Handout 3: Introduction to Nonparametric Methods. Density Estimation.

Master in Data Science for Decision Making
Barcelona School of Economics

Laura Mayoral

IAE and BSE

Barcelona, Winter 2025

1. Introduction

- The aim of statistical inference is

1. Introduction

- The aim of statistical inference is

to use data to infer an unknown quantity.

- trade-off between efficiency and generality

1. Introduction

- The aim of statistical inference is

to use data to infer an unknown quantity.

- trade-off between efficiency and generality

- this trade-off is controlled by the strength of assumptions that are made on the data generating process.

- Parametric inference favors efficiency.
  - Given a model (strong assumption!), parametric inference delivers a set of methods (point estimation, confidence intervals, hypothesis testing, etc.) tailored for such model.
    - Efficient if assumptions are true. Otherwise inconsistent.

- Nonparametric inference favors generality.

- Nonparametric methods: flexible estimation approaches
  - Given a set of minimal assumptions, it provides inferential methods that are consistent for broad situations, in exchange for losing efficiency for small or moderate sample sizes.

- Non-parametric econometrics is a huge field

- This lecture will provide an introduction of non-parametric methods in econometrics.

- Essential ideas are intuitive, but the concepts and technicalities involved get complicated fairly quickly.

Additional References

- For those of you interested in the topic (standard textbooks in graduate –PhD level– courses):
  - "Non-parametric Econometrics" by Pagan and Ullah
  - "Non-parametric Econometrics: Theory and Practice", by Li and Racine
  - "Nonparametric Econometrics: a Primer", by J.S. Racine

Some key differences between parametric and nonparametric methods

- Parametric methods

- need of "parametric" assumptions: examples, distribution of the data (for instance, normality) or about the shape of the relation among the variables under analysis (for instance, linearity).

- Can handle many variables.

- The focus is on estimating parameters.

- "Standard" asymptotic theory: (typically) $\sqrt{N}$-convergence.

Non-parametric methods

- They require making none of these assumptions.

- The focus in on estimating functions. Estimates are typically presented as graphs. (Then, it becomes essential to produce nicely formatted graphs, etc.)

- In practice, it cannot handle many variables ("curse of dimensionality").

- Typically different asymptotic theory, convergence is slower than $\sqrt{N}$

Semiparametric methods

- Semiparametric methods: compromise between parametric and non-parametric

- They aim at taking advantage of the advantages of the two worlds

- They place some structure in the data (i.e., a parametric part) but they are not fully-parametric; combine parametric and non-parametric methods

  - An example:

    - E.g.: $E(y|x,z) = x'\beta + g(z)$ where $g(.)$ is unspecified

    - reduces the dimension of the non-parametric component to one dimension

Takeaway

- Parametric and nonparametric methods have pros and cons.
- Therefore, they have traditionally considered be complements, not substitutes.
  - An example: you might estimate something parametrically, but then assess the validity of your assumptions by estimating it again nonparametrically, checking that results are consistent.

- Tradeoffs of Parametric methods
  - Pros:
    - theory is simpler
    - more efficient
    - they allow us to study the relationship between many variables.
  - Cons: accuracy depends on validity of assumptions

Takeaway, II

- Tradeoffs of Nonparametric methods:
  - Pros:
    - valid under very mild assumptions
  - Cons:
    - less efficient
    - theory is more complicated
    - curse of dimensionality: estimation becomes impossible as the number of regressors increase
    - Not fully "nonparametric": there are also parameters to be chosen, and the results will depend on the values chosen (i.e., as in parametric models!).

This handout: Density estimation and nonparametric regression

- We'll focus on two problems:
  - Density estimation: kernel density estimation.
  - Regression curve estimation: regression on a single scalar without imposing a functional form

Density Estimation: Roadmap

1. Introduction
2. Density Estimation.
3. Histograms
4. Nonparametric (kernel) density estimation

2. Density estimation

Problem: Given some observations from some variable X, we would like to obtain an estimate of its density.

Why?

Nonparametric density estimates may be used for:

- Exploratory data analysis.
- Estimating qualitative features of a distribution (e.g. unimodality, skewness, etc.).
- Specification and testing of parametric models (e.g., do my data look normal?).

- Constructing a nonparametric estimate of the conditional mean function (CMF) of Y given X:

$$\mu(x) = E(Y|X = x) = \int y f_{XY}(x, y)dy / f_X(x),$$

where $f_X(x) = \int_y f(x, y)dy$.

Given an estimate $\hat{f}(x, y)$ of the joint density $f(x, y)$ of (X, Y), the analogy principle suggests estimating $\mu(x)$ by

$$\hat{\mu}(x) = \int y \frac{\hat{f}(x, y)}{\hat{f}_X(x)} dy \tag{1}$$

where $\hat{f}_X(x) = \int \hat{f}(x, y)dy$.

- Many statistical problems involve estimation of a population parameter that can be represented as a statistical functional $T(f)$ of the population density f. In all these cases, if $\hat{f}$ is a reasonable estimate of f, then a reasonable estimate of $\theta = T(f)$ is $\hat{\theta} = T(\hat{f})$.

Alternative ways of estimating a density: Parametric vs. Nonparametric estimation

- Parametric density estimation:

- First step: choose a parametric density function (i.e., the normal, the exponential, the uniform, etc), and use data to estimate the parameters of this density.

- For example, if normality is assumed, then $X \sim N(\mu, \sigma^2)$.

- Use data to estimate μ and σ^2 using the approaches you already know, $X \sim N(\hat{\mu}, \hat{\sigma^2})$.

- Problem: what if X doesn't follow a normal distribution?

- Nonparametric density estimate:

1. Simplest approach: use a histogram by breaking data into bins and using the relative frequency within each bin.

  - Problem: a histogram is a step function, even if the variable is continuous.

2. Smooth nonparametric density estimate (kernel density estimate):

  - Use a histogram that is smoothed in several ways

3. Histograms

- A histogram is a nonparametric estimate of the density of X.

- Histogram is a step function defined over equally-spaced bins, where each step contains the fraction of observations falling in each bin.

- Formal derivation of the histogram:

- The density is the derivative of the cdf $F(x_0)$ (i.e. $f(x_0) = \frac{dF(x_0)}{dx}$). Then

$$\begin{aligned} f(x_0) &= \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} \\ &= \lim_{h \to 0} \frac{1}{2h} \Pr[x_0 - h < x < x_0 + h] \end{aligned}$$

- For a sample $x_i, i = 1, \ldots, N$ of size N, this suggests using the estimator

$$f_{\text{HIST}}(x_0) = \frac{1}{2h} \sum_{i=1}^{N} \frac{1(x_0 - h < x_i < x_0 + h)}{N}$$

(i.e., replace probability by relative frequency), where $1(.)$ is the indicator function defined as

$$1(A) = \begin{cases} 1 & \text{if } A \text{ is true } 0 \\ \text{otherwise} \end{cases}$$

- The estimator $f_{\text{HIST}}(x_0)$ is a histogram estimate centered at x_0 with bin width $2h$

- If f_{HIST} is evaluated over the support of x at equally spaced values of x, each $2h$ units apart, it yields a histogram.

