Computational Statistics & Data Analysis 6 (1988) 297-304
North-Holland

A p-subset property of L_1 and regression quantile estimates

Gilbert W. BASSETT, Jr.

Economics Department, University of Illinois, Chicago, Box 4348, Chicago, IL 60680, USA

Received November 1987

Abstract: The L_1 estimate for the p-variable linear model has the well known property of fitting p observations exactly. A less well known property is that certain subsets of p observations will not be fit by the L_1 estimate for any realization of the dependent variables. This property is shown to generalize to other regression quantiles and to the set of all regression quantiles. This identifies subsets of the data which seem to be unimportant. The analog of the property for the location submodel is a situation where one observation, say the first, would not be any quantile for any sample. The implications of the property for other estimates which are based on p-observation subsets are discussed, but the property is considered mainly because it seems strange.

Keywords: L_1 regression, Regression quantiles, Subset estimates.

1. Introduction

The L_1 estimate for the linear model has the well-known property of fitting a subset of the data exactly. If there are p explanatory variables (including the intercept) and if the design matrix has rank p then the L_1 estimate if it is unique will fit p of the observations; p residuals will be zero. More generally, the L_1 solution set is the convex hull of estimates each of which has the exact fit property. This property in the case of the median, the L_1 estimate when $p = 1$, corresponds to the fact that the median when unique is equal to one of the observations and in general the median set is the closed interval (viz. convex hull) with end-points equal to sample observations.

The set of *all* estimates which have the exact fit property has recently attracted some attention. In addition to L_1, there are other interesting statistics which have the p-subset property, there are well-known estimates which can be expressed as functions of the p-subset estimates, and there are estimates which are restricted for computational reasons to the p-subset domain.

The set of all p-subset estimates will be represented in the following way. Let $y \in R^n$ denote the data on the dependent variable and let X denote an n by p design matrix for a model which contains an intercept so that rows are $x_i = (1, z_i)$, $z_i \in R^{p-1}$. To avoid complications suppose rank $X = p$. Let h denote p distinct

0167-9473/88/$3.50 © 1988, Elsevier Science Publishers B.V. (North-Holland)

indices from $1, \ldots, n$. Then a p-subset estimate which exactly fits the observation indices in h is $b(h)$ where $X(h)b(h) = y(h)$ and where $X(h)$ is the submatrix of X with rows x_i, $i \in h$ and $y(h)$ is the subvector of y with elements $i \in h$. Restrict attention to only those such that, rank $X(h) = p$, so that $b(h) = X(h)^{-1}y(h)$ and let H be the set of all such h's. The set of all p-subset estimates given the data y and X is then denoted by S (the y and X hopefully being clear from the context) where

$$S = \left\{ b(h) = X(h)^{-1}y(h) \mid h \in H \right\}.$$

To simplify still further I will restrict attention throughout this paper to those $y \in R^n$ which are in general position. General position means that at most p of the (y_i, z_i) observations will fall on any hyperplane in R^p, or alternatively that $y_i - x_i b(h) = 0$ for only those $i \in h$. With y in general position the $b(h)$, $h \in H$, are all distinct.

The p-subset property for L_1 can now be stated with reference to the set of all p-subset estimates. Let $B^*(0.5)$ denote the set of solutions to: Minimize $\Sigma |d_i|$ subject to $X\beta + d = y$ with d, β unrestricted in sign. If $B^*(0.5)$ is a single point it is in S. If $B^*(0.5)$ is a nonsingleton set it is the convex hull of at most p + 1 points each of which is in S.

