Nonparametric and Semiparametric Estimation

for Non-Econometricians

Arthur Lewbel

Boston College

a mini-course, revised 2012

Goal: Understanding the basic ideas behind the econometrics revolution in nonparametric and semparametric methods.

This talk will focus on cross section methods. Think of data here as i.i.d. (independently, identically distributed observations).

Warning: This lecture will NOT be mathematically rigorous. Some of the definitions, formulas and derivations are incomplete (an expert might just call them wrong). My goal here will be to make the ideas and methods as clear as possible, favoring clarity of concepts over detailed technical accuracy.

Will not go through empirical applications, but will discuss relevant considerations for applying these estimators.

Examples:

Linear: $Y_i = X_i'\beta + e_i$
Parametric: $Y_i = G(X_i, \theta) + e_i$
finite dimensional parameter θ, β

Nonparametric: $Y_i = m(X_i) + e_i$
infinite dimensional parameter $m()$

Semiparametric: $\alpha = E\left(\frac{\partial m(X_i)}{\partial X}\right)$
Semiparametric: $Y_i = m(X_i) + Z_i'\gamma + e_i$
finite parameter of interest, infinite dimensional nuisance parameter

Why nonparametrics?

Statistical lit: don't need theories.

Most economic theories do not imply specific functional forms.

Can provide guidance for parametric models.

Why not nonparametrics?

Curse of Dimensionality

Why semiparametrics?

Sometimes overcomes the curse

Focus on features of interest.

More popular in econometrics than in statistics (we have finite dimensional features of interest).

X an r.v. (random variable, for now a scalar).
x a value X could take on.
X_i for $i = 1, ..., n$ are iid rv's.
x_i for $i = 1, ..., n$ observations, the sample.

X is drawn from a distribution function $F(X)$

each X_i is a r.v. with distribution F

$F(x) = \Pr(X \leq x)$, a function.

$F(x)$ = true distribution function of each X_i, evaluated at x.
If X continuous, $F(x)$ is a (usually) S shaped curve from 0 to 1.

$\widehat{F}(x)$ = empirical distribution function, evaluated at x.

$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$
= fraction of data less than x.

Graph $\widehat{F}(x)$ against x: a step function with n steps.

$\widehat{F}(x)$ is a nonparametric estimator of $F(x)$.

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$$

Is $\widehat{F}(x)$ an unbiased estimator of $F(x)$?

Recall what unbiasedness means:
Suppose we had an estimator $\widehat{\theta}$ of a vector θ.
Imagine we had very many data sets, instead of just one.
Calculate $\widehat{\theta}$ using each data set separately.
$\widehat{\theta}$ is unbiased if the average of the estimates $\widehat{\theta}$, averaged across the data sets, equals the true θ. This needs to hold for each element θ_j of θ.

The nonparametric estimator $\widehat{F}(x)$ is unbiased if the average of $\widehat{F}(x)$, averaged across an infinite number of data sets, equals the true $F(x)$, for every possible x.

Here x is like an index, it just refers to one 'element' of the function $F()$ (the element we happen to be estimating), just like j in θ_j indexes one element of a vector θ in a parametric model.

For every real number x, $\widehat{F}(x)$ is unbiased:

$$E\left[\widehat{F}(x)\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)\right]$$

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} E\left[1(X_i \leq x)\right] \\
&= E\left[1(X_i \leq x)\right] = E\left[1(X \leq x)\right] \\
&= \int_{X=-\infty}^{\infty} 1(X \leq x) f(X)\, dX = \int_{X=-\infty}^{x} f(X)\, dX \\
&= F(X)\,|_{X=-\infty}^{x} = \left[F(x) - F(-\infty)\right] \\
&= \left[F(x) - 0\right] = F(x)
\end{aligned}$$

Notice the X_i's are the random variables we are averaging over by taking the expectation. We are *not* averaging across x values. Again, x is like an index, it just refers to an 'element' of the function $F()$, just like j in θ_j indexes one element of a vector θ.

$$E\left[\widehat{F}(x)\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)\right] = F(x)$$

In the same way, can calculate

$$\begin{aligned} var\left[\widehat{F}(x)\right] &= E\left[\left(\widehat{F}(x) - F(x)\right)^2\right] \\ &= \left[\frac{1}{n}\sum_{i=1}^{n} E\left[1(X_i \leq x) - F(x)\right]\right]^2 = \frac{1}{n}F(x)\left[1 - F(x)\right] \end{aligned}$$

and $\widehat{F}(x)$ is an average, so by the central limit theorem (CLT),

$$\sqrt{n}\left(\widehat{F}(x) - F(x)\right) \rightarrow^d N\left[0, F(x)\left[1 - F(x)\right]\right]$$

$\widehat{F}(x)$ is root-n-CAN: consistent, asymptotically normal, at rate root-n. True at every point (every real number) x.

Now suppose we want to estimate $f(x)$, the pdf (probability density function) of X, at each point x.

If parameterize $f(x, \theta)$, do MLE.

Or can approximate $f(x)$ by a histogram:

For a small number (binwidth) h, look at

$$\frac{1}{n} \sum_{i=1}^{n} 1(x - h \leq x_i \leq x + h)$$

This is the fraction of observations in our sample near x (within distance h of x).

We will now construct a formal estimator, similar to the histogram.

$$\begin{aligned}\widehat{F}(x+h)-\widehat{F}(x-h) &= \frac{1}{n}\sum_{i=1}^{n}1(x-h\leq X_i\leq x+h)\\ &= \frac{1}{n}\sum_{i=1}^{n}1\left(|\frac{x-X_i}{h}|\ \leq 1\right)\end{aligned}$$

$$\begin{aligned}f(x) &= \frac{dF(x)}{dx}=\lim_{h\to 0}\frac{F(x+h)-F(x-h)}{2h}\\ &\approx \frac{F(x+h)-F(x-h)}{2h} \quad \text{for small } h\end{aligned}$$

Suggests the estimator, choose a small h and let:

$$\widehat{f}(x)=\frac{\widehat{F}(x+h)-\widehat{F}(x-h)}{2h}=\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}1\left(|\frac{x-X_i}{h}|\ \leq 1\right)$$

Estimator depends on h, so call it $\widehat{f}_h(x)$.

