Topics in Applied Econometrics for Public Policy

Master in Economics of Public Policy, BSE

Handout 1: Introduction to Non-Parametric Methods: Density estimators

Laura Mayoral

IAE and BSE

Barcelona, Spring 2023

1. Introduction

- Nonparametric methods: flexible estimation approaches
- minimal restrictions on the data generating process (DGP)
- We'll focus on two problems:
  - Density estimation: kernel density estimation.
  - Regression curve estimation: regression on a single scalar without imposing a functional form

- Non-parametric econometrics is a huge field

- These lectures 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 to study this methods more in depth:

  - "Non-parametric Econometrics" by Pagan and Ullah

  - "Non-parametric Econometrics: Theory and Practice" (the standard non-parametric econometrics textbook in graduate –PhD– level)

- Parametric methods: need of assumptions about the distribution of the distur- bance term (for instance, normality) or about the shape of the relation between the variables under analysis (for instance, linearity).

- Non-parametric methods: they require making none of these assumptions.

- Non-parametric methods are very useful to study relations between two variables.

- Including more and more variables in the analysis results in the errors. This is commonly known as the curse of dimensionality.

- Non-parametric estimates are typically presented as graphs because we're estimating functions, not parameters. Then, it becomes essential to produce nicely formatted graphs, etc.

Tradeoffs between parametric and nonparametric methods

- Parametric methods:
  - Pros: theory is simpler, they are more efficient, they allow us to study the relationship between many variables.
  - Cons: accuracy depends on validity of assumptions
- Nonparametric methods:
  - Pros: valid under very mild assumptions
  - Cons: less efficient, more complicated, curse of dimensionality as the number of regressors increase, there are also parameters to be chosen, and the results will depend on the values chosen (i.e., as in parametric models!).

Tradeoffs between parametric and nonparametric methods, II

- Since there are advantages and disadvantages in both types of methods, they are not substitutes, but complements

- Sometimes we might obtain a relationship using parametric methods but then check the validity of our assumptions estimating things again nonparametrically and checking that results are consistent.

- Another possibility: Semiparametric methods, compromise between parametric and non-parametric

- place some structure in the data but not fully-parametric; combine parametric and non-parametric methods

- Semi-parametric methods:

  - 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

- **Bootstrap:**

  - Bootstrap is often used to get standard errors.
  - More refined bootstraps can give better finite sample inference.

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.

- 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 $\hat{\theta} = T(\hat{f})$.

Why the density and not the distribution function?

- The distribution function (df) and the density function are equivalent ways of representing the distribution of X, but there may be advantages in analyzing a density:

- The graph of a density may be easier to interpret if one is interested in aspects such as symmetry or multimodality.

- Estimates of certain population parameters, such as the mode, are more easily obtained from an estimate of the density.

Alternative ways of estimating a density:

- Parametric density estimate:

- Assume a density and use estimated parameters of this density.

- For example, a normal density estimate: assume $X \sim N(\mu, \sigma^2)$ and use $N(\bar{X}, \sigma^2)$.

- Estimate parameters using the approaches you know.

- 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.

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

- 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 (replace probability by relative frequency):

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

where the indicator function is 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 range 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).

- 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.

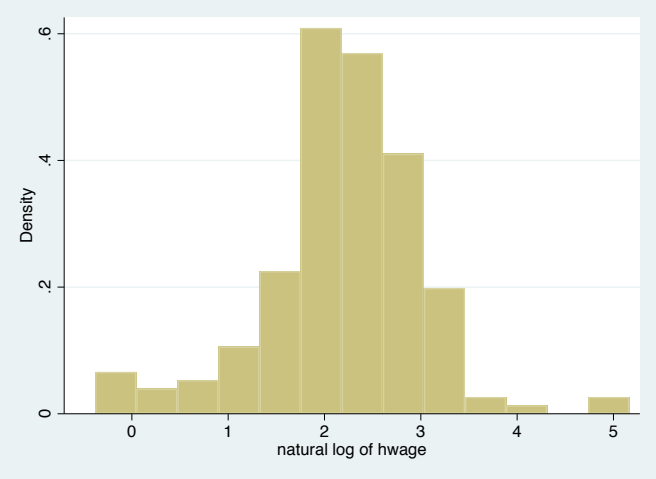
  - Stata: default is $\sqrt{N}$ for $N \leq 861$ and $10ln(N)/ln(10)$ for N>861

