Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Leverage and Breakdown in L_1 Regression

Steven P. Ellis & Stephan Morgenthaler

To cite this article: Steven P. Ellis & Stephan Morgenthaler (1992) Leverage and Breakdown in L_1 Regression, Journal of the American Statistical Association, 87:417, 143-148, DOI: 10.1080/01621459.1992.10475185

To link to this article: https://doi.org/10.1080/01621459.1992.10475185

Leverage and Breakdown in L_1 Regression

STEVEN P. ELLIS and STEPHAN MORGENTHALER*

In this article the notion of leverage of a design point when fitting a linear regression model is interpreted geometrically. In the case of least squares fitting, the leverage indicators based on the diagonal of the hat matrix are widely applied. By interpreting these hat matrix indicators geometrically, leverage can be generalized to groups of design points, as well as to other methods of fitting. The article introduces a leverage indicator that is appropriate for L_1 regression and discusses some aspects of this new diagnostic. It is shown that, in the case of L_1 regression, the leverage indicators have a precise interpretation. They tell us about the breakdown and/or the exactness of fit. As an application, the article considers the maximal possible breakdown value of L_1 regression and the choice of designs that achieve this maximum.

KEY WORDS: Design-dependent weights; Geometry of L_1 regression; Robust designs; Robustness.

1. INTRODUCTION

Methods of data analysis—if one understands this term in an appropriately broad sense—cannot be justified and evaluated on the basis of probabilistic models alone. Rather, probabilistic models must be interpreted as *leading cases* (Mallows and Tukey 1982), that is, as overly simple, yet useful, ideal situations. Next to probabilistic techniques, it is quite natural to evaluate data-analytic procedures by way of geometrical notions. Examples of such an approach include the well-known geometrical interpretation of least squares, the adaptation of robustness indicators like influence and breakdown to samples (Andrews et al. 1972, sec. 5E, Donoho and Huber 1983), the geometrical interpretations of correlation coefficients (Rodgers and Nicewander 1988), the use of distances to describe data (Ellis and Morgenthaler 1987), and others. This article examines the notion of a *leverage point* in this manner.

Let our data consist of a vector $\mathbf{y} \in \mathbb{R}^n$, the response variable, and vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{R}^n$, the explanatory variables. We assume a linear model

$$\mathbf{y} = \theta_1 \mathbf{x}_1 + \cdots + \theta_p \mathbf{x}_p + \boldsymbol{\epsilon}, \tag{1.1}$$

where the vector of errors $\boldsymbol{\epsilon}$ is assumed to have mean zero. *Leverage* is a diagnostic notion, and, as such, does not have a precise meaning. Vaguely stated, a design point far from the bulk of the others is called a leverage point. Leverage points should be distinguished from *influential points*. An observation taken at a leverage point has the potential to influence the fit, but it does not necessarily do so.

To be concrete, consider the least squares fit of (1.1). The vector of fitted values $\hat{\mathbf{y}}$ is found by projecting $\mathbf{y}$ orthogonally onto $F = \langle \mathbf{x}_1, \ldots, \mathbf{x}_p \rangle$, the linear subspace spanned by the explanatory variables. This projection is a linear mapping, $\hat{\mathbf{y}} = H\mathbf{y}$, with $H = (h_{ij})$ being the so-called hat matrix. In this situation, the diagonal elements of H are used as diagnostic indicators of leverage. Three reasons for this choice are given in the literature on linear model diagnostics:

$$\text{First, } \hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j,$$

$$\text{Second, } \operatorname{var}(y_i - \hat{y}_i) = (1 - h_{ii}) \operatorname{var}(\epsilon_1),$$

and

$$\text{Third, } h_{ii} = 1/n + \left(\text{Mahalanobis distance of } \begin{pmatrix} x_{1i} \\ \vdots \\ x_{pi} \end{pmatrix} \text{ from } \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} \right)^2 \Big/ (n-1).$$

In the last formula $\bar{x}_k$ denotes the mean of the kth explanatory variable. The second equation assumes that the errors are homoscedastic and uncorrelated, whereas the third equation is true only when we include a constant in our regression equation, that is, when the first explanatory vector $\mathbf{x}_1$ contains all 1's. Since H is symmetric and idempotent, the first equation shows that h_{ii} close to 1—Hoaglin and Welsch (1978) suggest the bound $h_{ii} > 2p/n$—implies that $\hat{y}_i$ must necessarily be close to y_i. The second equation shows the same probabilistically, since h_{ii} close to 1 implies a small variance for the ith residual. The third equation, finally, interprets h_{ii} with the help of a classical outlier detection tool. The first two equations show the potential of observations taken at leverage points to influence the least squares fit. The third equation shows that leverage is connected to outlyingness in the design.

