Microeconometrics: Non-Parametric Estimators

Ethan Kaplan

1 Introduction to Non-Parametric Estimators

- Suppose we want to estimate a highly non-linear relation between two variables. How would we do it?
  - Estimate relation with set of orthogonal functions?
    - High order polynomials.
    - Trigonometric functions.
    - Problems?
      - Very sensitive to outliers.
      - Non-local impact of outliers
  - Splines (linear, quadratic, ...)

* Divide X into I different sections.. $1, 2, ...I$

$$\min_{\hat{\alpha}_i, \hat{\beta}_i} \sum_i \left(Y_{it} - \hat{\alpha}_i - \hat{\beta}_i X_{it} \right)$$

$$s.t.\ \hat{\alpha}_{i-1} + \hat{\beta}_{i-1} X_I = \hat{\alpha}_i + \hat{\beta}_i X_I$$

* Linear, Quadratic, Quartic, Trigonometric

* More local impacts but non-differentiable

– Non-parametric estimators

* Require tons of data - especially problematic with high dimensionality of estimation

* Must choose how locally to estimate (band-width)

– Semi-parametric estimators

* Requires greater functional form assumptions

* Better at dealing with high dimensional estimation

2 Histograms

- The histogram is a probability mass function which is usually an approximation to the probability density function (pdf) of a random varialbe.

- To create a histogram for a variable X, divide up X into K parts [0, X_k) (could be equal portion of X-space or any other division of the X-space).

- Then the histogram is:

$$f(x_k) = \sum_j \frac{I\left(X_{k-1} \leq x_j < X_k\right)}{X_k - X_{k-1}}$$

  where $I(\cdot)$ is the indicator function.

- If the histograms are of equal length in X-space, then we can write the density as:

$$f(x_k) = \sum_j \frac{I\left(X_k - h \leq x_j < X_k + h\right)}{2h}$$

- In the limit as $h \longrightarrow 0$, if the density (pdf) is differentiable, then you will recover the density.

3 Kernel Density Estimation

- The kernel density estimator is a generalization of the histogram - it is in general smoother.

- The histogram density is for a sample from the population. Often the sample is a noisy estimate of the population. Therefore, kernel densities smooth the density estimates between points using functions called kernel functions.

- The value of the estimator at a point x_o is

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

where N is the number of total points being used in the estimation of the density.

- $K(\cdot)$ is called the kernel function and it is what smooths the density. It must satisfy 4 conditions

  1. $K(z)$ is symmetric around zero and continuous
  2. $\int K(z)\,dz = 1$, $\int zK(z)\,dz = 0$, and $\int |z| K(z)\,dz < \infty$
  3. Either
     (a) $\exists z_0$ such that $K(z_0) = 0$ $\forall z$ such that $|z| \geq z_0$
     (b) $\lim_{z\to\infty} |z| K(z) = 0$
  4. $\int z^2 K(z)\,dz = c < \infty$

- Usually kernel functions satisfy $(3a.)$ not just $(3b.)$

- Usually $\frac{\partial K}{\partial |z|} \leq 0$ so that the impact of data points z_k on the value of the non-parametric estimator at a point z_0 decline with distance between z_0 and z_k.

- h is called the bandwidth parameter; it roughly gives the size of the histogram bins.
  - Tradeoff: h large $\Longrightarrow$ density estimate is smoother
  - h small $\Longrightarrow$ less functional form bias

- Different kernels
  - Uniform: $\frac{1}{2} \cdot I\left(|z| < 1\right)$
  - Triangular: $\left(1 - |z|\right) \cdot I\left(|z| < 1\right)$
  - Epanechnikov: $\frac{3}{4}\left(1 - z^2\right) \cdot I\left(|z| < 1\right)$
  - Quartic: $\frac{15}{16}\left(1 - z^2\right)^2 \cdot I\left(|z| < 1\right)$

  - Gaussian: $(2\pi)^{-\frac{1}{2}} e^{-\frac{z^2}{2}}$

- Most popular: Epanechnikov and Uniform

  - Kernels with higher order terms fit better (lower bias)

  - Kernels with lower order terms are smoother

- The kernel density estimator is biased as $N \longrightarrow \infty$ keeping h fixed but not if $h \longrightarrow 0$ as $N \longrightarrow \infty$

  - Since inference is done with a fixed h, assymptotic statistical inference is complicated by an asymptotic bias term.

