Statistical Science
1998, Vol. 13, No. 4, 337–350

Instability of Least Squares, Least Absolute Deviation and Least Median of Squares Linear Regression

Steven P. Ellis

Abstract. Say that a regression method is "unstable" at a data set if a small change in the data can cause a relatively large change in the fitted plane. A well-known example of this is the instability of least squares regression (LS) near (multi)collinear data sets. It is known that least absolute deviation (LAD) and least median of squares (LMS) linear regression can exhibit instability at data sets that are far from collinear. Clear-cut instability occurs at a "singularity"—a data set, *arbitrarily* small changes to which can substantially change the fit. For example, the collinear data sets are the singularities of LS. One way to measure the extent of instability of a regression method is to measure the size of its "singular set" (set of singularities). The dimension of the singular set is a tractable measure of its size that can be estimated without distributional assumptions or asymptotics.

By applying a general theorem on the dimension of singular sets, we find that the singular sets of LAD and LMS are at least as large as that of LS and often much larger. Thus, *prima facie*, LAD and LMS are frequently unstable. This casts doubt on the trustworthiness of LAD and LMS as exploratory regression tools.

Key words and phrases: Collinearity, stability, singularity.

1. INTRODUCTION

Small changes to a nearly collinear data set can cause wild swings in the fitted least squares regression (LS) plane. (Recall that a data set is (multi)collinear if all the k-variate predictor vectors lie in a linear manifold (affine space) of dimension less than k. This paper considers linear regression as a multivariate technique, that is, with the predictors as well as the responses free to vary. An intercept is also included.) This is illustrated by Figure 1. The figure shows two nearly collinear synthetic data sets that are very close to each other. (All data sets discussed in this paper are listed in Ellis, 1997.) The solid lines are the fitted LS lines, which are drastically different even though the data sets are nearly identical. (We will discuss the dashed and dotted lines later.)

In this paper, we say that a statistic is "unstable" (or exhibits "instability") at a data set if a small movement of the data can produce a relatively large change in the statistic. So, "instability" is a form of sensitivity, but sensitivity to outliers is not our concern here. Thus, LS fitted planes (and, a fortiori, estimates) are unstable at nearly collinear data. We discuss presently regression methods that can be unstable at data that are far from collinear.

Since we prefer that the regression describe the overall pattern, not the last few significant digits, in the data, unstable behavior by a regression method is undesirable. That, in practice, measurement and computation are not infinitely precise makes instability more of a concern. Instability can greatly amplify even small measurement and rounding errors.

Two linear regression methods often recommended as robust or outlier-resistant alternatives to LS are least absolute deviation (LAD) or L_1-

Steven P. Ellis is with Department of Neuroscience, New York State Psychiatric Institute, Unit 42, 1051 Riverside Drive, New York, New York 10032 (e-mail: ellis@neuron.cpmc.columbia.edu) and Division of Biostatistics, Columbia University School of Public Health. The author began work on this paper while at the Department of Neurosciences at the University of Medicine and Dentistry of New Jersey.

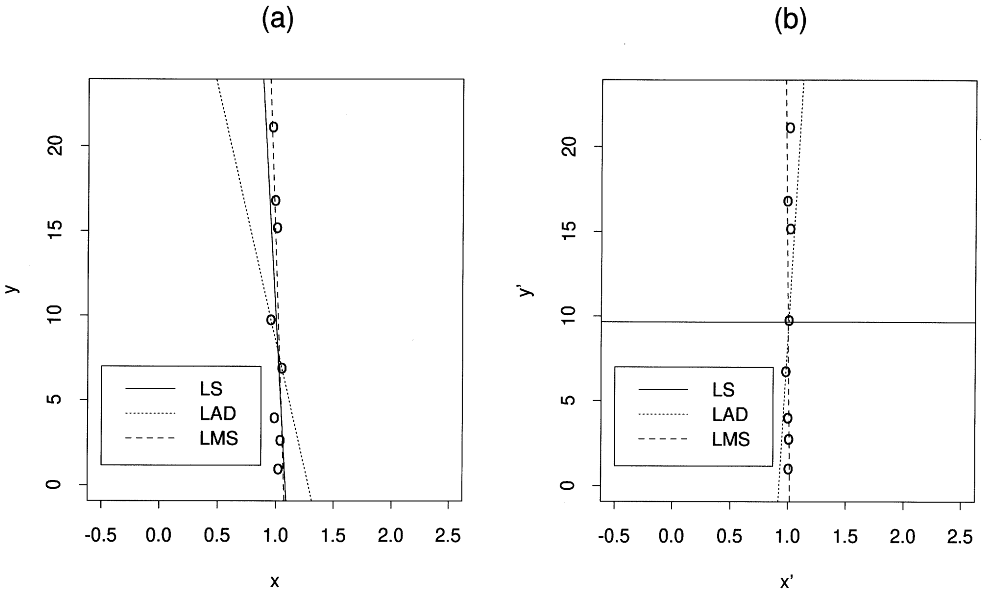

FIG. 1. *Two only slightly different, nearly collinear data sets*: (*solid lines*) *least squares lines*; (*dotted lines*) *least absolute deviation lines*; (*dashed lines*) *least median of squares lines.*

regression (Gentle, 1977; Bloomfield and Steiger, 1983; Dodge, 1987, 1992; and Birkes and Dodge, 1993) and least median of squares regression (LMS) (Hampel, 1975; Rousseeuw, 1984; and Rousseeuw and Leroy, 1987). Recall that a coefficient vector is an LAD estimate if it minimizes the sum of the absolute values of the residuals. A coefficient vector is an LMS estimate if it minimizes the median of the squared residuals.

LMS and LAD may reduce the danger from outliers, but they can be unstable at data sets that are far from collinear (see Ellis, 1991, Example 2.3; Hettmansperger and Sheather, 1992; and Rousseeuw, 1994). Davies (1993) and Ellis (1995a) offer some theoretical insight into the instability of LMS.

Figure 2 shows two real data sets at which LAD is unstable. (See Hettmansperger and Sheather,

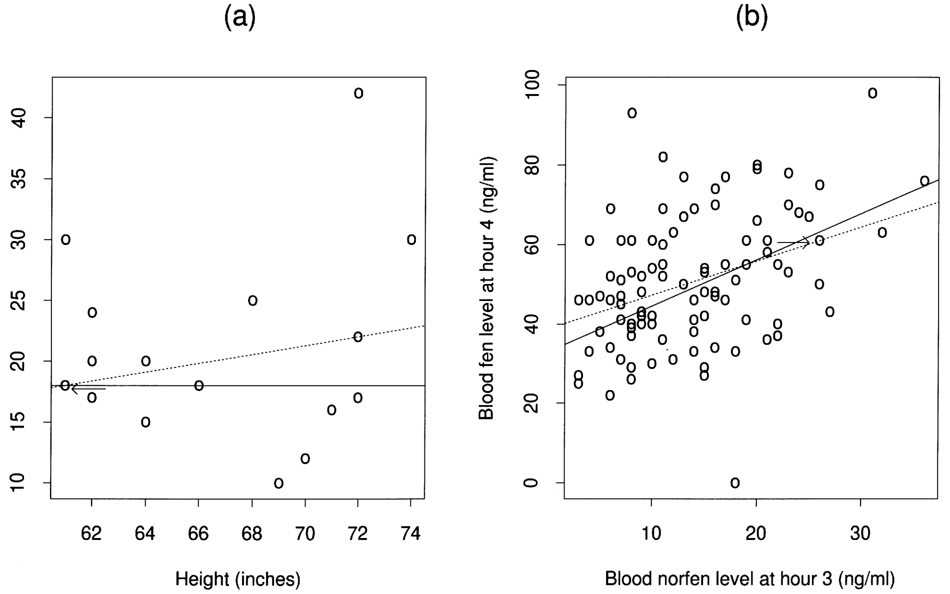

FIG. 2. (*Solid lines*) *LAD lines for data plotted*; (*dashed lines*) *LAD lines for data sets obtained by moving observations next to tips of arrows a tiny amount in the directions shown by the arrows.*

1992, for an LMS example.) These data come from the Mental Health Clinical Research Center for the Study of Suicidal Behavior at the New York State Psychiatric Institute. Figure 2a shows *age at first psychotherapy* versus *height* for a group of 16 research subjects. (This is not as odd a pairing of variables as it may at first appear. See Pine, Cohen and Brook, 1996.) Figure 2b shows *blood fenfluramine level at hour 4* versus *blood norfenfluramine level at hour 3* for a group of 99 research subjects in a fenfluramine challenge experiment. (See Mann et al., 1995, and Malone, Corbitt, Li and Mann, 1996.) In each plot, the solid line is the LAD fit to the data shown. The dashed line is the LAD line for a slightly different data set. (The LAD lines were computed using the S-PLUS function `l1fit`.) In each case, the perturbed data set differs from the original by a change in the x-value (predictor) in only one observation by a mere 1/20,000th of the interquartile range of the xs. The arrows in the figure indicate the observations that are perturbed and the direction of their perturbation. Note that these data sets are clearly far from collinear.

