CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Daniel Pacák

Least Absolute Deviations

Bachelor thesis

Prague 2017

Author: Daniel Pacák
Supervisor: prof. RNDr. Jan Ámos Víšek CSc.

Academic Year: 2016/2017

Bibliographic note

Daniel, Pacák. *Least Absolute Deviations*. Prague 2017. 37 pp. Bachelor thesis (Bc.) Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor prof. RNDr. Jan Ámos Víšek CSc.

Abstract

This is a theoretical study of the Least Absolute Deviations (LAD) fits. In the first part, fundamental mathematical properties of LAD fits are established. Computational aspects of LAD fits are shown and the Barrodale-Roberts Algorithm for finding LAD fits is presented. In the second part, the statistical properties of LAD estimator are discussed in the concept of linear regression. It is shown that LAD estimator is a maximum likelihood estimator if the error variables follow Laplace distribution. We state theorems establishing strong consistency and asymptotic normality of LAD estimator and we discuss the bias of LAD estimator. In the last section, we present the results of numerical experiments, where we numerically showed consistency of LAD estimator and discussed its behaviour under different distributions of error variables with comparison to the Ordinary Least Squares (OLS) estimator. Lastly, we looked at the behaviour of LAD and OLS estimators in the presence of corrupted observations.

Keywords

Least Absolute Deviations, Barrodale-Roberts, l_1 norm, l_1 regression, Laplace errors, iteratively reweighted least squares, robustness

Abstrakt

Toto je teoretická studie metody Nejmenších absolutních odchylek (NAD). V první části jsou uvedeny základní matematické vlastnosti metody NAD. Jsou představeny komputační aspekty metody NAD a Barrodaleův - Robertseův algoritmus, který se při komputaci této metody používá. Ve druhé části jsou diskutovány statistické vlastnosti metody NAD v kontextu linární regrese. Je ukázáno, že odhad metodou NAD je maximálně věrohodným odhadem, za předpokladu že chyby mají laplaceovo rozdělení. Jsou uvedeny věty, které dokazují silnou konzistenci a asimtotickou normalizu. Dále je diskutována nestranost odhadu metodou NAD. V poslední části jsou prezentovány výsledky numerických experimentů, ve kterých je numericky ukázána konzistenci odhadu metodou NAD. Dále je studováno chování tohoto odhadu za různých distribučních funkcí chyb v porovnání s metodou Nejmenších čtverců. Nakonec je uvedeno chování těchto dvou odhadů v přítomnosti chybných dat.

Klíčová slova

nejmenší absolutní odchylky, Barrodale-Roberts, l_1 norma, l_1 regrese, laplaceovi chyby, iteračně opakovaně vážené nejmenší čtverce, robustnost

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, 27 July 2017

Signature

Bachelor Thesis Proposal

Author: Daniel Pacák
Supervisor: prof. RNDr. Jan Ámos Víšek CSc.
Proposed Topic:: Least Absolute Deviations

Preliminary scope of work

Method of Least Absolute Deviations, or the so-called l_1 estimator, provides probably the most intuitive alternative to a widely used method of Least Squares, or the so-called l_2 estimator. l_1 is considered to be more robust than l_2, yet less stable and more computationally difficult. The aim of this thesis is to study the l_1 estimator, its usage in linear regression, its properties, its differences from l_2 and its robustness.

After the historical background and the introduction, the theory of l_1 estimator will be discussed. Various applications of l_1 will be presented including examples where l_1 is more appropriate then l_2 as an estimator e.g. when l_1 gives maximum likelihood estimation (MLE) whereas l_2 does not (i.e. under Laplace distribution of errors instead of Gaussian). In the next section the weight function will be introduced into the l_1 model to further improve its robustness. In the last section the simulations using appropriate statistical software will be conducted and the results will be discussed.

Research questions: What are the practical applications of l_1? Under what circumstances should l_1 be used instead of l_2

Preliminary outline

1. Introduction
2. Least Absolute Deviation
3. Least Weighted Absolute Deviation
4. Simulations of the models' robustness (breakdown point)
5. Conclusion

References

[1] Brennan, J. J. and Seiford L. M. *Linear programming and l_1 regression: A geometric interpretation.* Computational Statistics & Data Analysis 5, no. 4, 1987, pp. 263-276.

[2] Dodge, Y. *An introduction to L1-norm based statistical data analysis.* Computational Statistics & Data Analysis 5, no. 4, 1987, pp. 239-253.

[3] Dutter R., *Robust regression: Different approaches to numerical solutions and algorithms.* Res., Rep. no. 6, Fachgruppe für Statistik, Eidgen. Technische Hochschnule, Zurich, 1975.

[4] Ellis, S. P. *Instability of least squares, least absolute deviation and least median of squares linear regression.* Statistical Science, 1998, pp. 337-344.

[5] Gokarna R. A., *Study of laplace and related probability distributions and their applications.* . Graduate Theses and Dissertations, University of South Florida, 2006.

[6] Li, Y. and Arce, G. R. *A maximum likelihood approach to least absolute deviation regression.* EURASIP Journal on Advances in Signal Processing 2004, no. 12, 2004, pp. 1-8.

[7] Ling, S. *Self?weighted least absolute deviation estimation for infinite variance autoregressive models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, no. 3, 2005, pp. 381-393.

Contents

Introduction

Method of the Least Squares is probably both the best known and the most used method for finding fits for linear models. The method of the Least Absolute Deviations (LAD) is most likely its closest relative, at least conceptually speaking. From historical point of view, it is hard to tell which method is older. Hald (1986) showed that the first significant use of LAD seems to point at Galileo Galilei, who in 1632 suggested testing astronomical hypotheses by taking means of sums of absolute deviations of hypothesised values from observed one. Nevertheless, the invention of calculus allowed rise of such methods, which are easily manipulated with using calculus. Least Squares is one of them. On the other hand, calculus offer little to no help in dealing with LAD. That is probably why it fell out of popularity and did not re-emerged until the second half of the twentieth century, when the rise of computers and linear programming offer the ability to compute LAD fits efficiently.

The aim of this theses is to study the LAD fits. We will begin by taking a look at its most important mathematical properties, which we will all be needed in the later sections. Next, we will take a look at the computational aspect of LAD fits. Then, we will take a close look at the properties of LAD fits in statistics while comparing to the Ordinary Least Squares (OLS). We will conclude by numerical simulations where we will aim to test the theory presented in the previous section and we will again compare with the OLS.

Preliminaries

We will begin with some notes on notation and on some of the heavily used mathematical concepts. If we do not say differently, n and k stand for any natural number. Let $x, y \in \mathbb{R}^n, x = (x_1, \ldots, x_n)^T, y = (y_1, \ldots, y_n)^T$. We say $x = y$ if and only if $\forall i \in \{1, \ldots, n\} : x_i = y_i$. Also, we use symbol 0 without further specifications for a zero vector with any number of zero components, i.e. $0 = (0, 0, \ldots, 0)$ where the number of zeros is such number that other operations are well-defined in the given context.

If we say that some function is linear, we mean in in the "analytical" sense, not in the sense from linear algebra. That means we call linear any function which is affine. We say that matrix X has full column rank, if the rank of X equals to the number of columns. By operator ":=" we mean either the definition or redefinition i.e. if we write $x = 3$ (or $x := 3$) and later $x := 5$, then we first defined x to be 3 and then later redefined x to have the value of 5. This notation will come handy in the section dealing with algorithms.

Standard scalar product is a function $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by $\forall x, y \in \mathbb{R}^n, x = (x_1, \ldots, x_n)^T, y = (y_1, \ldots, y_n)^T : \langle x, y \rangle = \sum_{i=1}^n x_i y_i$. This is a very special case of a well known much general definition of scalar product that we will not need in this text. We will sometimes omit the word "standard" and say just "scalar product", but we will always mean the standard scalar product. The scalar product has the following properties which we will use throughout this text. For any $x, y, z \in \mathbb{R}^n, \lambda \in \mathbb{R}$ we have

$$\begin{array}{ll} (1) & \langle x, x \rangle > 0 \iff x \neq 0 \\ (2) & \langle \lambda x, y \rangle = \lambda \langle x, y \rangle = \langle x, \lambda y \rangle \\ (3) & \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle \\ (4) & \langle z, x + y \rangle = \langle z, x \rangle + \langle z, y \rangle \end{array}$$

(2), (3) and (4) imply that scalar product is linear in both arguments and this will be used heavily.

Let W be a finite set of vectors from $\mathbb{R}^k$. The orthogonal complement of W is the set $\{\delta \in \mathbb{R}^k : \langle \delta, x \rangle = 0 \text{ for all } x \in W\}$

Let $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. The l_2 norm of x or Euclidean norm of x is defined as

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

and l_1 norm of x is defined as

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

Let $p \in \{1, 2\}$. Then l_p norm has the following properties. For any $x, y \in \mathbb{R}^n, \lambda \in \mathbb{R}$ we have

$$\begin{aligned} &(1) \quad \|x\|_p = 0 \iff x = 0 \\ &(2) \quad \|\lambda x\|_p = |\lambda| \cdot \|x\|_p \\ &(3) \quad \|x + y\|_p \leq \|x\|_p + \|y\|_p \end{aligned}$$

where (3) is called triangular inequality. In the whole text, whenever we write $\|\cdot\|$ i.e. norm without the information which norm we mean, we always mean $\|\cdot\|_1$.

Let T be any set, then for any set x , the indicator function is defined as

$$\mathbb{1}(x) = \begin{cases} 1, & \text{if } x \subseteq T \\ 0, & \text{if } x \nsubseteq T \end{cases}$$

Let $\beta, \delta \in \mathbb{R}^k$, $f : \mathbb{R}^k \to \mathbb{R}$. Then the right hand side directional derivative of f at the point β in the direction of δ is defined as

$$f'_{\delta+}(\beta) = \lim_{h \to 0+} \frac{f(\beta + h\delta) - f(\beta)}{h}$$

and left hand sided directional derivatives is defined analogically.

1 Non-statistical properties

1.1 Problem definition

For some $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$, the well-known least squares fitting problem is defined as trying to find $\hat{x} \in \mathbb{R}^m$ such that

$$\|b - A\hat{x}\|_2 = \min_{x\in\mathbb{R}^m} \|b - Ax\|_2.$$

The least absolute deviation (LAD) fitting problem could be defined analogically, simply changing the Euclidean l_2 norm to the l_1 norm. We will, however, from the start use a different notation following the tradition in the linear regression.

