLS, the dot-dash blue is given by MM, and dashed red is given by the LS—parallel to LTS line in this case. The MM line is almost identical to LTS and LS lines.

The performance of the CGM in Section 4 (or rather any iterative procedure) severely depends on the initial given point $\boldsymbol{\beta}^0$. In light of its cyclic feature of the CGM for non-quadratic objective function (see page 195 of [15]) and our extensive empirical simulation experience, the performance of the $\boldsymbol{\beta}$ return by the CGM usually is not very different from (or better than) that of the initially selected $\boldsymbol{\beta}^0$. To achieve better performance for the WLS, we modified the LST of Zuo and Zuo [8] and utilized it as the initial $\boldsymbol{\beta}^0$ for CGM. Results for the three methods and different ns and ps and contamination levels ε are listed in Table 1.

Inspecting the table reveals that (i) LS is the fastest in all cases considered and the best performer for pure normal data sets, except the case $p = 20$ and $n = 200$, where WLS is even slightly more efficient. It, however, becomes the worst performer when there is contamination (except the $\varepsilon = 0.30$ cases, where the LTS and MM surprisingly become the worse performers. In theory, both MM and LTS can resist up to 50% contamination without breakdown). (ii) WLS has the smallest EMSE when there is contamination and this is true even with no contamination when $p = 20$ and $n = 200$. It is also the second fastest performer (except in the case $\varepsilon = 0.3$ and $p = 5$ or 10, where MM is faster). (iii) LTS is inferior to WLS in all cases and so is the MM (except it runs faster when $\varepsilon = 0.3$ and $p = 5$ or 10). (iv) MM performs better than LTS in TT and in EMSE (except when $p = 20$ and $\varepsilon = 0.0$, 0.10, or 0.20).

Table 1. EMSE, TT (s), and RE for MM, LTS, WLS, and LS based on all 1000 samples for various ns, ps, and contamination levels.

Normal Data Sets, Each with ε Contamination Rate								
p	**n**	**Method**	**EMSE**	**TT**	**RE**	**EMSE**	**TT**	**RE**
			$\varepsilon = 0\%$			$\varepsilon = 10\%$		
5	50	mm	0.3356	9.9427	0.9767	0.3357	9.8483	2.9876
		wls	0.3309	7.3604	0.9905	0.3324	9.4740	3.0178
		lts	0.3975	15.883	0.8246	0.3670	15.957	2.7326
		ls	0.3278	1.4243	1.0000	1.0030	1.2834	1.0000
			$\varepsilon = 20\%$			$\varepsilon = 30\%$		
5	50	mm	0.3565	9.8519	5.3673	8.4738	10.532	0.3311
		wls	0.3546	12.329	5.3951	0.3711	15.846	7.5618
		lts	0.6546	16.662	2.9228	27.223	17.026	0.1030
		ls	1.9132	1.3549	1.0000	2.8060	1.3472	1.0000
			$\varepsilon = 0\%$			$\varepsilon = 10\%$		
10	100	mm	0.2378	21.421	0.8839	0.2372	20.892	5.5816
		wls	0.2105	11.112	0.9985	0.2226	15.680	5.9499
		lts	0.2919	48.648	0.7201	0.2584	49.615	5.1245
		ls	0.2102	1.3298	1.0000	1.3242	1.2542	1.0000
			$\varepsilon = 20\%$			$\varepsilon = 30\%$		
10	100	mm	0.2410	20.669	10.244	5.1124	21.891	0.6979
		wls	0.2372	20.535	10.407	0.2600	29.146	13.724
		lts	0.2635	55.018	9.3714	40.403	64.803	0.0883
		ls	2.4691	1.2462	1.0000	3.5680	1.2626	1.0000
			$\varepsilon = 0\%$			$\varepsilon = 10\%$		
20	200	mm	0.2429	84.709	0.6564	0.2183	83.525	6.6713
		wls	0.1592	28.664	1.0021	0.1726	39.100	8.4390
		lts	0.2208	259.21	0.7224	0.2015	293.40	7.2261
		ls	0.1595	1.4936	1.0000	1.4564	1.4775	1.0000
			$\varepsilon = 20\%$			$\varepsilon = 30\%$		
20	200	mm	0.5299	78.387	5.1922	20.908	90.385	0.1899
		wls	0.1875	51.280	14.677	0.2126	71.148	18.672
		lts	0.1983	387.56	13.877	33.918	832.75	0.1170
		ls	2.7512	1.4566	1.0000	3.9694	1.4300	1.0000

Example 3 (Performance when $\boldsymbol{\beta}^0$ is given). *In the calculation of EMSE above, one assumes that $\boldsymbol{\beta}^0 = \mathbf{0}$ in light of regression equivariance of an estimator $\boldsymbol{t}$. In this example, we will provide $\boldsymbol{\beta}^0$ (for convenience, write it as $\boldsymbol{\beta}_0$) and calculate y_i using the formula $y_i = (1, \boldsymbol{x}_i^\top)\boldsymbol{\beta}_0^\top + e_i$, where we simulate $\boldsymbol{x}_i$ from a normal distribution with a zero mean vector and an identical covariance matrix. e_i follows a standard normal distribution.*

We set $p = 10$, $n = 100$ and $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, -1, -1, -1, -1, -1)$. There is a $\varepsilon = 10\%$ contamination for each of 1000 normal samples (generated as in Example 2) with the contamination scheme as follows: we randomly select $m = \lceil n\varepsilon \rceil$ points out of $\{\boldsymbol{z}_i, i \in \{1, \cdots, n\}\}$ and replace them by $(4.5, 4.5, \cdots, 4.5)^\top$. We then calculate the squared deviation (SD) $(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0)^2$ for each sample, the total time (TT) consumed by each procedures for all 1000 samples, and the relative efficiency (RE) (the ratio of EMSE of LS vs. EMSE of a procedure). The performance of three procedures for different criteria are displayed via the boxplot in Figure 6.

