CONSISTENT WINDOW ESTIMATION OF A NONLINEAR REGRESSION FUNCTION WITH MOMENT INEQUALITIES.

by

Clifford Spiegelman

FSU Statistics Report M413

May, 1977

THE FLORIDA STATE UNIVERSITY
DEPARTMENT OF STATISTICS
TALLAHASSEE, FLORIDA 32306
Abstract.

Under weak assumptions, Stone (1975a) has shown that a class of nearest neighbor estimates of a regression function have an integrated mean square error which converges to zero as the sample size increases to infinity. Under slightly stronger assumptions we show that a class of window estimates have the same property. The assumption of a pointwise smooth regression function is not used. We then give moment inequalities for the size of the integrated mean square error. As an application of these inequalities we show that rates of convergence for non-smooth functions such as paths of Brownian motion are easily obtainable.

Introduction and Model. Let \((\theta, X)\) be a random pair such that \(E\theta^2 < \infty\). For any estimate \(\delta(X)\) of \(\theta\) the statistician incurs loss \(L(\theta, \delta) = (\theta - \delta)^2\). The risk to the statistician is defined by \(R(\delta) = EL(\theta, \delta)\). It is well known that the Bayes estimator \(\delta^*(x) = E[\theta|x]\) has the property that \(R(\delta^*) = \min_{\delta} R(\delta)\).

Often the form of the joint distribution of \((\theta, X)\) is not known, even approximately, but from prior experience independent and identically distributed random variables \((\theta_i, X_i), i = 1, \ldots, n\) having the same distribution as \((\theta, X)\) are available. When the joint distribution of \((\theta, X)\) does not have a finite number of unknown parameters, the usual maximum likelihood procedures are inappropriate. Similar comments apply to ordinary least square procedures. However it has been shown by Benedetti (1975), Cover (1968), Johns (1957), Stone (1975a b) and Spiegelman (1976) that certain nonparametric estimates, \(\delta_n\), of \(\delta^*\) have the property that \(E(\delta_n - \delta^*)^2 \rightarrow 0\) as \(n \rightarrow \infty\).
Generally we may write $E(\delta_n - \theta)^2 = R(\delta^*) + E(\delta_n - \delta^*)^2$. Whence the estimates cited earlier are empirically Bayes optimal. The quantity $E(\delta_n - \delta^*)^2$ is of particular interest since it represents the extra expected loss we incur for not knowing the distribution of $(\theta, X)$.

Recently there has been much attention given to this problem. Some recent references are Benedetti (1975), Cover (1968), Stone (1975 a b), and Spiegelman (1976). Other references may be found in Stone (1975a). Stone (1975a) has shown that various classes of nearest neighbor type estimates are $L_q$ consistent under weak assumptions. It is shown here that a class of window estimates is $L_q$ consistent under slightly stronger conditions then those in Stone (1975a). The method of proving $L_q$ consistency uses a theorem in Stone (1975a).

Section 2. Results for window estimates. We now define a window estimate of a regression function. We let $b$ be a positive number and define $W_i = \begin{cases} 1 & \text{if } |x_i - x| < b, \\ 0 & \text{otherwise} \end{cases}$ our window estimate of $\delta^*$ is denoted by

$$\delta_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i W_i / 1 \vee \frac{1}{n} \sum_{i=1}^{n} W_i.$$  

$\delta_n$ is the average of those $\theta_i$ whose corresponding $x_i$ is within a distance of $b$ to $x$.

Most of what we derive here follows from the following elementary inequality. Let us take $g(x)$ such that $E|g(x)|^q < \infty$ for some $q \geq 1$.

Lemma 1. (Basic inequality). If $E|g(x)|^q < \infty$ for some $q \geq 1$ then

$$E \left[ \sum_{i=1}^{n} (g(x_i) - g(x)) W_i / 1 \vee \sum_{i=1}^{n} W_i \right]^q \leq E \left[ \left| \sum_{i=1}^{n} (g(x_i) - g(x)) W_i \right|^q \right] \left| W_i = 1 \right|.$$

Proof: From the Cauchy-Schwartz inequality we have
\[ E \left( \sum_{i}^{n} (g(x_i) - g(x)) W_i \right) / 1 \sum_{i}^{n} W_i \leq E \sum_{i}^{n} g(x_i) - g(x) \right| W_i / 1 \sum_{i}^{n} W_i. \]