Steps to construct a histogram

1. Select an initial observation t_0 and a bandwidth/window h

2. break data into bins of width $2h$, where the first bin is $(t_0, t_0 + 2h)$.

3. form rectangles of area the relative frequency = freq/N.

4. The height is freq/$(2Nh)$ (area = (freq/$(2Nh)$) $\cdot\, 2h$ = freq/N=relative frequency).

- The histogram estimate of $f(x_0)$, the density of x evaluated at x_0, is:

$$\hat{f}_{\text{HIST}}(x_0) = \frac{1}{2Nh} \sum_{i=1}^{N} 1(x_0 - h < x_i < x_0 + h)$$

Or:

$$\hat{f}_{\text{HIST}}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} \cdot 1(\left| \frac{x_i - x_0}{h} \right| < 1)$$

- It's easy to see that this is a proper density (always positive and integrates to 1).

An example (from Cameron and Trivedi)

- Histogram of log wage, N = 175 observations

  - Histograms can vary quite a lot depending on the number of bins.

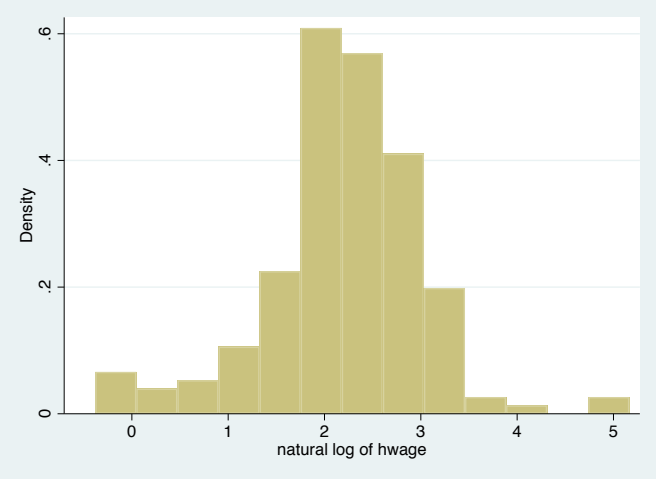
  - Number of bins in Stata: default is $\sqrt{N}$ for $N \leq 861$ and $10ln(N)/ln(10)$ for N>861

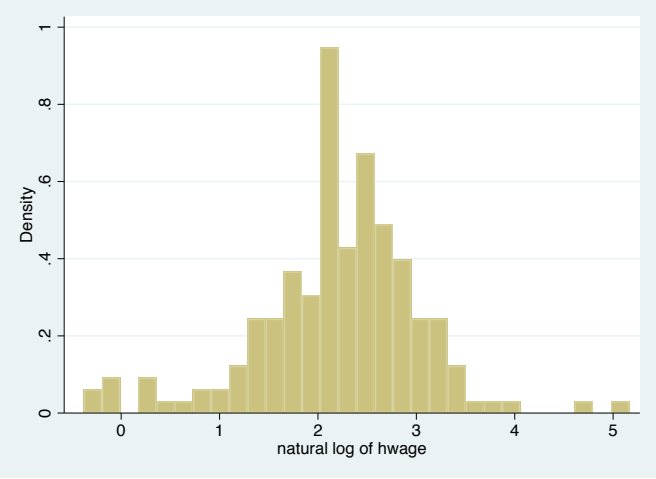
  - Two graphs: default and 30 bins, each of width 2h $\approx$ 0.20

  - STATA command:

histogram lnhwage, bin(30)

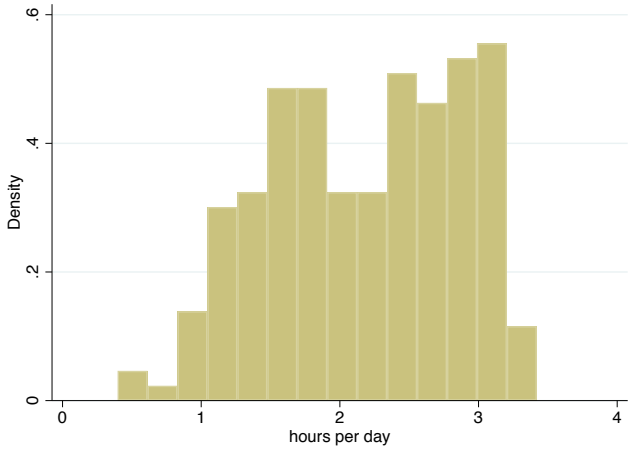
Histograms of log wage, with default number of bins and 30 bins.

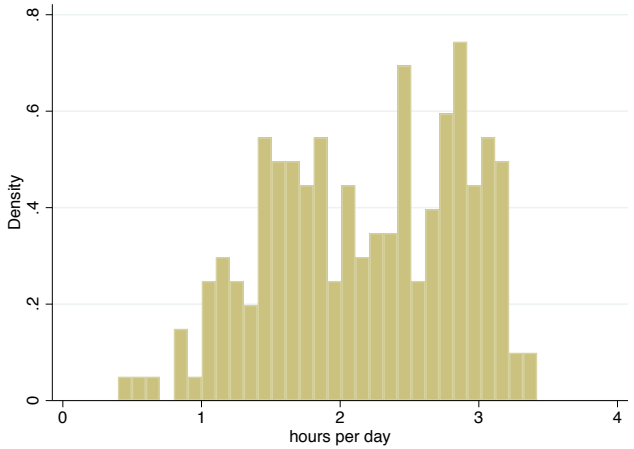


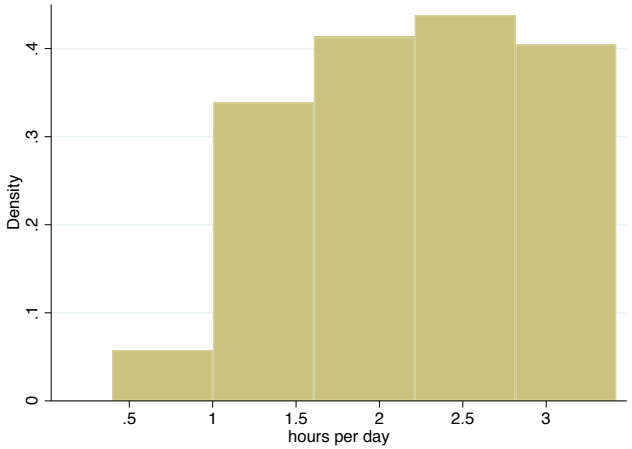


Stata code to generate and save some histograms

```
clear all

set seed -7

set obs 200

gen x = 5*uniform()+5

gen y = 2 + sin(x) + 0.25*rnormal()

la var y "hours per day"

la var x "wage"

hist y, graphregion(color(white))

graph export "hist1a.pdf", replace

hist y, bin(30) graphregion(color(white))

graph export "hist2a.pdf", replace

hist y, bin(5) graphregion(color(white))

graph export "hist3a.pdf", replace
```


Density
0
.2
.4
.6
0
1
2
3
4
hours per day


Density
0
.2
.4
.6
.8
0
1
2
3
4
hours per day


Density
0
.1
.2
.3
.4
.5
1
1.5
2
2.5
3
hours per day

Drawbacks of the Histogram

- Two main issues

  - Choice of number of bins (or, equivalently, choice of h)
    - Increasing J (reducing h) tends to give a histogram that is only informative about the location of the distinct sample points.
    - Reducing J (increasing h) eventually leads to a completely uninformative rectangle.
    - J may safely be increased if the sample size N also increases.
  - Lack of smoothness: The histogram is a step function (even for continuous variables!) with jumps at the end of each bin. Thus, it is impossible to incorporate prior information on the degree of smoothness of a density.