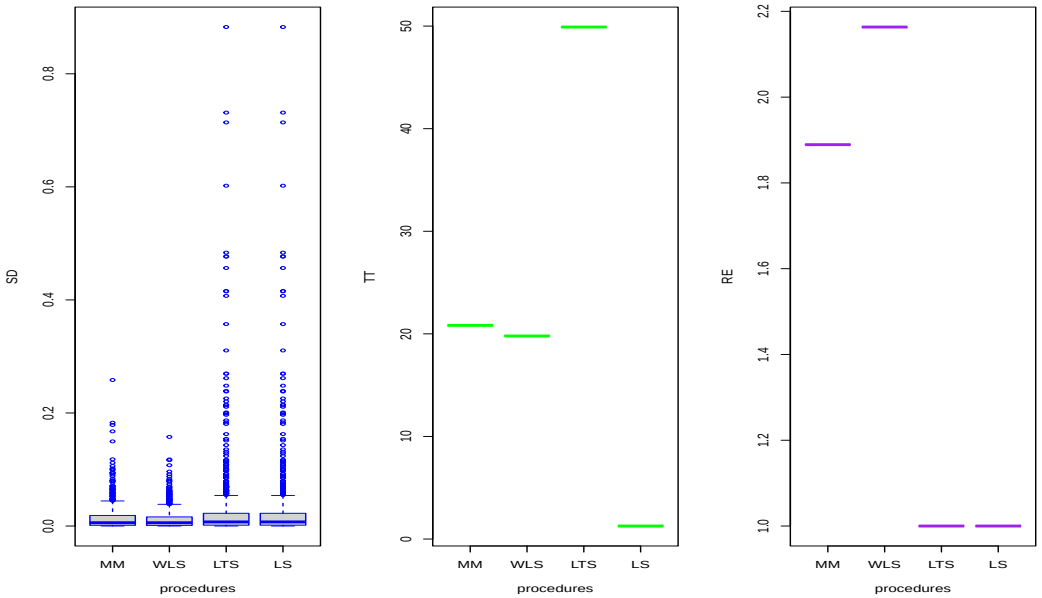

Figure 6. Performance of four procedures with respective to 1000 normal samples (points are highly correlated) with $p = 10$ and $n = 100$, each sample suffers 10% contamination.

Inspecting the figure reveals that (i) in terms of squared deviations, LTS and LS perform the same, where both have a wide spread and a high EMSE, whereas MM has a much smaller EMSE, and WLS has the smallest EMSE (in fact, the EMSE for the four (mm, wls, lts, ls) are (1.188640, 1.037962, 2.245551, 2.245551)). (ii) In terms of total time consumed, LS is the absolute winner, LTS is the absolute loser, and WLS is much better than LTS and slightly better than MM. (iii) In terms of relative efficiency, LTS is the loser (performs as bad as the LS), whereas WLS earns the trophy and is much more robust against 10% contamination. MM is the second best.

Up to this point, we have dealt with synthetic data sets. Next, we investigate the performance of MM, WLS, LTS, and LS with respect to real data sets in high dimension.

Example 4 (**Performance for a large real data set**). *Boston housing is a famous data set (see [17]) and studied by many authors with different emphasizes (transformation, quantile, nonparametric regression, etc.) in the literature. For a more detailed description of the data set, see http://lib.stat.cmu.edu/datasets/ accessed on 19 March 2024.*

The analysis reported here does not include any of the previous results but consists of just a straight linear regression of the dependent variable (median price of a house) on the thirteen explanatory variables as might be used in an initial exploratory analysis of a new data set. We have sample size $n = 506$ and dimension $p = 14$.

We assess the performance of the MM, the LST, the WLS, and the LS as follows. Since some methods depend on randomness, we run the computation $R = 1000$ times to alleviate the randomness. (i) We compute the $\widehat{\boldsymbol{\beta}}$ with different methods, and we do this 1000 times. (ii) We calculate the total time consumed (in seconds) by different methods for all replications and the EMSE (with true $\boldsymbol{\beta}_0$ being replaced by the sample mean of 1000 $\widehat{\boldsymbol{\beta}}$s from (i)), which is the sample variance of all $\widehat{\boldsymbol{\beta}}$s up to a factor 1000/999. The results are reported in Table 2.

Table 2. EMSE, TT (seconds), and RE for MM, LTS, WLS, and LS based on Boston housing real data set.

Performance Measure	MM	WLS	LTS	LS
EMSE	4.352446×10^{-5}	0.0000	4.619404×10^{1}	0.0000
TT	120.368098	161.465350	125.707603	1.487204
RE	0	NaN	0	NaN

Inspecting the table reveals that (i) WLS and LS produce the same $\widehat{\boldsymbol{\beta}}$ for each sample, so there is no variance, whereas this is not the case for MM and LTS. (ii) LS is the fastest runner followed by MM, LTS, and WLS. (iii) The relative efficiency of MM and LTS is 0% since the sample variance of LS is 0, whereas the RE of WLS and LS is undefined (not a number) since 0 appeared in the denominator. On the other hand, one can interpret WLS as being as good as LS in this case with RE 100%.

Example 5 ((Performance for a real data set which is known to contain outliers)**.** *We examine the data set of Buxton (1920) [18], which has been studied repeatedly in the literature, see Hawkins and Olive (2002) [19], Olive (2017) [20], Park, Kim, and Kim (2012) [21], Olive and Hawkins (2011) [22].*

We fit the different methods to the Buxton data, which is a 87 by 7 matrix (original row 9 was deleted), with height as the response variable and other four variables as predictor variables (two variables are excluded due the missing values) as Olive did. For more explanations , see Olive's website at http://parker.ad.siu.edu/Olive/buxton.txt accessed on 19 March 2024.