In the examples of Figure 2 and in the example in Figure 3a below, the perturbation consists of moving just one observation. In fact, I was able to recognize the instability of LAD at the data sets in Figure 2 by perturbing each observation in turn by a small amount and then fitting an LAD line. However, the notion of instability allows small perturbations of any number of the observations, even all of them. This makes it challenging, in general, to spot unstable behavior.

In Figure 2 and other examples in this paper, my purpose is to illustrate instability by showing what great effect truly tiny modifications in data can have. Of course, real data is of limited precision so very tiny perturbations may be unrealistic. However, one would expect that a large perturbation in the data would cause at least as large a displacement in the fitted regression as a tiny perturbation would. For example, in the data in Figure 2a, *height* and *age* are measured to the nearest inch and year, respectively. If, instead of a perturbation of less than 1/2000th of an inch as in Figure 2a, one changes the same x-value by 1 inch in the same direction, the LAD line also jumps.

More important, instability might make a regression method depend strongly on the precision of the data. For example, with measurements only accurate to the nearest inch or year, the discrepancy in slopes shown in Figure 2a cannot be seen, because round-off would yield the same vectors of predictors.

I stress that, at least for LS and LAD, variation of the predictors is essential for instability to occur. Consider the general linear model, $y = X\beta + \text{error}$, where y is a column vector of responses, X is a *fixed* design matrix *of full rank*, and β is the column vector of regression coefficients. First, consider the case in which X is a column vector of 1's. In that case, an estimate of β is a location estimate for the univariate data set y. In particular, the LAD estimates of β are precisely the medians of y. The LAD estimate

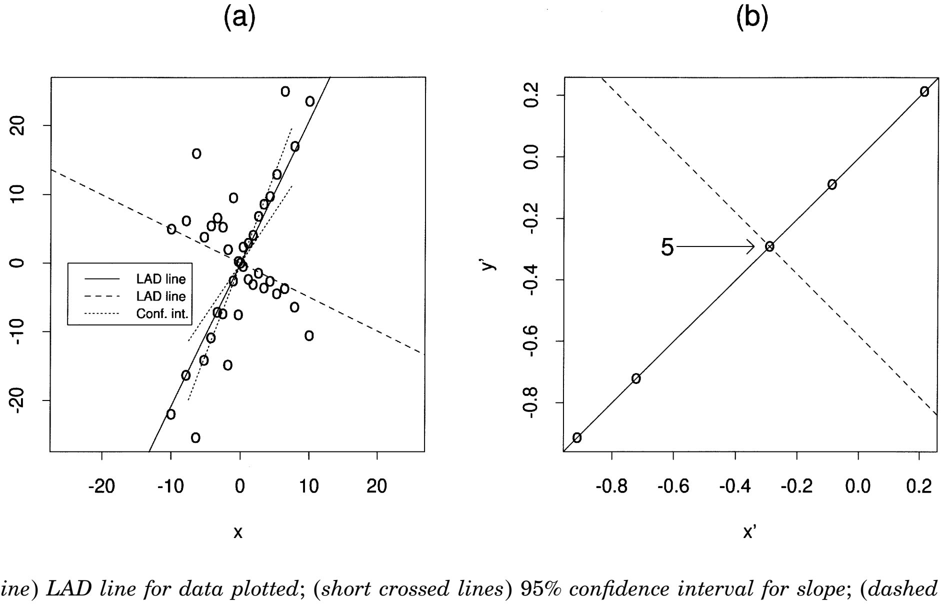

FIG. 3. (a) (*Solid line*) *LAD line for data plotted*; (*short crossed lines*) 95% *confidence interval for slope*; (*dashed line*) *LAD line for a slightly perturbed data set*; (b) (*solid line*) *LMS line for data shown* (*arrow indicates five coincident points*); (*dashed line*) *LMS line for a slightly perturbed data set.*

of β may not be unique, but suppose one adopts the usual convention and defines *the median* $\hat{\beta}(y)$ of y to be the midpoint of the set of medians, that is, the midpoint of the set of LAD estimates. Then $\hat{\beta}(y)$ is a Lipschitz function of y and, hence, exhibits no instability. (Recall that a function is "Lipschitz" if the distance between the values of the function at two points is bounded above by a constant multiple of the distance between the two points.) For general X, providing we generalize "midpoint of the set of LAD estimates" appropriately, the LAD estimate of β is still Lipschitz in y and, therefore, not unstable (Ellis, 1995b). The same is obviously true of the LS estimate of β. (I do not know if this is true of LMS.)

However, if there is an intercept in the model and the other predictors are allowed to vary freely, then the situation is completely different. Now LS, LAD and LMS will all exhibit instability. Moreover, as we will see in Section 3, there is no way to remedy the problem by careful choice of an estimate when it is not unique. (Probably the presence of an intercept is not essential here.)

To recapitulate, it is known that LS, LAD and LMS all exhibit instability. This paper looks more deeply into the phenomenon by asking, for each of the three methods, how common are data sets at which the method is unstable? Answering this question will suggest if instability is common enough to be of relevance to data analysis and will allow us to compare the three methods in terms of instability.

In this paper, we focus on the clear-cut form of instability found at "singularities." A "singularity" is a data set, *arbitrarily* small changes to which can substantially change the fit. (See Section 2 for the formal definition.) For example, the collinear data sets are the singularities of LS. A regression method will be unstable near its singularities. I expect, in fact, that for regression methods used in practice, the *only* place instability will occur is near a singularity. (For example, in Figure 1, each data set is near a collinear data set, i.e., a singularity of LS. In Figure 2, each data set is very near a singularity of LAD. See Ellis, 1997, Remark 2.4.)

We get information about how often instability of LS, LAD or LMS will occur by measuring, or at least bounding, the sizes of their "singular sets" (sets of singularities). The idea is that the more singularities there are, the more likely one is to get a data set near one of them and, therefore, encounter instability.

One way to measure the size of the singular set of a regression method is by its dimension ("degrees of freedom," in statistical parlance). For example, the dimension of a smooth curve is 1. The dimension of a smooth image of an open subset of a plane is 2, and so on, but one can assign a dimension to any metric space. In this paper, we bound below the dimensions of the singular sets of LAD and LMS and compare the bound to the (known) dimension of the singular set of LS (the collinear data sets). Our approach will require no asymptotics or distributional assumptions. We will find that the dimensions of the singular sets of LAD and LMS are always at least as large as that of LS and often larger. We will see that, surprisingly, the large size of the singular sets of LAD and LMS follows from the fact that, *in terms of fitted planes*, the two methods are quite stable near most collinear data sets! The methods of analysis applied here for LAD and LMS may also prove useful for studying other regression methods.

Thus, while LAD and LMS are more resistant to outliers than is LS, they are apparently more often unstable as well. This calls into question the superiority of LAD and LMS over LS as exploratory regression techniques. Data analysts should be on the alert for instability when using any of the three methods. Good methods for detecting instability in LAD and LMS regressions are needed.

Another issue is the size of the displacement in fit associated with instability. For example, the instability portrayed in Figure 2b is not very troubling because the displacement in fit caused by perturbing the data is small.

