A Mathematical Programming Approach for Improving the Robustness of Least Sum of Absolute Deviations Regression

Avi Giloni,[1] Bhaskar Sengupta,[2,*] Jeffrey S. Simonoff[3]

[1] *Sy Syms School of Business, Yeshiva University, 500 W 185 St, BH-428, New York, NY 10033, USA*

[2] *ExxonMobil Research and Engineering, 1545 Route 22 East, Annandale, NJ 08801, USA*

[3] *Leonard N. Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012, USA*

Received 26 July 2004; revised 22 May 2005; accepted 30 September 2005
DOI 10.1002/nav.20139
Published online 27 February 2006 in Wiley InterScience (www.interscience.wiley.com).

Abstract: This paper discusses a novel application of mathematical programming techniques to a regression problem. While least squares regression techniques have been used for a long time, it is known that their robustness properties are not desirable. Specifically, the estimators are known to be too sensitive to data contamination. In this paper we examine regressions based on Least-sum of Absolute Deviations (LAD) and show that the robustness of the estimator can be improved significantly through a judicious choice of weights. The problem of finding optimum weights is formulated as a nonlinear mixed integer program, which is too difficult to solve exactly in general. We demonstrate that our problem is equivalent to a mathematical program with a single functional constraint resembling the knapsack problem and then solve it for a special case. We then generalize this solution to general regression designs. Furthermore, we provide an efficient algorithm to solve the general nonlinear, mixed integer programming problem when the number of predictors is small. We show the efficacy of the weighted LAD estimator using numerical examples.
 Naval Research Logistics 53: 261–271, 2006

Keywords: algorithms; breakdown point; knapsack problem; nonlinear mixed integer programming; robust regression

1. INTRODUCTION

Consider the well-known statistical linear regression problem. We have n observations on some "dependent" variable y and some number $p \geq 1$ of "independent" variables $x_1, \ldots, x_p$, for each one of which we know n values as well. We denote

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} x_1^1 & \cdot & \cdot & \cdot & x_p^1 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ x_1^n & \cdot & \cdot & \cdot & x_p^n \end{pmatrix} = \begin{pmatrix} \mathbf{x}^1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}^n \end{pmatrix} = (\mathbf{x}_1, \ldots, \mathbf{x}_p),$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector of n observations, $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are column vectors with n components, and $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are row vectors with p components corresponding to the columns and rows of the $n \times p$ matrix $\mathbf{X}$, respectively. To rule out pathologies we assume throughout that the rank $r(\mathbf{X})$ of $\mathbf{X}$ is full, i.e., that $r(\mathbf{X}) = p$. If the regression model includes an intercept term, as is typical, $\mathbf{x}_1$ is a column of ones, but this is not required.

The statistical linear regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \ldots, \beta_p)$ is the vector of parameters of the linear model and $\boldsymbol{\varepsilon}^{\mathrm{T}} = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector of n random variables corresponding to the error terms in the asserted relationship. The superscript T denotes "transposition" of a vector or matrix throughout this work. In the statistical model, the dependent variable y is a random variable for which we obtain measurements or observations that contain some "noise" or measurement errors that are captured in the error terms $\boldsymbol{\varepsilon}$.

Although (1) gives the statistical model underlying the regression problem, the numerical problem faced is slightly different. For this, we write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}, \tag{2}$$

where given some parameter vector $\boldsymbol{\beta}$, the components r_i of the vector $\mathbf{r}^{\mathrm{T}} = (r_1, \ldots, r_n)$, are the *residuals* that result, given the observations $\mathbf{y}$, a fixed design matrix $\mathbf{X}$, and

*This work was done when the author was at Yeshiva University.
Correspondence to: A. Giloni (agiloni@ymail.yu.edu); B. Sengupta (Bhaskar.Sengupta@exxonmobil.com); J. S. Simonoff (jsimonof@stern.nyu.edu)

the chosen vector $\boldsymbol{\beta} \in \mathbb{R}^p$. It is well known that when the errors $\boldsymbol{\varepsilon}$ are normally (Gaussian) distributed, the least squares parameter estimator (which minimizes the ℓ_2-norm $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 = \sum_{i=1}^{n}(y_i - \mathbf{x}^i\boldsymbol{\beta})^2$ of the residuals) has many desirable properties, having the minimum variance among all linear unbiased estimators and (being the maximum likelihood estimator) achieving the minimum possible variance for all consistent estimators as the sample size becomes infinite.

Many other regression estimators, in addition to least squares, have been proposed in the statistical literature. These techniques have been introduced to improve upon least squares in some way. Among these techniques are those that are robust with respect to outliers, as it is known that least squares regression estimates are affected by wild observations. There have been several measures developed within the statistical literature that quantify the robustness of a regression estimator. In this paper, we focus on the *breakdown point* (c.f. [15]) to be formally defined in Section 2.

One of the earliest proposals for estimating regression parameters was regression performed using the ℓ_1-norm, also called Least-sum of Absolute Deviations (LAD). This regression problem can be solved using linear programming, hence, its interest in the operations research community. LAD regression has become more useful with the advent of interior point methods for solving linear programs and with the increase in computer processing speed [13]. Furthermore, LAD regression is more robust than least squares (c.f. [3]). As far back as the 1960s and 1970s, it was noticed that empirically, LAD outperformed least squares in the presence of fat tailed data (c.f. [16]). However, it is only more recently that the robustness properties of LAD regression have been theoretically determined. For regression problems where there may be outliers in the dependent variable, LAD regression is a good alternative to least squares. Furthermore, LAD can be utilized to demonstrate that least squares is accurate when indeed it is (if the LAD and least squares estimates are similar); this can be useful, since the least squares estimator is more efficient than LAD in the presence of Gaussian errors.

Although there has been much research on studying and quantifying the robustness properties of LAD regression, to the best of our knowledge there has not been any research on modifying the LAD estimator in order to improve upon its breakdown properties. Our contribution is that we model and provide a good solution to the problem of improving the robustness of LAD regression via mathematical programming techniques. Specifically, we study a nonlinear mixed integer program that can be used to improve the robustness of LAD regression. We demonstrate that the introduction of nonuniform weights can have a positive impact on the robustness properties of LAD regression. We develop an algorithm for determining these weights and demonstrate the usefulness of our approach through several numerical examples. Specifically, we develop an algorithm for choosing weights that can significantly improve the robustness properties of LAD regression. In order to study the weighted LAD (WLAD) regression problem, we use and apply linear and mixed integer programming techniques. Our studies indicate that WLAD regression should be seriously considered as a regression technique in many regression and forecasting contexts.

The structure of the paper is as follows. In Section 2, we introduce the LAD regression problem and summarize some of the pertinent research on LAD regression and its robustness properties. We show (in Section 3) that the problem of incorporating nonuniform weights can be formulated as a nonlinear mixed integer program. In Section 4, we demonstrate that this problem is equivalent to a mathematical program with a single functional constraint, resembling the knapsack problem. In Section 5, we discuss a special case of the weight determination problem for which an optimal solution can be obtained. Using the insights gained in Sections 3–5, we develop an algorithm (in Section 6) to solve the problem approximately and demonstrate that the algorithm significantly improves the robustness of the estimators through several numerical examples in Section 7.

