The Annals of Mathematical Statistics
1972, Vol. 43, No. 1, 84-88

JOINT ASYMPTOTIC DISTRIBUTION OF THE ESTIMATED REGRESSION FUNCTION AT A FINITE NUMBER OF DISTINCT POINTS

By EUGENE F. SCHUSTER

University of Texas at El Paso

As an approximation to the regression function m of Y on X based upon empirical data, E.A. Nadaraya and G.S. Watson have studied estimates of m of the form $m_n(x) = \sum Y_i k((x - X_i)/a_n)/\sum k((x - X_i)/a_n)$. For distinct points $x_1 \cdots, x_k$, we establish conditions under which $(na_n)^{\frac{1}{2}}(m_n(x_1) - m(x_1), \cdots, m_n(x_k) - m(x_k))$ is asymptotically multivariate normal.

1. Introduction and summary. Let (X, Y) be a bivariate random variable having a joint density function f and let g be the marginal density function of X. If $E_f Y$ is finite then the regression function m (of Y on X) may be defined as $m(x) = E[Y \mid X = x]$. As an approximation to m based upon empirical data, Nadaraya (1964) and Watson (1964) have considered estimates of the form

$$m_n(x) = \sum_{i=1}^n Y_i k((x - X_i)/a_n)/\sum_{i=1}^n k((x - X_i)/a_n)$$

where k is a univariate density function, $\{a_n\}$ is a sequence of positive numbers converging to zero and $(X_1, Y_1), \cdots, (X_n, Y_n)$ is a random sample of size n from f.

Nadaraya (1964) indicates that if Y is a bounded random variable and $na_n^2 \to \infty$, then $(na_n)^{\frac{1}{2}}(m_n(x) - Em_n(x))$ is asymptotically normal with mean zero and variance $E[Y^2 \mid X = x] \int k^2(u)\, du/g(x)$.[1] Of more interest is the asymptotic distribution of $(na_n)^{\frac{1}{2}}(m_n(x) - m(x))$. One would normally approach the asymptotic distribution of this statistic by attempting to establish that $(na_n)^{\frac{1}{2}}(Em_n(x) - m(x)) = o(1)$, from which one could conclude that the asymptotic distribution of $(na_n)^{\frac{1}{2}}(m_n(x) - m(x))$ is the same as that of $(na_n)^{\frac{1}{2}}(m_n(x) - Em_n(x))$. Instead, we approach the problem by proving that both the numerator and denominator of m_n are asymptotically normal and hence so is m_n. In fact for distinct points $x_1, \cdots, x_k$ we will establish conditions under which $(na_n)^{\frac{1}{2}}(m_n(x_1) - m(x_1), \cdots, m_n(x_k) - m(x_k))$ is asymptotically multivariate normal with mean vector zero and diagonal covariance matrix $C = [C_{ij}]$ with $C_{ij} = \text{Var}[Y \mid X = x_i] \int k^2(u)du/g(x_i)$. We note that this asymptotic variance disagrees with that of Nadaraya unless $m(x_i) = 0$. The fact that Nadaraya's result is incorrectly stated can be observed by noting that the

Received November 6, 1970.

[1] Whenever the integration extends over $(-\infty, \infty)$ no limits of integration will be given.

asymptotic distribution should be invariant with respect to translations of Y. If one applies Nadaraya's theorem to bivariate samples from each of (X, Y) and $(X, Y - c)$, for constant c, one obtains different asymptotic variances for the same statistic.

2. Statement of the theorem. We assume the kernel k and the sequence $\{a_n\}$ are chosen to satisfy the conditions:

(i) $k(u)$ and $|uk(u)|$ are bounded.
(ii) $\int uk(u)\,du = 0$.
(iii) $\int u^2 k(u)\,du < \infty$.
(iv) $\lim na_n^3 = \infty$ and $\lim na_n^5 = 0$.

For convenience we write $V[Y \mid X = x] = v(x)/g(x) - w^2(x)/g^2(x)$, where $g(x)$, $w(x)$, and $v(x)$ are defined by $\int f(x, y)\,dy$, $\int yf(x, y)\,dy$, and $\int y^2 f(x, y)\,dy$, respectively.

