Appendix: Non-parametric Econometrics

Stanislao Maldonado

Universidad del Rosario
stanislao.maldonado@urosario.edu.co

Impact Evaluation
Universidad del Rosario
March 21th, 2017

1. Motivation

- Standard regression models covered so far assumed that the relationship between y and X is linear
- Non-parametric approaches relax the linearity assumption but this advantage does come with costs
- We will cover in this lecture:
  - Densities estimation
  - Non-parametric regression
- Identification of causal effects using non-parametric techniques can be specially useful when treatment is multivalued and the researcher is not willing to impose functional forms

- Even when the goal is not to establish a causal relationship, non-parametric approaches can potentially provide a better approximation to the CEF when it is no linear

2. Estimating densities

- Assume a researcher wants to estimate the labor income distribution Y. Let's assume that Y is a random variable such that:

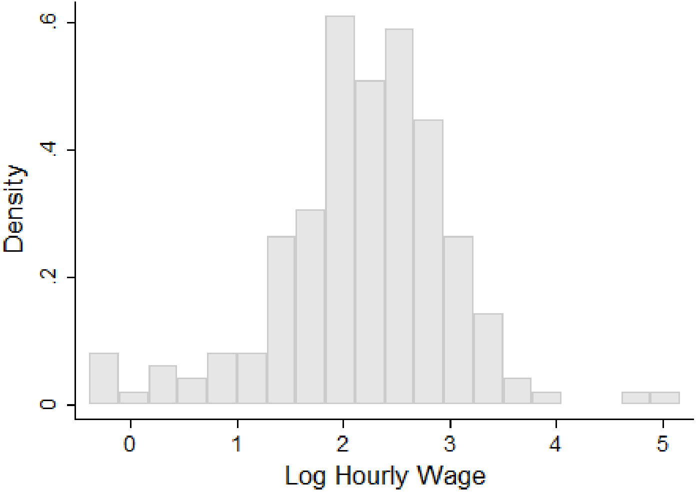
$$F(y) = Pr(Y \leq y) \tag{1}$$

$$f(y) = \frac{dF(y)}{dy} \tag{2}$$

- We want to estimate $f(y)$. In particular, we are interested in estimating $f(y_0)$ for an arbitrary point y_0

- What about using a simple histogram?
  - Divide the support of y into a bunch of evenly spaced bins
  - Count how many observations fall into each bin
  - Create a bar graph where the height of each bar is proportional to the number of observations that fall into the corresponding bin
- What is wrong about the histogram?
  - Unlike typical real density functions, it is discontinuous

Histogram for labor income in US(1993)



- Another option is to follow a parametric approach
  - Fit a theoretical distribution (Example: Pareto Distribution)

$$f(y;\alpha) = \begin{cases} \frac{\alpha y_m^{\alpha}}{y^{\alpha+1}} & \text{if } y \geq y_m \\ 0 & \text{if } y < y_m \end{cases}$$

  - $f(.)$ depends on α, so if an estimation of $\widehat{\alpha}$ were available, an estimation of $\widehat{f}(y^*, \widehat{\alpha})$ would be possible
  - The problem is parametric: estimating $\widehat{f}(y^*, \widehat{\alpha})$ is reduced to estimating $\widehat{\alpha}$
  - Problem: It requires to know $f(.)$, but in many situations is exactly what we wanted to know in the first place (income distribution)
- An estimator that is smooth as you move across different values of y would be typically preferred

- How do we get something like this?
  1. Take some window, say of radius 1 unit (total 2 units wide), and move it along the y-axis
  2. As we move it along, count how many observations fall within the window
  3. For any given point y^*, define the estimate of $f(y^*)$,$\widehat{f}(y^*)$, as the number of observations that fall within the moving window when it is centered at y^*
  4. Although this partially solves the smoothing problem, $\widehat{f}(y^*)$ will be somewhat discontinuous as observations drop into and out the window
  5. This can be solved by augmenting the moving window with a moving weighting function (known as *Kernel*) that places more weight on an observation as it moves closer to the center of the window

2.1 Kernel density estimation

- A kernel density estimator, introduced by Rosenblatt(1956), generalizes the histogram estimate by using an alternative weighting function:

$$\widehat{f}(y_0) = \frac{1}{n}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{Y_i - y_0}{h}\right) \tag{3}$$

- Where h is a parameter known as bandwidth and $K(.)$ satisfies the following conditions:
  1. $K(.)$ is symmetric around 0 and is continuous
  2. $K(z) \geq 0$
  3. $\int K(s)\mathrm{d}s = 1$
  4. $\int sK(s)\mathrm{d}s = 0$
  5. $\int s^2K(s)\mathrm{d}s = \mu < \infty$