On the other hand, Figure 3 shows synthetic examples of instability with very large displacements. In Figure 3a, the regression method is LAD. The short crossed dashed lines indicate a 95% confidence interval for the slope based on the asymptotic distribution of the LAD estimate (Bloomfield and Steiger, 1983, Theorem 1, page 64). (Specifically, the slopes of the short line segments are the endpoints of the interval.) The long dashed line is the LAD line for a data set only a tiny bit different from the one displayed.

This example suggests that confidence intervals based on standard asymptotics may not be good diagnostics for recognizing instability. On the other hand, it turns out that a simple bootstrap interval for the slope (not shown; see Efron and Tibshirani, 1993, Chapter 9) is so wide that it includes the slopes of both fitted LAD lines. (See Ellis, 1997, for details.)

In Figure 3b, the regression method is LMS. The data set plotted consists of nine points lying exactly on a line. The arrow indicates five of the nine points, which coincide. The solid line is the fitted LMS line to the data shown. The dashed line is the LMS line fitted to a slightly perturbed data set.

Now, instability does not have to be as extreme as that portrayed in Figure 3 to be a concern in data analysis. However, it turns out that restricting attention only to singularities with, roughly speaking, 90° displacements does not change the conclusions of this paper (Ellis, 1997, Lemma 1.4).

In Section 2, we formally define singularity and state a result bounding the dimensions of the singular sets of many multivariate procedures that fit planes to data. In Section 3, we apply this result to LS, LAD and LMS.

2. SINGULARITY AND PLANE-FITTING

Our strategy for studying the instability of LS, LAD and LMS will be to focus on the extreme form of instability I call "singularity." It can be defined for any statistical procedure, not just regression.

Let δ be a statistic defined on a sample space, X; δ may not be defined at literally every point of X. For example, if δ is LAD, LMS or LS, then $\delta(x)$ is the solution to an optimization problem (Section 3), and it might not be clear how to define $\delta(x)$ if the problem has no unique solution. However, assume δ is at least defined on some dense subset, X', of X. A point $x_0 \in X$ is a *singularity* of δ with respect to (w.r.t.) X') if $\lim_{x\to x_0,\, x\in X'} \delta(x)$ does not exist. (Strictly speaking, singularity is not the same thing as discontinuity.) A crucial point is that a singularity x_0 is a hazard to data analysis using δ even if one never gets x_0 as a data set. Small perturbations of data sets *near* x_0 can cause δ to change a lot. (For example, as mentioned above, in Figure 2 each data set is very near a singularity of LAD.) The set of all singularities of δ (w.r.t. X') is the *singular set* of δ (w.r.t. X'). Near its singular set, δ will be unstable.

We will treat regression as a form of "plane-fitting," the general class of statistical procedures that, for example, also includes principal components analysis. Consider fitting a k-dimensional plane ("k-plane") to n observations in $\mathbb{R}^p$($\mathbb{R}$ = reals). Write such a data set as an $n \times p$ matrix whose ith row is the ith observation. Let $\mathscr{Y}$ denote the set of all such matrices. We will assume $n > p > k > 0$.

Here, "plane" means "affine plane." So a plane does not have to pass through the origin. (Planes are in $\mathbb{R}^p$.) The set of all k-planes that do pass through the origin is the *Grassman manifold* $G(k, p)$. Thus, we can write any k-plane in the form $\xi + v$, where $v \in \mathbb{R}^p$ and $\xi \in G(k, p)$.

The manifold $G(k, p)$ is a compact $k(p-k)$-dimensional manifold (Boothby, 1975, pages 63–64). One metric generating the topology on $G(k, p)$ is the *angle* between two planes, defined as follows. Let $\xi, \zeta \in G(k, p)$. If x is a nonzero vector in $\mathbb{R}^p$, let $\angle(x, \zeta)$ be the angle between x and its orthogonal projection onto ζ. (If the projection of x onto ζ is 0, then $\angle(x, \zeta) = \pi/2$.) Define the angle,

$$\angle(\xi, \zeta) = \sup\{\angle(x, \zeta)\colon x \in \xi,\ x \neq 0\};$$

$\angle(\xi, \zeta)$ turns out to be a metric on $G(k, p)$ which generates its usual topology.

An important class of data matrices is the set $\mathscr{P}_k$ of all $Y \in \mathscr{Y}$ such that (s.t.) the rows of Y lie exactly on a unique k-plane. In other words, $Y \in \mathscr{Y}$, with rows $Y^1, \dots, Y^n$, is an element of $\mathscr{P}_k$ if and only if

$$\text{(2.1)}\quad \text{the matrix} \begin{pmatrix} Y^2 - Y^1 \\ \vdots \\ Y^n - Y^1 \end{pmatrix} \text{has rank exactly } k.$$

Thus, the data sets in $\mathscr{P}_k$ are those to which there is a perfect fit by a unique k-plane. For example, Figure 3b shows a data set in $\mathscr{P}_k$ in the case $k = 1$, $p = 2$, $n = 9$. If $Y \in \mathscr{P}_k$, let $\Xi(Y)$ be the element of $G(k, p)$ parallel to the k-plane on which the rows of Y lie. In the notation of (2.1), $\Xi(Y)$ is the vector space spanned by $Y^2 - Y^1, \dots, Y^n - Y^1$.

EXAMPLE 2.1. In linear regression, there are $p = k+1$ variables per observation (k predictors and one response). Suppose $Y \in \mathscr{Y}$ and write the rows of Y as $Y^i = (x_i, y_i)$, where x_i is a $1 \times k$ row vector of predictors and $y_i \in \mathbb{R}$ is the response ($i = 1, \dots, n$). Suppose these data are not collinear and there exists a p-vector $(\beta_0, \beta_1^T)^T$ ($\beta_0 \in \mathbb{R}$ and β_1 is $k \times 1$; "T" indicates transposition) s.t. $y_i = \beta_0 + x_i\beta_1$, $i = 1, \dots, n$. Then $Y \in \mathscr{P}_k$ and $\Xi(Y) \in G(k, p)$ is the k-dimensional subspace $\{(x, x\beta_1),\ x \in \mathbb{R}^k\}$.

On the other hand, if $x_2 - x_1, \dots, x_n - x_1$ do not span $\mathbb{R}^k$, then Y is collinear; Y can still be in $\mathscr{P}_k$, however, because (2.1) can still hold. In that case, $\Xi(Y)$ is the k-dimensional linear subspace, $V \times \mathbb{R}$, where V is the span of $x_2 - x_1, \dots, x_n - x_1$. (Note that a necessary condition that a collinear data set Y be in $\mathscr{P}_k$ is that V have dimension $k - 1$.)

If T is a technique for assigning k-planes to data sets in (a dense subset of) $\mathscr{Y}$, one can associate with it the map Φ that sends every $Y \in \mathscr{Y}$ s.t. $T(Y)$ is defined, to the k-dimensional subspace $\Phi(Y) \in G(k, p)$ parallel to $T(Y)$. For many of the T's used in data analysis, the map Φ turns out to be a "plane-fitter," defined as follows. Let $\mathscr{Y}'$ be a dense subset of $\mathscr{Y}$ s.t. $\mathscr{P}_k \cap \mathscr{Y}'$ is dense in $\mathscr{P}_k$. A map $\Phi\colon \mathscr{Y}' \to G(k, p)$ is a *plane-fitter* (on $\mathscr{Y}'$) if the following holds:

$$\text{(2.2)}\qquad \text{If } Y \in \mathscr{P}_k \cap \mathscr{Y}',\ \text{then } \Phi(Y) = \Xi(Y);$$

that is, Φ fits the "obvious" plane to Y. (In robustness terminology, the defining property of a plane-fitter is essentially that it have a positive “exact fit point” (Rousseeuw and Leroy, 1987, page 123).)

EXAMPLE 2.1 (Continued). Suppose T is a linear regression technique. If $Y \in \mathscr{Y}$ and $T(Y)$ is defined, then, for some $b_0 \in \mathbb{R}$ and $k \times 1$ vector b_1, the plane fitted to Y by T is the set $\{(x, b_0 + xb_1),\ x \in \mathbb{R}^k\}$. Then $\Phi(Y) = \{(x, xb_1),\ x \in \mathbb{R}^k\}$. I expect that for any regression technique T used in practice, Φ will be a plane-fitter.

