NONPARAMETRIC REGRESSION

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be indep. realizations of $(X, Y) \in \mathbb{R} \times \mathbb{R}$, where

$$Y = m(X) + \varepsilon,$$

with ε independent of X with $\mathbb{E}\varepsilon = 0$, $\mathbb{E}\varepsilon^2 = \sigma^2$.

So we have $m(x) = \mathbb{E}[Y \mid X = x]$.

Linear regression assumes

$$m \in \{ m : \mathbb{R} \to \mathbb{R} : m(x) = \beta_0 + \beta_1 x, \ \beta_0, \beta_1 \in \mathbb{R} \}$$

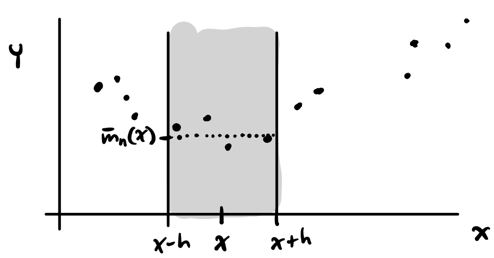
We will not assume any parametric form. Rather, we assume that m belongs to some class of functions of a certain smoothness.

There are a multitude of ways to estimate m nonparametrically!

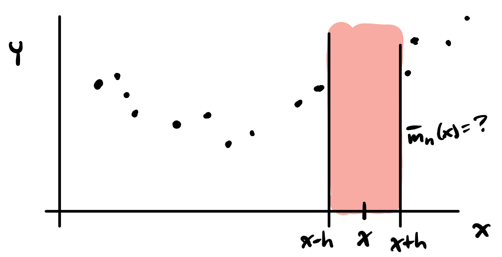
We first consider a local-averaging estimator

$$\bar{m}_n(x) = \frac{\sum_{i=1}^{n} Y_i \, \mathbb{1}(x - h \le X_i \le x + h)}{\sum_{j=1}^{n} \mathbb{1}(x - h \le X_j \le x + h)}.$$

At a point $x \in [0,1]$, $\bar{m}_n(x)$ is the average of Y_i values for which the corresponding X_i values are near x. It is the average of points within a moving window:



Note that we need the denominator $\sum_{i=1}^n \mathbb{1}(x-h \le X_i \le x+h) = \#\{X_i \in (x \pm h)\}$ to be positive for all x. Otherwise the estimator will be undefined:



A more general version of the local averaging estimator is the Nadaraya-Watson estimator

$$\hat{m}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i \, K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)},$$

when K is a kernel function like

$$K(u) = \mathbb{1}(|u| \le 1)$$
$$K(u) = (1-u^2)\,\mathbb{1}(|u| \le 1)$$
$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$$

Note that we do not need $\int_{\mathbb{R}} K(u)\,du = 1$ (as with KDEs).

Mean squared error of Nadaraya-Watson estimator

We consider MSE $\hat{m}_n^{NW}(x_0)$, $x_0 \in [0,1]$, when $m \in \text{Lipschitz}(L)$ on $[0,1]$.
(can shift/scale $X_1, \dots, X_n$ to be on $[0,1]$, so no generality is lost)

Let $$\text{MSE } \hat{m}_n^{NW}(x_0) = \underbrace{b_n^2(x_0)}_{\text{squared bias}} + \underbrace{\sigma_n^2(x_0)}_{\text{Variance}}, \quad \text{with}$$

$$b_n(x_0) = \mathbb{E}\,\hat{m}_n^{NW}(x_0) - m(x_0), \qquad \sigma_n^2(x_0) = \text{Var } \hat{m}_n^{NW}(x_0).$$

We will make the following assumptions on the kernel and on $X_1, \dots, X_n$:

(K1) Let $K: \mathbb{R} \to \mathbb{R}$ have support on $[-1,1]$ and satisfy $0 \le K(u) \le K_{max} < \infty \ \forall u \in \mathbb{R}$.

(D1) Let $X_1, \dots, X_n \in [0,1]$ be deterministic such that for some $n_0 > 0$

$$0 < f_{min} \le \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \le f_{max} < \infty \qquad \text{for all } x \in [0,1]$$

for all $n \ge n_0$.