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}1\left(\left|\frac{x-X_i}{h}\right| \leq 1\right)$$

Can write this as

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \qquad \text{where} \qquad K(u) = \frac{1}{2}1(|u| \leq 1)$$

so $K(u)$ equals a uniform density on $[-1,1]$.

This $\widehat{f}_h$ is a little ugly: Discontinuous in x. Also, gives equal weight to all observations within h of x, zero weight to those further.

Consider other $K(u)$ functions to get a nicer estimator. e.g., observations X_i closest to x are most informative about $f()$ at x, so could give those the most weight.

Nadayara-Watson Kernel density estimator:

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where you choose a kernel or window function $K(u)$ that is continuous almost everywhere with $\int_{-\infty}^{\infty} K(u)\,du = 1$ like a density. K is usually symmetric and has a mode at zero. h is called the binwidth or bandwidth.

Popular kernels are gaussian (normal) and Epanechikov (quadratic)

$$\begin{aligned} K(u) &= \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \\ K(u) &= \frac{3}{4}\left(1 - u^2\right) 1\left(|u| \leq 1\right) \end{aligned}$$

Is $\widehat{f}_h(x)$ unbiased like $\widehat{F}(x)$ was? Assume $K(u) = K(-u)$, a symmetric kernel. Also means a 'second order' kernel: $\int_{-\infty}^{\infty} u^p K(u)\,du = 0$ for all positive $p < 2$.

Again average over X_i, while x just indexes value of f are estimating.

$$\begin{aligned} E\left[\widehat{f}_h(x)\right] &= E\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right] \\ &= E\left[\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right] = \int_{X_i=-\infty}^{\infty} \frac{1}{h}K\left(\frac{x-X_i}{h}\right) f(X_i)\, dX_i \end{aligned}$$

Change of variables: $u = -\frac{x-X_i}{h}$, $X_i = x + uh$, Jacobian $= h$.Then

$$\begin{aligned} E\left[\widehat{f}_h(x)\right] &= \int_{u=-\infty}^{\infty} \frac{1}{h}K(u) f(x+uh)\, h du \\ &\approx \int_{u=-\infty}^{\infty} \frac{1}{h}K(u)\left[f(x) + uh\frac{df(x)}{dx} + \frac{u^2h^2}{2}\frac{d^2f(x)}{dx^2}\right] h du \\ &= f(x)\int_{u=-\infty}^{\infty} K(u)\, du + h\frac{df(x)}{dx}\int_{u=-\infty}^{\infty} uK(u)\, du \end{aligned}$$

$$+h^2\frac{d^2f(x)}{dx^2}\int_{u=-\infty}^{\infty} u^2K(u)\, du = f(x) + h^2 b(x)$$

$E\left[\widehat{f}_h(x)\right]$ is biased, bias is approximately $h^2 b(x)$

$$bias\left[\widehat{f}_h(x)\right] \approx h^2 b(x), \qquad b(x) = \frac{d^2 f(x)}{dx^2}\int_{u=-\infty}^{\infty} u^2 K(u)\, du$$

Similarly,

$$var\left[\widehat{f}_h(x)\right] \approx \frac{1}{nh} v(x), \qquad v(x) = f(x)\int_{u=-\infty}^{\infty} [K(u)]^2\, du$$

Mean squared error (MSE) trade off:
Bias small if h small, variance small if nh is big.
Can choose h to minimize MSE:

$$\begin{aligned} \text{MSE} &\approx h^4 b^2(x) + \frac{1}{nh} v(x) \\ \text{best } h &\approx \frac{v(x)}{4b^2(x)} n^{-1/5} \end{aligned}$$

Best $h \to 0$ as $n \to \infty$. But how choose h? Vary by x?
Two step estimation or optimal for parametric f or cross validation.

$\widehat{f}_h(x)$ is an average, so apply central limit theorem (CLT)

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

$$n^{1/2}\left[\widehat{f}_h(x) - E\left(\widehat{f}_h(x)\right)\right] \rightarrow^{d} N\left[0, var\left(\frac{1}{h}K\left(\frac{x-X}{h}\right)\right)\right]$$

$$n^{1/2}\left[\widehat{f}_h(x) - f(x) - h^2 b(x)\right] \approx N\left[0, \frac{1}{h}v(x)\right]$$

Problem: if $h \longrightarrow 0$ then $v(x)/h$ blows up. Fix: multiply by $h^{1/2}$

$$(nh)^{1/2}\left[\widehat{f}_h(x) - f(x) - h^2 b(x)\right] \approx N[0, v(x)]$$

Min MSE had h proportional to $n^{-1/5}$, so $(nh)^{1/2}$ is proportional to $n^{2/5}$.

Under some smoothness assumptions, $n^{2/5}$ is the fastest "rate of convergence" for nonparametric density estimation.

Note: for more rigor, need to keep track of and bound remainder terms, and use a CLT that allows h to depend on n.

Extensions: type of convergence:

Pointwise convergence (consistency): $plim\left|\widehat{f}_h(x) - f(x)\right| = 0$
Uniform convergence (consistency): $plim \sup_x \left|\widehat{f}_h(x) - f(x)\right| = 0$

Similarly have derived pointwise limiting distribution, means pointwise confidence intervals. can also construct uniform confidence intervals.

Extensions: Density derivatives:

$$\frac{d\widehat{f}_h(x)}{dx} = \frac{1}{nh}\sum_{i=1}^{n}\frac{dK\left(\frac{x-X_i}{h}\right)}{dx} = \frac{1}{nh^2}\sum_{i=1}^{n}K'\left(\frac{x-X_i}{h}\right)$$

Consistency requires $nh^2 \to \infty$ instead of $nh \to \infty$, must have $h \to 0$ slower, so get a slower optimal rate of convergence than for $\widehat{f}_h(x)$.