Definition 1. *Let $X \in \mathbb{R}^{n\times k}$, $y \in \mathbb{R}^n$. Then we call the problem of finding $\hat{\beta} \in \mathbb{R}^k$ such that*

$$\|y - X\hat{\beta}\|_1 = \min_{\beta\in\mathbb{R}^k} \|y - X\beta\|_1$$

the LAD fitting problem. Any such $\hat{\beta} \in \mathbb{R}^k$ we call a solution to LAD fitting problem or simply solution.

Therefore, the problem is to minimize the absolute deviation (AD) distance function $f : \mathbb{R}^k \to \mathbb{R}$ defined for all $\beta \in \mathbb{R}^k, \beta = (\beta_1, \ldots, \beta_k)^T$ as

$$f(\beta) = \|y - X\beta\|_1 = \sum_{i=1}^{n} |y_i - x_i\beta| = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{k} x_{ij}\beta_j|$$
$$= \sum_{i=1}^{n} |y_i - \langle x_i, \beta\rangle|.$$

where x_i is the i-th row of X.

1.2 Properties of solution

We are going to start with a Theorem describing some basic properties of the AD distance function.

Theorem 1. *Let f be AD distance function. Then f is continuous, convex and piecewise linear. Moreover, f can be written as*

$$f(\beta) = \sum_{i=1}^{2^n} l_i(\beta)\mathbb{1}_{L_i}(\beta)$$

where $l_i(\beta)$ is a liner function over L_i, $L_i \subseteq \mathbb{R}^k$, $L_i \cap L_j = \emptyset$, whenever $i \neq j$ and $\bigcup_{i=1}^{2^n} L_i = \mathbb{R}^k$.

The equation states, that f is "finitely" piecewise linear i.e. there are finitely many subsets of $\mathbb{R}^k$ over which f is liner.

Proof. Let us first prove continuity of f. Let $\beta^1, \beta^2 \in \mathbb{R}^k$. Then

$$\begin{aligned}
|f(\beta^1) - f(\beta^2)| &= \left| \sum_{i=1}^n |y_i - \langle x_i, \beta^1 \rangle| - \sum_{i=1}^n |y_i - \langle x_i, \beta^2 \rangle| \right| \\
&\leq \left| \sum_{i=1}^n (|y_i| + |\langle x_i, \beta^1 \rangle|) - \sum_{i=1}^n (|y_i| + |\langle x_i, \beta^2 \rangle|) \right| \\
&= \left| \sum_{i=1}^n (|\langle x_i, \beta^1 \rangle| - |\langle x_i, \beta^2 \rangle|) \right| \\
&\leq \sum_{i=1}^n |(|\langle x_i, \beta^1 \rangle| - |\langle x_i, \beta^2 \rangle|)| \\
&\leq \sum_{i=1}^n |\langle x_i, \beta^1 \rangle - \langle x_i, \beta^2 \rangle| = \sum_{i=1}^n |\langle x_i, \beta^1 - \beta^2 \rangle|.
\end{aligned}$$

First two inequalities follow from the triangle inequality. The last one follows from the so-called inverse triangle identity, which states that for any $x, y \in \mathbb{R}$ we have $||x| - |y|| \leq |x - y|$. We also used the linearity of the scalar product. Now, taking the limit of right hand side we have

$$\lim_{\beta^1 \to \beta^2} \sum_{i=1}^n |\langle x_i, \beta^1 - \beta^2 \rangle| = 0$$

and thus, by squeeze theorem, we also have $|f(\beta^1) - f(\beta^2)| \to 0$ as $\beta^1 \to \beta^2$. Which means that f is continuous. Now, let us prove convexity. Let again $\beta^1, \beta^2 \in \mathbb{R}^k$ and $t \in \langle 0,1 \rangle$. Then

$$\begin{aligned}
f(t\beta^1 + (1-t)\beta^2) &= \sum_{i=1}^{n} |y_i - \langle x_i, t\beta^1 + (1-t)\beta^2 \rangle| \\
&= \sum_{i=1}^{n} |y_i - t\langle x_i, \beta^1 \rangle - (1-t)\langle x_i, \beta^2 \rangle| \\
&= \sum_{i=1}^{n} |y_i t + (1-t)y_i - t\langle x_i, \beta^1 \rangle - (1-t)\langle x_i, \beta^2 \rangle| \\
&= \sum_{i=1}^{n} |t(y_i - \langle x_i, \beta^1 \rangle) + (1-t)(y_i - \langle x_i, \beta^2 \rangle)| \\
&\leq \sum_{i=1}^{n} \big(t|y_i - \langle x_i, \beta^1 \rangle| + (1-t)|y_i - \langle x_i, \beta^2 \rangle|\big) \\
&\leq \sum_{i=1}^{n} t|y_i - \langle x_i, \beta^1 \rangle| + \sum_{i=1}^{n}(1-t)|y_i - \langle x_i, \beta^2 \rangle| \\
&= tf(\beta^1) + (1-t)f(\beta^2)
\end{aligned}$$

Which proves that f is convex. Finally, let us prove the piecewise linearity. Let us consider the set $M = \{0,1\}^n$. Then $|M| = 2^n$. Let $\{m^i\}_{i=1}^{2^n}$ be a finite sequence of all subsets of M (i.e. all those subsets are vectors like $(\underbrace{1,0,0,1,\ldots,1}_{n-times})^T$). Now, let

$$L_i = \{\beta \in \mathbb{R}^k : (\forall j \in \{1,\ldots,n\} : y_j - x_j\beta > 0 \iff m^i_j = 1)\}.$$

It is easy to see, that L_i sets satisfy all the conditions given by the theorem. We define

$$l_i(\beta) = \sum_{j=1}^{n} 2(y_j - x_j\beta)m^i_j - \sum_{j=1}^{n}(y_j - x_j\beta) = \sum_{j=1}^{n}(y_j - x_j\beta)(2m^i_j - 1).$$

Clearly, for any given $i \in \{1,\ldots,n\}$, $l_i(\beta)$ is linear. We also have

$\forall \beta \in L_i : l_i(\beta) = f(\beta)$, because

$$\begin{aligned} f(\beta) &= \sum_{\{i:y_i - x_i\beta > 0\}} (y_i - x_i\beta) - \sum_{\{i:y_i - x_i\beta \leq 0\}} (y_i - x_i\beta) \\ &= 2 \cdot \left(\sum_{\{i:y_i - x_i\beta > 0\}} (y_i - x_i\beta) \right) - \sum_{\{i:y_i - x_i\beta > 0\}} (y_i - x_i\beta) - \\ &- \sum_{\{i:y_i - x_i\beta \leq 0\}} (y_i - x_i\beta \\ &= \sum_{\{i:y_i - x_i\beta > 0\}} 2(y_i - x_i\beta) - \sum_{i=1}^{n} (y_i - x_i\beta) \\ &= \sum_{i=1}^{n} 2(y_i - x_i\beta)\mathbb{1}_{\{\beta \in \mathbb{R}^k : y_i - x_i\beta > 0\}}(\beta) - \sum_{i=1}^{n} (y_i - x_i\beta) \end{aligned}$$

It is easy to see that the last is equals to $l_i(\beta)$ over any L_i. This gives

$$f(\beta) = \sum_{i=1}^{2^n} l_i(\beta)\mathbb{1}_{L_i}(\beta)$$

and the proof is complete. □

Let us first take a look at a simple one dimensional case. This will shed the first light on one of the most important characteristic of LAD fitting curve. The one dimensional LAD fitting problem can be seen as a trying to find $\beta \in \mathbb{R}$ such that a line $y = \beta x$ minimizes the sum of absolute values of residuals $r_i(\beta) = y_i - \beta x_i$. Theorem 2 will state that, unless $x_i = 0$ for all $i \in \{1, \ldots, n\}$, there exists a line such that at least one residual is zero, which means the line passes through point $[x_p, y_p]$ for some $p \in \{1, \ldots, n\}$.

Theorem 2. *Let f be the AD distance function with $k = 1$ i.e. $f : \mathbb{R} \to \mathbb{R}$ and $f(\beta) = \sum_{i=1}^{n} |y_i - \beta x_i|$ where $\forall i \in \{1, \ldots, n\} : y_i, x_i \in \mathbb{R}$. Let for some $i \in \{1, \ldots, n\}$ be $x_i \neq 0$. Then there exist an index $p \in \{1, \ldots, n\}$ such that*

$$\frac{y_p}{x_p} = \arg\min_{\beta \in \mathbb{R}} f(\beta)$$

Proof. Firstly, assume that $x_i \neq 0$ for all $i \in \{1, \dots, n\}$. Thanks to the absolute values we have

$$\lim_{\beta \to \infty} f(\beta) = \lim_{\beta \to -\infty} f(\beta) = +\infty. \tag{1}$$

Let $c_i = \frac{y_i}{x_i}$. Without a loss of generality we can assume $c_1 \leq c_2 \leq \dots \leq c_n$ (as we could just rename them). Let $K = \langle c_1 - 42, c_n + 42 \rangle$ be a closed interval. Then K is a compact set. And since f is by Theorem 1 continuous, by the extreme value theorem we have

$$\exists \hat{\beta} : \hat{\beta} = \arg\min_{\beta \in K} f(\beta).$$

Combining that f is linear on $(-\infty, c_1)$ and (1) gives that f in decreasing on $(-\infty, c_1)$ which means

$$\forall \beta \in (-\infty, c_1 - 42) : f(c_1 - 16) < f(\beta).$$

Analogically, we have

$$\forall \beta \in (c_n + 42, +\infty) : f(c_n + 16) < f(\beta).$$

Therefore $\hat{\beta}$ is a minimizer of f not only on K but also on $\mathbb{R}$. Now, it is easy to see that f is differentiable everywhere where $y_p - \beta x_p \neq 0$ for some p. If $f'(\hat{\beta})$ does not exist, we must have $y_p - \beta x_p = 0$ for some p and thus $\hat{\beta} = \frac{y_p}{x_p}$. If $f'(\hat{\beta})$ does exist, it must be equal to zero. In that case we have $\hat{\beta} \in \langle c_p, c_{p+1} \rangle$ for some p. Because f is linear over (c_p, c_{p+1}) it must have a constant derivative which is zero. Thus f must be constant over this interval. Because f is continuous, we have $f(\hat{\beta}) = f(c_p)$ and thus $c_p = \frac{y_p}{x_p}$ is also a minimizer of f.