There is yet another equation, which we will take as a starting point for our own investigation, namely,

$$\arccos((h_{ii})^{1/2}) = \text{smallest angle between a direction in } F = \langle \mathbf{x}_1, \ldots, \mathbf{x}_p \rangle \text{ and the } i\text{th coordinate vector } \mathbf{e}_i.$$

This last equation is verified most easily by computing the square of the cosine of the angle between the ith coordinate vector $\mathbf{e}_i$ and its orthogonal projection onto F,

$$h_{ii} = \frac{(\mathbf{e}_i^T (H\mathbf{e}_i))^2}{((H\mathbf{e}_i)^T (H\mathbf{e}_i))} = \mathbf{e}_i^T (H\mathbf{e}_i),$$

where the superscript T denotes transposition. We again made use of the fact that H is symmetric and idempotent to arrive

* Steven P. Ellis is Assistant Professor of Statistics and Biostatistics, Department of Statistics, University of Rochester, Rochester, NY 14627. Stephan Morgenthaler is Professor of Statistics, EPFL–DMA, 1015 Lausanne, Switzerland. The research for this article was supported in part by Swiss National Science Foundation Grant 21-26491.89, awarded to EPFL.

at the second equality in this equation. This formula for h_{ii} shows, once more, the potential for influence of observations taken at leverage points. A value of h_{ii} close to 1 implies a small minimal angle between F and the ith coordinate axis. This is "bad" in a resistance sense, because a translation of the observed responses $\mathbf{y}$ in the direction of the ith axis will be almost parallel to F and thus strongly "influences" the projection by H. Note in particular that, in the case where the vector $\mathbf{e}_i$ is itself an element of F, changing the data in that direction must change the projection by exactly the same amount. In this case we have $h_{ii} = 1$. The minimal angle can also be computed via

$$h_{ii} = \max_{\mathbf{f}\in F} \frac{(\mathbf{f}^T(P_E\mathbf{f}))^2}{(\mathbf{f}^T\mathbf{f})(\mathbf{f}^T P_E\mathbf{f})}$$

$$= \max_{\mathbf{f}\in F} \frac{f_i^2}{f_1^2 + \cdots + f_n^2},$$

where P_E denotes the orthogonal projection onto E, the linear space spanned by $\mathbf{e}_i$.

The interpretation of leverage via angles suggests generalizations in two directions. On one hand, we may speak of the leverage not of a single point, but rather of a group of points, say with indices $i_1, \ldots, i_m$. The generalized leverage indicator would be the squared cosine value of the minimal angle between two directions, one belonging to the linear space F, the other lying in $E_{i_1,\ldots,i_m}$, the linear space spanned by the coordinate vectors $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_m}$. The generalized leverage diagnostic that applies to the group of indices $\{i_1, \ldots, i_m\}$ can be written as

$$\max_{\mathbf{f}\in F} \frac{f_{i_1}^2 + \cdots + f_{i_m}^2}{f_1^2 + \cdots + f_n^2}. \tag{1.2}$$

This notion of leverage applicable to groups of points is explored in Clerc and Morgenthaler (1989).

The second possible generalization is to other methods of fitting. If, instead of minimizing the L_2 norm, we minimize the L_1 norm in calculating the vector of fitted values, then it would be natural to replace the diagnostic (1.2) by

$$\max_{\mathbf{f}\in F} \frac{|f_{i_1}| + \cdots + |f_{i_m}|}{|f_1| + \cdots + |f_n|}. \tag{1.3}$$

If we interpret leverage as the potential to influence a fit, it is to be expected that different methods of fitting require different leverage indicators. The indicator (1.3) is to be used in this sense for the least absolute deviation method.

In the next section, the leverage indicator (1.3) is examined more closely, and the relation to (1.2) is discussed. We will show that the leverage diagnostic (1.3) has a direct interpretation involving breakdown and exactness of fit. The breakdown of regression estimators is discussed in Rousseeuw and Leroy (1987, p. 10), where it was shown that the L_1-regression estimate breaks down when contaminating a single observation. This is true, however, only if it is allowable to change the response y_i as well as the corresponding explanatory variables $x_{1i}, \ldots, x_{pi}$ in an arbitrary way. In this sense, L_1 regression is as nonresistant as the least squares estimator. But the L_1 fit is often quite resistant, if only the responses y_i can be altered arbitrarily. The diagnostic indicator (1.3) tells us exactly how resistant. Our article shows that

> if, for an $m < n/2$, it is true for any choice of $i_1, \ldots, i_m$ that the maximum in (1.3) is smaller than 1/2, then the L_1 regression does not break down when any set of m responses $y_{i_1}, \ldots, y_{i_m}$ is changed arbitrarily.

This ease of interpretation is different from the last squares case, where the interpretation of the leverage diagnostic (1.2) remains quite obscure. If we restrict attention to the hat matrix diagonals, for example, it is possible to construct designs with highly influential observations that, nonetheless, have small h_{ii} values.

2. THE EXACT-FIT PROPERTY OF THE L_1 ESTIMATOR

Many regression estimators satisfy a minimal requirement that expresses the fact that the estimate uses only the data at hand, they are *equivariant*. Formally, equivariance means that, if we transform the responses $\mathbf{y}$ to $\mathbf{y}^* = \sigma(\mathbf{y} + \tau_1\mathbf{x}_1 + \cdots + \tau_p\mathbf{x}_p)$, then we obtain the new estimates by a simple transformation of the old ones, $\hat{\theta}_i(\mathbf{y}^*) = \sigma(\hat{\theta}_i(\mathbf{y}) + \tau_i)$, for $i = 1, \ldots, p$. In these transformations $\sigma > 0$ and $\tau_1, \ldots, \tau_p$ are scalars. A simple resistance property for such an estimator of linear model parameters was proposed by Rousseeuw (1984).

Definition. An equivariant regression estimator is said to have the *exact-fit property of order m,* if, whenever at most m of the responses $(y_1, \ldots, y_n)$ are nonzero, but otherwise arbitrary, and all the others are equal to zero, then the fit $(\hat{y}_1, \ldots, \hat{y}_n)$ is equal to zero. Related to this concept is the *exact-fit point,* which is the largest m for which the exact-fit property holds. This exact-fit point is somewhat easier to determine than the breakdown point, and there is a close connection between the two.