Extensions: Bias reduction

If f is smoother (more derivatives). recall

$$\begin{aligned} E\left[\widehat{f}_h(x)\right] &\approx \int_{u=-\infty}^{\infty} \frac{1}{h} K(u) \left[f(x) + uh\frac{df(x)}{dx} + \frac{u^2h^2}{2}\frac{d^2f(x)}{dx} \right] h du \\ &\approx f(x) + h^2 b(x) \end{aligned}$$

If $\int_{u=-\infty}^{\infty} u^p K(u)\, du = 0$ for $p = 1, 2, 3$ (a fourth order kernel), can Taylor expand to four terms, get $bias = E\left[\widehat{f}_h(x)\right] - f(x) \approx h^4 B(x)$. Variance is still $n^{-1}h^{-1}v(x)$, so best MSE is now rate $n^{-4/9}$. But: higher order kernels have negative $K(u)$ regions, are poorly behaved numerically unless n is huge.

Extensions: Multivariate Density Estimation

Joint density $f(y, x)$ of Y and X can be estimated as

$$\widehat{f}_h(y, x) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right) K\left(\frac{x - X_i}{h}\right)$$

For a J dimensional density, bias is still proportional to h^2, but variance is proportional to $n^{-1}h^{-J}$. So optimal MSE has h proportional to $n^{-1/(J+4)}$, and optimal rate of convergence is $n^{2/(J+4)}$.

The "curse of dimensionality"

For parametric models, the more parameters we estimate, the bigger the variance, but the rate stays root-n. Multiply n by 4, the standard errors are halved.

For nonparametrics, the higher the dimension of the function, the slower is the rate of convergence. To estimate a J dimensional function (without higher order kernels), must multiply n by $2^{((J+4)/2)}$ to halve the standard errors. E.g., with $J = 2$, you must multiply n by 8 to halve the standard errors.

Nonparametric Regression:
iid (Y_i, X_i). draws from joint density $f(Y, X)$
Let $m(X) = E(Y \mid X)$. Then $Y = m(X) + e$, $E(e \mid X) = 0$.

Goal, estimate $m(x)$ at any point x. Can, e.g., evaluate $\widehat{m}(x)$ at a fine grid of points x, and graph it. Local average estimator:

$$\begin{aligned} m(x) &= E(Y \mid X = x) \approx E(Y \mid x - h \leq X \leq x + h) \\ &\approx \frac{\sum_{i=1}^{n} Y_i 1(x - h \leq X_i \leq x + h)}{\sum_{i=1}^{n} 1(x - h \leq X_i \leq x + h)} = \frac{\sum_{i=1}^{n} Y_i 1\left(\left|\frac{x - X_i}{h}\right| \leq 1\right)}{\sum_{i=1}^{n} 1\left(\left|\frac{x - X_i}{h}\right| \leq 1\right)} \end{aligned}$$

If X is discrete, can let $h = 0$, otherwise let h be small nonzero.
Like a histogram above uses a uniform kernel. A kernel regression is:

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

which is a kernel function weighted average of Y_i. The closer X_i is to x, the larger the weight $K[(x - X_i)/h]$.

Another way to get the kernel regression:

$$m(x) = E(Y \mid X = x) = \int_{y=-\infty}^{\infty} y f_{y|x}(y \mid x)\, dy$$

$$= \int_{y=-\infty}^{\infty} y \frac{f_{y,x}(y,x)}{f_x(x)} dy = \frac{\int_{y=-\infty}^{\infty} y f_{y,x}(y,x)\, dy}{f_x(x)}$$

$$\widehat{m}_h(x) = \frac{\int_{y=-\infty}^{\infty} y \widehat{f}_{y,x}(y,x)\, dy}{\widehat{f}_x(x)} = \frac{\int_{y=-\infty}^{\infty} y \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{y-Y_i}{h}\right) K\left(\frac{x-X_i}{h}\right)}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

$$= \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \left[\frac{1}{h} \int_{y=-\infty}^{\infty} y K\left(\frac{y-Y_i}{h}\right) dy\right]}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

$$= \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \left[\frac{1}{h} \int_{u=-\infty}^{\infty} (hu + Y_i) K(u) h du\right]}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

and $$\frac{1}{h} \int_{u=-\infty}^{\infty} (hu + Y_i) K(u) h du = Y_i$$

Properties of kernel regressions:

$$m(x) = E(Y \mid X = x), \qquad \widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

h determines smoothness of $\widehat{m}(x)$, roughly, size of neighborhood around x over which data are averaged.

If $h = \infty$ then $\widehat{m}(x) = \overline{Y}$. Complete averaging.

If $h = 0$, then $\widehat{m}(x) = \frac{0}{0}$ if $x \neq X_i$ for some i, else $\widehat{m}(x) = Y_i$. No averaging.

With big h, $\widehat{m}(x)$ is close to a flat line at $\overline{Y}$. With a tiny h, $\widehat{m}(x)$ erratically jumps around, between almost going through each point (Y_i, X_i)

Properties of kernel regressions:

Like kernel densities, $bias = E\left[\widehat{m}_h(x)\right] - m(x) \approx h^2 B(x)$ for some $B(x)$,
$var\left[\widehat{m}_h(x)\right] \approx n^{-1}h^{-1}V(x)$,
$V(x) = \left(\int_{u=-\infty}^{\infty}\left[K(u)\right]^2 du\right) E\left(e^2 \mid X = x\right) / f_x(x)$

$$(nh)^{1/2}\left[\widehat{m}_h(x) - m(x) - h^2 B(x)\right] \approx N\left[0, V(x)\right]$$

Again optimal is h proportional to $n^{-1/5}$, so rate $(nh)^{1/2}$ is $n^{2/5}$.

As before, these give pointwise confidence intervals.

As before, possible to calculate a uniform confidence interval, a sleeve around $\widehat{m}_h(x)$.

As before, with more smoothness can use higher order kernels to converge faster in theory, usually numerically bad in practice unless n is huge.

