

# Lecture 1: Non-Parametric Estimation

---

Raul Riva

FGV EPGE

October, 2025

# Introduction

---

So far, you have concentrated on *parametric models*:

$$Y_i = g(X_i, \theta) + u_i$$

where  $g$  is a **known** function and  $\theta$  needs to be estimated;

- Example:  $g(X, \theta) = X'\theta$ , good and old OLS;

So far, you have concentrated on *parametric models*:

$$Y_i = g(X_i, \theta) + u_i$$

where  $g$  is a **known** function and  $\theta$  needs to be estimated;

- Example:  $g(X, \theta) = X'\theta$ , good and old OLS;
- But what happens if you don't know  $g$ ?
- If you have a structural model, theory alone might give you  $g$ ;

So far, you have concentrated on *parametric models*:

$$Y_i = g(X_i, \theta) + u_i$$

where  $g$  is a **known** function and  $\theta$  needs to be estimated;

- Example:  $g(X, \theta) = X'\theta$ , good and old OLS;
- But what happens if you don't know  $g$ ?
- If you have a structural model, theory alone might give you  $g$ ;
- Just assume some  $g$  that looks cute and call it a day?
- We can do better  $\implies$  non-parametric estimation;

Our goal will be estimating:

$$m(x) = \mathbb{E}[Y|X = x]$$

- Examples: expected wage given education, expected returns given exchange rates...
- For now we assume that both  $Y$  and  $X$  are scalars;
- We assume the researcher has access to a random sample  $\{(Y_i, X_i)\}_{i=1}^n$ ;

Our goal will be estimating:

$$m(x) = \mathbb{E}[Y|X = x]$$

- Examples: expected wage given education, expected returns given exchange rates...
- For now we assume that both  $Y$  and  $X$  are scalars;
- We assume the researcher has access to a random sample  $\{(Y_i, X_i)\}_{i=1}^n$ ;
- With OLS, you always assume that  $m(\cdot)$  is an affine function;
- Importantly: in OLS,  $\frac{\partial m(x)}{\partial x} = \text{constant}$ ;

## What if $X$ is discrete?

- Assume that  $X \in \{x_1, x_2, \dots, x_l\}$  for some  $l$ . What's the natural estimator?



## What if $X$ is discrete?

- Assume that  $X \in \{x_1, x_2, \dots, x_l\}$  for some  $l$ . What's the natural estimator?

$$\hat{m}(x) = \frac{\sum_{i=1}^n I_{X_i=x} \cdot Y_i}{\sum_{i=1}^n I_{X_i=x}}$$

- Example: binary  $X$ , a treatment;
- You will prove in the problem set that this estimator is consistent under mild conditions;
- How would you describe this estimator in words?

## What about the continuous case?

- Now assume  $X \in \mathbb{R} \implies$  the event  $\{X_i = x\}$  has zero probability;
- Well... if  $m(\cdot)$  is continuous... maybe observing points in a neighborhood of  $x$  is good enough!
- Problem: how large is a neighborhood?

## What about the continuous case?

- Now assume  $X \in \mathbb{R} \implies$  the event  $\{X_i = x\}$  has zero probability;
- Well... if  $m(\cdot)$  is continuous... maybe observing points in a neighborhood of  $x$  is good enough!
- Problem: how large is a neighborhood?

The **Binned Estimator**, given  $h > 0$  (called *bandwidth*), is defined as:

$$\hat{m}(x) = \frac{\sum_{i=1}^n I_{\{|X_i - x| \leq h\}} \cdot Y_i}{\sum_{i=1}^n I_{\{|X_i - x| \leq h\}}}$$

- How do you think  $h$  will play a role here? Small vs large  $h$ ?

## What about the continuous case?

Notice that we can also write

$$\hat{m}(x) = \sum_{i=1}^n w_i(x) \cdot Y_i, \quad w_i(x) \equiv \frac{I_{\{|X_i - x| \leq h\}}}{\sum_{j=1}^n I_{\{|X_j - x| \leq h\}}}$$

in which these  $w_i$ 's are weakly positive and  $\sum_{i=1}^n w_i(x) = 1$ .

- We can view this estimator as a weighted average of  $Y$ , with  $x$ -dependent weights;
- True or false: can we interpret these weights as a *probability distribution*?

