The Annals of Statistics
1973, Vol. 1, No. 3, 422-432

MULTIPLE ISOTONIC MEDIAN REGRESSION[1]

BY TIM ROBERTSON AND F. T. WRIGHT

University of Iowa

We consider the partial order on the unit square; $s_1 = (x_1, y_1) \ll s_2 = (x_2, y_2)$ if and only if $x_i \leqq y_i$ for $i = 1, 2$, and say that a real-valued function f is isotone if $s_1 \ll s_2$ implies that $f(s_1) \leqq f(s_2)$. Suppose that for each point, s, in the unit square we have a distribution with median $m(s)$ and $m(s)$ is isotone.

In this paper we propose an isotone estimator for m which we denote by $\hat{m}$ and give an algorithm for computing $\hat{m}$. Furthermore we show that if x_{ij} $(j = 1, \cdots, n_i)$ are observations at s_i $(i = 1, \cdots, k)$ then $\hat{m}$ minimizes $D(f) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |f(s_i) - x_{ij}|$ over all isotone functions f. The estimator is also shown to be consistent for m and some rates are given for this convergence. A brief discussion of isotone percentile regression is also given.

1. Introduction and summary. Suppose that for each point, $s = (x, y)$, in the unit square, $[0, 1] \times [0, 1]$ we have a distribution with mean $\mu(s)$ and median $m(s)$. Assume that we make observations on k of these distributions, corresponding to the points $s_1, s_2, \cdots, s_k$, and that the observations on the distribution at the observation point, s_i, are denoted by x_{ij} $(j = 1, 2, \cdots, n_i)$. Consider the problems of estimating $\mu(\cdot)$ and $m(\cdot)$ based on these observations.

Without any additional information about $\mu(\cdot)$ and $m(\cdot)$, estimates might be provided by $\mu^*(s_i) = \bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$ and $m^*(s_i)$, the median of all those observations at s_i. We adopt the convention of averaging the two middle items when the sample size is even. This ensures that when we pool two samples, the median of the pooled sample is between the medians of the individual samples (cf. Robertson and Waltman (1968)). This property of sample medians is used repeatedly in the proofs in Section 2 and we shall refer to it there as "apom" (averaging property of medians).

In certain problems we may have reason to believe that $\mu(\cdot)$ or $m(\cdot)$ has an isotone property. More specifically, we may believe that the value of the function under consideration at a point does not exceed its value at any point to the upper right of the original point. Technically speaking, introduce the partial ordering $\ll$ on $[0, 1] \times [0, 1]$ defined by: $s = (x, y) \ll s' = (x', y')$ if and only if $x \leqq x'$ and $y \leqq y'$. A function $\mu(\cdot)$ on $[0, 1] \times [0, 1]$ is said to be isotone with respect to this partial ordering if $\mu(s) \leqq \mu(s')$ whenever $s \ll s'$. We want to consider the case when the raw estimates do not enjoy this property and it is felt that some smoothing is required. (An analogous problem on the line is discussed by Cryer, Robertson, Wright and Casady (1972).)

Received October 1971; revised September 1972.

[1] This research was sponsored by the National Science Foundation under Grant Number GP-21073.

AMS 1970 *subject classifications*. Primary 62G05, 62J05; Secondary 60F15.

Key words and phrases. Isotone regression, median, percentile, consistency, σ-lattice.

Such a situation occurs in testing problems. Suppose x represents a child's grade level in school, y denotes his intelligence quotient, and $\mu(x, y)$ and $m(x, y)$ represent the average and median score, respectively, earned on a certain achievement test by those individuals in category (x, y). It seems reasonable to assume that $\mu(\cdot)$ and $m(\cdot)$ are isotone with respect to the partial ordering $\ll$. Even with relatively large sample sizes the raw estimates may not enjoy this property and in the past relatively *ad hoc* methods have been used for smoothing these estimates.

An estimate $\hat{\mu}(\cdot)$ of $\mu(\cdot)$ which is isotone, has been proposed and studied (cf. Brunk (1970) and Hanson, Pledger and Wright (1973)). We wish to investigate an analogous estimate $\hat{m}(\cdot)$ of $m(\cdot)$. Before we give the formulas for $\hat{\mu}(\cdot)$ and $\hat{m}(\cdot)$ we need to introduce some terminology and notation. Since we are only concerned with the value of our estimates at the observation points we will restrict our attention to $S = \{s_1, s_2, \cdots, s_k\}$. The domain of definition of those estimates may be extended to $[0, 1] \times [0, 1]$ in several ways, depending on what properties the investigator wishes his estimates to enjoy. A review of the results in Brunk (1965) or Robertson (1967) might be of value to the reader at this point.