Multivariate kernel regression

iid (Y_i, X_i, Z_i). draws from joint density $f(Y, X, Z)$
Let $m(X, Z) = E(Y \mid X, Z)$.

$$\widehat{m}_h(x, z) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right) K\left(\frac{z-Z_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) K\left(\frac{z-Z_i}{h}\right)}$$

For Y conditional on a J vector (above is $J = 2$) as with joint density estimation get bias still proportional to h^2, variance is proportional to $n^{-1}h^{-J}$, optimal MSE has h proportional to $n^{-1/(J+4)}$, and optimal rate of convergence is $n^{2/(J+4)}$.

The "curse of dimensionality" is back.

Bandwidth choice for kernel regression:
method of cross-validation.
If we chose h to minimize sum of squared errors $\sum_{i=1}^{n}\left(Y_i-\widehat{m}_h\left(X_i\right)\right)^2$, would get $h=0$.

Instead, let $\widehat{m}_{hi}\left(x\right)=$ kernel regression of Y on X, leaving out observation i, using bandwidth h.

For example:

$$\widehat{m}_{h1}\left(x\right)=\frac{\sum_{i=2}^{n}Y_iK\left(\frac{x-X_i}{h}\right)}{\sum_{i=2}^{n}K\left(\frac{x-X_i}{h}\right)}$$

Similarly have $\widehat{m}_{h2}\left(x\right)$, $\widehat{m}_{h2}\left(x\right)$, etc.,.
For each observation i, leave out data point i and see how well $\widehat{m}$ fits that data point. Error is $Y_i-\widehat{m}_{hi}\left(X_i\right)$. Cross validation is choose h to minimize sum of squared of these errors. h minimizes:

$$\sum_{i=1}^{n}\left(Y_i-\widehat{m}_{hi}\left(X_i\right)\right)^2$$

Extension: Varying bandwidth
Can let h vary by x. Where data are sparse (e.g., in the tails of the density of X), choose large h. where data are dense, choose small h.

Example is "k nearest neighbor" estimation:
at each x, choose h so that only the k observations having X_i closest to x are used.

Extension: Regression derivative:

$\frac{d\widehat{m}_{hi}(x)}{dx}$ is an estimator of $\frac{dm(x)}{dx} = \frac{dE(Y|X=x)}{dx}$.
If Y and X are log data, this is the elasticity of Y with respect to X at point x.

Extension: Local Linear estimation

Kernel regression same as regressing Y on a constant, using data near x:
For any x, the kernel regression estimator $\widehat{m}(x)$ equals the m that minimizes

$$\sum_{i=1}^{n} (Y_i - m)^2 K\left(\frac{x - X_i}{h}\right)$$

First order condition: $\sum_{i=1}^{n} -2 (Y_i - m) K\left(\frac{x-X_i}{h}\right) = 0$. Solve for m.

Instead regress Y on a constant and on X, using data near x:
let $\widehat{M}(x)$ and $\widehat{D}(x)$ equal the M and D that minimizes

$$\sum_{i=1}^{n} (Y_i - M - (x - X_i) D)^2 K\left(\frac{x - X_i}{h}\right)$$

can show $\widehat{M}(x) \longrightarrow E(Y \mid X = x)$, and $\widehat{D}(x) \longrightarrow \frac{dE(Y|X=x)}{dx} = \frac{dM(x)}{dx}$

Extension: local polynomial regression

Instead of fitting a different line near each point x (local linear) can fit a local quadratic, or cubic, or polynomial.

Why consider local linear or local polynomial?:
Speeds convergence rates, similar to higher order kernels.
Automatically provides derivative estimates
If true $m(x)$ is close to polynomial, then improves the fit.
Can behave better near boundaries of the data.

Downside: more complicated
can be more numerically unstable, more sensitive to outliers.

Other nonparametric regression estimators:

Series estimators.
If $m(x)$ is smooth enough, then there exists a_0, a_1, a_2, ... such that $m(x) = \sum_{j=0}^{\infty} a_j x^j$.
So define $\widehat{m}_J(x) = \sum_{j=0}^{J} \widehat{a}_j x^j$ .where $\widehat{a}_j$'s are estimated by ordinary least squares regression of Y_i on 1, X_i, X_i^2,...,X_i^J,

Let $J \to \infty$ and $n/J \to \infty$, so add terms slowly as $n \to \infty$. Then the average number of observations per coefficient $\to \infty$, and get $\widehat{m}_J(x)$ consistently estimates $m(x)$.

J is like a bandwidth, choice of J trades off bias and variance.

Careful interpreting: what each $\widehat{a}_j$ means depends on J. Best to focus on $\widehat{m}_J(x)$ or $\frac{d\widehat{m}_J(x)}{dx}$, not each $\widehat{a}_j$.

Why series estimators?
Matches what practitioners often do - fit a line if have a small data set, try adding quadratic terms with larger n.
Easy to do and understand: It's ordinary regression.

Why not do series estimation?
Asymptotic distribution theory not as well worked out as kernels.
Unlike kernels, that fit near each x locally, series can jump around: $\widehat{m}_h(X_i)$ gets closer to Y_i as $h \longrightarrow 0$. But $\widehat{m}_J(X_i)$ can jump arbitrarily far away from Y_i as J increases. (think how a fitted regression curve can move when you change specification from linear to quadratic, or quadratic to cubic).
Small changes in h mean small change in fit. Small changes in J can dramatically change the fit. So series can be more sensitive to choice of J.

Extension: multivariate series regressions:
$m(x,z) = E(Y \mid X = x, Z = z)$
$\widehat{m}_J(x) = \sum_{j=0}^{J} \sum_{k=0}^{J} \widehat{a}_{jk} x^j z^k$

Extension: other series regressions:
$\widehat{m}_J(x) = \sum_{j=0}^{J} \widehat{a}_j \phi_j(x)$
$\phi_1(x)$, $\phi_2(x)$, $\phi_3(x)$,... are fourier or other series (basis functions that span the space).