2. LAD REGRESSION AND BREAKDOWN

In the case of LAD regression, the general numerical problem (2) takes as the (optimal) parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ those that minimize the ℓ_1-norm $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n} |y_i - \mathbf{x}^i\boldsymbol{\beta}|$ of the residuals. It is well known that this problem can be formulated as the linear programming (LP) problem

$$\begin{aligned} \min \quad & \mathbf{e}_n^{\mathrm{T}}\mathbf{r}^+ + \mathbf{e}_n^{\mathrm{T}}\mathbf{r}^- \\ \text{such that} \quad & \mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y} \\ & \boldsymbol{\beta} \text{ free}, \ \mathbf{r}^+ \geq \mathbf{0}, \quad \mathbf{r}^- \geq \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{e}_n$ is the vector with all n components equal to 1. In (3) the residuals $\mathbf{r}$ of the general form (2) are simply replaced by a difference $\mathbf{r}^+ - \mathbf{r}^-$ of nonnegative variables, i.e., we require that $\mathbf{r}^+ \geq \mathbf{0}$ and $\mathbf{r}^- \geq \mathbf{0}$, whereas the parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ are "free" to assume positive, zero, or negative values. From the properties of linear programming solution procedures, it follows that for any solution inspected by the simplex algorithm, either $r_i^+ > 0$ or $r_i^- > 0$, but not both, thus giving $|r_i|$ in the objective function depending on whether $r_i > 0$ or $r_i < 0$ for any $i \in N$ where $N = \{1, \ldots, n\}$. Every optimal extreme point solution $\boldsymbol{\beta}^* \in \mathbb{R}^p$ of the LAD regression problem has the property that there exists a nonsingular $p \times p$ submatrix $\mathbf{X}_B$ of $\mathbf{X}$ such that

$$\begin{aligned} \boldsymbol{\beta}^* &= \mathbf{X}_B^{-1}\mathbf{y}^B, \\ \mathbf{r}^+ &= \max\{\mathbf{0}, \ \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\}, \\ \mathbf{r}^- &= -\min\{\mathbf{0}, \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\} \end{aligned}$$

where $|B| = p$ and $\mathbf{y}^B$ is the subvector of $\mathbf{y}$ corresponding to the rows of $\mathbf{X}_B$ (c.f. [3]).

The notion of the *breakdown point* of a regression estimator (due to Hampel [6]) can be found in [15] and is as follows. Suppose we estimate the regression parameters $\boldsymbol{\beta}$ by some technique τ from some data $(\mathbf{X}, \mathbf{y})$, yielding the estimate $\boldsymbol{\beta}^\tau$. If we contaminate m $(1 \leq m < n)$ rows of the data in a way so that row i is replaced by some arbitrary data $(\widetilde{\mathbf{x}}^i, \widetilde{y}_i)$, we obtain some new data $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$. The same technique τ applied to $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$ yields estimates $\boldsymbol{\beta}^\tau(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}})$ that are different from the original ones. We can use any norm $\| \cdot \|$ on $\mathbb{R}^p$ to measure the distance $\|\boldsymbol{\beta}^\tau(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}}) - \boldsymbol{\beta}^\tau\|$ of the respective estimates. If we vary over *all* possible choices of contamination then this distance either stays bounded or not. Let

$$b(m, \tau, \mathbf{X}, \mathbf{y}) = \sup_{\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}}} \|\boldsymbol{\beta}^\tau(\widetilde{\mathbf{X}}, \widetilde{\mathbf{y}}) - \boldsymbol{\beta}^\tau\| \tag{4}$$

be the maximum bias that results when we replace at most m of the original data $(\mathbf{x}^i, y_i)$ by arbitrary new ones. The breakdown point of τ is

$$\alpha(\tau, \mathbf{X}, \mathbf{y}) = \min_{1 \leq m < n} \left\{ \frac{m}{n} : b(m, \tau, \mathbf{X}, \mathbf{y}) \text{ is infinite} \right\},$$

i.e., we are looking for the minimum number of rows of $(\mathbf{X}, \mathbf{y})$ that if replaced by arbitrary new data make the regression technique τ break down. We divide this by n to get $\frac{1}{n} \leq \alpha(\tau, \mathbf{X}, \mathbf{y}) \leq 1$. In practice, $\alpha(\tau, \mathbf{X}, \mathbf{y}) \leq 0.5$, since otherwise it is impossible to distinguish between the uncontaminated data and the contaminated data. The breakdown point of LAD as well as least squares regression is $\frac{1}{n}$ or asymptotically 0, see e.g., [15]. Clearly, the larger the breakdown point, the more robust the regression estimator.

However, LAD regression is more robust than least squares in the following manner. The finite sample breakdown point of the LAD regression estimator is the breakdown point of LAD regression with a fixed design matrix $\mathbf{X}$ and contamination restricted only to the dependent variable $\mathbf{y}$, denoted by $\alpha(\tau, \mathbf{y}|\mathbf{X})$. The finite sample breakdown point, or conditional breakdown point, was introduced by Donoho and Huber ([1]). The finite sample breakdown point has been studied by many authors; see, e.g., [2], [7], [4], and [10]. Ellis and Morgenthaler ([2]) appear to be the first to mention that the introduction of weights can improve the finite sample breakdown point of LAD regression, but they only show this for very small data sets. Mizera and Müller ([11]) examine this question in more detail, showing that the predictors $\mathbf{X}$ can be chosen to increase the breakdown point of LAD.

In this paper, we use the notation and framework set forth by Giloni and Padberg ([4]). Let $N = \{1, \ldots, n\}$ and let U, L, Z be a mutually exclusive three-way partition of N such that $|U \cup L| = q$. Let $\mathbf{e}$ be an $n \times 1$ column vector of ones and let $\mathbf{X}_Z$ be the submatrix of $\mathbf{X}$ whose row indexes are in Z. Similarly, we define X_U and X_L. The subscripts of U and L on $\mathbf{e}$ denote a vector of ones of appropriate dimension. Giloni and Padberg ([4]) define the notion of q-stability of a design matrix as follows. $\mathbf{X}$ is *q-stable* if $q \geq 0$ is the largest integer such that

$$\mathbf{X}_Z\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0},$$

$$(-\mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U + \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L)\boldsymbol{\xi} + \mathbf{e}_Z^{\mathrm{T}}(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) \leq 0 \tag{5}$$

$$\boldsymbol{\xi} \neq \mathbf{0}, \qquad \boldsymbol{\eta}^+ \geq \mathbf{0}, \qquad \boldsymbol{\eta}^- \geq \mathbf{0} \tag{6}$$

is not solvable for any U, L, Z where $|U \cup L| = q$ and $Z = N - U - L$. They prove that a design matrix $\mathbf{X}$ is q-stable if and only if $\alpha(\ell_1, \mathbf{y}|\mathbf{X}) = \frac{q+1}{n}$. This is in direct contrast to least squares regression where its finite sample breakdown point is $\frac{1}{n}$ or asymptotically 0. (In other words, for any given design, only one component of the vector $\mathbf{y}$ need be contaminated in order that the maximal bias (4) based upon the least squares estimator goes to infinity.) They show that the finite sample breakdown point of LAD regression can be calculated by the following mixed integer program $MIP1$ (where $M > 0$ is a sufficiently large constant and $\varepsilon > 0$ is a sufficiently small constant):

$$MIP1 \quad \min \sum_{i=1}^{n} u_i + \ell_i = q + 1$$

$$\text{such that} \quad \mathbf{x}^i\boldsymbol{\xi} + \eta_i^+ - \eta_i^- + s_i - t_i = 0 \quad \text{for } i = 1, \ldots, n$$

$$s_i - Mu_i \leq 0, \quad t_i - M\ell_i \leq 0 \quad \text{for } i = 1, \ldots, n$$

$$\eta_i^+ + \eta_i^- + Mu_i + M\ell_i \leq M \quad \text{for } i = 1, \ldots, n$$

$$u_i + \ell_i \leq 1 \quad \text{for } i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \eta_i^+ + \eta_i^- - s_i - t_i \leq 0,$$

$$\sum_{i=1}^{n} s_i + t_i \geq \varepsilon$$

$$\boldsymbol{\xi} \text{ free}, \boldsymbol{\eta}^+ \geq \mathbf{0}, \boldsymbol{\eta}^- \geq \mathbf{0}, \mathbf{s} \geq \mathbf{0}, \mathbf{t} \geq \mathbf{0},$$

$$u_i, \ell_i \in \{0, 1\} \text{ for } i = 1, \ldots, n.$$

In our case, we are interested in WLAD regression. The WLAD regression problem also can be formulated as a linear program, as follows.

$$\min \sum_{i=1}^{n} w_i\left(r_i^+ + r_i^-\right)$$

$$\text{such that} \quad \mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y}$$

$$\boldsymbol{\beta} \text{ free}, \mathbf{r}^+ \geq \mathbf{0}, \mathbf{r}^- \geq \mathbf{0}. \tag{7}$$

Here, we assume that the residual associated with observation i is multiplied by some weight, w_i, where we assume that $0 < w_i \leq 1$ without restriction of generality. We note that

if we were to set $w_i = 0$, we would essentially "remove" observation i from the data. We do not permit this, although, if some (optimal) weight is sufficiently near 0, the user can choose to remove the observation from the data set.