- We now present a related method that tries to overcome these two problems.

Takeaway

- Densities are easy to visualize and interpret, making them ideal tools for data exploration of continuous random variables.

- Parametric and non-parametric estimators of the density.

- Histograms: non-parametric estimator of the density
  - Define equally-sized bins, compute relative frequencies of these bins.
  - All the points within the interval share the same value of the density

- Main problems of histograms
  - We need to specify the number of bins, shape might change considerably depending on this
  - Histograms are a step function, they are not smooth

4. Kernel density estimate

- An alternative way of estimating densities
- It can be thought as a "smooth" version of the histogram
- Intuition
- Recall the formula for the histogram

$$\hat{f}_{\text{HIST}}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} \cdot 1(\left| \frac{x_i - x_0}{h} \right| < 1)$$

- We could write this expression as

$$\hat{f}_{\text{HIST}}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} w\left(\frac{x_i - x_0}{h} \right), w(u) = \begin{cases} \frac{1}{2}, & \text{if } -1 \leq u < 1 \\ 0, & \text{otherwise} \end{cases}$$

- Interpretation:
  - $w(.)$ is the uniform density in the interval (-1,1)
  - this means we're giving equal weights to the observations

- The generalization is now obvious: replace $w(.)$ by an arbitrary density, $K(.)$

- A Kernel density estimate of $f(x_0)$, the density of x evaluated at x_0, is

$$\widehat{f(x_0)} = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x_i - x_0}{h})$$

- $K(\cdot)$ is called a kernel function
- $K(\cdot)$ is a density that is symmetric, and unimodal at zero.
- h is called the bandwidth or smoothing parameter

Kernel Function: Intuition

- Kernel function:

- assigns weights to data points based on their proximity to a target point, with the weights decreasing as distance increases.

- For each point, averages of the neighbouring (weighted) observations are computed, and this is how smooth estimates are obtained

- Key properties of the Kernel are inherited by the estimate of the density:

- If K is a proper density, then so is $\hat{f}$,

- and if K is differentiable up to order r, then so is $\hat{f}(z)$.

Kernel Function, II

More formally: The kernel function K $(\cdot)$ is a continuous function, symmetric around zero, that integrates to unity and satisfies additional boundedness conditions.

1. $K(z)$ is symmetric around 0 and is continuous.
2. $\int_{-\infty}^{\infty} K(z)dz = 1$, $\int_{-\infty}^{\infty} zK(z)dz = 0$, and $\int_{-\infty}^{\infty} |K(z)|dz < \infty$.
3. Either (a) $K(z) = 0$ if $|z| \geq z_0$ for some z_0, or (b) $|z|K(z) \to 0$ as $|z| \to \infty$.
4. $\int_{-\infty}^{\infty} z^2 K(z)dz = \kappa$, where κ is a constant.

(Note: In practice, kernel functions work better if they satisfy condition (3)(a) rather than just the weaker condition (3)(b).

Examples of Kernels

- Uniform (or box or rectangular): $\frac{1}{2} \times 1(|z| < 1)$
- Triangular (or triangle): $(1 - |z|) \times 1(|z| < 1)$
- Epanechnikov (or quadratic): $\frac{3}{4}(1 - z^2) \times 1(|z| < 1)$
- Quartic (or biweight): $\frac{15}{16}(1 - z^2)^2 \times 1(|z| < 1)$
- Triweight: $\frac{35}{32}(1 - z^2)^3 \times 1(|z| < 1)$
- Tricubic: $\frac{70}{81}(1 - |z|^3)^3 \times 1(|z| < 1)$
- Gaussian (or normal): $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$
- Fourth-order Gaussian: $\frac{1}{2}(3 - z)^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$
- Fourth-order quartic: $\frac{15}{16}(3 - 10z^2 + 7z^4) \times 1(|z| < 1)$
- etc.

Example 1

- Uniform kernel:

- uses the same weights as a histogram of bin width 2h

- but it produces a "running histogram" that is evaluated at all points in the sample, rather than using fixed bins.

Example 2

- Consider $K = \Phi$, where Φ denotes the density of the $\mathcal{N}(0,1)$ distribution

- The kernel estimate of $f(x_0)$ may be viewed as the average of N (N=sample size) points, where the weights are obtained from a N(0,1) distribution.

- Hence:

  - the uniform kernel estimate of $f(x_0)$ is based only on the observations that are within h distance from the evaluation point x_0 and assigns them a constant weight,

  - the Gaussian kernel estimate is based on all the observations but assigns them a weight that declines exponentially as the distance from the evaluation point increases

Implementation

- Two choices to be made: choice of h (bandwidth) and choice of kernel, $K(.)$

- Choice of h is much more important than choice of K(.). (we'll see why)

Choosing the bandwidth

- Why does this matter?

Choosing the bandwidth

- Why does this matter?

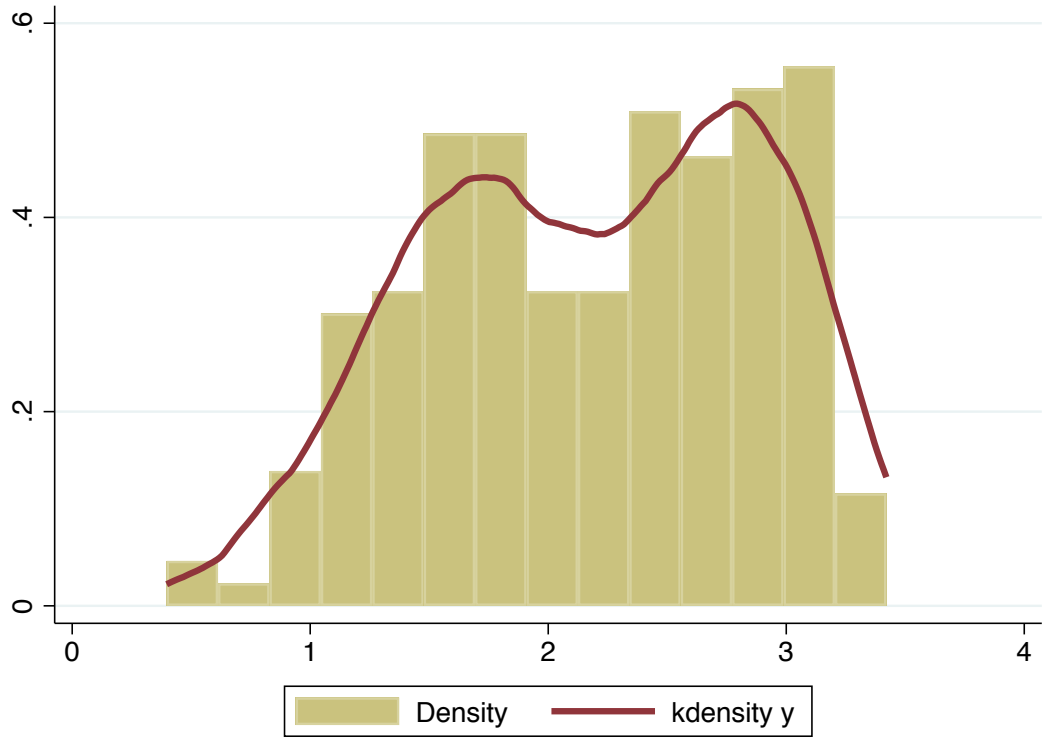
- The bandwidth h:

- controls the degree of smoothness or regularity of a density estimate.

