Nonparametric Estimation of a Regression Function by Means of Concomitants of Order Statistics

by

Gordon Johnston*

University of North Carolina at Chapel Hill
and
Texas Instruments Corporation

Abstract

We study the properties of nonparametric estimates of a regression function based on concomitants of order statistics (Yang (1977)).

Key Words: Nonparametric regression, density estimation, concomitants of order statistics

*The work of this author was supported by the Air Force Office of Scientific Research Contract AFOSR-75-2796.

1. Introduction.

S. S. Yang (1977) proposed as an estimation of the regression function $m(u) = E[Y|X = u]$ of a bivariate random vector (X,Y) the statistic M_n defined by

$$M_n(u) = (n\varepsilon_n)^{-1} \sum_{i=1}^{n} K\left(\frac{i/n - F_n(u)}{\varepsilon_n} \right) Y_{[i:n]}$$

Here $\{\varepsilon_n^{-1} K(x/\varepsilon_n)\}$ is a δ-function sequence of kernel type (Watson and Leadbetter (1964)) (X_i, Y_i), $i=1,\ldots,n$ are i.i.d. observations on (X,Y), F_n is the empirical distribution function (EDF) of the X-observations, and $Y_{[i:n]}$ is the Y-observation corresponding to the i-th order statistic of the X-observations, i.e., the i-th *concomitant* of the X-values n (see, e.g., Yong (1977)).

Our purpose here is to find conditions under which

$$(1.1) \qquad (n\varepsilon_n \log n)^{\frac{1}{2}} \left[\sup_{a \le u \le b} \left| \frac{(n\varepsilon_n)^{\frac{1}{2}} [M_n(u) - m(u)]}{[s(u) \int k^2(t)dt]^{\frac{1}{2}}} \right| - d_n \right]$$

$\xrightarrow{L}$ E as $n \to \infty$, where E is a random variable with density $e^{-2e^{-x}}$, $x > 0$, a, b, are constants, $\{\varepsilon_n\}$ and $\{d_n\}$ are appropriate real sequences and $s(u) = E[Y^2|X = a]$. Bickel and Rosenblatt (1973) proved a similar result for kernel estimates of a density function. A large sample confidence interval for $m(u)$, based on $M_n(u)$ is given, using (1.1).

We also give conditions under which

$$(1.2) \qquad (n\varepsilon_n)^{\frac{1}{2}} [M_n(u) - m(u)] \xrightarrow{L} N(0,\ s(u) \int k^2(t)dt) \text{ as } n \to \infty$$

for appropriate points u and sequence $\{\varepsilon_n\}$.

Our method of proof is to show that

$$(1.3) \quad (n\varepsilon_n \log n)^{\frac{1}{2}} \sup_{a \le u \le b} |M_n(u) - M_n^{**}(u)| \xrightarrow{P} 0,$$

where M_n^{**} is defined by

$$(1.4) \qquad M_n^{**}(u) = (n\varepsilon_n)^{-1} \sum_{i=1}^{n} Y_i K((F(X)_i) - F(u))/\varepsilon_n).$$

M_n^{**} is a special case of the regression function estimation proposed by Watson (1964). Johnston (1979) gives conditions under which (1.1) and (1.2) hold for M_n^{**} in place of M_n, and (1.1) and (1.2) will thus hold by virtue of (1.3).

2. Asymptotic Equivalence of M_n and M_n^{**}.

In this section we verify (1.3). The proof is given in the Appendix since it is rather technical and lengthy. Define

$$M_n^{*}(u) = (n\varepsilon_n)^{-1} \sum_{i=1}^{n} Y_i K((F_n(X_i) - F(u))/\varepsilon_n)$$

Then Lemma 2.1 gives conditions under which

$$(2.1) \qquad (n\varepsilon_n \log n)^{\frac{1}{2}} \sup_{a \le u \le b} |M_n^{*}(u) - M_n(u)| \xrightarrow{P} 0$$

$$(2.2) \qquad (n\varepsilon_n \log n)^{\frac{1}{2}} \sup_{a \le u \le b} |M_n^{**}(u) - M_n^{*}(u)| \xrightarrow{P} 0 ,$$

which together imply (1.3).