By applying the operators \( E[|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, W_1, \ldots, W_n|] \) to this last term we get \( E \sum_{i}^{n} |g(x_i) - g(x)|^q W_i / 1 \sum_{i}^{n} W_i = E \sum_{i}^{n} E \left[ |g(x_i) - g(x)|^q |X, W_i = 1\right] W_i / 1 \sum_{i}^{n} W_i. \) Since \( E \left[ |g(x_i) - g(x)|^q |X, W_i = 1\right] \) does not depend on \( i, 1 \leq i \leq n, \) \( E \sum_{i}^{n} E \left[ |g(x_i) - g(x)|^q |X, W_i = 1\right] W_i / 1 \sum_{i}^{n} W_i \leq \sum_{i}^{n} |g(x_i) - g(x)|^q |X, W_i = 1\right]. \) Q.E.D.

Under mild conditions we will show that \( \delta \) is an \( L_q \) consistent estimate for \( \delta^* \). Choose \( S \) to denote the set of mass points of \( F_X(t) \), the distribution function of \( X \). If \( S = \phi \), the empty set, we denote the density of \( X \) by \( f(x) \), when it exists.

**Theorem 1 (L_q consistency).** If 1) \( E[|\delta|^q < \infty \) for some \( q \geq 1, 2) b \to 0, \)

\( \text{nb} \to \infty, \) as \( n \to \infty, \) and either 3a) \( S = \phi \) or 3b) \( ||f||_{\infty} < \infty \) hold then .

\( E[\delta-\delta^*|^q \to 0 \) as \( n \to \infty. \)

**Proof:** Let \( g(x) = E[\theta|x]\). By Lemma \( E \sum_{i}^{n} (g(x_i) - g(x)) W_i \)

\( 1 \sum_{i}^{n} W_i \leq \sum_{i}^{n} |g(x_1) - g(x)|^q W_i = 1\].

We wish to obtain an upper bound for this expression. Let us take \( \{U_i\}_{i=1}^{\infty} \) to be an interval partition of the line such that \( U_{i-1} \) lies to the left of \( U_i \).

**Case 1:** \( ||f||_{\infty} < \infty \). We then have \( \sum_{i}^{n} E \left[ |g(x)|^q |X, W_i = 1\right] = \sum_{i}^{n} \int |g(x)|^q P(dx_1) \) P(dx_1). Since \( \{X \in U_{i-1}, W_i = 1\} \subset \{X \in U_1, X_1 \in U_1 \} \) the \( \sum_{i}^{n} \frac{P(W_i=1)}{P(W_i=1)} \) last term is bounded above by \( 3||f||_{\infty} E|g(x)|^q b/EW_1. \)

In order to bound \( 3||f||_{\infty} E|g(x)|^q b/EW_1 \) from above, it is necessary to bound \( EW_1 \) from below. We now proceed to do this. Clearly \( W_1 \geq \sum_{i}^{\infty} I_{U_1}(X_1)I_{U_i}(X) \)

so that \( EW_1 \geq \sum_{i}^{\infty} P^2(U_i) \). Let \( [[]] \) denote the greatest integer function. By partitioning the intervals \( (i, i+1] \) into \( [[1/b]] \)
subintervals we obtain \( \sum_{-\infty}^{\infty} \mathbb{P}^2(U_i) \geq b \sum_{i=-\infty}^{\infty} \mathbb{P}^2(i < x \leq i + 1) \). Whence

\[
1) \quad E[|g(x)|^q | W_1 = 1] \leq C_0 \ E|g(x)|^q,
\]

where \( C_0 \) is a finite constant.

Case 2: \( S \neq \emptyset \): By straightforward calculation we have

\[
E[|g(x)|^q | W_1 = 1] \leq \int \int \mathbb{P}(dx_1) \mathbb{P}(dx_2) / \sum_{t \in \mathbb{S}} (F(x(t)) - F(x(t_-)))^2. \quad \text{If } \{W_1 = 1\} \text{ is replaced by } \mathbb{R}^2 \text{ then we obtain}
\]

\[
2) \quad \int \int |g(x)|^q \mathbb{P}(dx_1) \mathbb{P}(dx_2) / \sum_{t \in \mathbb{S}} (F(x(t)) - F(x(t_-)))^2 \leq C_0 \ E|g(x)|^q,
\]

where \( C_0 \) is a constant that does not depend on \( g \). From this point on the proof is straightforward, and the details follow along the line of the proof of theorem 2 in Stone (1975).