We now note that since $(r_i^+ - r_i^-) = (y_i - \mathbf{x}^i\boldsymbol{\beta})$, and by the simplex method either $r_i^+ > 0$ or $r_i^- > 0$ but not both, then $|w_i(r_i^+ - r_i^-)| = w_i(r_i^+ + r_i^-)$. Therefore, if we were to transform our data by setting $(\widehat{\mathbf{x}}^i, \widehat{y}_i) = w_i(\mathbf{x}^i, y_i)$, then $(\widehat{y}_i - \widehat{\mathbf{x}}^i\boldsymbol{\beta}) = w_i(r_i^+ - r_i^-)$. In such a case, the linear program (7) can be reformulated as

$$\begin{aligned} \min \quad & \mathbf{e}_n^{\mathrm{T}}\mathbf{r}^+ + \mathbf{e}_n^{\mathrm{T}}\mathbf{r}^- \\ \text{such that} \quad & w_i\mathbf{x}^i\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = w_i y_i \quad \text{for } i = 1,\ldots,n \\ & \boldsymbol{\beta} \text{ free}, \mathbf{r}^+ \geq \mathbf{0}, \mathbf{r}^- \geq \mathbf{0}. \end{aligned}$$

This shows that WLAD regression can be treated as LAD regression with suitably transformed data. Therefore, the problem of determining the breakdown of WLAD regression with known weights corresponds to determining the breakdown of LAD regression with data $(\widehat{\mathbf{X}}, \widehat{\mathbf{y}})$.

The equivalence of WLAD regression and ordinary LAD regression on transformed data is also useful in other contexts. For example, LAD regression can be adapted to the nonparametric estimation of smooth regression curves (see [17] for a general discussion of nonparametric regression estimation), by local fitting of WLAD regressions ([18,19]). Formulations of the type in this section allowed Giloni and Simonoff ([5]) to derive the exact breakdown properties of local LAD regression at any evaluation point and to make recommendations concerning the best choice of weight function.

In the next section, we formulate the problem of determining the set of weights that maximizes the breakdown point of WLAD regression.

3. PROBLEM FORMULATION

The task of determining the weights that maximize the finite sample breakdown point of LAD regression is a complicated one. If one were to try to solve this problem by brute force, it would require the inspection of all or a large subset of all vectors (if this is possible) $\mathbf{w} \in \mathbf{R}^n$, the transformation of the data described above $(\widehat{\mathbf{x}}^i, \widehat{y}_i) = w_i(\mathbf{x}^i, y_i)$, and then the solution of $MIP1$ for each case. Instead, we formulate this problem as a nonlinear mixed integer program. For the solution methods we discuss below, we make the standard assumption that the design matrix $\mathbf{X}$ is in general position, i.e., every $p \times p$ submatrix of $\mathbf{X}$ has full rank. We do this in order to simplify the analysis that follows. However, we note that even in the case where this assumption is violated, the methodology that we develop is still valid. This is quite different from many so-called high breakdown regression techniques (ones where the breakdown point can be as high as 0.5), where the value of the breakdown point of the estimators depends upon whether the design matrix is in general position; see [15], p. 118. The mixed integer program is

$$\begin{aligned} NLMIP \quad & \max_{\mathbf{w}} \ \min \sum_{i=1}^{n}(u_i + \ell_i) \\ \text{such that} \quad & w_i\mathbf{x}^i\boldsymbol{\xi} + \eta_i^+ - \eta_i^- + s_i - t_i = 0 \\ & \qquad \text{for } i = 1,\ldots,n \quad (8) \\ & s_i - Mu_i \leq 0, \quad t_i - M\ell_i \leq 0 \\ & \qquad \text{for } i = 1,\ldots,n \quad (9) \\ & \eta_i^+ + \eta_i^- + Mu_i + M\ell_i \leq M \\ & \qquad \text{for } i = 1,\ldots,n \quad (10) \\ & u_i + \ell_i \leq 1 \quad \text{for } i = 1,\ldots,n \quad (11) \\ & \sum_{i=1}^{n} \eta_i^+ + \eta_i^- - s_i - t_i \leq 0 \quad (12) \\ & \sum_{i=1}^{n} s_i + t_i \geq \varepsilon \quad (13) \\ & w_i \leq 1 \quad \text{for } i = 1,\ldots,n \quad (14) \\ & \boldsymbol{\xi} \text{ free}, \boldsymbol{\eta}^+ \geq \mathbf{0}, \ \boldsymbol{\eta}^- \geq \mathbf{0}, \mathbf{s} \geq \mathbf{0}, \mathbf{t} \geq \mathbf{0}, \\ & \qquad u_i, \ell_i \in \{0,1\} \ \text{ for } i = 1,\ldots,n. \quad (15) \end{aligned}$$

Note that ξ_j is unrestricted for $j = 1,\ldots,p$. Also note that *NLMIP* is a nonlinear problem because of the first term in the left-hand side of (8). Since nonlinear mixed integer programs are extremely difficult to solve, we will resort to heuristics for finding good feasible solutions. In order to do so, we note that if the weights $\mathbf{w} \in \mathbf{R}^n$ are fixed, then *NLMIP* is just $MIP1$ with transformed data. Therefore, if we were to focus on only specific sets of weights, we could determine which set is the best by solving the different MIPs and choosing the set that had the largest objective function value among the MIPs considered. However, as mentioned by Giloni and Padberg ([4]), $MIP1$ can take a long time to solve for large data sets. We thus provide an alternative way of solving $MIP1$ that is quite efficient when p is small. We note that this alternative way is essentially what Mizera and Müller ([11]) proposed except that we develop it via a linear programming framework. The reasoning behind this alternative way is based upon the following proposition. We provide the proposition and proof here since it is useful in understanding our framework for selecting weights, which is new.

PROPOSITION 1: In order to determine the finite sample breakdown point of LAD regression, it is sufficient to consider $\binom{n}{p-1}$ candidate solutions for the vector $\boldsymbol{\xi} \in \mathbb{R}^p$ in (5) and (6).