We will see presently that even without any further assumptions, a plane-fitter Φ must have many singularities. The topology of $G(k, p)$ plays a crucial role in this phenomenon (Ellis, 1991, 1995a, 1996).

Let R be a regression method with associated plane-fitter Φ. Let $\mathscr{S}$ be the singular set of Φ. Any singularity of Φ is a singularity of R, so near $\mathscr{S}$ the regression method R will exhibit instability. Thus, the size of the collection of data sets in $\mathscr{Y}$ near $\mathscr{S}$ is a measure of the seriousness of instability of R. However, it is not immediately clear how “size” or “near” should be defined. Once that is settled, calculating the size will be difficult.

On the other hand, it is clear that the bigger $\mathscr{S}$ is, the bigger will be the size of the collection of data sets in $\mathscr{Y}$ near $\mathscr{S}$. The dimension of (i.e., “degrees of freedom” in) $\mathscr{S}$ is a tractable, if crude, measure of the size of $\mathscr{S}$ which can be related to the size of the collection of data sets in $\mathscr{Y}$ near $\mathscr{S}$. (See Ellis, 1995a.) Denote the dimension of $\mathscr{S}$ by $\dim \mathscr{S}$. [Technically, we have to specify what kind of dimension we mean by “dim.” In this paper, dim is Hausdorff dimension. It is defined for any metric space and gives the correct values for the dimensions of curves, surfaces, etc. See Falconer (1990), Morgan (1988) or Ellis (1995a), for its definition and properties. Actually, I expect the results presented here remain valid for any reasonable choice of dim.]

The following result gives lower bounds on the dimension of the singular set of a plane-fitter. In this paper, we apply it to LS, LAD and LMS. (We define “singular set” a little differently here than in Ellis, 1995a, but the difference is of no consequence.) The theorem states that if a plane-fitter has few singularities in $\mathscr{P}_k$, then it must have many of them elsewhere. (A plane-fitting technique can indeed have singularities in $\mathscr{P}_k$. For example, the collinear data sets, many of which lie in $\mathscr{P}_k$, are singularities of LS. Figure 3b shows a data set in $\mathscr{P}_k$ ($k = 1$) which is a singularity of LMS.)

THEOREM 2.2 (Ellis, 1995a, Theorem 2.2). *Let* Φ *be a plane-fitter. If* $\dim(\mathscr{S} \cap \mathscr{P}_k) < d \equiv \dim(\mathscr{P}_k) - 1$, *then* $\dim(\mathscr{S}) \geq np - 2$.

Note that, since there are only np degrees of freedom available in $\mathscr{Y}$, if $\dim(\mathscr{S}) \geq np - 2$, $\mathscr{S}$ is very high dimensional. It turns out that

$$d = nk + (k + 1)(p - k) - 1.$$

Since $n > p > k$, we have $np - 2 \geq d$. Thus, an immediate corollary of the theorem is, we always have $\dim(\mathscr{S}) \geq d$.

3. SINGULAR SETS OF LS, LAD AND LMS

Now we restrict attention to the regression case. Each of the n observations is a p-vector consisting of k predictors and a univariate response. So $p = k+1$. Continue to assume $n > p$ and $k > 0$. If $Y \in \mathscr{Y}$, generically write the ith row $Y^i = (x_i, y_i)$, where x_i is $1 \times k$ and $y_i \in \mathbb{R}$. Write regression coefficient vectors $b = (b_0, b_1^T)^T$ ($b_0 \in \mathbb{R}$ is the intercept, b_1 is $k \times 1$). As mentioned above, we will be interested in the case in which the predictors are not chosen in advance, but vary freely. (In particular, functional relationships among the predictors are not allowed. For example, if t is a predictor, t^2 may not be one.)

Let R = LS, LAD or LMS. If $Y \in \mathscr{Y}$, let $\Phi_R(Y)$ be the element of $G(k, p)$ parallel to the plane fitted to Y by R, providing it exists. As we shall see presently, Φ_R is a plane-fitter. Let $\mathscr{Y}'_R$ be the set of all noncollinear data sets in $\mathscr{Y}$ for which the appropriate minimization problem (see below) has a unique solution. (It is easy to see that, because of this uniqueness, the points of $\mathscr{Y}'_R$ cannot be singularities of R.) Let $\mathscr{S}_R$ be the singular set of Φ_R w.r.t. $\mathscr{Y}'_R$. For example, $\mathscr{S}_{\text{LS}}$ is precisely the set of collinear data sets. The dimension of the set of collinear data sets turns out to be d (Ellis, 1995a, Example 2.8); that is, $\dim(\mathscr{S}_{\text{LS}}) = d$, the smallest value possible, as remarked at the end of Section 2.

Recall the definitions of LAD and LMS. Let $Y \in \mathscr{Y}$. An LAD estimate for Y is any p-vector $\hat{\beta}$ s.t. $(b_0, b_1^T)^T = \hat{\beta}$ minimizes

$$\sum_{i=1}^{n} \left| y_i - b_0 - x_i b_1 \right|. \tag{3.1}$$

An LMS estimate for Y is any p-vector $\hat{\beta}$ s.t. $(b_0, b_1^T)^T = \hat{\beta}$ minimizes

$$\text{the median}\left\{(y_i - b_0 - x_i b_1)^2,\ i = 1, \ldots, n\right\}, \tag{3.2}$$

where “the median” is the midpoint of the interval of medians. In general, if $[n/2] \leq k$, where $[n/2]$ is the integer part of $n/2$, there is no unique LMS estimate.

The following is the main result of this paper. The basic idea of the proof is sketched below. See Ellis (1997) for details. Note that, as remarked in Section 1, how, or even whether, LAD (LMS) is defined at data sets where (3.1) [respectively, (3.2)] is not uniquely minimized is irrelevant to the theorem.

THEOREM 3.1. *Suppose* $R=$LAD *or LMS.* (*Assume* $[n/2] > k$ *in the LMS case.*) *Then* $\mathscr{Y}'_R$ *is dense in* $\mathscr{Y}$, $\mathscr{P}_k \cap \mathscr{Y}'_R$ *is dense in* $\mathscr{P}_k$, (2.2) *holds with* $\Phi = \Phi_R$ (*so* Φ_R *is a plane-fitter*) *and* $\dim(\mathscr{S}_R \cap \mathscr{P}_k) < d$. *Hence, by Theorem* 2.2, $\dim(\mathscr{S}_R) \geq np - 2$.

Thus, LAD and LMS have very many singularities. Instability is apparently an important problem for LAD and LMS. Because, in terms of dimension, most collinear data sets are in $\mathscr{P}_k$ (again, in terms of dimension) the vast bulk of the singularities of LAD and LMS are at noncollinear data sets.

Now, as observed at the end of Section 2, $np - 2 \geq d$. Moreover, in regression we typically have $n - k > 2$, which implies $np - 2 > d$. (In the regression case, $p = k + 1$ and $d = nk + k$.) Hence, by Theorem 3.1 and the fact that, as we just observed, $\dim(\mathscr{S}_{\text{LS}}) = d$, the dimensions of the singular sets of LAD and LMS are always at least as large as that of LS and typically strictly larger.

Some may object that it is not the fitted plane that is of interest, but rather the estimated coefficient vector, $\hat{\beta}_R$. However, the plane determined by a coefficient vector b is a continuous function of b. Therefore, any singularity of Φ_R is a singularity of $\hat{\beta}_R$.

Hettmansperger and Sheather (1992) speculate that LMS is unstable because (3.2) is not based on a norm. However, (3.1) is based on a norm and LAD is nevertheless often unstable.

In the $k = 1$ case the idea of the proof of Theorem 3.1 is easy to describe. First, consider LAD. We must show $\dim(\mathscr{S}_{\text{LAD}} \cap \mathscr{P}_k) < d$. Clearly, the noncollinear data sets in $\mathscr{P}_k$ are not singularities of LAD, so any singularity in $\mathscr{P}_k$ must be collinear. The data sets shown in Figure 1 are slight perturbations of a collinear data set Y in which the responses are distinct. The dotted lines in the plots in Figure 1 are the LAD lines. Notice that unlike the LS lines in Figure 1, both LAD lines are nearly vertical and, therefore, not far apart. In fact, Y is not a singularity of the fitted LAD *line*. (On the other hand, Y *is* a singularity of the LAD *slope*. Indeed, the LAD slopes, -29.8 in Figure 1a and 123.5 in Figure 1b, are quite far apart.)