We list in Table 3 the output of the methods (mm, lts, lms wls, ls, hbreg, and rmreg2), where the last two methods are proposed by Olive and Hawkins (2011) and Olive (2017) [20,22], respectively.

Table 3. Outputs of different methods based on Buxton data set.

Methods	Intercept	Head	Nasal	Bigonal	Cephalic
hbreg	1546.3737947	−1.1288988	6.1133570	−0.5871985	1.1263726
rmreg2	807.3303643	1.7963508	4.8262483	−0.1481552	3.9353752
wls	1437.3761729	−1.1107210	5.2669763	0.9199388	0.9766958
lts	1066.188018	−1.104774	6.476802	2.523815	2.623706
lms	449.515	−1.061	7.317	6.215	4.790
mm	1511.5503972	−1.1289155	6.5942674	−0.6341536	1.2965989
ls	1546.3737947	−1.1288988	6.1133570	−0.5871985	1.1263726

With great help from Dr. Olive, we were able to have the pairwise scatter plots of points of $(\hat{y}_i, y_i)$, namely, fitted values versus observed values and fitted values versus fitted values of different methods. The plot is given in Figure 7 (lms is omitted; it performs much the same as most other robust ones).

Inspecting Figure 7 reveals that there are five obvious outliers on response variable y. Further examining the data set confirms that observations 61:65 have unusual small response values from 18 to 19, while all others are in between 1500 and 1800 and have unusual, larger head length values. The first row of Figure 7 is $(\hat{y}_i, y_i)$ for different methods. It is seen that five out of six methods perform much the same, while rmreg2 performs remarkably different.

The latter produces much larger fitted values for the five outliers which might be interpreted as the method resisting the influence of the outliers while others cope with the five outliers and produce fitted values that are in the same order in magnitude as the observed values, which might be interpreted as these methods being heavily influenced by the five outliers.

To better understand the performance of the six methods, we produced a classic fitted value versus the standardized residuals plot in Figure 8, which clearly identifies five outliers and performance difference between the six methods (rmreg2 performs remarkably different from all others).

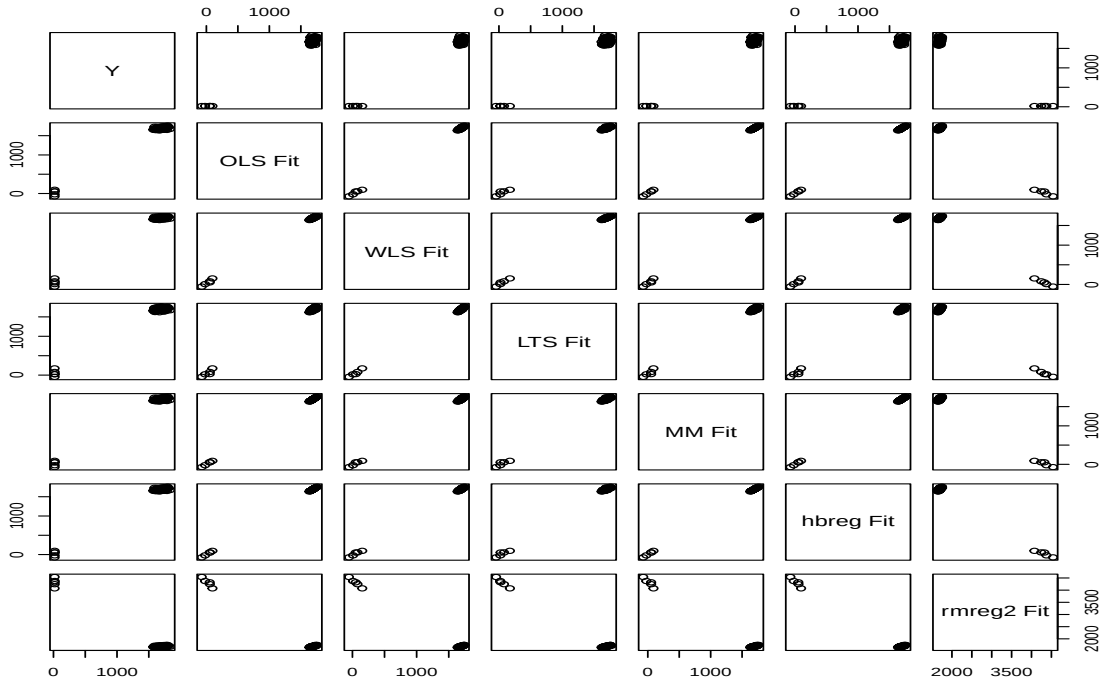

Figure 7. Pairwise plots of fitted values versus observed values and fitted values versus values for six different methods.