PROOF: Consider a three-way partition U_1, L_1, Z_1 of N where $U_1 \cap L_1 = \emptyset$, $Z_1 = N - U_1 - L_1$, and $q = |U_1 \cup L_1|$. Note that

$$\begin{aligned}&\mathbf{X}_{Z_1}\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0},\\ &(-\mathbf{e}_{U_1}^{\mathrm{T}}\mathbf{X}_{U_1} + \mathbf{e}_{L_1}^{\mathrm{T}}\mathbf{X}_{L_1})\boldsymbol{\xi} + \mathbf{e}_{Z_1}^{\mathrm{T}}(\boldsymbol{\eta}^+ + \boldsymbol{\eta}^-) \le 0\\ &\boldsymbol{\xi} \ne \mathbf{0}, \qquad \boldsymbol{\eta}^+ \ge \mathbf{0}, \qquad \boldsymbol{\eta}^- \ge \mathbf{0}\end{aligned}$$

has no solution if and only if the objective function value of the following optimization problem is strictly greater than zero.

$$\begin{aligned}OF = \quad & \min -\sum_{i\in U_1}\mathbf{x}^i\boldsymbol{\xi} + \sum_{i\in L_1}\mathbf{x}^i\boldsymbol{\xi} + \sum_{i\in Z_1}\eta_i^+ + \eta_i^-\\ \text{such that} \quad & \mathbf{X}_{Z_1}\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0}\\ & \boldsymbol{\xi} \ne \mathbf{0}, \quad \boldsymbol{\eta}^+ \ge \mathbf{0}, \quad \boldsymbol{\eta}^- \ge \mathbf{0}. \qquad (16)\end{aligned}$$

We note that since this problem includes a constraint of the form $\boldsymbol{\xi} \ne \mathbf{0}$, its study as a linear program is more difficult. Therefore, we argue as follows. If we were to assume that there exist $\boldsymbol{\xi}_0 \in \mathbb{R}^p$, $\boldsymbol{\xi}_0 \ne \mathbf{0}$, such that $OF < 0$, then consider $\boldsymbol{\xi} = \psi\boldsymbol{\xi}_0$ where $\psi > 0$ is some constant. If $i \in U_1 \cup L_1$, then the sign of $\mathbf{x}^i\boldsymbol{\xi}$ is the same as that of $\mathbf{x}^i\boldsymbol{\xi}_0$. If $i \in Z_1$, $\eta_i^+ + \eta_i^- \ge |\mathbf{x}^i\boldsymbol{\xi}|$ because of (16), and since we are minimizing, equality will hold. Therefore, $\boldsymbol{\xi} = \psi\boldsymbol{\xi}_0$ results in OF being multiplied by ψ. It follows that $OF < 0$ (actually $OF \to -\infty$ since we could let $\psi \to \infty$). It can be shown similarly if $OF = 0$ or $OF > 0$ that $\boldsymbol{\xi} = \psi\boldsymbol{\xi}_0$ does not change the sign of OF. Therefore, we set $\xi_j = \gamma$ where $\gamma > 0$ and without restriction of generality we let $j = 1$. This changes the optimization to the following linear program:

$$\begin{aligned}OF1 = \quad & \min -\sum_{i\in U_1}\mathbf{x}^i\boldsymbol{\xi} + \sum_{i\in L_1}\mathbf{x}^i\boldsymbol{\xi} + \sum_{i\in Z_1}\eta_i^+ + \eta_i^-\\ \text{such that} \quad & \mathbf{X}_{Z_1}\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- = \mathbf{0}\\ & \xi_1 = \gamma\\ & \boldsymbol{\xi} \text{ free}, \quad \boldsymbol{\eta}^+ \ge \mathbf{0}, \quad \boldsymbol{\eta}^- \ge \mathbf{0}.\end{aligned}$$

All basic feasible solutions to this linear program are of the form $\boldsymbol{\xi} = \binom{1\ \ \mathbf{0}}{\mathbf{X}_B}^{-1}\binom{\gamma}{\mathbf{0}}$ where $\mathbf{X}_B$ is a $(p-1)\times(p)$ submatrix of $\mathbf{X}_{Z_1}$ with rank $p-1$ and we assume that the square matrix $\binom{1\ \ \mathbf{0}}{\mathbf{X}_B}$ of order p is of full rank. (Note that if a row of $\mathbf{X}$ is of the form $(1\ \ \mathbf{0})$ then this would only reduce the number of possible basic feasible solutions.) By the fundamental theorem of linear programming (c.f. [12], Theorem 1), a solution with $OF1 \le 0$ exists if and only if a basic feasible solution exists with $OF1 \le 0$. Thus, if $OF1 \le 0$ then $OF \le 0$. We note that if an optimal solution to OF had $\xi_1 = \gamma < 0$ then it is possible that $OF \le 0$ when $OF1 > 0$ since $\gamma > 0$ in the linear program. However, this is no concern to us since we must consider all possible three-way partitions and we point out that switching the roles of U_1 and L_1 effectively changes the sign of ξ_1. In order to determine the q-stability of the design matrix $\mathbf{X}$ (and thus the finite sample breakdown point of LAD regression), we must consider all three-way partitions U, L, Z of N where $U \cap L = \emptyset$, $Z = N - U - L$, and $q = |U \cup L|$. Since there are at most n possible rows of $\mathbf{X}$ that could be included in any $\mathbf{X}_Z \subset \mathbf{X}$ there are at most $\binom{n}{p-1}$ possible subsets of all $\mathbf{X}_Z \subset \mathbf{X}$ that must be considered. Therefore, the proposition follows. □

4. AN EQUIVALENT PROBLEM

In this section, we show that the problem *NLMIP* is equivalent to the following problem:

$$\begin{aligned}EQMIP \quad & \max_{\mathbf{w}} \min_{\mathbf{z},\boldsymbol{\xi}} \sum_{i=1}^n z_i && (17)\\ \text{such that} \quad & \sum_{i=1}^n |w_i\mathbf{x}^i\boldsymbol{\xi}|z_i \ge 0.5\sum_{i=1}^n |w_i\mathbf{x}^i\boldsymbol{\xi}| && (18)\\ & w_i \le 1 \quad \text{for } i = 1,\ldots,n && (19)\\ & \boldsymbol{\xi} \ne 0, \quad z_i \in \{0,1\} \text{ for } i = 1,\ldots,n. && (20)\end{aligned}$$

To gain insight into this problem, let us make a substitution of $\alpha_i = w_i\mathbf{x}^i\boldsymbol{\xi}$ and let us also make a simplifying assumption that the α_i are known quantities. Then by a simple substitution of $z_i = 1 - \Gamma_i$, it is possible to show that EQMIP is equivalent to

$$\begin{aligned}& \max \sum_{i=1}^n \Gamma_i && (21)\\ \text{such that} \quad & \sum_{i=1}^n |\alpha_i|\Gamma_i \le 0.5\sum_{i=1}^n |\alpha_i| && (22)\\ & \Gamma_i \in \{0,1\} \quad \text{for } i = 1,\ldots,n. && (23)\end{aligned}$$

This problem is exactly the well-known knapsack problem, in which a weight $|\alpha_i|$ is associated with the ith object and the hiker is trying to maximize the number of objects carried, while limiting herself to half of the total weight of all of the objects. The solution to this problem is very trivial. We first order the objects by nondecreasing weights and keep selecting them one by one without exceeding half the total weight of all of the objects. In reality, of course, the weights $|\alpha_i|$ are not known to us and they are determined by several unknown variables ($\boldsymbol{\xi}$ and $\mathbf{w}$). Further complications arise because of the max–min nature of the objective function. Nevertheless, the demonstration of equivalence and the insight that we gain from understanding the problem in this

manner play an important role in the development of the algorithm to follow. In the following proposition, we show that the two problems, NLMIP and EQMIP, are equivalent in the sense that there is a simple method of constructing the optimal solution of one when the optimal solution of the other is known.

PROPOSITION 2: The problems *NLMIP* and *EQMIP* are equivalent.