Lemma 2.1 Suppose $\{\varepsilon_n^{-1} K(x/\varepsilon_n)\}$ is a δ-function sequence such that $(\log n)^{-1} (n\varepsilon_n^{\frac{1}{2}}) \to \infty$. K has bounded support and 3 bounded continuous derivatives on the support. Suppose $\int |K''(t)|dt < \infty$ and K and K' are of bounded variation.

Let (X,Y) be such that $E|Y| < \infty$, $g(u) = E[Y|X = F^{-1}(u)]$ has 2 bounded derivarives on [0,1] and $h(u) = E[(Y)|X = F^{-1}(u)]$ is bounded on [0,1].

Assume there exists a real sequence $\{a_n\}$ such that $a_n \to \infty$, $a_n^2 \log n/(n\varepsilon_n^3) \to 0$ and

$$n^{\frac{1}{2}} \int_{|y| > a_n} |y| dF^{Y}(y) \to 0 \text{ as } n \to \infty .$$

Then, for $0 < F(a) < F(b) < 1$, (2.1) and (2.2) hold. □

3. Applications.

We will assume throughout this section that the assumptions of Theorem 2.1 are in force. We first note that M_n^{**} may be written as

$$M_n^{**}(u) = (n\varepsilon_n)^{-1} \sum_{i=1}^{n} Y_i K((Z_i - F(u))/\varepsilon_n)$$

where

$$Z_i = F(X_i) \sim U(0,1).$$

According to Theorem 2.5.2 of Johnston (1979), under certain conditions,

$$(n\varepsilon_n)^{\frac{1}{2}} [M_n^{**}(u) - E(Y|Z = F(u))] \xrightarrow{L} N(0, E(Y^2|Z = F(u)) \int K^2(t)dt).$$

If we assume F to be strictly increasing, then

$$E(Y|Z = F(u)) = m(u)$$

and

$$E(Y^2|Z = F(u)) = s(u).$$

Thus we have, by virtue of (1.3)

$$(n\varepsilon_n)^{\frac{1}{2}}[M_n(u) - m(u)]$$

$$\xrightarrow{L} N(0, s(u) \int K^2(t)dt),$$

which completes the proof of normality of M_n. We note that this asymptotic variance differs from that of Yong (1977), Theorem 6.

If the conditions of Corollary 3.2.9 of Johnston (1979) hold, then

$$(3.1) \qquad (2\delta \log n)^{\frac{1}{2}} \left[\sup_{a \le u \le b} \left| \frac{(n\varepsilon_n)^{\frac{1}{2}}[M_n^{**}(u) - m(u)]}{[s(u) \int K^2(t)dt]^{\frac{1}{2}}} \right| - d_n \right] \xrightarrow{L} E,$$

where E is a random variable with density $e^{-2e^{-x}}$, $x > 0$. Here $\varepsilon_n = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{2}$ and d_n is the sequence of entering constants specified in Bickel and Rosenblatt (1973). By virtue of (1.3), (3.1) holds with M_n replacing M_n^{**}, as we wished to prove. Inverting (3.1) in the usual way yields an approximate $(1-\alpha) \times 100\%$ confidence band for $m(u)$ over the interval (a,b), based on $M_n(u)$:

$$M_n(u) \pm (n\varepsilon_n)^{-\frac{1}{2}}[s(u) \int K^2(t)dt]^{\frac{1}{2}} \left[d_n + \frac{c(\alpha)}{(2\delta \log n)^{\frac{1}{2}}}\right]$$

where

$$c(\alpha) = \log 2 - \log|\log (1-\alpha)|.$$

APPENDIX

Proof of Lemma 2.1.

We begin with the following preliminary lemma, which is very similar to Lemma 1 of Bhattacharyya (1967).

A1. Lemma Assume that $g(u) = E[Y|X = F^{-1}(u)]$ has r continuous derivatives on $[0,1]$, $r > 0$, and that K has bounded support and r bounded derivatives on the support. Then for a, b such that $0 < F(a) < F(b) < 1$,

$$\left|\varepsilon_n^{-(r+1)} \iint yK^{(r)}((F(x) - F(z))/\varepsilon_n)dF(x,y)\right| = O(1)$$

uniformly in $z \in [a,b]$ as $n \to \infty$.

