S-estimators for Robust Linear Regression

Glenn McGuinness

December 11, 2019

1 Background

1.1 Introduction

All statistical methods make assumptions about the structure of the data. These assumptions are intended to be good approximations of the real world. However, such assumptions are commonly violated in small and large ways and we would still like "good" performance. Many statistical methods are only commonly evaluated with a narrow class of "well-behaved" distributions, such as normal distributions. Unsurprisingly, many classical methods fail, sometimes spectacularly, when even a small portion of data follows a significantly different distribution.

Robust statistics is the study of statistical methods that perform well for data following a broader class of "bad" distributions than those considered in classical statistics. Data are frequently normally distributed about the centre, but have heavy tails. This type of data frequently gives rise to *outliers*. Outliers are atypical observations that lie far from the bulk of the data. A small proportion of such data can lead classical methods to have significantly different estimates from the same data with the outliers removed.

Outliers are particularly insidious for regression. Problematic outliers can often be managed with simple method in univariate data. For example, trimming off some upper and lower quartile of the data. However, in regression datasets, outliers may not appear atypical in any individual variable, but deviate from the predominant trend in the data. As the number of response variables grows, such outliers quickly become very difficult to identify with simple rules. Robust regression methods offer reliable performance when outliers can not be easily detected, avoiding the pitfalls inherent in working with data of unknown quality.

A common model for representing data with outliers, or contaminated data, is the Tukey-Huber contamination model [Alq+09]. In the Tukey-Huber contamination model, we assume that the data follow some "good" parametric distribution F_θ with probability $1-\epsilon$ and some "outlier" distribution G with probability ϵ. We consider the performance of our estimator $\hat{\hat{\theta}}$ in the context of a *contamination neighbourhood*

$$\mathcal{F}(F,\epsilon) = \{(1-\epsilon)F + \epsilon G : G \in \mathcal{G}\}. \tag{1}$$

For simplicity, $\mathcal{G}$ is frequently assumed to be the set of point mass distributions.

This report provides an overview of the robustness properties of *S-estimators for linear regression* [RY84], for which it is necessary to also review M-estimates of scale [Mar+19]. S-estimators are highly robust with a number of desirable properties. They are most frequently used as the initial estimate for MM-estimators [Yoh+87], a popular class robust regression estimators. First, we will present M-estimators of scale, which are a central element of S-estimators and their breakdown point and asymptotic efficiency, measures of robustness and performance, respectively. Second, we will do the same for S-estimators for linear regression. Finally, we will propose a new S-estimator as a possible new research direction.

1.2 Robustness Properties

So far, we have discussed robustness in a vague sense. Intuitively, a robust estimator must perform well when data deviates from a "good" distribution and not perform significantly worse than a classical estimator on "good" data. Common measures of robustness are developed to quantify this intuition. This section will present the formal definitions of properties that are commonly considered for robust estimator. In particular, this section will present the *breakdown point* to measure robustness and *efficiency* to measure performance.

We assume the distribution of the data lies in the contamination neighbourhood described in equation (1). The breakdown point (BP) can be defined asymptotically and for finite sample.

Definition 1.1. *Consider the parametric distribution family F_θ with parameter $\theta \in \Theta$. The* ***asymptotic contamination breakdown point*** *of estimator $\hat{\theta}$ at F, denoted by $\epsilon^*(\hat{\theta}, F)$ is the largest $\epsilon^* \in (0,1)$ such that for all $\epsilon < \epsilon^*$,*

$$\lim_{n\to\infty} \hat{\theta}((1-\epsilon)F + \epsilon F)$$

does not lie on the boundary of the parameter space Θ for all $G \in \mathcal{G}$. [Mar+19]

Unfortunately, it takes a very long time to collect an infinite number of observations. The *finite sample breakdown point* (FBP) may therefore be more useful in practice.

Definition 1.2. *Consider the parametric distribution family F_θ with parameter space Θ. Let $\partial\Theta$ denote the boundary of the parameter space. Let $\mathcal{X}_m$ be the set of all datasets of size n that have $n-m$ elements in common with x. The* ***replacement finite sample contamination breakdown point*** *of estimator $\hat{\theta}_n$ at x is*

$$\epsilon_n^*(\hat{\theta}_n, x) = \frac{m^*}{n},$$

where m^* is the largest integer such that $\hat{\theta}_n(x)$ is bounded and does not lie on the boundary of Θ. [Mar+19]

However, a high breakdown point is not sufficient for an estimator to be useful. For example, an estimator of $\hat{\theta} = 5$ has an asymptotic and finite sample breakdown point of $\epsilon^* = 1$. However, I expect you would not be able to convince a colleague that an always-5 estimator should be used to estimate the mean of a sample. Some measure of the variance of an estimator is also required.

Definition 1.3. *Consider the parametric distribution family F_Θ. Let $\hat{\theta}$ be a consistent estimator and $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \nu(\theta,\hat{}theta))$. Let $\hat{\Theta}$ be some class of estimators of θ. Let*

$$\nu_{min}(\theta) = \min\{\nu(\hat{\theta}, \theta) : \hat{\theta} \in \hat{\Theta}\}$$

. The ***asymptotic efficiency*** *of $\hat{\theta}$ at θ is defined as*

$$p_{eff}(\theta) = \frac{\nu_{min}(\theta)}{\nu(\theta)}.$$

Where $p_{eff}(\theta) \in (0, 1)$. [Mar+19]

Considered together, the BP and the asymptotic efficiency are useful properties when evaluating the robustness and performance of estimators.

2 M-estimates of Scale

Before we can consider S-estimators for linear regression, we must consider M-estimates of scale which are the critical element. In classical statistics, the degree of variation of data is usually discussed in terms of the sample variance. This is reasonable as the sample variance is the MLE of the variance of iid normally distributed data. However, if even a single observation is severely outlying, the sample variance can be wildly inaccurate. Instead, we may wish to use a different measure variation. Consider the multiplicative model

$$x_i = \sigma u_i \tag{2}$$

where the random variables u_i are independent and identically distributed with some distribution F. Let the distribution F have density f_0 and $\sigma \in \mathbb{R}_+$ be an unknown *scale* parameter.The distributions of x_i are in a scale family, with a density

$$\frac{1}{\sigma} f_0\left(\frac{x}{\sigma}\right).$$

To estimate σ, it is reasonable to first consider the maximum likelihood estimator (MLE). The MLE can be calculated by finding the value of $\hat{\sigma}$ that maximizes the log-likelihood. The first order conditions of the maximization of the log-likelihood are

$$\sum_{i=1}^{n} \rho(x_i/\hat{\sigma}) = 1 \tag{3}$$

where $\rho(x) = -x f_0'(x)/f_0(x)$. If the data are normally distributed, $\rho(t) = t^2$, giving

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

which is the sample standard deviation which we all know and love. However, notice that the size of $\rho(x_i/\hat{\sigma})$ grows rapidly the further the observation is from the bulk of the data. In fact, if even a single observation $x_i \to \infty$ then $\hat{\sigma}_{MLE} \to \infty$ so the sample variance has a BDP of zero. This results in outliers having an oversized impact on the scale estimate. Instead, it may be desirable to consider alternative estimator that does not allow outlying points to have such a large influence.