Definition. An equivariant regression estimator is said to *break down at order m,* if there exists a response vector $\mathbf{y}$ with the property that, with appropriate changes of $(m + 1)$ of the n components of $\mathbf{y}$, the corresponding vectors of fitted values can be made to be arbitrarily far apart. The *breakdown point* is the smallest order of breakdown.

Remark. Our definition of breakdown must, in each case, be complemented by a choice of distance, which makes precise what we mean by "far apart."

For equivariant estimators it can be shown that the breakdown point is at most equal to the exact-fit point (Rousseeuw and Leroy, 1987, p. 123). To prove equality, that is, to demonstrate that the exact-fit point cannot be bigger than the breakdown point, additional assumptions about the fitting procedure must be made.

Proposition 2.1. Let ρ be a metric on $\mathbb{R}^n$ such that the distance between two points depends only on the difference between the points. Suppose the equivariant map $\mathbf{y} \to \hat{\mathbf{y}}$ that takes a vector of responses into a vector of fitted values is uniformly continuous with respect to ρ. It then follows that the breakdown point and the exact-fit point are equal.

Proof. Suppose that the breakdown point m_b and the exact-fit point m_e satisfy the strict inequality $m_b < m_e$. Then

there exists a vector of responses $\mathbf{z}$ and indices $i_1, \ldots, i_{m_b+1}$, as well as a sequence of response vectors $(\mathbf{z}_k)_{k=1}^{\infty}$, such that (a) $\mathbf{z}$ and $\mathbf{z}_k$ differ only at positions $i_1, \ldots, i_{m_b+1}$, and (b) $\rho(\hat{\mathbf{z}}, \hat{\mathbf{z}}_k)$ is unbounded. Denote by $\mathbf{y}$ the vector defined by $y_{i_j} = z_{i_j}$ for $j = 1, \ldots, m_b$, with all the other components of $\mathbf{y}$ being equal to zero. Similarly, denote by $\mathbf{y}_k$ the vector obtained by setting all components of $\mathbf{z}_k$ equal to zero except those at positions $i_1, \ldots, i_{m_b+1}$. Since $m_e > m_b$, the fitted values corresponding to $\mathbf{y}$ and $\mathbf{y}_k$ are uniquely determined and equal to zero. Furthermore, we have $\rho(\mathbf{z}, \mathbf{y}) = \rho(\mathbf{z}_k, \mathbf{y}_k) = \delta$, because of (a) and the construction of $\mathbf{y}$ and $\mathbf{y}_k$. This last fact together with (b) contradicts uniform continuity.

The least squares fit is linear in the data $\mathbf{y}$ and uniformly continuous with regard to the L_2 norm. A more interesting example is the L_1 fit $\mathbf{y} \to \hat{\mathbf{y}}$, which is determined by the property $\|\mathbf{y} - \hat{\mathbf{y}}\|_1 \le \|\mathbf{y} - \mathbf{f}\|_1$, for all $\mathbf{f} \in F$, where $\|\cdot\|_1$ denotes the L_1 norm. This map is continuous with regard to the L_1 norm at any $\mathbf{y}$ for which a unique $\hat{\mathbf{y}}$ exists. To see why, consider two points $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, with $\|\mathbf{y} - \mathbf{z}\|_1 \le \delta$. Using the triangle inequality and the fact that $\|\mathbf{z} - \hat{\mathbf{z}}\|_1 \le \|\mathbf{z} - \mathbf{f}\|_1$ for all $\mathbf{f} \in F$, we have

$$\begin{aligned} \|\mathbf{y} - \hat{\mathbf{z}}\|_1 &\le \|\mathbf{z} - \hat{\mathbf{z}}\|_1 + \delta \\ &\le \|\mathbf{z} - \hat{\mathbf{y}}\|_1 + \delta \\ &\le \|\mathbf{y} - \hat{\mathbf{y}}\|_1 + 2\delta. \end{aligned} \tag{2.1}$$

This inequality, together with the uniqueness of the projection $\mathbf{y}$, proves that the mapping $\mathbf{y} \to \hat{\mathbf{y}}$ is continuous. The set $F \cap \{\mathbf{z} : \|\mathbf{y} - \hat{\mathbf{z}}\|_1 \le \|\mathbf{y} - \hat{\mathbf{y}}\|_1 + 2\delta\}$ is compact and shrinks to $\{\hat{\mathbf{y}}\}$ as δ tends to 0.

Since the unit ball of the L_1 norm is a polyhedron with flat faces, the possibility that the projection is not unique exists. In those cases we must define $\hat{\mathbf{y}}$ appropriately.

Proposition 2.2. Let $\mathbf{y} \to \hat{\mathbf{y}}$ be the equivariant L_1 projection onto the linear space F obtained by selecting the Steiner point of the solution set. This map is uniformly continuous.

