The Annals of Mathematical Statistics
1972, Vol. 43, No. 5, 1459–1469

MONOTONE MEDIAN REGRESSION[1]

BY J. D. CRYER, TIM ROBERTSON, F. T. WRIGHT AND ROBERT J. CASADY

University of Iowa

Suppose that for each real number t in [0, 1] we have a distribution with distribution function $F_t(\cdot)$, mean $\mu(t)$ and median $m(t)$ ($\mu(t)$ and $m(t)$ are referred to as regression functions). Consider the problems of estimating $\mu(\cdot)$ and $m(\cdot)$.

In this paper we propose and discuss an estimator, $\hat{m}(\cdot)$, of $m(\cdot)$ which is monotone. This estimator is analogous to the estimator $\hat{\mu}(\cdot)$ of $\mu(\cdot)$ which was explored by Brunk (1970) (Estimation of isotonic regression in *Nonparametric Techniques in Statistical Inference*, Cambridge University Press, 177–195). Rates for the convergence of $\hat{m}(\cdot)$ to $m(\cdot)$ are given and a simulation study, where $\hat{m}(\cdot)$, $\hat{\mu}(\cdot)$ and the least squares linear estimator are compared, is discussed.

1. Introduction and summary. Suppose that for each real number t in [0, 1] we have a distribution function $F_t(\cdot)$ with mean $\mu(t)$ and median $m(t)$. ($\mu(t)$ and $m(t)$ will be referred to as regression functions.) Consider the problems of estimating $\mu(\cdot)$ and $m(\cdot)$. The most common approach for estimating a regression function is to assume that it belongs to a family of functions which is characterized by a small number of parameters, such as the family of linear functions. The investigator then obtains an estimate of the regression function by estimating these parameters.

We wish to consider some "nonparametric" estimators. Suppose one has reason to believe that the function we are trying to estimate is non-decreasing. In this setting Brunk (1970) proposed and studied an estimator of the mean regression function $\mu(\cdot)$. In this paper we explore an analogous estimate of $m(\cdot)$. A detailed bibliography of earlier work concerning the estimation of monotone (isotone) parameters may be found in Brunk (1970). Pursuant to the article by Brunk we pose our problem within the following framework.

Let $\{t_n\}$ be a sequence of numbers in [0, 1], not necessarily distinct, to be called observation points. Let $\{Y_n(t_n)\}$ be a sequence of independent random variables such that the distribution function of $Y_n(t_n)$ is $F_{t_n}(\cdot)$. Based on the first n observations, the estimator of $\mu(t_j)$ studied by Brunk is given by

$$(1.1)\qquad \begin{aligned}\hat{\mu}_n(t_j) &= \max_{r\leqq t_j}\min_{s\geqq t_j} A\{Y_i(t_i): i \leqq n, r \leqq t_i \leqq s\}\\ &= \min_{s\geqq t_j}\max_{r\leqq t_j} A\{Y_i(t_i): i \leqq n, r \leqq t_i \leqq s\}\end{aligned}$$

where the symbol A is used to denote the arithmetic average of the set of random

Received October 19, 1971; revised February 18, 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. monotone regression, median, consistency.

variables described inside the braces. The equality of the two representations for $\hat{\mu}_n(t_j)$ given in (1.1) seems to be well known; however, a discussion of this equality can be found in Hanson, Pledger and Wright (1971). Motivated by these considerations we propose an estimator of the median regression function $m(\cdot)$ which is given by

$$\begin{aligned}(1.2)\qquad \hat{m}_n(t_j) &= \max\nolimits_{r\leqq t_j}\min\nolimits_{s\geqq t_j} M\{Y_i(t_i): i \leqq n,\, r \leqq t_i \leqq s\}\\ &= \min\nolimits_{s\geqq t_j}\max\nolimits_{r\leqq t_j} M\{Y_i(t_i): i \leqq n,\, r \leqq t_i \leqq s\}\end{aligned}$$

where M denotes the median of the set of random variables described inside the braces (we adopt the convention of averaging the two middle items when the sample size is even). The equality of the two representations for $\hat{m}_n(t_j)$ follows from the work of Robertson and Waltman (1968). The value of either $\hat{m}_n(t)$ or $\hat{\mu}_n(t)$ at values of t between observation points can be specified in several different ways depending on the properties the investigator wishes his estimate to enjoy. For example, one might define $\hat{m}_n(t)$ to be the value of $\hat{m}_n(\cdot)$ at the largest observation point, among the first n observation points, which is no larger than t. This function would not be continuous and if one believed the actual regression function is continuous then this might be unsatisfactory. Another possibility would be to adjoin the values at adjacent observation points with line segments. The resulting estimators would be continuous but not differentiable. We can see that there are many possibilities.