PROOF: Let us assume that the optimal solution of *EQMIP* is known. Let $\mathcal{Z}$ denote the subset of indices from the set $\{1,\ldots,n\}$ for which $z_i = 1$. We now construct a solution for *NLMIP* from the solution of *EQMIP* as follows:

$$\text{if } i \notin \mathcal{Z} \text{ and } \alpha_i < 0, \text{ then set } \eta_i^+ = -\alpha_i, \tag{24}$$

$$\text{if } i \notin \mathcal{Z} \text{ and } \alpha_i \geq 0, \text{ then set } \eta_i^- = \alpha_i, \tag{25}$$

$$\text{if } i \in \mathcal{Z} \text{ and } \alpha_i < 0, \text{ then set } s_i = -\alpha_i \text{ and } u_i = 1, \text{ and} \tag{26}$$

$$\text{if } i \in \mathcal{Z} \text{ and } \alpha_i \geq 0, \text{ then set } t_i = \alpha_i \text{ and } \ell_i = 1. \tag{27}$$

Set all other variables in *NLMIP* to zero. Let us denote the set of indices i satisfying (24)–(27) by $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$, and $\mathcal{G}_4$, respectively. Note that $\mathcal{G}_3 \cup \mathcal{G}_4 = \mathcal{Z}$ and $\mathcal{G}_1 \cup \mathcal{G}_2 = \mathcal{Z}^c$ (i.e., the complement of the set $\mathcal{Z}$). Let $\alpha = \sum_{i=1}^n |\alpha_i|$. Let $M = \max_{i\in\mathcal{Z}} |\alpha_i|$ and let $\varepsilon = \sum_{i\in\mathcal{Z}} |\alpha_i|$. We first show that $\varepsilon > 0$. If we assume the contrary, i.e., $\varepsilon = 0$, then it implies that the left-hand side of (18) is zero and, consequently, so is the right-hand side of (18) and the optimal value of the objective function is 0. This means that for each i, either $w_i = 0$ or $\mathbf{x}^i\boldsymbol{\xi} = 0$. Note that the assumptions that $\mathbf{X}$ is in general position and $\boldsymbol{\xi} \neq 0$ imply that at least one of the $\mathbf{x}^i\boldsymbol{\xi} \neq 0$. Then for that i, we can choose $w_i = 1$ and $z_i = 1$ and show that the optimum value of the objective can be increased to 1, which in turn implies that the solution could not be optimal. We now show that the solution constructed for *NLMIP* is feasible. It is fairly obvious from the construction of the solution that constraints (8)–(15) are satisfied, except the inequality in (12). To show that this holds, we note that

$$\begin{aligned}
&\sum_{i=1}^n (\eta_i^+ + \eta_i^- - s_i - t_i) \\
&\quad = \sum_{i\in\mathcal{G}_1} \eta_i^+ + \sum_{i\in\mathcal{G}_2} \eta_i^- - \sum_{i\in\mathcal{G}_3} s_i - \sum_{i\in\mathcal{G}_4} t_i && (28)\\
&\quad = \sum_{i\notin\mathcal{Z}} |\alpha_i| - \sum_{i\in\mathcal{Z}} |\alpha_i| \text{ (from } (24-27)) && (29)\\
&\quad = \sum_{i=1}^n |\alpha_i| - 2\sum_{i\in\mathcal{Z}} |\alpha_i| && (30)\\
&\quad \leq 0 \text{ (from (18))}. && (31)
\end{aligned}$$

So far we have shown that the optimal solution for *EQMIP* is feasible for *NLMIP*. To show that this solution is also optimal for *NLMIP*, we assume the contrary, i.e., assume that there is a solution to *NLMIP* that has a higher value of the objective function than the optimal solution to *EQMIP*. From this supposed solution of *NLMIP*, we construct a new solution for *EQMIP* by setting $z_i = u_i + \ell_i$. It is easy to verify that the new solution is feasible for *EQMIP* by observing the following: (a) An optimal solution to *NLMIP* has an equivalent solution, which is also optimal, in which at most one of the four variables η_i^+, η_i^-, s_i, and t_i for any i can be positive. (b) When η_i^+ or η_i^- is positive, the corresponding $u_i = \ell_i = 0$. (c) When s_i (or t_i) is positive, then $u_i = 1$ (or $\ell_i = 1$). (d) Using (a), (b), and (c), one reverses the arguments in (28)–(31) to show that the solution of *EQMIP* so obtained satisfies the constraint (18). With this construction, it is now easy to see that this solution has an optimal value of the objective function that is higher than that of the assumed optimal solution of *EQMIP*. This contradiction shows that the optimal solution of *EQMIP* must also be an optimal solution of the *NLMIP*.

To finish the proof, one must now show that a solution of *NLMIP* has an equivalent solution that is also an optimal solution of *EQMIP*. The construction of this is done by setting $z_i = u_i + \ell_i$, and the proof of optimality can be obtained by reversing the arguments given above. □

5. THE CASE OF UNIFORM DESIGN SIMPLE LINEAR REGRESSION

In this section, we discuss the special case of simple linear regression ($p = 2$) with a uniform design, i.e., $x_1^i = 1$ and $x_2^i = i$, for $i = 1,\ldots,n$ (there is no loss of generality in taking the values of x_2^i as the integers; any equispaced set of values yields the same weights). We refer to this problem as the *uniform simple linear* problem. We show how to obtain an exact solution for this problem. The problem stated in (17)–(20) without any restriction on w_i now becomes

$$USLMIP \quad \max_{\mathbf{w}} \min_{\mathbf{z},\boldsymbol{\xi}} \sum_{i=1}^n z_i \tag{32}$$

$$\text{such that} \quad \sum_{i=1}^n |w_i(\xi_1 + i\xi_2)| z_i \geq 0.5 \sum_{i=1}^n |w_i(\xi_1 + i\xi_2)| \tag{33}$$

$$\boldsymbol{\xi} \neq 0, \quad z_i \in \{0,1\} \text{ for } i = 1,\ldots,n. \tag{34}$$

A crucial result for determining the optimal values of w_i and $\boldsymbol{\xi}$ is given by the following lemma.

LEMMA 1: For the problem $USLMIP$, there is an optimum solution of w_i, which is symmetrical, i.e., $w_i = w_{n-i}$ for $i = 1, \ldots, n$.

PROOF: Define $\phi_i = \xi_1 + i\xi_2$ and let the optimal values of $\xi_1 = \delta_1$ and $\xi_2 = \delta_2$; this implies that $\phi_i = \delta_1 + i\delta_2$. Construct a new solution by selecting $\xi_1 = \delta_1 + n\delta_2$ and $\xi_2 = -\delta_2$. We refer to the function ϕ for the new solution as $\bar{\phi}$ to distinguish it from the original solution. Then $\bar{\phi}_i = \delta_1 + (n-i)\delta_2$, or $\phi_i = \bar{\phi}_{n-i}$. Note that as far as the regression problem is concerned, it is as if the two solutions differ only in the arrangements of the entries of the $\mathbf{x}\xi$ vector. This means that there is an optimal solution for which $w_i = w_{n-i}$ since a symmetric choice of w_i will cater to both choices of ϕ and $\bar{\phi}$. □

Based on the above result, we limit ourselves to symmetric linear functions for w_i. More explicitly, the functions we choose to examine are linear functions of the distance from the center of the range of $\mathbf{x}$, being of the form

$$w_i = \begin{cases} w^0 + i w^1 & \text{if } i \le n/2 \\ w^0 + (n-i)w^1 & \text{if } i > n/2. \end{cases}$$

We note that the problem defined in (32)–(34) is invariant in scaling for $\mathbf{w}$ and $\boldsymbol{\xi}$. Therefore, without loss of generality, we can impose two conditions such as $w^0 = \xi_1 = 1$. Thus, the simple linear problem reduces to the problem of finding just two parameters w^1 and ξ_2. Clearly, the two unknown quantities also depend on the value of n, the size of the problem. To remove this dependency, we convert the problem in (32)–(34) to an equivalent continuous problem in which the variable i is replaced with a variable x, where $0 \le x \le 1$. After solving this problem by conducting a search over the two unknown parameters, we reconvert the solution to the discrete case. The solutions obtained are $w^1 = 4/n$ and $\xi_2 = -1.35/n$. The optimal value of the objective function in the continuous version of the problem is $0.3005n$. This shows that the finite sample breakdown point of WLAD regression can reach over 30%. Note that if we simplified the problem by not considering the weights $\mathbf{w}$ (i.e., $w_i = 1$ for $i = 1, \ldots, n$), then the optimum breakdown for the continuous version of this problem is $0.25n$ (c.f. [2]). This implies that for the uniform simple linear problem, the breakdown can be increased by about 20% by a judicious choice of weights.