## What about the continuous case?

Notice that we can also write

$$\hat{m}(x) = \sum_{i=1}^n w_i(x) \cdot Y_i, \quad w_i(x) \equiv \frac{I_{\{|X_i - x| \leq h\}}}{\sum_{j=1}^n I_{\{|X_j - x| \leq h\}}}$$

in which these  $w_i$ 's are weakly positive and  $\sum_{i=1}^n w_i(x) = 1$ .

- We can view this estimator as a weighted average of  $Y$ , with  $x$ -dependent weights;
- True or false: can we interpret these weights as a *probability distribution*?
- How do these weights change with  $h$ ? Smoothly? Sharply?

## What about the continuous case?

Notice that we can also write

$$\hat{m}(x) = \sum_{i=1}^n w_i(x) \cdot Y_i, \quad w_i(x) \equiv \frac{I_{\{|X_i - x| \leq h\}}}{\sum_{j=1}^n I_{\{|X_j - x| \leq h\}}}$$

in which these  $w_i$ 's are weakly positive and  $\sum_{i=1}^n w_i(x) = 1$ .

- We can view this estimator as a weighted average of  $Y$ , with  $x$ -dependent weights;
- True or false: can we interpret these weights as a *probability distribution*?
- How do these weights change with  $h$ ? Smoothly? Sharply?
- One important drawback is that weights change discontinuously with  $h$ !
- $\hat{m}(x)$  is always a step function even if  $m(\cdot)$  is continuous!

## Kernels

---

- Ideally: points that are close to  $x$  should matter more for estimation, in a smooth way;
- It seems that we would like to have a “continuous way” of measuring distances;
- (Continuous) kernels are exactly what we need: they are fancy weights;

## Definition (Second-Order Kernel)

A second-order kernel function  $K(u)$  satisfies:

1.  $0 \leq K(u) \leq \bar{K} < \infty$ ;  $\implies$  the kernel is positive and bounded;
2.  $K(u) = K(-u)$ ;  $\implies$  the kernel is symmetric around zero;
3.  $\int_{-\infty}^{\infty} K(u) du = 1$ ;  $\implies$  this is like asking weights to sum up to 1;
4.  $\int_{-\infty}^{\infty} |u|^r K(u) du < \infty$  for positive integers  $r$ ;  $\implies$  “not too fat tails”;



## Examples?

Kernel	Formula	$R_K$
Rectangular	$K(u) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if }  u  < \sqrt{3} \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{2\sqrt{3}}$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$	$\frac{1}{2\sqrt{\pi}}$
Epanechnikov	$K(u) = \begin{cases} \frac{3}{4\sqrt{5}} \left(1 - \frac{u^2}{5}\right) & \text{if }  u  < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$	$\frac{3\sqrt{5}}{25}$
Triangular	$K(u) = \begin{cases} \frac{1}{\sqrt{6}} \left(1 - \frac{ u }{\sqrt{6}}\right) & \text{if }  u  < \sqrt{6} \\ 0 & \text{otherwise} \end{cases}$	$\frac{\sqrt{6}}{9}$

Figure 1: Examples of Kernels

## A More General Estimator

For a given bandwidth  $h > 0$ , the **Nadaraya-Watson (NW)** estimator is given by

$$\hat{m}(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \cdot Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

- Common choices of kernel: Gaussian and Epanechnikov;
- Choosing the bandwidth is more important than the kernel;

## A More General Estimator

For a given bandwidth  $h > 0$ , the **Nadaraya-Watson (NW)** estimator is given by

$$\hat{m}(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \cdot Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

- Common choices of kernel: Gaussian and Epanechnikov;
- Choosing the bandwidth is more important than the kernel;
- What happens when  $h \rightarrow 0$ ? And when  $h \rightarrow \infty$ ?

## A More General Estimator

For a given bandwidth  $h > 0$ , the **Nadaraya-Watson (NW)** estimator is given by

$$\hat{m}(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \cdot Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

- Common choices of kernel: Gaussian and Epanechnikov;
- Choosing the bandwidth is more important than the kernel;
- What happens when  $h \rightarrow 0$ ? And when  $h \rightarrow \infty$ ?
- You will prove that this estimator nests the binned estimator;

Questions?