  - Often densities don't have error bars on them

- Note that there are two types of convergence we can discuss since we discussing convergence to a density not just a paramter:

  - Convergence in distribution
  - Pointwise convergence
  - Most inference is pointwise

- One choice for an optimal bandwidth can come from minimizing mean integrated square error (between the density and the data).

- Two choices: kernel and bandwidth. Choice of kernel doesn't usually have a large impact on the estimation. Choice of bandwidth, however, is crucial.

4 Non-parametric Regression

- Can we use local regression methods to characterize the relationship between two variables as opposed

to the density of a variable and the variable itself? Yes!!! Its called non-parametric regression.

- Definition of the estimator:

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

where again $K(\cdot)$ is the kernel and h is the bandwidth.

- Basically you are averaging $Y(X_0)$ with $X's$ close to X_0 and in a weighted fashion.

- Special case of Local Weighted Average Estimator

$$\hat{m}(x_0) = \sum_{i=1}^{N} w_{i0,h} y_i$$

where $w_{io,h} = w(x_i, x_0, h)$

- K-Nearest Neighbor Estimator

$$\hat{m}(x_0) = \frac{1}{k}\left(y_{i-\left(\frac{k-1}{2}\right)} + \ldots + y_{i+\left(\frac{k+1}{2}\right)}\right)$$

- Generalization of kernel regression as local constant:

- Local linear regression estimator

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

then

$$\hat{m}(x) = \hat{a}_0 + \hat{b}_0(x - x_0)$$

- Regular kernel is local linear with b_0 constrained to be zero. We can generalize this approach to higher order polynomials.

- One particularlly popular kernel for non-parametric regression is: LOcally WEighted Scatterplot Smoothing (LOWESS) Estimator

$$K(z) = \frac{70}{81}\left(1 - |z|^3\right)^3 I(|z| < 1)$$

where

1. $h_{o,i}$ varies - it depends upon the distance of the point x_0 to the k^{th} nearest neighbor and

2. observations with large residuals, $(y_i - \hat{m}(x_i))$, are downweighted as in a quasi-GLS type estimator.

- Problems with non-parametric regression:

  - Requires alot of data, especially for multi-dimensional density estimation

5 Semi-parametric Regression

- Sometimes better to combine parametric and non-parametric - where along some dimensions you know the structure or where you don't care as much if you don't know the structure. Structure reduces the curse of dimensionality as with propensity score matching. This combination of parametric and non-parametric regression is called semi-parametric regression.

- Some semi-parametric estimators:

  - Partially Linear:

    $$E(Y|X,Z) = X\beta + \lambda(Z)$$

    parameters: β, non-parametric part: λ

  - Single Index:

    $$E(Y|X) = G(X\beta)$$

    parameters: β, non-parametric part: G

– Generalized Partial Linear:

$$E(Y|X,Z) = G(X\beta + \lambda(Z))$$

parameters: β, non-parametric parts: G, λ

6 Identification: IV + Non-parametrics

1. Almost nothing done here (a few recent papers such as by Blundell and Powell).

2. Hard because you need alot of data both for IV and for non-parametrics.

3. Even more difficult if you want your instruments to be non-parametric.

7 Overview

- Positive: Non-parametric methods can be very good data description techniques since they are very flexible.

- Negative: Require alot of data.

- Negative: Difficult to do inference.

- Negative: Difficult to get good indentification.

- Net: Often good complement (not substitute) for parametric analysis.