Proof. Consider first the case where a unique $\hat{\mathbf{y}}$ exists for each $\mathbf{y} \in \mathbb{R}^n$. Since $\{\mathbf{z} : \|\mathbf{y} - \mathbf{z}\|_1 \le \|\mathbf{y} - \hat{\mathbf{y}}\|_1 + 2\delta\}$ is an n-dimensional polyhedron, the set $F \cap \{\mathbf{z} : \|\mathbf{y} - \mathbf{z}\|_1 \le \|\mathbf{y} - \hat{\mathbf{y}}\|_1 + 2\delta\}$ is a p-dimensional polyhedron with a diameter proportional to δ for values of δ that are small compared to $\|\mathbf{y} - \hat{\mathbf{y}}\|_1$. Furthermore, this proportionality constant is bounded from above uniformly in $\mathbf{y}$ for large enough $\|\mathbf{y} - \hat{\mathbf{y}}\|_1$ because of the uniqueness of $\hat{\mathbf{y}}$. It is easy to verify uniform continuity in sets of the form $\{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \hat{\mathbf{y}}\|_1 \le c\}$ for finite c, so that the preceding comments imply uniform continuity on $\mathbb{R}^n$.

If the L_1 projection is not unique, the set of solutions $P(\mathbf{y}) = \{\mathbf{f} \in F : \|\mathbf{y} - \mathbf{f}\|_1 \le \|\mathbf{y} - \hat{\mathbf{y}}\|_1\}$ is a convex polyhedron of dimension $\le p$. For $\mathbf{z}$ with $\|\mathbf{y} - \mathbf{z}\|_1 \le \delta$, (2.1) shows that the set of solutions $P(\mathbf{z})$ is within a multiple of $P(\mathbf{y})$, and vice versa. Grünbaum (1967, p. 314) shows that the Steiner point is uniformly continuous.

Remarks. Ellis (1991) has proved uniform continuity for the centroid of the solution set. We conjecture that uniform continuity holds for any equivariant continuous L_1 projection.

It turns out that the leverage diagnostics that we discussed in Section 1 can be used to determine the order of the exactness-of-fit of the L_1-regression. As an illustration, consider the location problem.

Example. When we fit the model $y_i = \mu + \epsilon_i$, the space F is one-dimensional, spanned by the n-vector whose components are all equal to 1. In this case, for $m = 1$, the least squares leverage indicator and the L_1 leverage are identical and equal to $1/n$ (see 1.2 and 1.3). The generalizations to groups of m points leads to a leverage value of m/n in both cases. For this simple situation, we also understand the exactness-of-fit property. The L_1 fit of location, the median, has the exact-fit property of any order $m < n/2$. This is a special case of the result announced in the introduction. A bound of $1/2$ on the L_1-leverage indicator for groups of m points implies the exact-fit property of order m.

We now state the theorem of which this example is a special case.

Theorem 2.3. Let $i_1, \ldots, i_m$ be fixed indices, and let E be the linear space spanned by coordinate axes $i_1, \ldots, i_m$. Let $i_{m+1}, \ldots, i_n$ denote the complement set of these indices. Furthermore, let U be the L_1 unit ball, i.e., the set $\{\mathbf{y} \in \mathbb{R}^n : |y_1| + \cdots + |y_n| < 1\}$ and denote by ∂U its boundary. The following statements are equivalent:

1. The point in the linear subspace F closest to any $\mathbf{y} \in E$ is equal to $\mathbf{0}$, that is, $\Sigma_{i=1}^n |y_i| < \Sigma_{i=1}^n |y_i - f_i|$, for any $\mathbf{f} \in F$, $\mathbf{f} \neq \mathbf{0}$, and for any $\mathbf{y} \in E$.
2. For any choice of $\mathbf{e} \in \partial U \cap E$, the plane $F + \mathbf{e}$ parallel to F passing through $\mathbf{e}$ does not penetrate U, that is, $\{F + \mathbf{e}\} \cap U = \varnothing$.
3. For any $\mathbf{f} \in F$, $\mathbf{f} \neq \mathbf{0}$, we have $\Sigma_{j=1}^m |f_{i_j}| < \Sigma_{j=m+1}^n |f_{i_j}|$.

Proof. (1) $\Rightarrow$ (2): Suppose (2) is false. Then there exists at least one $\mathbf{e} \in E \cap \partial U$ and a corresponding $\mathbf{f} \in F$, $\mathbf{f} \neq \mathbf{0}$, with $\Sigma_{i=1}^n |f_i + e_i| < 1$. Choose $\mathbf{y} = \mathbf{e}$ to arrive at a contradiction, since $(-\mathbf{f}) \in F$ is closer to $\mathbf{y}$ than $\mathbf{0}$.

(2) $\Rightarrow$ (1): Take any $\mathbf{y} \in E$, $\mathbf{y} \neq \mathbf{0}$, and construct $\mathbf{e} = \mathbf{y}/\Sigma_{i=1}^n |y_i| \in E \cap \partial U$. Suppose there was an $\mathbf{f} \in F$, $\mathbf{f} \neq \mathbf{0}$, with $\Sigma_{i=1}^n |y_i - f_i| < \Sigma_{i=1}^n |y_i|$. This would imply $\Sigma_{i=1}^n |e_i - f_i^*| < 1$, where $f_i^* = f_i / \Sigma_{i=1}^n |y_i|$. This is a contradiction.

The equivalence of (2) and (3) is proved in the Appendix.

Remarks. (1) The generalization of this theorem to the case of a weighted L_1 norm is straightforward. The details are left to the reader.

(2) The theorem says that the exact-fit property, with regard to the indices $i_1, \ldots, i_m$, holds whenever $\Sigma_{j=1}^m |f_{i_j}| < \Sigma_{j=m+1}^n |f_{i_j}|$, for all $\mathbf{f} \in F$, $\mathbf{f} \neq \mathbf{0}$. This inequality is equivalent to $\Sigma_{j=1}^m |f_{i_j}| / \Sigma_{j=1}^n |f_i| < 1/2$, which shows the connection to (1.3) more clearly.