THEOREM. *Suppose $x_1, \cdots, x_k$ are distinct points and $g(x_i) > 0$ for $i = 1, 2, \cdots k$. If $E_f Y^3$ is finite and if g', w', v', g'' and w'' exist and are bounded, then $(na_n)^{\frac{1}{2}}(m_n(x_1) - m(x_1), \cdots, m_n(x_k) - m(x_k))^t$ converges in distribution to Z^* where Z^* is multivariate normal with mean vector 0 and diagonal covariance matrix $C = [C_{ij}]$ where*

$$C_{ii} = V[Y \mid X = x_i] \int k^2(u)du/g(x_i) \qquad (i = 1, 2, \cdots, k).$$

3. Proof of theorem. For simplicity we shall prove the theorem for the special case when $k = 2$. The method of proof remains valid in the more general case.

For brevity we define for $i = 1, 2, \cdots, n$ and $s = 1, 2,$:

$$U^*_{ni}(x_s) = k((x_s - X_i)/a_n)/a_n\,, \qquad U_{ni}(x_s) = (a_n)^{\frac{1}{2}}(U^*_{ni}(x_s) - EU^*_{ni}(x_s))\,,$$
$$V^*_{ni}(x_s) = Y_i\, U^*_{ni}(x_s)\,, \qquad V_{ni}(x_s) = (a_n)^{\frac{1}{2}}(V^*_{ni}(x_s) - EV^*_{ni}(x_s))\,,$$
$$U_n(x_s) = \textstyle\sum_{i=1}^n U_{ni}(x_s)\,, \qquad V_n(x_s) = \sum_{i=1}^n V_{ni}(x_s)\,,$$
$$W_{ni} = (U_{ni}(x_1), V_{ni}(x_1), U_{ni}(x_2), V_{ni}(x_2))\,,$$
$$(n)^{\frac{1}{2}}Z_n = (U_n(x_1), V_n(x_1), U_n(x_2), V_n(x_2))^t\,,$$
$$A = \int k^2(u)du \begin{bmatrix} g(x_1) & w(x_1) & 0 & 0 \\ w(x_1) & v(x_1) & 0 & 0 \\ 0 & 0 & g(x_2) & w(x_2) \\ 0 & 0 & w(x_2) & v(x_2) \end{bmatrix}.$$

Let Z be fourvariate normal with mean vector 0 and covariance matrix A. We first prove two lemmas.

LEMMA 1. *Suppose the density k satisfies the conditions* (i) *and* (ii) *above and suppose $na_n^3 \to \infty$. Let $E_f|Y|^3$ be finite and let g', w', and v' exist and be bounded. If $x_1 \neq x_2$ and $g(x_i) > 0$ for $i = 1, 2$, then Z_n converges in distribution to Z.*

PROOF. Using the Cramér-Wold Theorem (e.g., Theorem (xi) on page 103 of [3]), it will be sufficient to prove that $c \cdot Z_n{}^t$ converges in distribution to $c \cdot Z^t$ for any $c = (c_1, d_1, c_2, d_2)$ in R^4.

The following hold for $s = 1, 2$ and $r = 1, 2$ under the assumption that $s \neq r$ whenever s and r appear in the same expression:

$$EU_{ni}^2(x_s) = g(x_s) \int k^2(u)\, du + O(a_n) . \tag{1}$$

$$EV_{ni}^2(x_s) = v(x_s) \int k^2(u)\, du + O(a_n) . \tag{2}$$

$$EU_{ni}(x_s)V_{ni}(x_s) = w(x_s) \int k^2(u)\, du + O(a_n) . \tag{3}$$

$$EU_{ni}(x_s)U_{ni}(x_r) = O(a_n) . \tag{4}$$

$$EV_{ni}(x_s)V_{ni}(x_r) = O(a_n) . \tag{5}$$

$$EU_{ni}(x_s)V_{ni}(x_r) = O(a_n) . \tag{6}$$

We will be sketch the proofs of (1) and (4) to illustrate the method. To obtain (1), we see

$$EU_{ni}^2(x_s) = a_n[\int k^2(u)\, g(x_s - a_n u)\, du/a_n - (\int k(u)\, g(x_s - a_n u)\, du)^2] .$$

Since g' and $|yk(y)|$ are bounded and $\int |u|k(u)\, du$ is finite, it follows that

$$|\int k(u)\{g(x_s - a_n u) - g(x_s)\}\, du| \leqq \sup_x |g'(x)| a_n \int |u|k(u)\, du = O(a_n)$$

and

$$|\int k^2(u)\{g(x_s - a_n u) - g(x_s)\}\, du| \leqq \sup_x |g'(x)| a_n \int |u|k^2(u)\, du = O(a_n) .$$