# **Asymptotic Properties**

---

## Asymptotic Properties

- Do we have any guarantees that this methodology works?
- Yes, but flexibility will come many caveats! I want you to focus on **intuition**.

# Asymptotic Properties

- Do we have any guarantees that this methodology works?
- Yes, but flexibility will come many caveats! I want you to focus on **intuition**.

## Asymptotic Setup:

- We will assume that the sample size  $n \rightarrow \infty$ ;
- But we will also need  $h = h(n)$  to be converging towards zero  $\implies h \rightarrow 0$ ;
- But it cannot go to zero too fast  $\implies n \cdot h \rightarrow \infty$
- Is this intuitive? Why?



# Asymptotic Properties

- Do we have any guarantees that this methodology works?
- Yes, but flexibility will come many caveats! I want you to focus on **intuition**.

## Asymptotic Setup:

- We will assume that the sample size  $n \rightarrow \infty$ ;
- But we will also need  $h = h(n)$  to be converging towards zero  $\implies h \rightarrow 0$ ;
- But it cannot go to zero too fast  $\implies n \cdot h \rightarrow \infty$
- Is this intuitive? Why?
- The asymptotic theory will also be *pointwise*, i.e., for a fixed value of  $x$ ;
- The asymptotic distribution of our estimator will change as  $x$  changes;
- Very important: we will work through the case of an **interior point**  $x$ ;

# Asymptotic Theory

We start writing  $Y_i = m(X_i) + U_i$ , where  $\mathbb{E}[U_i|X_i] = 0$  and  $\sigma^2(x) \equiv \text{Var}(U_i|X_i = x)$ .

Let  $x \in \mathbb{R}$  and write  $Y_i = m(x) + (m(X_i) - m(x)) + U_i$

# Asymptotic Theory

We start writing  $Y_i = m(X_i) + U_i$ , where  $\mathbb{E}[U_i|X_i] = 0$  and  $\sigma^2(x) \equiv \text{Var}(U_i|X_i = x)$ .

Let  $x \in \mathbb{R}$  and write  $Y_i = m(x) + (m(X_i) - m(x)) + U_i$

Then we can write:

$$\begin{aligned} \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) m(x) \\ &\quad + \underbrace{\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) (m(X_i) - m(x))}_{\equiv \hat{\Delta}_1(x)} \\ &\quad + \underbrace{\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) U_i}_{\equiv \hat{\Delta}_2(x)} \end{aligned}$$

# Asymptotic Theory

- Assume  $x$  has some distribution with density  $f(\cdot)$ . Obviously, we assume  $f(x) > 0$ .
- Let  $f_n(x) \equiv \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$ ;
- You will prove in your problem set that  $f_n(x) \xrightarrow{p} f(x)$  as  $n \rightarrow \infty$ ;
- This is what we call the non-parametric kernel density estimator;
- This is just a fancy way of writing a histogram;
- For now, we will take this result for granted;

# Asymptotic Theory

- Assume  $x$  has some distribution with density  $f(\cdot)$ . Obviously, we assume  $f(x) > 0$ .
- Let  $f_n(x) \equiv \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$ ;
- You will prove in your problem set that  $f_n(x) \xrightarrow{p} f(x)$  as  $n \rightarrow \infty$ ;
- This is what we call the non-parametric kernel density estimator;
- This is just a fancy way of writing a histogram;
- For now, we will take this result for granted;
- The previous expression simplifies to

$$\hat{m}(x) - m(x) = \frac{1}{f_n(x)} \left[ \hat{\Delta}_1(x) + \hat{\Delta}_2(x) \right]$$

# Asymptotic Theory

We will show that  $\sqrt{nh} \cdot \hat{\Delta}_1(x)$  converges in probability and  $\sqrt{nh} \cdot \hat{\Delta}_2(x)$  has a limiting distribution.

# Asymptotic Theory

We will show that  $\sqrt{nh} \cdot \hat{\Delta}_1(x)$  converges in probability and  $\sqrt{nh} \cdot \hat{\Delta}_2(x)$  has a limiting distribution.