Let $\mathscr{L}$ be the collection of all upper layers in S (i.e., $\mathscr{L}$ is the collection of all subsets L of S having the property that $s_i \ll s_j$ together with $s_i \in L$ imply that $s_j \in L$). In order to simplify some of the notation used in the remainder of the paper we will henceforth only use the symbol L, with or without subscripts, primes, etc., to denote upper layers. It is easy to see that $\mathscr{L}$ is a complete σ-lattice. (Relationships between partial orderings and complete σ-lattices are discussed in Robertson (1967).) Let $R(\mathscr{L})$ be the collection of all $\mathscr{L}$-measurable functions on S (i.e., $R(\mathscr{L}) = \{g(\cdot); [g > a] \in \mathscr{L}$ for all real $a\}$). Define the decrement functions $D(\cdot)$ and $D'(\cdot)$ on the collection of all functions on S by:

$$D'(g) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [g(s_i) - x_{ij}]^2$$

and

$$D(g) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |g(s_i) - x_{ij}| .$$

If g is an estimate of $\mu(\cdot)$ or $m(\cdot)$ then D and D' measure how close g is to what we actually observed. For each nonempty subset S' of S let $A(S')$ denote the arithmetic average of all those observations corresponding to observation points in S' and let $M(S')$ be their median.

For each $L \in \mathscr{L}$ let $U(L)$ denote the collection of all proper sub-upper layers of L. An isotone estimate for μ is given by:

$$\hat{\mu}(s_j) = \max_{L \ni s_j} \min_{L' \in U(L)} A(L - L') . \tag{1.1}$$

It is shown in Hanson, Pledger and Wright (1973), or it follows from the fact that $\hat{\mu}$ can be represented as a conditional expectation given a σ-lattice, that $\hat{\mu}$ provides the nearest member of $R(\mathscr{L})$ to our observations in the sense that $\hat{\mu} \in R(\mathscr{L})$ and

$$D'(\hat{\mu}) \leqq D'(g) \tag{1.2}$$

for all $g \in R(\mathscr{L})$. In addition, it follows from properties of conditional expectation operators that

$$\sum_{i=1}^{k} (\hat{\mu}(s_i) - \mu(s_i))^2 n_i \leqq \sum_{i=1}^{k} (\bar{x}_i - \mu(s_i))^2 n_i \,. \tag{1.3}$$

In other words, $\hat{\mu}$ is closer to the unknown μ than the "raw" estimate provided by the unsmoothed averages $\bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$. Furthermore, $\hat{\mu}$ is a consistent estimator of μ and convergence rates have been studied (cf. Hanson, Pledger and Wright (1973)).

Motivated by the above considerations we propose the estimator $\hat{m}$ of m given by:

$$\hat{m}(s_j) = \max_{L \ni s_j} \min_{L' \in U(L)} M(L - L') \,. \tag{1.4}$$

In Section 2 we prove several theorems about $\hat{m}$ including the result which says that $\hat{m}$ provides the nearest point of $R(\mathscr{L})$ to our observations in the sense of $D(\cdot)$. More specifically, $\hat{m} \in R(\mathscr{L})$ and

$$D(\hat{m}) \leqq D(g) \tag{1.5}$$

for all $g \in R(\mathscr{L})$. The key to proving this result is a representation theorem which also simplifies computation of $\hat{m}$ considerably (a similar result holds for $\hat{\mu}$).

Note that if the distribution at s is a bilateral exponential distribution (i.e., the distribution is absolutely continuous with density function $f(x; s) = \frac{1}{2}e^{-|x-m(s)|}$) then (1.5) says that $\hat{m}$ provides the maximum likelihood estimate of m.

Section 3 contains some consistency results for the estimator $\hat{m}$ and gives rates of convergence for $P[|\hat{m}(t_k) - m(t_k)| > \varepsilon]$ to zero. In Section 4 we briefly discuss percentile regression and in Section 5 we indicate some problems for which we have not found a solution.

2. The estimate. In this section we explore certain properties of the estimate, $\hat{m}$, given by (1.4).

THEOREM 2.1. *If $s_i \ll s_j$ then $\hat{m}(s_i) \leqq \hat{m}(s_j)$* (i.e., $\hat{m} \in R(\mathscr{L})$).

PROOF. Consider the definitions of $\hat{m}(s_j)$ and $\hat{m}(s_i)$, note that since $s_i \ll s_j$ we know that $\{L;\, L \in \mathscr{L},\, s_i \in L\} \subset \{L;\, L \in \mathscr{L},\, s_j \in L\}$ and the desired result follows.

