JOURNAL OF RESEARCH of the National Bureau of Standards
Volume 83, No. 1, January–February 1978

# Norm Approximation Problems and Norm Statistics

**D. R. Shier and C. J. Witzgall**

**Institute for Basic Standards, National Bureau of Standards, Washington, DC 20234**

**(September 15, 1977)**

This paper explores a relation between various approximation problems (arising from fitting linear models to data) and corresponding statistical measures (norm statistics). It is established that for any optimal solution to an approximation problem defined with respect to a norm, the resulting residuals have zero as their norm statistic. This result holds whenever the underlying design matrix has a column of ones. An extension to the case of arbitrary design matrices is also considered.

Key words: Approximation; curve-fitting; $L_p$ problems; least squares; minimization; norm; residuals; statistic.

## 1. Motivation

In a paper[1] discussing alternative criteria to least squares for the fitting of linear models to data, Appa and Smith [1][2] derive certain properties of solutions to $L_1$ *approximation problems* (i.e., curve-fitting problems in which the sum of absolute deviations is minimized). In particular, Property 2 of [1] characterizes the sign pattern of the residuals $e_i = y_i - \hat{b}_0 - \sum_{j=1}^{m} \hat{b}_j x_{ij}$ corresponding to an optimal solution $(\hat{b}_0, \ldots, \hat{b}_m)$ to an $L_1$ approximation problem with independent variables $x_1, \ldots, x_m$ and dependent variable $y$. The result of Appa and Smith states that $|N_1 - N_2| \leq m + 1$, where $N_1$ and $N_2$ denote, respectively, the number of positive residuals and the number of negative residuals corresponding to any optimal $L_1$ solution.

This observation admits of a slight generalization [4]: namely, $|N_1 - N_2| \leq Z$, where $Z$ indicates the number of zero-valued residuals in the given optimal solution. (The assumption employed in [1] to eliminate degeneracy insures that $Z \leq m + 1$, and thus the result of Appa and Smith follows immediately from the above inequality.)

It is straightforward to show that $|N_1 - N_2| \leq Z$ is equivalent to the statement that the residuals in an optimal $L_1$ solution have a *median of zero*. Recall that a median of some set of observations is any value that exceeds at most half the observed numbers, and is exceeded by at most half the observed numbers. From this definition it immediately follows that a median of the numbers $u_1, \ldots, u_n$ (not necessarily distinct) is any value $\xi$ such that

$$N_1(\xi) + Z(\xi) \geq N_2(\xi) \tag{1}$$

and

$$N_2(\xi) + Z(\xi) \geq N_1(\xi), \tag{2}$$

where $N_1(\xi) = \text{card}\{i\colon u_i > \xi\}$, $N_2(\xi) = \text{card}\{i\colon u_i < \xi\}$, and $Z(\xi) = \text{card}\{i\colon u_i = \xi\}$. Hence, zero is a median of the residuals $e_1, \ldots, e_n$ if and only if $N_1 + Z \geq N_2$ and $N_2 + Z \geq N_1$. But the latter two inequalities are clearly equivalent to $|N_1 - N_2| \leq Z$.

The point to be emphasized here is that the sign pattern result[3] $|N_1 - N_2| \leq Z$ is equally a statement about zero being a median of certain residuals. Such a result brings to mind a related statement about the residuals for solutions to $L_2$ (least squares) approximation problems: namely, the mean of the residuals, derived from an optimal $L_2$ solution, is zero. Likewise for $L_\infty$ approximation problems (in which the object is

---

*AMS Subject Classification:* 62J05, 65D10, 90C50.

[1] This paper is also commented upon in the short communication [3] of Gentle et al.

[2] Figures in brackets indicate the literature references at the end of this paper.