- First we analyze  $\hat{\Delta}_2(x)$ . Notice that

$$\mathbb{E} \left[ \frac{1}{nh} \sum_{i=1}^n K \left( \frac{X_i - x}{h} \right) U_i \right] = \frac{1}{h} \mathbb{E} \left[ K \left( \frac{X_i - x}{h} \right) U_i \right] = 0$$

# Asymptotic Theory

We will show that  $\sqrt{nh} \cdot \hat{\Delta}_1(x)$  converges in probability and  $\sqrt{nh} \cdot \hat{\Delta}_2(x)$  has a limiting distribution.

- First we analyze  $\hat{\Delta}_2(x)$ . Notice that

$$\mathbb{E} \left[ \frac{1}{nh} \sum_{i=1}^n K \left( \frac{X_i - x}{h} \right) U_i \right] = \frac{1}{h} \mathbb{E} \left[ K \left( \frac{X_i - x}{h} \right) U_i \right] = 0$$

- We also see that:

$$\begin{aligned} \text{Var} [\hat{\Delta}_2(x)] &= \frac{1}{nh^2} \mathbb{E} \left[ \left( K \left( \frac{X_i - x}{h} \right) U_i \right)^2 \right] \\ &= \frac{1}{nh^2} \mathbb{E} \left[ K \left( \frac{X_i - x}{h} \right)^2 \sigma^2(X_i) \right] \\ &= \frac{1}{nh^2} \int_{-\infty}^{\infty} K \left( \frac{z - x}{h} \right)^2 \sigma^2(z) f(z) dz \end{aligned}$$



# Asymptotic Theory

- We perform a change of variable:  $u = \frac{z-x}{h}$
- We can only do this because we are assuming that  $x$  is in the interior of its support;
- The last integral becomes:

$$\frac{1}{nh^2} \int_{-\infty}^{\infty} K\left(\frac{z-x}{h}\right)^2 \sigma^2(z) f(z) dz = \frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 \sigma^2(x+hu) f(x+hu) du$$

# Asymptotic Theory

- We perform a change of variable:  $u = \frac{z-x}{h}$
- We can only do this because we are assuming that  $x$  is in the interior of its support;
- The last integral becomes:

$$\frac{1}{nh^2} \int_{-\infty}^{\infty} K\left(\frac{z-x}{h}\right)^2 \sigma^2(z) f(z) dz = \frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 \sigma^2(x+hu) f(x+hu) du$$

- We will assume that  $\sigma^2(\cdot)$  and  $f(\cdot)$  are continuous;
- For any continuous function  $g(\cdot)$  and fixed  $u$ , we have  $g(x+hu) = g(x) + o(1)$  as  $h \rightarrow 0$ ;

# Asymptotic Theory

- We perform a change of variable:  $u = \frac{z-x}{h}$
- We can only do this because we are assuming that  $x$  is in the interior of its support;
- The last integral becomes:

$$\frac{1}{nh^2} \int_{-\infty}^{\infty} K\left(\frac{z-x}{h}\right)^2 \sigma^2(z) f(z) dz = \frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 \sigma^2(x+hu) f(x+hu) du$$

- We will assume that  $\sigma^2(\cdot)$  and  $f(\cdot)$  are continuous;
- For any continuous function  $g(\cdot)$  and fixed  $u$ , we have  $g(x+hu) = g(x) + o(1)$  as  $h \rightarrow 0$ ;

$$\frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 \sigma^2(x+hu) f(x+hu) du = \frac{\sigma^2(x) f(x)}{nh} \int_{-\infty}^{\infty} K(u)^2 du + o\left(\frac{1}{nh}\right)$$

# Asymptotic Theory

- We define  $R(K) \equiv \int_{-\infty}^{\infty} K(u)^2 du$  as the *roughness* of the kernel;
- So, we have shown that

$$\text{Var} [\hat{\Delta}_2(x)] = \frac{\sigma^2(x)f(x)R(K)}{nh} + o\left(\frac{1}{nh}\right)$$

- Important to notice: the variance of this term only vanishes if  $nh \rightarrow \infty$ ;
- Fast  $h \rightarrow 0 \implies$  slow convergence of this variance towards zero;
- By the Lindberg-Feller CLT:

$$\sqrt{nh}\hat{\Delta}_2(x) \xrightarrow{d} N(0, \sigma^2(x)f(x)R(K))$$

- Notice how the asymptotic variance depends on  $x$ !