Example. Suppose we are interested in fitting a straight line through the origin using the design $x_1 = (1, 2, 3, 4)^T$.

The leverage indicators (1.2) and (1.3) are now different. When $m = 1$, their largest values are achieved for the index $i_1 = 4$, where they are equal to $16/30 = .53$ and $4/10 = .40$, respectively. It follows that the L_1 fit has the exact-fit property of order 1.

The example concerning the estimation of a location parameter illustrates another property of the diagnostics (1.2) and (1.3). If the design is such that it is close to being balanced, in the sense that the least squares diagnostics are not very different from each other, then one can deduce something about the exactness-of-fit of the L_1 regression. In the location case, we have perfect balance, and thus perfect agreement, between the two diagnostics. In this situation, the condition (3) of Theorem 2.3 can be replaced by a condition that involves the least squares diagnostics. In the next section, we generalize this idea.

3. A SUFFICIENT CONDITION FOR EXACTNESS–OF–FIT OF L_1 ESTIMATORS

We use the same notation as in Theorem 2.3. Let $i_1, \ldots, i_m$ ($m < n/2$) be the indices of the nonzero observations, and let $i_{m+1}, \ldots, i_n$ be the complement of that set of indices. Denote by E, $E^\perp$ the spaces spanned by the coordinate vectors $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_m}$ and its orthogonal complement, respectively.

3.1 The Minimal Angle Between F and E

We want to determine two directions, one parallel to E, the other parallel to F, such that the angle between the two is made as small as possible. This can be formalized as

$$\text{maximize } \mathbf{e}^T\mathbf{f},$$

subject to $\mathbf{e}^T\mathbf{e} = , \mathbf{f}^T\mathbf{f} = 1$, $\mathbf{e} \in E$, and $\mathbf{f} \in F$. A necessary condition on the solution vectors $\mathbf{e}$ and $\mathbf{f}$ is that $\mathbf{f}$ is a multiple of the orthogonal projection $H\mathbf{e}$ of $\mathbf{e}$ onto F, and vice versa. This implies that

$$\mathbf{e}^T\mathbf{f} = (\mathbf{e}^TH\mathbf{e})/\sqrt{\mathbf{e}^THH\mathbf{e}} = \sqrt{\mathbf{e}^TH\mathbf{e}}$$
$$= \sqrt{(P_E\mathbf{e})^TH(P_E\mathbf{e})} = \sqrt{\mathbf{e}^TP_E^THP_E\mathbf{e}},$$

where P_E denotes the orthogonal projection onto E. This shows that the minimal angle has squared cosine value of

$$(\cos \varphi(E, F))^2 = \lambda_{\max}(P_EHP_E), \tag{3.1}$$

the maximal eigenvalue of the symmetric, positive semidefinite matrix P_EHP_E. Equation (3.1) provides another way of expressing the diagnostic (1.2). We have

$$\lambda_{\max}(P_EHP_E) = \max_{\mathbf{f}\in F} \frac{f_{i_1}^2 + \cdots + f_{i_m}^2}{f_1^2 + \cdots + f_n^2}.$$

3.2 A Sufficient Condition for Exactness–of–Fit of Order M

Suppose we want to check whether the L_1 regression has the exact-fit property of order $m = 1$. If at most one of the responses is nonzero, we would have to check that, for any $\mathbf{y}$ of this form, $\Sigma_{i=1}^n|y_i| < \Sigma_{i=1}^n|y_i - f_i|$, $\forall \mathbf{f} \in F$, $\mathbf{f} \neq \mathbf{0}$. Let $\mathbf{y}$ be such a vector of responses and consider the sets $U_\Delta(\mathbf{y}) = \{\mathbf{z} : \Sigma_{i=1}^n|z_i - y_i| < \Delta\}$. We want to prove that, for all such $\mathbf{y}$, $U_\Delta(\mathbf{y})$ touches F for the first time at $\mathbf{0} \in F$ as Δ increases. It seems plausible that a sufficient condition can be based on angles measured in $\mathbb{R}^n$, since the set $U_\Delta(\mathbf{y})$ is a polyhedron. One would, in fact, hope that it suffices to check that the largest angle between $\mathbf{e}_i$ and any vector in the $(n-1)$-dimensional face of the polyhedron is smaller than the minimal angle between $\mathbf{e}_i$ and F.

In the general case, to determine whether the L_1 unit ball U "fits between E and F," we must calculate the maximal angle of opening of the polyhedron U. This is equal to the maximal angle between a vector $\mathbf{e} \in E$ with L_1 norm equal to 1, and a vector $(\mathbf{e} - \mathbf{e}^*)$, where $\mathbf{e}^* \in E^\perp$ also has L_1 norm equal to 1. We will show that, if this maximal angle is smaller than the minimal angle $\varphi(E, F)$, then, for any $\mathbf{y} \in E$, the closest point in F is the origin. The squared cosine value of the angle between $\mathbf{e}$ and $\mathbf{e} - \mathbf{e}^*$ is equal to $\mathbf{e}^T\mathbf{e}/(\mathbf{e}^T\mathbf{e} + (\mathbf{e}^*)^T\mathbf{e}^*)$, so that we obtain the following theorem.

