Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 3124–3140
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Robust weighted LAD regression

Avi Giloni[a], Jeffrey S. Simonoff[b,*], Bhaskar Sengupta[c,1]

[a] *Sy Syms School of Business, Yeshiva University, 500 West 185th Street, New York, NY 10033, USA*
[b] *Leonard N. Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012, USA*
[c] *Complex Systems Modeling, ExxonMobil Research and Engineering, 1545 Route 22 East, Annandale, NJ 08801, USA*

Received 13 February 2005; received in revised form 6 June 2005; accepted 7 June 2005
Available online 11 July 2005

Abstract

The least squares linear regression estimator is well-known to be highly sensitive to unusual observations in the data, and as a result many more robust estimators have been proposed as alternatives. One of the earliest proposals was least-sum of absolute deviations (LAD) regression, where the regression coefficients are estimated through minimization of the sum of the absolute values of the residuals. LAD regression has been largely ignored as a robust alternative to least squares, since it can be strongly affected by a single observation (that is, it has a breakdown point of $1/n$, where n is the sample size). In this paper we show that judicious choice of weights can result in a weighted LAD estimator with much higher breakdown point. We discuss the properties of the weighted LAD estimator, and show via simulation that its performance is competitive with that of high breakdown regression estimators, particularly in the presence of outliers located at leverage points. We also apply the estimator to several data sets.

Keywords: Breakdown point; Leverage points; Outliers; Robust regression

* Corresponding author. Tel.: +1 2129980452; fax: +1 2129954003.
E-mail addresses: agiloni@ymail.yu.edu (A. Giloni), jsimonof@stern.nyu.edu (J.S. Simonoff), Bhaskar.Sengupta@exxonmobil.com (B. Sengupta).

[1] This work was done while he was at Yeshiva University.

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.06.005

1. Introduction

The linear regression problem is certainly one of the most important data analysis situations (if not the most important). In this situation the data analyst is presented with n observations of a response variable y and some number p of predicting variables $x_1, \ldots, x_p$ satisfying

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

where $\boldsymbol{\epsilon}$ is a vector of errors and

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

$$\boldsymbol{\epsilon} = \begin{pmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

(note that usually $\mathbf{x}_1$ is a column of ones, but this is not required). We assume that $\mathbf{X}$ is of full rank (that is, $r(\mathbf{X}) = p$).

It is well-known that if the errors $\boldsymbol{\epsilon}$ are normally distributed with constant variance the optimal estimator of $\boldsymbol{\beta}$ is the ordinary least squares (OLS) estimator, based on minimizing the ℓ_2-norm $\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2 = \sum_{i=1}^n (y_i - \mathbf{x}^i\widehat{\boldsymbol{\beta}})^2$ of the residuals. Unfortunately, it is also well-known that the least squares estimator is very nonrobust, being highly sensitive to unusual observations in the y space (outliers) and $\mathbf{X}$ space (leverage points).

A way of quantifying this sensitivity is through the notion of the *breakdown point* of a regression estimator (Hampel, 1968). Suppose we estimate the regression parameters $\boldsymbol{\beta}$ by some technique τ from data $(\mathbf{X}, \mathbf{y})$, yielding the estimate $\boldsymbol{\beta}^\tau$. If we contaminate m $(1 \leqslant m < n)$ 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

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

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 then defined as

$$\alpha(\tau, \mathbf{X}, \mathbf{y}) = \min_{1 \leqslant m < n} \left\{ \frac{m}{n} : b(m, \tau, \mathbf{X}, \mathbf{y}) \text{ is infinite} \right\}, \tag{2}$$

the minimum proportion of rows of $(\mathbf{X}, \mathbf{y})$ that if replaced by arbitrary new data make the regression technique τ break down. The minimum possible value of the breakdown

point is $1/n$ and the maximum value is .5 (the latter value since otherwise it is impossible to distinguish between the uncontaminated data and the contaminated data). Clearly, the larger the breakdown point, the more robust is the regression estimator.

The least squares regression estimator has breakdown $1/n$, and many alternative estimators have been proposed to provide more robust regression estimation. One of the earliest proposals was regression performed through minimization of the ℓ_1 norm of the residuals, $\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_1 = \sum_{i=1}^{n} \left|y_i - \mathbf{x}^i\widehat{\boldsymbol{\beta}}\right|$, also called least-sum of absolute deviations (LAD) regression. Dielman (2005) gives an extensive review of research related to LAD regression. There is a good deal of empirical evidence going back more than 30 years that LAD regression is more robust than OLS in the presence of fat-tailed errors (see, e.g., Sharpe, 1971). Despite this, LAD regression has gained relatively little favor in the robustness statistical literature, for several perceived reasons:

1. LAD estimation is more than difficult than OLS estimation, because of the nondifferentiability of the criterion function. In fact, as Portnoy and Koenker (1997) pointed out, modern linear programming techniques have made the computation time of LAD regression competitive with, or even superior to, OLS even for extremely large data sets. This can be contrasted with high breakdown regression estimators, which are far more computationally intensive than OLS (Rousseeuw and Leroy, 1987; Hawkins and Olive, 2002). As an example of the gains these new methods can provide, LAD estimation for a data set with 100,000 observations and 100 variables takes more than 12 min using the Barrodale–Roberts modified simplex algorithm and only 30 s using the Frisch–Newton interior point algorithm on a Pentium 4 PC (3.2 GHz processor) using the `quantreg` package (Koenker, 2004).
2. The asymptotic theory for LAD regression is not as well-developed as for OLS regression. While this is true to a certain degree, it is also true for high-breakdown regression estimators.
3. The LAD regression estimator is not at all robust to observations with unusual predictor values; that is, it has a low breakdown point.

It is this final claim that is the starting point for this paper. In the next section, we discuss what is known about the breakdown of LAD regression, focusing on recent exact results. We discuss how the use of a weighted version of LAD regression can improve its robustness considerably, and describe an algorithm for choosing weights that leads to such improvements. In Section 3 we give the asymptotic properties of the resultant estimators, and examine finite-sample properties using Monte Carlo simulations. We apply the method to several data sets in Section 4. We conclude the paper with discussion of potential future work.