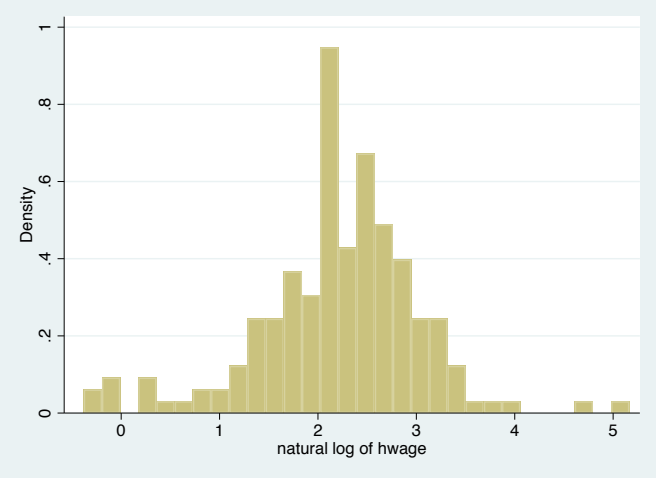
  - Two graphs: default and 30 bins, each of width 2h $\approx$ 0.20

  - STATA command:

histogram lnhwage, bin(30)

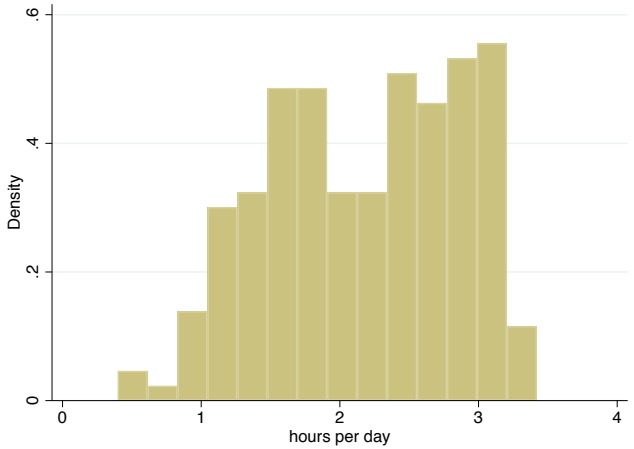
Histograms of log wage, with default number of bins and 30 bins.