Regression quantiles are linear model versions of the usual θth quantiles and, like L_1, which is the $\theta = 0.5$ regression quantile, each θth regression quantile has the p-subset property.[1] The θth regression quantile, $B^*(\theta)$, is defined as the solution to: Minimize $\Sigma\rho(d_i\colon \theta)$ subject to $X\beta + d = y$ with d, β unrestricted in sign, $\theta \in (0, 1)$ and $\rho(d_i\colon \theta) = \theta |d_i|$ or $(1 - \theta)|d_i|$ as $d_i \gtrless 0$. See Koenker and Bassett [9] for definitions and results on regression quantiles; see Ruppert and Carrol [11], Jurečková [8], Welsh [16] and the discussion, and Koenker and Portnoy [10] for generalizations of L-estimates which are based on regression quantiles.

The best-known estimate which can be expressed as a function of the p-subsets is least squares. That is, $\hat{\beta}$ the least squares estimate can be written as

$$\hat{\beta} = \sum_{h \in H} w(h)b(h), \quad w(h) = \frac{|X(h)|^2}{\sum_{h \in H} |X(h)|^2}$$

where $|X(h)|$ is the determinant of $X(h)$. This representation of least squares as a weighted average of p-subsets is due to Subrahmanyan [14]. It has been the motivation for recent proposals of Wu [17] and Shao and Wu [13]. For other recently proposed methods based on p-subsets see the comments by Atkinson and Welsh to Chatterjee and Hadi [4] and see Hawkins et al. [7] and Gray and Ling [5].

The least median of squares estimate proposed by Rousseeuw [11] is not itself in S but, because it is so computationally cumbersome, the recent proposals have

[1] The set of all regression quantiles; $B^*(\theta)$, $\theta \in (0, 1)$, is not equal to the set of all p-subsets and indeed in applications is far smaller. The relationship between the number of elements in S and $B^*(\theta)$, $\theta \in (0, 1)$, is an open question.

used a modified estimate which is defined on the restricted domain of p-subsets rather than on all of R^p; see Hampel et al. [6] and Atkinson [1]. The p-subset property appears here as an aid to computation rather than as an intrinsic property of the least median of squares statistic. [2]

There is a less well known p-subset property which originally stimulated my interest in L_1 and which, given the study of estimates based on p-subsets, may be of some current interest. Unlike the "goes through p points" property, this one has no median counterpart and I know of no good intuition why it should hold. The property may have some consequences for the small sample properties of L_1 and it might give some insights into other methods based on p-subsets, but it will be considered here mainly for its curiousity value. There are counterparts of the property for regression quantiles and these will also be considered. The property is described in the next section using simple examples and pictures. The appendix shows why the property holds.

2. Examples

A trivial property of the median is that any observation can be the median observation. I want to consider how this extends to L_1. We begin by mimicking what is often done in deriving the median's sampling density and consider the subset of the sample space such that a specified p-subset identifies the L_1 estimate.

Let X be fixed and select a given subset of p indices, the ones in h. Now consider all the $y \in R^n$ such that the L_1 estimate goes through the observations $i \in h$. That is, define

$$A(h\colon 0.5) = \{ y \in R^n \mid b(h) \in B^*(0.5) \}.$$

The dependence of A on the design X will be left out of the notation and will be hopefully clear from the context.

For example, consider the data presented in Figure 1 and Table 1. Let $h = (1, 4)$. Then the configuration of $y \in R^4$ given in the table is in $A((1, 4)\colon 0.5)$ because it corresponds to an L_1 estimate which goes through observations one and four. (Of course, this is not the only element of $A((1, 4)\colon 0.25)$. A general representation for the set A is given in the Appendix.)

For the same X matrix consider the $y \in R^4$ which would make the L_1 estimate go through, say, observations 1 and 3. That is, we are thinking about $A((1, 3)\colon 0.5)$. If your intuition is like mine and if you have not thought about it too much, you will probably expect the size of this set to be somehow proportional to

[2] The consequences of the domain restriction on the properties of the least median of squares are not clear. For example, with $p = 1$, the least median of squares is the midpoint of the shortest half of the sample and its breakdown properties are inherited from its being in the shortest half. But it can be shown that the least median of squares estimate when restricted to the p-subset ($p = 1$) domain need not fall in the shortest half of the sample!