  - Small values of h tend to produce estimates that are irregular, while large values of h correspond to very smooth estimates.

  - Setting h small will reduce bias
  - Setting h large will increase smoothness.

Example: histogram and kernel density estimate, STATA default values



STATA default is bandwidth=.21; Epanechnikov Kernel

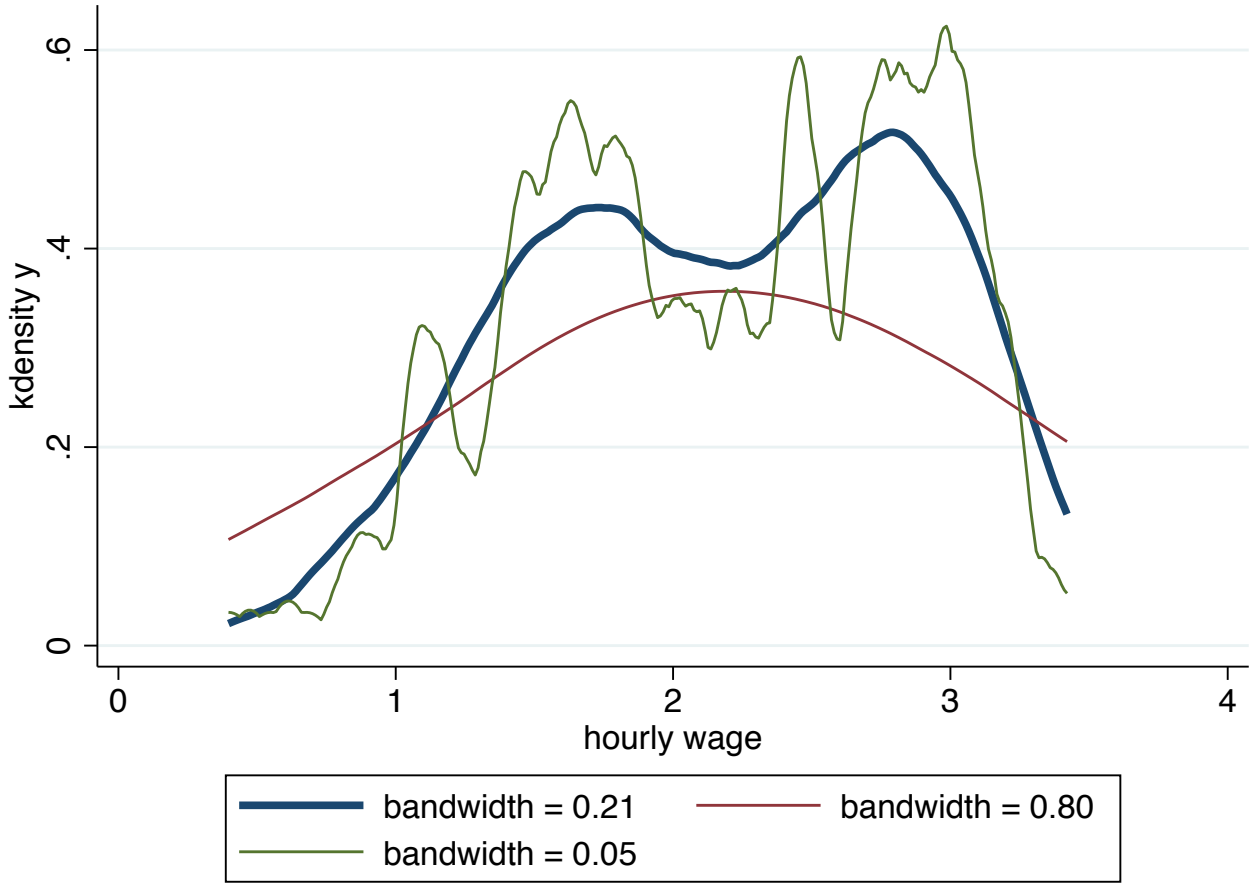
(STATA code) twoway (hist y, graphregion(color(white))) ///

——— (kdensity y, graphregion(color(white)) lwidth(thick))

graph export ”kernelhist.pdf”, replace

Example: kernel density estimation with different bandwidths

- Epanechnikov kernel, 0.21 is the STATA default



STATA code

```
twoway ///
——— (kdensity y, lwidth(thick) legend(label(1 ”bandwidth = 0.21”)))
///
——— (kdensity y, bw(.80) legend(label(2 ”bandwidth = 0.80”)))
///
——— (kdensity y, bw(.05) legend(label(3 ”bandwidth = 0.05”)))
///
, graphregion(color(white)) xtitle(hourly wage)
graph export ”kernel3.pdf”, replace
```

Bandwidth Choice: Optimal bandwidth

Choice of h is key

- Choice of h determines bias/smoothness (variance)
  - A large h over-smooths (less variance) but more bias
  - A small h under-smooths (more noise/variability) but potentially less bias.

→ Tradeoff between bias and variance

- A natural metric to evaluate performance of $\hat{f}(.)$: mean-squared error (MSE)

- MSE: (expectation is taken with respect to the density $f(x)$)

$$MSE[f(x_0)] = E[(f(x_0) - \hat{f}(x_0))^2]$$

- why is this measure appropriate?

- Recall that the MSE can be rewritten as the sum of the variance of the estimator and the square of the bias

- therefore by minimizing this quantity we take into account potential tradeoffs

- Notice that $MSE[f(x_0)]$ is a local criterion (defined for $x = x_0$)

- Let's make it global: define MISE, mean integrated squared error

- MISE: a global measure, it's the integral of the MSE over all values of x.

$$MISE(h) = \int MSE(\hat{f}(x_0))dx_0$$

Optimal bandwidth

Optimal bandwidth: minimizes MISE.

- Optimal bandwidth: obtained by differentiating MISE(h) with respect to h and solving for h. It yiedls:

$$h^* = \delta \left(\int f''(x_0)^2 dx_0 \right)^{-0.2} N^{-0.2}.$$

- A few considerations

- As $N \to \infty$, $h^* \to 0$. (Why? notice that $h^* = O(N^{-0.2})$).

- This means that as the sample grows, we should consider smaller and smaller intervals around each point (because as N grows, intervals become more "dense" in observations).

- The parameter δ is fully determined by the kernel employed. It's defined as:

$$\delta = \left(\frac{\int K(z)^2 dz}{(\int z^2 K(z) dz)^2} \right)^{-0.2}.$$

- The term $f''(x_0)$ is tricky: it's the second derivative of the function we try to estimate!

Estimation of the optimal bandwidth
Silverman's Plug-in Bandwidth estimate

- To obtain an estimate of the optimal bandwidth we should replace the "unknowns" in the previous formula by some estimates.