Extension: sieve estimators:
Let A_1, A_2,., A_J,.. be a sequence of vectors that get longer as J increases.
Let $\Phi_J(A_J, x)$ for $J = 1, 2, ...$ be a sequence of functions that get more complicated as J grows.
Assume there exists a sequence of values for A_J such that
$m(x) = \lim_{j\to\infty} \Phi_j(A_j, x)$.
$\widehat{m}_J(x) = \Phi_j\left(\widehat{A}_j, x\right)$ where $\widehat{A}_j$ obtained by regressing Y_i on $\Phi_j(A_j, x)$.

Example sieve: Series $\widehat{m}_J(x) = \Phi_j\left(\widehat{A}_j, x\right) = \sum_{j=0}^{J} \widehat{a}_j x^j$

Example sieve: Neural Networks:

$$\widehat{m}_J(x) = \sum_{j=0}^{J} \widehat{a}_j g\left(x'\widehat{B}_j\right) \quad \text{or} \quad \widehat{m}_J(x) = \sum_{k=0}^{K} \widehat{c}_k g\left(\sum_{j=0}^{J} \widehat{a}_{jk} g\left(x'\widehat{B}_j\right)\right)$$

above are 'single layer' and 'two layer' neural nets.
g is a 'squasher' function (like arctan, or a distribution function) that maps real line to 0-1 interval.
Each $g(\bullet)$ represents the action of a neuron.
Fits coefficients by nonlinear least squares. Regresses Y_i on $\sum_{j=0}^{J} a_j g\left(X_i' B_j\right)$

'Learning' is just updating the nonlinear least squares coefficients as n increases.

Semiparametric Estimation. Roughly, finite parameter (vector) of interest, and infinite dimensional other parameters (typically unknown functions).

Examples: Binary choice $Y = 1(X'\beta + e \geq 0)$
β, f_e unknown.

Linear regression $Y = X'\beta + e$ looks semiparametric with β, f_e unknown, but isn't, since can rewrite as $E(Y \mid X) = X'\beta$. Whether a problem is semiparametric depends on 'how much' the unknown infinite dimensional parameters affect the finite one.

Partly linear model: $Y = m(X) + Z'\gamma + e$
so $E(Y \mid X, Z) = m(X) + Z'\gamma$
γ, m unknown

Average derivative estimation: $Y = m(X) + e$, $E(Y \mid X) = m(X)$,
$\alpha = E\left(\frac{\partial m(X)}{\partial X}\right)$
α, m unknown

Many Semiparametric estimators can be written as averages of functions of nonparametric estimators (e.g., average derivative estimation).

Let $\widehat{m}_h(x)$ be a nonparametric (kernel or local polynomial) estimator of a function $m(x)$, with bandwidth h.

Assume (which we showed was true for nonparametric density and nonparametric regression estimation) that
$E[\widehat{m}_h(X) \mid X = x] \approx m(x) + h^2 B(x)$,
$var[\widehat{m}_h(X) \mid X = x] \approx n^{-1} h^{-1} V(x)$

Consider estimation of $\mu = E[g(X, m(X))]$, using
$\widehat{\mu} = n^{-1} \sum_{i=0}^{n} g[X_i, \widehat{m}_h(X_i)]$

Start with a simpler case: $\mu = E[m(X)]$, using $\widehat{\mu} = n^{-1} \sum_{i=0}^{n} \widehat{m}_h(X_i)$.

Averages of nonparametric estimators:

$\mu = E\left[m\left(X\right)\right]$, $\widehat{\mu} = n^{-1}\sum_{i=0}^{n}\widehat{m}_h\left(X_i\right)$ where

$E\left[\widehat{m}_h\left(X\right) \mid X = x\right] \approx m\left(x\right) + h^2 B\left(x\right)$,

$$\begin{aligned} E\left(\widehat{\mu}\right) &= E\left[n^{-1}\sum_{i=0}^{n}\widehat{m}_h\left(X_i\right)\right] = E\left[\widehat{m}_h\left(X\right)\right] = E\left[E\left(\widehat{m}_h\left(X\right) \mid X\right)\right] \\ &\approx \mu + E\left[h^2 B\left(X\right)\right] \approx \mu + h^2 B \quad \text{where } B = E\left[B\left(X\right)\right] \end{aligned}$$

So bias of $\widehat{\mu}$, as in $\widehat{m}_h\left(X\right)$, is of order h^2. What about variance?

For variance we have:

$$\begin{aligned} var\left(\widehat{\mu}\right) &= var\left[n^{-1}\sum_{i=0}^{n}\widehat{m}_h\left(X_i\right)\right] \approx n^{-1}var\left[\widehat{m}_h\left(X_i\right)\right] \\ &\approx n^{-1}var\left[m\left(X\right)+\left(\widehat{m}_h\left(X\right)-m\left(X\right)\right)\right] \\ &\approx n^{-1}\left(\sigma^2+c^2+r_h\right) \end{aligned}$$

where $\sigma^2 = var\left[m\left(X\right)\right]$, is like the variance of a true error term in a regression, c^2 is due to estimation error in $\widehat{m}_h\left(X\right)$ relative to the true $m\left(X\right)$, and r_h is a remainder term that depends on the bandwidth h.
Note: derivation of formula for c^2 involves U-statistic theory and influence functions.

Key point: Unlike $var\left[\widehat{m}_h\left(X\right)\right]$ which was order $n^{-1}h^{-1}$,
have $var\left(\widehat{\mu}\right)$ is of order n^{-1}. Averaging keeps bias but reduces variance.