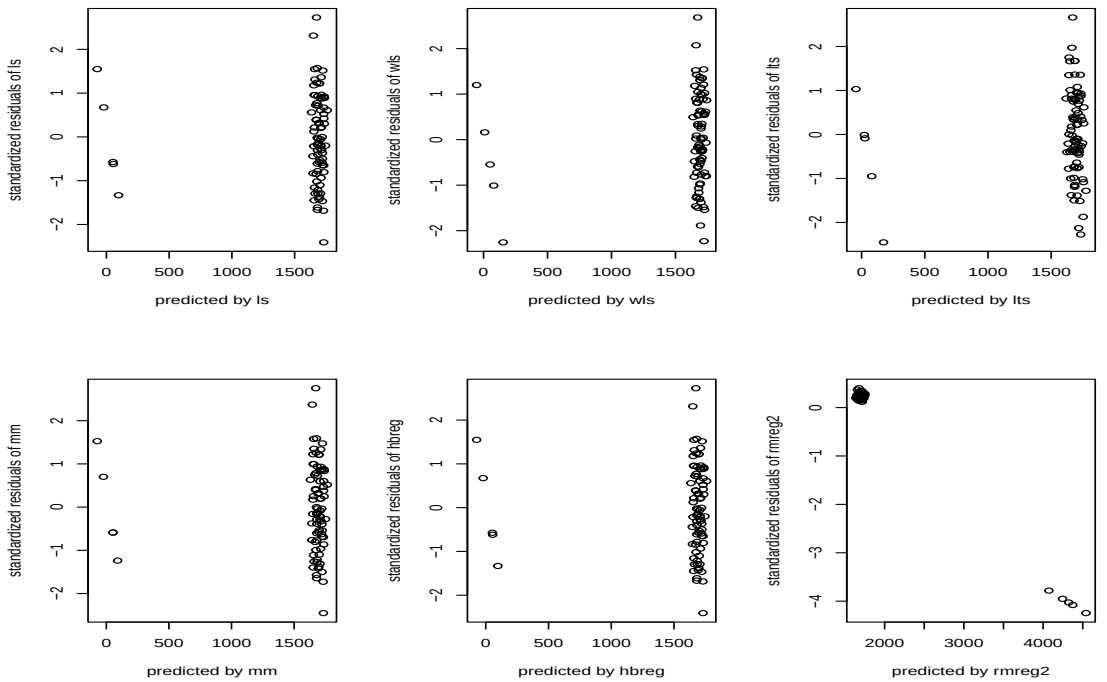

Figure 8. Fitted values versus standardized residuals plot for six different methods.

Furthermore, to better appreciate the hyperplanes induced from $\hat{\boldsymbol{\beta}}$ in Table 3 and to take the two-dimensional graphic visualization advantage, we look at the two-dimensional vertical cross-section of hyperplanes in the fifth dimension (restricted/project to y versus x_3 dimension) and plot the lines (intercept and head) based on Table 3 by different methods (they are different from the regression lines based on (x_3, y) by different methods) in Figure 9. From the Figure, we obtain a better understanding of the behavior of different methods. All seven lines but the one from rmreg2 have a negative slope.

Note that both hbreg and rereg2 functions output more than one solution. We chose hbreg$coef (which is identical to ls) and rmreg2$Bhat in this data set case. The lines from

hbreg, wls, and lts are almost parallel, while lines from mm and lms are also almost parallel to the majority but far away from the data cloud and should be discarded in this case. Similar plots with other variables could also be constructed.

Lines in Figure 9 are induced from the hyperplanes in Table 3 by projection to the (head, height)-dimension in the five-dimensional space. One naturally wonders: are they the same as the lines from direct regression on (head, height) by different methods? To appreciate the difference between two types of lines, we fit (head, height) [as (x, y)] with different methods, and the lines are given in Figure 10. Inspecting the figure reveals that all the lines perform the same but the line induced from rmreg2.

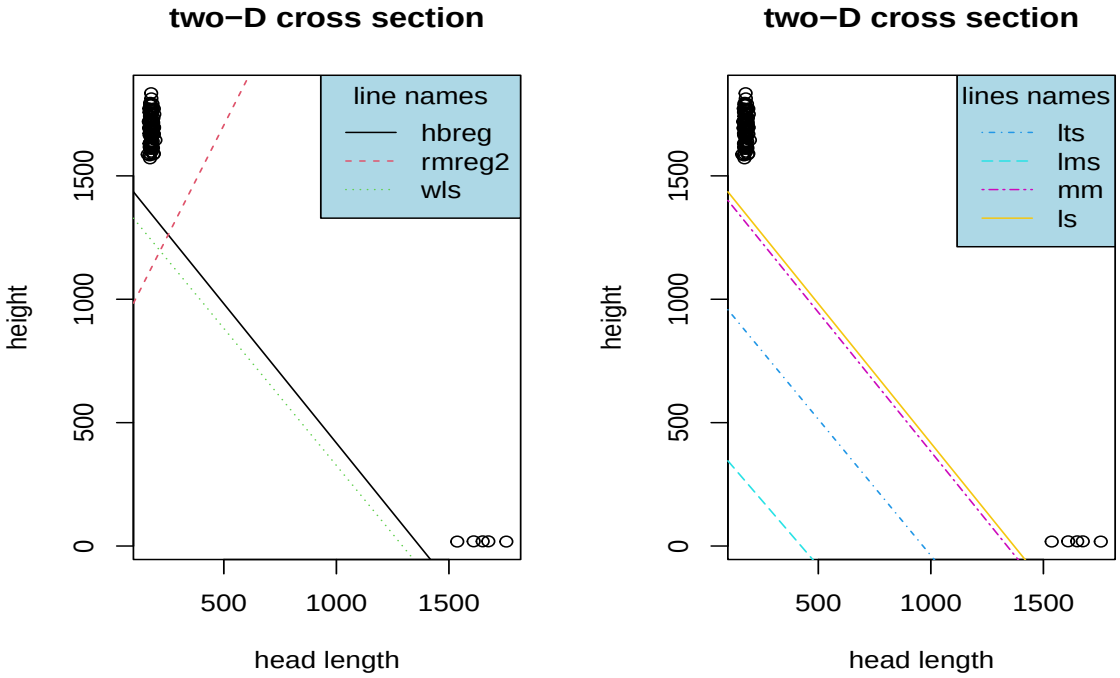

Figure 9. Restricted to (head length, height)-space, the two-dimensional vertical cross-section of hyperplances of seven different methods.