One might consider the class of estimators where the ρ function from (3) is not determined by the assumed distribution of the data. This is the approach taken by *M-estimators of scale*, where M stands for "Maximum likelihood type". An M-estimator of scale $\hat{\sigma}$ is any estimator satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \rho(x_i/\hat{\sigma}) = \delta, \quad 0 < \delta \leq 1. \tag{5}$$

Notice that the law of large numbers implies that $\frac{1}{n}\sum_{i=1}^{n} \rho(x_i/\hat{\sigma}) \to E(\rho(x_i/\hat{\sigma}))$ almost surely as $n \to \infty$. When choosing a ρ function, we wish to have consistency and good efficiency in addition to a high breakdown point. This can be done by choosing a bounded ρ function that is close to t^2 near the origin, but grows more slowly further away. One such ρ function is that corresponding to Tukey's biweight function, given by

$$\rho'(t) = \psi(t) = \begin{cases} t(1 - t^2/c^2)^2; & |t| < c \\ 0 & |t| \geq c \end{cases}. \tag{6}$$

We will also want to choose a ρ function such that σ converges to the standard deviation under normality. For Tukey's biweight function, this corresponds to $c = 1.56$ [Mar+19]. In this report we will only consider ρ functions with the following properties.

Definition 2.1. *We will consider ρ-functions that have the following properties:*

(R1) *$\rho(t)$ is continuously differentiable, symmetric, and non-decreasing in $|t|$*

(R2) *$\rho(0) = 0$*

(R3) $\rho(t)$ is bounded.

Continuous differentiability is often not required, but makes later proofs much simpler. Notably, $\rho(t) = |t|$ is not differentiable at zero.

2.1 Existence and Uniqueness

First, we will consider the existence and uniqueness of solutions of M-estimates of scale. For convenience, rewrite (5)

$$g(\sigma) = \sum_{i=1}^{n} \Psi(x_i, \sigma) = 0 \text{ where } \Psi(x, \sigma) = \rho(x/\sigma) - \delta. \tag{7}$$

The following theorem states that the M-estimate of scale under the regularity conditions exists and is unique.

Theorem 2.1. *Under the regularity conditions (R1)-(R3), we have*
(i) there exists at least one point $\hat{\sigma}$ where $g(\sigma) \geq 0$ for $\sigma > \hat{\sigma}$ and $g(\sigma) \leq 0$ for $\sigma < \hat{\sigma}$,

(ii) the set of such points is an interval,

(iii) $g(\hat{\sigma}) = 0$, and

(iv) $\hat{\theta}$ is unique.

Proof. This proof is based on that in [Mar+19], adapted for the regularity conditions (R1)-(R3). It is clear from (7) that

$$\lim_{\sigma \to 0} \Psi(x, \sigma) = -\delta < 0 < 1 - \delta = \lim_{\sigma \to \infty} \Psi(x, \sigma)$$

The monotonicity of $g(\sigma)$ implies (i) and that all values of σ for which $g(\sigma) = 0$ must be in an interval. If $g(\sigma) \neq 0$ for any σ, then the monotonicity of g implies that only one point exists satisfying (i), giving (ii). The continuity of $\rho(t)$ implies that $g(\sigma)$ is continuous in σ. As $g(\sigma)$ is continuous in σ, intermediate value theorem implies (iii). From (R1), $\rho(x/\sigma)$ is non-decreasing in σ, which implies (iv). □

2.2 Consistency

Let $x_1, \ldots, x_n$ be i.i.d with the parametric distribution F_{σ_F}. We will wish to show that $\hat{\sigma}_n$ converges in probability to σ_F. For convenience, define

$$\lambda_F(\sigma) = E_F \Psi(x, \sigma), \quad \hat{\lambda}_n(\sigma) = \frac{1}{n} \sum_{i=1}^{n} \Psi(x_i, \sigma). \tag{8}$$

Theorem 2.2. *There exists a σ_F such that $\lambda_F(\sigma_F) = 0$.*

Proof. This proof is based on that in [Mar+19], adapted to be specific to M-estimates of scale. As Ψ is bounded, Bounded Convergence Theorem implies,

$$\lim_{\sigma \to 0} E_F \Psi(x, \sigma) = E_F \lim_{\sigma \to 0} \Psi(x, \sigma) = -\delta < 0 < 1 - \delta = \lim_{\sigma \to 0} E_F \Psi(x, \sigma)$$

λ_F is continuous and monotonic in σ so, similar to the proof for Theorem 2.1, intermediate value theorem implies that there exists a σ_F such that $\lambda_F(\sigma_F) = 0$. □

Theorem 2.3. *If σ_F is unique, then $\hat{\sigma}_n \xrightarrow{p} \sigma_F$.*

Proof. This proof is based on that in [Mar+19], adapted to be specific to M-estimates of scale. By definition, $\lambda_F = E_F \Psi(x, \sigma)$, so the Law of Large Numbers implies that

$$\hat{\lambda}_n(\sigma_F - \epsilon) \xrightarrow{p} \lambda_F(\sigma_F - \epsilon) \text{ as } n \to \infty$$

using the same argument as the proof in Theorem 2.1, $\hat{\lambda}_n$ is non-increasing. Theorem 2.1 implies that σ_F is unique. Thus $\hat{\sigma}_n < \sigma_F - \epsilon$ for $\epsilon > 0$ implies that $\hat{\lambda}_n(\sigma_F - \epsilon) < 0$. Also, λ_F is decreasing and $\lambda_F(\sigma_F) = 0$ so $\lambda_F(\sigma_F - \epsilon) > 0$. Hence,

$$P(\hat{\sigma}_n < \sigma_F - \epsilon) \leq P(\hat{\lambda}_n(\sigma_F - \epsilon) < 0)$$

for all n. Therefore,

$$\lim_{n \to \infty} P(\hat{\sigma}_n < \sigma_F - \epsilon) \leq P(\lambda_F(\sigma_F - \epsilon) < 0) = 0.$$

Using similar arguments, we can show $P(\hat{\sigma}_n > \sigma_F + \epsilon) = 0$. Thus, by definition, $\hat{\sigma}_n \to \sigma_F$. □

2.3 Breakdown Point

In this section, the breakdown point of an M-estimator of scale will be derived. Due to length constraints, we will restrict our attention to the asymptotic breakdown point. Consider the contamination model in (1) in the case where $\mathcal{G}$ is the set of point mass distributions. Let $x_1, x_2, \ldots, x_n$ be i.i.d observations with distribution,

$$F_\epsilon = (1 - \epsilon)F + \epsilon G.$$

The scale estimator of this data is denoted as $\sigma_\epsilon \in [0, \infty]$.