2. LAD regression and breakdown

If contamination is permitted in both the predictor and response variables, the breakdown point of LAD is the same as that of OLS, $1/n$ (see, e.g., Rousseeuw and Leroy, 1987, p. 12). LAD regression is, however, more robust than least squares in the following sense. Define

the finite sample breakdown point of the LAD estimator to be 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 $\alpha(\tau, \mathbf{y}|\mathbf{X})$. This is the ordinary breakdown (2) for given predictor matrix $\mathbf{X}$. This is a natural criterion in the regression context, since standard regression theory proceeds based on conditioning on the observed values of the predictors. The finite sample breakdown point, or conditional breakdown point, was introduced in Donoho and Huber (1983). He et al. (1990), Mizera and Müller (2001), and Giloni and Padberg (2004) showed that the finite sample breakdown point of LAD regression can be greater than $1/n$, depending on the predictor values, with the second-named authors discussing how $\mathbf{X}$ can be chosen to increase the breakdown point of LAD. Giloni and Padberg (2004) showed that the breakdown can be calculated using a mixed integer program, and described an algorithm for solving it; Giloni et al. (2004) proposed an alternative algorithm similar to that of Mizera and Müller (2001) that is very efficient for large samples when p is small.

Ellis and Morgenthaler (1992) appear to be the first to explicitly show that the introduction of weights can improve the finite sample breakdown point of LAD regression (by downweighting observations that are far from the bulk of the data), but they only do this for very small data sets. Giloni et al. (2004) examined this question in more detail. The weighted LAD (WLAD) regression estimator is defined to be the minimizer of

$$\sum_{i=1}^{n} w_i |y_i - \mathbf{x}^i \boldsymbol{\beta}|. \tag{3}$$

This estimation problem can be formulated as a linear program,

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} w_i \left(r_i^+ + r_i^-\right) \\ \text{such that} \quad & \mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y} \\ & \boldsymbol{\beta} \textit{ free}, \quad \mathbf{r}^+ \geqslant \mathbf{0}, \quad \mathbf{r}^- \geqslant \mathbf{0}. \end{aligned} \tag{4}$$

Here, the residual r_i associated with observation i is multiplied by some weight, w_i, where we assume that $0 < w_i \leqslant 1$ without loss of generality. In the formulation of the linear program the residuals $\mathbf{r}$ are replaced by a difference $\mathbf{r}^+ - \mathbf{r}^-$ of nonnegative variables. Since $\left(r_i^+ - r_i^-\right) = \left(y_i - \mathbf{x}^i \boldsymbol{\beta}\right)$, and by the simplex method either $r_i^+ > 0$ or $r_i^- > 0$ but not both, then $\left|w_i \left(r_i^+ - r_i^-\right)\right| = w_i \left(r_i^+ + r_i^-\right)$. Therefore, transforming the data by setting $\left(\tilde{\mathbf{x}}^i, \tilde{y}_i\right) = w_i \left(\mathbf{x}^i, y_i\right)$, implies $\left(\tilde{y}_i - \tilde{\mathbf{x}}^i \boldsymbol{\beta}\right) = w_i \left(r_i^+ - r_i^-\right)$. This means that the linear program (4) can be reformulated as

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

That is, weighted LAD regression can be treated as LAD regression with suitably transformed data, and determining the breakdown of weighted LAD regression with known weights corresponds to determining the breakdown of LAD regression with data $\left(\tilde{\mathbf{X}}, \tilde{\mathbf{y}}\right)$.

The problem of choosing weights **w** to maximize the breakdown of the resultant weighted LAD estimator is a nonlinear mixed integer program. Giloni et al. (2004) showed that this problem is equivalent to a problem related to the knapsack problem, and solved a specific form of the problem for the simple regression case with a uniform design, resulting in a WLAD breakdown over 30% (LAD regression has a 25% breakdown point in this situation). They also proposed a simple weighting scheme for multiple regression that yields breakdown points over 20% for various two-predictor problems. We propose here an improved weighting scheme that is fast computationally and leads to higher breakdown values. As was noted by Ellis and Morgenthaler (1992), the goal is to downweight observations that are outlying in the predictor space (that is, are leverage points). Unfortunately, usual measures of leverage suffer from masking, in that multiple leverage points near each other can cause them to "hide" each other.

High-breakdown measures of leverage can be calculated, but these are computationally intensive. Rousseeuw and van Zomeren (1992) proposed a simple variation on this idea, identifying leverage points using high-breakdown one-dimensional measures of location and scale, omitting them, and using ordinary LAD regression on the remaining observations (a weighting scheme where all of the weights are either 0 or 1). Hubert and Rousseeuw (1997) proposed a weighted version of LAD regression (which they named the RDL_1 estimator) for regression models including both continuous and binary predictors, but it is also applicable to models based on only numerical predictors. The weights were based on the computationally intensive high breakdown minimum volume ellipsoid estimator (Rousseeuw, 1985). Defining $T(\mathbf{X})$ to be the center of the smallest ellipsoid containing half of $\mathbf{X}$ and $C(\mathbf{X})$ to be the covariance matrix implied by that ellipsoid, weights are based on the robust distances

$$RD\left(\mathbf{x}^i\right) = \sqrt{\left[\mathbf{x}^i - T(\mathbf{X})\right] C(\mathbf{X})^{-1} \left[\mathbf{x}^i - T(\mathbf{X})\right]'}$$

through the relation

$$w_i = \min\left(1,\, p/RD\left(\mathbf{x}^i\right)^2\right).$$

We will instead use conventional measures of leverage (avoiding the high computational burden), but will measure outlyingness relative to what is (hopefully) a clean subset of the data. The size ℓ of the clean subset should be large enough to include much of the data, but small enough so that it does not include outlying observations; we use $\ell = .6n$ here. The notion of identifying outlying observations by defining a clean subset of the data and then measuring the distance of observations relative to that subset was introduced by Rosner (1975) for univariate Gaussian data. The first application of this idea to multivariate and regression data appears to have been in Simonoff (1991). Other applications of this idea for multivariate and regression data can be found in Hadi (1992, 1994), Hadi and Simonoff (1993), Billor et al. (2000), and Atkinson and Riani (2000).

The definition of the clean subset used here is that used by Billor et al. (2000) as version 2 of their algorithm for finding a clean subset of a multivariate data set (p. 285), and as they note this is a robust method of choosing the clean subset. The subset is defined by first calculating $\left\|\check{\mathbf{x}}^i - \check{\mathbf{m}}\right\|$, $i = 1, \ldots, n$, where $\check{\mathbf{x}}^i$ is the ith row of the scaled predictor matrix

$\breve{\mathbf{X}}$ (the matrix $\mathbf{X}$ where each predictor is scaled to be in the range [0, 1] by subtracting the minimum value and then dividing by the maximum value), $\breve{\mathbf{m}}$ is the vector of coordinatewise medians of $\breve{\mathbf{X}}$, and $\|\cdot\|$ is the vector norm. Then, the ℓ observations with smallest distances are used as the clean subset, called $\mathbf{X}_S$.