We remark that these bounds allow \( g \) to be approximated by continuous functions with compact support, and our results for such functions are known (for example see Rosenblatt (1969)). Q.E.D.

Comment 1. For the case \( q=2 \), the reader may obtain the rest of the details of this proof by using the bounds calculated in theorem 2.

Comment 2. The following example was provided by J. Sacks and shows that the technique used in proving theorem 1 is not useful when \( \|f\|_{\infty} = \infty \).

Let \( \delta^*(x) = \mathbb{P}^q(x) \). Then \( E[\delta^*(x) | W_1 = 1] = \int f(x_1) \mathbb{P}^{q+1}(x) \ dx \ dx_1 / \{W_1 = 1\} \)

\[
\int f(x_1) \mathbb{P}^q(x) \ dx \ dx_1
\]

Suppose \( f(x) \) has \( \|f'(x)\|_{\infty} < \infty \). Then after dividing the top and bottom of the above expression by \( b \) and letting
b=0 we have \( \lim_{b \to 0} E[\delta^*(x)|W_1 = 1] = \int f^{q+2}(x) \, dx / \int f^2(x) \, dx. \) Then if \( f \in L_{q+1} \)

but not in \( L_{q+2} \) for all \( q > 1 \) we cannot hope to obtain a bound of the form

\[
E[\delta^*(x)|W_1 = 1] \leq C_0 \, E \delta^*(x).
\]

Theorem 2 (Fundamental inequality.) \( E(\delta_n - \delta^* )^2 \leq 20 \max \left( E[(\delta^*(x_1) - \\
\delta^*(x))^2 | W_1 = 1], \int_0^\infty (E[\theta_1^2 | W_1 = 1, x = t] / 1 V n \, P[1|X_1 - t| < b])P(dt) \right). \)

Amplification: Under the assumptions of theorem 1 both terms on the right hand side of this inequality converge to zero as \( n \to \infty \). Although this inequality looks complicated, various interpretable inequalities will follow from it.

Proof of theorem 2: By the \( C_1 \) inequality (Loeve (1963)) we have

\[
3) \quad E(\delta - \delta^*)^2 \leq 4 \left( E[\Sigma_1^n (\delta^*(x_1) - \delta^*(x))W_1 / 1 V \Sigma_1^n W_1] \right)^2 + \]

\[
E[\Sigma_1^n (\theta_1 - \delta^*(X_1))W_1 / 1 V \Sigma_1^n W_1]^2 + E\delta^2(x)I_{\{0\}}( \Sigma_1^n W_1 ).
\]

From lemma 1 \( E[\Sigma_1^n (\delta^*(x_1) - \delta^*(x))W_1 / 1 V \Sigma_1^n W_1]^2 \leq E[(\delta^*(X_1) - \delta^*(x))^2 | W_1 = 1]. \) By straightforward calculation on the second term on the right hand side of 3) we get

\[
E[\Sigma_1^n (\theta_1 - \delta^*(X_1))W_1 / 1 V \Sigma_1^n W_1]^2 = E[\Sigma_1^n E[(\theta_1 - \delta^*(X_1))^2 | W_1 = 1, X]W_1 / (1 V \Sigma_1^n W_1)]^2. \quad \text{For } i = 1, \ldots, n, E[(\theta_1 - \delta^*(X_1))^2 | W_1 = 1, X] = E[(\theta_1 - \delta^*(X_1))^2 | W_1 = 1, X]. \] Hence \( E[(\Sigma_1^n E[(\theta_1 - \delta^*(X_1))^2 | W_1 = 1, X]W_1 / (1 V \Sigma_1^n W_1)]^2 \leq E(E[(\theta_1 - \delta^*(X_1))^2 | W_1 = 1, X] / 1 V \Sigma_1^n W_1). \)