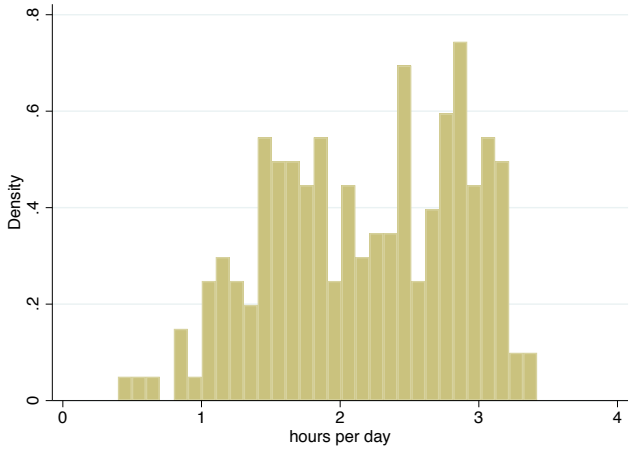


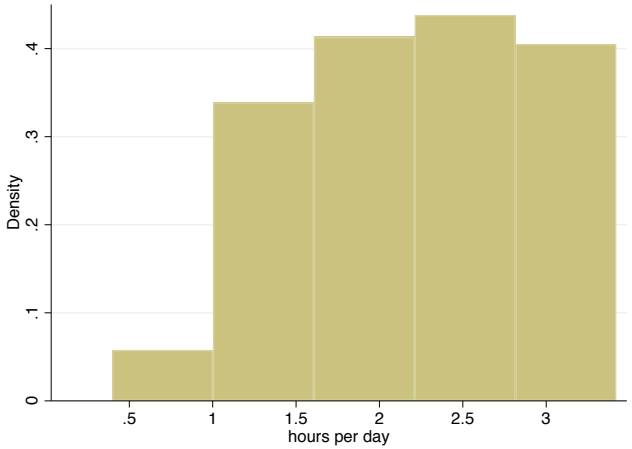


Stata code to generate and save some histograms

```
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

  - The results depend on the number of bins J or, equivalently, on the bin width $2h$.
    - 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.
  - The histogram is a step function 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

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

Notice that w(.) is the uniform density in the interval (-1,1), which means we're giving equal weights to the different observations

Now the generalization is 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

- Kernel function: a mathematical function that transforms data from one space to another.

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

- By computing these averages for each point in the sample we get smooth estimates.

Kernel Function, II

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

- If K integrates to one, then $\hat{f}$ also integrates to one.
- More generally, 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, III

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

- More formally, a kernel satisfies the following 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 (iii)(a) rather than just the weaker condition (iii)(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