The set of leverage values for an observation $\mathbf{x}^i$ relative to the clean subset is $h_i = \mathbf{x}^i \left(\mathbf{X}_S'\mathbf{X}_S\right)^{-1}\mathbf{x}^{i\prime}$, using the usual (least squares) notion of leverage. Since the leverage is proportional to the squared Mahalanobis distance (Chatterjee and Hadi, 1988, p. 102) the weight is taken to be $w_i = \sqrt{\min_j \left(h_j\right)/h_i}$; that is, it is inversely proportional to the distance from the clean subset. Since the results of Ellis and Morgenthaler (1992) imply that breakdown is related to distance, rather than squared distance, it seems reasonable to define the weights accordingly. All of the analyses in this paper are based on this weighting scheme.

As Billor et al. (2000) note, this method of choosing the clean subset is not affine invariant, but it is "almost invariant," in that the WLAD estimator based on these weights is itself invariant. Billor et al. (2000) also propose a method of choosing the clean subset based on Mahalanobis distances (their version 1) that is affine invariant, but they note that the method based on coordinatewise medians is more robust. Given the focus here on increasing the robustness properties of LAD estimation, we choose to sacrifice exact invariance for improved robustness.

WLAD regression is a particular example of generalized M (GM)-estimation (Mallows, 1975). It is known that the maximum asymptotic breakdown point of all GM-estimators is a decreasing function of p (Maronna et al., 1979), implying that the breakdown point of WLAD regression cannot be arbitrarily high for models with many predictors. In the next section we use Monte Carlo simulations to study the properties of WLAD regression, and show that despite this, WLAD estimation can be competitive with high breakdown estimation even for reasonably large values of p.

3. The properties of WLAD regression estimation

We begin this section with discussion of the asymptotic properties of the WLAD estimator. Consider again the regression model (1), and assume that the errors satisfy $\boldsymbol{\epsilon} = \boldsymbol{\Delta}\mathbf{u}$, where $\boldsymbol{\Delta} = \text{diag}\,(\delta_1, \ldots, \delta_n)$ is a matrix of known constants, and $\mathbf{u}$ are independent and identically distributed with cumulative distribution function F. Note that this structure allows for the possibility of heteroscedasticity in the errors. Assume that $\max \left\|\mathbf{x}^i\right\| = o\left(n^{1/4}\right)$, F is twice differentiable at 0, and $f(0) = F'(0) > 0$. Let $\hat{\boldsymbol{\beta}}_w$ be the WLAD estimator of $\boldsymbol{\beta}$. Let $\mathbf{W} = \text{diag}\,(w_1, \ldots, w_n)$, and assume that the weights are known positive values that satisfy $\max w_i\delta_i = \text{O}(1)$ and $\max\,(w_i\delta_i)^{-1} = \text{O}(1)$.

Theorem 1. *As* $n \to \infty$, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}\right)$ *is asymptotically p-variate normal with mean* $\mathbf{0}$ *and covariance matrix*

$$\mathbf{Q}^{-1}\left(\mathbf{X}'\mathbf{W}^2\mathbf{X}\right)\mathbf{Q}^{-1}\omega^2,$$

where $\mathbf{Q} = \lim_{n\to\infty} \mathbf{X}'\mathbf{W}\boldsymbol{\Delta}^{-1}\mathbf{X}/n$ *and* $\omega = [2f(0)]^{-1}$.

The proof of this theorem is given in Appendix. Note that in this theorem the heteroscedastic multipliers $\mathbf{\Delta}$ are taken as known, but under further regularity conditions they can be estimated from the data (Zhou and Portnoy, 1998). The theorem immediately implies the following result for the case of independent and identically distributed errors, where $\mathbf{\Delta} = \mathbf{I}$.

Corollary 1. *If the errors* $\boldsymbol{\epsilon}$ *are independent and identically distributed, as* $n \to \infty$, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}\right)$ *is asymptotically p-variate normal with mean* $\mathbf{0}$ *and covariance matrix*

$$\mathbf{Q}^{-1}\left(\mathbf{X}'\mathbf{W}^2\mathbf{X}\right)\mathbf{Q}^{-1}\omega^2,$$

where $\mathbf{Q} = \lim_{n\to\infty} \mathbf{X}'\mathbf{W}\mathbf{X}/n$ *and* $\omega = [2f(0)]^{-1}$.

The implication of this theorem is that confidence regions for $\boldsymbol{\beta}$ can be constructed based on $\hat{\boldsymbol{\beta}}_w$ using the estimated asymptotic covariance,

$$V\left(\hat{\boldsymbol{\beta}}_w\right) = \left(\mathbf{X}'\mathbf{W}\mathbf{\Delta}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{W}^2\mathbf{X}\right)\left(\mathbf{X}'\mathbf{W}\mathbf{\Delta}^{-1}\mathbf{X}\right)^{-1}\hat{\omega}^2,$$

where ω is estimated in some reasonable way. In the simulations that follow we use a kernel estimator to estimate $f(0)$, and hence ω.

We explore the finite-sample properties of WLAD regression using Monte Carlo simulations, performed using the R package (R Development Core Team, 2004). In addition to the WLAD estimator, we also report results for least squares, (unweighted) LAD, and MM estimators. The MM estimator (Yohai, 1987) uses an inefficient high-breakdown method as an initial estimate, but then uses M-estimation to improve efficiency while still maintaining a high breakdown point. We also included an M-estimator (Huber, 1973) and the least trimmed squares (LTS) high breakdown estimator (Rousseeuw, 1985) in the simulations, but the MM estimator consistently outperformed both, so we do not report those results here (we did not investigate the approach of Rousseeuw and van Zomeren, 1992, since they reported that its performance was inferior to that of the least median of squares estimator of Rousseeuw, 1984, and that estimator is known to be inferior to LTS). The (W)LAD estimators were constructed using the `quantreg` package (Koenker, 2004), while the LTS and MM estimators were constructed using the `MASS` package (Venables and Ripley, 2002). We examine various values of sample size n and number of predictors k (note that $p = k + 1$, as the models include an intercept term), and different outlier/leverage point proportion and position combinations. Predictors were generated multivariate normal (with each variable having mean 7.5 and standard deviation 4), with certain observations being modified to be leverage points in some situations, and 500 simulations replications were generated for each setting. All regression functions had intercept equal to 0 and all slopes equal to 5, with (constant) variance of the Gaussian errors equal to 1.