Now we analyze our other guy:  $\hat{\Delta}_1(x)$  using similar tricks:

$$\begin{aligned}\mathbb{E}[\hat{\Delta}_1(x)] &= \frac{1}{h} \mathbb{E} \left[ K \left( \frac{X_i - x}{h} \right) (m(X_i) - m(x)) \right] \\ &= \frac{1}{h} \int_{-\infty}^{\infty} K \left( \frac{z - x}{h} \right) (m(z) - m(x)) f(z) dz \\ &= \int_{-\infty}^{\infty} K(u) (m(x + hu) - m(x)) f(x + hu) du\end{aligned}$$

Now we analyze our other guy:  $\hat{\Delta}_1(x)$  using similar tricks:

$$\begin{aligned}\mathbb{E}[\hat{\Delta}_1(x)] &= \frac{1}{h} \mathbb{E} \left[ K \left( \frac{X_i - x}{h} \right) (m(X_i) - m(x)) \right] \\ &= \frac{1}{h} \int_{-\infty}^{\infty} K \left( \frac{z - x}{h} \right) (m(z) - m(x)) f(z) dz \\ &= \int_{-\infty}^{\infty} K(u) (m(x + hu) - m(x)) f(x + hu) du\end{aligned}$$

We add the assumptions that  $f \in \mathbb{C}^1$  and that  $m \in \mathbb{C}^2$  and expand them:

$$\begin{aligned}m(x + hu) - m(x) &= m'(x)hu + \frac{1}{2}m''(x)h^2u^2 + o(h^2) \\ f(x + hu) &= f(x) + f'(x)hu + o(h)\end{aligned}$$

Distribute the terms:

$$(m(x+hu)-m(x)) \cdot f(x+hu) = m'(x)f(x)hu + m'(x)f'(x)h^2u^2 + \frac{1}{2}m''(x)f(x)h^2u^2 + o(h^2)$$

# Asymptotic Theory

Distribute the terms:

$$(m(x+hu)-m(x)) \cdot f(x+hu) = m'(x)f(x)hu + m'(x)f'(x)h^2u^2 + \frac{1}{2}m''(x)f(x)h^2u^2 + o(h^2)$$

This will lead to:

$$\begin{aligned}\mathbb{E}[\hat{\Delta}_1(x)] &= h^2 \left( \int_{-\infty}^{\infty} u^2 K(u)^2 du \right) \left( m'(x)f'(x) + \frac{1}{2}m''(x)f(x) \right) + o(h^2) \\ &= h^2 \kappa_2 f(x) B(x) + o(h^2)\end{aligned}$$

where  $B(x) = \left( m'(x) \frac{f'(x)}{f(x)} + \frac{1}{2}m''(x) \right)$  and  $\kappa_2 \equiv \left( \int_{-\infty}^{\infty} u^2 K(u)^2 du \right)$



# Asymptotic Theory

Distribute the terms:

$$(m(x+hu)-m(x)) \cdot f(x+hu) = m'(x)f(x)hu + m'(x)f'(x)h^2u^2 + \frac{1}{2}m''(x)f(x)h^2u^2 + o(h^2)$$

This will lead to:

$$\begin{aligned}\mathbb{E}[\hat{\Delta}_1(x)] &= h^2 \left( \int_{-\infty}^{\infty} u^2 K(u)^2 du \right) \left( m'(x)f'(x) + \frac{1}{2}m''(x)f(x) \right) + o(h^2) \\ &= h^2 \kappa_2 f(x) B(x) + o(h^2)\end{aligned}$$

where  $B(x) = \left( m'(x) \frac{f'(x)}{f(x)} + \frac{1}{2}m''(x) \right)$  and  $\kappa_2 \equiv \left( \int_{-\infty}^{\infty} u^2 K(u)^2 du \right)$

Questions: is  $B(x)$  random? How does  $h$  impact  $\mathbb{E}[\hat{\Delta}_1(x)]$ ?