Now, if $M \subsetneq \{1, \dots, n\}$ and $x_i = 0$ for all $i \in M$ then we can find the minimizer of $\widetilde{f} = \sum_{i \in \{1,\dots,n\} \setminus M} |y_i - \beta x_i|$ which is just f minus a constant i.e. the minimizer of $\widetilde{f}$ is also a minimizer of f. And by the first part of a proof there is a minimizer of $\widetilde{f}$ in the form $\frac{y_p}{x_p}$ for some p. □

Note that the assumption "$\exists i \in \{1, \dots, n\} : x_i \neq 0$" is equivalent to "column rank of X is 1" where $X = (x_1, \dots, x_n)^T$. Analogical assumption will appear in Theorem 4, which will be a direct generalization of this theorem. The following Lemma will be used in a proof of Theorem 3.

Lemma 1. *Let A be a real matrix with a full column rank. Then matrix A^TA is invertible.*

Proof. Showing the invertibility of A^TA is equivalent to showing that if $A^TAx = 0$ for some vector x, then x is a zero vector. If $A^TAx = 0$ then also $x^TA^TAx = 0$. Thus we can write $0 = x^TA^TAx = (Ax)^T(Ax) = \langle Ax, Ax\rangle = \|Ax\|^2$ which gives $Ax = 0$ and since A has a full column rank, x must be a zero vector. □

We are now able to show the existence of solution. Firstly, note that if rank of X is zero, then X is a zero matrix and every fit is optimal. We will thus always assume that $\operatorname{rank}(X) > 0$.

Theorem 3. *Given LAD fitting problem, there always exists its solution.*

Proof. Let us first assume that $\operatorname{rank}(X) = k$. Thus $k \leq n$ and there exists k rows of X, say $x_1, \dots, x_r$ which are linearly independent. Then no non-zero $\beta \in \mathbb{R}^k$ can be orthogonal to all $x_1, \dots, x_r$ i.e. some scalar product of β and x_i has to be greater than zero for some i. Thus

$$\exists\delta \in \mathbb{R}, \delta > 0, \forall\beta \in \mathbb{R}^k, \|\beta\| = 1 : \sum_{i=1}^{n} |\langle x_i, \beta\rangle| \geq \delta.$$

Now, let $\beta^* \in \mathbb{R}^k, \beta^* \neq 0 \in \mathbb{R}^k$. Then let $\beta = \frac{\beta^*}{\|\beta^*\|}$. So we have $\|\beta\| = 1$ and also $\delta \leq \sum_{i=1}^n |\langle x_i, \beta\rangle| = \sum_{i=1}^n |\langle x_i, \frac{\beta^*}{\|\beta^*\|}\rangle| = \frac{1}{\|\beta^*\|}\sum_{i=1}^n |\langle x_i, \beta^*\rangle|$ which means

$$\exists\delta \in \mathbb{R}, \delta > 0, \forall\beta \in \mathbb{R}^k, \beta \neq 0 : \sum_{i=1}^{n} |\langle x_i, \beta\rangle| \geq \delta\|\beta\|.$$

Next, $\forall\beta \in \mathbb{R}^k, \beta \neq 0$ we can write

$$\begin{aligned} f(\beta) + f(0) = \sum_{i=1}^{n} |y_i - x_i\beta| + \sum_{i=1}^{n} |y_i| &= \sum_{i=1}^{n} (|y_i - x_i\beta| + |-y_i|) \\ \geq \sum_{i=1}^{n} |x_i\beta| = \sum_{i=1}^{n} |\langle x_i, \beta\rangle| &\geq \delta\|\beta\| \end{aligned}$$

If we take $\beta \in \mathbb{R}^k$ such that $\|\beta\| > \frac{2f(0)}{\delta}$ we get $f(\beta) > f(0)$. Let $K = \{\beta \in \mathbb{R}^k : \|\beta\| \leq \frac{2f(0)}{\delta}\}$. Then there is no minimizer of f outside K and, since K is compact and f is continuous, there exists a minimizer $\hat{\beta} \in K$ of f on K which therefore must be the minimizer of f on $\mathbb{R}^k$.

Now, suppose that $r = \text{rank}(X) < k$. Let $x_1, \ldots, x_r$ be linearly independent columns of X and $x_{r+1}, \ldots, x_n$ other columns of X. Consider two matrices $X_1 = (x_1, \ldots, x_r)$ and $X_2 = (x_{r+1}, \ldots, x_n)$ and two functions $f_1(\beta) = \|y - X_1\beta\|_1$ and $f_2(\beta) = \|y - X_2\beta\|_1$. By the first part of the proof there exists a minimizer $\hat{\beta} \in \mathbb{R}^r$ of f_1. Next, we find a matrix B such that $X_1 B = X_2$. This can be done as $B = (X_1^T X_1)^{-1} X_1^T X_2$ provided the inverse exists, which is true by Lemma 1. Without loss of generality assume that first r columns of X are linearly independent (otherwise we would just add an index which would store information which columns are independent). Then for all $\beta \in \mathbb{R}^k, \beta = (a^T, b^T)^T, a = (a_1, \ldots, a_r)^T, b = (b_1, \ldots, b_{k-r})^T$ we can write

$$\begin{aligned} f(\beta) = \|y - X\beta\|_1 &= \|y - X_1 a - X_2 b\|_1 = \|y - X_1 a - X_1 B b\|_1 \\ &= \|y - X_1(a - Bb)\|_1 = f_1(a - Bb) \end{aligned}$$

Hence minimizing f is equivalent to finding a and b such that $f_1(a - Bb)$ is minimal. If $\beta^* = (\hat{\beta}_1, \ldots, \hat{\beta}_r, 0, \ldots, 0) \in \mathbb{R}^k$, then $f(\beta^*) = f_1(\hat{\beta})$. And since $\hat{\beta}$ is the minimizer of f_1, $\beta^* \in \mathbb{R}^k$ is the minimizer of f.

□

We have just shown the existence of the solution to the LAD fitting problem. However, the proof was not constructive i.e. we do not have a formula for solution. The problem of finding the solution is much harder for LAD fitting problem then, for instance, for the Least Squares fitting problem and we will discuss this in more detail in the next section. The following theorem is fundamental for finding a solution. It is a direct generalization of Theorem 2

Theorem 4. *Let $r > 0$ be the rank of X. Then there exists a solution to the LAD fitting problem $\hat{\beta}$ such that there exist (at least) r different indexes $i_1, \ldots, i_r \in \{1, \ldots, n\}$ such that $y_{i_j} - x_{i_j}\hat{\beta} = 0$, $j = 1, \ldots, r$.*

Proof. Let $\hat{\beta}$ be any solution to the LAD fitting problem (which we know that exists by Theorem 3) and $p < r$ be the number of zero residuals. Without loss of generality let $\forall i \in \{1, \ldots, r\} : y_i -$

$x_i\hat{\beta} = y_i - \langle x_i, \hat{\beta}\rangle = 0$ i.e. the first r residuals are zero. Because $p < \operatorname{rank}(X)$ there exists vector $v \in \mathbb{R}^k$ such that $\forall i \in \{1, \dots, r\}$: $\langle x_i, v\rangle = 0$ and $\exists i \in \{r+1, \dots, n\} : \langle x_i, v\rangle \neq 0$. Let $\varphi(t) = \hat{\beta} + tv$. Then

$$\begin{aligned}
f(\varphi(t)) &= \sum_{i=1}^{n} |y_i - \langle x_i, \varphi(t)\rangle| = \sum_{i=1}^{n} |y_i - \langle x_i, \hat{\beta} + tv\rangle| \\
&= \sum_{i=1}^{n} |y_i - \langle x_i, \hat{\beta}\rangle - t\langle x_i, v\rangle| = \sum_{i=r+1}^{n} |y_i - \langle x_i, \hat{\beta}\rangle - t\langle x_i, v\rangle| \\
&= \sum_{i=r+1}^{n} |y_i^* - tx_i^*|
\end{aligned}$$

where $y_i^* = y_i - \langle x_i, \hat{\beta}\rangle$ and $x_i^* = \langle x_i, v\rangle$. Because we know that $\exists i \in \{r+1, \dots, n\} : x_i^* = \langle x_i, v\rangle \neq 0$, we can use Theorem 2 which gives the existence of index $j \in \{r+1, \dots, n\}$ such that $\hat{t} = \frac{y_j^*}{x_j^*}$ is the minimizer of $f(\varphi(t))$. Because it is a minimizer, we must have $f(\varphi(\hat{t})) \leq f(\varphi(0)) = f(\hat{\beta})$ and because $\hat{\beta}$ is a minimizer of f, $\varphi(\hat{t})$ is also a minimizer of f. And because $\varphi(\hat{t})$ is such a minimizer of f which has at least $r+1$ residuals equal to zero (residuals $\{1, \dots, r, j\}$ and because this argument can be used as long as the amount of zero residuals is less then r, the proof is complete. □

1.3 Directional derivatives

In this section, our goal will be to compute directional derivatives of the AD distance function f. Let $r_i = y_i - \langle \beta, x_i\rangle$ and $v_i = \langle \delta, x_i\rangle$ and let $h \in (0, \min\{\frac{|r_i|}{|v_i|}, i \in \{1, \dots, n\}\}$. If $\frac{r_i}{v_i} > 0$ for some i, then

$$\begin{aligned}
|r_i - hv_i| &= |v_i| \left| \frac{r_i}{|v_i|} - \frac{v_i}{|v_i|} h \right| = \\
&= \begin{cases} |v_i| \left| \frac{r_i}{v_i} - h \right| = |v_i| \left(\frac{r_i}{v_i} - h \right), & \text{if } v_i > 0, r_i > 0 \\ |v_i| \left| -\left(\frac{r_i}{v_i} - h \right) \right| = |v_i| \left(\frac{r_i}{v_i} - h \right), & \text{if } v_i < 0, r_i < 0 \end{cases}
\end{aligned}$$

thus $|r_i - hv_i| = |v_i| \left(\frac{r_i}{v_i} - h \right)$ whenever $\frac{r_i}{v_i} > 0$. Analogically, we get that $|r_i - hv_i| = -|v_i| \left(\frac{r_i}{v_i} - h \right)$ whenever $\frac{r_i}{v_i} < 0$.