- 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) Gaussian densities

- Unlike the uniform kernel, which takes constant positive value in the interval $[x_0 - h, x_0 + h)$ and vanishes outside this interval, the Gaussian kernel is always positive, reaches its maximum when $x_0 = x_i$, and tends to zero as $|x_0 - x_i| \to \infty$.

- 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

To estimate the density nonparametrically we need to choose h and K(.).

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

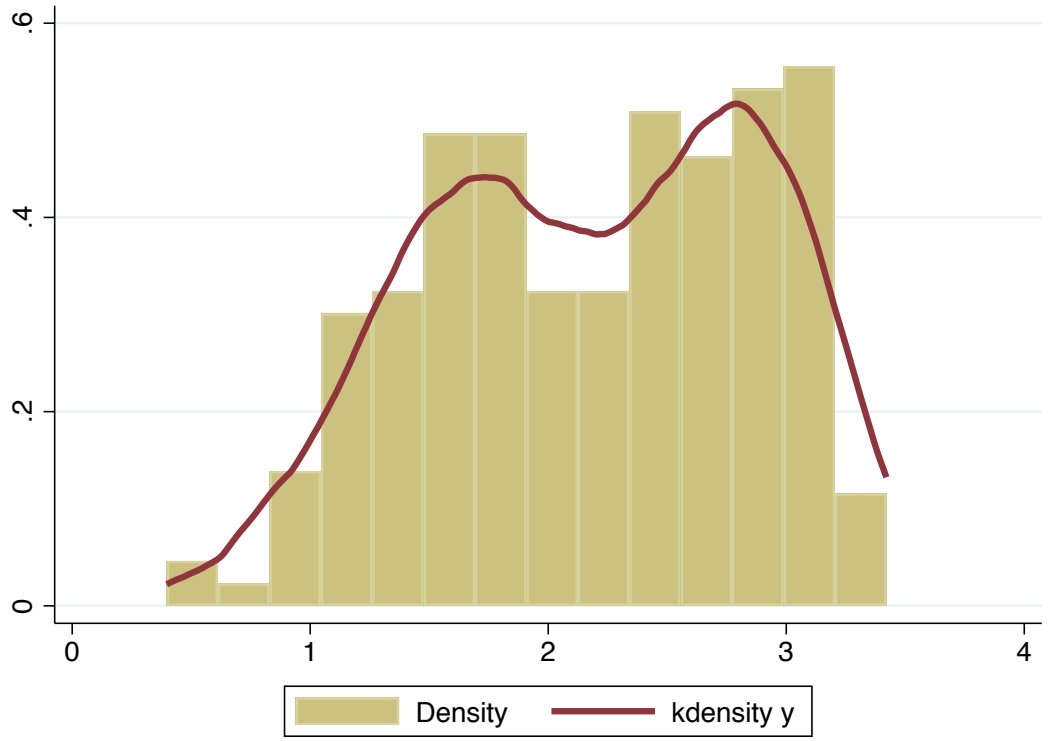
- The bandwidth h:

- 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

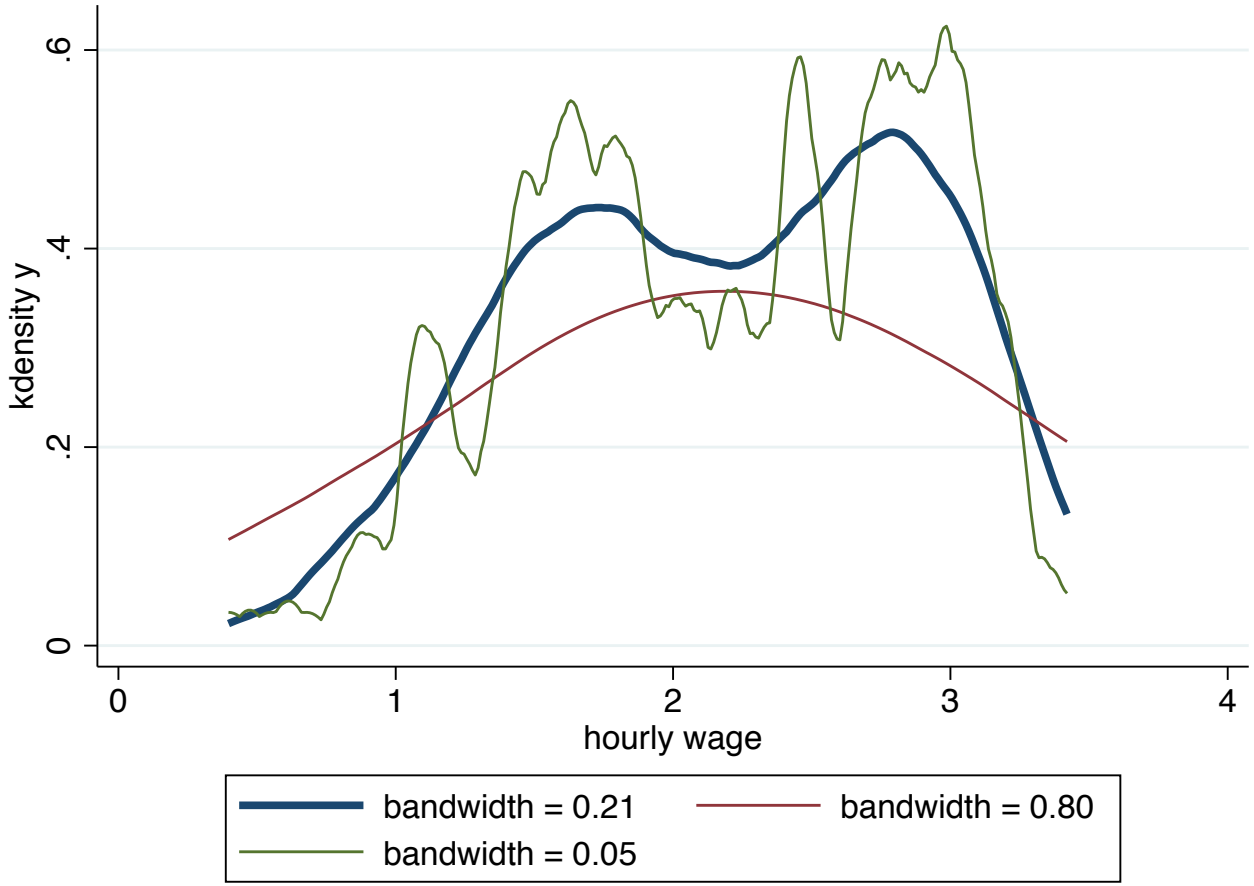
(STATA code) twoway (hist y, graphregion(color(white))) ///

——— (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 $f(x)$)

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

- Recall from your previous econometric courses that the MSE is the sum of the variance and the bias and therefore by minimizing this quantity we take into account potential tradeoffs

- Notice that this is a local criterion (defined for $x = x_0$)

- Let's make this a global criterion: define MISE, mean integrated squared error

- MISE: a global measure, integrate over all values of x.