It is easy to see why this is true. By Theorem 1 in Bloomfield and Steiger (1983, page 7), the LAD line for any data set in $\mathscr{Y}'_{\text{LAD}}$ must pass through at least two data points (e.g., see Figure 1; recall that singularity of LAD is defined in terms of $\mathscr{Y}'_{\text{LAD}}$). Since the responses in Y are distinct, in a slight perturbation of Y the vertical separation between any two observations will be relatively large compared to their horizontal separation. Therefore, any line passing through two data points will be vertical or nearly so. Thus, if the perturbed data set is in $\mathscr{Y}'_{\text{LAD}}$, the LAD line will be nearly vertical. Indeed, as $Y' \to Y$ through $\mathscr{Y}'_{\text{LAD}}$, $\Phi_{\text{LAD}}(Y')$ approaches vertical; that is, $\lim_{Y' \to Y,\, Y' \in \mathscr{Y}'_{\text{LAD}}} \Phi_{\text{LAD}}(Y')$ exists, so $Y \notin \mathscr{S}_{\text{LAD}}$. Thus, the only possible singularities of LAD in $\mathscr{P}_k$ are collinear with some responses equal. The space of all such data sets has lower dimension than the space $\mathscr{C}$ of all collinear data sets, but above we observed that $\dim(\mathscr{C}) = d$; that is, $\dim(\mathscr{S}_{\text{LAD}} \cap \mathscr{P}_k) < d$, so the hypothesis of Theorem 2.2 holds.

Next, consider the LMS case. A complication is that, as Figure 3b illustrates, not all elements of $\mathscr{S}_{\text{LMS}} \cap \mathscr{P}_k$ are collinear. For simplicity, still assume $k = 1$. It turns out that all noncollinear data sets in $\mathscr{P}_k$ which are singularities of LMS must be like the data in Figure 3b in that at least three observations must coincide, but the collection of all such data sets has dimension less than d.

It remains to consider collinear data in $\mathscr{P}_k$. The argument proceeds as in the LAD case. In fact, in Figure 1 the dashed lines are the LMS lines. They show virtually no change from one data set to the other. However, this time it is more useful to focus, not on the regression line itself, but on the two lines parallel to it and separated from it by vertical distances equal to the square root of the median of the squared residuals. One checks that an analog of Theorem 1 in Bloomfield and Steiger (1983, page 7) holds here: If $Y \in \mathscr{Y}'_{\text{LMS}}$ then these two lines, bracketing the LMS line, must, between them, pass through at least three data points. The rest of the argument proceeds as in the LAD case.

Thus, surprisingly, unlike LS fitted planes, LAD and LMS planes are typically stable near collinear data sets. Paradoxically, it follows that the two methods are often unstable at noncollinear data sets.

ACKNOWLEDGMENTS

I thank John Mann for permission to use the data plotted in Figure 2. Suggestions by the referees and the Editor improved the presentation.

REFERENCES

BIRKES, D. and DODGE, Y. (1993). *Alternative Methods of Regression*. Wiley, New York.

BLOOMFIELD, P. and STEIGER, W. L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms*. Birkhaüser, Boston.

BOOTHBY, W. M. (1975). *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, New York.

DAVIES, P. L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21** 1843–1899.

DODGE, Y., ed. (1987). *Statistical Data Analysis Based on the L_1-Norm and Related Methods*. North-Holland, Amsterdam.

DODGE, Y., ed. (1992). *L_1-Statistical Analysis and Related Methods*. North-Holland, Amsterdam.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

ELLIS, S. P. (1991). The singularities of fitting planes to data. *Ann. Statist.* **19** 1661–1666.

ELLIS, S. P. (1995a). Dimension of the singular sets of plane-fitters. *Ann. Statist.* **23** 490–501.

ELLIS, S. P. (1995b). A note on the smoothness of L_1-estimators for the linear model. *Sankhyā Ser. A* **57** 221–226.

ELLIS, S. P. (1996). On the size of singular sets of plane-fitters. *Utilitas Math.* **49** 233–242.

ELLIS, S. P. (1997). On the instability of least squares, least absolute deviation, and least median of squares linear regression, unpublished manuscript.

FALCONER, K. (1990). *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, New York.

GENTLE, J. E. (1977). Least absolute values estimation: an introduction. *Comm. Statist. B* **6** 313–328.

HAMPEL, F. R. (1975). Beyond location parameters: robust concepts and methods. *Bull. Internat. Statist. Inst.* **46** 375–382.

HETTMANSPERGER, T. P. and SHEATHER, S. J. (1992). A cautionary note on the method of least median squares. *Amer. Statist.* **46** 79–83.

MALONE, K. M., CORBITT, E. M., LI, S. and MANN, J. J. (1996). Prolactin response to fenfluramine and suicide attempt lethality in major depression. *British J. Psychiatry* **168** 324–329.

MANN, J. J., MCBRIDE, P. A., MALONE, K. M., DEMEO, M. and KEILP, J. (1995). Blunted serotonergic responsivity in depressed inpatients. *Neuropsychopharmacology* **13** 53–64.

MORGAN, F. (1988). *Geometric Measure Theory: A Beginner's Guide.* Academic Press, New York.

PINE, D. S., COHEN, P. and BROOK, J. (1996). Emotional problems during youth as predictors of stature during early adulthood: Results from a prospective epidemiologic study. *Pediatrics* **97** 856–863.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

ROUSSEEUW, P. J. (1994). Unconventional features of positive-breakdown estimators. *Statist. Probab. Lett.* **19** 417–431.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Comment

Stephen Portnoy and Ivan Mizera

1. INTRODUCTION

The ideas presented by Ellis are extremely thought-provoking, especially since the paper makes clear claims contradicting our understanding of LAD regression. As people who thoroughly enjoy puzzles, we were very eager to resolve these contradictions, and our remarks here will concentrate on the stability or instability of LAD regression estimators. The problem appears in the first sentence of the abstract: Ellis defines "unstable" to mean that a small change in the data can cause a large change in the estimator. Since LAD has bounded sensitivity while LS does not, this would seem to contradict the well-known robustness properties of LAD. We were also somewhat uneasy about the measure of stability proposed—the relative size of sets of measure zero would seem to have little applicability for typical statistical models. After considerable contemplation, we believe there are serious problems with the author's claims that LAD is less stable than LS and with interpreting the author's notion of singularity as anything like stability.

First, we note that as the LAD estimator is defined to be set-valued, it is continuous (in the appropriate sense) on exactly the same set of nonsingular designs where LS is continuous. Second, we also argue that sensitivity requires specifying what is "small" and what is "large." We believe that if "small" is with respect to variability in the data, and "large" is with respect to the standard error of the estimator, then the singularity used by Ellis does not indicate instability. In particular, we will show that the example in Figure 3 is *not* one where LAD is highly unstable, and the extent to which it may be more sensitive than LS is a reflection of the ability of conditional quantile analysis to find structure in the data that is missed by LS analysis. Last, we argue that the definition of stability in the paper seems to confuse two rather different concepts: design singularity and nonsingular nonuniqueness of the estimator (the case where the estimator is boundedly nonunique).

Stephen Portnoy is Professor, Department of Statistics, University of Illinois, 725 S. Wright St., Champaign, Illinois 61820 (e-mail: portnoy@stat.uiuc.edu). Ivan Mizera is Assistant Professor, Department of Probability and Statistics, Comenius University, Bratislava, Slovakia.

CONTINUITY OF SET-VALUED ESTIMATORS

Despite the formal distinction between singularity as introduced by Ellis (1995a) and discontinuity (in the usual sense of calculus), it is continuity which is of main theoretical concern here: if the limit exists at a data point, why not (re)define the estimator accordingly? According to its common definition as a minimizer, the LAD estimate is nonunique for some data; that is, it sometimes yields *a set of solutions* at some points. It turns out that these points are exactly the "unstable" ones. Does it mean that LAD is discontinuous at all these points?