- δ: it's fully determined by the chosen Kernel (see table for values)

Table 9.1. *Kernel Functions: Commonly Used Examples*[a]

Kernel	Kernel Function $K(z)$	δ
Uniform (or box or rectangular)	$\frac{1}{2} \times \mathbf{1}(\lvert z\rvert < 1)$	1.3510
Triangular (or triangle)	$(1 - \lvert z\rvert) \times \mathbf{1}(\lvert z\rvert < 1)$	–
Epanechnikov (or quadratic)	$\frac{3}{4}(1 - z^2) \times \mathbf{1}(\lvert z\rvert < 1)$	1.7188
Quartic (or biweight)	$\frac{15}{16}(1 - z^2)^2 \times \mathbf{1}(\lvert z\rvert < 1)$	2.0362
Triweight	$\frac{35}{32}(1 - z^2)^3 \times \mathbf{1}(\lvert z\rvert < 1)$	2.3122
Tricubic	$\frac{70}{81}(1 - \lvert z\rvert^3)^3 \times \mathbf{1}(\lvert z\rvert < 1)$	–
Gaussian (or normal)	$(2\pi)^{-1/2} \exp(-z^2/2)$	0.7764
Fourth-order Gaussian	$\frac{1}{2}(3 - z)^2(2\pi)^{-1/2} \exp(-z^2/2)$	–
Fourth-order quartic	$\frac{15}{32}(3 - 10z^2 + 7z^4) \times \mathbf{1}(\lvert z\rvert < 1)$	–

[a] The constant δ is defined in (9.11) and is used to obtain Silverman's plug-in estimate given in (9.13).

- The problematic term is $f''(x_0)$. Why?

- The problematic term is $f''(x_0)$. Why?

- it's the second derivative of the function why are trying to estimate!

- Proposed solution: Assume X is normal.

- For a known f (normal), it's easy to compute the second derivative which becomes: $\int f''(x_0)^2 dx_0 = 3/8\sqrt{(\pi)\sigma^5}) = .2116/\sigma^5$

- Under this assumption the optimal bandwidth becomes:

$$\hat{h}^* = 1.364\delta N^{-0.2}\sigma$$

where σ: standard deviation of X.

- A plug-in estimate: replace σ by the sample standard deviation of X, s.

Estimation of the optimal bandwidth
Silverman's Plug-in Bandwidth estimate, II

- A correction: if there are outliers, s can be too large. This can make h^* too large (oversmoothing).

- Solution: use an alternative estimator of σ valid for normal distributions: $\sigma = iqr/1.349$, where iqr is the interquartile range.

- In practice choose:

$$min(s, iqr/1.349)$$

- Silverman's plug-in estimate: then

$$\hat{h}^* = 1.364\delta N^{-0.2} min(s, iqr/1.349)$$

Last remarks on Plug-in estimators

- Despite the normality assumption needed to compute f'', plug-in estimates for h work well in practice, especially for symmetric unimodal densities, even if $f(x)$ is not the normal density.

- Nonetheless, one should also check by using variations such as twice and half the plug-in estimate.

- Stata uses this plug-in estimate

Example 2

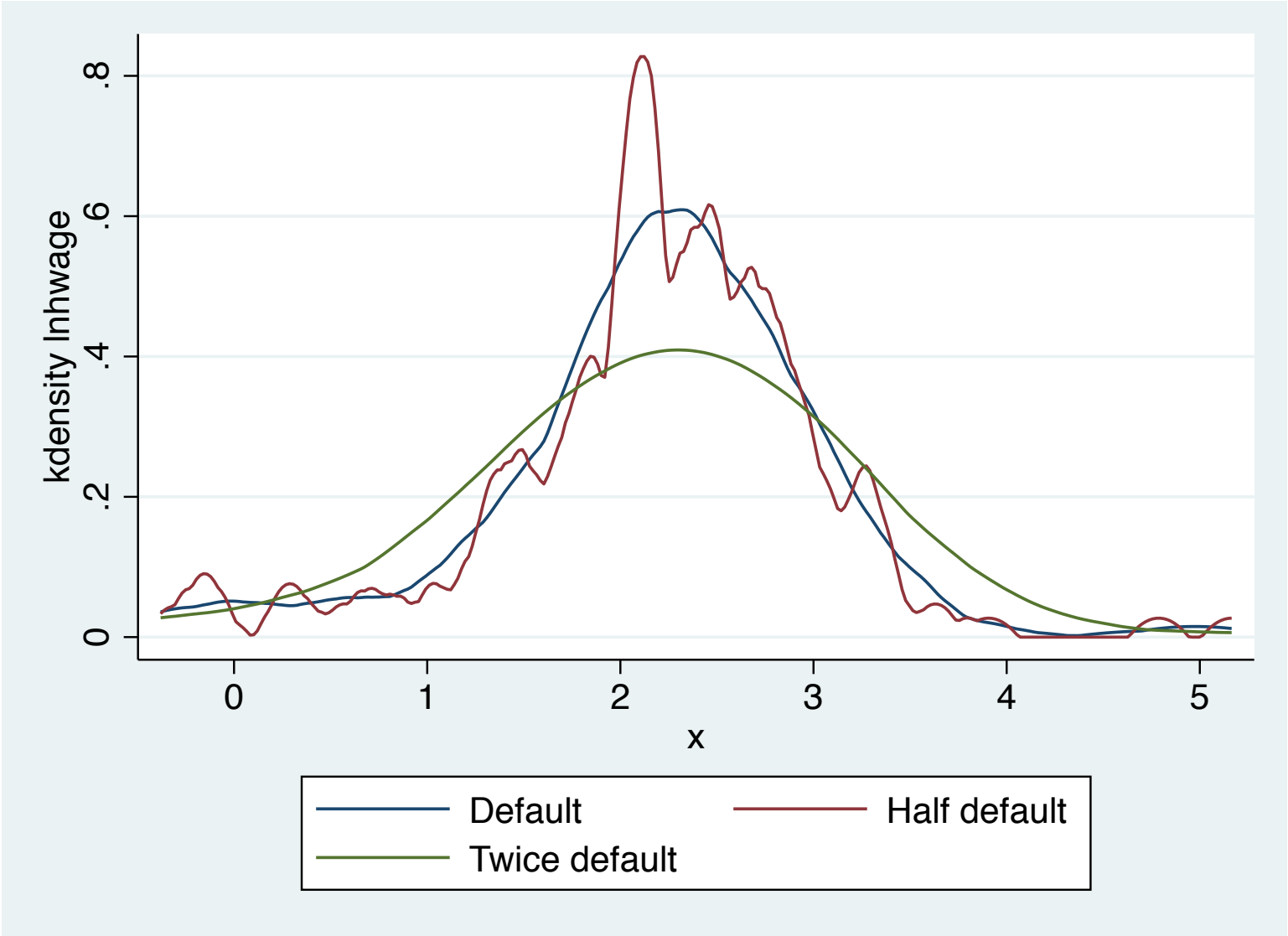
- Log wage data (same data used in Cameron & Trivedi).