The results of Robertson and Waltman (1968), interpreted in the present setting, show that $\hat{m}_n(\cdot)$ provides a maximum likelihood estimator when the distribution at t is the bilateral exponential with parameter $m(t)$ (i.e., this distribution is absolutely continuous with density function $f(x; m(t)) = 2^{-1}\cdot\exp[-|x - m(t)|]$). Another interesting way of interpreting this result is that $\hat{m}_n(\cdot)$ provides the nearest non-decreasing function to what we observed in a kind of L_1 sense. More specifically, if our observations are $y_1(t_1), y_2(t_2), \cdots, y_n(t_n)$ then

$$\sum_{i=1}^n |\hat{m}_n(t_i) - y_i(t_i)| \leqq \sum_{i=1}^n |g(t_i) - y_i(t_i)|$$

for any other non-decreasing function $g(\cdot)$ on [0, 1].

It also follows from the work of Robertson and Waltman (1968) that $\hat{m}_n(t_j)$ is a strongly consistent estimator of $m(t_j)$· when we keep the number of distinct observation points fixed and let the number at each grow large. Hanson, Pledger and Wright (1971) obtained convergence rates for the convergence of $\hat{\mu}_n(\cdot)$ to $\mu(\cdot)$ under a condition which implies that the sequence of observation points is dense in [0, 1]. These convergence rates imply that $\hat{\mu}_n(\cdot)$ is a strongly consistent estimator of $\mu(\cdot)$. In Section 2 we obtain similar convergence rates for convergence of $\hat{m}_n(\cdot)$ to $m(\cdot)$ under this same assumption about the sequence of observation points.

Simulation studies of "isotonized" estimators of population characteristics have been carried out by the authors and others (cf. Wegman (1970) and Hogg and Malmgren (1971)). Generally speaking, "isotonized" estimators do not perform as well as parametric estimators when the assumed parametric model is the correct one. (This is, of course, to be expected.) However, they perform better than the corresponding totally nonparametric estimators, that is, the raw sample means and medians, when the function to be estimated is monotone. In Section 3 we discuss the results of a simultaion study where we considered $\hat{\mu}_n(\cdot)$ and $\hat{m}_n(\cdot)$ as competitors and compared them to a least squares linear estimator of the regression function.

2. Consistency. In this section we investigate the rate of convergence for the estimator $\hat{m}_n(\cdot)$. Consistency results for isotonized estimators are known under two different hypotheses about the set, $\{t_n\}$, of observation points. The first and perhaps easier situation is when the number of observations at each of a fixed number of observation points gets large. In this case almost sure convergence of $\hat{\mu}_n(t_j)$ to $\mu(t_j)$ is an easy consequence of the strong law of large numbers. Similarly the almost sure convergence of $\hat{m}_n(t_j)$ to $m(t_j)$ follows from well-known results concerning the almost sure convergence of sample percentiles to population percentiles. It seems clear that convergence rates could easily be obtained under this hypothesis about $\{t_n\}$.

The other hypothesis about $\{t_n\}$ which has yielded convergence rates is one which loosely says that we have a large number of distinct observation points but perhaps only a few observations at each. We need some preliminary results which we give here for the sake of completeness.

LEMMA 2.1. *If the random variables $X_1, X_2, \cdots, X_n$ are independent and centered at expectations then for any $m \leqq n$, and positive number ε,*

$$P[\max_{m \leqq j \leqq n} S_j \geqq \varepsilon] \leqq e^{-\varepsilon t} E[e^{tS_n}]$$

for all $t \geqq 0$ $(S_j = \sum_{i=1}^{j} X_i)$.

PROOF. The proof of this lemma is essentially the same as the proof of the Kolmogorov inequality. Jensen's inequality is applied to the convex function e^{tx} to obtain $E(e^{tX}) \geqq 1$ when $E(X) = 0$.

The next lemma is a consequence of the results in Hanson (1967).

LEMMA 2.2. *If $\{X_n\}$ is a sequence of uniformly bounded random variables which are centered at expectations then there exists a positive constant c such that*

$$E[e^{X_n t}] \leqq e^{c \cdot t^2}$$

for all t.