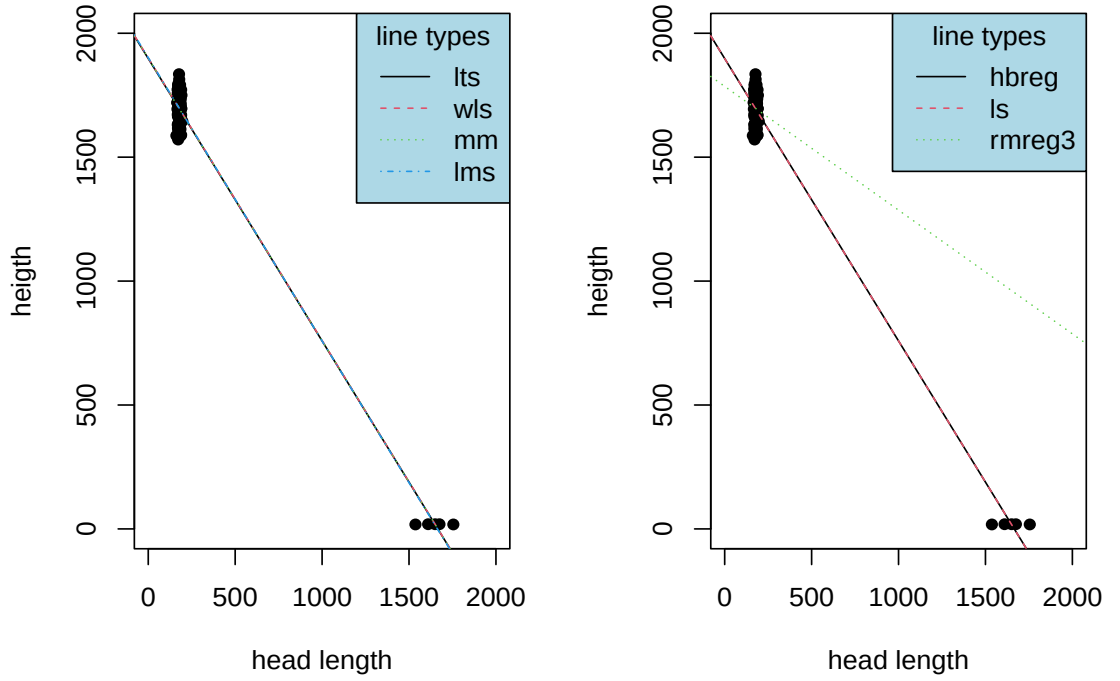

Figure 10. Regression lines based on (x = head length, y = height) by seven different methods.

6. Concluding Remarks

With a novel weighting scheme, the proposed weighted least squares estimator performs as efficiently as the classic least squares (LS) estimator for perfect normal data, while being more efficient than MM and much more efficient than the LTS estimator. It is much more robust than the LS when there is contamination or outliers (it is also more robust than MM and LTS when the contamination level is 30%). It performs as robustly as the LTS and the MM while being more efficient than MM and LTS when there are outliers. It possesses the best finite sample breakdown point robustness while achieving high efficiency and computability. It could serve as a robust alternative to the LTS and the MM in practice.

Author Contributions: Writing—original draft, Y.Z. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: Data are contained within the article

Acknowledgments: The authors thank Wei Shao for insightful comments and stimulus discussions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs of Main Results

Proof of Theorem 1. For a given $\boldsymbol{z}^{(n)}$ and any $\boldsymbol{\beta}$, write $M := \sum_{i=1}^{n} y_i^2 \geq \sum_{i=1}^{n} w(y_i^2/c^*)y_i^2 = O(\mathbf{0}_{p\times 1}, \boldsymbol{z}^{(n)})$. For a given $\boldsymbol{\beta} \in \mathbb{R}^p$, hereafter assume that $H_{\boldsymbol{\beta}}$ is the hyperplane determined by $y = \boldsymbol{w}^\top \boldsymbol{\beta}$, and let H_h be the horizontal hyperplane (i.e., $y = 0$, the $\boldsymbol{w}$-space).

Partition the space of $\boldsymbol{\beta}$s into two parts, S_1 and S_2, with S_1 containing all $\boldsymbol{\beta}$s such that $H_{\boldsymbol{\beta}}$ and H_h are parallel and S_2 consists of the rest of $\boldsymbol{\beta}$s so that $H_{\boldsymbol{\beta}}$ and H_h are not parallel.

If one can show that there are minimizers of $O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)})$ over S_i $i = 1, 2$, respectively, then one can have an overall minimizer.

Over S_1, $\boldsymbol{\beta} = (\beta_0, \mathbf{0}_{(p-1)\times 1}^\top)^\top$ and $r_i = y_i - \beta_0$. If the minimizer does not exist, then it means that any bounded β_0 cannot minimize $O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)})$, and the absolute value of the minimizer β_0^* must be greater than any $M^* > 0$. We seek a contradiction now. Denote the minimizer as $\boldsymbol{\beta}^* = (\beta_0^*, \mathbf{0}_{(p-1)\times 1}^\top)^\top$. Define $\boldsymbol{\beta}_1^* = (2\beta_0^*, \mathbf{0}_{(p-1)\times 1}^\top)^\top$, then it is readily seen that $r_i^2(\boldsymbol{\beta}_1^*) > r_i^2(\boldsymbol{\beta}^*)$ for large enough β_0^*. By lemma 1 below, one has $O(\boldsymbol{\beta}^*, \boldsymbol{z}^{(n)}) > O(\boldsymbol{\beta}_1^*, \boldsymbol{z}^{(n)})$. A contradiction is obtained.

Over S_2, denote by $l_{\boldsymbol{\beta}}$ the intersection part of $H_{\boldsymbol{\beta}}$ with the horizontal hyperplane H_h (we call it a hyperline, though it is $p-1$-dimensional). Let $\theta_{\boldsymbol{\beta}} \in (-\pi/2, \pi/2)$ be the acute angle between the $H_{\boldsymbol{\beta}}$ and H_h (and $\theta_{\boldsymbol{\beta}} \neq 0$). Consider two cases.

Case I. All $\boldsymbol{w}_i = (1, \boldsymbol{x}_i^\top)^\top$ with $r_i^2/c^* \leq c$ on the hyperline $l_{\boldsymbol{\beta}}$, where $r_i = y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta}$. Then, we have a vertical hyperplane that is perpendicular to the horizontal hyperplane H_h ($y = 0$) and intersect H_h at $l_{\boldsymbol{\beta}}$, But this contradicts **A1**. We only need to consider the other case.

