Nonparametric estimation

Giselle Montamat

Harvard University

Spring 2020

Nonparametrics

Review of methods that aim to estimate:

1. A density function, $f(x)$
   - Empirical distribution
   - Histogram
   - Kernel density estimators $\Rightarrow$ Tuning parameter: bandwidth h
2. A conditional expectation, $m(x) = E[Y|X = x]$
   - Bin scatter
   - Kernel regression $\Rightarrow$ Tuning parameter: bandwidth h
   - Series regression $\Rightarrow$ Tuning parameter: number of series p
   - Local polynominal regression $\Rightarrow$ Tuning parameters: h and p

Review of criteria for choosing optimal tuning parameter:

- Eye-ball it
- Plug-in method
- Cross-validation

Density estimation

Goal: estimate the density $f(x)$ of a random variable X using iid data $X_1, ..., X_N$.

Ideally, want the nonparametric estimate of a pdf to satisfy: $\hat{f}(x) \in [0, 1]$ and $\sum_x \hat{f}(x) = 1$.

If X_i discrete (and not many support points):

Empirical distribution:

$$\hat{f}(x) = \frac{1}{N} \sum_i 1\{X_i = x\}$$

That is, the empirical frequency of the points in the support of X.
Satisfies:

$$\sqrt{N}(\hat{f}(x_0) - f(x_0)) \xrightarrow{d} N(0, f(x_0)(1 - f(x_0)))$$

Density estimation

If X_i continuous:

1. **Histogram**: splits the continuous support of X into a finite number of bins.
   $\Rightarrow$ But binning throws away info...

2. **Kernel density estimators**: similar to histograms but they output a pdf and there's an optimal way to pick the bandwidth (bin size).

Density estimation: kernel

Note that pdf of X satisfies:

$$f(x_0) = \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h}$$

Plug-in principle then suggests:

$$\begin{aligned}
\hat{f}(x_0) &= \frac{1}{2h}\left[\frac{1}{N}\sum_i 1\{X_i \leq x_0 + h\} - \frac{1}{N}\sum_i 1\{X_i \leq x_0 - h\}\right] \\
&= \frac{1}{2h}\left[\frac{1}{N}\sum_i 1\{x_0 - h \leq X_i \leq x_0 + h\}\right] \\
&= \frac{1}{Nh}\left[\sum_i \frac{1}{2} 1\left\{\left|\frac{X_i - x_0}{h}\right| \leq 1\right\}\right] \\
&= \frac{1}{Nh}\sum_i K\left(\frac{X_i - x_0}{h}\right)
\end{aligned}$$

Where $K(.)$ is the kernel function and h the bandwidth.

In particular, this Kernel function is uniform, but there are other options...

Density estimation: kernel

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_i K\left(\frac{X_i - x_0}{h}\right)$$

Kernel function measures the proximity of X_i to x_0: whether $X_i \in [x_0 - h, x_0 + h]$ and, if so, weights according to how close to x_0.

Uniform kernel: assigns same weight $1/2$ to every $X_i \in [x_0 - h, x_0 + h]$.

$$K\left(\frac{X_i - x_0}{h}\right) = \begin{cases} 1/2 \text{ if } \left|\frac{X_i - x_0}{h}\right| \leq 1 \\ 0 \text{ if } \left|\frac{X_i - x_0}{h}\right| > 1 \end{cases}$$

Triangular kernel: assigns a positive weight to $X_i \in [x_0 - h, x_0 + h]$, and higher the closer to x_0.

$$K\left(\frac{X_i - x_0}{h}\right) = \begin{cases} \left(1 - \left|\frac{X_i - x_0}{h}\right|\right) & \text{if } \left|\frac{X_i - x_0}{h}\right| \leq 1 \\ 0 & \text{if } \left|\frac{X_i - x_0}{h}\right| > 1 \end{cases}$$

Density estimation: kernel

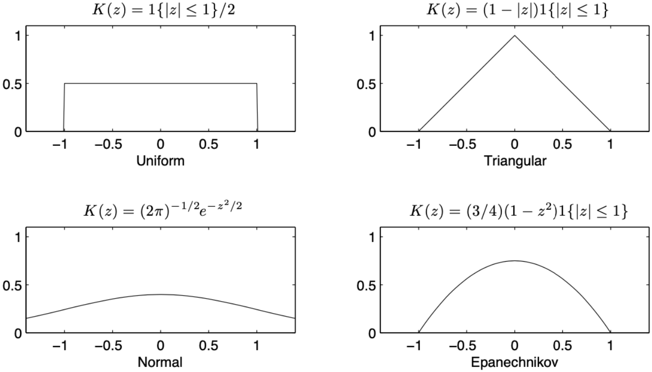
Epanechnikov kernel: assigns a positive weight to $X_i \in [x_0 - h, x_0 + h]$, and higher the closer to x_0.