- To illustrate how this estimator works, consider the case **triangle kernel**:
  - Suppose $h = 1$ and $K(z) = 1(|z| < 1)(1 - |z|)$(i.e., weight is a linear function of the distance)
  - At a given point y_0, we have:

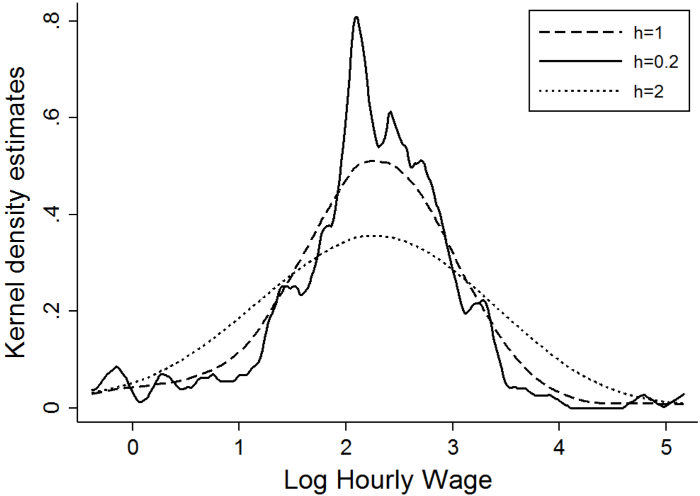
$$\widehat{f}(y_0) = \frac{1}{n}\sum_{i=1}^{N} 1(|Y_i - y_0| < 1)(1 - |Y_i - y_0|) \qquad (4)$$

  - Where $|Y_i - y_0|$ is the Euclidean distance metric

- $\widehat{f}(y_0)$ measures how close is on average each observation in the sample with respect to y_0
  - If Y_i falls more than one unit of distance away from y_0, it receives a weight of zero and has not input into the sum
  - If Y_i falls less than one unit of distance away from y_0, it receives a positive weight which is increasing as Y_i gets closer to y_0
  - Computing this summation at all possible values of support of the density leads to $\widehat{f}(y)$
- What about the role of h and $K(.)$?

- h plays a critical role in estimating the density:
  - What would happen if we choose a small h?
  - What would happen if we choose a large h?
- Trade-off between **bias** and **variance**!

Density estimates using a triangular Kernel with different h



- Less relevant is the role of $K(.)$
  - **Uniform** (or rectangular):

$$K(z) = \frac{1}{2}\mathbf{1}(|z| < 1) \tag{5}$$

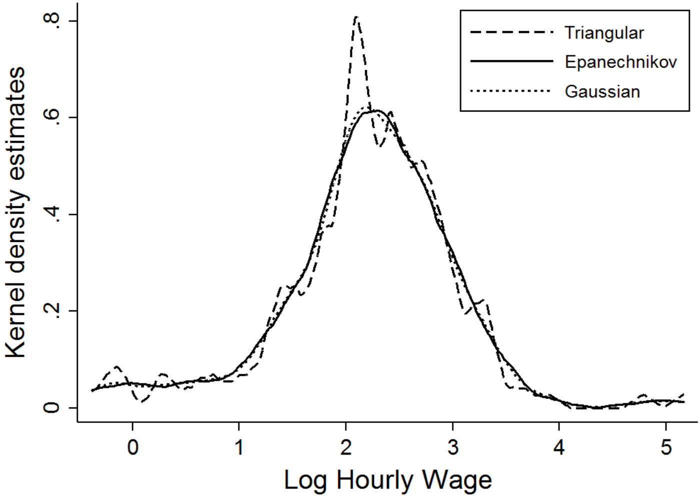
  - **Epanechnikov** (or quadratic)

$$K(z) = \frac{3}{4}(1 - z^2).\mathbf{1}(|z| < 1) \tag{6}$$

  - **Gaussian** (or normal)

$$K(z) = (2\pi)^{\frac{-1}{2}} exp(-z^2/2) \tag{7}$$

Density estimates using a different Kernel with $h = 0.2$



3. Non-parametric regression

- Assume a researcher wants to estimate the following model:

$$y_i = m(x_i) + u_i \tag{8}$$

- where $m(x_i)$ is an unknown function and $\mathbb{E}(u_i|x_i) = 0$ such that:

$$\mathbb{E}(y_i|x_i) = m(x_i) \tag{9}$$

- A researcher wants to estimate $m(x_i)$ using a sample (y_i, x_i)
- Recall the linear solution:

$$m(x_i) = \alpha + \beta x \tag{10}$$

- How to recover α and β without knowing the true functional form?