Let $V = \{i : v_i = 0\}, Z = \{i \notin V : r_i = 0\}, P = \{i \notin V : \frac{r_i}{v_i} > 0\}$ and $N = \{i \notin V : \frac{r_i}{v_i} < 0\}$. Clearly $V \cup Z \cup P \cup N = \{1, \ldots, n\}$ and V, Z, P, N are pairwise disjoint. We can thus write

$$\begin{aligned} f(\beta + h\delta) &= \sum_{i=1}^{n} |y_i - \langle \beta + h\delta, x_i \rangle| = \sum_{i=1}^{n} |y_i - \langle \beta, x_i \rangle - h\langle \delta, x_i \rangle| \\ &= \sum_{i=1}^{n} |r_i - hv_i| \\ &= \sum_V |r_i| + |h| \sum_Z |v_i| + \sum_P |v_i|(\frac{r_i}{v_i} - h) - \sum_N |v_i|(\frac{r_i}{v_i} - h). \end{aligned}$$

And also

$$\begin{aligned} f(\beta + h\delta) - f(\beta) =& \sum_V |r_i| + |h| \sum_Z |v_i| + \sum_P |v_i|(\frac{r_i}{v_i} - h) \\ & - \sum_N |v_i|(\frac{r_i}{v_i} - h) \\ & - \sum_V |r_i| - \sum_Z |r_i| - \sum_P |r_i| - \sum_N |r_i| \\ =& |h| \sum_Z |v_i| + \sum_P (\operatorname{sgn}(v_i) \cdot r_i - h|v_i|) \\ & - \sum_N (\operatorname{sgn}(v_i) \cdot r_i - h|v_i|) \\ & - \sum_Z 0 - \sum_P |r_i| - \sum_N |r_i| \\ =& |h| \sum_Z |v_i| - h \sum_P |v_i| + h \sum_N |v_i| \\ & + \sum_P (\operatorname{sgn}(v_i) \cdot r_i - |r_i|) \\ & + \sum_N (-\operatorname{sgn}(v_i) \cdot r_i - |r_i|). \end{aligned}$$

If $i \in P$ and $v_i > 0, r_i > 0$, then $\operatorname{sgn}(v_i) \cdot r_i - |r_i| = r_i - r_i = 0$.
If $i \in P$ and $v_i < 0, r_i < 0$, then $\operatorname{sgn}(v_i) \cdot r_i - |r_i| = -r_i + r_i = 0$.
If $i \in N$ and $v_i > 0, r_i < 0$, then $-\operatorname{sgn}(v_i) \cdot r_i - |r_i| = -r_i + r_i = 0$.
If $i \in N$ and $v_i < 0, r_i > 0$, then $-\operatorname{sgn}(v_i) \cdot r_i - |r_i| = -(-r_i) - r_i = 0$.

Hence $\operatorname{sgn}(v_i) \cdot r_i - |r_i| = 0$ in general and thus

$$f(\beta + h\delta) - f(\beta) = |h| \sum_Z |v_i| - h \sum_P |v_i| + h \sum_N |v_i|.$$

Now, we can easily get one sided directional derivatives in the direction of δ. We have

$$\begin{aligned} f'_{\delta^+}(\beta) &= \lim_{h \to 0+} \frac{f(\beta + h\delta) - f(\beta)}{h} \\ &= \lim_{h \to 0+} \frac{|h| \sum_Z |v_i| - h \sum_P |v_i| + h \sum_N |v_i|}{h} \\ &= \sum_Z |v_i| - \sum_P |v_i| + \sum_N |v_i| \end{aligned}$$

and

$$\begin{aligned} f'_{\delta^-}(\beta) &= \lim_{h \to 0+} \frac{f(\beta - h\delta) - f(\beta)}{-h} \\ &= \lim_{h \to 0+} \frac{|h| \sum_Z |v_i| + h \sum_P |v_i| - h \sum_N |v_i|}{h} \\ &= \sum_Z |v_i| + \sum_P |v_i| - \sum_N |v_i| \end{aligned}$$

There identities will be needed later. Also, we will need another expression for directional derivatives which we will calculate now. We know that $f(\beta + h\delta) = \sum_{i=1}^n |r_i(\beta) - h\langle \delta, x_i \rangle|$. Let $h > 0$ be small enough, that $\operatorname{sgn}(r_i(\beta) - h\langle \delta, x_i \rangle) = \operatorname{sgn}(r_i)$ for all i. Then

$$\begin{aligned} f(\beta + h\delta) - f(\beta) &= \sum_{i \in Z_\beta} |0 - h\langle \delta, x_i \rangle| + \sum_{i \notin Z_\beta} |r_i - h\langle \delta, x_i \rangle| \\ &\quad - \sum_{i \in Z_\beta} |0| - \sum_{i \notin Z_\beta} |r_i| \\ &= h \sum_{i \in Z_\beta} |\langle \delta, x_i \rangle| \\ &\quad + \sum_{i \notin Z_\beta} [\operatorname{sgn}(r_i)(r_i - h\langle \delta, x_i \rangle) - \operatorname{sgn}(r_i) r_i] \\ &= h \sum_{i \in Z_\beta} |\langle \delta, x_i \rangle| - h \sum_{i \notin Z_\beta} \operatorname{sgn}(r_i) \langle \delta, x_i \rangle \end{aligned}$$

And thus

$$f'_{\delta+}(\beta) = \sum_{i \in Z_\beta} |\langle \delta, x_i \rangle| - \sum_{i \notin Z_\beta} \operatorname{sgn}(r_i) \langle \delta, x_i \rangle.$$

Also, note that if $e^j \in \mathbb{R}^k$ is the j-th unit coordinate of $\mathbb{R}^k$ i.e. $e^j = (0, \dots, 0, 1, 0, \dots, 0)^T$ where 1 is the j-th component, we have

$$f'_{e^j+}(\beta) = \sum_{i \in Z_\beta} |x_{ij}| - \sum_{i \notin Z_\beta} \operatorname{sgn}(r_i) x_{ij}.$$

In the next, we will use the following notation. Let $\beta \in \mathbb{R}^k$. Then $Z_\beta = \{i : r_i(\beta) = 0\}$. We say, that β is an extreme point, if $\operatorname{span}(\{x_i : i \in Z_\beta\}) = \mathbb{R}^k$ (and thus $|Z_\beta| \geq k$). We say, that β is a degenerate point, if $|Z_\beta| > k$.

Theorem 5. *Let $\beta \in \mathbb{R}^k$. Then β is a minimizer of LAD distance function if and only if $\forall \delta \in \mathbb{R}^k : f'_{\delta+}(\beta) \geq 0$ i.e.*

$$\forall \delta \in \mathbb{R}^k : \sum_P |\langle \delta, x_i \rangle| - \sum_N |\langle \delta, x_i \rangle| \leq \sum_Z |\langle \delta, x_i \rangle|,$$

where $V = \{i : \langle \delta, x_i \rangle = 0\}, Z = \{i \notin V : \frac{y_i - \langle \beta, x_i \rangle}{\langle \delta, x_i \rangle} = 0\}, P = \{i \notin V : \frac{y_i - \langle \beta, x_i \rangle}{\langle \delta, x_i \rangle} > 0\}$ and $N = \{i \notin V : \frac{y_i - \langle \beta, x_i \rangle}{\langle \delta, x_i \rangle} < 0\}$. Moreover, β is a unique minimizer if and only if the inequality above is strict.

The proof of the first part of this theorem can be found in any advanced calculus book, while the “i.e.” part follows from the equation for direction derivatives presented above. In practice, it is of course impossible to check directional derivatives for all directions. But if we a have non-degenerate extreme point, it is much easier to decide whether it is a minimizer. The following theorem does just that.

Theorem 6. *Let $\beta \in \mathbb{R}^k$ be a non-degenerate extreme point. Then β is a minimizer of LAD distance function if and only if $\forall i \in Z_\beta$: $f'_{\delta_i+}(\beta) \geq 0$, for all $\delta_i \in \mathbb{R}^k, \delta_i \neq 0$ satisfying $\forall j \in Z_\beta, j \neq i$: $\langle \delta_i, x_j \rangle = 0$. Moreover, β is a unique minimizer if and only if the inequality above is strict.*

Proof of this Theorem can be found in Bloomfield and Steiger (1983). A few notes about the theorem should be said. Because β

is an extreme point, we have $\mathrm{span}(\{x_i : i \in Z_\beta\}) = \mathbb{R}^k$. Because it is non-degenerate, we have $|Z_\beta| = k$. Thus, δ_i is the direction of a "line" in $\mathbb{R}^k$ which is an orthogonal compliment of $\{x_i : i \in Z_\beta, i \neq j\}$.

2 Finding a solution

In this section we will discuss how the solution to the LAD fitting problem is found. Since AD distance function is not smooth i.e. it does not have continuous partial derivatives, no useful formula for solution can be written down. This is the first obvious difference between LAD and the Least Squares To find a solution to the LAD fitting problem we will need to use one of iterated algorithms.

Theorem 4 implies that under the assumption of X having the full column rank k, to find a solution to the LAD fitting problem we could do the following: for every subset $S = \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$, find $\beta \in \mathbb{R}^k$ such that $\forall j \in S : r_j(\beta) = 0$. Then the solution of LAD fitting problem $\hat{\beta}$ is such β from those computed which has minimal $f(\beta)$. This would lead to an algorithm which would have to solve $\binom{n}{k}$ systems of k linear equations. Computations needed for this algorithm grow too fast as n grows larger i.e. this algorithm cannot be uses efficiently. However, the property of LAD fitted line given by Theorem 4 is used by all the most used algorithms for finding the solution to the LAD fitting problem, only all those algorithms are using this property in a more clever way. In this section, we will describe one of those algorithms.

All the best-known algorithms for finding LAD fit are iterated and all are using Theorem 4. Firstly (in the first phase) they find some extreme point $c_1 \in \mathbb{R}^k$. Then, they iterate towards optimum by a sequence of extreme fits $c_2, c_3, ...$ which satisfies $f(c_1) \geq f(c_2) \geq f(c_3) \geq \ldots$. Also, two subsequent c_j, c_{j+1} only differ at one residual. More precisely, if $\forall i \in R_j \subset \{1, \ldots, n\} : r_i(c_j) = 0$ and $\forall i \in R_{j+1} \subset \{1, \ldots, n\} : r_i(c_{j+1}) = 0$ (i.e. R_j is a set of indexes of those residual, which are zero for c_j fit) then algorithms satisfies that by going from extreme fit c_j to c_{j+1} and thus from R_j to R_{j+1}, only one element in R_j is replaced by one other element in $R_{j+1} \setminus R_j$. All the most used algorithms differ only at three things. Firstly, how they find the first extreme fit c_1. Secondly, how they chose which element in R_j to replace. Thirdly, how they chose the replacement from $R_{j+1} \setminus R_j$. We will now take a look at one of these algorithms, known as the Barrodale-Roberts (BR) algorithm