- Computing  $\text{Var}(\hat{\Delta}_1(x))$  is a boring computation;
- See section 19.26 from Hansen's book. He can show that:

$$\text{Var}(\hat{\Delta}_1(x)) = o\left(\frac{1}{nh}\right)$$

- Computing  $\text{Var}(\hat{\Delta}_1(x))$  is a boring computation;
- See section 19.26 from Hansen's book. He can show that:

$$\text{Var}(\hat{\Delta}_1(x)) = o\left(\frac{1}{nh}\right)$$

- This will imply that  $\text{Var}(\sqrt{nh}\hat{\Delta}_1(x)) = o(1)$
- The main result is that:

$$\sqrt{nh} \left( \hat{\Delta}_1(x) - h^2 \kappa_2 f(x) B(x) \right) \xrightarrow{p} 0$$

as long as  $nh^5 = O(1)$  (why do we need this?);

# Asymptotic Theory - Main Result

Please see all the technical conditions on Hansen's book (Chapter 19);

## Theorem (Asymptotic Distribution of the NW Estimator)

*Under regularity conditions, for interior  $x$ ,*

$$\sqrt{nh} (\hat{m}(x) - m(x) - h^2 \kappa_2 B(x)) \xrightarrow{d} N \left( 0, \frac{\sigma^2(x) R(K)}{f(x)} \right)$$

*where  $B(x) = \left( m'(x) \frac{f'(x)}{f(x)} + \frac{1}{2} m''(x) \right)$ .*

# Asymptotic Theory - Main Result

Please see all the technical conditions on Hansen's book (Chapter 19);

## Theorem (Asymptotic Distribution of the NW Estimator)

*Under regularity conditions, for interior  $x$ ,*

$$\sqrt{nh} (\hat{m}(x) - m(x) - h^2 \kappa_2 B(x)) \xrightarrow{d} N \left( 0, \frac{\sigma^2(x) R(K)}{f(x)} \right)$$

*where  $B(x) = \left( m'(x) \frac{f'(x)}{f(x)} + \frac{1}{2} m''(x) \right)$ .*

- Why is the convergence happening at the rate  $\sqrt{nh}$  and not just  $\sqrt{n}$ ?
- True or false: is there always a finite sample bias here?

Questions?

# The Bias-Variance Trade-off

- If  $n$  is very large:

$$(\hat{m}(x) - m(x)) \approx N\left(h^2 \kappa_2 B(x), \frac{\sigma^2(x) R(K)}{nh \cdot f(x)}\right)$$

- What is the effect of  $h$  on the mean and variance?

# The Bias-Variance Trade-off

- If  $n$  is very large:

$$(\hat{m}(x) - m(x)) \approx N \left( h^2 \kappa_2 B(x), \frac{\sigma^2(x) R(K)}{nh \cdot f(x)} \right)$$

- What is the effect of  $h$  on the mean and variance?
- Fundamental trade-off: we can reduce the bias, at the expense of variance;
- Or you can get low variance... and a huge bias!



# The Bias-Variance Trade-off

- If  $n$  is very large:

$$(\hat{m}(x) - m(x)) \approx N\left(h^2 \kappa_2 B(x), \frac{\sigma^2(x) R(K)}{nh \cdot f(x)}\right)$$

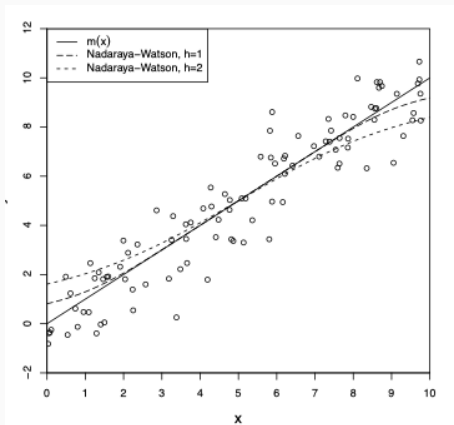
- What is the effect of  $h$  on the mean and variance?
- Fundamental trade-off: we can reduce the bias, at the expense of variance;
- Or you can get low variance... and a huge bias!
- The Epanechnikov kernel is the one that minimizes the mean-squared-error of this estimation across a large class of kernels;
- The efficiency loss when using the Gaussian kernel is minimal;
- Honestly, let the Gaussian kernel be your default in empirical research;

## How to pick the bandwidth?

- The optimal bandwidth will depend on moments of the data and derivatives of  $m$ ;
- That is unknown in practice;
- Ideally, your results should be robust to different bandwidths (within reason);
- A popular way to choose a bandwidth is leave-one-out cross-validation;
- Exactly as the Machine Learning literature does!
- See Hansen's book for details and the theory behind it;