Theorem 2.4. *Under the regularity conditions (R1)-(R3) and letting* $\lim_{t \to \infty} \rho(t) = 1$. *The asymptotic breakdown point of M-estimates of scale is,*

$$\epsilon^* = min(\delta, 1 - \delta)$$

Proof. This proof is based on that in [Mar+19], adapted to be specific to M-estimates of scale. Recall that $\sigma_\epsilon = \lim_{n \to \infty} \hat{\sigma}_n(F_\epsilon)$ is the implicit solution to $E_{F_\epsilon}(\rho(x/\hat{\sigma}_\epsilon) - \delta) = 0$

Letting $G = \delta_{x_0}$,

$$(1 - \epsilon)E_F(\rho(x/\hat{\sigma}_\epsilon)) - \delta) + \epsilon(\rho(x_0/\hat{\sigma}_\epsilon) - \delta) = 0 \tag{9}$$

We will find the breakdown point by finding an upper and lower bound of ϵ^* and showing that they are equal. Let $\epsilon < \epsilon^*$. This implies that $\hat{\sigma}_\epsilon$ is bounded. Hence

$$\lim_{x_0 \to \infty} \rho(x_0/\hat{\sigma}_\epsilon) - \delta = 1 - \delta \text{ and } \lim_{x_0 \to 0} \rho(x_0/\hat{\sigma}_\epsilon) - \delta = -\delta \tag{10}$$

Let $x_0 \to \infty$ in (15). By (R2), $\rho(t) \geq 0$ so

$$0 = (1 - \epsilon)E_F(\rho(x/\hat{\sigma}_\epsilon)) - \delta) + \epsilon(1 - \delta) \geq (1 - \epsilon)(-\delta) + \epsilon(1 - \delta) \implies \epsilon \leq \delta \tag{11}$$

Now, let $x_0 \to 0$ in (15). We have , $\rho(t) \leq 1$ so

$$0 = (1 - \epsilon)E_F(\rho(x/\hat{\sigma}_\epsilon)) - \delta) + \epsilon(-\delta) \leq (1 - \epsilon)(1 - \delta) + \epsilon(-\delta) \implies \epsilon \leq 1 - \delta \tag{12}$$

Hence, $\epsilon \leq \min(\delta, 1 - \delta)$. Now, let $\epsilon > \epsilon^*$. This implies that there exists a subsequence G_n such that $\hat{\sigma}_{\epsilon,n}$ is unbounded. Suppose this subsequence contains a subsequence such that $\hat{\sigma}_{\epsilon,n}$ tends to ∞. Then $\rho(x/\hat{\sigma}_\epsilon) \to 0$ so clearly $\rho(x/\hat{\sigma}_{\epsilon,n}) - \delta \leq 1 - \delta$ for each x. Then (15) implies

$$0 \leq (1 - \epsilon) \lim_{n \to \infty} E_F(\rho(x/\hat{\sigma}_{\epsilon,n})) - \delta) + \epsilon(1 - \delta). \tag{13}$$

Under regularity conditions, $\rho(x/\hat{\sigma}_{\epsilon,n})$ is bounded, so Bounded Convergence Theorem implies

$$0 \leq (1 - \epsilon)(-\delta) + \epsilon(1 - \delta) \implies \epsilon \geq \delta. \tag{14}$$

Similarly, as $\rho(x/\hat{\sigma}_{\epsilon,n}) - \delta \geq -\delta$ for each x (15) and Bounded Convergence Theorem imply

$$0 \geq (1 - \epsilon)(1 - \delta) + \epsilon(-\delta) \implies \epsilon \geq 1 - \delta. \tag{15}$$

Hence, $\epsilon \geq \min(\delta, 1 - \delta)$. The upper and lower bound are equal so $\epsilon^* = \min(\delta, 1 - \delta)$. □

2.4 Asymptotic Normality and Efficiency

Asymptotic efficiency is used to measure the performance of an estimator for uncontaminated data. The formal definition of asymptotic efficiency is given in Definition 1.3. Hence, we will show that an M-estimate of scale is asymptotically normal, find the asymptotic variance, and calculate the asymptotic efficiency. We will use the notation from (7) and in (8,

Theorem 2.5. *Consider the set of observations $x_1, x_2, \ldots, x_n$ which are iid and $x_i \sim F$ with variance σ_F^2 and mean zero. Let $A = E(x, \sigma_F)^2$ and $B = \lambda'(\sigma_F)$. Assume that $A < \infty$. Then*

$$\sqrt{n}(\hat{\sigma}_n - \sigma_F) \xrightarrow{d} N(0, \nu)$$

where $\nu = A/B^2$. Further, if $\dot{\Psi}(x, \sigma) = \partial\Psi(x, \sigma)/\partial\sigma$ is dominated by some function $K(x)$ for all σ such that $\mathbb{E}K(x) < \infty$ then $B = \mathbb{E}\dot{\Psi}(x, \sigma)$

Proof. This proof is based on that in [Mar+19], adapted to be specific to M-estimates of scale. By assumption, $\dot{\Psi}(x, \sigma)$ is dominated by $K(x)$ so Dominated Convergence Theorem implies that $B = \mathbb{E}_F\dot{\Psi}(x, \sigma)$, proving the second statement. To prove the first statement, we will perform a second Taylor expansion of $\dot{\Psi}(x, \hat{\sigma}_n)$ about the true value, σ_F, and find the asymptotic distribution of the result. The Taylor expansion at σ_F is given by

$$\Psi(x, \hat{\sigma}_n) = \Psi(x, \hat{\sigma}_F) + (\hat{\sigma}_n - \sigma_F)\dot{\Psi}(x, \hat{\sigma}_n) + 0.5(\hat{\sigma}_n - \sigma_F)^2\ddot{\Psi}(x, \sigma_i)$$