2.1 The Barrodale-Roberts Algorithm

The BR algorithms has two phases. The first phase has a goal to find some extreme point. We start at $c_0 = 0 \in \mathbb{R}^k$. First phase has at most k steps. It produces a sequence $c_1, c_2, \ldots, c_k$ where c_k is the extreme point. Throughout the algorithm, we have set $\{\delta^1, \ldots, \delta^k\}$ of "edge" directions. At the start, we let $\forall j : \delta^j = e^j$. The first step goes as follows. We compute all the directional derivatives i.e. for all j we compute

$$f'_{\delta^j+}(0) = \sum_{i \in Z} |x_{ij}| - \sum_{i \notin Z} \operatorname{sgn}(r_i(0)) x_{ij}$$

and also compute

$$f'_{-\delta^j+}(0) = \sum_{i \in Z} |x_{ij}| + \sum_{i \notin Z} \operatorname{sgn}(r_i(0)) x_{ij},$$

where $Z = \{i : y_i = 0\}$. Because f is convex, for any j, we have $f'_{\delta^j+}(0)$ and $f'_{-\delta^j+}(0)$ either both non-negative, or they have opposite signs. We will assume there is at least one negative direction derivative. The case when all directional derivatives are non negative will be asserted later. We then choose such $p \in \{1, \ldots, k\}$ for which $f'_{\delta^j+}(0)$ (or $f'_{-\delta^j+}(0)$) is the most negative. We then call δ^p active and we let $\tau_1 = p$ for future use. We then move from c_0 to c_1 by minimizing f along $t\delta^p$. We thus need to find

$$\operatorname*{arg\,min}_{t \in \mathbb{R}} f(t\delta^p) = \operatorname*{arg\,min}_{t \in \mathbb{R}} \sum_{i=1}^{n} |y_i - t\langle \delta^p, x_i \rangle| = \operatorname*{arg\,min}_{t \in \mathbb{R}} \sum_{i=1}^{n} |y_i - t x_{ip}|.$$

We already know (Theorem 2) that solution is $t = t_q = \frac{y_q}{x_{qp}}$ for some $q \in \{1, \ldots, n\}$. We should state that finding such q is equivalent to finding weighted median which is a basic problem solved using linear programming. We then let $c_1 = c_0 + t_q\delta^p = 0 + t_q\delta^p = t_q\delta^p$ and we have $r_q(c_1) = y_q - \langle c_1, x_q \rangle = 0$. We let $\sigma_1 = q$ for future use.

Now, we would like to repeat this step i.e. to find another "steepest" direction. But if we would do that just as above, we would not have $r_q(c_2) = 0$. We will need to modify our directions $\delta^1, \ldots, \delta^k$. We want

$$r_q(c_1 + t\delta^j) = y_q - \langle c_1, x_q \rangle - t\langle \delta^j, x_q \rangle = 0 \tag{1}$$

to hold for all j such that $j \neq p$ and all $t \in \mathbb{R}$. To do this we find $\delta^1, \ldots, \delta^k$ such that

$$\langle \delta^j, x_q \rangle = \begin{cases} 1, & \text{if } j = p \\ 0, & \text{if } j \neq p \end{cases}$$

which we do by

$$\delta^j := \begin{cases} \frac{\delta^p}{x_{qp}}, & \text{if } j = p \\ \delta^j - \frac{x_{qj}}{x_{qp}} \delta^p, & \text{if } j \neq p \end{cases}$$

This choice assures (1) holds for all $j, j \neq p$ and also assures that $\delta^1, \ldots, \delta^k$ are still linearly independent (which is actually needed for the next Theorem 7 to hold). In the next step of the first phase we basically repeat step one. But when finding the steepest edge using the directional derivatives, we will not include direction δ^p.

By induction, let us describe $(j+1)$-th step of the first phase. At the start of this step, we have $\{\sigma_1, \ldots, \sigma_j\}$, the set of indexes of those residuals, which are zero for c_j, i.e. $r_{\sigma_i}(c_j) = 0, i = 1, \ldots, j$. Also, we have $\{\tau_1, \ldots, \tau_j\}$, the set of indexes of those $\delta^1, \ldots, \delta^k$ which have already been used i.e. we have indexes of the active directions. Finally, we have linearly independent directions $\delta^1, \ldots, \delta^k$ satisfying

$$\forall j \in \{1, \ldots, k\} \; \forall l \in \{1, \ldots, j\} : \langle \delta^j, x_{\sigma_l} \rangle = \begin{cases} 1, & \text{if } j = \sigma_l \\ 0, & \text{if } j \neq \sigma_l \end{cases} \tag{2}$$

At the start of the $(j+1)$-th step we will, for all i such that $i \notin \{\tau_1, \ldots, \tau_j\}$ calculate

$$f'_{\delta_i \pm}(c_j) = \sum_{l \in Z_{c_j}} |\langle \delta_i, x_l \rangle| \mp \sum_{l \notin Z_{c_j}} \operatorname{sgn}(r_l(c_j)) \langle \delta_i, x_l \rangle. \tag{3}$$

Because f is convex, for any i we have $f'_{\delta^i +}(c_j)$ and $f'_{-\delta^i +}(c_j)$ either both non-negative, or they have opposite signs.

We will now assert the case that $\forall i \notin \{\tau_1, \ldots, \tau_j\} : f'_{\delta^i +}(c_j) \geq 0, f'_{-\delta^i +}(c_j) \geq 0$. In this case we are actually done, as says the following theorem proof of which can be found in Bloomfield & Steiger (1983).

Theorem 7. *In the BR algorithm, if at any step we have* $\forall i \notin \{\tau_1, \ldots, \tau_j\} : f'_{\delta^i +}(c_j) \geq 0, f'_{-\delta^i +}(c_j) \geq 0$*, then* c_j *is the minimizer of* f.

Otherwise we can choose p such that the directional derivative is most negative along line $t\delta^p$. We let $c_{j+1} = c_j + t_q \delta^p$, where

$$t_q = \arg\min_{t \in \mathbb{R}} \sum_{i=1}^{n} |y_i - \langle c_j, x_i \rangle - t\langle \delta^p, x_i \rangle| = \frac{y_q - \langle c_j, x_q \rangle}{\langle \delta_p, x_q \rangle}$$

for some q. Again, we want (2) to hold in the next step. We thus let

$$\delta^i := \begin{cases} \frac{\delta^p}{\langle \delta^p, x_q \rangle}, & \text{if } i = p \\ \delta^j - \frac{\langle \delta^j, x_q \rangle}{\langle \delta^p, x_q \rangle}\delta^p, & \text{if } i \neq p \end{cases} \tag{4}$$

and move to another step. The first phase is finished, when we compute c_j and $j = k$. At that point, c_k is the extreme fit.

The goal of the second phase is to find the minimum of f. In every step, we compute directional derivatives $f'_{\delta_i \pm}(c_j)$ (where $j \geq k$) just as in the (3), but we do that for all i (not just for $i \notin \{\tau_1, \ldots, \tau_j\}$). If all the directional derivatives are non-negative at c_j, c_j is minimizer of f by Theorem 7. Otherwise, as before, we move to $c_{j+1} = c_j + t_q \delta^p$, where δ^p is the direction of steepest downhill edge and

$$t_q = \frac{y_q - \langle c_j, x_q \rangle}{\langle \delta^p, x_q \rangle}$$

is the minimizer of $f(t\delta^p) = \sum_{i=1}^{n} |y_i - \langle c_j, x_i \rangle - t\langle \delta^p, x_q \rangle|$. This changes one zero residual for another i.e. c_{j+1} is still an extreme point. Next we update $\{\delta^1, \ldots, \delta^k\}$ as in (4) and repeat. The algorithm must find solution by theorems from the fist section. There are only finitely many extreme points, one of them is minimizer and BR algorithm always moves closer to optimum (i.e. downhill). If it cannot move downhill, i.e. when all the direction derivatives in the direction of the edge vectors are non-negative, we are at the optimum by Theorem 7. We should point out, that the BR algorithm can be rewritten using linear programming, which is the usual way to treat LAD fitting problem as, for instance, in Brennan and Seiford (1987).

2.2 Iteratively reweighted least squares

In this section, we will briefly discuss the method of finding an approximate solution to the LAD fitting problem. This method is called Iteratively reweighted least squares (IRLS for short). The Barrodale-Roberts Algorithm and other algorithms for finding LAD fit using linear programming are superior to IRLS in a way that they give an exact solution whereas IRLS produces only an approximate. The advantage of IRLS is its much easier implementation and that it can be used to more general problem then just LAD fitting line. Here, we will only show its usage for LAD fits.

Theorem 8. *Let X and y be as in LAD fitting problem and let $\forall i \in \{1, \dots, n\} : w_i \in \mathbb{R}, w_i > 0$. Then*

$$\arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} w_i (y_i - x_i \beta)^2 = (X^T W X)^{-1} X^T W y,$$

where $W \in \mathbb{R}^{n \times n}, W = (w_{ij})_{i,j=1}^{n}$ and

$$w_{ij} = \begin{cases} w_i, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

Proof. Because $g(\beta) \in C^2(\mathbb{R}^k)$ we can find minimum by standard analytical methods. We have

$$\frac{\partial g}{\partial \beta_1}(\beta) = -2 \sum_{i=1}^{n} w_i x_{i1} (y_i - \sum_{j=1}^{k} x_{ij} \beta_j)$$

Taking all partial derivatives and setting them to zero we get (after rearrangement) system of equations

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{k} w_i x_{i1} x_{ij} \beta_j &= \sum_{i=1}^{n} w_i x_{i1} y_i \\ \sum_{i=1}^{n} \sum_{j=1}^{k} w_i x_{i2} x_{ij} \beta_j &= \sum_{i=1}^{n} w_i x_{i2} y_i \\ &\vdots \\ \sum_{i=1}^{n} \sum_{j=1}^{k} w_i x_{ik} x_{ij} \beta_j &= \sum_{i=1}^{n} w_i x_{ik} y_i \end{aligned}$$

witch is equivalent to

$$X^T W X \beta = X^T W y.$$

Invertibility of $X^T W X$ can be shown similarly as in Lemma 1 and fact that $(X^T W X)^{-1} X^T W y$ is indeed a minimum is a matter of technical checking of the sufficient condition for minimum. □

The IRLS algorithms repeatedly solves

$$\operatorname*{arg\,min}_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} w_i^{(t)} (y_i - x_i \beta)^2$$

where t is index of step of the algorithm. It produces sequence $\beta^0, \beta^1, \ldots$ which can be shown to converge to $\operatorname{arg\,min}_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} |y_i - x_i \beta|$ i.e. to the LAD fit, where the only condition for convergence is the existence of inverse to $X^T W X$. At the start we set $w_i^{(0)} = 1, i = 1, \ldots, n$ and compute

$$\beta^0 = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} w_i^{(0)} (y_i - x_i \beta)^2.$$

In the $(j+1)$-th step, we let

$$w_i^{(j+1)} = \frac{1}{|y_i - \langle \beta^{(j)}, x_i \rangle|}$$
$$\beta^{j+1} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} w_i^{(j+1)} (y_i - x_i \beta)^2$$

As stated above, finding approximate LAD fits is only one of the applications of IRLS and using it for LAD fit actually requires additional treatment. Namely $w_i^{(j+1)} = \frac{1}{|y_i - \langle \beta^{(j)}, x_i \rangle|}$ might be undefined (division by zero). We could, for instance, use

$$w_i^{(j+1)} = \frac{1}{\max\{\varepsilon, |y_i - \langle \beta^{(j)}, x_i \rangle|\}}$$

for some $\varepsilon > 0$. Or as discussed by Schlossmacher (1973) we could delete those data points, which have close to zero residuals. In this case we would stop the algorithm when all residuals corresponding to not deleted points did not change enough from the previous iteration.