Applying a Central Limit Theorem gives:

$$n^{1/2}\left[\widehat{\mu}-\mu-h^2B\right] \approx N\left(0,\sigma^2+c^2+r_h\right)$$

More generally, let $\mu = E\left[g\left(X, m\left(X\right)\right)\right]$, $\widehat{\mu} = n^{-1}\sum_{i=0}^{n} g\left[X_i, \widehat{m}_h\left(X_i\right)\right]$.

$$n^{1/2}\left(\widehat{\mu} - \mu\right) = n^{-1/2}\sum_{i=0}^{n} g\left[X_i, m\left(X_i\right)\right] - \mu + \left[\widehat{m}_h\left(X_i\right) - m\left(X_i\right)\right]\frac{\partial g_i}{\partial m_i} + r$$

For remainder $r \to^p 0$ needs h such that $n^{-1/2}\left[\widehat{m}_h\left(X_i\right) - m\left(X_i\right)\right]^2 \to^p 0$, so $n^{1/4+\varepsilon}\left(\widehat{m}_h\left(X_i\right) - m\left(X_i\right)\right)$ bounded in probability (i.e. a faster than $n^{1/4}$ rate of convergence). Recall had rate $n^{2/5}$, so $n^{1/4}$ is feasible.

Then above looks like a weighted average of $\widehat{m}_h\left(X_i\right)$, so as before get

$$n^{1/2}\left[\widehat{\mu} - \mu - h^2 B_g\right] \approx N\left(0, \sigma_g^2 + c_g^2 + r_n\right)$$

for a remainder term r_n. Averaging keeps bias but reduces variance.

With $\mu = E\left[g\left(X, m\left(X\right)\right)\right]$, $\widehat{\mu} = n^{-1}\sum_{i=0}^{n} g\left[X_i, \widehat{m}_h\left(X_i\right)\right]$

If $n^{1/2}h^2 \longrightarrow 0$, $1/nh$ bounded, remainder terms $\longrightarrow 0$, then

$$n^{1/2}\left(\widehat{\mu} - \mu\right) \approx N\left(0, \sigma_g^2 + c_g^2\right)$$

Where σ_g^2 is variance of $g\left(X, m\left(X\right)\right)$, (ordinary averaging variance) and c_g^2 is variance from estimation error in $\widehat{m}_h$.

Key feature of many semiparametric estimators: Focusing on the finite parameters of interest, to get to faster than nonparametric rates of convergence. Getting to root - n generally requires:
1. Averaging
2. Low bias and controlled variance (e.g., may require high order kernels and undersmoothing - extra small bandwidth).

General difficulty in many semiparametric estimators: appropriate selection of bandwidth, smoothing parameters or nonparametric nuisance function estimator.

A Semiparametric Example: Index Model Estimators

Index Models: $E(Y \mid X) = g(X'\beta)$ so $Y = g(X'\beta) + \varepsilon$

Parametric: $g()$ known, or parameterized as $g(X'\beta, \theta)$

Examples: linear regression $g(X'\beta) = X'\beta$, probit $g(X'\beta) = \Phi(X'\beta)$, tobit, generalization of Box-Cox transformation.

If g is completely unknown then $X'\beta$ is only identified up to location and scale.

If e.g., we replaced $X'\beta$ with $X'\widetilde{\beta} = a + X'\beta b$ for any a, and any $b > 0$, then get a new g where the model looks the same:
$Y = g(X'\beta) + \varepsilon = \widetilde{g}\left(X'\widetilde{\beta}\right) + \varepsilon$

A common feature of many semiparametric models is identification and estimation only up to location and/or scale normalizations. β up to location and scale gives marginal rates of substitution.

Index Models: $E(Y \mid X) = g(X'\beta)$ so $Y = g(X'\beta) + \varepsilon$

Powell Stock Stoker (1989) weighted average derivatives:

Let $m(X) = E(Y \mid X)$. $\frac{\partial m(X)}{\partial X} = \frac{\partial g(X'\beta)}{\partial X'\beta}\beta$
Let $w(X) = f(X)^2$ a weighting function chosen for technical convenience.

$E\left[w(X)\frac{\partial m(X)}{\partial X}\right] = E\left[w(X)\frac{\partial g(X'\beta)}{\partial X'\beta}\beta\right] = \kappa\beta$, (scalar κ) is β up to scale.

So let $\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)\widehat{\frac{\partial m(X_i)}{\partial X}}$, can write as a function of kernel density estimates.

Advantages: Weighted average derivative $\widehat{\beta}$ interpretable even if not an index model.

Disadvantages: Requires high dimensional nuisance function estimation, all elements of X must be continuous.

Index Models: $E(Y \mid X) = g(X'\beta)$ so $Y = g(X'\beta) + \varepsilon$

Ichimura (1991): If we knew β, could estimate g by nonparametric regression. If we knew g, could estimate β by nonlinear least squares.

For any $\widetilde{\beta}$ define $g_{\widetilde{\beta}}(s) = E\left(Y \mid X'\widetilde{\beta} = s\right)$. Given any vector $\widetilde{\beta}$, Let $\widehat{g}_{\widetilde{\beta}}(s)$ be the fitted value of a one dimensional kernel regression of Y on S, where $S = X'\widetilde{\beta}$.

Let $\widehat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \widehat{g}_{\beta}(X_i'\beta)\right]^2$

Numerically can kernel estimate $\widehat{g}$ given a $\widehat{\beta}$, then least squares estimate $\widehat{\beta}$ given $\widehat{g}$, and iterate to convergence.

Advantages: Uses a least squares criteria, can have some discrete X elements, has only a one dimensional nonparametric component.

Disadvantages: Can be numerically unstable with multiple local minima.

Index Models: $E(Y \mid X) = g(X'\beta)$ so $Y = g(X'\beta) + \varepsilon$

Maximum Rank Correlation (Han 1987, Sherman 1993):

Assume $g()$ is monotonically increasing. If $X_i'\beta > X_j'\beta$ then $\Pr(Y_i > Y_j) > \Pr(Y_j > Y_i)$. Estimate $\widehat{\beta}$ to maximize how often $X_i'\beta - X_j'\beta$ has the same sign as $Y_i > Y_j$.

$$\widehat{\beta} = \arg\max_\beta \sum_{i=1}^{n} \sum_{j=1}^{i} 1(Y_i > Y_j)\, 1\left(X_i'\beta - X_j'\beta\right)$$

Advantages: No kernel or other nonparametric estimation needed (though limiting distribution does need it). Can have some discrete X elements.