Fig. 1 summarizes the results of simulations where $n = 40$ and $k = 2$. Each bar's height represents $n \times MSE$, where MSE is the mean squared error of the slope estimate, separated by predictor, with the shaded portion corresponding to squared bias and the unshaded portion corresponding to variance. Although both predictors were generated to have variance equal to 16, by random chance the second predictor had much lower variability (sample variance 10.1), resulting in higher values of $n \times MSE$ for that predictor's slope estimate for all

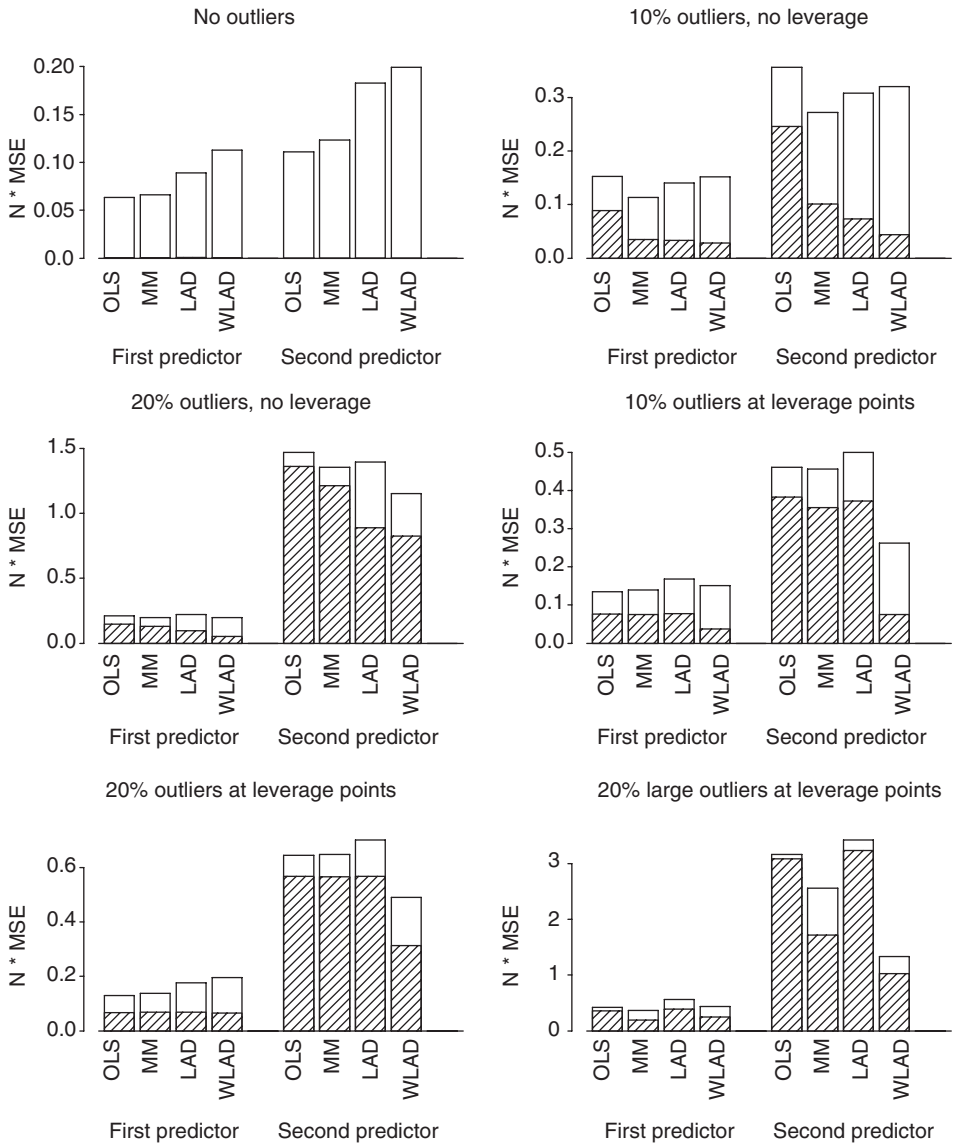

Fig. 1. Results of simulations with $n = 40$ observations and two predictors. Bars represent $n \times$ mean squared error for each estimator, separated into squared bias (shaded) and variance (unshaded) parts.

methods. As can be seen in the figure, the difference in variability of the predictors affects the relative performance of the methods.

When there are no outliers, all of the estimators are (virtually) unbiased, and relative efficiency drives the results. As expected, OLS is most efficient, with the MM estimator close. The LAD estimators are less efficient, with WLAD having highest variability (as

would be expected, since observations with predictor values farthest from the center are downweighted, and it is these observations that increase efficiency). Relative performance changes markedly in the presence of outliers however, with the nonrobust OLS estimator no longer effective. The first five plots refer to outliers with mean 3 standard deviations from the true expected value, while the last ("large outliers") refers to outliers with mean 7 standard deviations from the true value. The last three plots refer to situations where the observations with outliers first had their predictor values adjusted to make them leverage points (the predictor values were centered roughly 4 standard deviations from the predictor mean). Given that the weighting scheme of the WLAD estimator is designed to downweight leverage points, it is not surprising to see much better performance for WLAD in this situation compared to LAD (in fact, the breakdown point of LAD is 17.5% while that of WLAD is 22.5%, so deteriorating performance for LAD in the 20% outliers case would be expected). That breakdown is not the entire story, however, is clear from the much better performance of WLAD compared to MM, especially when the outliers are at leverage points. This is being driven by much lower squared bias, although the variance of WLAD is also lower when there are 20% large outliers at the leverage points.

Fig. 2 gives corresponding results where $n = 100$ and $k = 6$. Although it is not feasible to calculate exact breakdown values for the (W)LAD estimators, upper bounds on those values can be determined (by running the mixed integer program for many iterations, but not to convergence), and they are 16% (for LAD) and 20% (for WLAD), respectively, in this case for the designs with 20% leverage points. Despite the fact that the breakdown point for WLAD is not greater than the observed percentage of generated outlier observations (implying that breakdown can occur), it is still an effective estimator, once again outperforming the MM estimator in the presence of leverage points. Once again, the estimator exhibits good bias properties, although that is outweighed by high variance when the data do not contain leverage points. Fig. 3 gives results for $n = 1000$ and $k = 20$. Even in this case, where the large number of predictors implies a maximum breakdown point of WLAD less than 2% for the design with 20% leverage points, the performance of WLAD is similar to that seen earlier, with the estimator outperforming the MM estimator for large outliers in the presence of leverage points. Thus, it is apparent that the goal of increasing the breakdown point of the estimator is a reasonable one, even when there are more outliers than the breakdown value. This will also be evident in several of the data examples discusses in the next section.