3.1 Kernel regression

- Suppose that for a distinct value of the regressor, say x_0, there are N_0 observations on y_i. An estimator for $m(x_0)$ is the sample average for these N_0 values
- Problem: this estimator might be too noisy in finite samples if N_0 is too small
- A way to solve this issue is looking not only to values of y_i for x_0, but also the observed values of y_i when x is close to x_0

- This leads to the **local weighted average estimator**:

$$\widehat{m}(x_0) = \sum_{i=1}^{N} w(x_i, x_0, h).y_i = \frac{\frac{1}{Nh}\sum_{i=1}^{N} 1\left(\frac{x_i - x_0}{h}\right) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} 1\left(\frac{x_i - x_0}{h}\right)} \quad (11)$$

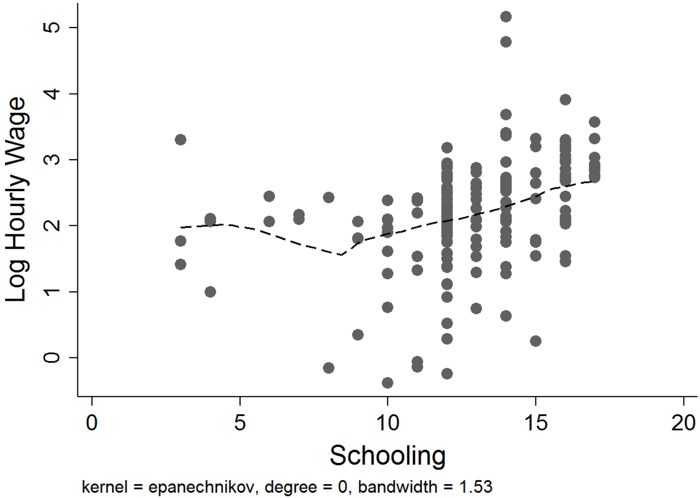
- Problem: this estimator gives equal weights to all observations close to x_0

- This leads to the **Kernel regression estimator**:

$$\widehat{m}(x_0) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)} \tag{12}$$

- This estimator is also known as the **Nadaraya-Watson**
- h play a similar role as in the case of densities

Example: Kernel regression



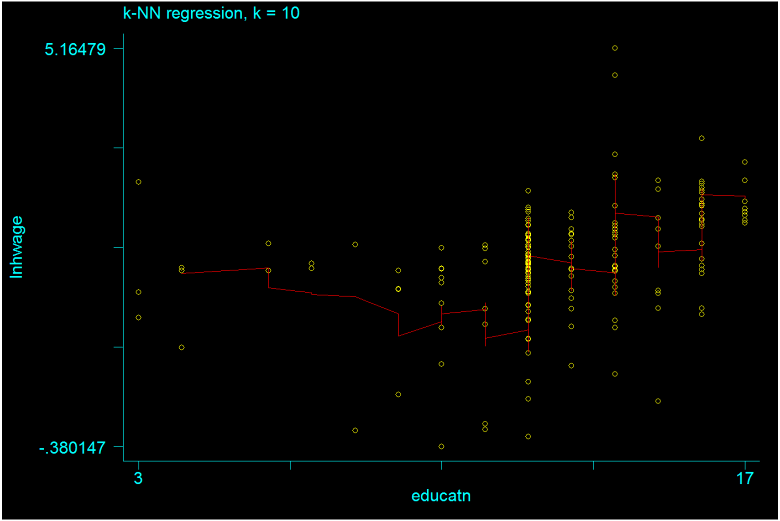
3.2 Nearest neighbors regression

- Consider alternative weights. For instance, the researcher could fix the number of observations around a given point to be included in the analysis
- **k-nearest neighbor estimator** is the equally weighted average of y values for the k observations of x_i closest to x_0

$$\widehat{m}_{k-NN}(x_0) = \frac{1}{k}\sum_{i=1}^{N} 1(x_i \in N_k(x_0)).y_i \tag{13}$$

- This estimator is a Kernel estimator with uniform weights

Example: Nearest neighbors regression (k=10)