. with $\sigma_i \in (\hat{\sigma}_n, \sigma_F)$. Taking this expression at each x_i and taking the average across all i, we get

$$\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i, \hat{\sigma}_F) + (\hat{\sigma}_n - \sigma_F)\frac{1}{n}\sum_{i=1}^{n}\dot{\Psi}(x_i, \hat{\sigma}_n) + \frac{1}{2}(\hat{\sigma}_n - \sigma_F)^2\frac{1}{n}\sum_{i=1}^{n}\ddot{\Psi}(x_i, \sigma_i) = 0$$

$$= A_n + B_n(\hat{\sigma}_n - \sigma_F) + C_n(\hat{\sigma}_n - \sigma_F)^2$$

Solving for $\sqrt{n}(\hat{\sigma}_n - \sigma_F)$, we find

$$\sqrt{n}(\hat{\sigma}_n - \sigma_F) = -\sqrt{n}A_n/(B_n + (\hat{\sigma}_n - \sigma_F)C_n)$$

First, consider $(\hat{\sigma}_n - \sigma_F)C_n$. From Theorem 2.2, we have $\Psi(x_i, \hat{\sigma}_F)$ are iid with mean 0. Theorem 2.3 states that $\hat{\sigma}_n \xrightarrow{p} \sigma_F$ so Slutzky's Theorem implies $\hat{\sigma}_n - \sigma_F \xrightarrow{p} 0$. By assumption, C_n is bounded, so Slutzky's Theorem implies $(\hat{\sigma}_n - \sigma_F)C_n \xrightarrow{p} 0$. Second, consider B_n. The Law of Large Numbers implies that $B_n \xrightarrow{p} B$. Finally, Central Limit Theorem implies that $A_n \xrightarrow{d} N(0, A)$. Hence, Slutzky's Theorem implies,

$$\sqrt{n}(\hat{\sigma}_n - \sigma_F) \xrightarrow{d} N(0, A/B^2)$$

□

Using this result, the asymptotic efficiency can be easily found.

Theorem 2.6. *Consider the set of normal distributions with mean zero. The asymptotic efficiency of the M-estimator of scale, $\hat{\sigma}$ is given by,*

$$p_{eff}(\sigma) = \nu_{MLE}(\sigma)/\nu(\sigma) = 2B^2/\sigma^2 A$$

Proof. The asymptotic variance of the MLE of the standard deviation under normality is well known to be $\hat{\sigma}_{MLE} = \sigma^2/2$. The asymptotic efficiency is simple to calculate from Theorem 2.5. □

3 S-estimators for Linear Regression

A useful robust linear regression method must perform well for data with a distribution given by a contamination neighbourhood. It must perform well when the data is entirely uncontaminated ($\epsilon = 0$) and when it is not. Therefore, it is reasonable to first consider classical estimators and then try to improve their robustness. Consider the regression dataset $(y_i, \mathbf{x_i})$ with $\mathbf{x}_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$ iid observations follwing the linear regression model

$$y_i = \mathbf{x_i}'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{16}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of coefficients to be estimated and $\varepsilon_i \sim (1-\epsilon)F + \epsilon G$ is the error, which is independent of $\mathbf{x}_i$. The case where $\epsilon = 0$ and $F \sim N(0, \sigma^2)$ is assumed when deriving ordinary least squares (OLS). The OLS estimate is the maximum likelihood estimate is given by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1} r_i(\boldsymbol{\beta})^2 \tag{17}$$

where $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ is the residual of observation i. Similar to the M-estimate of scale, one approach to creating a robust estimator is to replace the square penalty with a more general ρ function. This is called an M-estimator for linear regression.

M-estimators for linear regression have high asymptotic efficiency, but require that the data be normalized with a robust estimate of the scale of the residuals, $\hat{\sigma}$. However, estimating the scale of the residuals requires an estimate of $\boldsymbol{\beta}$. Obviously, this presents a problem. One might consider making a joint estimate for both $\boldsymbol{\beta}$ and $\hat{\sigma}$. However, such solutions are not robust to "leverage points", which are observations that are outlying in both the predictors and the responses [Mar+19].

An alternative approach is to notice that (17) can also be interpreted to be minimizing an estimate of the scale of the residuals, namely the variance. A reasonable robust estimator might therefore estimate the $\boldsymbol{\beta}$ which minimizes some robust estimate of the scale of the residuals. Indeed, that is the approach taken by S-estimators for linear regression [RY84].

Specifically, S-estimators minimize the M-estimate of scale of the residuals, described in Section 2. In addition to in Definition 2.1, we will consider ρ function with the following properties,

(R4) There exists $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant beyond c with $\rho(c) \geq 1$.

(R5) $K = \mathbb{E}_\Phi[\rho] = \frac{1}{2}\rho(c)$.

The restrictions are slightly more strict than the standard definition of ρ functions corresponding to redescending M-estimates [Mar+19], but are sufficient to include many common ρ functions. It should be mentioned that, although $\rho(c)$ is not one, M-estimates of scale defined with (R3) and (R4) have an asymptotic breakdown point of 0.5.

Definition 3.1. *Consider the regression dataset $(y_i, \mathbf{x_i})$ with $\mathbf{x}_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$ iid observations.* ***S-estimates for linear regression*** *are defined by*

$$\underset{\boldsymbol{\beta}}{\arg\min}\ s(r_1(\boldsymbol{\beta}), r_2(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta})) \tag{18}$$

where $s(r_1, \ldots, r_n)$ is an M-estimate of scale defined by (5).

We will assume that all observations where $_i = 0$ have been removed, as these points contain no information on $\boldsymbol{\beta}$ and that no more than $\left[\frac{n}{2}\right]$ points lie on a $(p-1)$ dimensional subspace of $(y = 0)$ that passes through the origin. Also, when we say the observations are in the general position when any p points give a unique value for $\hat{\boldsymbol{\beta}}$. These restrictions greatly simplify the calculation of the breakdown point.

The following Lemma is an important intermediate result. By constructing bounds for $s(r_1, \ldots, r_n)$ using $\text{median}_{1\leq i\leq n}(|r_i|)$, we can use existing results for medians to prove properties of S-estimators.

Lemma 1. *For each ρ satisfying conditions (R1) to (R3) and for each $n \in \mathbb{N}$ there exists $\alpha, \beta \in \mathbb{R}_+$ such that the s given by equation (5) satisfies*

$$\alpha \underset{1\leq i\leq n}{\text{median}}(|r_1|, \ldots, |r_n|) \leq s(r_1, \ldots, r_n) \leq \beta \underset{1\leq i\leq n}{\text{median}}(|r_1|, \ldots, |r_n|) \tag{19}$$

Proof. The proof expands upon the proof in [Rou84a]. The inequality can be verified for $\alpha = 1/c$ and $\beta = 1/\rho^{-1}(\rho(c)/(n+1))$ for n odd and $\beta = 2/\rho^{-1}(\rho(c)/(n+2))$ for n even.

