Recitation 9: Local polynomials their asymptotics

Seung-hun Lee

Columbia University
Introduction to Econometrics II Recitation

April 4th, 2022

Kernel density regression

Kernel density estimates are CAN....

- Consistency: Kernel estimator achieves pointwise consistency, or consistent at particular point $y = y_0$ if both the bias and the variance disappear at that point
  - Stronger version with uniform continuity: $\hat{f}_n(y)$ is consistent at all values of y if $\hat{f}_n(y)$ uniformly converges to $f(y)$
- Asymptotically normal: Estimator is recentered differently, but still normal
  - We know $E[\hat{f}_n(y)] = f(y) + \frac{1}{2}h^2(f''(y)) \int_{-\infty}^{\infty} u^2 K(u)du$ and $Var(\hat{f}_n(y)) = \frac{1}{nh} f(y) \int_{-\infty}^{\infty} K^2(u)du$
  - Thus the CLT leads to

$$\sqrt{nh}\left(\hat{f}_n(y) - f(y) - \frac{1}{2}h^2(f''(y)) \int_{-\infty}^{\infty} u^2 K(u)du\right) \sim N\left(0, f(y) \int_{-\infty}^{\infty} K^2(u)du\right)$$

...Just not in a typical manner

$$\sqrt{nh}\left(\hat{f}_n(y) - f(y) - \frac{1}{2}h^2(f''(y))\int_{-\infty}^{\infty} u^2K(u)du\right) \sim N\left(0, f(y)\int_{-\infty}^{\infty} K^2(u)du\right)$$

- Effective sample is nh: If h is optimally chosen, effective sample size is $O(n^{4/5})$
- $\frac{1}{2}h^2(f''(y))\int_{-\infty}^{\infty} u^2K(u)du$: Cannot ignore them since bias and standard errors move at same rate (parametrics: bias converged to 0 quicker)
- Further conditions (not as significant as the two above): f is twice continuously differentiable and $\int_{-\infty}^{\infty} u^2K(u)du$ is constant to pin the value of bias

Confidence intervals are centered with bias correction!

- We define a 95% pointwise confidence interval at particular y value as

$$CI = \left[\hat{f}_n(y) - b(y) - 1.96\sqrt{\frac{1}{nh}\hat{f}_n(y)\int_{-\infty}^{\infty} K^2(u)du},\ \hat{f}_n(y) - b(y) + 1.96\sqrt{\frac{1}{nh}\hat{f}_n(y)\int_{-\infty}^{\infty} K^2(u)du}\right]$$

  where $b(y) = \hat{f}_n(y) - \frac{1}{2}h^2(f''(y))\int_{-\infty}^{\infty} u^2K(u)du$
- The problem is also complicated further due to the existence of $f''(y)$
- Alternative: Confidence bands are computed for $f(y)$ over all possible values of y, which results in a wider confidence intervals than the pointwise ones

One dimensional kernel was already hard. How about a multiple?

- Now assume that y is not necessarily a scalar, but of dimension d.
- Then we can use a d-dimensional kernel K and estimate $f(y)$ with

$$\hat{f}_n(y) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right)$$

  where K can be a d-product of a uni-dimensional kernels.
- We are not necessarily confined to using a same bandwith for all d kernels
  - If variance of y_1 is larger than y_2 you may want higher h_1 relative to h_2
  - Sphericize them if necessary if you suspect correlation between kernels (like GLS)
- But this is only a minor problem: Optimal bandwidth is even more complicated

Optimal bandwidth: Solving again for AMISE

- Curse of dimensionality has to do with bandwidth costs
- Variance has a different leading term (leading term)

$$
\begin{aligned}
E[\hat{f}_n(y)^2] &= E\left[\frac{1}{n^2h^{2d}}\left(\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right)\right)^2\right] \\
&= E\left[\frac{1}{n^2h^{2d}}\left(\sum_{i=1}^{n} K^2\left(\frac{y-y_i}{h}\right) + 2\sum_{i<j} K\left(\frac{y-y_i}{h}\right)K\left(\frac{y-y_j}{h}\right)\right)\right] \\
&= \frac{1}{nh^{2d}}\int_{-\infty}^{\infty} K^2\left(\frac{y-t}{h}\right)f(t)dt + \frac{n(n-1)}{n^2h^{2d}}\left(\int_{-\infty}^{\infty} K\left(\frac{y-t}{h}\right)f(t)dt\right)^2
\end{aligned}
$$