[3] It is also easy to show that when $n$ is odd, a slightly stronger result obtains: $|N_1 - N_2| \leq Z - 1$. Indeed, since $N_1 + N_2 + Z = n =$ odd, the parity (even, odd) of $N_1 + N_2$, and thus $N_1 - N_2$, is the same as the parity of $Z - 1$. Accordingly, $|N_1 - N_2| \leq Z$ is equivalent to $|N_1 - N_2| \leq Z - 1$, when $n$ is odd.

to minimize the maximum absolute deviation), it is known that the midrange [6] of the residuals in an optimal $L_\infty$ solution is zero. One wonders whether these facts might not be separate manifestations of a general relationship between approximation problems and corresponding statistical measures. Such a general relationship indeed exists and will be explored in the subsequent sections. The proof of this relationship is extremely simple, simpler than the proofs for the special $L_1$ and $L_2$ cases we have found in the literature. The results of this paper therefore provide both simplification and unification.

## 2. Norm Approximation Problems

Suppose that $n$ sets of observations are available on a single dependent variable $y$ and $m \geq 0$ independent variables $x_1, \ldots, x_m$. Such observations can be arranged in a column vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ and an $n \times m$ matrix $X = (x_{ij})$, where $y_i, x_{i1}, \ldots, x_{im}$ represent observations in the $i$th set. Then the $L_p$ *approximation problem* [2], $1 \leq p \leq \infty$, is that of finding values $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_m$ that minimize

$$\left[\sum_{i=1}^{n} \left|y_i - b_0 - \sum_{j=1}^{m} b_j x_{ij}\right|^p\right]^{1/p} \tag{3}$$

over all $b_0, b_1, \ldots, b_m$. For the case $p = 1$, the problem is that of minimizing the sum of the absolute values of the deviations by choice of parameters $b_0, b_1, \ldots, b_m$. When $p = 2$, the above formulation presents the familiar problem of curve-fitting by least squares. In the case $p = \infty$, the objective function in (3) becomes $\max_i \left|y_i - b_0 - \sum_{j=1}^m b_j x_{ij}\right|$, and we have the linear Chebyshev approximation problem. Every such $L_p$ approximation problem can in fact be formulated [2] as a mathematical programming problem with a convex objective function and linear constraints.

A problem more general than that described by the objective function (3) is the *weighted* $L_p$ *approximation problem*, where $1 \leq p < \infty$. Given nonnegative weights $w_1, \cdots, w_n$, this problem concerns finding parameter values $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_m$ to minimize

$$\left[\sum_{i=1}^{n} w_i \left|y_i - b_0 - \sum_{j=1}^{m} b_j x_{ij}\right|^p\right]^{1/p}. \tag{4}$$

The inclusion of weights in the above may reflect, for example, identical observations as well as differing degrees of confidence (or measures of importance) to be attached to the observed data points.

An even more general approximation problem can be formulated in the present context with respect to any *norm*. A norm $N(\mathbf{x})$ is defined on vectors $\mathbf{x}$ and is assumed to have the following properties [5]:

$$N(\mathbf{x}) > 0 \text{ unless } \mathbf{x} = \mathbf{0},$$

$$N(\lambda\mathbf{x}) = \lambda N(\mathbf{x}), \text{ for } \lambda \geq 0,$$

$$N(\mathbf{x} + \mathbf{y}) \leq N(\mathbf{x}) + N(\mathbf{y}).$$

Let $\mathbf{b} = (b_1, \ldots, b_m)^T$ and form the residuals $\mathbf{e} = \mathbf{y} - b_0\mathbf{1} - X\mathbf{b}$, where $\mathbf{1} = (1, \ldots, 1)^T$. Then the *norm approximation problem* is that of finding $(\hat{b}_0, \hat{\mathbf{b}})$ to minimize

$$N(\mathbf{e}) = N(\mathbf{y} - b_0\mathbf{1} - X\mathbf{b}). \tag{5}$$

The objective function (3) is a special case of (5) with $N(\mathbf{e}) = N(e_1, \ldots, e_n) = [\sum_{i=1}^n |e_i|^p]^{1/p}$, while (4) is also a special case with $N(\mathbf{e}) = [\sum_{i=1}^n w_i|e_i|^p]^{1/p}$.

It can readily be shown that $N(\mathbf{e})$ is a convex function of $(b_0, \mathbf{b})$, and thus the approximation problem described by (5) is well behaved: any local minimum to this problem is also guaranteed to be a global minimum.