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

Optimal bandwidth

Optimal bandwidth: minimizes MISE.

- Differentiating MISE(h) with respect to h and solving for h yields:

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

- The parameter δ depends on the kernel function used:

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

Values of δ (from Cameron and Trivedi)

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).

Computing the optimal bandwidth: Silverman's Plug-in Bandwidth estimate

- It's a simple formula that depends on N (sample size) and s (standard deviation of X).

- Assume X is normal, then it's easy to compute $\int f''(x_0)^2 dx_0 = 3/8\sqrt{(\pi)}\sigma^5) = .2116/\sigma^5$

$$h^* = 1.363\delta N^{-0.2} s$$

where s: standard deviation of X.

- A correction: if there are outliers, s can be too large. This can make h^* too large. In practice rather than using s it's used

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

iqr: interquartile range

- 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

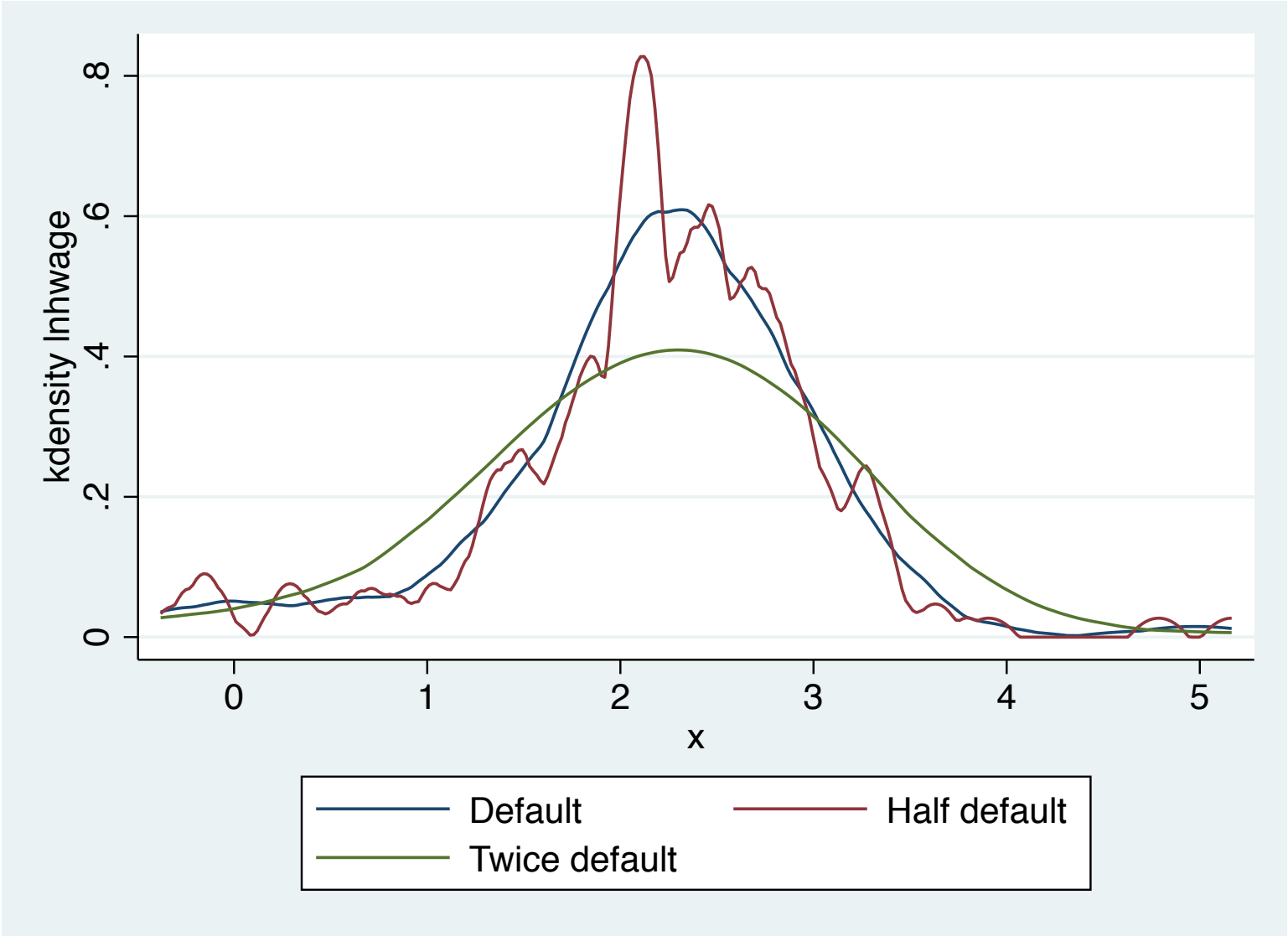
- Log wage data (same data used in Cameron & Trivedi).

- Kernel density of log wage, three values of h

- Stata's default kernel: epanechnikov

- default $h = 0.9m/N^{0.2}$, 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)
```