For the remainder of this section we assume that $m(\cdot)$ is continuous on $[0, 1]$ and that for each positive number ε,

$$\inf_k F_{t_k}(m(t_k) + \varepsilon) - \tfrac{1}{2} > 0 \tag{2.1}$$

and

$$\tfrac{1}{2} - \sup_k F_{t_k}(m(t_k) - \varepsilon) > 0 \,. \tag{2.2}$$

It is of interest to note that (2.1) and (2.2) are satisfied when the random variables $Y_j(t_j) - m(t_j)$ are identically distributed, their common distribution function G has median 0 and

$$G(-\varepsilon) < \tfrac{1}{2} < G(\varepsilon)$$

for all $\varepsilon > 0$.

THEOREM 2.3. *If for each nondegenerate closed subinterval I of* $(0, 1)$ *we have*

$$\liminf_{n\to\infty} [n^{-1} \cdot \textit{cardinality of } \{j: j = n, t_j \in I\}] > 0 \tag{2.3}$$

then for each observation point t_k in $(0, 1)$ *and each positive number ε, there exists a positive constant c and a number ρ* $(0 \leqq \rho < 1)$ *such that*

$$P[|\hat{m}_n(t_k) - m(t_k)| > \varepsilon] \leqq c \cdot \rho^n . \tag{2.4}$$

PROOF. Fix k and suppose s_0 is an arbitrary number between t_k and one. Then, using the definition of $\hat{m}_n(\cdot)$, obvious properties of medians and maximums and finally the assumption that $m(\cdot)$ is non-decreasing we can write:

$$\begin{aligned}
\hat{m}_n(t_k) - m(t_k) &= \min_{s \geqq t_k} \max_{r \leqq t_k} M\{Y_j(t_j); j \leqq n, r \leqq t_j \leqq s\} - m(t_k) \\
&\leqq \max_{r \leqq t_k} M\{Y_j(t_j); j \leqq n, r \leqq t_j \leqq s_0\} - m(t_k) \\
&= \max_{r \leqq t_k} M\{Y_j(t_j) - m(s_0); j \leqq n, r \leqq t_j \leqq s_0\} \\
&\qquad + m(s_0) - m(t_k) \\
&\leqq \max_{r \leqq t_k} M\{Y_j(t_j) - m(t_j); j \leqq n, r \leqq t_j \leqq s_0\} \\
&\qquad + m(s_0) - m(t_k) .
\end{aligned}$$

Fix n at least as large as k, let ε be an arbitrary positive number and choose s_0, between t_k and one, so that $m(s_0) - m(t_k) < \varepsilon/2$. Relabeling the first n observation points and the associated random variables, if necessary, assume that $k = 1$ and that $t_1 \leqq t_2 \leqq \cdots \leqq t_l \leqq s_0$, $t_1 > t_{l+1} \geqq t_{l+2} \geqq \cdots \geqq t_h \geqq 0$ and $s_0 < t_{h+1} \leqq t_{h+2} \leqq \cdots \leqq t_n$. Letting $Z_j = Y_j(t_j) - m(t_j)$ we have:

$$\begin{aligned}
\hat{m}_n(t_1) - m(t_1) &\leqq \max_{l \leqq j \leqq h} M\{Z_1, Z_2, \cdots, Z_j\} + \varepsilon/2 \\
&\leqq \max_{l \leqq j \leqq n} M\{Z_1, Z_2, \cdots, Z_j\} + \varepsilon/2 .
\end{aligned}$$

Now let $W_i = I_{(-\infty, \varepsilon/2]}(Z_i)$ and observe that $M\{Z_1, Z_2, \cdots, Z_j\} > \varepsilon/2$ implies that $\sum_{i=1}^{j} W_i \leqq j/2$. Thus, using the fact that $E(W_i) = F_{t_i}(m(t_i) + \varepsilon/2)$ we can write:

$$\begin{aligned}
[\hat{m}_n(t_1) - m(t_1) > \varepsilon] &\subset [\max_{l \leqq j \leqq n} M\{Z_1, Z_2, \cdots, Z_j\} > \varepsilon/2] \\
&\subset [\min_{l \leqq j \leqq n} (\textstyle\sum_{i=1}^{j} W_i - j/2) \leqq 0] \\
&\subset [\min_{l \leqq j \leqq n} \{\textstyle\sum_{i=1}^{j} (W_i - F_{t_i}(m(t_i) + \varepsilon/2)) \\
&\qquad + \textstyle\sum_{i=1}^{j} F_{t_i}(m(t_i) + \varepsilon/2) - j/2\} \leqq 0] \\
&\subset [\min_{l \leqq j \leqq n} \{\textstyle\sum_{i=1}^{j} (W_i - E(W_i)) \\
&\qquad + j(\inf_i F_{t_i}(m(t_i) + \varepsilon/2) - \tfrac{1}{2})\} \leqq 0] .
\end{aligned}$$