THEOREM 2.2. *If $M_i \leqq M_j$ whenever $s_1 \ll s_j$ then the value of $\hat{m}(\cdot)$ at each observation point, s_i, is given by M_i.*

PROOF. Let $L(s_i)$ be the smallest upper layer containing s_i. By hypothesis, $M_i \leqq M_j$ for all j such that $s_j \in L(s_i)$. It then follows, using "apom," that $M_i \leqq M(L(s_i) - L')$ for all L' in $U(L(s_i))$. Thus

$$M_i \leqq \min_{L' \in U(L(s_i))} M(L(s_i) - L')$$

and it follows that $M_i \leqq \hat{m}(s_i)$ from the definition of $\hat{m}(s_i)$. On the other hand, suppose $s_i \in L$ and let $L_0' = \{s_j = (x_j, y_j);\, x_j > x_i \text{ or } y_j > y_i\}$. Then, $L_0' \in U(L)$ and $M_i \geqq M_j$ for all j with the property that $s_j \in L - L_0'$. It now follows that

$$M_i \geqq M(L - L_0') \geqq \min_{L' \in U(L)} M(L - L') \,.$$

The desired conclusion obtains since L was an arbitrary upper layer containing s_i.

If one uses the decrement function, $D(\cdot)$, as a tool for measuring the "goodness" of estimators $m(\cdot)$ then the property given for $\hat{m}(\cdot)$ in Theorem 2.2 is a desirable one. This is so because if $h(s_i) = M_i$ is $\mathscr{L}$-measurable then $h(\cdot)$ provides a best estimate, in the above sense.

Now let $L_1 = S$ and choose L_2 in $U(L_1)$ so that $\min_{L' \in U(L_1)} M(L_1 - L') = M(L_1 - L_2)$.

REMARK 2.3. If L_2' is another member of $U(L_1)$ with the property that $\min_{L' \in U(L_1)} M(L_1 - L') = M(L_1 - L_2')$ then $L_2 \cap L_2'$ also has that property.

PROOF. First observe that $L_2 \cap L_2'$ is also a member of $U(L_1)$. Now if $L_1 - (L_2 \cup L_2') = \varnothing$ then $L_1 - (L_2 \cap L_2') = (L_1 - L_2) + (L_1 - L_2')$ and the desired result follows, using "apom" and the fact that $M(L_1 - L_2) = M(L_1 - L_2')$. On the other hand, suppose $L_1 - (L_2 \cup L_2')$ is not empty. Then $L_2 \cup L_2' \in U(L_1)$ so that using the hypothesis that L_2 provides a minimum, $L_1 - L_2 = (L_1 - (L_2 \cup L_2')) + (L_2' - L_2)$ and "apom" we get

$$M(L_2' - L_2) \leqq M(L_1 - L_2) \leqq M(L_1 - (L_2 \cup L_2')) . \tag{2.1}$$

Next, using the fact that L_2' provides a minimum, $L_1 - (L_2 \cap L_2') = (L_1 - L_2') + (L_2' - L_2)$ and "apom" we get

$$M(L_2' - L_2) \geqq M(L_1 - (L_2 \cap L_2')) \geqq M(L_1 - L_2') . \tag{2.2}$$

Combining (2.1) and (2.2) we can infer that

$$M(L_1 - (L_2 \cap L_2')) \leqq M(L_1 - L_2) ,$$

from which the desired result follows.

Using the above result and the facts that S is finite and $\mathscr{L}$ is closed under intersections we know that there is a smallest member of $U(L_1)$ with the property prescribed for L_2. Assume that L_2 is chosen to be this smallest member of $U(L_1)$.

Now choose $L_3 \in U(L_2)$ "as small as possible" so that

$$\min_{L' \in U(L_2)} M(L_2 - L') = M(L_2 - L_3) .$$

Continuing in this fashion we obtain a sequence $L_1, L_2, \cdots, L_H$ of upper layers such that $L_1 \supset L_2 \supset \cdots \supset L_H$ and

$$\min_{L' \in L_i} M(L_i - L') = M(L_i - L_{i+1}) .$$

($L_{H+1} = \varnothing$). Our construction procedure must terminate since S is finite and $L_{i+1} \in U(L_i)$ reqires that $L_i - L_{i+1}$ is not empty.

THEOREM 2.4. *The value of $\hat{m}(\cdot)$ at any observation point in $L_i - L_{i+1}$ is given by $M(L_i - ML_{i+1})$.*

PROOF. Suppose $s_j \in L_{i+1}$. Then

$$\begin{aligned} \hat{m}(s_j) &= \max_{L \ni s_j} \min_{L' \in U(L)} M(L - L') \\ &\geqq \min_{L' \in U(L_i)} M(L_i - L') = M(L_i - L_{i+1}) . \end{aligned} \tag{2.3}$$

Next assume that L is an arbitrary upper layer containing s_j. Then for $h \leqq i$,