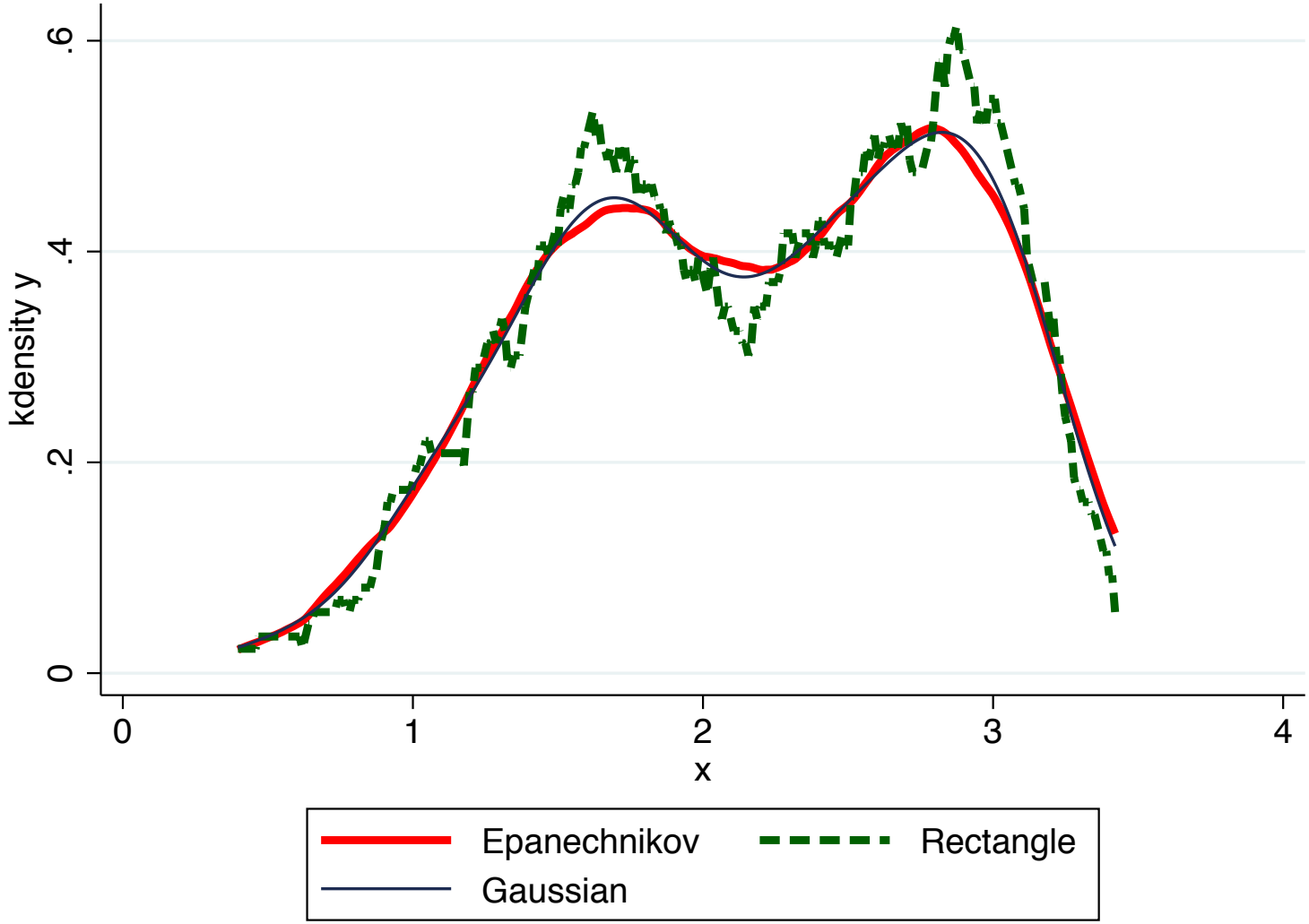
Choice of Kernel

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

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

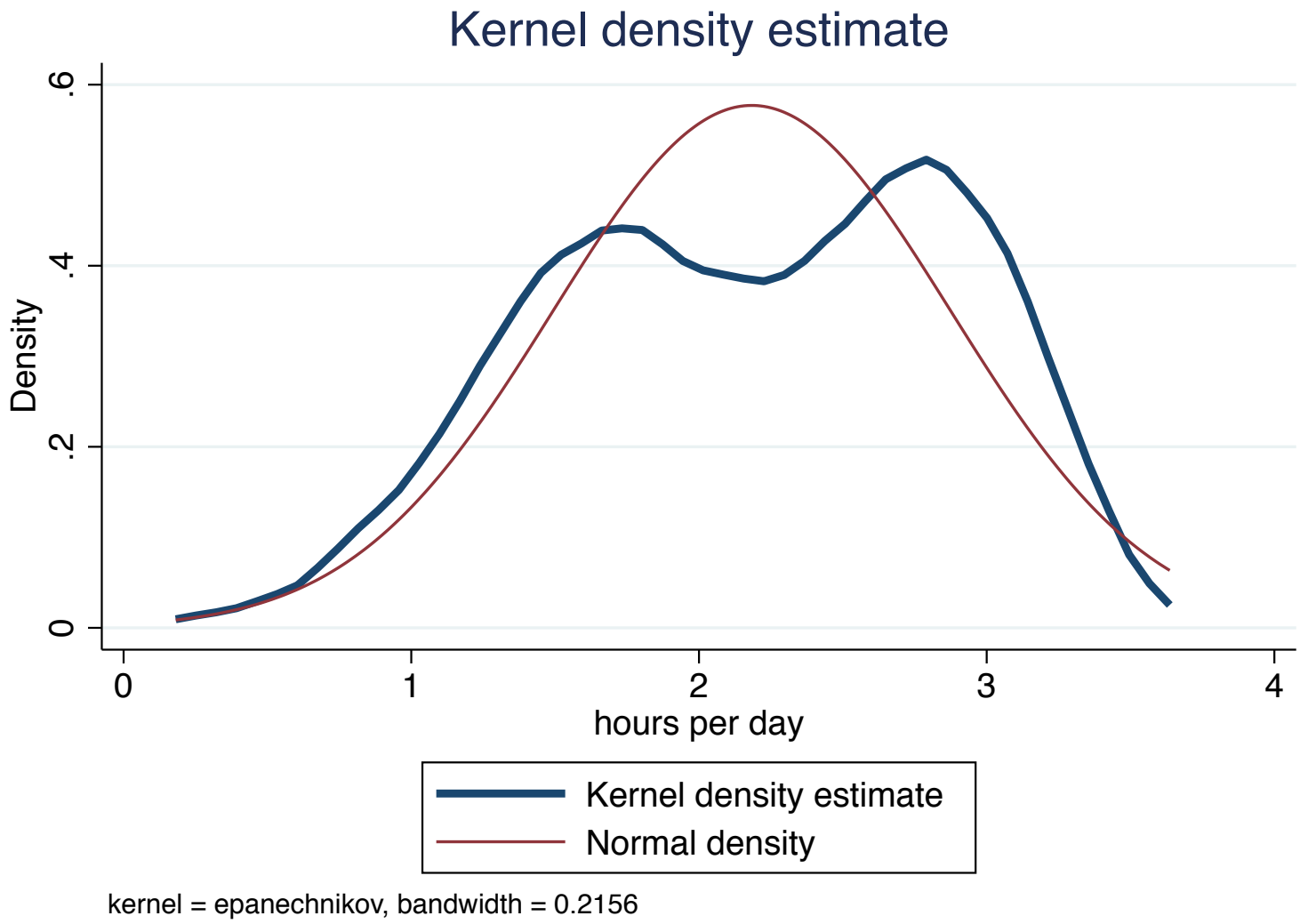
- Optimal Kernel: Epanechnivkov kernel, though the advantage is small.

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

Consistency

- $\widehat{f(x)}$ is a function, not a parameter $\rightarrow$ difference between point-wise convergence and uniform convergence

- In simple terms:

- pointwise convergence means that the sequence of functions converges to the limit function point by point (for all points).

- uniform convergence is a stronger form of convergence. It means that the sequence of functions converges to the limit function uniformly (at the same rate) over the entire domain, which ensures that the limit function is the same at all points.

- Definition:
- Pointwise convergence: $\widehat{f(x_0)} - f(x_0) \xrightarrow{p} 0$, for all points x_0
- Uniform converge: $\sup_{x_0} \left|\widehat{f(x_0)} - f(x_0)\right| \xrightarrow{p} 0$

- The kernel estimator is consistent (pointwise and uniformly) provided h is chosen "appropriately"

- The Kernel estimator is pointwise consistent if $h \to 0$ and $Nh \to \infty$ as $N \to \infty$

- The Kernel estimator is uniformly consistent if $h \to 0$ and $Nh/lnN \to \infty$ (h needs to be larger in this case)

- 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)

- 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

Asymptotic Normality

- The kernel density estimate is asymptotically normal. But:

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

- This slower convergence rates has several negative implications
  - Converge to the normal distribution is slower, which means that in small samples the normal critical values might not very accurate.
  - Although consistent, the kernel estimator is biased and this bias shows up in the asymptotic distribution.