## 3. Norm Statistics

The discussion in section 1 indicated that certain statistics (namely, the median, mean and midrange) were useful in describing properties of certain $L_p$ approximation problems. Namely, the residuals of an optimal $L_1$ solution have a median of zero, the residuals of an $L_2$ solution have a mean of zero, and the residuals of an $L_\infty$ solution have a midrange of zero. Moreover, it is well known that these three statistics themselves solve appropriate one-dimensional $L_p$ approximation problems.

For example, the median of a set of values $u_1, \ldots, u_n$ is a value $v$ that minimizes $\sum_{i=1}^n |u_i - v|$ over all possible $v$. That is, a median solves an $L_1$ approximation problem with one parameter. Similarly, the mean of $u_1, \ldots, u_n$ minimizes $\sum_{i=1}^n |u_i - v|^2$, and thus also $[\sum_{i=1}^n |u_i - v|^2]^{1/2}$. Accordingly, the mean solves a one-parameter $L_2$ problem. Finally, the midrange minimizes $\max_i |u_i - v|$, an $L_\infty$ approximation problem, again with one parameter. As suggested by the above examples, we define a *p-statistic* of $u_1, \ldots, u_n$ to be a value $v$ that minimizes

$$\left[\sum_{i=1}^{n} |u_i - v|^p\right]^{1/p},$$

where $1 \le p \le \infty$. This definition follows that given by Rice and White [7], who refer to such a value as an "$L_p$ estimate." In similar fashion, a *weighted p-statistic* of $u_1, \ldots, u_n$ is defined to be a value $v$ that minimizes

$$\left[\sum_{i=1}^{n} w_i |u_i - v|^p\right]^{1/p},$$

where the nonnegative weights $w_i$ are given and $1 \le p < \infty$. Such a concept generalizes, for example, the idea of a weighted mean or a weighted median.

Finally, let $N$ be a norm as defined in section 2. Then a *norm statistic*, or an *N-statistic*, for $\mathbf{u} = (u_1, \ldots, u_n)^T$ is defined to be a value $v$ that minimizes $N(\mathbf{u} - v\,\mathbf{1})$. Clearly, the concept of an $N$-statistic includes as special cases both $p$-statistics and weighted $p$-statistics.

## 4. Norm Approximation Problems and N-Statistics

This section contains the main result relating $N$-statistics and norm approximation problems.

THEOREM: *Let* $(\hat{b}_0, \hat{\mathbf{b}})$ *be an optimal solution to the norm approximation problem (5), and let* $\mathbf{e} = \mathbf{y} - \hat{b}_0 \mathbf{1} - X\hat{\mathbf{b}}$. *Then zero is an* N*-statistic for the residuals* $\mathbf{e}$.

PROOF:
$$\begin{aligned} N(\mathbf{e} - 0\cdot\mathbf{1}) &= N(\mathbf{e}) \\ &= N(\mathbf{y} - \hat{b}_0\mathbf{1} - X\hat{\mathbf{b}}) \\ &\le N(\mathbf{y} - [\hat{b}_0 + v]\,\mathbf{1} - X\hat{\mathbf{b}}) && \text{for all } v \\ &= N(\mathbf{y} - \hat{b}_0\mathbf{1} - X\hat{\mathbf{b}} - v\,\mathbf{1}) && \text{for all } v \\ &= N(\mathbf{e} - v\,\mathbf{1}) && \text{for all } v. \end{aligned}$$

The third line above holds because $(\hat{b}_0, \hat{\mathbf{b}})$ minimizes (5). The resulting inequality $N(\mathbf{e} - 0\cdot\mathbf{1}) \le N(\mathbf{e} - v\,\mathbf{1})$, for all $v$, shows that 0 minimizes $N(\mathbf{e} - v\,\mathbf{1})$, and so 0 is an $N$-statistic for $\mathbf{e}$. This completes the proof.