$L_h - L_{h+1} = (L_h - (L \cup L_{h+1})) + ((L \cap L_h) - L_{h+1})$. Considering separately the cases when $L_h - (L \cup L_{h+1})$ is empty and otherwise we obtain

$$M((L \cap L_h) - L_{h+1}) \leqq M(L_h - L_{h+1}) \tag{2.4}$$

$h = 1, 2, \cdots, i$. This result, together with the fact that

$$L - L_{i+1} = \sum_{h=1}^{i} ((L \cap L_h) - L_{h+1})$$

gives:

$$M(L - L_{i+1}) \leqq \max_{1 \leqq h \leqq i} M((L \cap L_h) - L_{h+1}) . \tag{2.5}$$

Now $L_{i-1} - L_{i+1} = (L_{i-1} - L_i) + (L_i - L_{i+1})$ so

$$M(L_{i-1} - L_i) \leqq M(L_{i-1} - L_{i+1}) \leqq M(L_i - L_{i+1}) . \tag{2.6}$$

$L_{i-2} - L_{i+1} = (L_{i-2} - L_{i-1}) + (L_{i-1} - L_{i+1})$ so $M(L_{i-2} - L_{i-1}) \leqq M(L_{i-2} - L_{i+1}) \leqq M(L_{i-1} - L_{i+1})$. Combining this with (2.6) we get

$$M(L_{i-2} - L_{i-1}) \leqq M(L_i - L_{i+1}) . \tag{2.7}$$

Continuing in this fashion we obtain

$$M(L_h - L_{h+1}) \leqq M(L_i - L_{i+1}) \tag{2.8}$$

for $h = 1, 2, \cdots, i$. Combining (2.4) and (2.8) we obtain $M((L \cap L_h) - L_{h+1}) \leqq M(L_i - L_{i+1})$ for $h = 1, 2, \cdots, i$ which together with (2.5) gives $M(L - L_{i+1}) \leqq M(L_i - L_{i+1})$. Now $L \cap L_{i+1} \in U(L)$ since $s_j \in L - L_{i+1}$ so $\min_{L' \in U(L)} M(L - L') \leqq M(L_i - L_{i+1})$. Since L was arbitrary it follows that

$$\hat{m}(s_j) \leqq M(L_i - L_{i+1}) . \tag{2.9}$$

(2.3) and (2.9) give the desired result.

Since this argument depends only on the averaging property of medians an analogous result can be obtained for $\hat{\mu}$. Theorem 2.4 provides for us the key to showing that $\hat{m}(\cdot)$ has the desired minimizing property. More importantly, perhaps, this theorem provides an algorithm for computing $\hat{m}(\cdot)$. Suppose we have n^2 observation points lying in an n by n grid. The number of upper layers is $m = \binom{2n}{n}$. Thus, if we intended to use the definition for $\hat{m}(\cdot)$ in our computations then to find the value of $\hat{m}(\cdot)$ at an s_j somewhere in the middle we would be required to compute approximately $m/2$ minimums each of which involves computing $m/2$ medians. This task may be impossible even for a modern computer when n is 15 or larger. Theorem 2.4 clearly provides a better method for computing $\hat{m}(\cdot)$. We are currently in the process of studying the estimate $\hat{m}(\cdot)$ in a problem which arises in estimating achievement levels for given IQ and grade levels.

COROLLARY 2.5. *Other representations for the value of* $\hat{m}(\cdot)$ *at observation points are given by*:

$$\begin{aligned} \hat{m}(s_j) &= \max_{L \ni s_j} \min_{L' \ni s_j} M(L - L') \\ &= \min_{L' \ni s_j} \max_{L \ni s_j} M(L - L') . \end{aligned} \tag{2.10}$$

PROOF. To prove either of these let $m^*(s_j)$ be the appropriate function on the right-hand side of (2.10) and argue that if $s_j \in L_i - L_{i+1}$ then $m^*(s_j) = M(L_i - L_{i+1})$ (see Theorem 2.4).

We are now in a position to show that $\hat{m}$ solves our extreme value problem. We first argue that there is a solution.

LEMMA 2.6. *There is a minimizing point $g(\cdot)$ in $R(\mathscr{L})$.*

PROOF. Let $R(\mathscr{L})^*$ be that subset of $R(\mathscr{L})$ whose points have values bounded below by $\min(M_1, M_2, \cdots, M_k)$ and above by $\max(M_1, M_2, \cdots, M_k)$. Using the fact that if we move the value of $g(\cdot)$ at s_i closer to M_i we do not increase the value of $D(\cdot)$, we can see that for each point $g(\cdot)$ in $R(\mathscr{L}) - R(\mathscr{L})^*$ there is a point $\hat{g}(\cdot)$ in $R(\mathscr{L})^*$ such that $D(g) \geqq D(\hat{g})$. Hence we may restrict our attention to $R(\mathscr{L})^*$. But this set is a closed and bounded subset of Euclidean k-space and $D(\cdot)$ is continuous so that it follows that a minimizing point exists.