- The leading term is $\frac{1}{nh^{2d}}\int_{-\infty}^{\infty} K^2\left(\frac{y-t}{h}\right)f(t)dt \simeq \frac{1}{nh^d}\int K^2(-u)f(y)du$ since u is actually d-dimensional.
- Thus, the variance is now $O\left(\frac{1}{nh^d}\right)$.

Optimal bandwidth now leads to even slower convergence

- Bias is at the same rate

$$
\begin{aligned}
E[\hat{f}_n(y)] &= E\left[\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right)\right] \\
&= \frac{1}{h^d}\int_{-\infty}^{\infty} K\left(\frac{y-t}{h}\right) f(t)dt \\
&= \int_{-\infty}^{\infty} K(-u)f(y+uh)du \ (\because u = \frac{t-y}{h} \text{ and } K(\cdot) \text{ is an } d\text{-dimensional multi-kernel})
\end{aligned}
$$

  apply same procedure as the univariate kernel to get bias$= \frac{1}{2}\int_{-\infty}^{\infty} K(u)h^2u^2f''(y)du$

- What this means is that $AMISE = Ah^4 + \frac{B}{nh^d}$ and the optimal h is solved as

$$
h = \left(\frac{Bd}{4An}\right)^{1/(4+d)}
$$

- h will be in $n^{-\frac{1}{4+d}}$, bias and standard errors are in $n^{-\frac{2}{4+d}}$ (even slower!)

What does it all mean? LARGE observations are needed

- Larger observations needed to achieve similar precisions as in lower-level kernel density estimation
- Rigorously, the sparseness problem becomes bigger with larger dimensions - fewer observations around y receive substantial weight (or an empty space phenomenon)
- Silverman documents that the required sample size to accurately estimate density of a standard normal at 0 rises drastically - 4 for univariate to 842,000 in 10-dimensional

Local polynomial regression

Conditional expectations from nonparametric approach is possible

- Given (y_i, x_i), we are attempting to capture $E[g(y,x)|x] = m(x)$ for some $g(y,x)$
- Frequently, we let $g(y,x) = y$
- Conditional expectation is written as

$$
\begin{aligned}
E[y|x] &= \int y f_{Y|X}(y|x) dy \\
&= \int y \frac{f_{Y,X}(y,x)}{f_X(x)} dy \ (\because \text{Relation between conditional, marginal, and joint pdf}) \\
&= \frac{\int y f_{Y,X}(y,x) dy}{\int f_{Y,X}(y,x) dy} \ (\because f_X(x) = \int f_{Y,X}(y,x) dy)
\end{aligned}
$$

- Key? Replace $f_{Y,X}$ with its kernel estimator (For easy life, let both be a scalar)

Numerator to a kernel density estimate

- Numerator

$$
\begin{aligned}
\int y\hat{f}(y,x)dy &= \int y \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) K\left(\frac{y-y_i}{h}\right) dy \\
&= \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \int yK\left(\frac{y-y_i}{h}\right) dy \\
&= \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \int (y_i + sh) K(s)\, hds \ \left(\because s = \frac{y-y_i}{h}\right) \\
&= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i \ \left(\because \int K(s)ds = 1, \int sK(s)ds = 0\right)
\end{aligned}
$$

Denominator and the complete set

- Denominator

$$\begin{aligned}\int \hat{f}_{Y,X}(y,x)dy &= \int \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)K\left(\frac{y-y_i}{h}\right)dy \\ &= \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\int K\left(\frac{y-y_i}{h}\right)dy \\ &= \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\int K\left(s\right)hds = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\end{aligned}$$

- Thus, the estimator for the conditional expectation becomes

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

So what is $\hat{m}(x)$ exactly?