Note: (K1) excludes the beloved Gaussian kernel, but it makes our proof simpler!

Result: Under (K1) and (D1), if $m \in \text{Lipschitz}(L)$ on $[0,1]$, we have

$$\text{MSE}\ \hat{m}_n^{NW}(x_0) \le h^2 \cdot L^2 + \frac{\sigma^2}{nh} \frac{K_{max}}{f_{min}} \qquad \text{for all } x_0 \in [0,1],$$

provided $n \ge n_0$.

Proof: First write

$$\hat{m}_n^{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)} = \sum_{i=1}^{n} W_{ni}(x) Y_i,$$

with

$$W_{ni}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}.$$

Now we consider the variance:

$$\hat{\sigma}_n^2(x_0) = \text{Var}\left[\sum_{i=1}^{n} W_{ni}(x_0) Y_i\right]$$

$$= \sum_{i=1}^{n} W_{ni}^2(x_0)\, \sigma^2$$

$$\le \sum_{i=1}^{n} W_{ni}(x_0) \cdot W_{ni}(x_0) \cdot \sigma^2 \qquad (K(u) \ge 0 \ \forall u)$$

$$\le \sup_{1\le j\le n} W_{nj}(x_0) \sum_{i=1}^{n} W_{ni}(x_0) \, \sigma^2 \qquad \left(\sum_{i=1}^{n} W_{ni}(x_0) = 1 \right)$$

$$= \sup_{1\le j\le n} W_{nj}(x_0)\, \sigma^2$$

$$\le \sup_{1\le j\le n} \left(\frac{K\left(\frac{X_j - x_0}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)} \sigma^2 \right)$$

$$\le \frac{\frac{1}{nh} K_{max}}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)} \cdot \sigma^2$$

$$\le \frac{1}{nh} \cdot \frac{K_{max}}{f_{min}} \cdot \sigma^2 .$$

Now the bias:

$$b_n(x_0) = \mathbb{E} \sum_{i=1}^{n} W_{ni}(x_0) Y_i - m(x_0)$$

$$= \sum_{i=1}^{n} W_{ni}(x_0)\, m(X_i) - m(x_0)$$

$$= \sum_{i=1}^{n} W_{ni}(x_0) \left[m(X_i) - m(x_0) \right]. \qquad \left(\sum_{i=1}^{n} W_{ni}(x_0) = 1 \right)$$

From here we write

$$|b_n(x_0)| \le \sum_{i=1}^{n} W_{ni}(x_0)\, L\, |X_i - x_0| \qquad \left(m \in \text{Lipschitz}(L) \text{ on } [0,1] \right)$$

K supported on [-1, 1]

$$= \sum_{i=1}^{n} W_{ni}(x_0)\, L\, |X_i - x_0|\, \mathbb{1}\left(|X_i - x_0| \le h \right)$$

$$\le \sum_{i=1}^{n} W_{ni}(x_0)\, L \cdot h$$

$$= h \cdot L . \qquad \square \qquad \boxed{4}$$

Exercise: Find optimal bandwidth and then optimal MSE bound for $\hat{m}_n^{NW}(x_0)$.

$$\frac{\partial}{\partial h}\left(\text{bound on MSE } \hat{m}_n^{NW}(x_0)\right) = 2h\cdot L^2 - \frac{\sigma^2}{nh^2}\frac{K_{max}}{f_{min}} = 0$$

$$\iff \quad h = n^{-\frac{1}{3}}\left(\frac{\sigma^2}{L^2}\cdot\frac{K_{max}}{2\cdot f_{min}}\right)^{1/3} =: h_{opt}.$$

Under this choice of h, we have

$$\text{MSE } \hat{m}_n^{NW}(x_0) \le \left[n^{-\frac{1}{3}}\left(\frac{\sigma^2}{L^2}\cdot\frac{K_{max}}{2\cdot f_{min}}\right)^{1/3}\right]^2\cdot L^2 + \frac{\sigma^2}{n}\frac{K_{max}}{f_{min}}\, n^{1/3}\left(\frac{L^2\ 2\cdot f_{min}}{\sigma^2\cdot K_{max}}\right)^{1/3}$$

$$= n^{-2/3}L^{2/3}\left\{\left(\frac{\sigma^2 K_{max}}{2 f_{min}}\right)^{2/3} + \left(\frac{\sigma^2\ K_{max}}{f_{min}}\right)^{2/3} 2^{1/3}\right\}.$$