Disadvantages: Numerically difficult - nondifferentiable objective function, can be unstable with multiple local maxima. Requires monotonic g.

How to compare Semiparametric estimators? Three possible ways are:

1. Required assumptions (e.g., error properties, normalizations, smoothness and boundary conditions).

2. Rates of convergence (pointwise or uniform).

3. If root-n, nearness to the Semiparametric Efficiency Bound.

Semiparametric Efficiency Bounds (Chamberlain 1987):

Extension of the parametric model Cramer-Rao lower bound.

Let P be a semiparametric model that has a finite dimensional parameter β.

Let Q be any parametric model (called a submodel) that is a special case of a semiparametric model P.

Let $V_Q = var\left(\widehat{\beta}\right)$ when $\widehat{\beta}$ is efficiently estimated based on model Q.

The semiparametric efficiency bound V for β in P is the maximum V_Q over all possible submodels Q of P.

Semiparametric Efficiency Bounds (Chamberlain 1987):

Example: P could be the binary choice model $Y = 1\left(X'\beta + e > 0\right)$ with unknown distribution e independent of X.

Q could be probit and in that case V_Q would be the variance of β based on probit maximum likelihood. Q could also be logit, or could be the binary choice model with *any* parameterized distribution for e.

The semiparametric efficiency bound V would be the largest V_Q among all models $Y = 1\left(X'\beta + e > 0\right)$ with the e distribution parameterized.

Semiparametric estimation can't do better than the worst parametric special case, since that worst case could be the true model. So a semiparametric estimator $\widehat{\beta}$ with variance V is the best (most efficient) possible.

If can't get root-n, then the efficiency bound is infinite.

Semiparametric Binary Choice Models

$Y = 1\left(X'\beta + e > 0\right)$. Instead of probit or logit where e is known to be normal or logistic, suppose e distribution is unknown.
Scale is arbitrary. Location obtained by assuming e mean or median zero.
β gives marginal rates of substitution. Location needed for reservation prices, e.g.

If only want choice probabilities, can just do nonparametric regression: $E\left(Y \mid X\right) = \Pr\left(Y = 1 \mid X\right)$.

If we know e independent of X, for faster convergence rate can do linear index estimation for β, and then do one dimensional nonparametric regression $E\left(Y \mid X'\beta\right)$.

Specific binary choice semiparametric estimators either allow weaker assumptions than e independent of X, or are more efficient than using general index model estimators for β.

Binary Choice Models $Y = 1\left(X'\beta + e > 0\right)$:

Klein and Spady (1993), based on Cosslett (1983):

Assume e independent of X.

Like Ichimura, for any $\widetilde{\beta}$ define $g_{\widetilde{\beta}}(s) = E\left(Y \mid X'\widetilde{\beta} = s\right)$. Given any vector $\widetilde{\beta}$, Let $\widehat{g}_{\widetilde{\beta}}(s)$ be the fitted value of a one dimensional kernel regression of Y on S, where $S = X'\widetilde{\beta}$.

For true β, $g_{\beta}(s) = E\left(Y \mid X'\beta = s\right) = \Pr\left(-e \leq s\right) = F_{-e}(s)$

If we knew F_{-e}, could do maximum likelihood. So instead do MLE using the estimate $\widehat{g}$:

Let $\widehat{\beta} = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i \ln\left(\widehat{g}_{\beta}\left(X_i'\beta\right)\right) + \left(1 - Y_i\right)\ln\left(1 - \widehat{g}_{\beta}\left(X_i'\beta\right)\right)\right]$

$Y = 1\left(X'\beta + e > 0\right)$ Klein and Spady (1993):

$$g_{\widetilde{\beta}}(s) = E\left(Y \mid X'\widetilde{\beta} = s\right)$$

Let $\widehat{\beta} = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i \ln\left(\widehat{g}_{\beta}\left(X_i'\beta\right)\right) + \left(1 - Y_i\right)\ln\left(1 - \widehat{g}_{\beta}\left(X_i'\beta\right)\right)\right]$

Advantages: Can have some discrete X elements. rate root-n convergence. Attains the semiparametric efficiency bound given independent errors. Automatically get choice probability estimates $\widehat{g}_{\beta}\left(X_i'\beta\right)$.

Disadvantages: Requires kernel and bandwidth choice. Can be numerically unstable with multiple local minima, or $\widehat{g}$ outside of $[0, 1]$. Doesn't permit heteroskedasticity. Does not identify location (Given Klein-Spady estimates, Lewbel 1997 provides an estimator for location based on $E(e) = 0$, and gives moments of e).

Binary Choice Models $Y = 1(X'\beta + e > 0)$:

Maximum Score Estimation (Manski 1975, 1985; Cavanaugh 1987):

Assume $median(e \mid X) = 0$. If $X_i'\beta > 0$ then $\Pr(Y_i = 1) > \Pr(Y_i = 0)$.

Estimate $\widehat{\beta}$ to maximize the number of correct predictions of Y_i

$\widehat{\beta} = \arg\max_\beta \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)\,1(X_i'\beta > 0)$

Advantages: No kernel or other nonparametric estimation needed. Can have some discrete X elements. Can handle some heteroskedasticity.

Disadvantages: Numerically difficult - nondifferentiable objective function. Converges at rate $n^{1/3}$ to a wierd, nonnormal distribution. Only get choice probabilities at the median, unless $e \perp X$. Large variance relative to Klein and Spady (since more general error assumption)

Rate is not root-n even though there is averaging, because it is a local average. Only data in the neighborhood of the median determine the estimate.

Examples: Binary Choice Models $Y = 1\left(X'\beta + e > 0\right)$:

Smoothed Maximum Score (Horowitz 1992):

Again assume $median\left(e \mid X\right) = 0$.
Instead of $\widehat{\beta} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(2Y_i - 1\right) 1\left(X_i'\beta > 0\right)$ do

$$\widehat{\beta} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(2Y_i - 1\right) K\left(X_i'\beta / h\right)$$

where K is any smooth distribution function over the real line and $h \longrightarrow 0$ is a bandwidth.