- Kernel density of log wage, three values of h

- Stata's default kernel: epanechnikov

- default $h = 1.364\delta m/N^{0.2}$, m=min (std. dev. (x), interquartil-erange (x)/1.349), which yiels h=0.21

- Other values of h, h=.07 (oversmooths), 0.21 (default) and .63 (undersmooths)


kdensity lnhwage
0
.2
.4
.6
.8
0
1
2
3
4
5
x
Default
Half default
Twice default

- Stata code:

```
kdensity lnhwage, bw(0.21)

graph twoway (kdensity lnhwage, bw(0.21)) ///

(kdensity lnhwage, bw(0.07) clstyle(p2)) ///

(kdensity lnhwage, bw(0.63) clstyle(p3)), legend( label(1 "Default") ///

label(2 "Half default") label(3 "Twice default") ) scale(1.1)
```

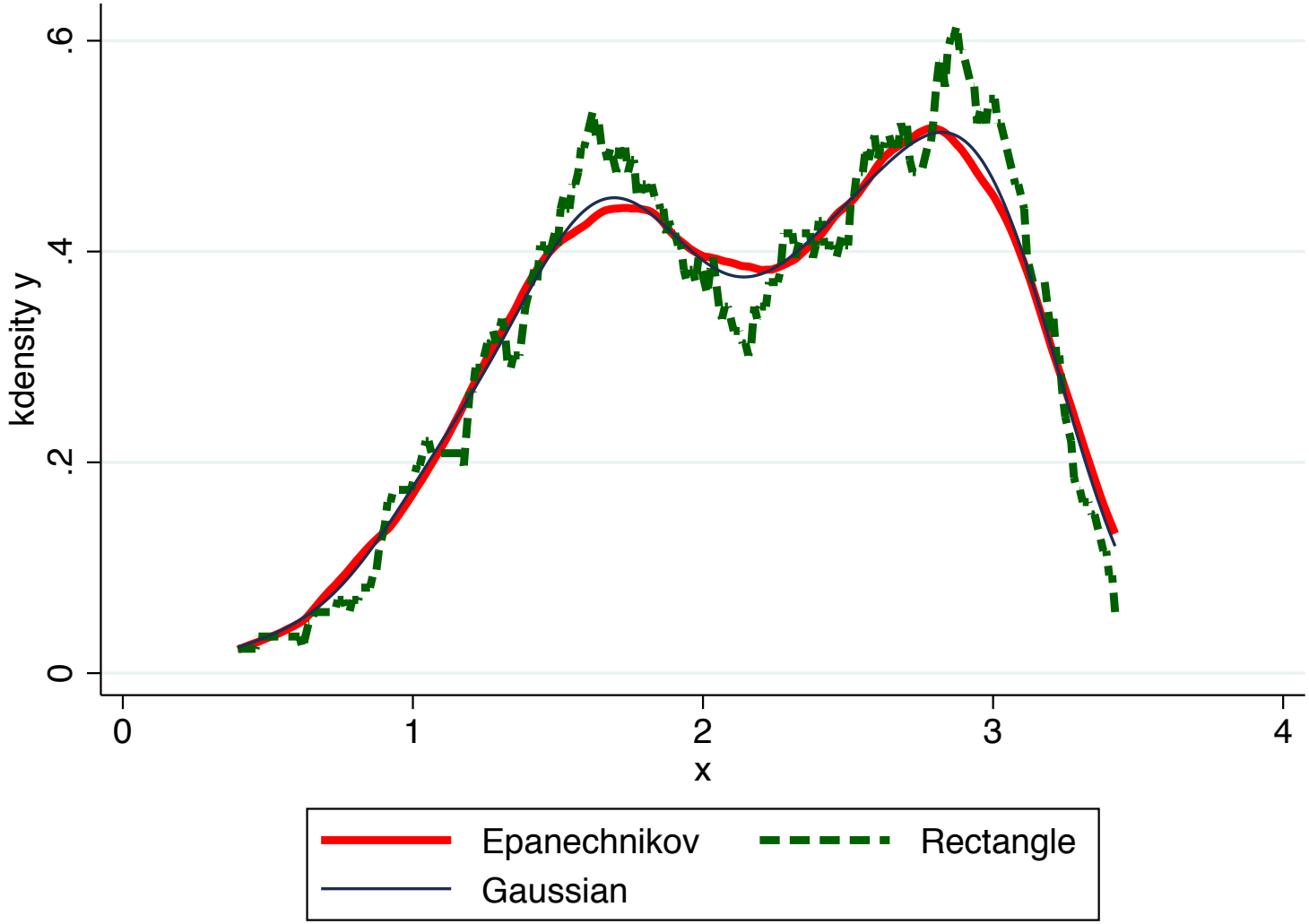
Choice of Kernel

- MISE(h^*) varies little across kernels.

- This means that after choosing h^*, choice of kernel is not that important

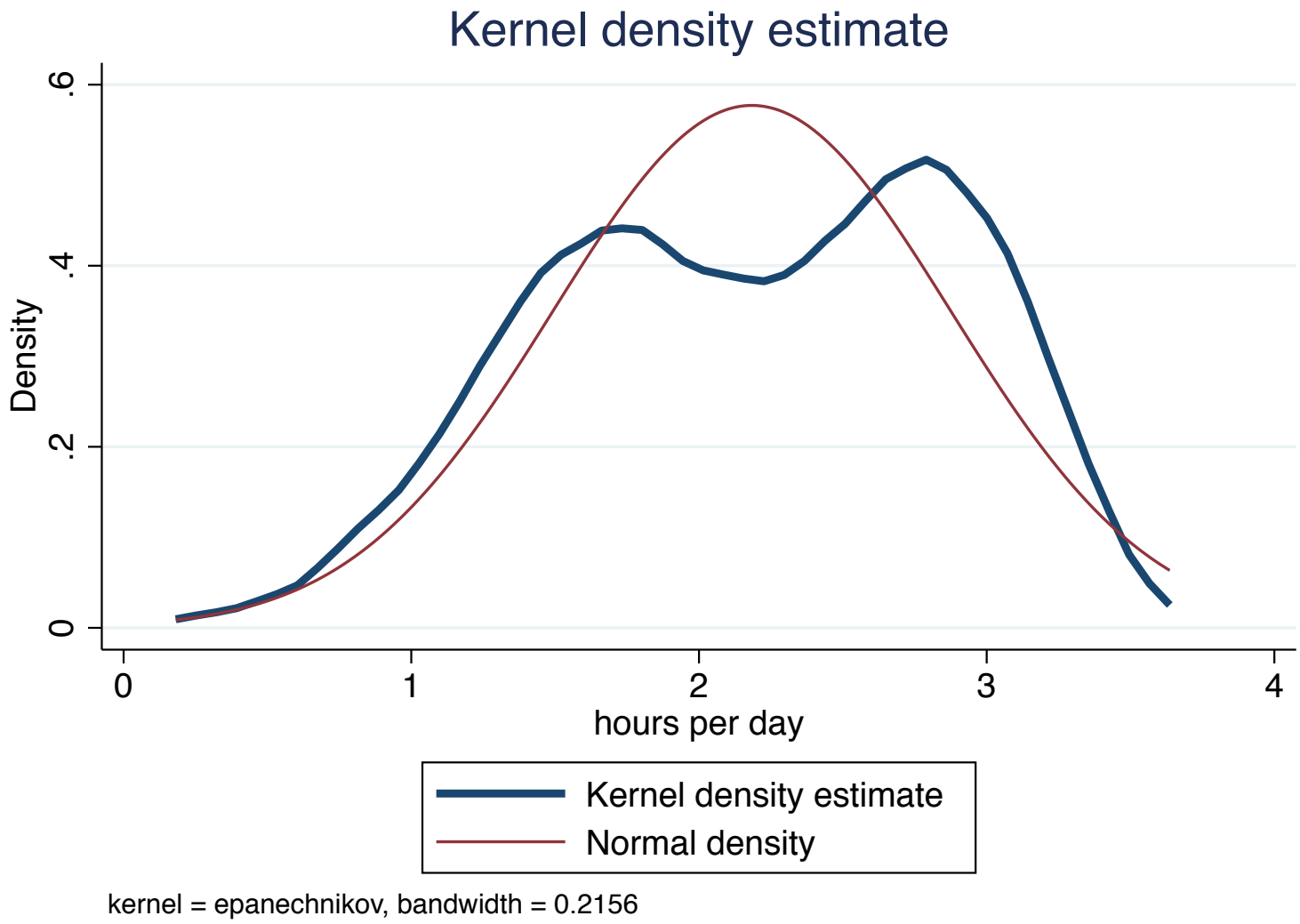
- Optimal Kernel: Epanechnivkov kernel, though the advantage is small.