First, consider $\sum_{i=1}^{n} \rho(cr_i/\operatorname{median}_i |r_i|)$. We have $|r_i|/\operatorname{median}_i |r_i| \geq 1$ for $[n/2]$ residuals. This implies

$$\sum_{i=1}^{n} \rho\left(c\frac{r_i}{\operatorname{median}_i |r_i|}\right) \geq [n/2]\rho(c)$$

as ρ is symmetric. (R5) gives $\sum_{i=1}^{n} \rho(r_i/s) = n\rho(c)/2$ and ρ is non-decreasing, so this implies that $\alpha \operatorname{median}_{1\leq i\leq n}(|r_1|, \ldots, |r_n|) \leq s(r_1, \ldots, r_n)$ for $\alpha = 1/c$.

Now, we will prove the upper inequality. Let n be odd. Similar to before, consider

$$\sum_{i=1}^{n} \rho\left(\rho^{-1}\left(\frac{\rho(c)}{n+1}\right)\frac{r_i}{\operatorname{median}_i |r_i|}\right).$$

Obviously, $\rho^{-1}(\rho(c)/(n+1)) < c$ so this implies that $\rho(r_i/\beta \operatorname{median}_i |r_i|) \leq \rho(c)$ for $n+1/2$ observations. Hence,

$$\sum_{i=1}^{n} \rho\left(\rho^{-1}\left(\frac{\rho(c)}{n+1}\right)\frac{r_i}{\operatorname{median}_i |r_i|}\right) \leq \frac{n+1}{n}\rho\left(\rho^{-1}\left(\frac{\rho(c)}{n+1}\right)\right) + \frac{(n-1)\rho(c)}{2n} = \rho(c)/2$$

which implies $s \leq \beta \operatorname{median}_i |r_i|$ for $\beta = 1/\rho^{-1}(\rho(c)/(n+1))$. Verification for n even is very similar, except $n/2$ points have $r_i/\operatorname{median}_i |r_i| > 1$.

□

3.1 Breakdown Point of S-estimators

In this section, we will consider the finite breakdown ((1.2)) and observe the limit as $n \to \infty$. Finding the BDP of S-estimators is more complicated than for M-estimates of scale as S-estimators are defined as the minimizer of an implicitly defined robust scale. However, medians are commonly used in robust statistics so Lemma 2.2 provides a useful means to extend the properties of medians to M-estimates of scale and thus S-estimators. The property that we will need is given by Theorem 1 from [Rou84a], which is as follows,

Lemma 2. *The least median of squares estimator is given by* $\hat{\boldsymbol{\beta}}_{LMS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \operatorname{median}_i r_i^2$. *If* $p > 1$ *and the observations are in the general position, the breakdown point of the LMS is* $([n/2] - p + 2)/n$ *[Rou84a].*

The proof of this Lemma is given in [Rou84a]. To use Lemma 1, this result must be extended to the minimizer of the median of the absolute values of the residuals.

Lemma 3. *The least median of absolute (LMA) estimator is given by* $\arg\min_{\hat{\boldsymbol{\beta}}} \operatorname{median}_i |r_i|$. *If* $p > 1$ *and the observations are in the general position, the breakdown point of the LMS is* $([n/2] - p + 2)/n$ *[Rou84a].*

Proof. Obviously, $|r_i|^2 = r_i^2$ so as $g(t) = t^2$ is convex and $h(t) = |t|$ is a positive function, the solution to $\arg\min_{\hat{\boldsymbol{\beta}}} \operatorname{median}_i r_i^2$ is also the solution to the LMA estimator. Hence, the BDP of the LMA estimator is given by Lemma 2. □

Now that the intermediate results are proven, we can consider the breakdown point of S-estimators.

Theorem 3.1. *Under the given conditions on* ρ, *if* $p > 1$ *and the observations are in the general position an S-estimator has a finite sample breakdown point of* $\epsilon_n^* = ([n/2] - p + 2)/n$.

Proof. This proof will use Lemma 3 and Lemma 1 to prove that the finite sample breakdown point of an S-estimator is equal to ϵ_n^*. First, consider a sample where $m < m^* = ([n/2] - p + 2)$ observations in a sample are replaced with arbitrary values. Lemma 3 implies that $\hat{\boldsymbol{\beta}}_{LMA} = \arg\min_{\hat{\boldsymbol{\beta}}} \operatorname{median}_i |r_i|$ is bounded. Hence, Lemma 1 implies that $s(r_1(\hat{\boldsymbol{\beta}}_{LMA}), \ldots, r_n(\hat{\boldsymbol{\beta}}_{LMA})$ is bounded. By definition, $s(r_1(\hat{\boldsymbol{\beta}}), \ldots, r_n(\hat{\boldsymbol{\beta}})) \leq s(r_1(\hat{\boldsymbol{\beta}}_{LMA}), \ldots, r_n(\hat{\boldsymbol{\beta}}_{LMA})$ which implies that the scale is bounded above at $\hat{\boldsymbol{\beta}}$.Also, $\alpha \operatorname{median}_i |r_i(\hat{\boldsymbol{\beta}}_{LMA})|$ is the smallest lower bound and is non-zero by Lemma 3.

Let C be the set of indices with arbitrary observations and D be the remaining observations. First, assume that the norm of the S-estimator $||\hat{\boldsymbol{\beta}}||$ is unbounded. If $||\hat{\boldsymbol{\beta}}||$ is unbounded and s is bounded, then $r_i \to \infty$ for $i \in D$. When $r_i \to \infty$, $\rho(r_i) \to \rho(c)$. By assumption, $\#(D) = n - m \geq n - ([n/2] - p + 2)$. Hence, by definition of s and recalling that $K = \rho(c)/2$, we can show

$$\sum_{i=1} n\rho(r_i) = nK \geq ([n/2] + p - 2)\rho(c) > nK$$

which is a contradiction. Hence, $||\hat{\boldsymbol{\beta}}||$ is bounded for $m \leq^*$.