THEOREM 2.7. *$\hat{m}$ minimizes $D(\cdot)$ in $R(\mathscr{L})$* (i.e., $\sum_{i=1}^{k} \sum_{j=1}^{n_i} |\hat{m}(s_i) - x_{ij}| \leqq \sum_{i=1}^{k} \sum_{j=1}^{n_i} |g(s_i) - x_{ij}|$ *for all g in $R(\mathscr{L})$*).

PROOF. The argument uses the representation for $\hat{m}$ given by Theorem 2.4. We consider separately the case where $H = 1$ (i.e., $\hat{m}(s_j) = M(L_1) = \min_{L' \in U(L_1)} M(L_1 - L')$). Suppose that this is the case. Assume that $g(\cdot)$ is any minimizing function, and relabeling, if necessary, suppose that $g(s_1) \leqq g(s_2) \leqq \cdots \leqq g(s_k)$. Choose $i(j)$ to be the jth integer such that $g(s_{i(j)}) < g(s_{i(j)+1})$ ($j = 1, 2, \cdots, p-1$, $i(p) = k$). If g is constant on S, we could easily complete the argument. The argument when $H = 1$ is now divided into three cases depending on the relation between the values of $g(\cdot)$ and $M(L_1)$. We give only one; the other cases are similar.

Assume that $M(L_1) \leqq g(s_1) \leqq g(s_2) \leqq \cdots \leqq g(s_k)$. We argue that there is a constant minimizing function. If $i(1) = k$ we have the desired result. Otherwise, let $L_2' = \{s_{i(p-1)+1}, s_{i(p-1)+2}, \cdots, s_k\}$. Since $g \in R(\mathscr{L})$, $L_2' \in \mathscr{L}$ so that $L_2' \in U(L_1)$. Thus $M(L_1) \leqq M(L_1 - L_2')$ and using "apom" we obtain

$$M(L_2') \leqq M(L_1) \leqq M(L_1 - L_2') .$$

Since g is constant on L_2' we can move its value closer to $M(L_2')$, the median of the observations in L_2', without increasing the value of $D(\cdot)$ at g or destroying the isotone property of g. Move it down to $g(s_{i(p-1)})$. Continuing in this fashion we obtain a constant minimizing function.

Arguing in a similar fashion we can obtain a constant minimizing function in the case that $g(s_1) \leqq g(s_2) \leqq \cdots \leqq g(s_k) \leqq M(L_1)$ and the case that $g(s_1) \leqq g(s_2) \leqq \cdots \leqq g(s_{i(q)}) \leqq M(L_1) < g(s_{i(q)+1}) = \cdots \leqq g(s_k)$. Hence in any eventuality there is a constant minimizing function, $\hat{g}$. It follows from well-known properties of $D(\cdot)$ that $D(\hat{m}) \leqq D(\hat{g})$.

Now consider the general case. The argument proceeds by induction on the number k of distinct observation points. The desired result is clearly true when

$k = 1$. Suppose the result is true for any number of distinct observation points less than or equal to k and that we have $k + 1$ distinct observation points $s_1, s_2, \cdots, s_{k+1}$. Relabeling, if necessary, assume that $\{s_1, s_2, \cdots, s_p\} = L_1 - L_2$ and $\{s_{p+1}, s_{p+2}, \cdots, s_{k+1}\} = L_2$. If $p = k + 1$ the result follows from our first case. By supposing that $s_1, s_2, \cdots, s_p$ were the only observation points it follows from our first case that

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} |\hat{m}(s_i) - x_{ij}| \leqq \sum_{i=1}^{p} \sum_{j=1}^{n_i} |g(s_i) - x_{ij}| \tag{2.11}$$

for all $g \in R(\mathscr{L})$. Then assuming that $s_{p+1}, s_{p+2}, \cdots, s_{k+1}$ are the only observation points, recalling the procedure for constructing $L_1, L_2, \cdots, L_H$ and using the induction hypothesis it follows that

$$\sum_{i=p+1}^{k+1} \sum_{j=1}^{n_i} |\hat{m}(s_i) - x_{ij}| \leqq \sum_{i=p+1}^{k+1} \sum_{j=1}^{n_i} |g(s_i) - x_{ij}| \tag{2.12}$$

for all g in $R(\mathscr{L})$. The desired result follows by combining (2.11) and (2.12).

3. Consistency results. In this section we prove a theorem and two corollaries which show that $\hat{m}$ is consistent for m and which give rates for this convergence. It is convenient for us to think of the observation points as a sequence $\{t_j : j = 1, 2, \cdots\}$ whose elements are not necessarily distinct. We assume that associated with each t_j is a random variable Y_j with the following properties.

(3.1) The random variables Y_j for $j = 1, 2, \cdots$ are independent.