$$K\left(\frac{X_i - x_0}{h}\right) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{X_i - x_0}{h}\right)^2\right) & \text{if } \left|\frac{X_i - x_0}{h}\right| \leq 1 \\ 0 & \text{if } \left|\frac{X_i - x_0}{h}\right| > 1 \end{cases}$$

Normal kernel: assigns a positive weight even to observations outside of $[x_0 - h, x_0 + h]$, and higher the closer to x_0.

$$K\left(\frac{X_i - x_0}{h}\right) = (2\pi)^{-1/2} e^{-\frac{1}{2}\left|\frac{X_i - x_0}{h}\right|^2}$$

Density estimation: kernel



Common kernels are symmetric density functions with mean zero. For such a kernel, the estimated density satisfies $\hat{f}(.) \geq 0$ and $\int \hat{f}(x)dx = 1$.

Density estimation: kernel

Can show bias and variance are:

$$b(\hat{f}(x_0)) \equiv E[\hat{f}(x_0)] - f(x_0) = \frac{h^2}{2} f''(x_0) \int z^2 K(z) dz + O(h^4)$$

$$Var(\hat{f}(x_0)) = \frac{1}{Nh} f(x_0) \int K(z)^2 dz + o\left(\frac{1}{Nh}\right)$$

Where $z = \frac{x - x_0}{h}$ and variance and expectation taken wrt X_i.

Notice variance-bias trade-off wrt h: small h (higher flexibility of model, "less smooth") reduces bias but increases variance.

$$MSE(\hat{f}(x_0)) = Var(\hat{f}(x_0)) + b(\hat{f}(x_0))^2$$

Note: MSE is a function of x_0. Epanechnikov kernel minimizes the MSE.

Density estimation: kernel

Consistency: If $N \to \infty$, $h \to 0$ and $Nh \to \infty$:

$$b(\hat{f}(x_0)) \to 0 \;\; ; \;\; Var(\hat{f}(x_0)) \to 0 \;\; ; \;\; \hat{f}(x_0) \xrightarrow{p} f(x_0)$$

Asymptotic normality: If $N \to \infty$, $h \to 0$ and $Nh \to \infty$:

$$\sqrt{Nh}(\hat{f}(x_0) - f(x_0) - b(x_0)) \xrightarrow{d} N\left(0, f(x_0)\int K(z)^2 dz)\right)$$

If, in addition, $\sqrt{Nh}b(x_0) \to 0$:

$$\sqrt{Nh}(\hat{f}(x_0) - f(x_0)) \xrightarrow{d} N\left(0, f(x_0)\int K(z)^2 dz)\right)$$

Condition satisfied if $\sqrt{Nh^5} \to 0$ (ie, h is small enough: "undersmoothing")

Density estimation: kernel

Choice of bandwidth h implies variance-bias trade-off:

- Large h: $\hat{f}(x_0)$ is smoother (low model flexibility). Low variance, high bias
- Small h: $\hat{f}(x_0)$ more jagged (high model flexibility). High variance, low bias

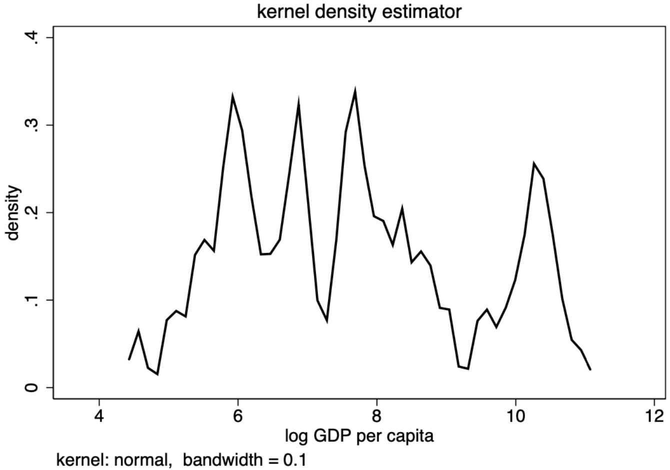
Optimal choice of h? Options:

1. Eye-ball it.
2. Plug-in methods.
3. Rules of thumb.
4. Cross-validation.

Density estimation: kernel

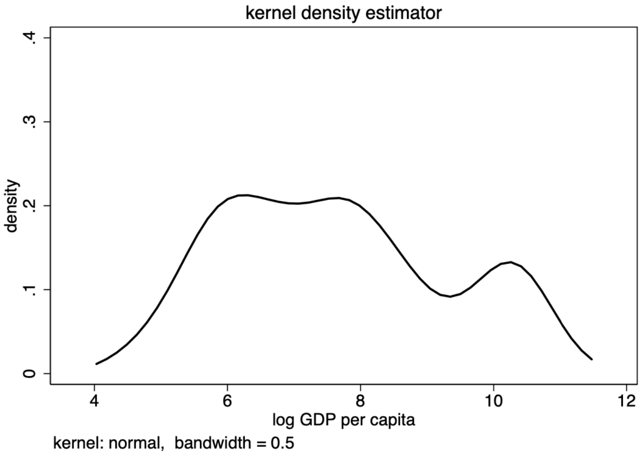
Example: world income per capita distribution.

Small h:



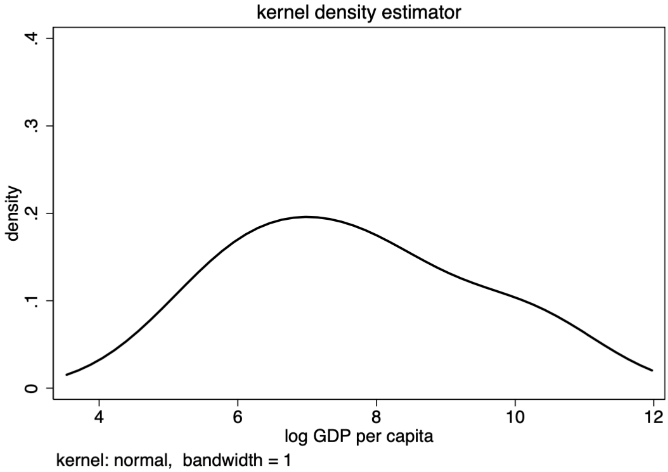
Density estimation: kernel

Medium h:



Density estimation: kernel

Large h:



Density estimation: kernel

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_i K\left(\frac{X_i - x_0}{h}\right)$$

Integrated mean squared error:

$$IMSE(\hat{f}) = \int MSE(\hat{f}(x_0))dx$$

Note: remember integrated risk under quadratic loss? The risk (which under quadratic loss is the MSE) was a function of θ, and the integrated risk integrated over θ. Well, this is the same idea, with x_0 as θ.

$$h^* = arg\min_h \; IMSE(\hat{f})$$

Result depends on f'', which we don't know, and $K(.)$, which we choose.

Density estimation: kernel

Plug-in method: estimate f'' using a first-pass bandwidth and then plug-in to the formula for f^*. But then need to find optimal bandwidth for this first pass, etc, etc.

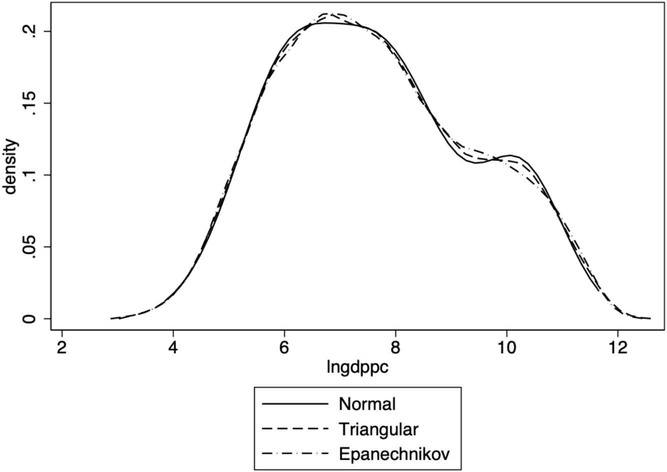
Rule of thumb: assume f is normal ("normal reference rule").