In order to deal with this last term we need to take a technical digression.
Let $U = (X-b, X+b)$. Given $X$, $\Sigma^m_1 W_1 = \Sigma^m_1 I_U(x_1)$ is $b(n,p)$ where $p = P(x \in U)$. By straightforward calculation $E[1 \ V \ \Sigma^m_1 W_1]^{-1} \leq 2 \ E_0^{n}(1/j+1)^{\frac{n}{j}} p^j (1-p)^{n-j} \leq 2[p(n+1)]^{-1}$. By application of this inequality and the Cauchy-Schwarz inequality we have $E[((\theta_1 - \delta^*(x_1))^2|W_1 = 1, X = t]/1 \ V \ \Sigma^m_1 W_1) \leq 2 \int_{-\infty}^{\infty} (E[\theta_1^2|W_1 = 1, X = t]) \ P_X(dt)$. Finally the term $E \delta^2(x) I_{\{0\}}(\Sigma^m_1 W_1)$ is dealt with. We have $E \delta^2(x) I_{\{0\}}(\Sigma^m_1 W_1) = E \delta^2(x) (1 \ V \ \Sigma^m_1 W_1 - \Sigma^m_1 W_1/1 \ V \ \Sigma^m_1 W_1)$. This last term is less than or equal to $E \delta^2(1 \ V \ \Sigma^m_1 W_1)^{-1}$. By the same arguments as before $E \delta^2(1 \ V \ \Sigma^m_1 W_1)^{-1} \leq 2 \int_{-\infty}^{\infty} (E[\theta_1^2|W_1 = 1, X = t]) \ P_X(dt)$. Q.E.D.

Some moment inequalities are now given for the expression $E[((\delta^*(x_1) - \delta^*(x))^2|W_1 = 1]$. Define $W' = \begin{cases} W_1 & \text{if } x = x_1 \\ 0 & \text{otherwise.} \end{cases}$

Lemma 2. If $||\delta^*||_{\infty} = 1$ and $S \neq \phi$ then $E[((\delta^*(x_1) - \delta^*(x))^2|W_1 = 1] \leq 4EW'/\int_{t \in \mathbb{S}} F(t) - F(t)^{-2}$.

Proof: By definition $E[((\delta^*(x_1) - \delta^*(x))^2|W_1 = 1] = \int_{\{W_1 = 1\}} (\delta^*(x_1) - \delta^*(x))^2 P_X(dx_1) P_X(dx)/EW_1$. By assumption $||\delta^*(x_1) - \delta^*(x)||_{\infty} \leq 2$ and $EW_1 \geq \int_{t \in \mathbb{S}} (F(t) - F(t)^-) > 0$. The result follows by applying these two bounds. Q.E.D.

Suppose that $(\delta^*(x_1) - \delta^*(x))^2 \leq k(x)\Delta(|x_1-x|)$ where $\Delta(0) = 0$ and $\Delta$ is a monotonically nondecreasing function on the positive half line. Then we have:

Lemma 3: $E[((\delta^*(x_1) - \delta^*(x))^2|W_1 = 1] \leq \begin{cases} 3Ek(x)\Delta(b)(\frac{\Sigma_{t \in \mathbb{S}} (F(t) - F(t^-))^2}{\int_{t \in \mathbb{S}} F(t^-)^2})^{-1} & \text{if } S \neq \phi \\ 3||f||_{\infty} b\Delta(b)(EW_1)^{-1} & \text{if } X \text{ has a density } f. \end{cases}$
Proof: By supposition we have \( E[(\delta^*(x) - \delta^*(x_1))^2 | W_1 = 1] \leq \Delta(b) \) \( E[k(x) | W_1 = 1] \). The result now follows from the bounds obtained in the proof of Theorem 1; (see inequalities 1) and 2).

We now deal with the second term of our fundamental inequality.

Lemma 3: If \( ||\theta||_\infty < \infty \), and \( E|x|^{1+t} = 1 \) for some \( t > 0 \) then
\[
\int_0^\infty (E[\theta^2 | W_1 = 1, x = t]) / 1 V n P(|x_1 - t| < b) \) \( P_x(dt) \leq ((2s_0)^n + 1/b^{1+t} \int_0^\infty \lambda^{1+t}d\lambda) \) where \( s_0 = b^{-1}n^{-1/1+t} \).
\]

Proof: Let us take \( \{U_i\}_{i=-\infty}^\infty \) to be an interval partition of the real line such that each \( U_i \) has length \( b \), and \( U_{i-1} \) lies to the left of \( U_i \).