Proof. Note that

$$\varepsilon_n^{-(r+1)} \iint yK^{(r)}((F(x) - F(z))/\varepsilon_n)\, dF(x,y)$$

$$= \varepsilon_n^{-(r+1)} EYK^{(r)}((F(X) - F(z))/\varepsilon_n)$$

$$= \varepsilon_n^{-(r+1)} \int m(x)K^{(r)}((F(x) - F(z))/\varepsilon_n)\, dF(x)$$

$$= \varepsilon_n^{-(r+1)} \int_0^1 g(u)K^{(r)}((u-F(z))/\varepsilon_n)\, du.$$

Now write

$$\varepsilon_n^{-(r+1)} g(u)K^{(r)}((u-F(z))/\varepsilon_n)$$

$$= \varepsilon_n^{-1}\, g^{(r)}(u)K((u-F(z))/\varepsilon_n)$$

$$- \frac{d}{du} \sum_{s=0}^{r-1} \varepsilon_n^{-(s+1)} g^{(r-s-1)}(u)K^{(s)}((u-F(z))/\varepsilon_n).$$

Hence

$$\sup_z \left|\varepsilon_n^{-(r+1)} \int_0^1 g(u)K^{(r)}((u-F(z))/\varepsilon_n)\, du\right|$$

$$\le \sup_z \left|\varepsilon_n^{-1} \int_0^1 g^{(r)}(u)K((u-F(z))/\varepsilon_n)\, du\right|$$

$$+ \sup_z \left| \left[\sum_{s=0}^{r-1} \varepsilon_n^{-(s+1)} g^{(r-s-1)}(u) K^{(s)}((u-F(z))/\varepsilon_n) \right]_{u=0}^{1} \right|$$

The second term above is zero for large n since the argument of $K^{(s)}$ is eventually outside the support of K. Write

$$\sup_z \left| \varepsilon_n^{-1} \int_0^1 g^{(r)}(u) K((u-F(z))/\varepsilon_n)\, du \right|$$

$$= \sup_z \left| \int_{-F(z)/\varepsilon_n}^{(1-F(z))/\varepsilon_n} K(v) g^{(r)}(\varepsilon_n v + F(z))\, dv \right|$$

$$\leq \sup_t |g^{(r)}(t)| \int |K(v)|\, dv < \infty . \qquad \square$$

We now proceed with the proof of Lemma 2.1. It is convenient to rewrite

$$M_n(u) = \varepsilon_n^{-1} \iint y K((F_n(x) - F_n(u))/\varepsilon_n)\, dF_n(x,y),$$

and similarly for M_n^* and M_n^{**}. Thus, letting $Z_n(x,y) = F_n(x,y) - F(x,y)$, we may write

$$M_n^*(u) - M_n(u)$$

$$= \varepsilon_n^{-1} \iint y \left[K\left(\frac{F_n(x)-F(u)}{\varepsilon_n} \right) - K\left(\frac{F_n(x)-F_n(u)}{\varepsilon_n} \right) \right] dZ_n(x,y)$$

$$+ \varepsilon_n^{-1} \iint y \left[K\left(\frac{F_n(x)-F(u)}{\varepsilon_n} \right) - K\left(\frac{F_n(x)-F_n(u)}{\varepsilon_n} \right) \right] dF(x,y)$$

$= J_1 + J_2$, say. We first show $(n\varepsilon_n \log n)^{\frac{1}{2}} |J_2| \xrightarrow{P} 0$. Since, by assumption, K has 3 continuous derivatives, we may write (by expanding $K((F_n(x) - F_n(u))/\varepsilon_n)$ about $(F_n(x) - F(u))/\varepsilon_n$)

$$J_2 = \varepsilon_n^{-2} [F_n(u) - F(u)] \iint y K' \left(\frac{F_n(x)-F(u)}{\varepsilon_n} \right) dF(x,y)$$

$$+ \varepsilon_n^{-3} [F_n(u) - F(u)]^2 \iint y K'' \left(\frac{F_n(x)-F(u)}{\varepsilon_n} \right) dF(x,y)$$

$$+ \varepsilon_n^{-4} [F_n(u) - F(u)]^3 \iint yK''' \left(\frac{F_n(x) + w_n(u)}{\varepsilon_n} \right) dF(x,y)$$

$$= J_2^{(1)} + J_2^{(2)} + J_2^{(3)}, \text{ say, where } w_n(u) \text{ is between } F_n(u) \text{ and } F(u).$$