(3.2) The distribution function and median of Y_j are F_j and $m(t_j)$, respectively.

(3.3) For each $\varepsilon > 0$, $\inf_j F_j(m(t_j) + \varepsilon) - \frac{1}{2} > 0$ and $\frac{1}{2} - \sup_j F_j(m(t_j) - \varepsilon) > 0$.

It is also more appropriate in this section to state our results in terms of functions on $[0, 1] \times [0, 1]$. Thus we now adopt the following terminology: an upper layer is any subset L of $[0, 1] \times [0, 1]$ with the property that $s' \in L$ whenever $s \ll s'$ and $s \in L$. Thus, for a fixed value of n the proposed estimator at an observation point t_j with $j \leqq n$ can be written

$$\hat{m}_n(t_j) = \max_{L \ni t_j} \min_{L' \ni t_j} M_n(L - L') \tag{3.4}$$

where $M_n(L - L')$ is the median of all those observations at observation points t_i with $i \leqq n$ and $t_i \in L - L'$ (cf. 2.10). We again agree to use the symbol L when referring to an upper layer and agree to average the two middle items when the sample size is even. (This latter convention is not necessary for our consistency results.) The value of $\hat{m}_n(\cdot)$ at points other than $t_1, t_2, \cdots, t_n$ is arbitrary except that $\hat{m}_n(\cdot)$ should be isotone with respect to the partial ordering given in Section 1.

For any subset J of $[0, 1] \times [0, 1]$ we define $N_n(J)$ to be the number of observation points among the first n observation points which are in J. Also for each positive integer η we let $I_{ij} = I_{ij}(\eta) = [(i - 1)/\eta, i/\eta] \times [(j - 1)/\eta, j/\eta]$ for $i, j = 1, 2, \cdots, \eta$.

THEOREM 3.1. *Assume* $m(\cdot)$ *is continuous on* $[0, 1] \times [0, 1]$; *for each non-degenerate rectangle* $J \subset [0, 1] \times [0, 1]$

$$\liminf_{n\to\infty} N_n(J)/n > 0$$

and that

(3.5) *there exists a constant* M *such that for any positive integer* η *there exists an* $n(\eta)$ *such that* $N_n(I_{ij}) \leqq M\eta^{-2}n$ *for all* $n \geqq n(\eta)$ *and* $i, j \leqq \eta$.

Then for each $t_k \in (0, 1) \times (0, 1)$ *and each* $\varepsilon > 0$ *there exist positive constants* C *and* $\zeta < 1$ *such that*

$$P[|\hat{m}_n(t_k) - m(t_k)| > \varepsilon] \leqq C\zeta^n . \tag{3.6}$$

Before the proof of the theorem is given we note that (3.5) relates the proportion of the observations taken in the special squares I_{ij} to the area of I_{ij}. This condition was used in Hanson, Pledger and Wright (1973) to obtain consistency results for $\hat{\mu}$.

PROOF. The first part of the proof of Theorem 3.1 is similar to the initial part of the proof of Theorem 3.3 in Cryer, Robertson, Wright and Casady (1972) and the last part is just like the last part of the proof of Theorem 6 of Hanson, Pledger and Wright (1973). We shall omit much of the detail. Let $\varepsilon > 0$. We will show that there exist positive constants C and $\zeta < 1$ such that $P[\hat{m}_n(t_k) - m(t_k) > \varepsilon] \leqq C\zeta^n$. The other half of the proof is similar. Choose s such that $t_k \ll s \ll (1, 1)$ and $m(s) - m(t_k) < \varepsilon/2$ and define $L_0 = \{s' : s' \ll s\}^c$. Now using (2.10) it follows that:

$$\begin{aligned} \hat{m}_n(t_k) - m(t_k) &\leqq \max_{L \ni t_k} M_n(L - L_0) - m(t_k) \\ &\leqq \max_{L \ni t_k} M_n(L - L_0) - m(s) + \varepsilon/2 . \end{aligned} \tag{3.7}$$

However, for each L which contains t_k the median of $\{Y_j : j \leqq n \text{ and } t_j \in L - L_0\} - m(s)$ is the median of $\{Y_j - m(s) : j \leqq n \text{ and } t_j \in L - L_0\}$ which is bounded above by the median of $\{Y_j - m(t_j) : j \leqq n \text{ and } t_j \in L - L_0\}$. Define $Z_j = Y_j - m(t_j)$; $M_n^*(L - L_0)$ to be the median of $\{Z_j : j \leqq n \text{ and } t_j \in L - L_0\}$; and $W_j = I_{(-\infty, \varepsilon/2]}(Z_j)$. For notational convenience we set $A(L, n) = \{j : j \leqq n \text{ and } t_j \in L - L_0\}$ and we observe that if $M_n^*(L - L_0) > \varepsilon/2$ then $\sum_{j \in A(L,n)} W_j \leqq N_n(L - L_0)/2$ which implies that

$$[N_n(L - L_0)]^{-1} \sum_{j \in A(L,n)} (W_j - EW_j) + \inf_j F_j(m(t_j) + \varepsilon/2) - \tfrac{1}{2} \leqq 0 .$$