Next, consider a sample where $m \geq ([n/2] - p + 2) = m^*$ observations in a sample are replaced with arbitrary values. For this case, we will need an additional result, given by Corollory 1 in [*Rou84a*], called the exact fit property. The exact fit property is given by,

Lemma 4. *If $p > 1$ and there exists some $\boldsymbol{\beta}$ such that at least $n - [n/2] + p - 1$ observations are satisfy $y_i = \mathbf{x}_i'\boldsymbol{\beta}$ exactly and are in the general position, then the LMS solution equals $\boldsymbol{\beta}$, whatever the observations are [Rou84a].*

By Lemma 3, this property also holds for the LMA estimator. Using this result, we can construct a data set such that $||\hat{\boldsymbol{\beta}}||$ is unbounded. First, define some plane H such that $p - 1$ observations lie on this plane. Next, place the $([n/2] - p + 2)$ arbitrary observations on H. A total of $([n/2] - p + 2) + p - 1 = [n/2] + 1 > n/2$ points lie on H, so a total of $[n/2] + 1$ residuals are zero. Hence, $\text{median}_i\, |r_i| = 0$ and thus Lemma 1 implies that $s = 0$. When $s \to 0$, $\rho(r_i/s) \to \rho(c)$ for finite r_i. Assume $||\hat{\boldsymbol{\beta}}||$ is bounded. Then

$$nK = \sum_{i=1}^{n} \rho(r_i/s) \geq ([n/2] + 1)\rho(c) > nK$$

which is a contradiction. Hence, $||\hat{\boldsymbol{\beta}}||$ is unbounded for $m > m^*$, completing the proof. □

We can see that for fixed p that as $n \to \infty$ that $\epsilon_n^* \to 0.5$. This proof assumes a finite sample size, so an additional proof is needed to find the asymptotic breakdown point. However, such a proof also finds breakdown point of 0.5 [RY84].

3.2 Consistency of S-estimators

Consistency for S-estimators follows from the consistency of general M-estimators [Mar+19]. A general M-estimator of regression and scale is given by the following definition [MY81],

Definition 3.2. *Consider the regression dataset $(y_i, \mathbf{x}_i)$ with $i \in \{1, \ldots, p\}$ and $\mathbf{x}_i \in \mathbb{R}^p$ of iid random variables follwing the linear regression model $y_i = \mathbf{x}_i' + \varepsilon_i$ where ε_i are iid random variables independent of $\mathbf{x}_i$ following the parametric distribution family F_σ. An M-estimator of regression and scale $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ is defined as the solution to the system of equations,*

$$\sum_{i=1}^{n} \Psi(\mathbf{x}_i, (y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma)\mathbf{x}_i = \mathbf{0} \tag{20}$$

$$\sum_{i=1}^{n} \chi((y_i - \mathbf{x}_i'\boldsymbol{\beta})/\sigma) = 0 \tag{21}$$

Considering Definition 3.1, it is clear that an S-estimator is a generalized M-estimator of regression and scale when $\psi(u, v) = -\rho'(v)$ and $\chi(u) = \rho(u) - \delta$ where (20) gives the first order conditions and (21) gives the M-estimate of scale of the residuals. We can now consider the consistency of S-estimators.

Theorem 3.2. *Consider the regression dataset $(y_i, \mathbf{x}_i)$ with $i \in \{1, \ldots, p\}$ and $\mathbf{x}_i \in \mathbb{R}^p$ of iid random variables follwing the linear regression model $y_i = \mathbf{x}_i' + \varepsilon_i$ where ε_i are iid random variables independent of $\mathbf{x}_i$ following the parametric distribution family F_σ. If the conditions (R1)-(R5) hold as well as (i) $\rho'(t)/t$ is non-increasing for $t > 0$ and (ii) $E_H[||x||] < \infty$ where $x \sim H$ and H has a density. Then the S-estimator for linear regression $(\hat{\sigma}, \hat{\boldsymbol{\beta}}) \xrightarrow{a.s.} (\sigma, \boldsymbol{\beta})$.*

Proof. As S-estimators are a type of general M-estimate, Theorem 3.1 in [MY81] implies that the S-estimators consistent under conditions (R1)-(R5). □

3.3 Asymptotic Normality of S-estimators

In practice, when comparing the asymptotic efficiency of two robust estimators, on the asymptotic variance. This is because therefore the minimum asymptotic variance in the efficiency ν_{min} is the same in the efficiencies of both estimators. Hence, only the variance of the asymptotic distribution will considered.

Similar to the consistency of S-estimators, the asymptotic normality of S-estimators is given by the fact that S-estimators are a type of general M-estimate for regression and scale. We can thus find the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ as follows [RY84].

Theorem 3.3. *For simplicity, consider $\theta_0 = 0$ and $\sigma_0 = 1$. If conditions (i) and (ii) from Theroem 3.2 hold and (iii) ρ' is differentiably everywhere but a finite number of points, $|\rho''|$ is bounded and $\int \rho'' d\Psi > 0$, where Ψ is the normal distribution and (iv) $\mathbb{E}_H[\mathbf{x}'\mathbf{x}]$ is nonsingular and $E_H[||\mathbf{x}||^3] < \infty$ then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(0, \mathbb{E}_H[\mathbf{x}'\mathbf{x}]^{-1} \int (\rho')^2 d\Psi / (\int \rho' d\Psi)^2) \text{ and} \tag{22}$$

$$\sqrt{n}(\hat{\sigma}_n - \sigma) \xrightarrow{d} N(0, \int (\rho(t) - K)^2 d\Psi(t) / (\int t\rho(t) d\Psi(t))^2) \tag{23}$$

Proof. S-estimators are a type of general M-estimate so this follows directly from Theorem 4.1 of [MY81]. □

3.4 Computation of S-estimators

S-estimators are defined as the implicit solution to (3.1). Solving this solution requires solving a non-convex optimization problem. As a result, there is not guarantee that common iterative methods, such as gradient descent, will converge to a global minima. S-estimators are generally calculated by choosing a good initial estimate for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ and performing iterative updates until a local minimum is reached [Mar+19]. These methods are discussed in detail Section 5.7.1.1 and Section 9.2 of "Robust Statistics with Examples in R" by Maronna *et al.* (2019) [Mar+19].

4 Open questions and research directions

So far, we have only considered S-estimators for linear regression. S-estimates for linear regression is defined as the minimizer of a robust estimate of scale. However, many other regression estimators are also defined as minimizers of some estimate of scale. Similarly, we may wish to consider replacing a non-robust estimate of scale with a robust estimate of scale to get a robust estimator. Such methods are very common and are also referred to as S-estimators. In this section, we will discuss a new research direction, S-estimators for sparse linear regression. We will still consider the linear regression model defined in (16).

Robust and sparse estimation for linear models with high multicollinearity is commonly required in fields such as chemometrics and bioinformatics. Improvements in sensor and computer technology has led to the number of variables p to grow much larger than the number of observations n. However, the number of variables correlated with the response s is still frequently very small relative to the number of observations, requiring sparsity. Sparse methods can improve estimator variance when p/n is large and s/n is small by selecting a subset of variables to include in the model.