Now, expanding $K' \left(\frac{F_n(x) - F(u)}{\varepsilon_n} \right)$ about $(F(x) - F(u))/\varepsilon_n$ yields

(A1) $$(n\varepsilon_n \log n)^{\frac{1}{2}} \sup_u |J_2^{(1)}|$$

$$\leq (n\varepsilon_n \log n)^{\frac{1}{2}} \varepsilon_n^{-2} \sup_u |F_n(u) - F(u)|$$

$$\times \left\{ \left| \iint yK' \left(\frac{F(x) - F(u)}{\varepsilon_n} \right) dF(x,y) \right| \right.$$

$$+ \left| \iint \left[\frac{F_n(x) - F(x)}{\varepsilon_n} \right] yK'' \left(\frac{F(x) - F(u)}{\varepsilon_n} \right) dF(x,y) \right|$$

$$\left. + \left| \iint \left[\frac{F_n(x) - F(x)}{\varepsilon_n} \right] yK''' \left(\frac{v_n(x,a)}{\varepsilon_n} \right) dF(x,y) \right| \right\}$$

where $v_n(x,u)$ is between $F_n(x) - F(u)$ and $F(x) - F(u)$.

Using the fact that $\sup_u |F_n(u) - F(u)| = O_p(n^{-\frac{1}{2}})$ and applying Lemma A1 implies that the first term on the RHS of inequality A1 goes to zero. For the second term, note that

$$\varepsilon_n^{-1} \iint \left| yK'' \left(\frac{F(x) - F(u)}{\varepsilon_n} \right) \right| dF(x,y)$$

$$= \varepsilon_n^{-1} \int_0^1 h(t) \left| K'' \left(\frac{t - F(u)}{\varepsilon_n} \right) \right| dt$$

$$= \int_{-F(u)/\varepsilon_n}^{(1-F(u))/\varepsilon_n} |K''(v)| h(\varepsilon_n v + F(u)) \, dv,$$

which is a bounded sequence since h is bounded and K'' has bounded supports.

Thus the second term on the RHS of (A1) is equal to

$(n\varepsilon_n \log n)^{\frac{1}{2}} \varepsilon_n^{-2} O_p(n^{-1}) O(1)$, which converges to zero in probability if $(n\varepsilon_n \log n)^{\frac{1}{2}}/n\varepsilon_n^2 \to 0$, i.e., if $(n\varepsilon_n^3)(\log n)^{-1} \to \infty$, which is true by assumption. For the third term on the RHS of (A1) note

$$\int \left| yK''' \left(\frac{v_n(x,u)}{\varepsilon_n} \right) \right| dF(x,y)$$

$$\leq \sup_v |K'''(v)| E|Y| < \infty.$$

Thus the third term is a $(n\varepsilon_n \log n)^{\frac{1}{2}} \varepsilon_n^{-4} O_p(n^{-3/2})$ sequence, and converges to zero in probability since $(\log n)^{-1} n\varepsilon_n^{7/2} \to \infty$. Similar arguments apply to $J_2^{(2)}$ and $J_2^{(3)}$, and we have shown $(n\varepsilon_n \log n)^{\frac{1}{2}} \sup |J_2| \xrightarrow{P} 0$.