It follows from (2.1) that there exists a positive number δ such that

$$\inf_i F_{t_i}(m(t_i) + \varepsilon/2) - \tfrac{1}{2} > \delta$$

so that

$$\begin{aligned}[\hat{m}_n(t_1) - m(t_1) > \varepsilon] &\subset [\min_{l \leqq j \leqq n} \{\textstyle\sum_{i=1}^{j} (W_i - E(W_i)) + j\delta\} \leqq 0] \\ &\subset [\min_{l \leqq j \leqq n} \{j^{-1} \textstyle\sum_{i=1}^{j} (W_i - E(W_i))\} \leqq -\delta] \\ &= [\max_{l \leqq j \leqq n} \{j^{-1} \textstyle\sum_{i=1}^{j} (E(W_i) - W_i)\} \geqq \delta] \\ &\subset [\max_{l \leqq j \leqq n} \{\textstyle\sum_{i=1}^{j} (E(W_i) - W_i)\} \geqq l \cdot \delta] \, . \end{aligned} \tag{2.5}$$

Now the random variables $E(W_i) - W_i$ are independent, identically distributed, centered at expectations and uniformly bounded so that by taking probabilities in (2.5) and applying Lemma 2.1 and then Lemma 2.2 we conclude that

$$\begin{aligned} P[\hat{m}_n(t_1) - m(t_1) > \varepsilon] &\leqq e^{-l \cdot \delta \cdot y} E[e^{yS_n}] \\ &\leqq e^{-l \cdot \delta \cdot y} e^{n_1 c_1 \cdot y^2} \end{aligned}$$

for some positive constant c_1 and for all nonnegative y (S_n is defined in the obvious fashion). Now l is the number of observations in $[t_1, s_0]$ so that by (2.3) there exists a positve constant c_2 such that $n \cdot c_2 \leqq l$ for n sufficiently large. Thus

$$P[\hat{m}_n(t_1) - m(t_1) > \varepsilon] \leqq \exp[-n \cdot c_2 \cdot \delta \cdot y + n \cdot c_1 \cdot y^2]$$

for all positive y. Letting y have the value $\delta \cdot c_2/2 \cdot c_1$ we conclude that:

$$P[\hat{m}_n(t_1) - m(t_1) > \varepsilon] \leqq \exp[-n \cdot c_2{}^2 \cdot \delta^2/4c_1] = \alpha^n$$

where $0 \leqq \alpha < 1$. A similar argument, using (2.2) rather than (2.1) gives

$$P[\hat{m}_n(t_1) - m(t_1) < -\varepsilon] \leqq \gamma^n$$

for sufficiently large n ($0 \leqq \gamma < 1$). The desired result can be easily obtained using these two conclusions.

COROLLARY 2.4. *Assume the hypotheses of Theorem* 2.3. *Then for* $0 < a < b < 1$ *and* $\varepsilon > 0$, *there exist constants* $c > 0$ *and* $0 \leqq \rho < 1$ *such that*

$$P[\sup_{a \leqq t \leqq b} |\hat{m}_n(t) - m(t)| \geqq \varepsilon] \leqq c \cdot \rho^n \, . \tag{2.6}$$

It follows that

$$P[\lim_{n \to \infty} \sup_{a \leqq t \leqq b} |\hat{m}_n(t) - m(t)| = 0] = 1 \, . \tag{2.7}$$

PROOF. In order to see that (2.6) follows from Theorem 2.3, choose l observation points such that

$$0 < t_{k(1)} < a < t_{k(2)} < \cdots < t_{k(l-1)} < b < t_{k(l)} < 1 \qquad \text{and}$$

$$m(t_{k(i+1)}) - m(t_{k(i)}) < \varepsilon/2 \tag{2.8}$$

$i = 1, 2, \cdots, l - 1$. It follows from Theorem 2.3 that for each $t_{k(i)}$ ($i = 1, 2, \cdots, l$), there exist constants $c_i = c(t_{k(i)}) > 0$ and $0 \leqq \rho(t_{k(i)}) = \rho_i < 1$ such that for n sufficiently large:

$$P[|\hat{m}_n(t_{k(i)}) - m(t_{k(i)})| < \varepsilon/2] \leqq c_i \cdot \rho_i{}^n \, .$$

Hence

$$P[\max_{i \leqq l} |\hat{m}_n(t_{k(i)}) - m(t_{k(i)})| > \varepsilon/2] \leqq c \cdot \rho^n \, .$$

with $c = \sum_{i=1}^{l} c_i$ and $\rho = \max_{i \leqq l} \rho_i$. However, using (2.8) we conclude that

$$[\sup_{a \leqq t \leqq b} |\hat{m}_n(t) - m(t)| > \varepsilon] \subset [\max_{i \leqq l} |m_n(t_{k(i)}) - m(t_{k(i)})| > \varepsilon/2]$$

and (2.6) follows. The conclusion (2.7) follows from (2.6) and the Borel–Cantelli Lemma.