3 Statistical properties

Let $y \in \mathbb{R}, x, \beta \in \mathbb{R}^k$ and ϵ random variable. In this section, we will use the Least Absolute Deviations to estimate the linear model

$$y = \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon = \langle \beta, x \rangle + \varepsilon \tag{M}$$

with the given random sample $(y_i, x_i)_{i=1}^n, x_i = (x_{i1}, \ldots, x_{ik})$, where $x_{i1}, \ldots, x_{ik} \in \mathbb{R}$ are observed independent variables, $\beta_1, \ldots, \beta_k \in \mathbb{R}$ are non random unobserved coefficients, ϵ_i are unobserved random disturbance variables (also called error variables or errors for short) and y_i are observed dependant variables. We sometimes write model (M) as

$$y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = \langle \beta, x_i \rangle + \varepsilon_i. \tag{M}$$

Our object will be to estimate $\beta = (\beta_1, \ldots, \beta_k)^T$. Let us start by defining the Least Absolute Deviation (LAD) estimator of β in model (M).

Definition 2. *The LAD estimator of β in model (M) is defined as*

$$\hat{\beta}_{LAD} = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n |y_i - \langle \beta, x_i \rangle|.$$

Thus all the mathematical background from the previous parts of this text apply.

3.1 MLE under Laplace errors

In this part, we will show that if the errors ε_i are drawn from Laplace distribution with median $\mu = 0$, then the LAD estimator is the Maximum likelihood estimator for our model (M).

Definition 3. *Random variable has a Laplace(μ, s) distribution, if its probability density function is*

$$f(x) = \frac{1}{\sqrt{2}s} \exp\left(-\frac{\sqrt{2}|x-\mu|}{s}\right), \ x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ is a location parameter and $s > 0$ is a scale parameter.

It is not hard to show that if $X \sim$ Laplace(μ, s) then Median(X) $= \mathrm{E}X = \mu$.

Theorem 9. *If in the model (M) it holds that* $\forall i \in \{1, \dots, n\} : \varepsilon_i \sim$ *Laplace*$(0, s)$*, then LAD estimator of* β *of model (M) is equal to the Maximum likelihood estimator of* β *of model (M) i.e.* $\hat{\beta}_{LAD} = \hat{\beta}_{MLE}$.

Proof. If we have $\forall i \in \{1, \dots, n\} : \varepsilon_i = y_i - x_i\beta \sim Laplace(\mu, s)$, the likelihood function is

$$\begin{aligned} L(\beta, \mu, s) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2}s} \exp\left(-\frac{\sqrt{2}|y_i - x_i\beta - \mu|}{s}\right) \\ &= \frac{1}{(\sqrt{2}s)^n} \exp\left(\sum_{i=1}^{n}\left(-\frac{\sqrt{2}}{s}|y_i - x_i\beta - \mu|\right)\right) \\ &= \frac{1}{(\sqrt{2}s)^n} \exp\left(-\frac{\sqrt{2}}{s}\sum_{i=1}^{n}|y_i - x_i\beta - \mu|\right) \end{aligned}$$

where the exponent term is maximized whenever $\sum_{i=1}^{n} |y_i - x_i\beta - \mu|$ is minimized. Thus, if $\mu = 0$, we have

$$\hat{\beta}_{MLE} = \underset{\beta \in \mathbb{R}^k}{\arg\min} \sum_{i=1}^{n} |y_i - x_i\beta|$$

and therefore $\hat{\beta}_{MLE} = \hat{\beta}_{LAD}$. □

From the proof we see that $\hat{\beta}_{MLE}$ and $\hat{\beta}_{LAD}$ can be different when $\mu \neq 0$. As $\hat{\beta}_{LAD}$ is the maximum likelihood estimator under the presence of Laplace error, under this condition it could be recommended to use LAD estimator instead of OLS estimator in linear regression. The behaviour of both estimators will be tested under the presence of Laplace errors in the last section of this text. Some examples of models, where Laplace errors seem more likely then normal are given for instance in Gokarna (2006). Some examples are the modeling of detector relative efficiencies, extreme wind speeds, position errors in navigation, stock return, the Earth's magnetic field and wind shear data.

3.2 Consistency and asymptotic normality

In this section we will state theorems which establishes the strong consistency of LAD estimator and its asymptotic behaviour. Since we have no explicit formula for $\hat{\beta}_{LAD}$, the proof of these theorems are far harder then for instance in the least squares case. In general, proofs of these theorems utilizes Ergodic theory and consequently, are beyond the scope of this text. Throughout this chapter we will assume that $(x_i, y_i)_{i=1}^n$ is stationary and ergodic sequence which obeys model (M). Even the proper definition of "stationary ergodic sequence" is beyond the scope of this text. We can think of this sequence as a generalization of random sample. The proof of the following theorems can be found in the Bloomfield and Steiger (1983).

Theorem 10. *Let $(x, y)^T \in \mathbb{R}^{k+1}$ be an integrable random vector, $y = \langle\beta, x\rangle + \varepsilon$. Let x and ε be independent, ε have an unique zero median and let*

$$P(\langle\beta, x\rangle = 0) = 1 \implies \beta = 0$$

Then $\hat{\beta}_{LAD} \to \beta$ almost surely.

Fact that x and ϵ are independent implies $\mathrm{E}(\varepsilon|x) = \mathrm{E}(\varepsilon)$ a.s. and ε having the unique median at zero substitutes for usual $\mathrm{E}(\varepsilon) = 0$ used to establish consistency of the Ordinary Least Squares. The following theorem establishes asymptotic normality if LAD estimator.

Theorem 11. *Let $(x, y)^T \in \mathbb{R}^{k+1}$ be a random vector, $y = \langle\beta, x\rangle + \varepsilon$. Let x and ε be independent. Let the matrix of second movements of x C be positive definite. Let y have positive and continuous density around zero. Let ε have an unique zero median. Then*

$$\sqrt{n}(\hat{\beta}_{LAD} - \beta) \to N_{LAD} \text{ in distribution,}$$

where N_{LAD} is the normal vector with zero mean and covariance matrix $\frac{1}{(2f(0))^2}C^{-1}$

3.3 Bias

Theorem from the previous part assure strong consistency of the LAD estimator. This in turn implies that the LAD estimator is asymptotically unbiased, but the way we defined LAD estimator makes it in general biased. We defined the LAD estimator as any minimizer of f and we know that there might not be a unique solution which might cause the estimator to be biased. Fortunately, this can be fixed. We know that there is no analytical formula for solution and that all solutions thus must be computed using iterated algorithms. The LAD estimator could also be defined as an output of certain algorithm. We could, for instance, say that given data set $(x_i, y_i) \in \mathbb{R}^k$, the LAD estimator of β in model (M) is the output of The Barrodale-Roberts algorithm with the given set used as input. But the BR algorithm as we stated it does only find solution, it does not try to find such solution, which would be an unbiased estimator. Therefore, question is whether there is an algorithm which would yield an unbiased LAD estimator. Indeed, there are such algorithms introduced in Sielken and Heartely (1973). They rely heavily on methods of linear programming and their description is beyond the scope of this text.

3.4 Robustness

It is commonly said, that LAD estimator is more robust then for instance OLS estimator. In this part we will discuss its robustness properties. Firstly, we should discuss why the “robustness property” of estimator is desired. Informally, we call estimator robust, if it is able to resist corruptions in the data i.e. it is able to somewhat filter those observation, which seem corrupt. The robustness property of the estimator has not bean considered an important property of an estimator for a large part of history of statistics. One reason for that is that in the past, statistics were done “on paper” i.e. without the use of computer. That implied limited sample sizes and in a small sample size one can manually check for the presence of errors in observations. Nowadays, when sample sizes are in thousands, this would be impractical or impossible. Henceforth, it is desirable that an estimator is able to do this on its own. In the next we will assume that matrix X defined as always has full column rank.

The most common measure of robustness is the concept of Breakdown points which is introduced in the next definition.

Definition 4. *Let T be a family of estimators defined for all sample sizes and let $x = \{x_1, \dots, x_n\}$ be a given data set. Let Z_m be the set of all data sets of size m. We say that estimator T breaks down at x for a contamination of size m if*

$$\sup_{z \in Z_m} \|T(x) - T(x \cup z)\| = +\infty,$$

i.e. we are able to make $T(x \cup z)$ arbitrary far from $T(x)$. Let m^* be the smallest amount of contamination for which T breaks at x i.e.

$$m^* = \min_{m \in \mathbb{N}} \left\{ m : \sup_{z \in Z_m} \|T(x) - T(x \cup z)\| = +\infty \right\}.$$

The breakdown point of T at x is defined as

$$\varepsilon^*(T, x) = \frac{m^*}{n + m^*}.$$

Clearly, the best would be $\varepsilon^* = \frac{1}{2}$ which occurs when $m^* = n$. The worst is the case of $m^* = 1$ i.e. $\varepsilon^* = \frac{1}{n+1}$ which occurs when adding just one contamination point causes arbitrary large changes in T. It is easy to show that OLS estimator has this property.

Theorem 12. *For a given data set $(x_i, y_i)_{i=1}^n \in \mathbb{R}^{k+1}$, the OLS estimator of β in model (M) has the breakdown point $\frac{1}{n+1}$ i.e.*

$$\varepsilon^*(\hat{\beta}_{OLS}, (x_i, y_i)_{i=1}^n) = \frac{1}{n+1}.$$

Moreover, let K be any subset of $\mathbb{R}^k$. Then adding point (x_{n+1}, y_{n+1}), where $x_{n+1} \in K$ and $y_{n+1} \in \mathbb{R}$ is arbitrary far from zero can cause arbitrary large $\|\hat{\beta}_{OLS}((x_i, y_i)_{i=1}^n) - \hat{\beta}_{OLS}((x_i, y_i)_{i=1}^{n+1})\|$.