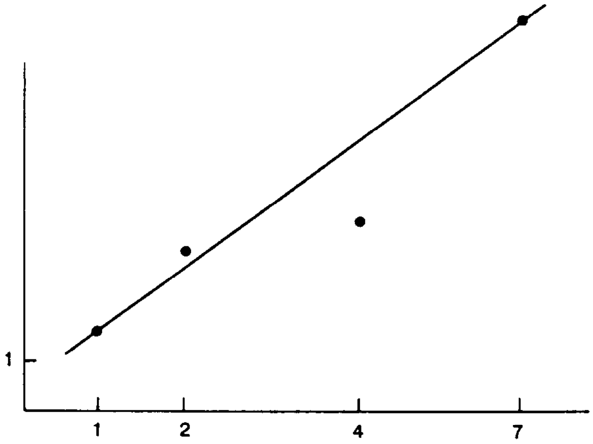

Fig. 1. L_1 estimate.

$|X(h)|$. This intuition is vaguely based on the Subrahmanyan formula for the least squares estimate and on the simple $p = 1$ case where any observation is equally likely to be the median because all the x_i's are equal. The design points one and three with z_i equal to one and four are rather far apart relative to other design points (they have fairly high leverage) and the L_1 estimate, like other M-estimates, might be expected to be somewhat influenced by these observations. Most people who think this way find it somewhat surprising when they find out that $A((1, 3){:}\ 0.5)$ is empty. That is, you cannot make the L_1 estimate go through observations one and three; no configuration of $y \in R^4$ will get the L_1 residuals equal to zero at observations one and three. The possibility that some p-subsets would never yield the L_1 estimate was conjectured by Lester Taylor [15] based on Monte Carlo experiments in the early 1970's. The fact that certain p-subsets would never determine L_1 was demonstrated in Bassett [2].

For the data of Table 1, there are $\binom{4}{2} = 6$ elements of H, but as can be verified by using the formulas given in the Appendix the subsets (1, 2), (1, 3), and (2, 3) will never determine the L_1 estimate. From the representation in the Appendix it

Table 1
Data for Figure 1

Observation	y	X matrix
(1)	1.5	$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 7 \end{bmatrix}$
(2)	3	
(3)	3.5	
(4)	7.5	

$B^*(0.5) = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$

does not seem that this finding depends on n being small or on special features of X. From L_1's view, there exist particular subsets which will never give a good fit, and for reasons which are not intuitively obvious, can be ignored.

The above can be extended in an obvious way to regression quantiles. That is, for a given X, define

$$A(h\colon\ \theta)=\{y\in R^n\,|\,b(h)\in B^*(\theta)\}.$$

(In the $p=1$ case, $A(h\colon\ \theta)$ is just the $y\in R^n$ which makes the hth observation the θth ordinary sample quantile.) Depending on X it turns out that the above $\theta=0.5$ property holds for other θ's as well. That is, there can exist h's such that $A(h\colon\ \theta)=\emptyset$, the empty set.

The fact that $A(h\colon\ \theta)$ can be empty suggests a connection between θ and h which is completely absent from the $p=1$ case. That is, certain p-subsets are more likely, some less likely, and some can never show up as θth quantiles. It is not at all obvious why these properties should hold.

The relation between the p-subsets and θ is highlighted by considering:

$$C(h)=\{\theta\in(0,1)\,|\,A(h\colon\ \theta)\neq\emptyset\}.$$

This is the set of θ's such that the p-subset prescribed by h can possibly (any $y\in R^n$) correspond to a θth regression quantile. In the $p=1$ case this set is just, $C(h)=(0,1)$, for all h. What this says is the trivial fact that there is some $y\in R^n$ which will make any given observation, say the first, any given θth sample quantile.