COROLLARY 2.5. *Under the hypotheses of Theorem* 2.3 *it follows that*

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

PROOF. From Theorem 2.3 it follows that there exists a countable dense subset T (namely, the observation points excluding zero and one) such that

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

Using the fact that $\hat{m}_n(t)$ is nondecreasing and the assumption that $m(t)$ is continuous, one can argue that

$$[\lim_{n\to\infty} \hat{m}_n(t) = m(t) \text{ for all } t \in T] \subset [\lim_{n\to\infty} \hat{m}_n(t) = m(t) \text{ for all } t \in (0, 1)]$$

and the desired result follows.

We note that the techniques used here are similar to those of Hanson, Pledger and Wright (1971). However, the results are easier to obtain because we are able to convert statements about medians into statements about sums of independent Bernoulli random variables. One might wonder if a weaker hypothesis than (2.3) would suffice for medians. It can be seen by examining the proof of Theorem 2 of Hanson, *et al.* (1971) that a similar example can be constructed for medians to show that strong consistency does not follow if one simply assumes that the set of observation points is dense in [0, 1]. However, we can prove a theorem and corollary from which we can conclude that our estimator is weakly consistent when the set of observation points is dense in [0, 1]. For the sake of completeness we restate Lemma 3 of Hanson, Pledger and Wright (1971).

LEMMA 2.6. *Let F be a real valued function on* $[0, \infty)$ *such that* $F(y) \to 0$ *as* $y \to \infty$ *and* $\int_0^\infty y|dF(y)| < \infty$. *Then for each* $\varepsilon > 0$ *there exists a positive integer* M *such that if* $\{X_i : i = 1, 2, \cdots\}$ *is an independent sequence of random variables such that* $EX_i = 0$ *and* $P[|X_i| \geqq y] \leqq F(y)$ *for all* i *and* $y \geqq 0$, *then*

$$P[\sup_{M \leqq n} n^{-1}|X_1 + \cdots + X_n| \geqq \varepsilon] \leqq \varepsilon . \tag{2.9}$$

THEOREM 2.7. *If the set of observation points is dense in* [0, 1], *then for each observation point* $t_k \in (0, 1)$ *and each* $\varepsilon > 0$, *we have*

$$P[|\hat{m}_n(t_k) - m(t_k)| > \varepsilon] \to 0 \qquad \text{as } n \to \infty .$$

PROOF. Using the techniques and the definition of W_i given in the proof of Theorem 2.3, we see that for $n \geqq k$

$$[\hat{m}_n(t_k) - m(t_k) > \varepsilon] \subset [\max_{l \leqq j \leqq n} j^{-1} \sum_{i=1}^{j} (E(W_i) - W_i) \geqq \delta]$$

where $l = \mathrm{card}(\{j\colon j \leqq n,\ t_k \leqq t_j \leqq s_0\})$. Let $\eta > 0$. Since the W_i's are uniformly bounded there clearly exists an F as hypothesized in Lemma 2.6. So corresponding to $\gamma = \min(\eta, \delta)$ there exists a positive integer M such that

$$P[\sup_{j \geqq M} j^{-1} \cdot \textstyle\sum_{i=1}^{j} (E(W_i) - W_i) \geqq \gamma] \leqq \gamma\ .$$

So for $l \geqq M$

$$\begin{aligned} &P[\max_{l \leqq j \leqq n} j^{-1} \cdot \textstyle\sum_{i=1}^{j} (E(W_i) - W_i) \geqq \delta] \\ &\qquad \leqq P[\sup_{M \leqq j} j^{-1} \cdot \textstyle\sum_{i=1}^{j} (E(W_i) - W_i) \geqq \gamma] \leqq \gamma \leqq \eta\ . \end{aligned}$$

Since the set $\{t_k : k = 1, 2, \cdots\}$ is dense in $[0, 1]$, we see that $l \to \infty$ as $n \to \infty$. Hence $P[\hat{m}_n(t_k) - m(t_k) > \varepsilon] \to 0$ as $n \to \infty$. A similar argument shows that $P[\hat{m}_n(t_k) - m(t_k) < -\varepsilon] \to 0$ as $n \to \infty$.