Proof. Let $x_{n+1} \in K$. Let $\hat{\beta}_{OLS}^n$ be OLS estimator of β in model (M) with only points $(x_i, y_i)_{i=1}^n$ and let $\hat{\beta}_{OLS}^{n+1}$ be OLS estimator of β in model (M) with points $(x_i, y_i)_{i=1}^{n+1}$. Let

$$g(y_{n+1}) := \hat{\beta}_{OLS}^n - \hat{\beta}_{OLS}^{n+1}.$$

Let $X_n \in \mathbb{R}^{n \times k}$ be a matrix with x_i as a i-th row. Let $\tilde{y}_n = (y_1, \dots, y_n)^T$ and let $\widetilde{X_n} = (X_n^T X)^{-1} X_n^T$. Then $\hat{\beta}_{OLS}^n = \widetilde{X_n} \tilde{y}_n$ or equivalently

$$\hat{\beta}_{OLS}^n = \begin{pmatrix} y_1 \tilde{x}_{11} + \cdots + y_n \tilde{x}_{1n} \\ y_1 \tilde{x}_{21} + \cdots + y_n \tilde{x}_{2n} \\ \vdots \\ y_k \tilde{x}_{k1} + \cdots + y_n \tilde{x}_{kn} \end{pmatrix}$$

thus

$$g(y_{n+1}) = \begin{pmatrix} y_1 \tilde{x}_{11} + \cdots + y_n \tilde{x}_{1n} - y_1 \tilde{x}_{11} - \cdots - y_n \tilde{x}_{1n} - y_{n+1} \tilde{x}_{1,n+1} \\ y_1 \tilde{x}_{21} + \cdots + y_n \tilde{x}_{2n} - y_1 \tilde{x}_{21} - \cdots - y_n \tilde{x}_{2n} - y_{n+1} \tilde{x}_{2,n+1} \\ \vdots \\ y_1 \tilde{x}_{k1} + \cdots + y_n \tilde{x}_{kn} - y_1 \tilde{x}_{k1} - \cdots - y_n \tilde{x}_{kn} - y_{n+1} \tilde{x}_{k,n+1} \end{pmatrix}$$

and because all $\tilde{x}_{ij}$ does not depend on y_{n+1}, we can write (if we let $0 \cdot +\infty = 0$)

$$\lim_{y_{n+1} \to +\infty} g(y_{n+1}) = \begin{pmatrix} -\operatorname{sgn}(\tilde{x}_{1,n+1}) \cdot (+\infty) \\ -\operatorname{sgn}(\tilde{x}_{2,n+1}) \cdot (+\infty) \\ \vdots \\ -\operatorname{sgn}(\tilde{x}_{k,n+1}) \cdot (+\infty) \end{pmatrix}.$$

Also, this vector cannot be a zero vector, because X has full column rank. We thus must have

$$\lim_{y_{n+1} \to +\infty} \|g(y_{n+1})\| = +\infty.$$

□

The second part of the theorem states, that given dataset (x_i, y_i), even if we take K in such a way, that it contains cluster of points $x_i, i = 1, \dots, n$ and "not much more" (i.e. whenever we take x_{n+1} from K, it will be close to another point x_i for some i), sufficiently large y_{n+1} still "breaks" OLS estimator. The LAD estimator also has the worst breakdown point as will be shown in Theorem 13, but this "moreover" property is not there i.e. when we want to break LAD for sure, we need to force not only y_{n+1} to be large, but also x_{n+1}.

Theorem 13. *For a given data set $(x_i, y_i)_{i=1}^n \in \mathbb{R}^{k+1}$, the LAD estimator of β in model (M) has the breakdown point $\frac{1}{n+1}$ i.e.*

$$\varepsilon^*(\hat{\beta}_{LAD}, (x_i, y_i)_{i=1}^n) = \frac{1}{n+1}.$$

Proof. Let $(x_i, y_i)_{i=1}^n \in \mathbb{R}^{k+1}$ be given. Let

$$F = \{S \subseteq \{1, \dots, n\} : |S| = k, \operatorname{span}\{x_i : i \in S\} = \mathbb{R}^k\}$$

and let

$$\Delta(S) = \{\delta \in \mathbb{R}^k : \|\delta\| = 1, \langle \delta, x_i \rangle = 0 \\ \text{for } i \in I \text{ where } I \text{ is the subset of } S, |I| = k - 1\}$$

$\Delta(S) = \{\delta \in \mathbb{R}^k : \|\delta\| = 1, \langle \delta, x_i \rangle = 0$ for $i \in I$ where I is the subset of S, $|I| = k - 1$ } }. This forces $\operatorname{span} \Delta(S) = \mathbb{R}^k$. Finally, let

$$t = \min_{\|x_{n+1}\|=1} \left\{ \min_{S \in F} \left\{ \max_{\delta \in \Delta(S)} \{|\langle \delta, x_{n+1}|\rangle\} \right\} \right\}.$$

We must have $t > 0$ since otherwise there would exist $x_{n+1} \in \mathbb{R}^k, \|x_{n+1}\| = 1$ and $S \in F$ such that $\forall \delta \in \Delta(S) : \langle \delta, x_{n+1} \rangle = 0$ but that contradicts to $\operatorname{span}\{\Delta(S)\} = \mathbb{R}^k$. Let us choose $x_{n+1} \in \mathbb{R}^k$ such that

$$\|x_{n+1}\| > \frac{1}{t}\sum_{i=1}^{n}\|x_i\|.$$

We will show that for any $y_{n+1} \in \mathbb{R}$ the LAD fit through points $(x_i, y_i)_{i=1}^{n+1}$ must pass through $[x_{n+1}, y_{n+1}]$. Let $y_{n+1} \in \mathbb{R}$. For contradiction let us assume that we have LAD fit $\hat{\beta}_{LAD}$ using points $(x_i, y_i)_{i=1}^{n+1}$ which satisfies $y_{n+1} - \langle \hat{\beta}_{LAD}, x_{n+1} \rangle \neq 0$. This fit must pass through k points from $(x_i, y_i)_{i=1}^{n}$. Let it be points $(x_i, y_i), i \in S$ for some $S \in F$. By the definition of t we find $\delta \in \Delta(S)$ such that$|\langle \delta, x_{n+1} \rangle| \geq t\|x_{n+1}\|$. Because $1 = \|\delta\| = |\delta_1| + \cdots + |\delta_k|$, for all i we get $\delta_i \leq 1$. Hence, for all $i \in \{1, \ldots, n\}$ we have $|\langle \delta, x_i \rangle| = |\delta_1 x_{i1} + \cdots + \delta_k x_{ik}| \leq |\delta_1||x_{i1}| + \cdots + |\delta_k||x_{ik}| \leq |x_{i1}| + \cdots + |x_{ik}| = \|x_i\|$. We can thus write

$$|\langle \delta, x_{n+1} \rangle| \geq t\|x_{n+1}\| > \sum_{i=1}^{n}\|x_i\| \geq \sum_{i=1}^{n}|\langle \delta, x_i \rangle|.$$

By Theorem 5 we must have $f'_{\delta\pm}(\hat{\beta}_{LAD}) \geq 0$. We know that

$$f'_{\delta+}(\hat{\beta}_{LAD}) = \sum_{Z_{n+1}}|\langle \delta, x_i \rangle| - \sum_{P_{n+1}}|\langle \delta, x_i \rangle| + \sum_{N_{n+1}}|\langle \delta, x_i \rangle|$$

where Z, P, N are the same as in Theorem 5 (for instance $P_{n+1} = \{i \in \{1, \ldots, n+1\} : y_i - \langle \delta, x_i \rangle > 0\}$).

If $y_{n+1} - \langle \hat{\beta}_{LAD}, x_{n+1} \rangle > 0$, then $n + 1 \in P$ and we have

$$\begin{aligned} f'_{\delta+}(\hat{\beta}_{LAD}) &= \sum_{Z_n}|\langle \delta, x_i \rangle| - \sum_{P_n}|\langle \delta, x_i \rangle| - |\langle \delta, x_{n+1} \rangle| + \sum_{N_n}|\langle \delta, x_i \rangle| \\ &< |\langle \delta, x_{n+1} \rangle| - |\langle \delta, x_{n+1} \rangle| - |\langle \delta, x_{n+1} \rangle| + |\langle \delta, x_{n+1} \rangle| = 0, \end{aligned}$$

i.e. $f'_{\delta+}(\hat{\beta}_{LAD}) < 0$ and that is a contradiction to $\hat{\beta}_{LAD}$ being LAD fit.

If $y_{n+1} - \langle \hat{\beta}_{LAD}, x_{n+1} \rangle < 0$, then

$$\begin{aligned} f'_{\delta-}(\hat{\beta}_{LAD}) &= \sum_{Z_n} |\langle \delta, x_i \rangle| + \sum_{P_n} |\langle \delta, x_i \rangle| - \sum_{N_n} |\langle \delta, x_i \rangle| - |\langle \delta, x_{n+1} \rangle| \\ &< |\langle \delta, x_{n+1} \rangle| + |\langle \delta, x_{n+1} \rangle| - |\langle \delta, x_{n+1} \rangle| - |\langle \delta, x_{n+1} \rangle| = 0, \end{aligned}$$

and again, that is a contradiction. We thus must have $y_{n+1} - \langle \hat{\beta}_{LAD}, x_{n+1} \rangle = 0$. Because y_{n+1} was arbitrary, we can take it arbitrary large, and that forces $\|\hat{\beta}_{LAD}\|$ to be arbitrary large, thus

$$\lim_{y_{n+1} \to \infty} \|\hat{\beta}_{LAD}((x_i, y_i)_{i=1}^{n}) - \hat{\beta}_{LAD}((x_i, y_i)_{i=1}^{n+1})\| = +\infty.$$

□

The differences between Theorem 12 and Theorem 13 could also be viewed that almost always, corrupted observation causes at least some issue for OLS estimator. In the LAD case, errors in observations might not cause an error (i.e. when only y_{n+1} is corrupted and x_{n+1} is not). It could also be intuitively understood, that unless LAD fit passes through error observation, it should not be effective in a big way. We will test how LAD compares against OLS in the chapter dealing with the simulations.