3.3 Local linear and local polynomial regression

- Kernel regression is a local constant estimator (it assumes $m(x_0)$ equals a constant in the local neighborhood of x_0)
- We can instead let $m(x_0)$ be linear in the local neighborhood of x_0:

$$m(x_0) = a_0 + b_0(x_i - x_0) \tag{14}$$

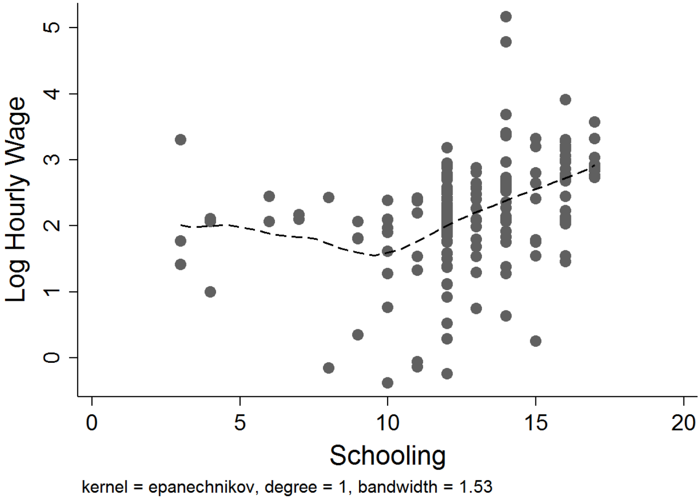
- The **local linear regression estimator** minimizes:

$$\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)(y_i - a_0 - b_0(x_i - x_0))^2 \tag{15}$$

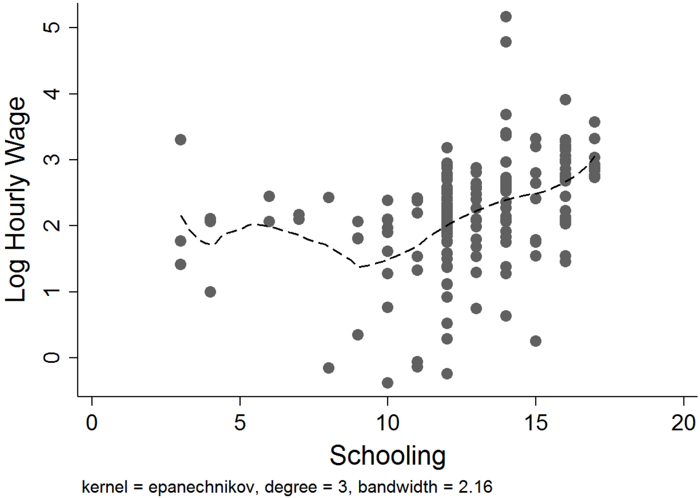
- This can be easily generalized to the case of **local polynomial estimator of degree p** which minimizes:

$$\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)(y_i - a_0 - a_1(x_i - x_0) - \ldots - a_p \frac{(x_i - x_0)^p}{p!})^2 \quad (16)$$

Example: Local linear regression



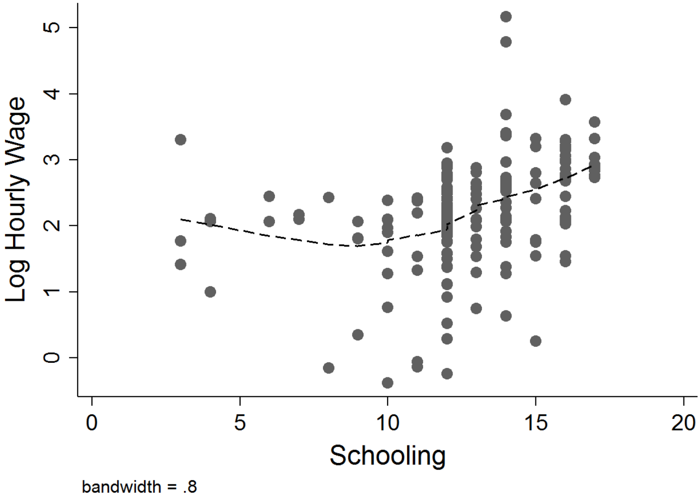
Example: Local polynomial regression (degree 3)



3.4 Locally weighted scatterplot smoothing (LOWESS) regression

- This is a popular variant of local polynomial estimation in (16):
    - Uses a variable bandwidth determined by the distance from x_0 to the k-th nearest neighbor
    - Uses a tricubic kernel
    - Downweights observations with large residuals

Example: LOWESS regression



4. The curse of dimensionality

- Non-parametric techniques typically work well with univariate distributions or with low-dimensional multivariate distributions
- It is hardly the case in most empirical research that scholars work with few covariates
- There is a curse of dimensionality:
  - Sparsity of the data grows exponentially with the number of covariates
  - The more dimensions, the less data to estimate conditional expectations
- There is an important trade-off to evaluate: a researcher can relax the linearity assumption, but it comes with the cost of the curse of dimensionality