The Least absolution shrinkage and selection operator (Lasso) [Tib96] and Elastic Net [05] are commonly used sparse regularized regression methods. Lasso penalizes the L_1 norm of the estimators to produce a sparse solution. However, Lasso is only able to select at most n variables. Also, when groups of highly correlated variables are present in the data Lasso tends to select only one member of the group. To address these problems, Elastic Net penalizes a linear combination of the L_1 and the L_2 norms of the estimators. As a result, Elastic Net can be tuned to have different levels of sparsity and, unlike Lasso, can select more variables than n.

However, while both Lasso and Elastic Net address the issue of sparsity, they do not address the issue of outliers. In many applications, there can be sources of outlying data, including faulty equipment and mistakes during data transcription, making a robust method desirable. However, existing work on robust regularized regression has been limited. The first published highly robust regularized regression method was RLARS [KAZ07], a modification of Least Angle Regression (LARS). This method replaces the sample correlation in the LARS method with a robust correlation. An extension of RLARS to include a variant of the Lasso modification to LARS is described in the thesis summary section. Later work by [ACG13] extended the Least Trimmed Squares estimator [Rou84b] to include L_1 penalization of the estimators. This method is demonstrated to have good robustness properties, but can only be tuned for either high robustness or high efficiency under normality, but not both.

Recently, [Chr+19] proposed the Split Regularized Regression estimator (SplitReg). SplitReg includes a linear combination of the L_1 and L_2 penalties to encourage sparsity while also splitting variables between several models, penalizing similarity between models. Theoretical results [CVZ18] have shown SplitReg reduces estimator variance when data has some collinearity. Simulation studies and real-data applications have demonstrated improved predictive performance over the base estimator, Elastic Net, over a range of correlation structures, signal-to-noise ratios, and dimensionalities. This suggests that robust regularized regression method that splits variables between models like SplitReg is a good candidate to outperform existing robust regularized regression methods.

Therefore, the development of a robust split regularized regression estimator could offer significant improvements. The first step would be to develop a split regularized Elastic Net S-estimator by using a robust square scale function in the place of the sum of squares residual error. Consider the case where we have G models in our estimator. The coefficients will be a $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$ is a pxG matrix, with the coefficients for each model $i \in 1, 2, \ldots, G$ being in the corresponding column of $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}_g$ be column g of $\boldsymbol{\beta}$. The proposed estimator is defined as the solution to

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{p x G}}{\arg\min} \sum_{g=1}^{G} \left(\sigma(\mathbf{r}(\boldsymbol{\mu}, \boldsymbol{\beta}_g))^2 + \lambda_s \Big(\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}_g||_2^2 + \alpha||\boldsymbol{\beta}_g||_1) \Big) + \frac{\lambda_d}{2} \sum_{h \neq g}^{G} P_d(\boldsymbol{\beta}_g, \boldsymbol{\beta}_h) \right) \tag{24}$$

where σ is an M-estimate of scale of the residuals, λ_s scales the sparsity penalty and α tunes the combination of the L_1 and L_2 penalties, corresponding to elastic net, and λ_d scales the diversity penalty and P_d is diversity penalty, penalizing similarity models. This is the equivalent to summing the PENSE loss function [Fre+17] across the G models with the addition of a diversity penalty. The most significant obstacle in this work would be the computational complexity, as both PENSE and SplitReg are already computationally intensive.

A Exercises

Two exercises are presented in this section. The first exercises presents a straightforward robustness and consistency proof. Breakdown point proofs can quickly become complex, so the available options were limited, but I think that the median absolute deviation is a good choice. The breakdown point derivation and consistency proof for the median absolute deviation are simple, but require a clear understanding of the problems. The second exercise is an existence proof from literature [CRH94] that, while short, requires some thought.

Exercise 1

Question

The definition of the median absolute deviation (MAD) estimator of scale is given by,

$$\hat{\sigma} = c \operatorname*{median}_i |x_i - \operatorname*{median}_i x_i| \tag{25}$$

for some $c \in \mathbb{R}$. For simplicity, in this problem we will assume that $\text{median}_i\, x_i = 0$, n is odd, and $x_i \neq x_j$ for all $i \neq j$.

(i) What is the finite sample breakdown point?

(ii) Assume $x \sim N(0, \sigma^2)$. When is $\hat{\sigma}$ a consistent estimator of the standard deviation?

Solutions

(i)
Consider a set of observations with $m = (n-1)/2$ arbitrary values. By definition, for n odd,

$$\#\{i : |x_i| < \operatorname*{median}_i |x_i|\} = (n-1)/2.$$

Hence, if $|x_i| < M$ for some $M \in \mathbb{R}$ for $(n+1)/2$ observations, then $\text{median}_i\, |x_i| < M$. The non-arbitrary values are bounded by the largest observation plus one, which implies $m^* >= (n-1)/2$.
Next, consider a dataset with $(n+1)/2$ observations replaced with arbitrary values. If we take all of these values to be equal to x_0 and take $x_0 \to \infty$, then $\text{median}_i\, x_i \to \infty$, which implies $m^* <= (n-1)/2$. Hence, $\epsilon^* = m^*/n = 1 - (1/2)n$, which goes to 0.5 as $n \to \infty$.
(ii)
We will first show that $\text{median}_i\, |x_i| \xrightarrow{p} F^{-1}_{|X|}(0.5)$ for all $\varepsilon > 0$ for some $c \in \mathbb{R}_+$. Let ν be the true median and $\varepsilon \in \mathbb{R}\ \{0\}$. Then,

$$P(c \operatorname*{median}_i |x_i| > \nu + \varepsilon)\ (*)$$

is the probability that at least $(n+1)/2$ points exceed $\nu + \varepsilon$. This corresponds to a binomial random variable $N_n > \text{Binom}(n, p)$ where $p = P(X > \nu + \varepsilon) \neq 0.5$. Hence

$$\begin{aligned}
(*) &= P(N_n \geq (n+1)/2) && (26)\\
&= P(N_n - np \geq (n+1)/2 - np) && (27)\\
&\leq P(N_n - np \geq n(0.5 - p)). && (28)\\
& && (29)
\end{aligned}$$

As $Var(N_n) < \infty$, the Extended Markov Inequality implies that

$$(*) \leq \frac{p(1-p)}{n(0.5-p)}.$$

We have $p \neq 0.5$. This implies that

$$\frac{p(1-p)}{n(0.5-p)} \to 0 \text{ as } n \to \infty$$

and thus $P(|c\,\text{median}_i\,|x_i| - \nu| + \varepsilon) = 0$. We assume that $x_i \sim N(0, \sigma^2)$ for all i, so we know that $|x_i|$ is the half normal distribution. It is well known that the median of the half-normal distribution is $\sqrt{2}\text{erf}^{-1}(0.5)\sigma$ where erf is the error function. Hence, Slutzky's Theorem implies that $c\hat{\sigma} \xrightarrow{p} \sigma$ for $c = 1/(\sqrt{2}\text{erf}^{-1}(0.5))$.