Theorem 3.1. Let E be the space spanned by the coordinate vectors $\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_m}$. If

$$\lambda_{\max}(P_EHP_E) < \frac{1}{m+1}, \tag{3.2}$$

then we can conclude that

$$\|\mathbf{y}\|_1 < \|\mathbf{y} - \mathbf{f}\|_1,$$

for any $\mathbf{y} \in E$ and any $\mathbf{f} \in F$ with $\mathbf{f} \neq \mathbf{0}$.

Proof. First, note that the minimal value of the ratio

$$\mathbf{e}^T\mathbf{e}/(\mathbf{e}^T\mathbf{e} + (\mathbf{e}^*)^T\mathbf{e}^*)$$

is achieved when $\mathbf{e}^*$ has maximal L_2 norm and when $\mathbf{e}$ has minimal L_2 norm. This implies that all nonzero components of $\mathbf{e}$ are equal to $1/m$, whereas only one of the components of $\mathbf{e}^*$ is nonzero (and equal to 1). The minimal value of the ratio is hence equal to $1/(m+1)$.

Now, suppose (3.2) holds and there exists $\mathbf{f} \in F$, $\mathbf{f} \neq \mathbf{0}$, with $\|\mathbf{y} - \mathbf{f}\|_1 \leq \|\mathbf{y}\|_1$. Without loss of generality we may assume $i_j = j$, for $j = 1, \ldots, m$, and $\|\mathbf{y}\|_1 = 1$. Then there exists $\mathbf{z}$ such that $\|\mathbf{z}\|_1 \leq 1$ and $\mathbf{f} = \mathbf{y} + \mathbf{z}$. Without loss of generality we may assume that $\|\mathbf{z}\|_1 = 1$. Hence there exists $\mathbf{u} \in E$, $\mathbf{v} \in E^\perp$, such that $\|\mathbf{u}\|_1 = \|\mathbf{v}\|_1 = 1$ and, for some $\tau \in [0, 1)$, $\mathbf{z} = \tau\mathbf{u} + (1-\tau)\mathbf{v}$. [If $\tau = 1$, then $\mathbf{f} \in E$, contradicting (3.2).] Let $\mathbf{w} = \mathbf{y} + \tau\mathbf{u}$ and $\rho = 1 - \tau$, so $\mathbf{f} = \mathbf{w} - \rho\mathbf{v}$.

First we show $\rho^{-1}\|\mathbf{w}\|_1 \geq 1$, so, in particular, $\mathbf{w} \neq \mathbf{0}$. Suppose not. Write $\mathbf{u} = u_1\mathbf{e}_1 + \cdots + u_m\mathbf{e}_m$. Then

$$1 > \rho^{-1}\sum_{i=1}^m |y_i + \tau u_1| \geq \rho^{-1}\sum_{i=1}^m (|y_i| - \tau|u_1|) = 1.$$

(The equality at the end holds because $\|\mathbf{u}\|_1 = \|\mathbf{y}\|_1 = 1$.) *Contradiction.*

Let $\psi = \|\mathbf{w}\|_1^{-1}$. Let $\mathbf{e} = \psi\mathbf{w}$, $\mathbf{e}^* = \mathbf{v}$, so $\mathbf{e} \in E$ and $\mathbf{e}^* \in E^\perp$, with $\|\mathbf{e}\|_1 = \|\mathbf{e}^*\|_1 = 1$. Let $h(t) = \cos(t\mathbf{e} + (1-t)\mathbf{w}, t(\mathbf{e} - \mathbf{e}^*) + (1-t)(\mathbf{w} - \rho\mathbf{v}))$, $0 \leq t \leq 1$. A routine calculation shows that h is decreasing on $[0, 1]$ providing that $\rho \leq 1/\psi$. However, $\rho^{-1}\|\mathbf{w}\|_1 \geq 1$, so $\psi \leq 1/\rho$. Thus, cos

$\varphi(E, F) \geq \cos(\mathbf{w}, \mathbf{f}) = h(0) \geq h(1) = \cos(\mathbf{e}, \mathbf{e} - \mathbf{e}^*)$, contradicting (3.2).

Remarks. (1) It is again possible to generalize this theorem to the case of a weighted L_1 norm.

(2) A special case of (3.2) may be of particular interest. If all diagonal elements of the least squares hat matrix H are smaller than $1/2$, then the L_1 regression based on the same design has the exact-fit property of order 1.

(3) The sufficient condition (3.2) is not necessary. The simple linear regression with design $x_1 = (-1, 0, 1, 3, 3)^T$ provides a counterexample. The largest diagonal element of H is the first entry, which is equal to .578. And yet the L_1 regression has the exact-fit property of order 1.

4. APPLICATIONS

It is well known that M estimators of regression have a worst case breakdown that is inversely proportional to $(1 + p)$. Since the L_1 estimates are M estimators, this must be true for those estimates. It is easy to check, with the help of Theorem 2.3, that the worst case breakdown point of L_1 regression in any dimension is, in fact, equal to zero. This is achieved in designs where

$$|x_{ij}| \geq \sum_{k \neq j} |x_{ik}|,$$

for some variable x_i and some observation j. But these designs are unrealistic, and one wonders what the breakdown point of L_1 regression would be like in other, more typical, situations.