We also examined the usefulness of the asymptotic distribution derived in Theorem 1 as a tool for inference, by examining the properties of hypothesis tests based on the approximate normal distribution. Construction of such tests requires estimating $\omega = [2f(0)]^{-1}$, which is done here using a kernel density estimate (see, e.g., Simonoff, 1996, Section 3.1), with the amount of smoothing chosen using the bandwidth selector of Sheather and Jones (1991). The adequacy of the approximation was evaluated by determining the average rejection proportions (empirical sizes) of Wald (Gaussian-based) tests for each coefficient of the actual null value at a nominal .05 level and then averaging over all slopes, so the goal would be tests with size close to .05.

Results for the simulation situations previously examined are summarized in Table 1. We give results for the actual tests, and also for tests where the coefficient estimates are recentered at the null value, so that the bias of the estimators does not affect performance.

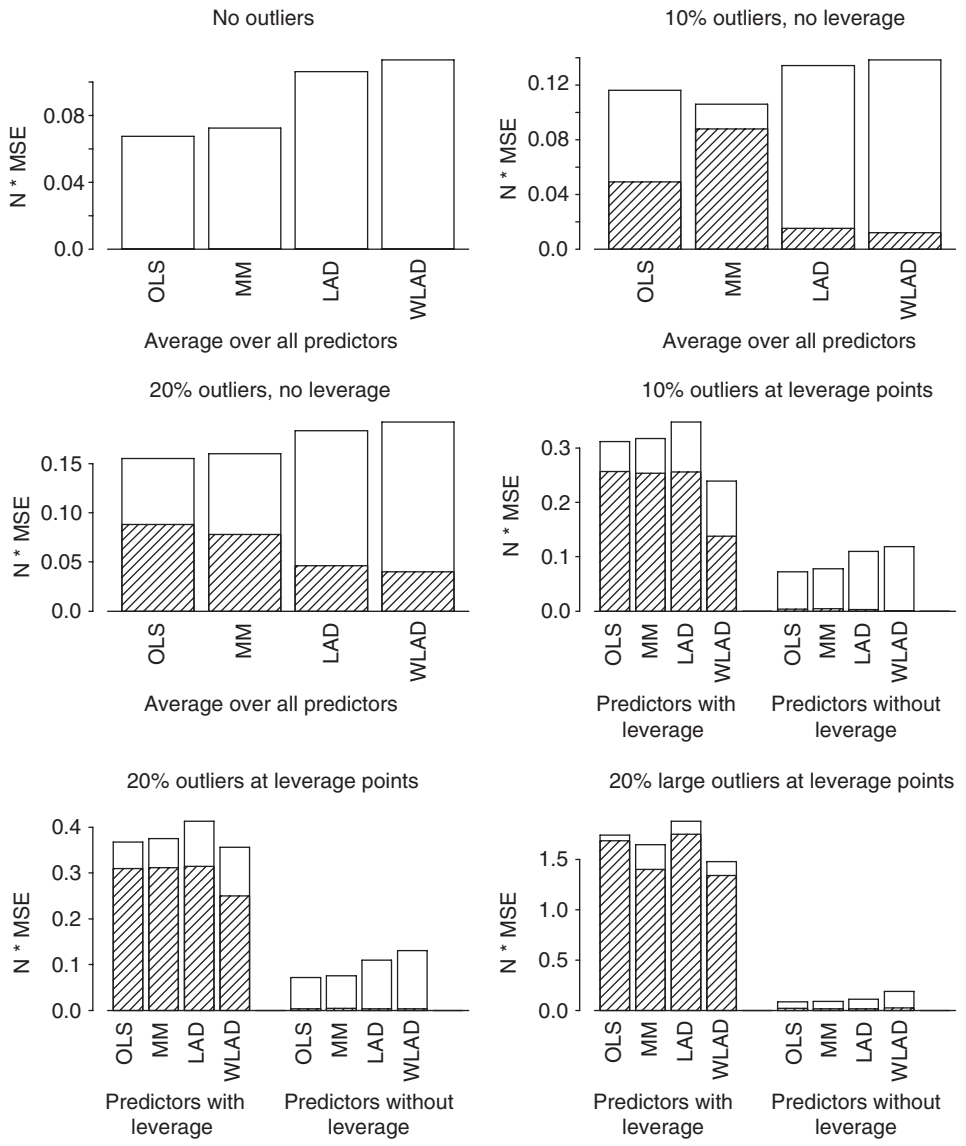

Fig. 2. Results of simulations with $n = 100$ observations and six predictors. Bars represent $n \times$ mean squared error for each estimator, separated into squared bias (shaded) and variance (unshaded) parts.

For the $n = 1000$ case with leverage points, separate average empirical sizes are given for the predictors with and without leverage points for the uncorrected tests, since their performances are very different. It can be seen that while the bias seen earlier (especially in the leverage point case) results in very anticonservative tests (which would presumably also be true for the other estimators, since they are even more biased), when this bias is corrected,