Example: Kernel comparison



Do the data look normal?

- Add a normal distribution to the previous plot



STATA code

```
twoway (kdensity y, epan legend(label(1 Epanechnikov)) lcolor(red)
lw(thick)) ///

——— (kdensity y, rectangle legend(label(2 Rectangle)) lpattern(dash)
lw(thick) ///

lcolor(dkgreen)) ///

——— (kdensity y, gaussian legend(label(3 Gaussian)) lcolor(dknavy))
///

, graphregion(color(white))

graph export "kernel_comparison.pdf", replace

kdensity y, normal graphregion(color(white)) lwidth(thick)

graph export "kernel4.pdf", replace
```

Asymptotic properties

- How well does the kernel density estimator behaves?
  - Answering this question for a finite N is very difficult
  - Solution: consider the behavior for very large N (infinity) and derive properties in this case.
- Two basic properties

Asymptotic properties

- How well does the kernel density estimator behaves?
  - Answering this question for a finite N is very difficult
  - Solution: consider the behavior for very large N (infinity) and derive properties in this case.

- Two basic properties
  - Consistency: if $N \to \infty$, is the estimator "close" to the true value?
  - Asymptotic distribution: Provided the above is true, how can we carry out hypotheses testing?

Consistency

Notice that we're not estimating parameters, we're estimating a whole function. We can define two types of consistency, pointwise and uniform.

Pointwise Consistency:

- The kernel estimator is pointwise consistent at a particular point $(x = x_0)$, if both the bias and variance converge to zero.
- then, $\hat{f}(x_0) \xrightarrow{p} f(x_0)$ for each x_0
- This is achieved if $h \to 0$ and $Nh \to \infty$.
- Meaning:
- As N gets large, we should be focusing on very narrow windows around x_0 (h small)
- But we need sufficient points to compute the average at each point x_0 in order to have a consistent estimate (Nh large)

Consistency, II

Uniform Convergence:

- For estimation of $f(x)$ at all values of x, the stronger condition of uniform convergence is needed.
- Uniform convergence is defined as:

$$\sup_{x_0} \left| \hat{f}(x_0) - f(x_0) \right| \xrightarrow{p} 0,$$

  where this condition can be shown to occur if $\frac{Nh}{\ln N} \to \infty$.
- This requires h larger than for pointwise convergence

■ Uniform convergence is more stringent than pointwise convergence. h can be larger to achieve uniform convergence.

Last remarks:

- Important: these are conditionals needed to establish the asymptotic theory

- But for implementation they don't really provide a guideline about how to choose h

Unbiasedness

- The kernel density estimator is not unbiased in finite samples:

$$E(\hat{f}_N(x_0)) = f(x_0) + b_N(x_0), b(x_0) = 2h^2 f(x_0) \int z^2 K(z) dz$$

but notice if $h \to 0$, this tends to zero so it's asymptotically unbiased.

- However, we will see this term ($b(x_0)$) again when we describe the asymptotic distribution.

Asymptotic Normality

- What is the asymptotic distribution of the kernel density estimator?

Asymptotic Normality

- What is the asymptotic distribution of the kernel density estimator?

- In nonparametric statistics the distributions are often "non-standard" (i.e., non normal, χ^2, t or F).

- But: The kernel density estimator is asymptotically normal:

$$\sqrt{Nh}(\widehat{f(x_0)} - f(x_0) - b(x_0))) \xrightarrow{d} N\left(0, \int_{-\infty}^{\infty} f(x_0)K(z)^2 dz\right). \qquad (2)$$

and recall that $b(x_0) = 2h^2 f(x_0) \int z^2 K(z)dz = O(h^2)$

Asymptotic Normality, II

- Although normal, the distribution is different from the ones you are used to seeing in two respects:

  - The rate of convergence is not $\sqrt{N}$, as usual, but smaller ($\sqrt{Nh}$)

  - The bias term $b(x_0)$ appears in the asymptotic distribution.

Slower convergence rate: $\sqrt{Nh}$

- Recall that $h \to 0$, therefore, $\sqrt{N}h$ is smaller than $\sqrt{N}$, the usual rate of convergence of parametric estimators.

- What are the implications of slower convergence rates?

  - Converge to the normal distribution is slower, which means that in small samples the normal critical values might not very accurate.

  - Why? There are terms in the distribution function that converge to zero at a slow rate, so a very large N is needed for them to become negligible.

The bias term: $\sqrt{Nh}b(x_0)$

- What's the behaviour of the term $\sqrt{Nh}b(x_0)$ that appears in the asymptotic distribution?

- Recall that $N \to \infty$, and $b(x_0) = O(h^2)$, so $b(x_0) \to 0$ if $h \to 0$, therefore the answer is not obvious (i.e., the limit of the product of ∞ and zero is not obvious).

- Short answer: it depends.

- Medium-length answer: depending on the speed of convergence of h to zero, $\sqrt{Nh}b(x_0)$ can go to zero, it can converge to a bounded quantity or it can go to infinity!

- Why? consider a few examples

Example 1. Choose h as the optimal predictor, h^*

Recall:

- $h^* = 1.364\delta N^{-0.2}\sigma$, which implies that $h^* = O(N^{-0.2})$
- Recall that $b(x_0) = O((h^*)^2) = O(N^{-0.4})$, i.e.
- Interpretation: if the optimal bandwidth h^* is employed, then $b(x_0)$ tends to zero at a rate $N^{\cdot 4}$
- Hence, the term $\sqrt{Nh}b(x_0) = O(N^{0.4})O(N^{-0.4}) = O(1)$
- Interpretation: if h^* is employed, then $\sqrt{Nh}b(x_0)$ converges to a bounded quantity, different from zero.
- Therefore: the asymptotic distribution is not centered around the true value $f(x_0)$ but around $f(x_0) + b(x_0)$!

Example 2: undersmoothing

- Recall that undersmoothing implies choosing a value of h smaller than h^*.