4.1 Lattice Designs

Consider the case of q explanatory variables, together with a constant and the design supported by the 2^q lattice points $\{-1, 1\}^p$. In this case we have $p = q + 1$. Suppose that each observation is replicated k times, so that we have a total of $n = k \cdot 2^q$ observations. In the case $q = 1$, this leads to the well-known D-optimal design for simple linear regression, which clearly has breakdown point equal to $k/2 - 1 = n/4 - 1$ for k even. An element $\mathbf{f} \in F$ has components $a_0 \pm a_1 + \cdots \pm a_p$, where all combinations of signs are replicated k times and where $a_0, \ldots, a_p$ are arbitrary reals. The diagnostic (1.3) is maximized for $m \leq k$ by choosing all the indices $i_1, \ldots, i_m$ at the same design point

$$\max_{a_0, \ldots, a_p} \frac{m \cdot |a_0 + a_1 + \cdots + a_p|}{k \cdot \sum |a_0 \pm a_1 \pm \cdots + a_p|},$$

where the sum in the denominator is over the 2^q assignments of signs. For $k < m \leq 2k$, the diagnostic is maximized by adding a second cluster of observations at some other design point. When $2k < m \leq 3k$, we start to add points from a third cluster, and so on. The question whether the breakdown point remains equal to $(n/4 - 1)$ for $q > 1$ is answered in the affirmative if we can show that

$$\max_{a_0, \ldots, a_p} \frac{\sum |a_0 + a_1 + a_2 \pm a_3 \cdots \pm a_p|}{\sum |a_0 \pm a_1 \pm \cdots \pm a_p|}$$

is less than or equal to $1/2$, the value of the diagnostic when we put $a_0 = a_1 = a_2 = 1$ and $a_3 = \cdots = a_p = 0$. Note that, in the numerator, we have added exactly $n/4$ components of $\mathbf{f}$. If the ratio is bounded by $1/2$, it must be strictly less than $1/2$ if any one of the terms in the numerator is dropped. By Theorem 2.3, this implies that the breakdown point is $n/4 - 1$. The claimed inequality follows from the inequality

$$|a_0 + a_1 + a_2 + c| \leq |a_0 + a_1 - a_2 + c| + |a_0 - a_1 + a_2 + c| + |a_0 - a_1 - a_2 + c|,$$

for all values of a_0, a_1, a_2, and c. This example shows the existence of high-dimensional designs having large breakdown point.

4.2 The Use of Weights to Enhance the Breakdown Point

Consider the simple linear regression design that takes two observations at $x = 0$, two observations at $x = 1$, and one observations at $x = 3$. The equally weighted L_1 regression has breakdown point $m = 0$ since the maximal value of

$$\frac{|a_0 + 3a_1|}{2|a_0| + 2|a_0 + a_1| + |a_0 + 3a_1|}$$

is larger than $1/2$. If we use a weighted L_1 norm, for example, giving weight $1/3$ to the observation at $x = 3$, the breakdown point increases to $m = 1$. To see why, note that only the diagnostic

$$\max_{a_0, a_1} \frac{|a_0 + 3a_1|/3}{2|a_0| + 2|a_0 + a_1| + |a_0 + 3a_1|/3}$$

needs to concern us, since both of the other points are replicated. This maximum is less than $1/2$.

This example shows that asymmetric designs can, to a limited extent, be corrected for low breakdown values due to outlying design points. To do so, one has to downweight the observations far from the bulk of the data.

4.3 Optimum Designs

Another question of interest is to investigate the upper limits of the breakdown point of L_1 regression. We have seen that the breakdown point depends on the design, and the question becomes one of finding optimum designs having highest breakdown value.

Among univariate designs supported by $\{-1, 1\}$, the one that puts half the observations at each of the two points is optimal and has breakdown value $(n/4 - 1)$. Symmetric designs supported by three points—chosen without loss of generality as $\{-1, 0, 1\}$—are characterized by the fraction p_1 of observations taken at -1 and 1. An elementary calculation shows that the optimal choice is $p_1 = 1/4$ and leads again to a breakdown value of $(n/4 - 1)$. Looking at this limited set of examples, it seems plausible that the breakdown value of $(n/4 - 1)$ is an upper bound, and that the upper bound is reached by many different designs.

Proposition 4.1. Suppose we are fitting a linear regression including a constant. It follows that the L_1 breakdown point of any design with $n = 4k$ observations is bounded by $k - 1 = n/4 - 1$.

Proof. Without loss of generality we may restrict ourselves to the case of univariate regression, since increasing the dimension can only decrease the breakdown point. Suppose our design consists of $x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$, with $n = 4k$. To have breakdown point larger than or equal to $k = n/4$ the following two inequalities must simultaneously be satisfied for all values of a_0 and a_1:

$$|a_0 + a_1x_1| + \cdots + |a_0 + a_1x_k| < |a_0 + a_1x_{k+1}| + \cdots + |a_0 + a_1x_{4k}|$$

and

$$|a_0 + a_1x_{3k+1}| + \cdots + |a_0 + a_1x_{4k}| < |a_0 + a_1x_1| + \cdots + |a_0 + a_1x_{3k}|.$$

Substituting $a_0 = -a_1x_{3k+1}$, and $a_0 = -a_1x_k$, respectively, leads to the inequalities

$$(x_{3k+1} - x_k) + \cdots + (x_{3k+1} - x_1) < |x_{4k} - x_{3k+1}| + \cdots + |x_{k+1} - x_{3k+1}|$$

and

$$(x_{4k} - x_k) + \cdots + (x_{3k+1} - x_k) < |x_{3k} - x_k| + \cdots + |x_1 - x_k|.$$