As is frequently the case, the truth is not that simple. In our setting, a natural question is: can a small change in the data drive the LAD far away *from the original set of solutions*? (This is the essence of the formal notion of "upper semicontinuity" of set-valued mappings; see Rockafellar and Wets, 1998).

The answer for LAD is *no*. Let us first view LAD as a functional M on the space of the distribution functions for the data; the specific estimator is then $M(F_n)$, where F_n is the empiric distribution of the data (generally, both x and y for regression). If the design space is bounded, we observe the above-mentioned continuity, with respect to weak convergence. We know that the restriction on design space is inevitable, since weak continuity is closely related to qualitative robustness—and we know that LAD is qualitatively robust only with respect to arbitrary departures in y, not in x (see Mizera, 1998, for more details).

Ellis considers LAD simply as a function of the data. In this case, the same type of continuity holds as a consequence of the previous continuity of the LAD functional. This was proved separately by Dupačová (1992), who moreover gave a modulus of continuity. Since at the points with unique LAD our continuity reduces to ordinary continuity, we may claim that LAD is continuous on exactly the same set of nonsingular designs as LS.

It is important to remark that these continuity properties do not arise automatically with the introduction of the set-valued framework; the behavior of LMS, for instance, is far from that straightforward (see again Mizera, 1998).

At singular points, the LAD solutions are unbounded, as are those of LS; and LAD is then continuous at these designs in the same sense as LS: any limit point of solutions for perturbed data tends to a solution for the original data (Rockafellar and Wets, 1998, call this outer semicontinuity).

We agree that the insistence on picking a unique solution from the set of LAD solutions may lead to discontinuity. However, even this does not happen if xs are kept fixed (the setting appropriate in "designed" or "ANOVA" situations): as shown by Ellis (1995b) (see also Ellis and Morgenthaler, 1992), the centroid or Steiner point of the solution set is *uniformly continuous* as a function of ys when xs are constant.

MEDIAN NONUNIQUENESS AND STABILITY

In comparing estimators, it is important to consider their functional representation, not only for applying statistical theory (e.g., classical sensitivity analysis from robustness theory), but also to emphasize that estimators estimate different quantities. These population quantities coincide only in very restrictive homoscedastic models (with symmetric error distributions). In heteroscedastic cases, the values of $M(F)$ for different functionals M at the true model need not coincide; see Portnoy and Welsh (1992) for examples of just how much conditional means and conditional medians can differ. Thus, a statistical theory suggesting conditional medians should not be estimated would not be well received by users who find such parameters natural in their specific problems.

The use of the functional approach has a bonus here: precisely the same type of instability presented in the examples occurs for the one-sample median in even sample sizes. In particular, vanishingly small perturbations of the empirical distribution function can move the median to either endpoint of its interval of definition. There are two reasons why this is not generally considered a cogent criticism of the median. First, it can have no inferential effect: confidence intervals for the population median must contain the entire interval of median solutions. Contrary to the author's claim, this is also the case in Figure 3, as will be shown below. Second, median statistical analysis is most appropriate when the data is close to that generated by a model with a continuous positive density at the median. In such cases, the difference between successive order statistics near the median is of order $\mathscr{O}_p(1/n)$, and thus sensitivity must be rather small. Results in Portnoy (1991a) suggest that this also holds for multiple regression, although standard asymptotics only provides $\mathscr{O}_p(n^{-1/2})$.

ON FIGURE 1.3

An especially troublesome feature of the paper was the plot of the lines in Figure 3 corresponding to the 95% confidence interval for the slope. In analogy with the one-sample case, it is possible to use pairs of regression quantiles to get (simultaneous) confidence intervals for a conditional quantile (Zhou and Portnoy, 1996, 1998). Although these are not directly applicable to confidence intervals on specific coefficients, they would yield confidence bands that would have to include fitted values for *all* solutions for moderately small x-values. Thus, the extremely similar confidence lines in Figure 3 seem highly suspect. It seems clear from the plot that data are from a mixture of two rather different regression lines, and there is rather clear heteroscedasticity. It seemed likely that the LAD was in fact nonunique, and that the two lines in the plot (for the original data and the perturbed data) were the regression quantiles at successive breakpoints (up to perturbation). If this were so, then any confidence set for the LAD would have to cover the entire solution set (i.e., all convex combinations of both lines).

Fortunately, Roger Koenker has developed regression quantile software that is available from Statlib and from his web page:

```
http://www.econ.uiuc.edu/~roger/research
```

Use of the "rq" program in S-PLUS verified the guesses above and provided confidence intervals for the data of Figure 3. The two regression lines were $y = 0.000005 - 0.49574x$ and $-0.00087 + 2.08440x$. The default 95% confidence bounds for the intercept and slope were $(-3.988, 3.610)$ and $(-0.699, 2.211)$. This last interval is much larger than that indicated on Figure 3, and clearly includes both slopes. Here one might argue that the change in the LAD estimate, although not statistically significant, is greater than that of the LS estimator. This possibly greater sensitivity, however, seems to indicate a useful sensitivity of regression quantile methods to important data features to which LS is insensitive. That is, the nonuniqueness and wide confidence intervals might suggest the possibility of a mixture model for the conditional median, while a naive LS analysis (at least one not looking at the data) gives a slope estimate of 0.803 with a SE of 0.309, which would indicate a significant regression with a slope corresponding to neither part of the mixture.

To try to clarify the discrepancy between these results and the confidence intervals Ellis presents, we would like to suggest some places where errors may have occurred. First, the error density seems to be quite small at the median, and thus estimation of the sparsity function may be especially problematic. More important, we believe Ellis used the variance estimate for the i.i.d. model. In heteroscedastic cases, the appropriate variance is given by a "sandwich": $(X'DX)^{-1}X'X(X'DX)^{-1}$, where D is a diagonal matrix of the density evaluated at each observation (see Portnoy, 1991b). The use of the i.i.d. variance estimate is not even consistent in heteroscedastic cases. Koenker (1994) presents several methods for generating asymptotically legitimate confidence intervals for specific coefficients in heteroscedastic cases, and he recommends the one based on inverting a regression quantile rank test for a coefficient. This is the default in the software described above. It is interesting to note that Koenker's software permits the specification of the confidence interval estimate based on the i.i.d. model sparsity estimate. This method uses the Hall–Sheather optimal estimate (with bandwidth $n^{-1/3}$) and gives $(1.335, 2.834)$ as an interval for the slope. Although this estimate is also invalidated by heteroscedasticity, it is still much larger than that of Ellis. Use of a sparsity estimate different from the optimal Hall–Sheather approach may have also contributed to the discrepancy.

It is also legitimate to use the (x, y) bootstrap in heteroscedastic cases. The example here seems to be quite similar to that plotted in the figures in Spady (1991). The bootstrap distribution appears to have four modes: one at each of the solid and dashed lines in Figure 3 and two less obvious ones. This multimodality might suggest inaccuracies in the naive percentile method; but using the percentile (x, y) bootstrap with 1,000 replications generated a 95% confidence interval for the slope of $(-0.783, 2.274)$, corroborating the rank method. Last, it should also be noted that the difficulty of forcing nonuniqueness for a given quantile makes this data highly artificial.

DESIGN SINGULARITY AND NONUNIQUENESS OF LAD

Finally consider the two types of singularity that are combined in the dimension measure used by Ellis. Design singularity is when the design matrix is not of full rank. It is not a problem of sensitivity or stability but of parameter identifiability. Identifiability, of course, has a long and extensive (if not always illustrious) history, especially in ANOVA. As far as we know, identifiability has never been considered as a problem of stability (e.g., to be solved by a statistical choice of an estimation method), but as a problem in interpretation of parameters (whose solution rightfully belongs in the domain of the application). This form of singularity is exactly the same for all estimators. In terms of statistical stability, nearly singular cases have little information concerning the parameters, and confidence intervals must be rather large. In fact, it is easy to check that, in simple linear regression, changes in the Studentized estimate of the slope coefficient can be bounded independently of the design. In particular, if (x, y) and (x^*, y^*) are two pairs of vectors of observations with $\|x - x^*\| \leq \epsilon\|x\|$, then the difference in Studentized slopes is bounded independent of x and x^*; namely,

$$\big|\,\|x - \bar{x}\|(x - \bar{x})'y - \|x^* - \bar{x}^*\|(x^* - \bar{x}^*)'y^*\big| \\ \leq (1 - \epsilon)\big(\|y - y^*\| + \epsilon(\|y^*\| + \|y\|)\big).$$

We believe a similar bound holds for the LAD estimator, but we do not know of a formal version of this result.