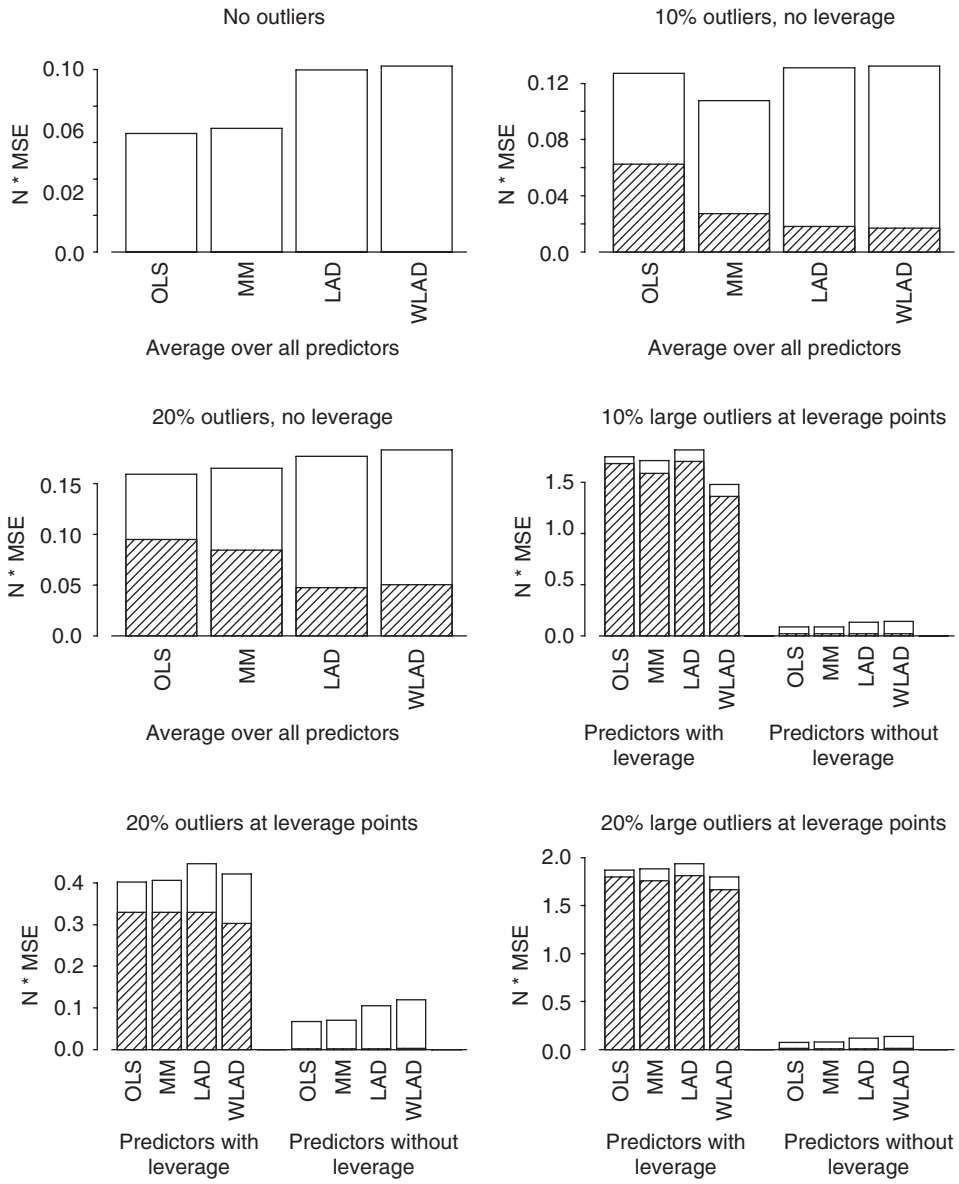

Fig. 3. Results of simulations with $n = 1000$ observations and 20 predictors. Bars represent $n\times$ mean squared error for each estimator, separated into squared bias (shaded) and variance (unshaded) parts.

the tests have size reasonably close to .05. Thus, the evidence suggests that the assumption of a Gaussian distribution for the coefficient estimator, and the implied estimates of standard errors of the coefficient estimates, are reasonable even for reasonably small samples.

Table 1
Average empirical size of tests for slope coefficients (nominal size $\alpha = .05$)

n	k	Outlier structure	Uncorrected tests	Bias-corrected tests
40	2	None	.067	.068
		10%, no leverage	.092	.059
		20%, no leverage	.206	.050
		10%, leverage	.132	.079
		20%, leverage	.218	.073
		20% large outliers, leverage	.325	.071
100	6	None	.073	.073
		10%, no leverage	.065	.058
		20%, no leverage	.061	.036
		10%, leverage	.123	.074
		20%, leverage	.139	.069
		20% large outliers, leverage	.334	.048
1000	20	None	.057	.057
		10%, no leverage	.060	.044
		20%, no leverage	.063	.031
		10% large outliers, leverage	.904, .053	.040
		20%, leverage	.366, .060	.055
		20% large outliers, leverage	.927, .056	.045

4. Application to real data sets

In this section we apply the WLAD estimator to several well-known data sets from the robustness and outlier identification literature. The Hertzsprung–Russell stars data (Rousseeuw and Leroy, 1987, p. 27) consist of measurements of the logarithm of light intensity versus logarithm of effective surface temperature for 47 stars. The data are given in the leftmost plot in Fig. 4. Although there is a direct relationship between the two variables for most of the observations, four stars have low temperature with high light intensity (these are so-called "red giant" stars). The OLS fit (dotted line) is drawn to these outliers, as is the LAD fit (dashed line), but the WLAD fit (solid line) follows the general pattern of the points well (note that the LAD estimate actually passes through one of the outliers). It might be thought that the improved performance of WLAD over LAD is because of its higher breakdown point, but this is not, in fact, the case. The breakdown point of LAD here is 10.6% (5/47), so the four observed outliers are not enough to break down the estimator. This can be seen in the middle plot of Fig. 4, where the four outliers have had their responses adjusted upwards by 10. The OLS line continues to follow the points, but the LAD line is virtually unchanged (and no longer passes through any of the outliers), because it has not broken down (not surprisingly, the WLAD line is unchanged). Thus, WLAD provides a better fit than LAD even when LAD has not broken down. On the other hand, the improved breakdown point of WLAD (which is $14/47 = 29.8\%$) becomes apparent if two more values are perturbed upwards (the rightmost plot of Fig. 4). With six outliers LAD has broken

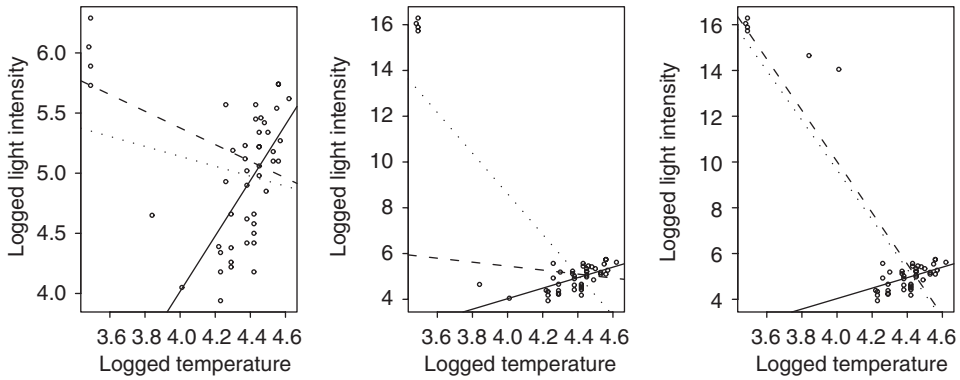

Fig. 4. OLS (dotted line). LAD (dashed line) and WLAD (solid line) regression fits for Hertzsprung–Russell stars data. Left plot refers to original data, middle plot refers to data with original outliers adjusted upwards, and right plot refers to data with two additional observations adjusted upwards.

down, and tracks the unusual points in a similar way to OLS (once again passing through one of them), while WLAD still follows the bulk of the points.