- If $K(.)$ normal:
$$h^* = \frac{1.059\sigma}{N^{1/5}}$$
- If $K(.)$ triangular:
$$h^* = \frac{2.576\sigma}{N^{1/5}}$$
- If $K(.)$ Epanechnikov:
$$h^* = \frac{2.345\sigma}{N^{1/5}}$$

$\hat{f}(x)$ is typically fairly insensitive to the choice of kernel, as long as the optimal bandwidth is used for each kernel.

Density estimation: kernel

Normal reference rule with different kernel functions:



Density estimation: kernel

Note that $\sqrt{N(h^*)^5} \to c > 0$, so bias doesn't dissapear in the asymptotic distribution. Would need a bandwidth smaller than these h^*, aka require *undersmoothing*.

Silvernman's rule of thumb:
The normal reference rule may *oversmooth* bimodal distributions. For a normal kernel, Silverman proposes to reduce the factor 1.059 to 0.9 and to use the minimum of two estimators of σ:

$$h^* = \frac{0.9min\{\hat{\sigma}, I\hat{Q}R/1.349\}}{N^{1/5}}$$

Where $\hat{\sigma}$ is the sample standard error and IQR is the interquartile range (and for a normal distribution $\sigma = IQR/1.349$).

Conditional expectation estimation

Goal: estimate $m(x) = E[Y|X = x]$ without taking a strong stand on the functional form of $m(x)$.

If X_i discrete (and not many support points):

Bin scatter:

1. Group the data points $X_1, ..., X_N$ into a finite number S of bins (like histogram).
2. Compute the average outcome Y in each bin.
3. Plot the average outcomes against the midpoint of each bin.

(Like regressing Y on S indicator functions that indicate if X_i is in the corresponding bin).

But if X continous, binning throws away info, doesn't yield an estimate of $m(x)$ for every possible x and not obvious how to pick the bins.

Conditional expectation estimation: kernel regression

If X_i continuous:

Kernel regression (Nadaraya-Watson):

It is weighted average:

$$\hat{m}(x_0) = \sum_i \underbrace{\frac{K\left(\frac{X_i - x_0}{h}\right)}{\sum_j K\left(\frac{X_j - x_0}{h}\right)}}_{\equiv w_i} Y_i$$

Where the weights w_i sum to 1, and observations closer to x_0 get larger weights.

Conditional expectation estimation: kernel regression

Say $X \in \mathbb{R}^k$.

Consistency: If $N \to \infty$, $h \to 0$ and $Nh^k \to \infty$ (+ regularity conditions):

$$\hat{m}(x_0) \xrightarrow{p} m(x_0) = E[Y|X = x_0]$$

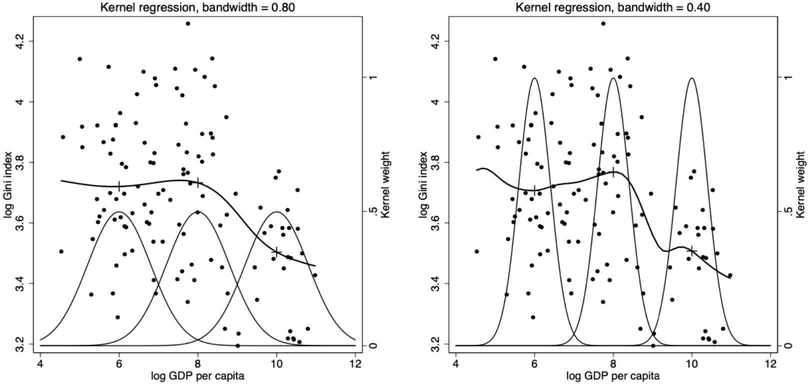
Asymptotic normality: If $N \to \infty$, $h \to 0$, $Nh \to \infty$ AND $Nh^{k+4} \to 0$ (which guarantees that the bias goes to 0, aka "undersmoothing"):

$$\sqrt{Nh^k}(\hat{m}(x_0) - m(x_0)) \xrightarrow{d} N\left(0, \frac{\sigma^2(x_0)}{f(x_0)} \int K(z)^2 dz\right)$$

Conditional expectation estimation: kernel regression

h is the tuning/smoothing parameter:

- Large h: regression is smoother (lower model flexibility)
- Small h: regression is more wiggly (higher model flexibility)



Optimal h? Options: eye-ball it, plug-in methods, cross-validation.