We now turn to J_1. Let $\{a_n\}$ be a sequence as specified in the hypotheses and write

$$J_1 = \varepsilon_n^{-1} \int_{|y|>a_n} \int yG_n(x,u) Z_n(dx,dy)$$

$$+ \varepsilon_n^{-1} \int_{|y|\leq a_n} \int yG_n(x,u) Z_n(dx,dy)$$

$$= J_1^{(1)} + J_1^{(2)}, \text{ say, where, for convenience, we write}$$

$$G_n(x,u) = K\left(\frac{F_n(x) - F(u)}{\varepsilon_n} \right) - K\left(\frac{F_n(x) - F_n(u)}{\varepsilon_n} \right)$$

Using integration by parts, write

$$J_1^{(2)} = \varepsilon_n^{-1} \int_{|y|\leq a_n} \int Z_n(x,y)\, dyG_n(dx,u)$$

$$+ \lim_{t\to\infty} \varepsilon_n^{-1} \int_{-a_n}^{a_n} G_n(t,u)\, yZ_n(t,dy)$$

$$- \lim_{t\to-\infty} \varepsilon_n^{-1} \int_{-a_n}^{a_n} G_n(t,u)\, yZ_n(t,dy)$$

$$+ \varepsilon_n^{-1} a_n \int Z_n(x,a_n)\, G_n(dx,u)$$

$$+ \varepsilon_n^{-1} a_n \int Z_n(x,-a_n)\, G_n(dx,u)$$

$$= I_1 + I_2 + I_3 + I_4 + I_5, \text{ say.}$$

Since $Z_n(-\infty,y) = 0$ for each n and y, it is easily ascertained that $I_3 = 0$ for each n (e.g. Natanson (1964), p 233). Similarly,

$$I_2 = I_2(u) = \lim_{t\to\infty} G_n(t,u)\, \varepsilon_n^{-1} \int_{-a_n}^{a_n} y dQ_n(y)$$

where

$$Q_n(y) = \lim_{t\to\infty} Z_n(t,y) = F_n^{Y}(y) - F^{Y}(y).$$

Now

$$\int_{-a_n}^{a_n} y dZ_n(y) = n^{-1} \sum_{i=1}^{n} \left\{ Y_i I_{[-a_n,a_n]}(Y_i) - EYI_{[-a_n,a_n]}(Y) \right\} = O_p(n^{-\frac{1}{2}})$$

as $n \to \infty$ by standard central limit theorem arguments. Further, using the mean value theorem,

$$\lim_{t\to\infty} G_n(t,u) = K\left(\frac{1 - F(u)}{\varepsilon_n} \right) - K\left(\frac{1 - F_n(u)}{\varepsilon_n} \right)$$

$$= \frac{F_n(u) - F(u)}{\varepsilon_n} K'\left(\frac{1 + q_n(u)}{\varepsilon_n} \right) = \varepsilon_n^{-1} O_p(n^{-\frac{1}{2}})$$

uniformly in u, where $q_n(u)$ is between $F_n(u)$ and $F(u)$.

Thus we have

$$(n\varepsilon_n \log n)^{\frac{1}{2}} \sup_n |I_2(u)| = (n\varepsilon_n \log n)^{\frac{1}{2}} \varepsilon_n^{-2} O_p(n^{-1}) \to 0$$

since $n\varepsilon_n^3/\log n \to \infty$.

For I_4, note that

$$\left|\int Z_n(x,a_n)\, G_n(dx,u)\right|$$

$$\leq \sup_x |Z_n(x,a_n)|\, V[G_n(\cdot\,, u)],$$

Where V[ ] denotes total variation over R. Now

$$\sup_x |Z_n(x,a_n)| = O_p(n^{-\frac{1}{2}})$$

and it is easily verified, using the mean value theorem, that

$$V[G_n(\cdot\,, u)] = \varepsilon_n^{-1} O_p(n^{-\frac{1}{2}})$$

uniformly in u. Thus

$$(n\varepsilon_n \log n)^{\frac{1}{2}} \sup_u |I_4(u)|$$

$$= a_n (n\varepsilon_n \log n)^{\frac{1}{2}} \varepsilon_n^{-2} O_p(n^{-1}) \xrightarrow{P} 0$$

since $a_n^2 \log n/n\varepsilon_n^3 \to 0$ by assumption. A similar argument applies to show

$$(n\varepsilon_n \log n)^{\frac{1}{2}} \sup_u |I_5(u)| \xrightarrow{P} 0.$$

For I_1, note that

$$\left|\int_{|y|\leq a_n} \int Z_n(x,y)\, dy G_n(dx,u)\right|$$

$$\leq \sup_{x,y} |Z_n(x,y)| V[yG_n(x,u)] ,$$

where V denotes here the total variation in (x,y) over $R \times [-a_n, a_n]$. As before,

$$\sup_{x,y} |Z_n(x,y)| = O_p(n^{-\frac{1}{2}})$$

and

$$V[yG_n(x,u)] = a_n \varepsilon_n^{-1} O_p(n^{-\frac{1}{2}}) \text{ uniformly in } u.$$