- Consider for instance, $h = O(N^{-0.3})$. This tends to 0, and quicker (it's then smaller).

- Then, what's the limit of $\sqrt{Nh}b(x_0)$:

$$\sqrt{Nh}b(x_0) = O(N^{0.35})O(N^{-0.6}) \rightarrow ?$$

Example 3: oversmoothing

- Choose now $h = O(N^{-.1})$

- h tends to zero at a slower rate than h^*, therefore, it's larger: oversmoothing.

- Derive what happens in this case with the previous product.

Confidence intervals

- Kernel density estimates are usually presented without confidence intervals.

- But it is possible to construct pointwise confidence intervals for $f(x_0)$, where pointwise means evaluated at a particular value of x_0.

- How to do it:
  - Select a number of evaluation points x_0, that are evenly distributed over the range of x
  - plot these along with the estimated density curves.
  - (In principle) a 95% confidence interval for $f(x_0)$:

$$f(x_0) \in \left(\widehat{f(x_0)} - b(x_0) \pm 1.96 \times \sqrt{\frac{1}{\sqrt{Nh}}\widehat{f(x_0)} \int K(z)^2 dz}\right).$$

But...some problems

- Problem 1:

The bias: $b(x_0)$ is part of the confidence interval!

- Solutions

1. Ignore the bias. But then then the CI is not centered (not a great solution)

2. Estimate $f(x_0)$ but use a different h (undersmoothing) to estimate the CI

3. Estimate the bias and correct for it. But the estimate is noisy, so this is not great.

See Cameron and Trivedi for additional details.

4. Bootstrapping

- Problem 2: Slow convergence rates$\Rightarrow$ Normal critical values not very accurate for finite N.

- Solution: Use bootstrap

Note: But...What is boostrap?

- Alternative to using asymptotic distribution

- Uses resampling from a given data set to estimate the sampling distribution of a statistic (i.e., finite sample distribution).

- In practice, bootstrap

1) draws a random sample with replacement from the original data set,

2) calculates the statistic of interest (density at x_0 in this case);

3) repeats this many times to generate a large number of resamples.

- The distribution of the statistic across these resamples is used to estimate the sampling distribution of the statistic.

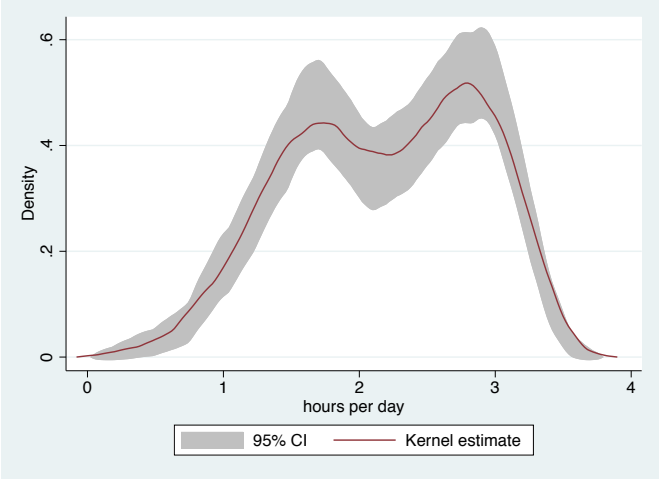
Confidence Bands in STATA

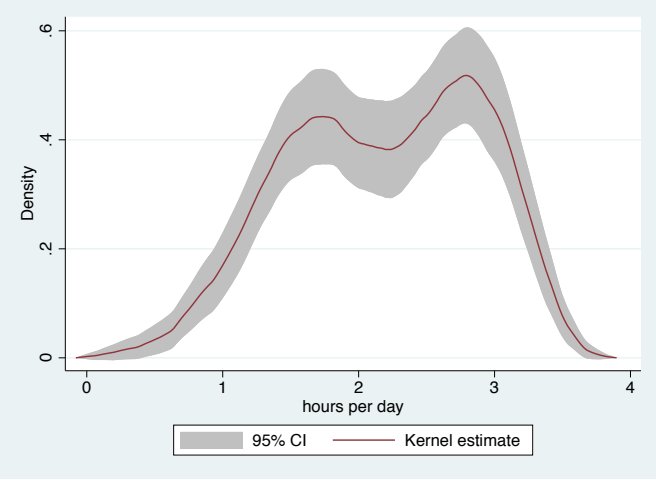
- The native command in STATA, Kdensity, doesn't allow you to do that
- But you can install Kdens

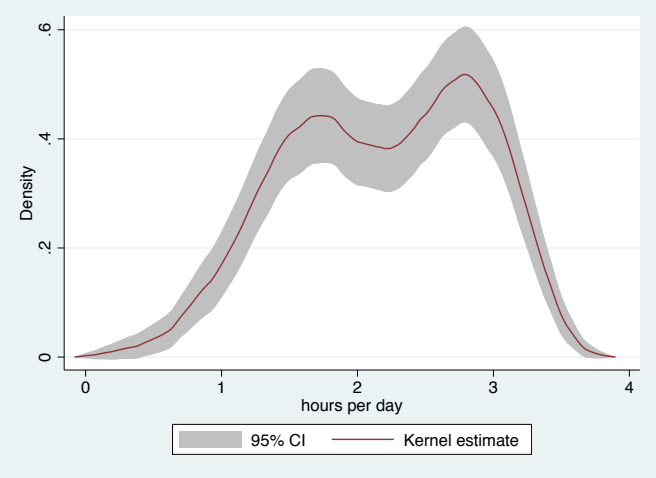
ssc install kdens

Examples: three types of confidence bands

- order: Asymptotic distribution, bootstrap and jacknife







STATA code

```
ssc install kdens

kdens y, ci vce(bootstrap, reps(200))

kdens y, ci usmooth

kdens y, ci vce(jackknife)
```

Multivariate Kernel density estimation:

- Conceptually it's straightforward to extend univariate nonparametric methods to multivariate settings.

- Consider the multivariate variable x, which now is a vector of dimension $k \times 1$. Then

$$\hat{f}(x_0) = \frac{1}{Nh^k} \sum_{i=1}^{N} K(\frac{x_i - x_0}{h})$$

- In practice it is problematic for at least two reasons:

1. A practical issue:

- Nonparametric methods are typically represented graphically.

- How to represent the results of a nonparametric analysis involving two or more variables is an important practical problem.

2. curse-of-dimensionality problem. when one-dimensional non-parametric methods are generalized to higher dimensions, their statistical properties deteriorate very rapidly because of the so-called

- The curse-of-dimensionality refers to the fact that the volume of data required to maintain a tolerable degree of statistical precision grows much faster than the number of variables under examination.

- For these reasons, simple generalizations of one-dimensional methods to the case of more than two or three variables tend to produce results that are difficult to represent and are too irregular, unless the size of the available data is very large.

Takeaway

- Nonparametric methods for density estimation
- Kernel estimates are easy to compute and work well in practice
- Choosing h, the bandwidth, is very important
- Once h is chosen, the choice of kernel is less important
- Statistical properties are worse that parametric methods: biases, lower rates of converge...
- Biased confidence bands if asymp. distribution is used. Other methods available
- Curse of dimensionality, when multiple variables are considered