We summarize the above by giving a uniform result:

Result: Under (K1) and (D1), with $h = \alpha n^{-1/3}$ for some $\alpha > 0$, we have

$$\sup_{m\in \text{Lipschitz}(L) \text{ on } [0,1]}\ \sup_{x\in[0,1]} \mathbb{E}\left[\hat{m}_n^{NW}(x) - m(x)\right]^2 \le n^{-2/3}\cdot C,$$

for $n \ge n_0$, when $C > 0$ depends on K_{max}, L, σ^2, and f_{min}.

Comparison to MSE in simple linear regression:

What would the MSE bound be in parametric regression?

Let $\hat{m}_n^{Lin}(x)$ be the estimator given by

$$\hat{m}^{Lin}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \quad (\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1\in\mathbb{R}}{\arg\min} \sum_{i=1}^n \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2$$

and define the class of linear functions on an interval $T \subset \mathbb{R}$ as

$$\mathrm{Lin}(T) = \{ m : T \to \mathbb{R} \; : \; m(x) = ax+b, \; a, b \in \mathbb{R} \}.$$

We will make the following assumption:

(D1 lin) Let $x_1, \ldots, x_n \in [0,1]$ be deterministic such that for some $n_0 > 0$

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \geq S_{\min} > 0$$

for all $n \geq n_0$.

<u>Result:</u> Under (D1 lin), for $n \geq n_0$, we have

$$\sup_{m \in \mathrm{Lin}([0,1])} \; \sup_{x \in [0,1]} \mathbb{E} \left[\hat{m}_n^{\mathrm{Lin}}(x_0) - m(x_0) \right]^2 \leq n^{-1} \left[1 + \frac{1}{S_{\min}} \right] \sigma^2 .$$

<u>Proof:</u> For every $x_0 \in [0,1]$, $\hat{m}_n^{\mathrm{Lin}}(x_0)$ is <u>unbiased,</u> since

$$\mathbb{E}\, \hat{m}_n^{\mathrm{Lin}}(x_0) = \beta_0 + \beta_1 x_0 = m(x_0) .$$

Next, we have

$$\mathrm{Var}\, \hat{m}_n^{\mathrm{Lin}}(x_0) = \mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0 \right)$$

$$= \left[\frac{1}{n} + \frac{(x_0 - \bar{x}_n)^2}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} \right] \sigma^2$$ (See my STAT 513 notes)

$$= \frac{1}{n} \left[1 + \frac{(x_0 - \bar{x}_n)^2}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2} \right] \sigma^2$$

$$\leq \frac{1}{n}\left[1 + \frac{1}{S_{min}}\right]\sigma^2. \qquad \left((D1 \text{lin}) \text{ and } (x_0 - \bar{x}_n) \leq 1\right)$$

Since this holds for all $m \in Lin([0,1])$, the proof is complete.

<u>Summary:</u> The MSE of the parametric estimator is smaller than that of the non parametric estimator. But parametric models might be misspecified (this is what motivates non parametric statistics).

Confidence Interval for $m(x_0)$ at $x_0 \in [0,1]$:

To construct a C.I. for $m(x_0)$ at some $x_0 \in [0,1]$, we would like a result like

$$\left(\hat{m}_n^{NW}(x_0) - m(x_0)\right)/\vartheta_n(x_0) \longrightarrow N(0,1) \text{ in dist. as } n \to \infty$$

for some sequence $\vartheta_n(x_0)$. We will break this into two pieces.

$$\left(\hat{m}_n^{NW}(x_0) - m(x_0)\right)/\vartheta_n(x_0) = \underbrace{\left(\hat{m}_n^{NW}(x_0) - \mathbb{E}\,\hat{m}_n^{NW}(x_0)\right)/\vartheta_n(x_0)}_{\text{Asymptotically Normal?}} + \underbrace{\left(\mathbb{E}\,\hat{m}_n^{NW}(x_0) - m(x_0)\right)/\vartheta_n(x_0)}_{\text{Vanishing bias term?}}$$

<u>Remark:</u> Note that for consistency ($MSE\ \hat{m}_n^{NW}(x_0) \to 0$) we need $h \to 0$ <u>and</u> $nh \to \infty$.