The solution to the uniform simple linear problem after normalizing the weights such that $w_i \le 1$ for $i = 1, \ldots, n$ is

$$w_i = \begin{cases} \frac{1+i\left(\frac{4}{n}\right)}{1+\lfloor \frac{n}{2} \rfloor \left(\frac{4}{n}\right)} & \text{if } i \le n/2 \\ \frac{1+(n-i)\left(\frac{4}{n}\right)}{1+\lfloor \frac{n}{2} \rfloor \left(\frac{4}{n}\right)} & \text{if } i > n/2. \end{cases}$$

Thus, the selected weights range from approximately $\frac{1}{3}$ to 1. Our algorithm for the determination of general weights given in the next section is based upon this solution for the uniform simple linear problem.

6. WEIGHTS FOR GENERAL REGRESSION DESIGNS

In this section, we describe a general weight-determination algorithm and describe in more detail how *NLMIP* can be solved once the weights are chosen. The key idea in generalizing the weights is to note that the symmetric linear weights have the property that the weights decrease linearly with the distance from the "center" of the range of the predictor. The proposed general weights have a corresponding property, decreasing linearly with the sum of the distances for the predictors from the coordinatewise median of the observation, where the distances are scaled by the range of each predictor. The median is chosen as the center of the data since the LAD estimator for univariate location is the median. This yields the following algorithm for choosing the weights.

Algorithm for Choosing Weights

Step 1. Let $z_j^i = (x_j^i - \min_i\{x_j^i\})/\max_i\{x_j^i\}$, $j = 1, \ldots, p$ (if $\mathbf{x}_1$ is a column of ones, set $z_1^i = 0$ for all i). For $j = 1, \ldots, p$, let m_j denote the median of the entries z_j^i for $i = 1, \ldots, n$.

Step 2. Let $r_i = \sum_{j=1}^{p} |z_j^i - m_j|$.

Step 3. Let $r_i = (r_i - \min_v r_v)/(\max_v r_v - \min_v r_v)$.

Step 4. Let $w_i = 1 - \frac{2}{3} r_i$.

Note that steps 3 and 4 of the algorithm guarantee that these weights are (virtually) identical to those derived earlier for the uniform simple linear case when the data take that form.

Once the weights are chosen, the results of Section 3 provide a way of approximately solving *EQMIP* and hence *NLMIP*. The details are as follows.

Algorithm for solving *EQMIP*

Step 1. Let $\mathcal{N}$ denote the set $\{1, \ldots, n\}$. Let $\mathcal{K}_k$ denote all of the subsets of $\mathcal{N}$, such that $|\mathcal{K}_k| = p - 1$ for $k = 1, \ldots, r$. Clearly, $r = n!/((p-1)!(n-p+1)!)$. For $k = 1, \ldots, r$, solve the following system of linear equations for the unknown $p \times 1$ vector $\boldsymbol{\xi}^k$:

$$w_i \mathbf{x}^i \boldsymbol{\xi}^k = 0 \text{ for } i \in \mathcal{K}_k$$

and $\xi_1^k = 1$. Note that this system is guaranteed to have a unique solution, based on the assumptions on $\mathbf{x}^i$. Let

$$\alpha_i^k = w_i \mathbf{x}^i \boldsymbol{\xi}^k.$$

Step 2. Reorder the elements $|\alpha_i^k|$ for $i = 1, \ldots, n$ in decreasing order so that $|\alpha_i^k|$ now denotes the ith order statistic. Identify the smallest index $m^*(k)$ satisfying

$$\sum_{i=1}^{m^*(k)} |\alpha_i^k| \ge 0.5 \sum_{i=1}^{n} |\alpha_i^k|.$$

Step 3. Find $m^* = \min_k m^*(k)$. Let k^* be the value of k for which this is minimum. The solution is given by the following:

if $i > m^*$ and $\alpha_i^{k^*} < 0$, then set $\eta_i^+ = -\alpha_i^{k^*}$,
if $i > m^*$ and $\alpha_i^{k^*} \geq 0$, then set $\eta_i^- = \alpha_i^{k^*}$
if $i \leq m^*$ and $\alpha_i^{k^*} < 0$, then set $s_i = -\alpha_i^{k^*}$ and $u_i = 1$, and
if $i \leq m^*$ and $\alpha_i^{k^*} \geq 0$, then set $t_i = \alpha_i^{k^*}$ and $\ell_i = 1$.

Set all other variables to zero.

7. EXAMPLES

In this section we illustrate the benefits of weighting using the proposed algorithm. We first make some brief comments about computation times for the weighted LAD estimator. Figure 1 gives CPU times (in seconds) on a Pentium 4 PC (3.2-GHz processor) for a range of sample sizes (1000, 5000, 10000, and 50000) and number of predictors (2, 20, 50, and 100), plotted on a log–log scale. Figure 1 gives times for WLAD estimation using the Barrodale–Roberts modified simplex algorithm (solid line), WLAD estimation using the

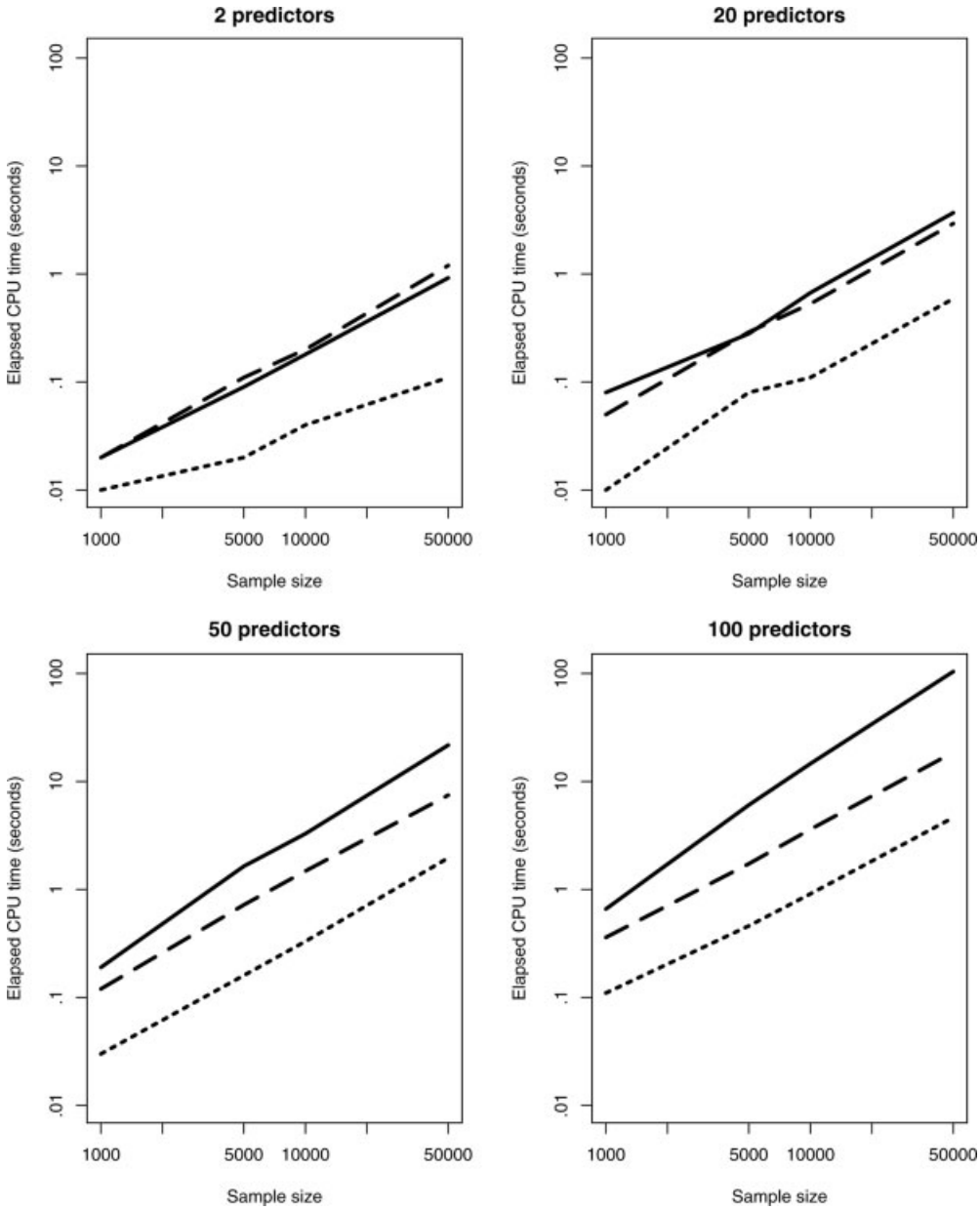

Figure 1. Timings for WLAD estimation using simplex method (solid line), WLAD estimation using interior point method (dashed line), and least squares estimation (dotted line).