Substituting $x_j - x_i$ by $(x_j - x_{j-1} + \cdots + x_{i+1} - x_i)$ $(i < j)$, we can rewrite the first of these inequalities as

$$\begin{aligned}&(x_2 - x_1) + 2(x_3 - x_2) + \cdots + k(x_{k+1} - x_k)\\ &+ (k-1)(x_{k+2} - x_{k+1}) + \cdots + (x_{2k} - x_{2k-1})\\ &< (x_{4k} - x_{4k-1}) + 2(x_{4k-1} - x_{4k-2})\\ &+ \cdots + k(x_{3k+1} - x_{3k})\\ &+ (k-1)(x_{3k} - x_{3k-1}) + \cdots + (x_{2k+2} - x_{2k+1}).\end{aligned}$$

Since the other inequality implies just the opposite, we arrive at a contradiction, which proves the proposition.

APPENDIX: PROOF OF THEOREM 2.3

It remains to be shown that, in Theorem 2.3, the third statement is equivalent to the first two.

Proof. (2) $\Rightarrow$ (3): Let $\mathbf{f} \in F$ be such that

$$\sum_{j=1}^{m} |f_{ij}| > \sum_{j=m+1}^{n} |f_{ij}|. \tag{A.1}$$

Choose $\mathbf{e} \in \partial U \cap E$ such that the components $e_{i_1} \neq 0, \ldots, e_{i_m} \neq 0$ and such that the signs of e_{i_j} agree with the signs of f_{i_j}, for $j = 1, \ldots, m$. From (A.1) we get the inequality

$$\sum_{j=1}^{m} f_{ij} \operatorname{sign}(e_{ij}) > \sum_{j=m+1}^{n} |f_{ij}|,$$

which implies that for $\eta < 0$ and $|\eta|$ sufficiently small,

$$\sum_{j=1}^{m} |e_{ij} + \eta f_{ij}| = \sum_{j=1}^{m} |e_{ij}| + \eta \sum_{j=1}^{m} f_{ij} \operatorname{sign}(e_{ij}) < 1 - |\eta| \sum_{j=m+1}^{n} |f_{ij}|.$$

This in turn shows that

$$\sum_{j=1}^{m} |e_{ij} + \eta f_{ij}| < 1$$

for negative η's with sufficiently small value of $|\eta|$. We conclude that the line $\{\mathbf{e} + \eta\mathbf{f} : \eta \in \mathbb{R}\}$ penetrates U.

(3) $\Rightarrow$ (2): Suppose the line $\{\mathbf{e} + \eta\mathbf{f} : \eta \in \mathbb{R}\}$ penetrates U. Either for positive or for negative η's with sufficiently small value of $|\eta|$, this means that

$$\sum_{j=1}^{n} |e_{ij} + \eta f_{ij}| < 1. \tag{A.2}$$

If $e_{i_j} \neq 0$ and if $|\eta|$ is sufficiently small, we have

$$|e_{i_j} + \eta f_{i_j}| = |e_{i_j}| + \eta f_{i_j} \operatorname{sign}(e_{i_j}).$$

Suppose that $e_{i_1}, \ldots, e_{i_l}$ $(l \leq m)$ are the only nonzero components of $\mathbf{e}$. It follows that, for sufficiently small values of $|\eta|$, we have the following chain of equalities

$$\sum_{j=1}^{n} |e_{ij} + \eta f_{ij}| = \sum_{j=1}^{l} |e_{ij}| + \eta \sum_{j=1}^{l} f_{ij} \operatorname{sign}(e_{ij}) + |\eta| \sum_{j=l+1}^{n} |f_{ij}|.$$

The inequality (A.2), therefore, implies

$$\sum_{j=1}^{l} \eta f_{ij} \operatorname{sign}(e_{ij}) < -|\eta| \sum_{j=l+1}^{n} |f_{ij}|,$$

which shows that

$$-\sum_{j=1}^{l} |f_{ij}| \leq \operatorname{sign}(\eta) \sum_{j=1}^{l} f_{ij} \operatorname{sign}(e_{ij}) < -\sum_{j=l+1}^{n} |f_{ij}|.$$

This shows that (A.1) holds.

[*Received April 1990. Revised August 1991.*]

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), *Robust Estimates of Location*, Princeton: Princeton University Press.

Clerc Bérod, A., and Morgenthaler, S. (1989), *A Close Look at the Hat Matrix*, Technical Report, EPF, DMA, Lausanne.

Donoho, D. L., and Huber, P. J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich L. Lehmann*, Belmont, CA: Wadsworth, pp. 157–184.

Ellis, S. P. (1991), personal communication.

Ellis, S. P., and Morgenthaler, S. (1987), "A Geometric Paradigm for Exploratory Data Analysis," in *Proceedings of the 19th Symposium of the Interface*, Alexandria, VA: American Statistical Association, pp. 289–295.

Grünbaum, B. (1967), *Convex Polytopes*, New York: John Wiley.

Hoaglin, D. C., and Welsch, R. E. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17–22.

Mallows, C. L., and Tukey, J. W. (1982), in "Some Recent Advances in Statistics," *An Overview of Techniques of Data Analysis, Emphasizing Its Exploratory Aspects*, New York: Academic Press, pp. 111–172.

Rodgers, J. L., and Nicewander, A. (1988), "Thirteen Ways to Look at the Correlation Coefficient," *The American Statistician*, 42, 59–66.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.