For the Normality part, we will need the following result:

<u>Corollary to Lindeberg CLT:</u>

For seq. of iid rvs $\xi_1, \xi_2, \ldots$ with mean 0 and variance 1 and a seq. of real numbers $a_1, a_2, \ldots$ that satisfy

$$\max_{1 \leq i \leq n} |a_i| \Big/ \left[\sum_{j=1}^n a_j^2\right]^{1/2} \to 0 \quad \text{as} \quad n \to \infty,$$

we have

$$\frac{\sum_{i=1}^n a_i \xi_i}{\left[\sum_{j=1}^n a_j^2\right]^{1/2}} \longrightarrow N(0,1) \quad \text{in distribution as } n \to \infty.$$

We now present a theorem:

Theorem: Let K have support on $[-1,1]$ and satisfy $0 \leq K(u) \leq K_{max} < \infty$ for all $u \in \mathbb{R}$ and let $X_1, \ldots, X_n$ be deterministic such that for any sequence of bandwidths h_n satisfying $h_n \to 0$ and $nh_n \to \infty$ there exists an $n_0 > 0$ such that

$$(\star) \quad \frac{1}{nh_n} \sum_{i=1}^n K^2\left(\frac{X_i - x}{h_n}\right) \geq M > 0 \quad \text{and} \quad \underset{\text{needed?}}{0 < f_{min} \leq} \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right) \leq f_{max} < \infty$$

for all $x \in [0,1]$ for all $n \geq n_0$. Then for $m \in \text{Lipschitz}(L)$ on $[0,1]$ and any sequence $h_n \to 0$ with $nh_n \to \infty$, we have

$$① \quad \frac{\hat{m}_n^{NW}(x_0) - \mathbb{E}\hat{m}_n^{NW}(x_0)}{\sigma \left[\sum_{i=1}^n W_{ni}^2(x_0) \right]^{1/2}} \to \mathcal{N}(0,1) \text{ in dist. as } n \to \infty$$

$$② \quad \frac{\left| \mathbb{E}\hat{m}_n^{NW}(x_0) - m(x_0) \right|}{\sigma \left[\sum_{i=1}^n W_{ni}^2(x_0) \right]^{1/2}} \leq \frac{n^{1/2} h_n^{3/2}}{\sigma} \frac{f_{max}}{M} L \quad \text{for all } n \geq n_0.$$

Proof of ①: Note that we can write

$$\hat{m}_n^{NW}(x_0) - \mathbb{E}\hat{m}^{NW}(x_0) = \sum_{i=1}^n W_{ni}(x_0) Y_i - \sum_{i=1}^n W_{ni}(x_0) \mathbb{E} Y_i$$

$$= \sum_{i=1}^n W_{ni}(x_0) \varepsilon_i,$$

where $W_{ni}(x_0) = K\left(\frac{X_i - x_0}{h_n}\right) \Big/ \sum_{j=1}^n K\left(\frac{X_j - x_0}{h_n}\right)$.

By the Corollary to the Lindeberg C.L.T., it is suff. to show that

$$\max_{1 \leq i \leq n} |W_{ni}(x_0)| \Big/ \left[\sum_{j=1}^n W_{ni}^2(x_0) \right]^{1/2} \to 0 \quad \text{as } n \to \infty.$$