Conditional expectation estimation: kernel regression

Cross-validation:

Idea: choose h to minimize an estimate of the out-of-sample error.

$$h_{CV} = arg\min_{h} \; CV(h)$$

$$h_{CV} = arg\min_{h} \; \frac{1}{J}\sum_{j} \phi^{j}(h)$$

$$h_{CV} = arg\min_{h} \; \frac{1}{J}\sum_{j} \frac{1}{|I_j|}\sum_{i \in I_j}(Y_i - \hat{m}_{-j}(X_i))^2$$

$\hat{m}_{-j}$ is the kernel regression estimator that excludes observations in fold j.

Note that, if $J = N$:

$$h_{CV} = arg\min_{h} \; \frac{1}{N}\sum_{i}(Y_i - \hat{m}_{-i}(X_i))^2$$

$\hat{m}_{-i}$ is the kernel regression estimator that excludes observation i from the sample.

Exercise: Pset 11, Exercise 2 asks you to show that $CV(h)$ is indeed an unbiased estimator of the out-of-sample error.

Conditional expectation estimation: series regression

Series regression:

$$\hat{m}(x_0) = \hat{b}_0 + \hat{b}_1 x_0 + ... + \hat{b}_p x_0^p$$

Where:

$$\hat{b} = arg \min_{b_0,\ldots,b_p} \sum_i (Y_i - b_0 - b_1 X_i - b_2 X_i^2 ... - b_p X_i^p)^2$$

That is, it fits a polynominal of X_i of order p.

Note 1: Stome-Weirstrauss approximation theorem: any continuous $m(x)$ can be well approximated by linear combinations of polynomials over compact sets.

Note 2: this notation assumes for simplicity that x is scalar, but can extend to case where it has higher dimension.

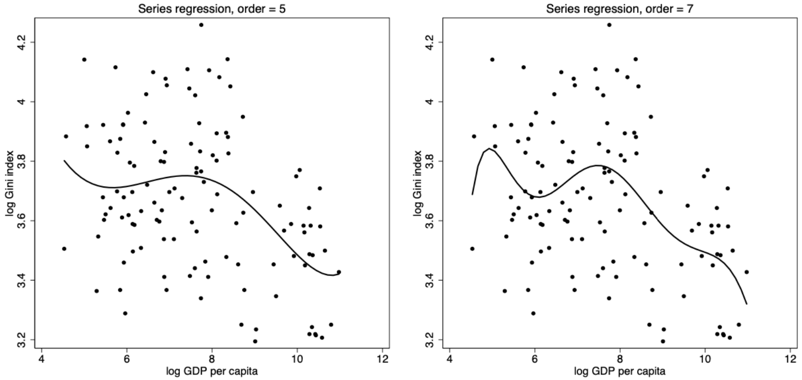
Conditional expectation estimation: kernel regression

Consistency: If $p \to \infty$ as $N \to \infty$ (and true $m(x)$ is smooth):

$$\hat{m}(x_0) \xrightarrow{p} m(x_0) = E[Y|X = x_0]$$

Note: p depends on N, denote as p_N. It is the tuning parameter:

- Small p_N: regression is smoother (lower model flexibility)
- Large p_N: regression is jagged (higher model flexibility)



Conditional expectation estimation: series regression

A combination of kernel regression and series regression...

Local polynominal regression:
For each x_0, compute $\hat{m}(x_0) = \hat{b}_0 + \hat{b}_1 x_0 + ... + b_p x_0^p$

$$\hat{b} = arg \min_{b_0,...,b_p} \sum_i K\left(\frac{X_i - x_0}{h}\right)(Y_i - b_0 - b_1 X_i - b_2 X_i^2 ... - b_p X_i^p)^2$$

That is, fit a polynominal regression locally around each point x_0.

- Kernel regression is particular case of local polynomial regression that uses $p = 0$.
- Series regression is a particular case of local polynomial regression that uses constant kernel.

Note: again, this notation assumes for simplicity that x is scalar, but can extend to case where it has higher dimension.

Conditional expectation estimation: series regression

Special case:

Local linear regression:

For each x_0, compute $\hat{m}(x_0) = \hat{b}_0 + \hat{b}_1 x_0$

$$\hat{b} = \arg\min_{b_0, b_1} \sum_i K\left(\frac{X_i - x_0}{h}\right)(Y_i - b_0 - b_1 X_i)^2$$