Exercise 2

The least quartile difference estimator for linear regression (LQD-estimator) [CRH94] is defined as,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, Q_n(r_1, \ldots, r_n),$$

$$Q_n = \{|r_i - r_j|; i < j\}_{\binom{h_p}{2}:\binom{n}{2}}$$

where r_i is the residual of the ith case and $h_p = (n + p + 1)/2$. Q_n is the $\binom{h_p}{2}$ order statistic of the set $\{|r_i - r_j|; i < j\}$ among the $\binom{n}{2}$ elements in the set.
The LQD-estimator is a special case of a Generalized S-estimator, defined as the minimizer of a more general class of functions of residuals. Assume that none of the differences $(\mathbf{x}_i - \mathbf{x}_j, y_i - y_j)$ lie on the same vertical hyperplane in $\mathbb{R}^p$. A vertical hyperplane is defined as a plane passing through $(\mathbf{0}, 0)$ and $(\mathbf{0}, 1)$.
Prove the existence of the LQD estimator. $Q_n(\boldsymbol{\beta})$ is continuous, so it is sufficient to show that there exists a compact ball $\mathcal{B}$, such that for any $\boldsymbol{\beta}_1 \in \mathcal{B}$ and $\boldsymbol{\beta}2 \in \mathbb{R}^{p+1}$ $\mathcal{B}$ that $Q_n(\boldsymbol{\beta}_1) < Q_n(\boldsymbol{\beta}_2)$.
Hint: For the ith order statistic of some sequence $r_1, r_2, \ldots, r_n$, if $\#\{j : r_j > M\} = n - i + 1$ for some $M \in \mathbb{R}$ then $r_{(i)} > M$.

Solutions

This solution relies heavily on the proof in [CRH94]. It is about as simple as an existence proof gets, so I believe it is a reasonable exercise. The key for this proof is to consider $Q_n(\boldsymbol{\beta})$ at a $\boldsymbol{\beta}$ for which we have a known bound and then construct a ball around this point for which $|r_i - r_j|$ exceeds the bound for at least $(\binom{n}{2} - \binom{h_p}{2} + 1)$ observations (as Q_n is an order statistic).
Let $M = \max_{i<j} |y_i - y_j|$. This corresponds to the maximum difference between residuals for $||\beta|| = 0$. The reverse triangle inequality implies that,

$$|r_i - r_j| \geq ||x_i'\boldsymbol{\beta} - x_j'\boldsymbol{\beta}| - |y_i - y_j||.$$

Let $\boldsymbol{\theta} = \boldsymbol{\beta}/||\boldsymbol{\beta}||$.

$$= ||x_i'\boldsymbol{\theta} - x_j'\boldsymbol{\theta}|||\boldsymbol{\beta}|| - |y_i - y_j||.(*)$$

The fact that the observations are in the general position now becomes important. By definition of the general position, there exists no $p-1$ dimensional subspace of the plane $y = 0$ on which $\binom{h_p}{2}$ points lie. As a result,

$$\inf_{i<j} |x_i'\boldsymbol{\theta} - x_j'\boldsymbol{\theta}| = \delta > 0.$$

If the observations were not in the general position, this would not be true. Let $||\boldsymbol{\beta}|| = 2M/\delta + 1$,

$$(*) \geq |x_i'\boldsymbol{\theta} - x_j'\boldsymbol{\theta}|||\boldsymbol{\beta}|| - M$$

$$\geq (2M/\delta + 1)\delta - M > M$$

for $(\binom{n}{2} - \binom{h_p}{2} + 1)$ observations. Let $\mathcal{B}(2M/\delta + 1)$ be ball of radius $2M/\delta + 1$ about $\boldsymbol{\beta} = \mathbf{0}$. Thus for all $\boldsymbol{\beta}_1 \in \mathcal{B}(2M/\delta + 1)$ and $\boldsymbol{\beta}_2 \in (\mathcal{B}(2M/\delta + 1))^C$,

$$Q_b(\boldsymbol{\beta}_1) < Q_b(\boldsymbol{\beta}_2).$$

As $Q_n(\boldsymbol{\beta})$ is continuous, this implies that a minimum of Q_n exists and the minimum is within the ball $\mathcal{B}(2M/\delta + 1)$.

References

[ACG13] A. Alfons, C. Croux, and S. Gelper. "SPARSE LEAST TRIMMED SQUARES REGRESSION FOR ANALYZING HIGH-DIMENSIONAL LARGE DATA SETS". In: *The Annals of Applied Statistics* 7.1 (2013), pp. 226–248.

[Alq+09] F. Alqallaf et al. "Propagation of outliers in multivariate data". In: *The Annals of Statistics* 37.1 (2009), pp. 311–331.

[Chr+19] A.-A. Christidis et al. "Split Regularized Regression". In: *Technometrics* (2019), pp. 1–9.

[CVZ18] A. Christidis, S. Van Aelst, and R. Zamar. "Split regression modeling". In: *arXiv preprint arXiv:1812.05678* (2018).

[CRH94] C. Croux, P. J. Rousseeuw, and O. Hössjer. "Generalized S-estimators". In: *Journal of the American Statistical Association* 89.428 (1994), pp. 1271–1281.

[Fre+17] G. V. C. Freue et al. "PENSE: A Penalized Elastic Net S-Estimator". In: (2017).

[KAZ07] J. A. Khan, S. V. Aelst, and R. H. Zamar. "Robust Linear Model Selection Based on Least Angle Regression". In: *Journal of the American Statistical Association* 102.480 (2007), pp. 1289–1299.

[MY81] R. A. Maronna and V. J. Yohai. "Asymptotic behavior of general M-estimates for regression and scale with random carriers". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 58.1 (1981), pp. 7–20.

[Mar+19] R. A. Maronna et al. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[05] "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[Rou84a] P. J. Rousseeuw. "Least median of squares regression". In: *Journal of the American statistical association* 79.388 (1984), pp. 871–880.

[Rou84b] P. J. Rousseeuw. "Least Median of Squares Regression". In: *Journal of the American Statistical Association* 79.388 (1984), pp. 871–880.

[RY84] P. Rousseeuw and V. Yohai. "Robust regression by means of S-estimators". In: *Robust and nonlinear time series analysis*. Springer, 1984, pp. 256–272.

[Tib96] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288.

[Yoh+87] V. J. Yohai et al. "High breakdown-point and high efficiency robust estimates for regression". In: *The Annals of Statistics* 15.2 (1987), pp. 642–656.