We have

$$\frac{\max_{1\le i\le n} |W_{ni}(x_0)|}{\left[\sum_{j=1}^n W_{ni}^2(x_0)\right]^{1/2}} = \frac{\max_{1\le i\le n} K\left(\frac{X_i - x_0}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x_0}{h_n}\right)} \Bigg/ \left(\sum_{i=1}^n \frac{K^2\left(\frac{X_i - x_0}{h_n}\right)}{\left[\sum_{j=1}^n K\left(\frac{X_j - x_0}{h_n}\right)\right]^2}\right)^{1/2}$$

$$= \frac{\max_{1\le i\le n} K\left(\frac{X_i - x_0}{h_n}\right)}{\left[\sum_{j=1}^n K^2\left(\frac{X_j - x_0}{h_n}\right)\right]^{1/2}}$$

$$\le \frac{K_{max}}{\sqrt{nh_n}\,\underbrace{\left[\frac{1}{nh_n}\sum_{j=1}^n K^2\left(\frac{X_j - x_0}{h_n}\right)\right]^{1/2}}_{\ge M \quad \forall n \ge n_0}} \to 0 \quad \text{as } n \to \infty,$$

by (✿) and because $nh_n \to \infty$.

Proof of ②: By $K(u) = 0$ for $u \notin [-1,1]$ and $m \in \text{Lipschitz}(L)$, we have

$$\left|\mathbb{E}\, \hat{m}_n^{NW}(x_0) - m(x_0)\right| \le h_n \cdot L. \quad (\text{previous work})$$

Also, for all $n \ge n_0$ we have

$$\frac{1}{\left[\sum_{i=1}^n W_{ni}^2(x_0)\right]^{1/2}} = \frac{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)}{\left[\sum_{j=1}^n K^2\left(\frac{X_j - x_0}{h}\right)\right]^{1/2}} = \sqrt{nh_n}\,\frac{\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)}{\left[\frac{1}{nh_n}\sum_{j=1}^n K^2\left(\frac{X_j - x_0}{h}\right)\right]^{1/2}} \le \sqrt{nh_n}\,\frac{f_{max}}{M},$$

by (✿). Combining these bounds, we obtain ②.

Remark: The condition (✿) just means that the $X_1, \dots, X_n$ are "nicely" spread across the interval $[0,1]$, not too concentrated around any one point and not to sparse over any interval.

Corollary: Under the conditions of the Theorem, we have

$$\frac{\left[\hat{m}_n^{NW}(x_0) - m(x_0)\right]}{\sigma\left[\sum_{i=1}^{n} W_{ni}^2(x_0)\right]^{1/2}} \to N(0,1) \quad \text{in dist. as } n \to \infty$$

provided $\quad n^{1/2} h_n^{3/2} \to 0.$

Exercise: Determine whether we can build a valid C.I. for $m(x_0)$ under the MSE-optimal choice of h_n.

Solution: The MSE-optimal choice of h_n is $h_n = \alpha n^{-1/3}$ for some $\alpha > 0$.
Under this choice, $n^{1/2} h_n^{3/2} = n^{1/2}\left(n^{-1/3}\right)^{3/2} = 1$, so the bias does not vanish.

Remark: We must choose h_n to be smaller than the MSE-optimal bandwidth in order to build valid C.I.s. This is called "undersmoothing".

Estimating σ^2:

Let $(X_{(1)}, Y_{(1)}), \dots, (X_{(n)}, Y_{(n)})$ represent the data reordered such that

$$X_{(1)} < X_{(2)} < \dots < X_{(n)}.$$

Consider the variance estimator:

$$\hat{\sigma}_n^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(Y_{(i+1)} - Y_{(i)}\right)^2$$

Result: If $X_1, \dots, X_n \in [0,1]$, $\mathbb{E}\,\varepsilon_1^4 = \mu_4 < \infty$, $m \in \text{Lipschitz}(L)$ on $[0,1]$, then

$$\hat{\sigma}_n^2 \to \sigma^2 \quad \text{in probability}$$

as $\quad n \to \infty.$

Proof:

We have

$$\hat{\sigma}_n^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left[\left(m(X_{(i+1)}) + \varepsilon_{(i+1)} \right) - \left(m(X_{(i)}) + \varepsilon_{(i)} \right) \right]^2$$

$$= \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(\varepsilon_{(i+1)} - \varepsilon_{(i)} \right)^2 + \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left[m(X_{(i+1)}) - m(X_{(i)}) \right]^2$$

$$+ 2 \cdot \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(\varepsilon_{(i+1)} - \varepsilon_{(i)} \right) \left[m(X_{(i+1)}) - m(X_{(i)}) \right].$$

The first term has expectation σ^2 and the third term has expectation 0.