COROLLARY 2.8. *If the set of observation points is dense in* $[0, 1]$, *then for any* $0 < a < b < 1$ *and any* $\varepsilon > 0$

$$P[\sup_{a \leqq t \leqq b} |\hat{m}_n(t) - m(t)| > \varepsilon] \to 0 \qquad \textit{as } n \to \infty\ .$$

PROOF. This result follows from Theorem 2.7 just as (2.6) follows from Theorem 2.3.

3. Comparisons among estimators.

For certain distributions the mean and median functions are identical. In such instances we might consider the estimators of these regression functions provided by (1.1) and (1.2) as competitors and it is meaningful to compare their properties. An assumption which is frequently made is that the regression function is linear or at least approximately so. Hence another competitor is provided by:

$$\hat{l}(t) = \hat{a} + \hat{b}t \tag{3.1}$$

where $\hat{a}$ and $\hat{b}$ are the usual least squares estimators. In this section we discuss some conclusions which are based mostly on a simulation study of these three estimates and illustrate these remarks with two examples.

As measures of fit we considered both mean squared error and mean absolute error. We computed these, both pointwise at the observation points and as a global measure summed them over the distinct observation points. We also obtained the pointwise bias of the estimators.

For the least squares estimator, these quantities are easily obtained from standard results:

$$\begin{aligned} E[\hat{l}(t)] = (t - \bar{t}) \cdot [\textstyle\sum_{j=1}^{n} (t_j - \bar{t})^2]^{-1} \sum_{j=1}^{n} (t_j - \bar{t})\mu(t_j) \\ + n^{-1} \cdot \textstyle\sum_{j=1}^{n} \mu(t_j)\ , \end{aligned} \tag{3.2}$$

$$\mathrm{Var}[\hat{l}(t)] = \sum_{j=1}^{n} \left[\frac{(t - \bar{t})(t_j - \bar{t})}{\sum_{j=1}^{n} (t_j - \bar{t})^2} + n^{-1} \right]^2 \cdot \mathrm{Var}[Y_j(t_j)]\ , \tag{3.3}$$

where $\bar{t} = n^{-1} \sum_{j=1}^{n} t_j$. If, in addition, the Y's are normally distributed we may use:

$$E[|\hat{l}(t) - \mu(t)|] = (2v/\pi)^{\frac{1}{2}} \cdot e^{-b^2/2v} + 2b \cdot [\Phi(b/v^{\frac{1}{2}}) - (\tfrac{1}{2})] \tag{3.4}$$

where $v = \mathrm{Var}[\hat{l}(t)]$, $b = \text{bias} = E[\hat{l}(t)] - \mu(t)$ and Φ is the standard normal distribution function. Unfortunately, these moment computations are mathematically intractable for our nonparametric estimators so that we resorted to simulation techniques.

Bias. In general, both of the nonparametric estimators are biased at the extreme observation points. For example, if $q = \min\{t_1, t_2, \cdots, t_n\}$ then it follows from the definition of $\hat{\mu}(q)$ that

$$\begin{aligned}\hat{\mu}(q) &= \min\nolimits_{s \geqq q} A\{Y_j(t_j); j \leqq n, q \leqq t_j \leqq s\} \\ &\leqq A\{Y_j(t_j); j \leqq n, t_j = q\}\,.\end{aligned}$$

However, $A\{Y_j(t_j); j \leqq n, t_j = q\}$ is an unbiased estimator of $\mu(q)$ so that $E[\hat{\mu}(q)] \leqq \mu(q)$. Furthermore, this inequality is strict except in the rare circumstance that the distribution at any observation point t_j with $t_j > q$ does not overlap the distribution at q. In fact, if there is substantial overlap among the distributions at $t = q$ and other observation points then we would expect $E[\hat{\mu}(q)]$ to be much smaller than $\mu(q)$. In an analogous fashion, $\hat{\mu}(\cdot)$ is biased high for t's at the other extreme. Similar remarks can be made about $E[\hat{m}(t)]$, at least for symmetric distributions.