The other form of singularity is the bounded nonuniqueness of the LAD (or LMS) estimators. However, this is just a reflection of the fact that such estimators are set-valued. As shown above, the sets tend to be small (at least in a statistical sense). Furthermore, the LAD is continuous as a set-valued function in any case. Because of continuity, this form of nonuniqueness is not necessarily a case of instability; and so it does not seem reasonable to combine these two concepts.

CONCLUSIONS

In summary, the LAD estimator does not appear to exhibit extreme forms of sensitivity or instability. Its bounded nonuniqueness has long been accepted in one-sample problems, and there appear to be no qualitative differences in this aspect arising in the multiple regression problem. Contrary to the claims of Ellis, statistical methods need not be unstable or sensitive near a singularity, neither in absolute terms at points of bounded nonuniqueness, nor in a statistical sense at points of design singularity. We believe that analysis of the size of a change induced by a change in the data is intrinsically a metric property, and that it should be measured by some relative of a modulus of continuity (preferably expressed in terms of the natural statistical variation in the data). The dimension measure of Ellis does not appear to bear on any such metric criteria.

ACKNOWLEDGMENTS

The first author's research is supported in part by NSF Grant 97-03758. The second author's research is supported in part by Slovak GAS Grant 1/4196/97.

Rejoinder

Steven P. Ellis

I thank Portnoy and Mizera for their thought-provoking comments. Amazingly, they seem to deny the very existence of the instability phenomenon in LAD. Hopefully, this exchange will help clarify the issues.

Let me restate the problem in the LAD case. Each panel in Figure 2 in the paper shows two real data sets that are virtually identical but whose LAD lines are quite distinguishable. This is undesirable because a statistical data summary should detect important structures in the data, not trivial ones.

CONTINUITY AND SCALE

Portnoy and Mizera point out that LAD estimates are not, in general, unique. They assert that, when viewed as a set-valued function, LAD is actually continuous. From this they conclude that LAD is not unstable. In the Appendix to this rejoinder I show that the claim that LAD, regarded as a set-valued function, is literally continuous is dubious. However, as a practical matter worrying about the nonuniqueness of LAD is unnecessary. The collection of data sets for which the LAD estimate is nonunique has Lebesgue measure 0. Therefore, under a reasonable interpretation of the assumption I make in the paper that the predictors as well as the responses are free to vary, one will never get a data set with nonunique LAD estimates. Such data sets are important (they are the singularities!), but we do not have to worry about how, or even if, LAD is defined at them. So we need focus attention only on those noncollinear data sets at which the LAD estimate is unique. In the paper I call the collection of all such data sets $\mathscr{Y}'_{\text{LAD}}$. If LAD performs poorly on $\mathscr{Y}'_{\text{LAD}}$, then, a fortiori, it will perform poorly on the collection $\mathscr{Y}$ of all data sets or on the collection of all noncollinear data sets. As a further simplification, take the number k of regressors to be 1. If $Y \in \mathscr{Y}'_{\text{LAD}}$, let $b(Y)$ be the slope of the, necessarily unique, LAD line for Y.

In this simple context, the message of Portnoy and Mizera concerning continuity appears to be: b is continuous on $\mathscr{Y}'_{\text{LAD}}$. So, for example, if we perturb the original data set in each panel in Figure 2 by a small enough amount, the LAD lines of the original and perturbed data sets will nearly agree. However, for these data sets, that perturbation would have to be smaller than 1/20,000 of the IQR of x. I doubt if most statisticians would find this acceptable. In general we have the following:

(∗) Given any $\delta > 0$ one can find lots of data sets $Y \in \mathscr{Y}'_{\mathrm{LAD}}$ whose LAD slopes change by a relatively large amount after moving Y by less than δ units.

(Hint: Look near the singular set.) In particular, b is not *uniformly* continuous on $\mathscr{Y}'_{\mathrm{LAD}}$. A worrisome feature of (∗) is that it holds even when the scale of Y is specified.

Let $Y \in \mathscr{Y}'_{\mathrm{LAD}}$ and let $Y' \in \mathscr{Y}'_{\mathrm{LAD}}$ be a data set near Y. The instability of LAD at Y has to do with the distance between $b(Y)$ and $b(Y')$ compared to that between Y and Y'. Portnoy and Mizera suggest that we should measure the distance between Y and Y' relative to the spread in Y and the distance between $b(Y)$ and $b(Y')$ relative to the standard error (SE) of $b(Y)$. Actually, the distances should be measured in units appropriate for the practical problem at hand. Measuring distance(Y, Y') in terms of the spread in Y is often sensible. (I use that scale myself in Figure 2.) However, measuring distance$[b(Y), b(Y')]$ relative to the SE of $b(Y)$ represents the same mistake as regarding the p-value as a measure of the importance of an effect (Freedman, Pisani, Purves and Adhikari, 1991, Chapter 29, Section 3). In particular, an SE that is gigantic in practical terms is a poor yardstick.

In any case, no matter what (reasonable) scale we choose for measuring perturbations of data and slope, the LAD slope b is not uniformly continuous. I give an analytical example of this in the Appendix to this rejoinder. In fact, in the example I even implement Portnoy and Mizera's suggestion and measure perturbations in both Y and $b(Y)$ relative to estimates of spread.

The question of scale is part of a larger problem. How should instability be measured? (See Belsley, 1991, Chapter 11 for a general examination of this issue.) We have just seen that if one measures it in relative terms, that is, by

$$\text{distance}[b(Y), b(Y')]/\text{distance}[Y, Y'],$$

then using any reasonable, even data dependent, scales, the impact of perturbation can be arbitrarily large.

However, the relative change,

$$\text{distance}[b(Y), b(Y')]/\text{distance}(Y, Y'),$$

is not the only measure of interest. The absolute change in the slope, $|b(Y) - b(Y')|$, is also important. As a practical matter, who cares if a microscopic change in Y leads to a wildly disproportionate, but still tiny, change in $b(Y)$? In the paper I take a first step toward addressing this question by mentioning that if one restricts attention to displacements in the fitted line of about 90°, the conclusions of the paper still hold. Even in SE units, a jump in the LAD line of 90° will usually mean a large change in the slope.

QUANTIFICATION

I agree with Portnoy and Mizera that instability is ultimately a "metric property." In my paper I focused on the most extreme situations: In relative terms, a regression method is *infinitely* unstable at a singularity; in absolute terms, 90° is as far apart as two lines through the origin can be. For such extreme instability, one need not be fussy about how instability is measured.

The notion of singularity describes the behavior of a regression method in metric terms "in the limit" as the data approach the singularity. Another metric property is the volume (or probability) of the collection of data sets within δ of the singular set $\mathscr{S}$ (and within a ball centered at the origin, say). The Hausdorff dimension of $\mathscr{S}$ describes the limiting behavior of this volume as $\delta \downarrow 0$. (See Ellis, 1995a.)

More careful investigation of the instability problem will require quantification of the nonlimiting behavior of regression methods. When this quantification is done, I expect the region of high instability will be found to be near the singular set. In the case of LAD and LMS, that region will be large.

FIGURE 3A

My purpose in creating Figure 3a was to show an example of a data set at which the amplitude of the instability of LAD is very large. It was also intended to show that we cannot count on confidence intervals based on standard asymptotics to serve as diagnostics of instability. On the other hand, I mention in the text that the bootstrap interval did detect the instability.

Portnoy and Mizera do make a valid criticism of the example in Figure 3a. Indeed, in constructing the interval portrayed in the plot, I *did* use the variance estimate for the i.i.d. model. I should have pointed that out in the text. For the record, I computed the confidence interval based on the asymptotic normal distribution (Bloomfield and Steiger, 1983, Theorem 1, page 64) using (5.5), page 137 in

Venables and Ripley (1994) to calculate a bandwidth for a Gaussian kernel estimate of the density of the residuals at 0. I make no claims that this is optimal, only that it is a fair interpretation of "standard asymptotics."