Notice that in the proof above, we did not at all need the norm properties of $N$. As a matter of fact, $N$ could have been an arbitrary function; in this case, the theorem applies to a global solution (if it exists) to a very general approximation problem.

## 5. Arbitrary Design Matrices

A further generalization of the above theorem is possible for weighted $L_p$ approximation problems. The extension of interest allows an arbitrary "design matrix," where a column of 1's is not necessarily imposed.

In such a problem, the object is to find $\hat{\mathbf{b}} = (\hat{b}_0, \ldots, \hat{b}_m)$ such that

$$\left[\sum_{i=1}^{n} w_i \, |y_i - \sum_{j=0}^{m} b_j x_{ij}|^p\right]^{1/p} \tag{6}$$

is minimized.

EXTENSION: *Let* $\hat{\mathbf{b}}$ *be an optimal solution to (6), and let* $\mathbf{e} = \mathbf{y} - X\hat{\mathbf{b}}$. *Then zero is a weighted* p-*statistic* ($1 \le p < \infty$) *for the values* $\{e_i/x_{i0}: x_{i0} \neq 0, i = 1, \ldots, n\}$ *with weights* $w_i|x_{i0}|^p$.

PROOF:
$$\begin{aligned}
\sum_{i=1}^{n} w_i|e_i - 0 \cdot x_{i0}|^p &= \sum_{i=1}^{n} w_i \, |e_i|^p \\
&= \sum_{i=1}^{n} w_i|y_i - \sum_{j=0}^{m} \hat{b}_j x_{ij}|^p \\
&= \sum_{i=1}^{n} w_i|y_i - \hat{b}_0 x_{i0} - \sum_{j=1}^{m} \hat{b}_j x_{ij}|^p \\
&\le \sum_{i=1}^{n} w_i|y_i - [\hat{b}_0 + v]x_{i0} - \sum_{j=1}^{m} \hat{b}_j x_{ij}|^p \\
&= \sum_{i=1}^{n} w_i|e_i - v x_{i0}|^p.
\end{aligned}$$

Thus, if we define $T = \{i: x_{i0} \neq 0\}$, the above inequality gives

$$\sum_{i \epsilon T} w_i|e_i - 0 \cdot x_{i0}|^p \le \sum_{i \epsilon T} w_i|e_i - v x_{i0}|^p,$$

or

$$\sum_{i \epsilon T} w_i|x_{i0}|^p \, |\frac{e_i}{x_{i0}} - 0|^p \le \sum_{i \epsilon T} w_i|x_{i0}|^p \, |\frac{e_i}{x_{i0}} - v|^p.$$

Upon taking the $p$th root ($1 \le p < \infty$) of both sides, we conclude that zero is a weighted $p$-statistic for $\{e_i/x_{i0}: x_{i0} \neq 0\}$ with weights $w_i|x_{i0}|^p$.

Notice that in the proof above, the choice of the first column, corresponding to the $x_{i0}$'s, is clearly arbitrary. Any column of the design matrix can be used with similar result.

## 6. References


[1] Appa, G., and Smith, C., On $L_1$ and Chebyshev estimation, Mathematical Programming **5** (1973), pp. 73–87.

[2] Barrodale, I., and Roberts, F. D. K., Applications of mathematical programming to $L_p$ approximation, *in* Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, Eds., (Academic Press, New York, 1970), pp. 447–464.

[3] Gentle, J. E., Sposito, V. A., and Kennedy, W. J., On some properties of $L_1$ estimators, Mathematical Programming **12** (1977), pp. 139–140.

[4] Sposito, V.A., Kennedy, W. J., and Gentle, J.E., Useful generalized properties of $L_1$-estimators, to appear in Mathematical Programming.

[5] Householder, A. S., The approximate solution of matrix problems, J. Assoc. Comput. Mach. **5** (1958), pp. 205–243.

[6] Kendall, M. G., and Stuart, A., The Advanced Theory of Statistics, Vol. 1 (Charles Griffin and Co., London, 1963).

[7] Rice, J. R., and White, J. S., Norms for smoothing and estimation, SIAM Review **6** (1964), pp. 243–256.