Thus we have

$$EU_{ni}^2(x_s) = g(x_s) \int k^2(u)\, du + O(a_n) .$$

As for (4), suppose $x_2 > x_1$. Let $\delta = x_2 - x_1$ and $\delta_n = \delta/a_n$. Then

$$\begin{aligned}
EU_{ni}(x_1)U_{ni}(x_2) &= \int k((x_1 - u)/a_n)k((x_2 - u)/a_n)g(u)\, du/a_n + O(a_n) \\
&= \int k(u)k(\delta_n + u)g(x_1 - a_n u)\, du + O(a_n) \\
&= \int_{|u|<\delta_n/2} k(u)k(\delta_n + u)g(x_1 - a_n u)\, du \\
&\qquad + \int_{|u|\geqq\delta_n/2} k(u)k(\delta_n + u)g(x_1 - a_n u)\, du + O(a_n) \\
&\leqq \sup_{|u|<\delta_n/2} k(\delta_n + u) \cdot \int k(z)g(x_1 - a_n z)\, dz \\
&\qquad + \sup_{|u|\geqq\delta_n/2} k(u) \cdot \int k(\delta_n + z)g(x_1 - a_n z)\, dz + O(a_n) \\
&\leqq \sup_{|u|>\delta_n/2} k(u) \cdot O(1) \\
&\qquad + \sup_{|u|\geqq\delta_n/2} k(u) \cdot \int k(z)g(x_2 - a_n z)\, dz + O(a_n) \\
&\leqq 2 \sup_{|u|\geqq\delta_n/2} k(u) \cdot O(1) + O(a_n) \leqq 4\delta_n{}^{-1} \\
&\qquad \times \sup_{|u|\geqq\delta_n/2} |uk(u)| \cdot O(1) + O(a_n) = O(a_n)
\end{aligned}$$

which was to be shown.

Now let $\sigma_n{}^2 = \text{Var}\,(c \cdot Z_n{}^t)$ so that by (1)–(6) above, we have

$$\sigma_n^2 = \int k^2(u)\,du \cdot \sum_{s=1}^{2} [c_s^2 g(x_s) + d_s^2 v(x_s) + 2c_s d_s w(x_s)] + O(a_n) .$$

Put $\rho_{ni}^3 = E\,|(c \cdot W_{ni})/n^{-\frac{1}{2}}|^3$ and $\rho_n^3 = \sum_{i=1}^{n} \rho_{ni}^3$ so that

$$\begin{aligned} \rho_n^3 &= n^{-\frac{1}{2}} E\,|c \cdot W_{n1}|^3 \leqq n^{-\frac{1}{2}}\,|c|^3\, E\,|W_{n1}|^3 \\ &\leqq 8n^{-\frac{1}{2}}\,|c|^3 \max_{s=1,2} \{E\,|U_{n1}(x_s)|^3,\ E\,|V_{n1}(x_s)|^3\} . \end{aligned}$$

Since g', w' v' and k are bounded and $E_f|Y|^3$ is finite it follows by arguments similar to those above that

$$E\,|U_{ni}(x_s)|^3 = O(a_n^{-\frac{1}{2}}) \qquad \text{and} \qquad E\,|V_{ni}(x_s)|^3 = O(a_n^{-\frac{3}{2}})$$

$(s = 1, 2)$ so that $\rho_n^3 = O(a_n^{-\frac{3}{2}} n^{-\frac{1}{2}})$.

Since $g(x)v(x) - w^2(x) = g^2(x)V[Y \mid X = x]$ we can deduce that A is positive definite whenever $g(x_1) > 0$ and $g(x_2) > 0$. Thus for $c \neq 0$

$$\lim_{n\to\infty} \sigma_n^2 = cAc^t > 0$$

since cAc^t is a quadratic form associated with the positive definite matrix A. Hence it follows that $\lim_{n\to\infty} \rho_n/\sigma_n = 0$ (recall that $na_n^3 \to \infty$) whenever $c \neq 0$.

An application of the Berry-Esséen Theorem on page 288 of [1] now completes the proof.