Advantages: Numerically easier than maximum score, and rate between $n^{1/3}$ and $n^{1/2}$ (the smoother f_e is, the more data we locally average). Can have some discrete X elements. Can handle some heteroskedasticity.

Disadvantages: Slower than root-n. Can only recover choice probability estimates if $e \perp X$. Requires choice of a bandwidth and a kernel type function. Large variance relative to Klein and Spady (since more general error assumption).

Binary Choice Models $Y = 1\left(X'\beta + e > 0\right)$.

Lewbel (2000): Rewrite as $Y = 1\left(V + Z'\gamma + e > 0\right)$
where V is an exogenous regressor with real line support.

Assume e is independent of V, conditional on Z, and $E\left(eZ\right) = 0$. Let $f_V\left(V \mid Z\right)$ be the conditional density of V given Z.

Let

$$\widetilde{Y}_i = \frac{Y_i - 1\left(V_i > 0\right)}{\widehat{f}_V\left(V_i \mid Z_i\right)}$$

Let $\widehat{\gamma}$ be an ordinary least squares linear regression of $\widetilde{Y}$ on Z. Lewbel (2000) shows $\widehat{\gamma} \to \gamma$.

Lewbel (2000). Rewrite binary choice as $Y = 1\left(V + Z'\gamma + e > 0\right)$

$$\widetilde{Y}_i = \frac{Y_i - 1\left(V_i > 0\right)}{\widehat{f}_V\left(V_i \mid Z_i\right)}, \qquad \widehat{\gamma} = \left[\sum_{i=1}^{n}\left(Z_i Z_i'\right)\right]^{-1} \sum_{i=1}^{n}\left(Z_i \widetilde{Y}_i\right)$$

Advantages: Can have some discrete X elements. Estimates location along with the other parameters. Allows for some heteroskedasticity ($var\left(e\right)$ can depend on Z, not V). Given very thick tailed V, is rate root-n and attains the semiparametric efficiency bound given its error assumptions. No numerical optimization required. Can be immediately extended to handle endogenous regressors by instrumental variables. Extends to binary choice fixed effects panels (Honore and Lewbel 2002).

Disadvantages: Requires kernel, bandwidth, high dimension $\widehat{f}_V$. Requires a 'special' (exogenous, real line support, preferably thick tailed) regressor V. Has large variance relative to Klein and Spady (since more general error assumption). Finite sample bias numerically sensitive to V - empirically best when $var\left(V\right)$ is large relative to variance of $var\left(Z'\gamma\right)$.

How Lewbel (2000) works:

Have $Y = 1\left(V + Z'\gamma + e > 0\right)$ where V is an exogenous regressor with real line support. Assume e is independent of V, conditional on Z, and $E\left(eZ\right) = 0$. $f_V\left(V \mid Z\right)$ is the conditional density of V given Z.

Let $S = -\left(Z'\gamma + e\right)$. Then $\gamma = -E\left(ZZ'\right)^{-1} E\left(ZS\right)$, so want to estimate $E\left(ZS\right)$. Now $Y = 1\left(S < V\right)$ so $E\left(Y \mid V, Z\right) = \Pr\left(Y = 1 \mid V, Z\right) = \Pr\left(S < V \mid V, Z\right) = \Pr\left(S < V \mid Z\right) = F_S\left(V \mid Z\right)$.

$$E\left(ZS\right) = E\left[ZE\left(S \mid Z\right)\right] = E\left[Z\int S\frac{\partial F_s\left(S \mid Z\right)}{\partial S}dS\right]$$

Change variables S to V:

$$E\left(ZS\right) = E\left[Z\int V\frac{\partial F_s\left(V \mid Z\right)}{\partial V}dV\right]$$

$$E(ZS) = E\left[Z\int V\frac{\partial F_s(V \mid Z)}{\partial V}dV\right]$$

Now do integration by parts

$$\begin{aligned}
E(ZS) &= E\left[-Z\int [F_s(V \mid Z) - 1(V > 0)]\, dV\right] \\
&= -E\left[Z\int [E(Y \mid V, Z) - 1(V > 0)]\, dV\right] \\
&= -E\left[Z\int \frac{E(Y \mid V, Z) - 1(V > 0)}{f_V(V \mid Z)} f_V(V \mid Z)\, dV\right] \\
&= -E\left[ZE\left(\frac{E(Y \mid V, Z) - 1(V > 0)}{f_V(V \mid Z)} \mid Z\right)\right] \\
&= -E\left[ZE\left(\widetilde{Y} \mid Z\right)\right] = -E\left(Z\widetilde{Y}\right)
\end{aligned}$$

So $\gamma = -E(ZZ')^{-1} E(ZS) = E(ZZ')^{-1} E\left(Z\widetilde{Y}\right)$, which is what we wanted because $S = -(Z'\gamma + e)$.

Semiparametric binary choice extensions:

Endogenous regressors: Control function estimators, "special regressor" estimators. An overview of some of these is in “Simple Estimators for Binary Choice Models with Endogenous Regressors,” by Dong and Lewbel.

Other extensions include:
semiparametric panel data binary choice models,
selection models
treatment models
ordered choice models

Many, many other semiparametric models and estimators exist.

Some good references:

Engle, R.F. and D. L. McFadden (1994) "Handbook of Econometrics, vol. IV," North-Holland., chapters:
36: "Large Sample Estimation and Hypothesis Testing," by Newey, W.K., and McFadden,
38: "Applied Nonparametric Methods," by Hardle, W. and Linton, O.
41: "Estimation of Semiparametric Models," by Powell, J.,

Nonparametric Econometrics by Adrian Pagan and Aman Ullah

Semiparametric Regression for the Applied Econometrician (Themes in Modern Econometrics) by Adonis Yatchew.

Nonparametric Econometrics: Theory and Practice by Qi Li and Jeff Racine.

Nonparametric and Semiparametric Models by Wolfgang Härdle, Marlene Muller, Stefan Sperlich, and Axel Werwatz.