## What about the boundary?

- Suppose  $X_i$  comes from a bounded distribution, for example;
- Let's say  $X \sim U[0, 10]$  and  $Y|X = x \sim N(x, 1)$ ;



**Figure 2:** Bias at the boundary

# The Local Linear Estimator

- Notice that the Nadaraya-Watson estimator also satisfies:

$$\hat{m}(x) = \arg \min_{c \in \mathbb{R}} \sum_{i=1}^n K \left( \frac{X_i - x}{h} \right) \cdot (Y_i - c)^2$$

- It's also called the local-constant estimator;

# The Local Linear Estimator

- Notice that the Nadaraya-Watson estimator also satisfies:

$$\hat{m}(x) = \arg \min_{c \in \mathbb{R}} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \cdot (Y_i - c)^2$$

- It's also called the local-constant estimator;
- What is stopping us from fitting a linear function here?

## Definition (Local-Linear Estimator)

For each  $x$ , solve the following optimization problem:

$$(\hat{\beta}_0(x), \hat{\beta}_1(x)) = \arg \min_{(b_0, b_1)} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) (Y_i - b_0 - b_1(X_i - x))^2$$

The local-linear estimator of  $m(x)$ , denoted by  $\hat{m}(x)_{LL}$ , is the local intercept  $\hat{\beta}_0(x)$ .

## How do they compare?

- Deriving the asymptotic distribution requires similar tricks. See Hansen's book;
- **Important:** both have the same asymptotic variance;
- **Also Important:** the bias term for the local-linear estimator is  $\frac{h^2\kappa_2}{2}m''(x)$

$$\sqrt{nh} \left( \hat{m}(x)_{LL} - m(x) - h^2\kappa_2 \frac{m''(x)}{2} \right) \xrightarrow{d} N \left( 0, \frac{\sigma^2(x)R(K)}{f(x)} \right)$$

- What does it imply if  $m(\cdot)$  is, in fact, linear?
- The local-linear estimator is much better close to the boundary! Why?

Questions?

# Curse of Dimensionality

- So far, we dealt with scalar  $x$ .
- What if  $X_i \in \mathbb{R}^p$  with  $p > 1$ ?
- In that case we use multivariate kernels and measure “distances” as

$$K\left(\frac{X_1 - x}{h_1}\right) \cdot K\left(\frac{X_2 - x}{h_2}\right) \cdots K\left(\frac{X_p - x}{h_p}\right)$$

where  $(h_1, \dots, h_p)$  are potentially different bandwidths;

- Math gets more involved but the same type of results are obtained;



# Curse of Dimensionality

- So far, we dealt with scalar  $x$ .
- What if  $X_i \in \mathbb{R}^p$  with  $p > 1$ ?
- In that case we use multivariate kernels and measure “distances” as

$$K\left(\frac{X_1 - x}{h_1}\right) \cdot K\left(\frac{X_2 - x}{h_2}\right) \cdots K\left(\frac{X_p - x}{h_p}\right)$$

where  $(h_1, \dots, h_p)$  are potentially different bandwidths;

- Math gets more involved but the same type of results are obtained;
- There is a super important caveat: convergence will happen at rate  $\sqrt{n \cdot h_1 \cdot h_2 \cdots h_p}$
- In case you use the same  $h$ , we will need that  $\sqrt{nh^p} \rightarrow \infty$ . This is **very** slow;
- In practice, if  $p > 4$  you are screwed  $\implies$  open the Machine Learning toolbox then;

## Confidence Bands

- You might be interested in making inference about  $m(x)$  since you did all the math to get the distribution...
- Keep in mind: any sort of confidence band you draw is **pointwise**;
- Usually, we cheat compute the 95% confidence interval as

$$\hat{m}(x) \pm 1.96 \cdot \sqrt{\frac{\hat{\sigma}^2(x)R(K)}{nh\hat{f}_n(x)}}$$

where you can use the residuals from your fit and the same kernel to compute  $\hat{\sigma}^2(x)$ ;