Frisch–Newton interior point algorithm (dashed line), each based on the `quantreg` package ([8]), and least squares estimation (dotted line), using the statistical package R ([14]). Calculation of weights in all cases took less than 5 s (and usually less than 1 s), so the times for WLAD estimation come primarily from the LAD optimization calculations. As can be seen, while (weighted) LAD estimation is, as expected, more computationally intensive than least squares, use of the interior point algorithm makes it quite feasible even for large problems; the total time is less than 20 s for the largest problem examined, and it takes only roughly four times as much time as least squares when there are 100 predictors.

We now determine the breakdown points of (weighted) LAD estimators for 25 one-, two-, and four-predictor designs that cover a wide range of possible patterns. The one- and two-predictor designs are based on $n = 500$ observations, while the four-predictor designs are based on $n = 100$ observations. In the one-predictor case, we generate n observations from either a uniform (on the interval $(0, 1)$), exponential (with mean 1) or a normal distribution, as these represent varying degrees of nonuniformity in the design. We also consider the situation where a certain percentage (either 10 or 20%) of the observations are adjusted by adding $N(4, 0.5^2)$ values to the original predictor values. The existence of unusual values for the predictors (called leverage points in the statistics literature) is of particular interest, since it is well known that LAD regression is very sensitive to leverage points. Recall that, as noted earlier, the least squares estimator has breakdown point $1/n$ for all designs.

We also examine two-predictor and four-predictor (multiple) regressions by generating two predictors using the six possible combinations of uniform, exponential, and normal distributions for two predictors and designs where all predictors were uniform or all were exponential for four predictors. We also examine the effects of leverage points by adjusting 20% of the observations as described above.

The results are given in Table 1. For each design the breakdown point (expressed as a percentage of the total sample size) is given for LAD, WLAD based on applying the algorithm for choosing weights once, and WLAD based on applying the algorithm for choosing weights iteratively. Iterating is done by treating the weighted design $\mathbf{WX}$ (where $\mathbf{W}$ is a diagonal matrix of weights) as the current design and reapplying the algorithm. At each iteration, the breakdown point is determined, and iteration continues until the breakdown point stops increasing (the total number of iterations in each case is given in parentheses in Table 1).

Obviously LAD is a special case of WLAD, so equal weights can be used (that is, ordinary LAD) if the breakdown after weighting is lower than that of the LAD estimator,

Table 1. Breakdown points for various designs.

Design	Leverage(%)	LAD breakdown point (%)	WLAD breakdown point (%) (1 iteration)	WLAD breakdown point (%) (No. of iterations)
Exponential		14.0	17.8	26.6 (4)
Exponential	10	11.2	14.6	27.0 (4)
Exponential	20	14.6	17.2	27.0 (4)
Normal		23.0	27.0	30.2 (2)
Normal	10	14.8	22.4	29.6 (2)
Normal	20	15.0	20.0	29.8 (2)
Uniform		24.6	29.8	29.8 (1)
Uniform	10	7.2	10.4	27.4 (3)
Uniform	20	11.8	14.8	26.8 (3)
Exponential/exponential		14.0	20.6	23.6 (4)
Exponential/exponential	20	13.4	20.0	21.6 (4)
Exponential/normal		14.0	16.2	19.6 (4)
Exponential/normal	20	13.4	16.6	21.2 (2)
Exponential/uniform		14.0	17.4	21.0 (3)
Exponential/uniform	20	11.6	13.8	21.0 (4)
Normal/normal		22.8	25.4	25.8 (2)
Normal/normal	20	13.2	17.4	21.4 (2)
Normal/uniform		23.0	25.2	25.2 (1)
Normal/uniform	20	11.6	13.6	21.6 (3)
Uniform/uniform		23.4	25.4	25.4 (1)
Uniform/uniform	20	11.0	13.2	20.6 (3)
Uniform × 4		20.0	22.0	22.0 (1)
Uniform × 4	20	11.0	13.0	18.0 (3)
Exponential × 4		13.0	15.0	18.0 (3)
Exponential × 4	20	13.0	17.0	17.0 (1)

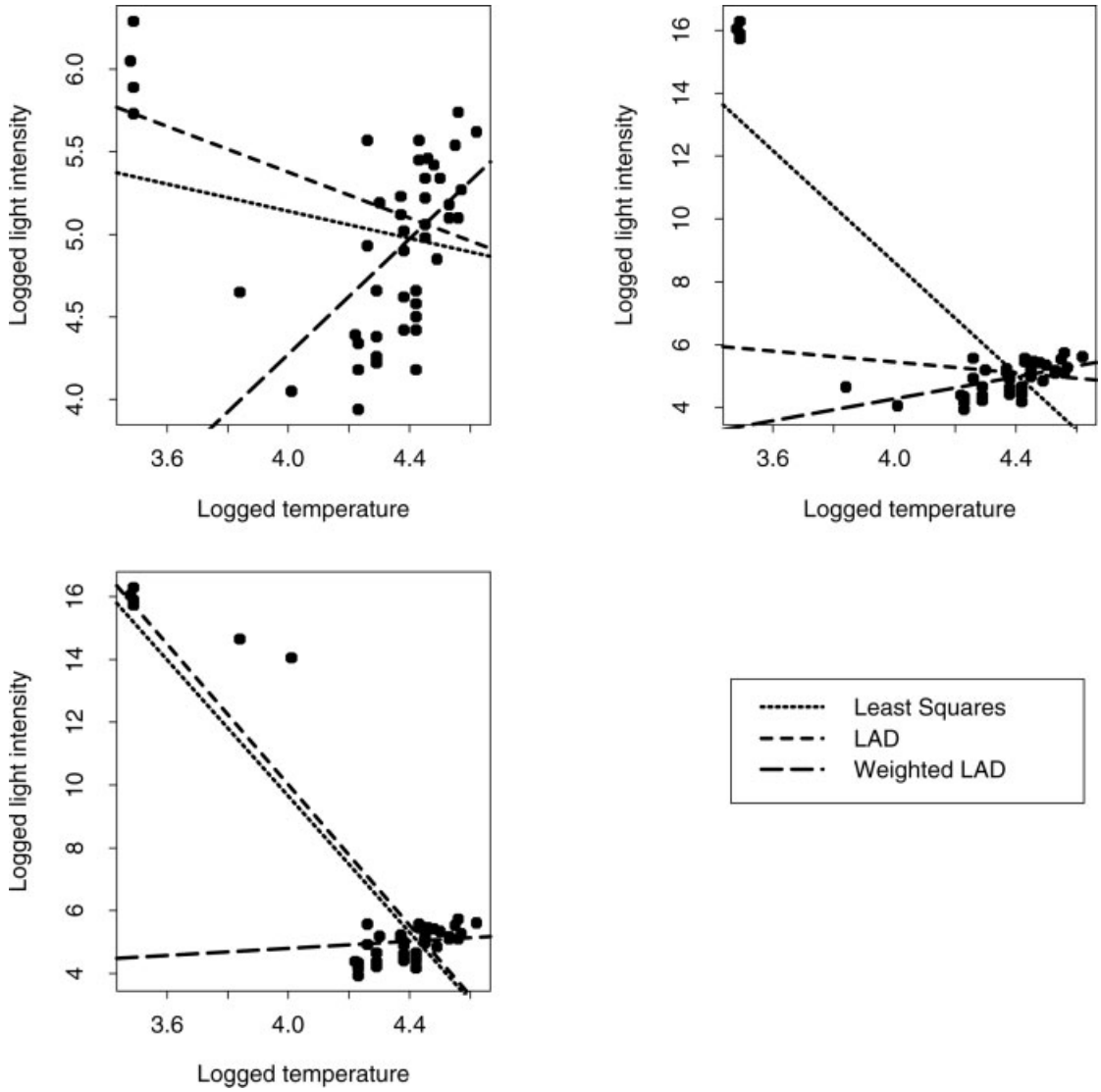

Figure 2. Stars data and modifications, with three regression lines.