3.5 Least Weighted Absolute Deviations

Let x_i, y_i be as in the LAD fitting problem. The Least Weighted Absolute Deviations (LWAD for short) fitting problem is the problem of finding the minimizer of function $g : \mathbb{R}^k \to \mathbb{R}$ given by

$$g(\beta) = \sum_{i=1}^{n} w_i(\beta) |y_i - \langle \beta, x_i \rangle|,$$

where $w_i : \mathbb{R}^k \to \mathbb{R}$ are given functions. The idea is to define LWAD estimator of β in model (M) as the solution to LWAD fitting problem with w_i chosen in such a way, that the statistical properties of LWAD estimator are better then for the standard LAD estimator. Namely, the idea is to improve robustness. We call w_i the weight functions and we choose them in such a way that data which seems corrupted (i.e. are too large in most cases) have lesser weight then non-corrupted data. This should make the estimator resistant to outliers and to (bad) leverage points. The same idea is used in many other estimators, for instance weighted least squares. Among all estimator using some sort of weight function, LWAD seems very unpopular. The reason for that seems even harder computational difficulties than for the standard LAD estimator. Also, it seems no apparent advantages of LWAD exist over more popular estimator.

4 Simulations

In this section we present the result of simulations we conducted to illustrate the behaviour of LAD estimator in linear regression. We considered liner model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

and a random samples $(x_{i1}, x_{i2}, x_{i3}, y_i)_{i=1}^n$ of different sizes n. The experiments were performed as follows. Firstly, the distribution of ε was chosen. Next, we divided interval $\langle -5, 5 \rangle$ equidistantly to n intervals I_i. Middle point of I_i was assigned to x_i for $i = 1, \dots, n$. Then x_{i2}, x_{i3} were generated from normal distribution with zero mean and standard deviation 3. Then, $\beta_0, \beta_1, \beta_2, \beta_3$ were chosen from uniform distribution on interval $(-10, 10)$. Lastly, we generated ε_i from the chosen distribution and taken $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ for $i = 1, \dots, n$. Then, we ran a regression of (x_{i1}, x_{i2}, x_{i3}) on y_i using the same model and obtained OLS estimator and LAD estimator. For LAD estimator we used the BR algorithm. This was repeated m number of times (10 000 times or 20 000 times actually).

The distributions of ε we considered were as follows: Standard Normal; Laplace with $\mu = 0, s = 1$; standard Logistic, with cumulative distribution function

$$F(x) = \frac{1}{1 + e^{-x}},\ x \in \mathbb{R},$$

and lastly the special case of symmetric Pareto distribution with density

$$g(x) = \frac{1}{1 + |x|^3},\ x \in \mathbb{R}.$$

4.1 Consistency

The first experiment we conducted meant to demonstrate the consistency of LAD estimator. We ran the experiment using normal errors $m = 20\ 000$ times for each $n = 10, 50, \dots, 10\ 000$. Each time we obtained LAD estimator $\hat{\beta}$ and calculated the deviation of $\hat{\beta}$ from the true value $\beta = (\beta_1, \beta_1, \beta_3, \beta_4)^T$. In the Table 1 we present the maximums of such deviations.

n	$\max\{\|\beta - \hat{\beta}\|\}$
10	3.5654
50	0.7773
150	0.4201
500	0.2345
1500	0.1280
2500	0.1012
10 000	0.0529

Table 1: $m = 20\ 000$

We can clearly see, that with the increasing number of observations within a sample, the maximum deviation of LAD estimator of β deviates less and less from the true velue of β. That is in line with consistency.

4.2 Comparison with OLS

In the next experiment, we compared the LAD estimator to the OLS estimator under different distributions of ε and under different sample sizes. In the Table 2,3 and 4 we present the means of $\|\beta - \hat{\beta}\|$ where $\hat{\beta}$ is either LAD estimator or OLS estimator.

	Normal	Laplace	Logistic	Pareto
OLS	0.354	0.502	0.668	1.525
LAD	0.427	0.509	0.752	1.006

Table 2: $n = 10$, $m = 10\ 000$

	Normal	Laplace	Logistic	Pareto
OLS	0,130	0,343	0,248	1.509
LAD	0,168	0,297	0,272	0.830

Table 3: $n = 50$, $m = 10\ 000$

Our results are as expected. Under normal errors LAD seems inferior to the OLS estimator. This is in line with the the Gauss-Markov theorem. If the errors follow Laplace distribution our data suggest LAD has superiority over OLS. That is in line with Theorem 9 which ensures that LAD is the maximum likelihood estimator.

	Normal	Laplace	Logistic	Pareto
OLS	0,076	0,104	0,138	1.522
LAD	0,095	0,086	0,150	0.806

Table 4: $n = 150$, $m = 10\ 000$

Given the results and the theory, one should safely recommend the usage of LAD over OLS in the case of Laplace errors. If the ε follow standard logistic distribution (which can be though of as normal distribution with heavier tales), we still see the superiority of OLS, but this superiority is not as large as if errors have normal distribution. From the Table 4, we have that under normal errors, LAD estimator of β has 1.25 times larger (i.e. larger by 25%) mean deviation from the true value of β then the OLS estimator. Under Logistic errors, this falls down to LAD having 1.088 times larger (i.e. larger by 8.8%) mean deviation form the true value. This might suggest LAD is better resistant to larger errors (which are implied by logistic function having heavier tales). Lastly, under errors drawn from symetric Pareto distribution, which can be thought as Laplace distribution with heavier tales, LAD has, as expected, advantage over OLS. As was the case with the logistic function, the heavier tales of density of errors influence OLS in a bigger way then LAD. Under the Laplace errors, OLS had about 21% large mean deviations form true value then LAD. Under the Pareto errors, this grows to 89%.

4.3 Corrupted data

In the third and final experiment, our goal was to test the resistance of the LAD and the OLS estimators to the presence of corrupted data. We ran the same experiment as was the previous one, but we only consider Normal errors and Laplace errors. Also, we corrupted the percentage of date. This was done in the following way. We randomly selected data points to corrupt. Then we found the radius r of s sphere in $\mathbb{R}^k$ such that all non-corrupted data point lie within this sphere. Then we assigned a new values to the corrupted data points. The new values were assigned in such a way that the corrupted data points were randomly (based on uniform distribution) placed on a sphere in $\mathbb{R}^k$ with the radius $5r$ (or $50r$). We present the resulting mean $\|\beta - \hat{\beta}\|$ in the tables 5 and 6. In the tables, the percentage point means the percentage of corrupted data, the $5r$ or $50r$ stand for the radius of sphere with corrupted points.

	5%, $5r$	10%, $5r$	5%, $50r$	10%, $50r$
OLS	1,557	2,125	10,779	17,784
LAD	0,308	0,539	1,071	1,006

Table 5: Normal errors, $n = 150$, $m = 10\ 000$

	5%, $5r$	10%, $5r$	5%, $50r$	10%, $50r$
OLS	1,482	2,121	9,427	18,052
LAD	0,276	0,534	1,088	1,017

Table 6: Laplace errors, $n = 150$, $m = 10\ 000$

The results for the OLS estimator is with line Theorem 12 - the OLS estimator is always influenced by the presence of corrupted data. The LAD estimator is also influenced, but at the much lower level. This might suggest, that if the data are subject to corruption, using LAD might prove to be better then using OLS. Of course, there are plenty of techniques for treating this type of corruption which might in turn make least standard regression better. But here we wanted to test the robustness of LAD estimator and our data are in line with often used statement, that LAD estimator is robust.

Conclusions

We have shown the most important properties of Least Absolute Deviations and the key ideas behind this method. We have started with basic properties of solution which demonstrated the difficulties involving LAD fits. Non-smoothness of AD distance function f rendered the usage of calculus mostly ineffective. On the other hand, the fact that f was piecewise linear allow the extensive use of linear algebra. Most important result was the proof of existence of extreme LAD fit. This existence was later used in the section where we described The Barrodale-Roberts Algorithm which we later used in the section dealing with statistical simulations. After the brief example of method for finding approximate solutions, Iteratively Reweighted Least Squares, we moved towards the section where we described the statistical properties of the LAD fit with focus on linear regression. We defined the LAD estimator for linear model and showed it arises as the maximum likelihood estimator if the error term follow Laplace distribution. Later, in experimental section, we demonstrated that under Laplace error, the LAD estimator has an advantage over the Ordinary Least Squares and suggested its usage under the assumption of Laplace errors. We also talked about consistency and asymptotic normality of LAD estimator. We presented theorems establishing these properties. We later successfully tested its consistency. Then we talked about how LAD estimator is in general biased, but also explained that the unbiased LAD estimator can be obtained by improving the algorithm used for its computation. Lastly, we discussed the robustness of LAD estimator. We showed the important difference in the breakdown points of LAD and OLS. The robustness of LAD estimator was tested in the experimental section and we saw its advantages over OLS in this regard.

References

[1] Bloomfield, P., Steiger, W. *Least Absolute Deviations: Theory, Applications and Algorithms.* Birkhäuser Basel, Progress in Probability, 1983.

[2] Brennan, J. J. and Seiford L. M. *Linear programming and l_1 regression: A geometric interpretation.* Computational Statistics & Data Analysis 5, no. 4, 1987, pp. 263-276.

[3] Dielman, T. E. *Least absolute value regression: recent contributions.* Journal of Statistical Computation and Simulation 75, no. 4, 2005, pp.263-286.

[4] Dodge, Y. *An introduction to L1-norm based statistical data analysis.* Computational Statistics & Data Analysis 5, no. 4, 1987, pp. 239-253.

[5] Dutter R., *Robust regression: Different approaches to numerical solutions and algorithms.* Res., Rep. no. 6, Fachgruppe für Statistik, Eidgen. Technische Hochschnule, Zurich, 1975. Ellis, S. P. *Instability of least squares, least absolute deviation and least median of squares linear regression.* Statistical Science, 1998, pp. 337-344.

[6] Gilbert B. Jr. and Koenker R. *Asymptotic theory of least absolute error regression.* Journal of the American Statistical Association 73, no. 363, 1978, pp. 618-622.

[7] Gokarna R. A., *Study of laplace and related probability distributions and their applications.* . Graduate Theses and Dissertations, University of South Florida, 2006.

[8] Hald, A. *Galileo's Statistical Analysis of Astronomical Observations* International Statistical Review Vol. 54, No. 2, Aug., 1986, pp. 211-220

[9] Lei, D., Anderson I. J. and Cox., M. *A robust algorithm for least absolute deviations curve fitting.* Huddersfield Univ Leeds (The United Kingdom), 2001.

[10] Li, Y. and Arce, G. R. *A maximum likelihood approach to least absolute deviation regression.* EURASIP Journal on Advances in Signal Processing 2004, no. 12, 2004, pp. 1-8.

[11] Ling, S. *Self?weighted least absolute deviation estimation for infinite variance autoregressive models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, no. 3, 2005, pp. 381-393. Schlossmacher, E.J. *An Iterative Technique for Absolute Deviations Curve Fitting*. Journal of the American Statistical Association, 68, 1973, pp. 857-859