Consider the design in Table 2 and $h=(1,2)$. This p-subset can be a θth regression quantile for any θ except for those θ in the intervals $(\frac{12}{40},\frac{13}{40})$ and $(\frac{27}{40},\frac{28}{40})$. Why should these "$\theta$-holes" show up given this X and this p-subset? I do not have a good answer. (It can be shown that the holes are always symmetric about 0.5; $\theta\in C(h)$ iff $(1-\theta)\in C(h)$.

For the final example I want to consider the possibility that $C(h)$ itself might be empty. That is, is it possible for there to be a p-subset which would not be

Table 2
A design with regression quantile holes

			Observation
$X=$	1	-10	1
	1	2	2
	1	3	3
	1	5	4
$h=(1,2)$			

Result: There does not exist any $y\in R^4$ such that $b(h)$ is the θth regression quantile where:
$\theta\in(\frac{12}{40},\frac{13}{40})$
$\theta\in(\frac{27}{20},\frac{28}{40})$.

Conversely, for any other $\theta\in(0,1)$ there is some $y\in R^4$ such that $b(h)$ is the θth regression quantile.

Table 3
A design with a p-subset which never corresponds to a θth regression quantile

	Observation
$X=\begin{bmatrix}1&1\\1&2\\1&4\\1&10\end{bmatrix}$	1 2 3 4
$h=(1,3)$	

Result: There does not exist any $y\in R^4$ such that $b(h)$ is a θth regression quantile for any $\theta\in(0,1)$.

associated with any θth quantile for any $y\in R^n$? This would seem to identify a very bad p-subset. The $p=1$ analog of such a situation would be that some observation, say the first, could not be any θth quantile, $\theta\in(0,1)$, for any $y\in R^n$. Such an observation, if one existed, would seemingly be irrelevant and contain no information.

Consider the design given in Table 3. Take $h=(1,3)$. It turns out that, for this h, $C(h)=\emptyset$. That is, no θth regression quantile line, $\theta\in(0,1)$, will go through observations one and three for any $y\in R^n$. This property is not expected and, again, there is no clear intuition for it. From the characterization given in the appendix, it does not seem that this particular example is special and that the property could hold for other X's. From the viewpoint of the regression quantiles, the h where $C(h)=\emptyset$ holds seems to define an irrelevant p subset.

3. Conclusion

It is important to recall that the asymptotics for L_1 and the other θth regression quantiles show that they do in fact correspond to linear model versions of the median and ordinary θth sample quantiles; see Koenker and Bassett [4] and Bassett and Koenker [3]. The examples presented here show that in small samples there exist properties of the L_1 and θth regression quantiles which have no median or ordinary quantile counterparts. While they may have some implications for proposed estimates which are based on p-subsets, the properties have been considered mainly because they seem strange.

Appendix

Let $y\in R^n$ and X be an $n\times p$ matrix with a first column of ones. I will assume throughout that y is in general position and rank $X=p$. A $b(h)\in B^*(\theta)$ is said to be unique if it is the only element of $B^*(\theta)$.

Let sgn(e) be the sign function ($+1$, 0, -1 as $e\gtreqless 0$). Let I^* denote an $(n-p)$ diagonal matrix whose diagonal entries are $+1$ or -1. Let D denote the set of all such matrices. Let $\bar{h}$ be the $(n-p)$ indices not in h.

The first result below gives the conditions such that $b(h)$ is in $B^*(\theta)$. This follows directly from the directional derivative characterization of the solution set for $\rho(b\colon \theta)$; see Koenker and Bassett [8]. This result is then used to describe $A(h\colon \theta)$ and the conditions under which $C(h)$ is empty.

Result 1. *$b(h) \in B^*(\theta)$ if and only if*

$$m_i \in [-\theta, 1-\theta], \quad i = 1, \dots, p,$$

where m_i is the ith component of

$$\sum_{j \in \bar{h}} \left[\theta - 0.5 + 0.5 \operatorname{sgn}(y_j - x_j b(h))\right] x_j X(h)^{-1}. \tag{A1}$$

$b(h)$ is unique when $m_i \in (-\theta, 1, -\theta)$ for all $i \in p$.