Portnoy and Mizera point out that the sensitivity of LAD indicates a "useful sensitivity... to important data features." However, one must not form a stereotype of singularities as data sets that have important data features like the data set in Figure 3a of the paper. The nearly singular data sets in Figure 2 are ordinary-looking data sets without important data features. Figure 2 also points up the fact that, besides having a "useful sensitivity... to important data features," LAD also has an annoying sensitivity to unimportant data features.

COLLINEARITY

In the paper, I only mention collinearity in reference to LAD as part of describing the proof of Theorem 3.1. In general, the collection of noncollinear singularities of LAD dwarfs the collection $\mathscr{C}$ of collinear data sets, so, except for discussing the proof of the theorem, there is not much reason to discuss the stability of LAD near $\mathscr{C}$.

Portnoy and Mizera remark that, normalized by the standard error, the LS slope is stable near $\mathscr{C}$. I refer the reader to my earlier comments about using the standard error as a unit of displacement of slope.

APPENDIX

Continuity of LAD as a Set-valued Function

Portnoy and Mizera claim that if we regard LAD as a set-valued function, then it is, in fact, "continuous." For example, they write, "LAD is continuous on exactly the same set of nonsingular designs as LS." I will interpret "nonsingular" here to mean "noncollinear." Let $\tilde{\mathscr{Y}} \equiv \mathscr{Y} \setminus \mathscr{C}$ denote the set of noncollinear data sets. For the sake of completeness, let us explore the possibility that LAD, regarded as a set-valued function, is literally continuous on $\tilde{\mathscr{Y}}$. To keep things very simple, I will consider simple linear regression ($k = 1$) with $n = 5$ observations. If $Y \in \tilde{\mathscr{Y}}$ is a data set, let $B(Y)$ denote the collection of the slopes in all solutions to the LAD minimization problem (see expression (3.1) in the paper). Then $B(Y)$ is a bounded closed interval, though typically of length 0.

The issue here is continuity of B. Let X be the space of all bounded closed intervals. So each bounded closed interval is a point of X and B maps $\tilde{\mathscr{Y}}$ into X. In order for B to be continuous, X must be equipped with some topology. Which topology? It is easy to prove the following.

Proposition 1. *If X is equipped with a topology that makes B continuous then that topology does not satisfy the T_1 separation property.*

Recall that "X does not satisfy the T_1 separation property" means that there are distinct points $x_0, x_1 \in X$ such that every neighborhood of x_0 contains x_1 (Simmons, 1963, page 130). Now, the T_1 separation property is probably the most basic separation property a topological space can have. So if X lacks that property, it is a very strange space. In particular, metrizability of X is out of the question. Such a strange space is unlikely to have any meaning for regression. Thus, it appears that B is not literally continuous in any meaningful sense. (On the other hand, Portnoy and Mizera mention *semi*continuity. I can accept semicontinuity of B, but not continuity.)

Proof of Proposition 1. Remember that I am taking $k = 1$ and $n = 5$. Consider the data set $Y_{s,t} = ((1 + s, 1 + s)^T, (-1, 1)^T, (-1, -1)^T, (1 + t, -1 - t)^T, (0, 0)^T)^T$, for some s, t. Thus, Y consists of the origin together with the corners of the square with sides of length 2 centered at the origin with the two rightmost corners possibly displaced diagonally. Temporarily set $t = 0$. If $s > 0$, then $B(Y_{s,0}) = [1, 1] = \{1\}$; $B(Y_{0,0}) = [-1, 1]$ ($Y_{0,0}$ is a singularity of LAD). Write $x_1 = \{1\}$ and $x_0 = [-1, 1]$. So $x_1 \neq x_0$.

Denote the topology of X by τ. Let s_m, $m = 1, 2, \ldots$, be a sequence of positive numbers decreasing to 0. Then $Y_{s_m,0} \to Y_{0,0}$ as $m \to \infty$. Since, by hypothesis, B is continuous, $x_1 = B(Y_{s_m,0}) \to B(Y_{0,0}) = x_0$. Thus, if $U \in \tau$ is any neighborhood of x_0, then for m sufficiently large $x_1 = B(Y_{s_m,0}) \in U$. That is, any neighborhood of x_0 contains x_1. □

Example of Arbitrarily Close Data Sets with Far-apart LAD Solutions

I will reuse the data set from the preceding proof, but this time I will make use of t. Let $s, t > 0$. Then, as before, the LAD line of $Y_{s,0}$ is unique and has slope 1. Similarly, the LAD line of $Y_{0,t}$ is unique and has slope -1. In particular, $Y_{s,0}, Y_{0,t} \in \mathscr{Y}'_{\mathrm{LAD}}$.

Let $\delta > 0$ be given and let $s = t = \delta/4$. Then in the Euclidean norm, $|Y_{s,0} - Y_{0,t}| = \delta/2 < \delta$. However, the slopes of the LAD lines of $Y_{s,0}$ and $Y_{0,t}$ differ by 2, independent of δ.

Actually, this example illustrates the scale invariance of the instability of the LAD slope b measured in relative terms. Let $s_Y(Y)$ be a measure of spread in $Y \in \mathscr{Y}$. Let $s_b(Y)$ be an estimate of the SE of b for $Y \in \mathscr{Y}'_{\mathrm{LAD}}$. Suppose that s_Y and s_b are sufficiently well behaved that $s_b(Y)$ converges to a finite limit as $Y \to Y_{0,0}$ through $\mathscr{Y}'_{\mathrm{LAD}}$ and s_Y is defined, finite and continuous at $Y_{0,0}$. Also suppose $s_Y(Y_{0,0}) > 0$. (For example, the standard deviations and interquartile ranges of the predictors and responses in $Y_{0,0}$ are all positive. Both eigenvalues of the covariance matrix of $Y_{0,0}$ are positive.) The relative change in b corresponding to the change from $Y_{\delta/4,0}$ to $Y_{0,\delta/4}$, measured in s_b, s_Y units, is

$$\frac{|b(Y_{\delta/4,0}) - b(Y_{0,\delta/4})|/s_b(Y_{\delta/4,0})}{|Y_{\delta/4,0} - Y_{0,\delta/4}|/s_Y(Y_{\delta/4,0})}.$$

This goes to ∞ as $d \downarrow 0$.

ADDITIONAL REFERENCES

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley, New York.

Dupačová J. (1992). Robustness of L_1 regression in the light of linear programming. In *L_1-Statistical Analysis and Related Methods* (Y. Dodge, ed.) 47–61. North-Holland, Amsterdam.

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

Freedman, F., Pisani, R., Purves, R. and Adhikari, A. (1991). *Statistics*, 2nd ed. Norton, New York.

Koenker, R. (1994). Confidence intervals for regression quantiles. In *Asymptotic Statistics: Proceedings of the 5th Prague Symposium* (P. Mandl and M. Hušková, eds.) 349–359. Springer, Berlin.

Mizera, I. (1998). On continuity: resistance and qualitative robustness. Preprint. (Available at http://www.uniba.sk/~mizera.)

Portnoy, S. (1991a). Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J. Sci. Statist. Comput.* **12** 867–883.

Portnoy, S. (1991b). Asymptotic behavior of regression quantiles in nonstationary, dependent cases. *J. Multivariate Anal.* **38** 100–113.

Portnoy, S. and Welsh, A. (1992). Exactly what is being modelled by the systematic component of a heteroscedastic linear regression. *Statist. Probab. Lett.* **13** 253–258.

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational Analysis*. Springer, Berlin.

Simmons, G. F. (1963). *Introduction to Topology and Modern Analysis*. McGraw-Hill, New York.

Spady, R. (1991). Saddlepoint approximations for regression models. *Biometrika* **78** 879–889.

Venables, W. N. and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer, New York.

Zhou, Q. and Portnoy, S. (1996). Direct use of regression quantiles to construct confidence sets for linear models. *Ann. Statist.* **24** 287–306.

Zhou, Q. and Portnoy, S. (1998). Statistical inference on heteroscedastic linear models based on regression quantiles. *J. Nonparametr. Statist.* **9** 239–260.