Case II. Otherwise, define

$$\delta = \frac{1}{2} \inf\{\tau, \text{such that } N(l_{\boldsymbol{\beta}}, \tau) \text{ contains all } \boldsymbol{w}_i \text{ with } r_i^2/c^* \leq c\},$$

where $N(l_{\boldsymbol{\beta}}, \tau)$ is the set of points in $\boldsymbol{w}$-space such that each distance to the $l_{\boldsymbol{\beta}}$ is no greater than τ. Clearly, $0 < \delta < \infty$ (since $\delta = 0$ has been covered in **Case I** and $\delta \leq \max_i\{\|\boldsymbol{w}_i\|\} < \infty$, where i satisfies $r_i^2/c^* \leq c$, and the first inequality follows from the fact that the hypotenuse is always longer than any legs).

We now show that when $\|\boldsymbol{\beta}\| > (1+\eta)\sqrt{M}/\delta$, where $\eta > 1$ is a fixed number, then

$$O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)}) = \sum_{i=1}^{n} w(r_i^2/c^*) r_i^2(\boldsymbol{\beta}) > M \geq O(\mathbf{0}_{p\times 1}, \boldsymbol{z}^{(n)}). \tag{A1}$$

That is, for the solution of the minimization of (5). One only needs to search over the ball $\|\boldsymbol{\beta}\| \leq (1+\eta)\sqrt{M}/\delta$, a compact set. Note that $O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)})$ is continuous in $\boldsymbol{\beta}$ since $r_i(\boldsymbol{\beta})$ and $w(r_i^2/c^*)$ are. Then, the minimization problem certainly has a solution over the compact set.

The proof is complete if we can show (A1) when $\|\boldsymbol{\beta}\| > (1+\eta)\sqrt{M}/\delta$. It is not difficult to see that there is at least one i_0 such that $r_{i_0}^2/c^* \leq c$ and $\boldsymbol{w}_{i_0} \notin N(l_{\boldsymbol{\beta}}, \delta)$ since otherwise, it contradicts the definition of δ above. Note that $\theta_{\boldsymbol{\beta}}$ is the angle between the normal vectors $(-\boldsymbol{\beta}^\top, 1)^\top$ and $(\mathbf{0}^\top, 1)^\top$ of hyperplanes $H_{\boldsymbol{\beta}}$ and H_h, respectively. Then, $|\tan\theta_{\boldsymbol{\beta}}| = \|\boldsymbol{\beta}\|$ (see [8]) and (see Figure A1)

$$|\boldsymbol{w}_{i_0}^\top \boldsymbol{\beta}| > \delta |\tan\theta_{\boldsymbol{\beta}}| = \delta\|\boldsymbol{\beta}\| > (1+\eta)\sqrt{M}.$$

Now, we have

$$|r_{i_0}(\boldsymbol{\beta})| = |\boldsymbol{w}_{i_0}^\top \boldsymbol{\beta} - y_{i_0}| \geq \big||\boldsymbol{w}_{i_0}^\top \boldsymbol{\beta}| - |y_{i_0}|\big| > (1+\eta)\sqrt{M} - |y_{i_0}|. \tag{A2}$$

Therefore,

$$\begin{aligned} O(\boldsymbol{\beta}, \boldsymbol{z}^{(n)}) = \sum_{j=1}^{n} w(r_j^2/c^*) r_j^2(\boldsymbol{\beta}) &\geq w(r_{i_0}^2/c^*) r_{i_0}^2(\boldsymbol{\beta}) = r_{i_0}^2(\boldsymbol{\beta}) \\ &> \Big((1+\eta)\sqrt{M} - |y_{i_0}|\Big)^2 \geq \Big((1+\eta)\sqrt{M} - \sqrt{M}\Big)^2 \\ &= \eta^2 M > M \geq O(\mathbf{0}_{p\times 1}, \boldsymbol{z}^{(n)}). \end{aligned}$$