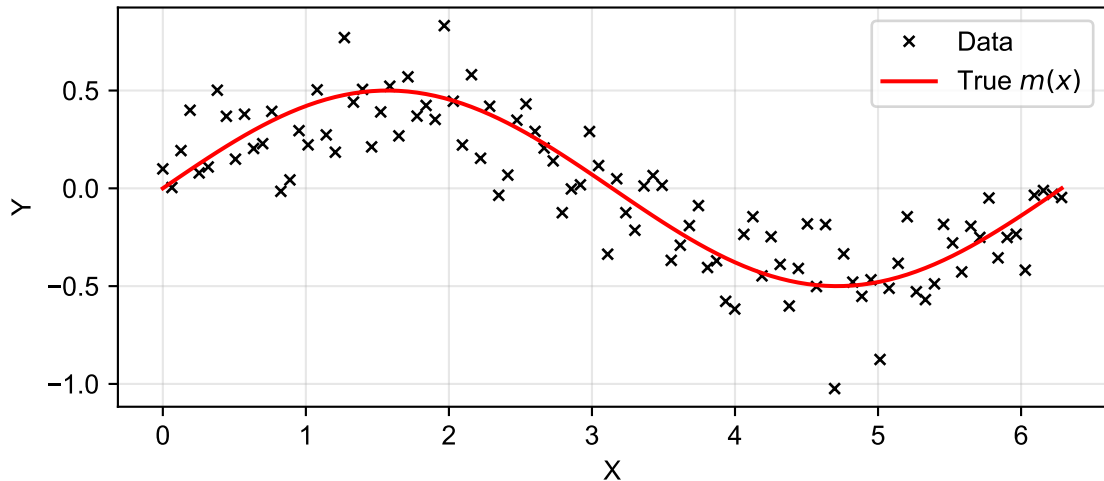
- Why is this cheating?

## Example

---

## A Quick Simulation

$$Y_i = \sin(X_i) + \varepsilon_i, \varepsilon_i \sim N(0, 0.2)$$

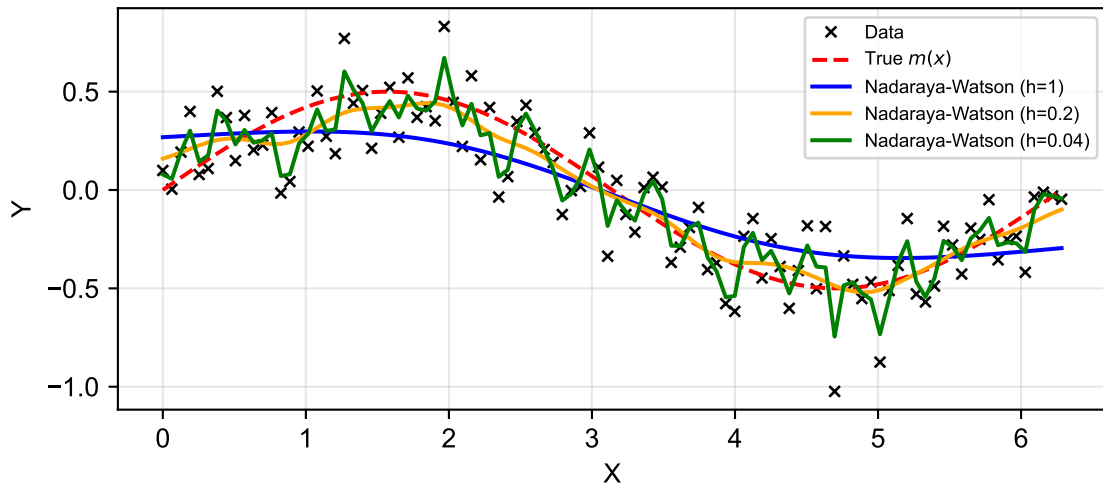


It's super easy to implement the Nadaraya-Watson estimator:

```
def gaussian_kernel(x, h):  
    return (1 / (h * np.sqrt(2 * np.pi))) * np.exp(-0.5 * (x / h) ** 2)  
  
def nw_estimator(X, Y, x, h):  
    K = gaussian_kernel((X - x) / h, 1)  
    return np.sum(K * Y) / np.sum(K)
```

## Different Bandwidths

Nadaraya-Watson Estimator with Different Bandwidths



**Questions?**

- Chapter 19 from *Econometrics*
- Chapter 17 from *Probability and Statistics for Economists*