Since \( E[\theta^2 | W_1 = 1, x = t] \leq 1 \), we have \( \int_0^\infty (E[\theta^2 | W_1 = 1, x = t]) / 1 V n P(|x_1 - t| < b) \) \( P_x(dt) \leq \int_0^\infty P_x(U_i^*) / 1 V P_x(U_i) \).

Choose \( \{U_i^*\}_{i=1}^\infty \) to be a rearrangement of \( \{U_i\}_{i=-\infty}^\infty \) such that \( P(U_i^*) \geq P(U_i^*) \). If \( X^* \) is a random variable such that \( P((i-1)b << x^* \leq ib) = P_x(U_i^*) \), then \( E|x|^t = 1 \). By straightforward calculation
\[
1 \geq E|x|^t \geq \int_{U_i} |x|^t dF_X(x) \geq ((i-1)b)^{1+t} P_x(U_i^*). \]
Clearly a/1 V na \( \leq 1/n \) for \( a \geq 0 \). Whence we have \( \int_0^\infty P_x(U_i^*) / 1 V P_x(U_i) = \int_0^\infty P_x(U_i^*) / 1 V P_x(U_i^*) \).

By calculation we obtain \( \int_0^\infty 1/n + \int_{s_0}^{\infty} P(U_i^*) \leq 2s_0/n + 1/b^{1+t} \int_0^\infty \lambda^{1+t}d\lambda \).

The minimum of this last expression is attained at \( s_0 = n^{-1/1+t} b^{-1} \).

Comment 2. Other bounds similar in form to lemma 3 are obtainable when \( E|\theta|^q < \infty \), \( q > 2 \). They follow by straightforward application of the Cauchy Schwartz inequality and Lemma 3.

Example 1: Let \( \delta^*(x) = x \), and \( \theta + Q \), where \( X \) has a density \( f \) bounded below by a positive number \( \eta \) and \( ||f||_\infty < \infty \). We suppose that \( Q \) is independent of \( X \) and \( ||\theta||_\infty < \infty \). It can be shown, by applying Theorem 2,
that the best choice of $b$ is $b = 0(n^{-1/3})$. For this choice $E(\delta_n - \delta)^2 = O(n^{-2/3})$.

Example 2: Let $D(t)$ be Brownian motion on $[0,1]$, and $d(t)$ be any fixed path of $D(t)$. It is known that $d(t)$ is non-differentiable a.e.. Let $\delta^*(x) = d(x)$ and $\theta = \delta^*(x) + \Omega$. We suppose that $Q$ and $D$ are independent of $X$, and $X$ has a density $f$ which is bounded below by a positive number $n$ and $||f||_\infty < \infty$. We also assume that $||\theta||_\infty < \infty$. It can be shown that the best choice of $b$ is $b = 0(n^{-1/4})$ and $E(\delta_n - \delta)^2 = O(n^{-1/2})$. This follows by first applying Fubini's theorem to the term $b E \left[ \frac{|D(x_1) - D(x_2)|^2}{|x_1 - x|} \mid W_1 = 1 \right]$ and then applying Theorem 2.

Section 3. Local least square estimates. (Stone 1975b) has shown that under many situations we would be better off by estimating $\delta^*$ with the linear least square estimate of $\delta^*(x)$ obtained from the nearest neighbors of $X$. We now extend our results to include linear least square estimates of $\delta^*(x)$ formed from observations having $W_i = 1$.

Define $\overline{X} = \sum_1^n X_i W_i / 1 + \sum_1^n W_i$. Define $k_i$ to be the number of the observations $X_1, \ldots, X_n$ between $X$ and $X + (-1)^i b$, $i = 1, 2$. If $k_1 k_2 \geq 1$ let $\delta_n$ be the linear least square estimate of $\delta^*$, otherwise take $\delta_n = 0$. Define $M(x_i) = a + c(x_i - \overline{x})$ to be the projection of $\delta^*$ on $1$ and $x_i$ with respect to $P(dX_1 \mid W_1 = 1, X)$. Define $\sigma^2(x_1) = E[(\theta_1 - m(x_1))^2 \mid X_1, X]$.