Result 1 can be written in matrix notation as

$$-\theta i \leqslant i'[(\theta - 0.5)I + 0.5I^*]\, X(\bar{h}) X(h)^{-1} \leqslant (1-\theta) i \tag{A2}$$

where I^* has diagonal entries equal to $\operatorname{sgn}(y_i - x_i b(h))$, i denotes a conformable vector of ones, and the inequalities are read component-wise. Since y enters (A2) only through the I^* term we have the following.

Result 2. *If* (A2) *does not hold for any $I^* \in D$ then $b(h) \notin B^*(\theta\colon y, X)$ for any $y \in R^n$.*

That this can happen can be verified by using the data in Table 2.

Rewriting equation (A2) to combine the θ terms gives

$$0 \leqslant \theta i' X X(h)^{-1} + i'[0.5I^* - 0.5I]\, X(\bar{h}) X(h)^{-1} \leqslant i. \tag{A3}$$

Since the set $C(h)$ consists of the θ such that some I^* in D makes (A3) valid we have

Result 3. *$C(h)$ is empty if* (A3) *fails for every θ in* (0, 1) *and I^* in D.*

That this can happen can be verified by plugging the values from Table 3 into (A3) and testing over θ, I^*.

References

[1] A.C. Atkinson, Masking unmasked, *Biometrika* 73 (3) (1986) 533–41.

[2] G.W. Bassett, Jr., Some properties of the least absolute error estimator, Ph.D. Thesis, University of Michigan (1973).

[3] G.W. Bassett, Jr., and R. Koenker, Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association* 73 (363) 618–622.

[4] S. Chatterjee and A.S. Hadi, Influential observations, high leverage points and others in linear regression, *Statistical Science* 1 (3) (1986) 379–416.

[5] J.B. Gray and R.F. Ling, *K* clustering as a detection tool for influential subsets in regression, *Technometrics* **26** (4) (1984) 305–329.

[6] F.R. Hampel, E.M. Ronchetti, et al. *Robust Statistics: The Approach Based on Influence Functions* (John Wiley, New York, 1986).

[7] D.M. Hawkins, D. Bradu and G.V. Kass, Location of several outliers in multiple-regression data using elemental sets, *Technometrics* **26** (3) (1984) 197–208.

[8] J. Jurečková, Regression quantiles and trimmed least squares estimator under a general design, *Kybernetika* 20 (1984) 345–357.

[9] R. Koenker and G. Bassett, Regression quantiles, *Econometrica* 46(1) (1978) 33–50.

[10] R. Koenker and S. Portnoy, L-estimation for linear models, *Journal of the American Statistical Association* **82** (1987) 851–1857.

[11] P. Rousseeuw, Least median of squares regression, *Journal of the American Statistical Association* **79** (1984) 388.

[12] D. Ruppert and R. Carroll, Trimmed least squares estimation in the linear model, *Journal of the American Statistical Association* **75** (1980) 93–101.

[13] J. Shao and C.F.J. Wu, Some general theory for the jacknife, *Tech. Report No. 797*, November, University of Wisconsin (1986).

[14] M. Subrahmanyan, A property of simple least squares estimate, *Sankhyā* B **34** (1972) 355–56.

[15] L.D. Taylor, Estimation by minimizing the sum of absolute errors, in: P. Zarembka (Ed.), *Frontiers in Econometrics* (Academic Press, New York, 1974).

[16] A.H. Welsh, The trimmed mean in the linear model, *The Annals of Statistics* **15** (1) (1987).

[17] C.F.J. Wu, Jacknife, bootstrap and other researching methods in regression analysis, *The Annals of Statistics* **14** (4) (1986) 1261–1295.