The second term satisfies

$$\frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left[m(X_{(i+1)}) - m(X_{(i)}) \right]^2 \leq \frac{1}{2(n-1)} \sum_{i=1}^{n-1} L^2 \cdot \left(X_{(i+1)} - X_{(i)} \right)^2 \quad (m \in \text{Lipschitz}(L))$$

$$\leq \frac{L^2}{2(n-1)}. \quad (X_1, \ldots, X_n \in [0,1])$$

Therefore $\quad \mathbb{E}\, \hat{\sigma}_n^2 - \sigma^2 \to 0 \quad$ as $\quad n \to \infty$.

Moreover

$$\operatorname{Var} \hat{\sigma}_n^2 = \frac{1}{4(n-1)^2} \left[\sum_{i=1}^{n-1} \operatorname{Var}\left(\left(Y_{(i+1)} - Y_{(i)} \right)^2 \right) + \sum_{|i-j|=1} \operatorname{Cov}\left(\left(Y_{(i+1)} - Y_{(i)} \right)^2, \left(Y_{(j+1)} - Y_{(j)} \right)^2 \right) \right]$$

$$\leq \frac{1}{4(n-1)^2} \left[\sum_{i=1}^{n-1} \operatorname{Var}\left(\left(Y_{(i+1)} - Y_{(i)} \right)^2 \right) + \sum_{|i-j|=1} \sqrt{ \operatorname{Var}\left(\left(Y_{(i+1)} - Y_{(i)} \right)^2 \right) \operatorname{Var}\left(\left(Y_{(j+1)} - Y_{(j)} \right)^2 \right) } \right]$$

(See Appendix)

$$\leq \frac{1}{4(n-1)^2} \left[(n-1) + 2(n-1) \right] \Big[\underbrace{2\left(2\mu_4 + 6\sigma^4 \right) + 16\sigma^2 L^2}_{\text{Upper bound for } \operatorname{Var}\left(\left(Y_{(i+1)} - Y_{(i)} \right)^2 \right)\ \forall i.} \Big]$$

$$\to 0 \quad \text{as} \quad n \to \infty.$$

Since the bias and variance go to zero, we have consistency. $\square$

Note: Other estimators exist, but this one is nice because it does not depend on a bandwidth choice.

Corollary: Under the conditions of the theorem and $\mathbb{E}\,\varepsilon_i^4 < \infty$,

$$P\left(m(x_0) \in \left(\hat{m}_n^{NW}(x_0) \pm z_{\alpha/2}\, \hat{\sigma}_n^2 \sqrt{\sum_{i=1}^n W_{ni}^2(x_0)} \right) \right) \to 1-\alpha$$

as $n \to \infty$, provided $n^{1/2} h_n^{3/2} \to 0$.

Local-Polynomial Estimators

We now introduce a generalization of the N-W estimator and consider its performance over $m \in \mathcal{H}(\beta, L)$ on $[0,1]$.

This estimator fits a polynomial function to the data locally.

We have in mind, for x close to x_0, the approximation

$$m(x) \approx a_0 + a_1(x-x_0) + \dots + a_\ell \frac{(x-x_0)^\ell}{\ell!}$$

for some coefficients $a_0, a_1, \dots, a_\ell$.

We thus define $\hat{m}_n^{LP}(x)$ as the first element of the vector

$$(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_\ell)(x) = \underset{a_0, a_1, \dots, a_\ell \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^n \left[Y_i - \left(a_0 + a_1(X_i - x) + \dots + a_\ell \frac{(X_i - x)^\ell}{\ell!} \right) \right]^2 K\left(\frac{X_i - x}{h} \right).$$

That is, we fit a local order-ℓ polynomial approximation

$$\hat{a}_0(x) + \hat{a}_1(x)(\cdot - x) + \dots + \hat{a}_\ell(x) \frac{(\cdot - x)^\ell}{\ell!}$$

to the function m at x and evaluate it at x, obtaining $\hat{m}_n^{LP}(x) = \hat{a}_0$.