providing an effective lower bound on the breakdown point of WLAD estimation. The WLAD estimator is a member of the class of generalized M-estimators, for which it is known that the asymptotic (infinite sample size) breakdown cannot exceed $c_p/(c_p + 1)$, where c_p satisfies the recursion

$$c_p = (2/\pi)/[(p-1)c_{p-1}]$$

with $c_1 = 1$, for certain contamination models and spherically symmetric predictors ([9]). This provides upper bounds (assuming infinite sample sizes) on the breakdown point of 0.39, 0.33, and 0.27, respectively, for the one-, two-, and four-predictor cases examined here. The breakdown points were computed exactly in Table 1, but in the situation where a user is faced with designs too large to determine the breakdown exactly, we suggest utilizing the heuristic for finding a good upper bound of the breakdown suggested by Giloni and Padberg ([4]) and then iterating until the upper bound no longer increases (of course, another possibility would be to just iterate once).

Table 1 makes clear the benefits of weighting. Just one iteration of the weighting algorithm typically increases the breakdown point 3–5 percentage points and as much as 6–7 percentage points when there are leverage points. Even more impressively, iterating the weighting algorithm leads to gains in breakdown of at least 5 percentage points in most cases and as much as 10–15 percentage points in many, sometimes approaching the asymptotic upper bound noted earlier. The WLAD estimator is much more robust than the LAD estimator and therefore more trustworthy as a routine regression tool.

We conclude with discussion of a well-known data set from the robust regression literature, the so-called Stars data set ([15], p. 27). Figure 2 contains three graphs. At the top left is a scatter plot of the original data, which is a plot of the logarithm of the light intensity versus the logarithm of the temperature at the surface of 47 stars in the star cluster CYG OB1. The plot (called a *Hertzsprung–Russell star chart*) also includes the least squares regression line, the LAD regression line, and the WLAD regression line using the weights defined by our weight selection algorithm, including iteration. There are four obvious outliers in this data set all with logged temperature approximately equal to 3.5. These outlying data points are what are referred to as "red giants," as opposed to the rest of the stars, which are considered to lie in the "main sequence." It is apparent that the least squares line is drawn toward the red giants, which is not surprising, given its breakdown point of $1/47$. By contrast, the WLAD line is unaffected by the outliers and goes through the main sequence. The WLAD line has breakdown $10/47$ if only one iteration of the weighting algorithm is used and increases to $13/47$ with two iterations.

It is interesting to note, however, that the LAD line is also drawn toward the outliers, despite the fact that the LAD line has not broken down (its breakdown point is $5/47$, so the four outliers are not enough to break it down). This illustrates that weighting can be helpful if outliers occur at leverage points, even if the LAD estimator has not broken down. The top right plot reinforces that LAD has not broken down. In this plot the light intensities of the four giants have been further contaminated (by adding 10 to the values). The least squares line follows the outliers, but the LAD line is virtually identical to what it was in the original plot, since it has not broken down (although it still does not follow the main sequence, as the WLAD line still does). In the third plot of the figure, two more stars have their light intensities contaminated. Since there are now six outliers, the LAD line breaks down and is as poor as the least squares line, while the weighted LAD line is still resistant to the outliers.

8. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this paper, we have demonstrated that WLAD regression is a regression technique whose robustness properties can be studied by mathematical programming methods. We have developed a computationally feasible method for calculating the optimum weights and have demonstrated that the optimum breakdown can be significantly increased by a judicious choice of weights. These results leave open the statistical properties of WLAD regression, which would be fruitful topics for further research. These include study of the asymptotic properties of WLAD regression, its small-sample properties, and investigation of whether nonlinear weights might lead to better performance.

ACKNOWLEDGMENTS

We thank two anonymous referees and the Associate Editor for comments and suggestions that helped improve the paper.

REFERENCES

[1] D.L. Donoho and P.J. Huber, "The notion of breakdown point," A Festschrift for Erich Lehmann, P. Bickel, K. Doksum, and J.L. Hodges (Editors), Wadsworth, Belmont, CA, 1983, pp. 157–184.

[2] S.P. Ellis and S. Morgenthaler, Leverage and breakdown in L_1-regression, J Am Statist Assoc 87 (1992), 143–148.

[3] A. Giloni and M. Padberg, Alternative methods of linear regression, Math Comput Model 35 (2002), 361–374.

[4] A. Giloni and M. Padberg, The finite sample breakdown point of ℓ_1-regression, SIAM J Optimiz 14 (2004), 1028–1042.

[5] A. Giloni and J.S. Simonoff, The conditional breakdown properties of robust local polynomial estimators, J Nonparametric Statist 17 (2005), 15–30.

[6] F.R. Hampel, Contributions to the theory of robust estimation, Ph.D. Thesis, University of California, Berkeley, 1968.

[7] X. He, J. Jureckova, R. Koenker, and S. Portnoy, Tail behavior of regression estimators and their breakdown points, Econometrica 58 (1990), 1195–1214.

[8] R. Koenker, quantreg: Quantile Regression, R package version 3.70, 2004 (http://www.econ.uiuc.edu/roger/research/rq/rq.html).

[9] R.A. Maronna, O.H. Bustos, and V.J. Yohai, "Bias- and efficiency-robustness of general M-estimators for regression with random carriers," Smoothing techniques for curve estimation, T. Gasser and M. Rosenblatt (Editors), Springer-Verlag, New York, 1989, pp. 91–116.

[10] I. Mizera and C.H. Müller, Breakdown points and variation exponents of robust M-estimators in linear models, Ann Statist 27 (1999), 1164–1177.

[11] I. Mizera and C.H. Müller, "The influence of the design on the breakdown points of ℓ_1-type M-estimators," MODA6—Advances in model-oriented design and analysis, A. Atkinson, P. Hackl, and W. Müller (Editors), Physica-Verlag, Heidelberg, 2001, pp. 193–200.

[12] M. Padberg, Linear optimization and extensions, Springer-Verlag, Berlin, 1995.

[13] S. Portnoy and R. Koenker, The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators, Statist Sci 12 (1997), 279–300.

[14] R Development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2004 (`http://www.R-project.org`).

[15] P.J. Rousseeuw and A.M. Leroy, Robust regression and outlier detection, Wiley, New York, 1987.

[16] W.F. Sharpe, Mean-absolute-deviation characteristic lines for securities and portfolios, Manage Sci 18 (1971), B1–B13.

[17] J.S. Simonoff, Smoothing methods in statistics, Springer-Verlag, New York, 1996.

[18] F.T. Wang and D.W. Scott, The L_1 method for robust nonparametric regression, J Am Statist Assoc 89 (1994), 65–76.

[19] K. Yu and M.C. Jones, Local linear quantile regression, J Am Statist Assoc 93 (1998), 228–237.