We now set $\delta = \inf_j F_j(m(t_j) + \varepsilon/2) - \frac{1}{2}$. Assumption (3.3) ensures that $\delta > 0$. Hence, using (3.7), if $\hat{m}_n(t_k) - m(t_k) > \varepsilon$, then

$$\max_{L \ni t_k} [N_n(L - L_0)]^{-1} \sum_{j \in A(L,n)} (EW_j - W_j) \geqq \delta . \tag{3.8}$$

The random variables $EW_j - W_j$ for $j = 1, 2, \cdots$ are independent and centered at their means. Since they are uniformly bounded there exist constants D and λ such that

$$P\{|W_j - EW_j| \geqq y\} \leqq De^{-\lambda y}$$

for all j and all $y > 0$. The remainder of the proof is just like the proof of Theorem 6 of Hanson, Pledger and Wright (1973) beginning at (43).

COROLLARY 3.2. *Let* $0 < a < b < 1$ *and assume the hypotheses of Theorem* 3.1. *For any* $\varepsilon > 0$ *there exist positive constants* C *and* $\zeta < 1$ *such that*

$$P[\sup_{t \in [a,b]\times[a,b]} |\hat{m}_n(t) - m(t)| > \varepsilon] \leqq C\zeta^n . \tag{3.9}$$

Also

$$P[\lim_{n\to\infty} \sup_{t \in [a,b]\times[a,b]} |\hat{m}_n(t) - m(t)| = 0] = 1 . \tag{3.10}$$

PROOF. Conclusion (3.9) follows from (3.6) just as (27) follows from (42) of Hanson, Pledger and Wright (1973). Conclusion (3.10) follows from (3.9) by the Borel–Cantelli Theorem.

COROLLARY 3.3. *Assume the hypothesis of Theorem* 3.1. *Then*

$$P[\lim_{n\to\infty} \hat{m}_n(t) = m(t) \text{ for all } t \in (0, 1) \times (0, 1)] = 1 . \tag{3.11}$$

PROOF. Since a countable intersection of sets with probability one is again a set with probability one, (3.11) follows from (3.9).

4. Percentile regression. Fix p with $0 < p < 1$ and define the $100p$-percentile of a distribution function F (the right continuous version) to be

$$\xi_F = \min \{t: F(t) \geqq p\} .$$

REMARK 4.1. If $0 < \alpha < 1$, F and G are distribution functions and $H = \alpha F + (1 - \alpha)G$, then ξ_H is between ξ_F and ξ_G.

PROOF. We assume that $\xi_F \leqq \xi_G$. We show that $\xi_H \leqq \xi_G$ by arguing that $H(\xi_G) \geqq p$ and the other half of the proof is similar. The desired conclusion follows from

$$H(\xi_G) = \alpha F(\xi_G) + (1 - \alpha)G(\xi_G) \geqq \alpha F(\xi_F) + (1 - \alpha)G(\xi_G) \geqq p .$$

We agree that the $100p$ sample percentile is the $100p$-percentile of the empirical distribution function. From Remark 4.1 and the fact that the empirical distribution function of the pooled sample can be expressed as a convex combination of the individual empiricals, we see that the sample percentile of a pooled sample is between the two percentiles of the individual samples.

Suppose for each point, s, in the unit square that $p(s)$ is the $100p$-percentile of the distribution associated with s and suppose that $p(s)$ is isotone with respect to the partial order given in Section 1. As in Section 1, let x_{ij} $(j = 1, \cdots, n_i)$ be observations at s_i $(i = 1, \cdots, k)$. We define $\hat{p}$ analogous to (1.1) and (1.4), that is,

$$\hat{p}(s_j) = \max_{L \ni s_j} \min_{L' \in U(L)} P(L - L') \tag{4.1}$$

where $P(L - L')$ is the $100p$ sample percentile of the collection $\{x_{ij}: i = 1, \cdots, n_i$ and $s_i \in L - L'\}$.

With the definitions, results analogous to Theorems 2.1, 2.2, 2.4 and Corollary 2.5 of Section 2 hold for $\hat{p}$. In Section 3, if one modifies the hypotheses and conclusions of Theorem 3.1, Corollary 3.2 and Corollary 3.3 and replaces the sample median by either the $[n \cdot p]$th or the $[n \cdot p] + 1$th order statistic then analogous results for $\hat{p}$ can be obtained by similar arguments.

We are grateful to Professor R. V. Hogg for pointing out that the above definition for percentiles gives the needed averaging property.