Remark: The N-W estimator is the LP estimator of order $\ell = 0$.

Exercise: (i) Find matrices $\mathbb{U}_x$ and $\mathbb{K}_x$ and the vector $\underset{\sim}{Y}$ such that

$$\underset{\sim}{\hat{a}}(x) := (\hat{a}_0, \hat{a}_1, \dots, \hat{a}_\ell)(x) = \left(\mathbb{U}_x^T \mathbb{K}_x \mathbb{U}_x \right)^{-1} \mathbb{U}_x^T \mathbb{K}_x \underset{\sim}{Y}$$

(ii) Show that we may write

$$\hat{m}_n^{LP}(x) = \sum_{i=1}^n W_{ni}^*(x)\, Y_i$$

with

$$W_{ni}^*(x) = \frac{1}{nh}\, e_1^T \left(\frac{1}{nh} \mathbb{U}_x^T \mathbb{K}_x \mathbb{U}_x \right)^{-1} \underbrace{\mathbb{U}_{x,i}}_{\text{row } i \text{ of } \mathbb{U}_x} \cdot K\left(\frac{X_i - x}{h} \right).$$

Solution: Let

$$\mathbb{U}_x = \begin{bmatrix} 1 & \cdots & (X_1 - x) & \cdots & \frac{1}{2}(X_1 - x)^2 & \cdots & \frac{1}{\ell!}(X_1 - x)^\ell \\ \vdots & & \vdots & & \vdots & & \vdots \\ 1 & \cdots & (X_n - x) & \cdots & \frac{1}{2}(X_n - x)^2 & \cdots & \frac{1}{\ell!}(X_n - x)^\ell \end{bmatrix}$$

$$\mathbb{K}_x = \operatorname{diag}\left(K\left(\frac{X_1 - x}{h}\right), \dots, K\left(\frac{X_n - x}{h}\right) \right).$$

We present a result from Tsybakov's book, which depends on these assumptions:

(LP1) There exists a $\lambda_0 > 0$ and $n_0 > 0$ such that

$$\lambda_{min}\left(\frac{1}{nh}\, \mathbb{U}_x^T \mathbb{K}_x \mathbb{U}_x\right) \geq \lambda_0 > 0 \quad \text{for all } n \geq n_0.$$

[Points $X_1, \dots, X_n$ distributed such that no interval in $[0,1]$ stays empty.]

(LP2) There exists a $a_0 > 0$ such that for every $A \subset [0,1]$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \in A) \leq a_0 \left(Leb(A) \vee \frac{1}{n} \right).$$

[Points $X_1, \dots, X_n$ not <u>too</u> concentrated in any one small interval]

(LP3) K with support on $[-1,1]$ and $0 \leq K(u) \leq K_{max} < \infty \quad \forall\, u \in \mathbb{R}$.

<u>Theorem 2</u>: Let $X_1, \dots, X_n$ be deterministic such that (LP1), (LP2), and (LP3) hold under $h = \alpha n^{-1/(2\beta+1)}$. Let $m \in \mathcal{H}(\beta, L)$ on $[0,1]$. Then

$$\sup_{m \in \mathcal{H}(\beta,L)} \; \sup_{x \in [0,1]} \mathbb{E}\left[\hat{m}_n^{LP}(x) - m(x)\right]^2 \leq n^{-\frac{2\beta}{2\beta+1}} \cdot C$$

for all $n \geq n_0$, when $\hat{m}_n^{LP}(x)$ is of order $\ell = \beta - 1$. The constant C depends only on $\beta, L, \lambda_0, a_0, K_{max}, \sigma^2$, and α.

In practice, the N-W and local linear ($\ell = 1$) polynomial estimators are used very often. Making ℓ larger can lead to numerical issues...

MSE of N-W estimator under bounded 2nd derivative

Much of the nonparametric literature is written under the settings of the following theorem (filched from Larry Wasserman's lecture notes).

Let $\hat{m}_n^{LP\text{-}1}(x)$ denote the LP estimator with $\ell = 1$ ("local linear").