We suppose that $\sigma^2(x_1)$ is locally bounded by an integrable function $k(x)$, i.e. for some integrable function $k(x)$, we have $\sigma^2(x_1)W_i \leq k(X)W_i$. Under the additional assumptions of theorem 1 we have:

**Theorem 5:** $E(\delta_n(x) - \delta^*(x))^2 \to 0$ as $n \to \infty$.

**Proof:** We have $E(\delta_n - m)^2 \leq 2E(\hat{a} - a)^2 + 2E(x - \overline{x})^2 k(x) / \sum_1^n W_i$. Clearly $(x - \overline{x})^2 / \sum_1^n W_i \leq 1/k_1 k_2$. Since $k_1$ and $k_2 \to \infty$ as $n \to \infty$, $E(\delta_n - m(x))^2 \to 0$ as $n \to \infty$. 

We must show that $E(\delta(x) - m(x))^2 \to 0$ as $n \to \infty$. In order to achieve this let $X_1$ be a random variable having distribution $P(dx_1|W_1 = 1, x)$ given $x$.

Then

$$E(\delta(x) - m(x))^2 \leq 2E(\delta^*(x) - \delta^*(x_1))^2 + 2E(\delta^*(x_1) - m(x))^2.$$  

Now

$$E[(\delta^*(x) - \delta^*(x_1))^2] = E[E[(\delta^*(x) - \delta^*(x_1))^2|W_1 = 1, x]] = E[(\delta^*(x) - \delta^*(x_1))^2|W_1 = 1].$$

We bound $E(\delta^*(x_1) - m(x))^2$ above by $2E(\delta^*(x_1) - m(x))^2 + 2E(m(x_1) - m(x))^2$. From the definition of $m$ and $x_1$ we have $E(\delta^*(x_1) - m(x_1))^2 \leq E(\delta^*(x_1) - E[\delta^*(x_1)|x])^2$. Once again we add and subtract $\delta(x)$. By straightforward calculations $E(\delta^*(x_1) - E[\delta^*(x_1)|W_1 = 1, x])^2 \leq 4E[(\delta^*(x_1) - \delta^*(x))^2|W_1 = 1]$.

Finally we deal with the term $E(m(x_1) - m(x))^2$. From the definition of $x_1$, $E(m(x_1) - m(x))^2 = E[(m(x_1) - m(x))^2|W_1 = 1]$. By application of the pythagorean theorem $E[\delta^2(x_1)|W_1 = 1] \leq E[(\delta^*(x))^2|W_1 = 1]$. In particular this implies that $m$ is a continuous function of $\delta^*$ with respect to the $L_2$ norm.

Hence by standard approximation techniques we may take $\delta^*$ to be twice continuously differentiable with compact support. Then $E[(m(x_1) - m(x))^2|W_1 = 1] \to 0$ as $n \to \infty$.

It only remains to note that from 1) and 2) in the proof of Theorem 1, $\delta^*$ may be taken to be continuous with compact support. Hence

$$E[(\delta^*(x_1) - \delta(x))^2|W_1 = 1] \to 0$$

as $n \to \infty$. Q.E.D.

Conclusions: Under mild conditions we have shown that window estimates are $L_q$ consistent estimates of a regression function. We have also given moment bounds for $E(\delta_n - \delta)^2$. We end this discussion with a list of open problems.

1. How serious is the assumption that $(\theta_i, X_i)$ $i = 1, n$ be i.i.d.?
2. What are the weak convergence properties of \( L(\delta_n - \delta^*) \) for non smooth \( \delta^* \) and suitable functionals \( L \)?

3. If \( X \) is a Banach valued random variable do similar results hold?

4. (G. Meeden). In the usual empirical Bayes formulation we know \( F_0(x) \) and observe \( X_1, \ldots, X_n \). How much better off are we in our formulation of the problem?

An interesting and extended application of the results of this paper may be found in Spiegelman (1976).

Acknowledgement. This paper was motivated by discussions held with his major professor J. Sacks. The author thanks G. Meeden for helping with the organization of the manuscript.
REFERENCES


Stone, C. (1975a). Consistent weighted average estimators of a regression function, manuscript.