Thus

$$(n\varepsilon_n \log n)^{\frac{1}{2}} \sup_u |I_1(u)|$$

$$= a_n \varepsilon_n^{-2} (n\varepsilon_n \log n)^{\frac{1}{2}} O_p(n^{-1}) \xrightarrow{P} 0$$

since $a_n^2 \log n / n\varepsilon_n^3 \to 0$ by assumption.

As the final step in the proof, we must verify that $(n\varepsilon_n \log n)^{\frac{1}{2}} \sup_u |J_1^{(1)}| \xrightarrow{P} 0$. Note that

$$\text{(A2)} \qquad \varepsilon_n |J_1^{(1)}| \le \left| \int_{|y|>a_n} \int yG_n(x,u)\, dF_n(x,y) \right|$$

$$+ \left| \int_{|y|>a_n} \int yG_n(x,u)\, dF(x,y) \right|.$$

For the first term, note

$$\left| \int_{|y|>a_n} \int yG_n(x,u)\, dF_n(x,y) \right|$$

$$\le \sup_{x,u} |G_n(x,u)| \int_{|y|>a_n} |y|\, dF_n^Y(y).$$

As before,

$$\sup_{x,u} |G_n(x,u)| = \varepsilon_n^{-1} O_p(n^{-\frac{1}{2}}),$$

and

$$\int_{|y|>a_n} |y| \, dF_n^Y(y) = n^{-1} \sum_{i=1}^{n} |Y_i| \, I_{(a_n,\infty)}(|Y_i|) .$$

Now, by the Markov inequality, for any $\varepsilon > 0$

$$P\left\{ \left| \sqrt{n} \int_{|y|>a_n} |y| \, dF_n^Y(y) \right| > \varepsilon \right\}$$

$$\leq \varepsilon^{-1} E \left| \sqrt{n} \int_{|y|>a_n} |y| dF_n^Y(y) \right|$$

$$= \sqrt{n} \int_{|y|>a_n} |y| \, dF^Y(y) \to 0$$

by assumption, and thus

$$\int_{|y|>a_n} |y| \, dF_n^Y(y) = O_p(n^{-\frac{1}{2}}).$$

A similar argument applies to the second integral on the RHS of (A2) and we thus have

$$(n\varepsilon_n \log n)^{\frac{1}{2}} \sup_u |J_1^{(1)}(u)|$$

$$(n\varepsilon_n \log n)^{\frac{1}{2}} \varepsilon_n^{-2} O_p(n^{-1}) \xrightarrow{P} 0$$

since $n\varepsilon_n^3/\log n \to \infty$ by assumption. □

The proof of (2.2) follows a similar pattern, and we omit the details.

ACKNOWLEDGEMENT

The author thanks R. J. Carroll for many helpful suggestions.

REFERENCES

Bhattacharya, P. K. (1967), "Estimation of a probability density function and its derivatives", Sankhya, Series A, 29, pp 373-382.

Bickel, P. J. and Rosenblatt, M. (1973), "On some global measures of the deviation of density function estimates", Ann Math Statist 1, pp 1071-1095.

Johnston, G. J. (1979), Ph.D. Dissertation, UNC at Chapel Hill, Dept of Statistics, unpublished.

Natanson, I. P. (1964), "Theory of functions of a real variable", Vol I, Ungar.

Watson, G. S. (1964), "Smooth regression analysis", Sankhya, Series A, 26, pp 359-372.

Watson, G. S. and Leadbetter, M. R. (1964), "Hazard analysis II", Sankhya, Series A, 26, pp 101-116.

Yang, S. S. (1977), "Linear functions of concomitants of order statistics", Technical Report No 7, Dept of Math, MIT.