That is, we have certified (A1). □

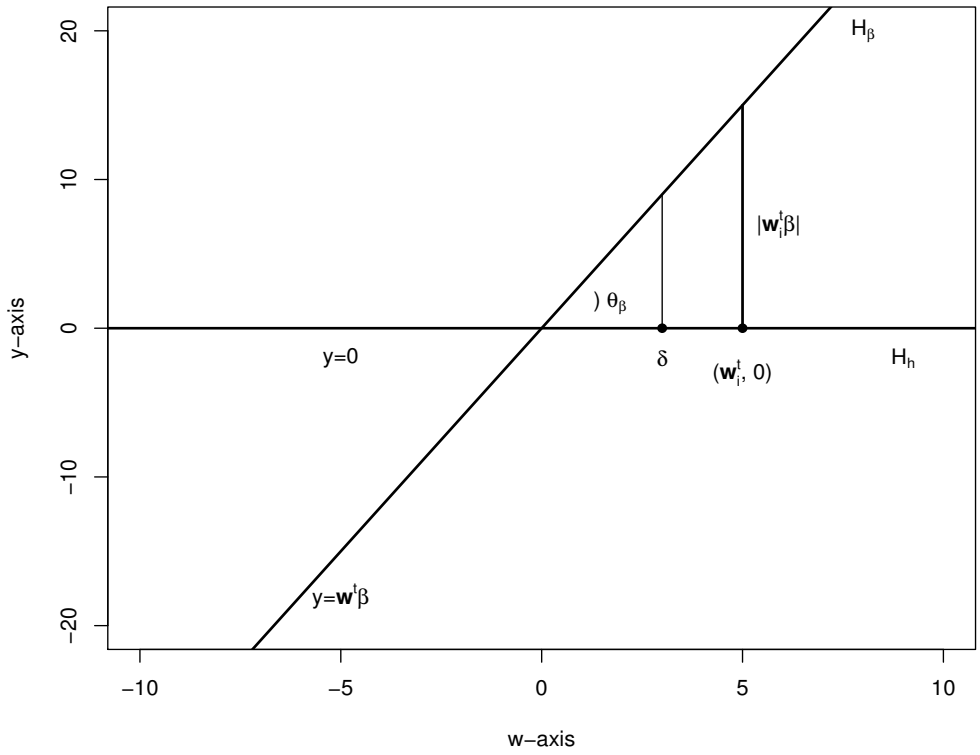

Figure A1. A two-dimensional vertical cross-section (that goes through points $(\boldsymbol{w}_i^t, 0)$ and $(\boldsymbol{w}_i^t, \boldsymbol{w}_i^t\boldsymbol{\beta})$) of a figure in $\mathbb{R}^p$ ($\boldsymbol{w}_i^t = \boldsymbol{w}_i^\top$). Hyperplanes H_h and $H_{\boldsymbol{\beta}}$ intersect at hyperline $l_{\boldsymbol{\beta}}$ (which does not necessarily pass through $(\mathbf{0}, 0)$, here just for illustration). The vertical distance from point $(\boldsymbol{w}_i^t, 0)$ to the hyperplane $H_{\boldsymbol{\beta}}$, $|\boldsymbol{w}_i^t\boldsymbol{\beta}|$, is greater than $\delta|\tan(\theta_{\boldsymbol{\beta}})|$.

Proof of Lemma 1. Write $w(r^2/c^*)r^2 = c^* w(r^2/c^*)r^2/c^* := c^* w(x^2)x^2$, where $x = |r|/\sqrt{c^*}$ and $x^2 = r^2/c^* > c$. It suffices to show that $w(x^2)x^2$ is strictly decreasing in x (this intuitively is clear from Figure 2), or equivalently, to show that the derivative of $w(x^2)x^2$ is negative. A straightforward calculus derivation yields

$$\Big(w(x^2)x^2\Big)' = 2x/(1-e^{-k})\Big(e^{-k(1-c/x^2)^2}\big(1-2kc/x^2(1-c/x^2)\big) - e^{-k}\Big).$$

Now it suffices to show that

$$\Big(e^{-k(1-c/x^2)^2}\big(1-2kc/x^2(1-c/x^2)\big) - e^{-k}\Big) < 0.$$

Or, equivalently, it suffices to show that

$$e^{k\left((1-c/x^2)^2-1\right)} > 1-2kc/x^2(1-c/x^2).$$

For convenience, write $t := c/x^2$. Then, $t \to 0$ as $x^2 \to \infty$. Now, we want to show that

$$e^{-tk(2-t)} > 1-2kt(1-t). \tag{A3}$$

A straightforward Taylor expansion of $e^x = 1 + x + x^2/2! + x^3/3! + \cdots$ to the left-hand side (LHS) of (A3) yields

$$\begin{aligned}
e^{-tk(2-t)} &= 1 + (-2kt+kt^2) + (-2kt+kt^2)^2/2 + (-2kt+kt^2)^3/3! + (-2kt+kt^2)^4/4! + \cdots \\
&> 1 + (-2kt+kt^2) + (-2kt+kt^2)^2/2 + (-2kt+kt^2)^3/3! \\
&= 1 - 2kt(1-t) - kt^2 + (-kt(2-t))^2/6 + 2(-kt(2-t))^2/6 + (-kt(2-t))^3/6 \\
&= 1 - 2kt(1-t) + kt^2\Big(k(2-t)^2/6 - 1\Big) + k^2t^2(2-t)^2\big(2-kt(2-t)\big)/6 \\
&> 1 - 2kt(1-t)
\end{aligned} \tag{A4}$$

where the first inequality follows from the fact that

$$\frac{(-kt(2-t))^{2n}}{(2n)!} + \frac{(-kt(2-t))^{2n+1}}{(2n+1)!} = \frac{(-kt(2-t))^{2n}(2n+1-kt(2-t))}{(2n+1)!} > 0,$$

for $n \geq 2$ and small enough t. And the last inequality in (A4) follows the facts (i) $k(2-t)^2/6 - 1 > 0$ (if $t < 2 - \sqrt{6/k}$) and (ii) $2 - kt(2-t) > 0$ (if $t < 1 - \sqrt{1-2/k}$).

Combining (A4) with (A3), we complete the proof. □

Proof of Theorem 3. It suffices to treat the case $p > 1$, and furthermore by Theorem 4 on p. 125 of [9], it is sufficient to show that $m = \lfloor (n-p)/2 \rfloor$ contaminating points are not enough to break drown $\widehat{\boldsymbol{\beta}}_{wls}$. Assume it is otherwise. This implies that either

(I) $|\widehat{\boldsymbol{\beta}}^n_{wls}((\mathbf{z}^{(n)}_m)_j)_1| \to \infty$ and $\|\widehat{\boldsymbol{\beta}}^n_{wls}((\mathbf{z}^{(n)}_m)_j)_2\|$ is finite, or

(II) $\|\widehat{\boldsymbol{\beta}}^n_{wls}(\mathbf{z}^{(n)}_m)_j)_2\| = \Big|\tan\Big(\theta_{\widehat{\boldsymbol{\beta}}^n_{wls}(\mathbf{z}^{(n)}_m)_j)}\Big)\Big| \to \infty$,

along a sequence of $(\mathbf{z}^{(n)}_m)_j$ as $j \to \infty$, where the subscripts 1 and 2 correspond to the intercept and non-intercept terms, respectively, as in the case $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^\top)^\top$ in $\mathbb{R}^p$. We seek a contradiction for both cases. For description simplicity, write $\boldsymbol{\beta}_j := \widehat{\boldsymbol{\beta}}^n_{wls}((\mathbf{z}^{(n)}_m)_j)$

Case **(I)**. For simplicity, write $\boldsymbol{\beta}_j = (\beta_1, \boldsymbol{\beta}_2^\top)^\top$ and $\boldsymbol{\beta}_{jj} = (2^m\beta_1, \boldsymbol{\beta}_2^\top)^\top$. Then, it is readily seen that $r_i^2(\boldsymbol{\beta}_j) < r_i^2(\boldsymbol{\beta}_{jj})$ for large positive m. In light of Lemma 1, one has that $O(\boldsymbol{\beta}_j) > O(\boldsymbol{\beta}_{jj})$; a contradiction is obtained.