Theorem 3: Let $X_1, \dots, X_n$ have a continuous, differentiable density f, $f(x) > 0$, and suppose m'' is continuous and bounded. Then if $h_n \to 0$ and $nh_n \to \infty$, we have

$$\text{Bias } \hat{m}_n^{NW}(x) = \frac{h^2}{2}\left(m''(x) + \frac{2\, m'(x) f'(x)}{f(x)}\right) \int_{\mathbb{R}} u^2 K(u)\, du + o_p(h^2)$$

$$\text{Bias } \hat{m}_n^{LP\text{-}1}(x) = \frac{h^2}{2} m''(x) \int_{\mathbb{R}} u^2 K(u)\, du + o_p(h^2),$$

while $\text{Var } \hat{m}_n^{NW}(x)$ and $\text{Var } \hat{m}_n^{LP\text{-}1}(x)$ are given by

$$\frac{1}{nh} \frac{\sigma^2}{f(x)} \int_{\mathbb{R}} K^2(u)\, du + o_p\left((nh)^{-1}\right)$$

The proof of this result is more cumbersome than that of our result under Lipschitz smoothness, but it offers some insights.

The optimal rate is $O\left(n^{-4/5}\right)$ under these settings (better than under Lipschitz smoothness, due to existence of 2 derivatives).

We see how the curvature $m''(x)$ contributes to the bias.

Also how a scarcity of X values (small $f(x)$) can inflate both variance and bias.

Appendix:

For $Y_i = m(X_i) + \varepsilon_i$, $i = 1, \dots, n$, $\varepsilon_1, \dots, \varepsilon_n$ iid with $\mathbb{E}\varepsilon_1 = 0$, $\mathbb{E}\varepsilon_1^2 = \sigma^2$, $\mathbb{E}\varepsilon_1^4 = \mu_4 < \infty$ and $m \in \text{Lipschitz}(L)$ on $[0,1]$, $X_1, \dots, X_n \in [0,1]$ we have

$$
\begin{aligned}
\mathrm{Var}\left(\left(Y_{(i+1)} - Y_{(i)}\right)^2\right) &= \mathrm{Var}\left[\left(m(X_{(i+1)}) - m(X_{(i)}) + \varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^2\right] \\
&= \mathrm{Var}\Big[\left(m(X_{(i+1)}) - m(X_{(i)})\right)^2 + \left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^2 \\
&\qquad + 2\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)\left(m(X_{(i+1)}) - m(X_{(i)})\right)\Big] \\
&= \mathrm{Var}\left[\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^2 + 2\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)\left(m(X_{(i+1)}) - m(X_{(i)})\right)\right] \\
&\le 2\,\mathrm{Var}\left[\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^2\right] + 2\,\mathrm{Var}\left[2\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)\left(m(X_{(i+1)}) - m(X_{(i)})\right)\right] \\
&\le 2\,\mathbb{E}\left[\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^4\right] + 8\,\underbrace{\mathbb{E}\left[\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^2\right]}_{2\sigma^2}\,\underbrace{\left|m(X_{(i+1)}) - m(X_{(i)})\right|^2}_{\substack{\le L^2 |X_{(i+1)} - X_{(i)}|^2 \\ \le L^2 \\ (X_1, \dots, X_n \in [0,1])}} \\
&\le 2\left(2\mu_4 + 6\sigma^4\right) + 16\,\sigma^2 L^2,
\end{aligned}
$$

for all $i = 1, \dots, n$, since

$$
\begin{aligned}
\mathbb{E}\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^4 &= \mathbb{E}\left[\varepsilon_{(i+1)}^4 - 4\varepsilon_{(i+1)}^3 \varepsilon_{(i)} + 6\,\varepsilon_{(i+1)}^2 \varepsilon_{(i)}^2 - 4\varepsilon_{(i+1)} \varepsilon_{(i)}^3 + \varepsilon_{(i)}^4\right] \\
&= \mu_4 + 6\,\sigma^2 \cdot \sigma^2 + \mu_4 \\
&= 2\mu_4 + 6\sigma^4.
\end{aligned}
$$