Hawkins et al. (1984) constructed an artificial three-predictor data set with 75 observations, where outliers were placed at cases 1–10. The LAD estimator has breakdown point 10.7% (8/75) and breaks down, with the fitted regression hyperplane going through one of the outlier points (it goes through cases 5, 20, and 32). In contrast the WLAD estimator, with breakdown point 20% (15/75) works well, going through cases 18, 25, and 30, with each of the outlier observations having absolute residual at least 2.3 times that of any of the clean points. The fitted WLAD regression is

$$Y = -0.446 + 0.159X_1 + 0.090X_2 - 0.032X_3,$$

with z-statistics for the three slopes being 1.15, 0.78, and -0.37, respectively. That is, there is little evidence for any relationship here, which is consistent with the way the data were constructed, as none of the t-tests for the three predictors in an OLS fit on observations 11–75 are statistically significant (in contrast to the fit on all of the observations, where the outliers result in variables X_2 and X_3 being significant predictors).

The final data set examined here is the modified wood gravity of Rousseeuw (1984). These data are based on a real data set (with $n = 20$ and $k = 5$), but were modified to have outliers at cases 4, 6, 8, and 19. Both LAD and WLAD have the same breakdown point ($15\% = 3/20$), but despite this, while LAD performs poorly (passing through the outlier case 8), WLAD performs well, with each of the four outlier cases having absolute residual more than 7.7 times that of any of the clean points. Thus, the WLAD weighting is beneficial even when breakdown is not improved. The fitted WLAD regression is

$$Y = 0.387 + 0.321X_1 - 0.422X_2 - 0.541X_3 - 0.336X_4 + 0.523X_5,$$

with z-statistics for the five slopes being 8.50, -2.64, -15.18, -6.32, and 7.79, respectively. An OLS fit on the clean data also identifies predictors 1, 3, 4, and 5 as being most important, although in that case the coefficient for variable 2 is not statistically significant (in contrast

to an OLS fit on the entire data set, where variables 4 and 5 are insignificant, because of the effect of the outliers).

5. Conclusion

In this paper we have proposed a weighted version of LAD regression designed to increase the breakdown of the estimator that is easy to compute and has performance competitive with high breakdown estimators, particularly in the presence of leverage points. These weighting ideas also apply to other estimators. Given the good bias properties of the WLAD estimator, it is reasonable to wonder if a similar weighting scheme used for a more efficient GM-estimator, such as that of Krasker and Welsch (1982), would reduce the variance while preserving robustness, and be even more effective than WLAD. The LAD estimator is a special case of regression quantile estimators (Koenker and Bassett, 1978; Koenker, 2000), which have been shown to be useful in highlighting interesting structure in regression problems, including nonnormality and heteroscedasticity in the error distribution (in this context, LAD would be considered the median regression estimator). Antoch and Jurečková (1985) and de Jongh et al. (1988) described GM-estimators for quantile regression (each constructed differently from the WLAD estimator discussed here, although the former was based on the hat matrix), and it would be interesting to see if weights such as those described here would result in estimators that are more resistant to the effects of unusual observations. There is some evidence that, at least for very small samples without leverage points, Wald hypothesis tests can be beaten by likelihood ratio, Lagrange multiplier, or bootstrap tests under some circumstances (Dielman and Pfaffenberger, 1992; Dielman and Rose, 1995, 1996), and alternatives to density estimation-based methods for estimating ω in the testing context have been suggested (Sheather and McKean, 1987); investigation of application of these approaches to WLAD testing would be worthwhile.

Acknowledgements

The authors would like to thank Steve Portnoy for helpful discussion of this material, and an anonymous referee for helpful comments on an earlier draft of the paper.

Appendix

Consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Phi}_n \mathbf{u},$$

where $\boldsymbol{\Phi}_n = \operatorname{diag}(\phi_1, \ldots, \phi_n)$, $\boldsymbol{\beta} \in \Re^p$, and $\mathbf{u}$ is a vector of independent and identically distributed errors with cumulative distribution function F. Assume that the ϕ_is are known values satisfying $\max \phi_i = \mathrm{O}(1)$ and $\max \phi_i^{-1} = \mathrm{O}(1)$, and that $\max \|\mathbf{x}^i\| = o\left(n^{1/4}\right)$.

Assume that F is twice differentiable at 0, and $f(0)=F'(0)>0$. Let $\mathbf{Q}_n=\mathbf{X}'\mathbf{\Phi}_n^{-1}\mathbf{X}/n=\mathbf{Q}+\mathrm{O}\left(n^{-1/4}\log n\right)$. Let $\hat{\boldsymbol{\beta}}$ be the least absolute deviation (LAD) estimator of $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}\left|y_i-\mathbf{x}^i\boldsymbol{\beta}\right|.$$

Lemma 1. *As* $n\to\infty$,

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\frac{n^{-1}\mathbf{Q}_n^{-1}}{f(0)}\sum_{i=1}^{n}\mathbf{x}^{i\prime}\Psi(u_i)+O\left((\log n/n)^{3/4}\right),$$

where $\Psi(x)=I(x<0)-.5$.

Proof. This result follows from an adaptation of the proof of Zhou and Portnoy (1998, Theorem 2.1) (hereafter ZP). In that theorem the multipliers ϕ_i are estimated based on a linear function of a parameter vector $\boldsymbol{\gamma}$, so the results quoted here are based on taking $\hat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}$ in that proof. Let $\mathbf{W}_n(\mathbf{t})=\sum_{i=1}^{n}\mathbf{x}^{i\prime}\Psi(u_i)$. By Lemma A.1 of ZP,

$$\mathbf{W}_n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\mathrm{O}\left(n^{-3/4}\right). \tag{5}$$

Let $M_n=c_0n^{-1/2}(\log n)^{1/2}$. Then, by Lemma A.2 of ZP,

$$\sup_{\|\mathbf{t}\|\leqslant M_n}\|\mathbf{W}_n(\mathbf{t})-\mathbf{W}_n(\mathbf{0})+f(0)\mathbf{Q}\mathbf{t}\|=\mathrm{O}_p\left(n^{-3/4}(\log n)^{3/4}\right). \tag{6}$$

Substituting $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ for $\mathbf{t}$ in (6) and then substituting (5) into (6) gives

$$\mathbf{W}_n(0)=f(0)\mathbf{Q}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\mathrm{O}_p\left(n^{-3/4}(\log n)^{3/4}\right).$$

The result of the lemma then follows. □

The next lemma uses this result to establish the asymptotic distribution of the LAD estimator under known heteroscedasticity.

Lemma 2. *As* $n\to\infty$, $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ *is asymptotically p-variate normal with mean* $\mathbf{0}$ *and covariance matrix*

$$\mathbf{Q}^{-1}(\mathbf{X}'\mathbf{X})\mathbf{Q}^{-1}\omega^2,$$

where $\mathbf{Q}$ *is defined as in Lemma* 1 *and* $\omega=[2f(0)]^{-1}$.