Case **(II)**. This case implies there is a sequence of hyperplanes induced from $\widehat{\boldsymbol{\beta}}^n_{wls}((\mathbf{z}^{(n)}_m)_j)$ that tend to the eventual vertical position as $j \to \infty$. Denote by H_j those hyperplanes. Let H_j intercept with the horizontal hyperplane H_h at ℓ_j, the hyperlines (or the common part of H_j and H_h).

For simplicity, write the minimizer $\boldsymbol{\beta}_j = (\beta_1, \boldsymbol{\beta}_2^\top)^\top := \widehat{\boldsymbol{\beta}}^n_{wls}((\mathbf{z}^{(n)}_m)_j)$. Introduce a new hyperplane determined by $\boldsymbol{\beta}_{jj} = (\alpha\beta_1, \kappa\boldsymbol{\beta}_2^\top)^\top$ ($\kappa > 1$ is a positive integer). This $\boldsymbol{\beta}_{jj}$ amounts to tilting H_j (corresponding to $\boldsymbol{\beta}_j$) along ℓ_j to a more vertical position H_{jj} (corresponding to $\boldsymbol{\beta}_{jj}$). Note that it is possible that there are no data points touched during the titling process except those originally on the H_j since both hyperplanes are almost vertical. It is readily seen that $r_i^2(\boldsymbol{\beta}_{jj}) > r_i^2(\boldsymbol{\beta}_j) \to \infty$ except those points $(\boldsymbol{x}_i^\top, y_i))^\top$ that originally lie on the ℓ_j with a zero residual. By Lemma 1, $O(\boldsymbol{\beta}_j) > O(\boldsymbol{\beta}_{jj})$, a contradiction is reached. □

References

1. Huber, P.J. Robust estimation of a location parameter. *Ann. Math. Statist.* **1964**, *35*, 73–101. [CrossRef]
2. Rousseeuw, P.J. Least median of squares regression. *J. Amer. Statist. Assoc.* **1984**, *79*, 871–880. [CrossRef]
3. Rousseeuw, P.J.; Yohai, V.J. Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*; Lecture Notes in Statist; Springer: New York, NY, USA, 1984; Volume 26, pp. 256–272.
4. Yohai, V.J. High breakdown-point and high efficiency estimates for regression. *Ann. Statist.* **1987**, *15*, 642–656. [CrossRef]
5. Yohai, V.J.; Zamar, R.H. High breakdown estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.* **1988**, *83*, 406–413. [CrossRef]
6. Rousseeuw, P.J.; Hubert, M. Regression depth (with discussion). *J. Am. Statist. Assoc.* **1999**, *94*, 388–433. [CrossRef]
7. Zuo, Y. On general notions of depth for regression. *Stat. Sci.* **2021**, *36*, 142–157. [CrossRef]
8. Zuo, Y.; Zuo, H. Least sum of squares of trimmed residuals regression. *Electron. J. Stat.* **2023**, *17*, 2416–2446. [CrossRef]
9. Rousseeuw, P.J.; Leroy, A. *Robust Regression and Outlier Detection*; Wiley: New York, NY, USA, 1987.
10. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
11. Müller, C. Redescending M-estimators in regression analysis, cluster analysis and image analysis. *Discuss. Math. Stat.* **2004**, *24*, 59–75. [CrossRef]
12. Zuo, Y. Projection-based depth functions and associated medians. *Ann. Statist.* **2003**, *31*, 1460–1490. [CrossRef]
13. Donoho, D.L.; Huber, P.J. The notion of breakdown point. In *A Festschrift foe Erich L. Lehmann*; Bickel, P.J., Doksum, K.A., Hodges, J.L., Jr., Eds.; Wadsworth: Belmont, CA, USA, 1983; pp. 157–184.
14. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
15. Edgar, T.F.; Himmelblau, D.M.; Lasdon, L.S. *Optimization of Chemical Processes*, 2nd ed.; McGraw-Hill Chemical Engineering Series; McGraw-Hill: New York, NY, USA, 2001.
16. Gilbert, J.C.; Nocedal, J. Global Convergence Properties of Conjugate Gradient Methods for Optimization. *Siam J. Optim.* **1992**, *2*, 21–42. [CrossRef]
17. Harrison, D.; Rubinfeld, D.L. Hedonic prices and the demand for clean air. *J. Environ. Econ. Manag.* **1987**, *5*, 81–102. [CrossRef]
18. Buxton, L.H.D. The Anthropology of Cyprus. *J. R. Inst. Great Br. Irel.* **1920**, *50*, 183–235. [CrossRef]
19. Hawkins, D.M.; Olive, D.J. Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm, (with discussion). *J. Am. Stat. Assoc.* **2002**, *97*, 136–159. [CrossRef]
20. Olive, D.J. *Robust Multivariate Analysis*; Springer: New York, NY, USA, 2017.
21. Park, Y.; Kim, D.; Kim, S. Robust Regression Using Data Partitioning and M-Estimation. *Commun. Stat. Simul. Comput.* **2012**, *8*, 1282–1300. [CrossRef]
22. Olive, D.J.; Hawkins, D.M. Practical High Breakdown Regression. 2011. Available online: http://www.math.siu.edu/olive/pphbreg.pdf (accessed on 19 March 2024).