In order to illustrate these comments we considered the case when $\mu(t) = 4t^4$ and the distribution at t was the bilateral exponential with variance one. We took

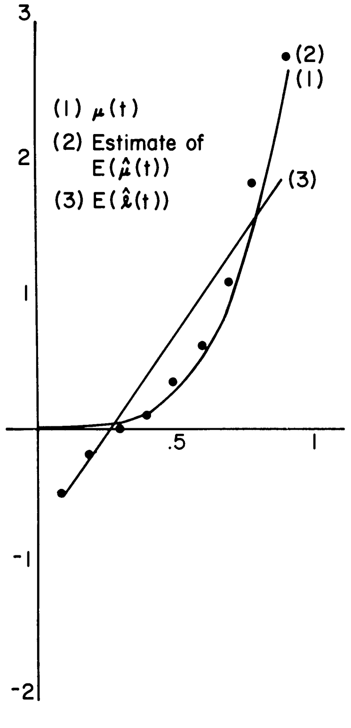

FIG. 1. Bias.

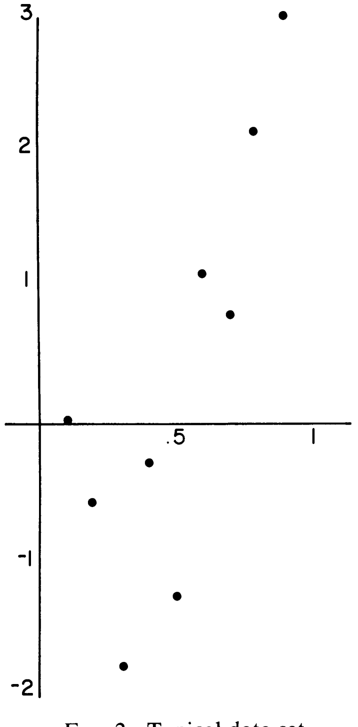

FIG. 2. Typical data set.

one observation at each of nine equally spaced observation points; .1, .2, $\cdots$, .9. In Fig. 1 we have plotted $\mu(t)$ together with $E[\hat{l}(t)]$ (obtained from (3.1)) and an estimate of $E[\hat{\mu}(t)]$ based on 10,000 iterations of the sampling experiment. Our estimate of $E[\hat{\mu}(t)]$ is plotted only at the observation points. We did not plot the corresponding estimate of $E[\hat{m}(t)]$ because it was almost identical to $E[\hat{\mu}(t)]$.

Observe that $E[\hat{\mu}(t)]$ is much lower than $\mu(t)$ at values of t close to zero. On the other hand, $E[\hat{\mu}(t)]$ is very close to $\mu(t)$ at values of t near one. This is true because the distributions at adjacent observation points close to zero are overlapping but the distributions at observation points to the right of $\frac{1}{2}$ do not overlap greatly because the mean function rises abruptly there.

Error. As is well known, the least squares estimator, $\hat{l}(\cdot)$, perfoms very well when $\mu(\cdot)$ is linear or even approximately linear. On the other hand, one might argue that when $\mu(\cdot)$ is very nonlinear this fact could be recognized and some other, more appropriate, parametric model would be chosen. However, in Fig. 2 we have plotted what we consider to be a typical data set when $\mu(t) = 4t^4$ and the distribution at t is the bilateral exponential with variance one. Some investigators might attempt to fit a line to such data even in the absence of prior knowledge that the mean function is (at least approximately) linear. In such cases the weaker assumption of monotonicity would appear to be more prudent. In our simulation study we compared the errors (both mean squared and mean absolute) of $\hat{\mu}(\cdot)$, $\hat{m}(\cdot)$ and $\hat{l}(\cdot)$ for this mean function.

With $\mu(t) = 4t^4$ and the distribution at t normal and bilateral exponential, we varied the number of observation points and the number of observations at each. As expected, $\hat{\mu}(\cdot)$ did slightly better than $\hat{m}(\cdot)$ in the normal case and vice-versa in the other case. In this nonlinear situation even when the number of observations is small, $\hat{m}(\cdot)$ and $\hat{\mu}(\cdot)$ have smaller mean absolute error than

TABLE 1

Bilateral exponential distribution with $\mu(t) = 4t^4$ and $\sigma^2 = 1$

t	Mean Squared Error				Mean Absolute Error		
	$\hat{l}(\cdot)$		$\hat{\mu}(\cdot)$	$\hat{m}(\cdot)$	$\hat{l}(\cdot)$	$\hat{\mu}(\cdot)$	$\hat{m}(\cdot)$
	exact	est.	est.	est.	est.	est.	est.
.1	.624	.619	.666	.653	.636	.575	.556
.2	.304	.301	.302	.280	.428	.411	.385
.3	.181	.178	.239	.211	.326	.375	.343
.4	.209	.205	.239	.206	.367	.378	.345
.5	.297	.293	.272	.237	.460	.402	.371
.6	.337	.333	.318	.289	.489	.434	.413
.7	.274	.271	.396	.375	.417	.483	.472
.8	.267	.269	.527	.517	.405	.553	.549
.9	.964	.976	.769	.767	.836	.637	.638
Total	3.457	3.445	3.489	3.535	4.364	4.248	4.072