Proof. This is a direct consequence of Lemma 1. The Central Limit Theorem implies asymptotic normality, and since

$$V\left(\sum_{i=1}^{n}\mathbf{x}^{i\prime}\Psi(u_i)\right)=\left(\mathbf{X}'\mathbf{X}\right)/4$$

the result follows. □

These results then allow us to derive the asymptotic distribution of the WLAD estimator, as follows.

Proof of Theorem 1. This follows from Lemma 2. Let $\tilde{\mathbf{X}} = \mathbf{WX}$ and $\tilde{\mathbf{y}} = \mathbf{Wy}$. Then the weighted LAD estimation problem is equivalent to unweighted estimation for the model

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\epsilon} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\Delta}\boldsymbol{u}.$$

Lemma 2 implies that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}\right)$ is asymptotically p-variate normal with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{Q}^{-1}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)\mathbf{Q}^{-1}\omega^2,$$

where $\mathbf{Q} = \lim_{n\to\infty} \tilde{\mathbf{X}}'(\mathbf{W}\boldsymbol{\Delta})^{-1}\tilde{\mathbf{X}}/n$. Substituting $\mathbf{WX}$ for $\tilde{\mathbf{X}}$ gives the result. □

References

Antoch, J., Jurečková, J., 1985. Trimmed least squares estimator resistant to leverage points. Comput. Statist. Quart. 4, 329–339.

Atkinson, A., Riani, M., 2000. Robust Diagnostic Regression Analysis. Springer, New York.

Billor, N., Hadi, A.S., Velleman, P.F., 2000. BACON: blocked adaptive computationally efficient outlier nominators. Comput. Statist. Data Anal. 34, 279–298.

Chatterjee, S., Hadi, A.S., 1988. Sensitivity Analysis in Linear Regression. Wiley, New York.

de Jongh, P.J., de Wet, T., Welsh, A.H., 1988. Mallows-type bounded-influence-regression trimmed means. J. Amer. Statist. Assoc. 83, 805–810.

Dielman, T.E., 2005. Least absolute value regression: recent contributions. J. Statist. Comput. Simulation 75, 263–286.

Dielman, T.E., Pfaffenberger, R.C., 1992. A further comparison of tests of hypotheses in LAV regression. Comput. Statist. Data Anal. 14, 375–384.

Dielman, T.E., Rose, E.L., 1995. A bootstrap approach to hypothesis testing in least absolute value regression. Comput. Statist. Data Anal. 20, 119–130.

Dielman, T.E., Rose, E.L., 1996. A note on hypothesis testing in LAV multiple regression: a small sample comparison. Comput. Statist. Data Anal. 21, 463–470.

Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: Bickel, P., Doksum, K., Hodges, J.L. (Eds.), A Festschrift for Erich Lehmann. Wadsworth, Belmont, CA, pp. 157–184.

Ellis, S.P., Morgenthaler, S., 1992. Leverage and breakdown in L_1-regression. J. Amer. Statist. Assoc. 87, 143–148.

Giloni, A., Padberg, M., 2004. The finite sample breakdown point of ℓ_1-regression. SIAM J. Optim. 14, 1028–1042.

Giloni, A., Sengupta, B., Simonoff, J.S., 2004. A mathematical programming approach for improving the robustness of LAD regression, unpublished manuscript.

Hadi, A.S., 1992. Identifying multiple outliers in multivariate data. J. Roy. Statist. Soc. Ser. B 54, 761–771.

Hadi, A.S., 1994. A modification of a method for the detection of outliers in multivariate samples. J. Roy. Statist. Soc. Ser. B 56, 393–396.

Hadi, A.S., Simonoff, J.S., 1993. Procedures for the identification of multiple outliers in linear models. J. Amer. Statist. Assoc. 88, 1264–1272.

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

Hawkins, D.M., Olive, D.J., 2002. Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. J. Amer. Statist. Assoc. 97, 136–159.

Hawkins, D.M., Bradu, D., Kass, G.V., 1984. Location of several outliers in multiple regression using elemental subsets. Technometrics 26, 197–208.

He, X., Jurečková, J., Koenker, R., Portnoy, S., 1990. Tail behavior of regression estimators and their breakdown points. Econometrica 58, 1195–1214.

Huber, P.J., 1973. Robust regression: asymptotics, conjectures, and Monte Carlo. Ann. Statist. 1, 799–821.

Hubert, M., Rousseeuw, P.J., 1997. Robust regression with both continuous and binary regressors. J. Statist. Plann. Inference 57, 153–163.

Koenker, R., 2000. Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. J. Econometrics 95, 347–374.

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

Koenker, R., Bassett Jr., G., 1978. Regression quantiles. Econometrica 46, 33–50.

Krasker, W.S., Welsch, R.E., 1982. Efficient bounded-influence regression estimation. J. Amer. Statist. Assoc. 77, 595–604.

Mallows, C.L., 1975. On some topics in robustness. Technical Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Maronna, R.A., Bustos, O.H., Yohai, V.J., 1979. Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In: Gasser, T., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics, vol. 757, Springer, Berlin, pp. 91–116.

Mizera, I., Müller, C.H., 2001. The influence of the design on the breakdown points of ℓ_1-type M-estimators. In: Atkinson, A., Hackl, P., Müller, W. (Eds.), MODA6—Advances in Model-Oriented Design and Analysis. Physica-Verlag, Heidelberg, pp. 193–200.

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

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

Rosner, B., 1975. On the detection of many outliers. Technometrics 17, 221–227.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79, 871–880.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications, vol. B. Dordrecht, Reidel, pp. 283–297.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Rousseeuw, P.J., van Zomeren, B., 1992. A comparison of some quick algorithms for robust regression. Comput. Statist. Data Anal. 14, 107–116.

Sharpe, W.F., 1971. Mean-absolute-deviation characteristic lines for securities and portfolios. Management Sci. 18, B1–B13.

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. Ser. B 53, 683–690.

Sheather, S.J., McKean, J.W., 1987. A comparison of testing and confidence interval methods for the median. Statist. Probab. Lett. 6, 31–36.

Simonoff, J.S., 1991. General approaches to stepwise identification of unusual values in data analysis. In: Stahel, W., Weisberg, S. (Eds.), Directions in Robust Statistics and Diagnostics: Part II. Springer, New York, pp. 223–242.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with *S*. fourth ed.. Springer, New York.

Yohai, V.J., 1987. High breakdown-point and high efficiency robust estimates for regression. Ann. Statist. 15, 642–656.

Zhou, K.Q., Portnoy, S.L., 1998. Statistical inference on heteroscedastic models based on regression quantiles. J. Nonparametric Statist. 10, 239–260.