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

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

- How the two problems mentioned above affect the computation of CI:

- Problem 1: 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.

- Problem 2:

The CI above CANNOT BE COMPUTED! notice the bias: $b(x_0)$ (unknown)

- Solutions

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

2. undersmoothing (small h)

3. use higher order Kernels (fourth order Gaussian/cubic)

4. Estimate the bias and correct for it.

See Cameron and Trivedi for additional details.

- Bootstrapping and all the options above can be combined!

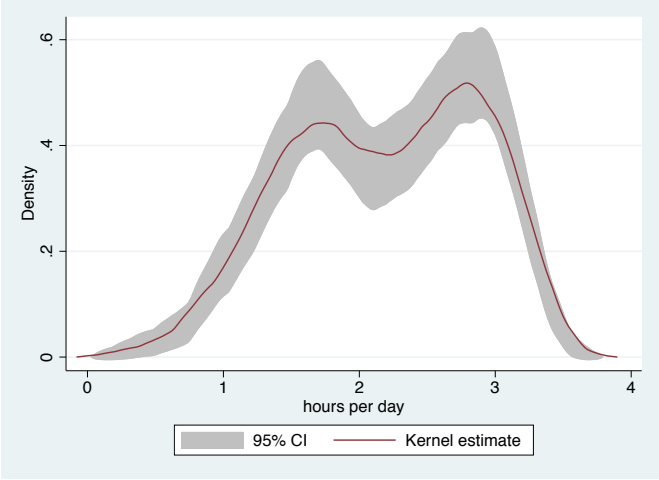
Confidence Bands in STATA

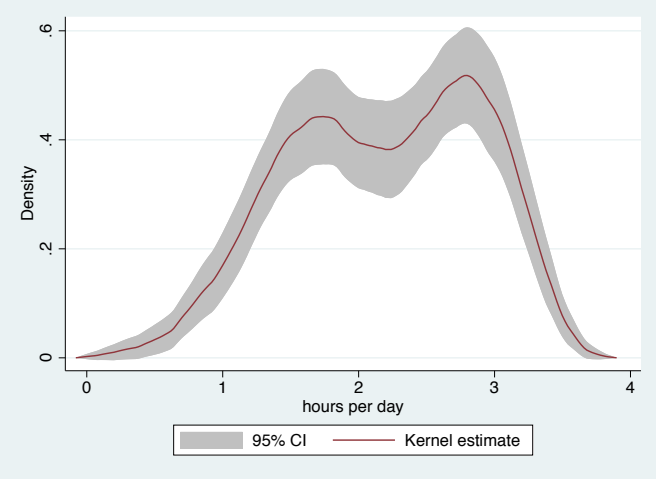
- The native command in STATA, Kdensity, doesn't allow you to do that
- But you can install Kdens

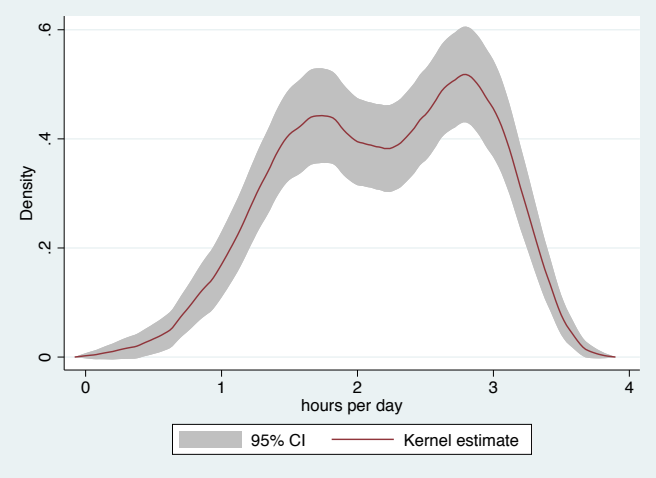
ssc install kdens

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)
```

Joint density estimation:

The curse-of-dimensionality problem

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

- In practice 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. More importantly: when one-dimensional nonparametric methods are generalized to higher dimensions, their statistical properties deteriorate very rapidly because of the so-called curse-of-dimensionality problem

- This 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

- This handout: 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