5. Some unresolved questions. There remain several interesting questions for which we have been unable to provide answers. The first question which comes to mind is suggested by (1.2) and (1.3). These two say that $\hat{\mu}$ was simultaneously nearest to what we observed and "in-between" the raw estimate and the unknown μ. Theorem 2.7 guarantees that $\hat{m}$ is nearest to our observations but we have no result analogous to (1.3) telling us that it is closer to the unknown m than the raw estimate. This type of result has not even been obtained in the linear case (cf. Robertson and Waltman (1968)). In the linear case a result analogous to the following was shown: There exists a measure α on the collection of all subsets of S with the property that $\hat{m} = E_\alpha(h \mid \mathscr{L})$, where $h(s_i) = M_i$. Such a result would imply that

$$\sum_{i=1}^k (\hat{m}(s_i) - m(s_i))^2 \alpha(\{s_i\}) \leqq \sum_{i=1}^k (M_i - m(s_i))^2 \alpha(\{s_i\}) .$$

However, it is not clear what this kind of result would mean since nice characterizations of α are unknown and α depends not only on the sample sizes but on the values of the observations. One might conjecture that

$$\sum_{i=1}^k |\hat{m}(s_i) - m(s_i)| n_i \leqq \sum_{i=1}^k |M_i - m(s_i)| n_i .$$

However, Malmgren (1972) has provided a counterexample.

In the consistency results the assumption that $\liminf_{n\to\infty} N_n(J)/n > 0$ for all non-degenerate rectangles $J \subset [0, 1] \times [0, 1]$ is a direct analogue of the condition used in Brunk (1970) and Cryer, *et al.* (1972) to obtain strong consistency of the analogous mean and median functions on the unit interval. In Hanson, *et al.* (1973) it was shown that something more than $\{t_k : k = 1, 2, \cdots\}$ dense in $[0, 1]$ was needed to obtain strong consistency. Is the above kind of condition what is needed to prove strong consistency for mean and median regression functions on $[0, 1]$ and $[0, 1] \times [0, 1]$? In Hanson, *et al.* (1973) and Cryer, *et al.* (1972) weak consistency results were obtained for mean and median regression functions on $[0, 1]$. Can the assumptions in Theorem 3.1 and Theorem 5 of Hanson, *et al.* (1973) be weakend and still obtain weak consistency? Condition (3.5) was also used in Hanson, *et al.* (1973), but has not been investigated.

Makowski (1971) obtained some law of the iterated logarithm type convergence rates for estimators of isotone regression functions on the line. Can similar results be obtained here?

Finally, it seems as though asymptotic distribution theory for "isotonized" estimators has been left relatively untouched. Some results (cf. Rao (1966) and

Brunk (1970)) have been obtained for the asymptotic distribution of the estimator at a point. It would seem that similar results could be obtained here. More desirable yet, can one obtain asymptotic distributions for more global type measures of the difference between the estimate and the true function like $\sup_t |\hat{m}(t) - m(t)|$, $\int_{[0,1]} [\hat{m}(t) - m(t)]^2\, dt$ or $\int_{[0,1]} |\hat{m}(t) - m(t)|\, dt$?

REFERENCES

[1] BRUNK, H. D. (1965). Conditional expectation given a σ-lattice and applications. *Ann. Math. Statist.* **36** 1339–1350.

[2] BRUNK, H. D. (1970). Estimation of isotonic regression. *Non Parametric Techniques in Statistical Inference*. Cambridge Univ. Press. 177–195.

[3] CRYER, J. D., ROBERTSON, TIM, WRIGHT, F. T. and CASADY, R. J. (1972). Isotonic median regression. *Ann. Math. Statist.* **43** 1459–1469.

[4] HANSON, D. L., PLEDGER, GORDON and WRIGHT, F. T. (1973). On consistency in monotonic regression. *Ann. Statist.* **1** 401–421.

[5] MAKOWSKI, GARY GEORGE (1971). Laws of the iterated logarithm for maximums of absolute values of partial sums of permuted random variables. Ph. D. dissertation, Univ. of Iowa.

[6] MALMGREN, EDWARD (1972). Contributions to the estimation of ordered parameters. Ph. D. thesis, Univ. of Iowa.

[7] PRAKASA RAO, B. L. S. (1966). Asymptotic distributions in some non-regular statistical problems. Technical Report No. 9, Department of Statistics, Michigan State Univ.

[8] ROBERTSON, TIM (1967). On estimating a density which is measurable with respect to a σ-lattice. *Ann. Math. Statist.* **38** 482–493.

[9] ROBERTSON, TIM and WALTMAN, PAUL (1968). On estimating monotone parameters. *Ann. Math. Statist.* **39** 1030–1039.

DEPARTMENT OF STATISTICS
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52240