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

- Effectively we are putting weight $\frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$ on each observation y_i.
- This is called a **local constant estimation** or **Nadaraya-Watson estimator**: When we assume $y_i = a + e_i$ for some constant a, weigh each observation by its kernel density $K\left(\frac{x-x_i}{h}\right)$ and solve the following minimization problem

$$\hat{f}(x) = \arg\min_{a} \frac{1}{nh} \sum_{i=1}^{n} (y_i - a)^2 K\left(\frac{x - x_i}{h}\right)$$

The first order condition on a yields the following results

$$\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right) = a \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \implies a = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

Asymptotics of the NW estimator: Building blocks

- Note that $y_i = m(x_i) + e_i = m(x) + (m(x_i) - m(x)) + e_i$
- Assume $m(x_i) = E[y|x_i]$ is twice continuously differentiable, $f(x)$ is a pdf that is once continuously differentiable, and $E[e_i|x_i] = 0$, $E[e_i^2|x_i = x] = \sigma^2(x) < \infty$
- We can express the numerator of $\hat{m}(x)$ as

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(m(x) + (m(x_i) - m(x)) + e_i) \\ &= \hat{f}(x)m(x) + \hat{m}_1(x) + \hat{m}_2(x)\end{aligned}$$

- $\hat{m}(x) = \frac{\hat{f}(x)m(x)+\hat{m}_1(x)+\hat{m}_2(x)}{\hat{f}(x)} = m(x) + \frac{\hat{m}_1(x)}{\hat{f}(x)} + \frac{\hat{m}_2(x)}{\hat{f}(x)}$

Asymptotics bias of the NW estimator

- We need to know $E[\hat{m}_1(x)]$ and $E[\hat{m}_2(x)]$
- Since $E[e_i|x_i] = 0$, then $E\left[e_i K\left(\frac{x-x_i}{h}\right)\right] = 0$ and $E[\hat{m}_2(x)] = 0$
- For $E[\hat{m}_1(x)]$, we work with

$$
\begin{aligned}
E[\hat{m}_1(x)] = E\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(m(x_i) - m(x))\right] &= \frac{1}{h}\int K\left(\frac{s-x}{h}\right)(m(s) - m(x))f(s)ds \\
&= \int K(u)\left(m(x+hu) - m(x)\right)f(x+hu)du \ \left(\because u = \frac{s-x}{h}\right) \\
&= \int K(u)\left(uhm'(x) + \frac{1}{2}u^2h^2m''(x)\right)(f(x) + uhf'(x))du \ (\because \text{two Taylor expansions!}) \\
&= \int h^2 m'(x)f'(x)u^2K(u)\,du + \int \frac{h^2}{2}m''(x)f(x)u^2K(u)du + \dots.
\end{aligned}
$$

- ... includes higher order terms than h^2

Asymptotics bias has complexities with m and f functions

- If we have $\hat{f}(x)$ that is a consistent density estimator for $f(x)$, we can now write

$$E[\hat{m}(x)] = m(x) + \int h^2 \frac{m'(x)f'(x)}{f(x)} u^2 K(u)\, du + \int \frac{h^2}{2} m''(x) u^2 K(u) du$$

- Therefore, the bias is still of order h^2, but the exact expression becomes

$$E[\hat{m}(x)] - m(x) = h^2 \left(\frac{m'(x)f'(x)}{f(x)} + \frac{m''(x)}{2} \right) \int u^2 K(u) du$$

Asymptotics variance of the NW estimator

- We have

$$
\begin{aligned}
var(\hat{m}_2(x)) &= E\left[\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) e_i\right)^2\right] = \frac{1}{nh^2} E\left[\left(K\left(\frac{x_i - x}{h}\right) e_i\right)^2\right] \ (\because IID) \\
&= \frac{1}{nh^2} E\left[\left(K\left(\frac{x_i - x}{h}\right) \sigma(x_i)\right)^2\right] \ (\because \text{condition on } x) = \frac{1}{nh^2}\int K\left(\frac{s - x}{h}\right)^2 \sigma^2(s) f(s) ds \\
&= \frac{1}{nh}\int K(u)^2 \sigma^2(x + hu) f(x + hu) du \simeq \frac{1}{nh}\int K(u)^2 \sigma^2(x) f(x) du
\end{aligned}
$$