$\hat{l}$ and all three estimators have about the same mean squared error. (One would expect the nonparametric estimators to outperform $\hat{l}(\cdot)$ when the number of observations is large and $\mu(\cdot)$ is nonlinear since they are consistent estimators and $\hat{l}(\cdot)$ is not.)

In Table 1 we have given the results of one such simulation study. Again, the distribution at t was the double exponential with mean $\mu(t) = 4t^4$ and variance one. We took nine observation points with one observation at each. The estimates of the various moments of the estimators were based on 10,000 iterations of the procedure. We tabulated the mean squared error of $\hat{l}(\cdot)$, $\hat{\mu}(\cdot)$, and $\hat{m}(\cdot)$. In order to give the reader a feel for the sampling error involved, we tabulated for $\hat{l}(\cdot)$ both the exact values obtained from (3.2) and (3.3) and estimates obtained from the simulation. Mean absolute error estimates are also given. Note that all of the estimators do much better in the middle than in the tails.

TABLE 2

Normal distribution with $\mu(t) = 4t$ and $\sigma^2 = 1$

t	Mean Squared Error				Mean Absolute Error			
	$\hat{l}(\cdot)$		$\hat{\mu}(\cdot)$	$\hat{m}(\cdot)$	$\hat{l}(\cdot)$		$\hat{\mu}(\cdot)$	$\hat{m}(\cdot)$
	exact	est.	est.	est.	exact	est.	est.	est.
.1	.378	.379	.686	.694	.490	.489	.643	.649
.2	.261	.262	.425	.445	.408	.407	.520	.531
.3	.178	.179	.382	.412	.336	.336	.495	.511
.4	.128	.128	.365	.391	.285	.285	.486	.502
.5	.111	.111	.359	.386	.266	.265	.480	.495
.6	.128	.126	.363	.387	.285	.284	.484	.497
.7	.178	.175	.384	.409	.336	.333	.497	.512
.8	.261	.256	.433	.450	.408	.405	.524	.534
.9	.378	.371	.676	.687	.490	.486	.640	.646
Total	2.000	1.987	4.073	4.261	3.304	3.290	4.769	4.877

As a final comparison we present in Table 2 the results from a case most favorable to the usual least squares regression estimator. Here the distribution at t was taken to be normal with mean, $4t$ and variance one with one observation at each of nine points. Mean squared error and mean absolute error are given, as in Table 1, with the additon of exact results from (3.4).

In summary, what do we recommend to the statistician who might consider using $\hat{\mu}(\cdot)$ or $\hat{m}(\cdot)$? First of all, if there is reason to believe that $\mu(\cdot)$ is nearly linear then it is difficult to find a better estimator than $\hat{l}(\cdot)$. On the other hand, if one is not sure, then $\hat{\mu}(\cdot)$ and $\hat{m}(\cdot)$ do provide, in the case that the regression functions are monotonic, some protection against the possibility that $\mu(\cdot)$ or $m(\cdot)$ is very nonlinear.

Finally, $\hat{m}(\cdot)$ is more difficult to calculate than $\hat{\mu}(\cdot)$ but these calculations pose no problem for a high speed computer. Both can be calculated using techniques which involve pooling of observations at adjacent observation points (cf. Robertson and Waltman (1968)).

REFERENCES

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

[2] HANSON, D. L. (1967). Some results relating moment generating functions and convergence rates in the law of large numbers. *Ann. Math. Statist.* **38** 742–750.

[3] HANSON, D. L., PLEDGER, GORDON and WRIGHT, F. T. (1971). On consistency in monotonic regression. Technical Report No. 45, Mathematical Sciences, Univ. of Missouri.

[4] HOGG, ROBERT V. and MALMGREN, EDWARD G. (1971). Nonparametric regression using a Bayesian treatment of Brunk-type estimates. Unpublished manuscript.

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

[6] WEGMAN, EDWARD J. (1970). Nonparametric probability density estimation: II. A comparison of density estimation methods. Unpublished manuscript.

DEPARTMENT OF STATISTICS
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52240