Let us write

$$\begin{aligned} Z_n^* = a_n^{\frac{1}{2}} n^{-\frac{1}{2}} \Big(&\sum_{i=1}^{n} [U_{ni}^*(x_1) - g(x_1)],\ \sum_{i=1}^{n} [V_{ni}^*(x_1) - w(x_1)], \\ &\sum_{i=1}^{n} [U_{ni}^*(x_2) - g(x_2)],\ \sum_{i=1}^{n} [V_{ni}^*(x_2) - w(x_2)]\Big)^t . \end{aligned}$$

LEMMA. 2. *Suppose* $\int uk(u)\,du = 0$, $\int u^2 k(u)\,du$ *is finite and* $na_n^5 \to 0$. *If* g'' *and* w'' *exist and are bounded then, under the conditions of Lemma 1,* Z_n^* *converges in distribution to* Z.

PROOF. Let $B_n = (g(x_1) - EU_{n1}^*(x_1),\ w(x_1) - EV_{n1}^*(x_2),\ g(x_2) - EU_{n1}^*(x_2),\ w(x_2) - EV_{n1}^*(x_2))^t$. Since $\int uk(u)\,du = 0$, $\int u^2 k(u)\,du$ is finite and g'' is bounded, it follows that

$$\begin{aligned} |EU_{n1}^*(x_i) - g(x_i)| &= |\int k(u)\{g(x_i - a_n u) - g(x_i)\}\,du| \\ &\leqq \sup_x |g''(x)|\, a_n^2 \int u^2 k(u)\,du/2 = O(a_n^2) \quad (i = 1, 2). \end{aligned}$$

Similarly $|EV_{n1}^*(x_i) - w(x_i)| = O(a_n^2)$ so that $B_n = O(a_n^2)$. Then $Z_n - Z_n^* = (na_n)^{\frac{1}{2}} B_n = O(na_n^5)^{\frac{1}{2}} = o(1)$ since $na_n^5 \to 0$. The desired result now follows from standard large sample theory and Lemma 1.

We are now in a position to complete the proof of the theorem. Let the function H from R^4 to R^2 be defined by

$$H(y_1, y_2, y_3, y_4) = (H_1(y_1, y_2, y_3, y_4),\ H_2(y_1, y_2, y_3, y_4))^t ,$$

where $H_1(y_1, y_2, y_3, y_4) = y_2/y_1$, and $H_2(y_1, y_2, y_3, y_4) = y_4/y_3$ and let $\theta = (g(x_1), w(x_1), g(x_2), w(x_2))$. Let us now write $Z_n^* = (na_n)^{\frac{1}{2}}(T_n - \theta)^t$ where $T_n = (T_{n1}, T_{n2}, T_{n3}, T_{n4})$ with

$$T_{n1} = \sum_{i=1}^{n} U_{ni}^{*}(x_1)/n\,, \qquad T_{n2} = \sum_{i=1}^{n} V_{ni}^{*}(x_1)/n\,,$$
$$T_{n3} = \sum_{i=1}^{n} U_{ni}^{*}(x_2)/n\,, \qquad T_{n4} = \sum_{i=1}^{n} V_{ni}^{*}(x_2)/n\,,$$

Then the Mann-Wald Theorem (e.g., Theorem (ii) on page 321 of [3]), with $(n)^{\frac{1}{2}}$ replaced replaced by $(na_n)^{\frac{1}{2}}$ may be applied, together with Lemma 2, to, conclude that $(na_n)^{\frac{1}{2}}(H(T_n) - H(\theta))$ converges in distribution to Z^* where Z^* is $N(O, DAD^t)$ and where D is the matrix of partial derivatives of H, evaluated at θ. It is readily verified that $DAD^t = C$, and that

$$H(T_n) - H(\theta) = (m_n(x_1) - m(x_1)\,,\, m_n(x_2) - m(x_2))^t$$

completing the proof.

Acknowledgment. The research reported here was performed as part of the Ph.D. dissertation of the author at the University of Arizona under the direction of Professor P.K. Bhattacharya, whose guidance and many suggestions are gratefully acknowledged.

REFERENCES

[1] LOÈVE, M. (1963). *Probability Theory* 3rd ed. Van Nostrand, Princeton.
[2] NADARAYA, E.A. (1964). On estimating regression. *Theor. Probability Appl.* **9** 141-142.
[3] RAO, C.R. (1965). *Linear Statistical Inference and Its Applications.* Wiley, New York.
[4] WATSON, G.S. (1964). Smooth regression analysis. *Sankhya. Ser. A* **26** 359-372.