- $var(\hat{m}_1(x))$, it is actually $O\left(\frac{h^2}{nh}\right)$ a smaller order than $O\left(\frac{1}{nh}\right)$
- Thus, variance is

$$
\frac{1}{nh}\int K(u)^2 \sigma^2(x) f(x) du / f^2(x) = \frac{1}{nh}\frac{\sigma^2(x)}{f(x)}\int K(u)^2 du
$$

Asymptotic distribution of NW

- As a result, the asymptotic distribution of the local constant estimator is

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2\left(\frac{m'(x)f'(x)}{f(x)} + \frac{m''(x)}{2}\right)\int u^2 K(u)du\right) \sim N\left(0, \frac{\sigma^2(x)}{f(x)}\int K(u)^2\, du\right)$$

- NW estimators become inaccurate as $f(x)$ is small or at the boundary value
- Moreover, the estimate also becomes volatile with $\sigma^2(x)$ - the variance of the Y conditional on X
- Bandwidth choice is tougher

Something better: Local linear estimation

- In mathematical expression, we solve

$$\min_{a,b} \frac{1}{nh} \sum_{i=1}^{n} (Y_i - a - b(x - x_i))^2 K\left(\frac{x - x_i}{h}\right)$$

  and obtain that $\hat{a} = \hat{g}$ and $\hat{b}$ is an estimate of $\frac{\partial g(x)}{\partial x}$
- Lesser bias if true DGP is linear and better performance at the boundary
- Easier expression for asymptotics: We have explicitly modeled the linear term by controlling for $x - x_i$, and coefficient on this will estimate $m'(x)$. $(y_i = m(x_i) + e_i \simeq m(x) + m'(x)(x_i - x) + e_i)$
- The end result for the asymptotic distribution of local linear estimator is

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2 \frac{m''(x)}{2} \int u^2 K(u) du\right) \sim N\left(0, \frac{\sigma^2(x)}{f(x)} \int K(u)^2 du\right)$$

Seminonparametric regressions

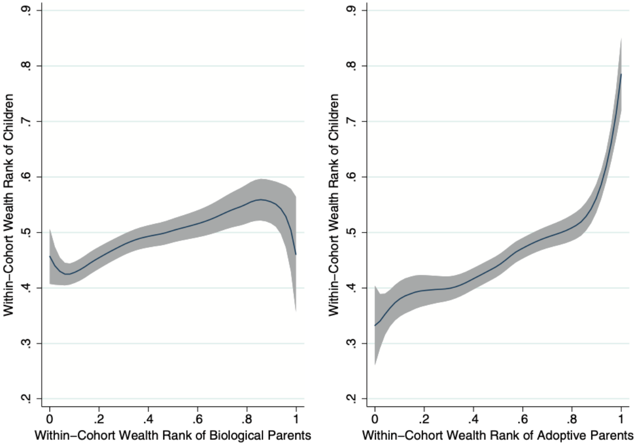
- Idea: If we are sure that $f(y)$ can be characterized by $f_{m,\sigma}$, we can choose a family of positive functions $P_\theta^1, P_\theta^2, \ldots (\because$ Weierstrass approximation thm) and maximize over

$$\sum_{i=1}^{n} \log f_{m,\sigma}(y_i) P_\theta^M(y_i)$$

  - Mixture of normals is a special case
- Series estimation: Run a linear regression $y_i = \sum_{k=1}^{M} P_k(x_i)\theta_k + \epsilon_i$, where P_k is an orthonormal basis and the $\sum_{k=1}^{M} P_k(x_i)\theta_k$ part is a series approximation to $g(x)$
- Sieve estimation: Similar to series estimation, where the choice of the basis is data-dependent (Splines or polygonals)

Black et al 2019 Restud: Linearity between parental and child wealth

- Relationship of intergenerational wealth correlation is linear, with stronger role by the adoptive parents than biological



Brückner, Ciccone ECMA 2011: Rainfall, income, and democracy

- Rainfall shock is positively correlated with income and is followed by an improvement in institutions
  - Negative rainfall shocks (drought) $\rightarrow$ drop in income $\rightarrow$ protests and institutional change