Notes On Nonparametric Regression Estimation

James L. Powell
Department of Economics
University of California, Berkeley

The Nadaraya-Watson Kernel Regression Estimator

Suppose that $z_i \equiv (y_i, x_i'$ is a $(p+1)$-dimensional random vector that is jointly continuously distributed, with y_i being a scalar random variable. Denoting the joint density function of z_i as $f_{y,x}(y,x)$, the conditional mean $g(x)$ of y_i given $x_i = x$ (assuming it exists) is given by

$$\begin{aligned} g(x) &\equiv E[y_i|x_i = x] \\ &= \frac{\int y \cdot f_{y,x}(y,x)dy}{\int f_{y,x}(y,x)dy} \\ &= \frac{\int y \cdot f_{y,x}(y,x)dy}{f_x(x)}, \end{aligned}$$

where $f_x(x)$ is the marginal density function of x_i. If $\hat{f}_{y,x}(y,x)$ is the kernel density estimator of $f_{y,x}(y,x)$, i.e.,

$$\hat{f}_{y,x}(y,x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^{p+1}} \tilde{K}\left(\frac{y-y_i}{h}, \frac{x-x_i}{h}\right)$$

for some $(p+1)$-dimensional kernel function $\tilde{K}(v,u)$ satisfying $\int \tilde{K}(v,u)dvdu = 1$, then an analogue estimator for $g(x) = E[y_i|x_i = x]$ would substitute the kernel estimator $\hat{f}_{y,x}$ for $f_{y,x}$ in the expression for $g(x)$. Further assuming that the first "moment" of $\tilde{K}$ is zero,

$$\int \begin{pmatrix} u \\ v \end{pmatrix} \tilde{K}(v,u)dvdu = 0$$

(which could be ensured by choosing a $\tilde{K}$ that is symmetric about zero with bounded support), this analogue estimator for $g(x)$ can be simplified to

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

where

$$K(u) \equiv \int \tilde{K}(v,u)dv.$$

The estimator $\hat{g}(x)$, known as the *Nadaraya-Watson kernel regression estimator*, can be written as a weighted average

$$\hat{g}(x) \equiv \sum_i w_{in} \cdot y_i,$$

where

$$w_{in} \equiv \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$$

has $\sum_i w_{in} = 1$. Since $K(u) \to 0$ as $\|u\| \to \infty$ (because K is integrable), it follows that $w_{in} \to 0$ for fixed h as $\|x - x_i\| \to \infty$, and also that $w_{in} \to 0$ for fixed $\|x - x_i\|$ as $h \to 0$; hence $\hat{g}(x)$ is a "locally-weighted average" of the dependent variable y_i, with increasing weight put on observations with values of x_i that are close to the target value x as $n \to \infty$.

For the special case of $p = 1$ (i.e., one regressor) and $K(u) = 1\{|u| \leq 1/2\}$ (the density of a $Uniform(-1/2, 1/2)$ variate), the kernel regression estimator $\hat{g}(x)$ takes the form

$$\frac{\sum_{i=1}^{n} 1\{x - h/2 \leq x_i \leq x + h/2\} \cdot y_i}{\sum_{i=1}^{n} 1\{x - h/2 \leq x_i \leq x + h/2\}},$$

an average of y_i values with corresponding x_i values within $h/2$ of x. This estimator is sometimes called the "regressogram," in analogy with the histogram estimator of a density function at x.

Derivation of the conditions for consistency of $\hat{g}(x)$, and of its rate of convergence to $g(x)$, follow the analogous derivations for the kernel density estimator. Indeed, $\hat{g}(x)$ can be written as

$$\hat{g}(x) = \frac{\hat{t}(x)}{\hat{f}(x)},$$

where $\hat{f}(x)$ is the usual kernel density estimator of the marginal density of x_i, so the conditions for consistency of the denominator of $\hat{g}(x)$ – i.e., $h \to 0$ and $nh^p \to \infty$ as $n \to \infty$ – have already been established, and it is easy to show the same conditions imply that

$$\hat{t}(x) \to^p t(x) \equiv g(x)f(x).$$

The bias and variance of the numerator $\hat{t}(x)$ are also straightforward extensions of the corresponding formulae for the kernel density estimator $\hat{f}(x)$; here

$$\begin{aligned} E[\hat{t}(x)] &= E\left[\frac{1}{nh^p}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\cdot y_i\right] \\ &= E\left[\frac{1}{nh^p}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\cdot g(x_i)\right] \\ &= \int \frac{1}{h^p} K\left(\frac{x-z}{h}\right) g(x)f(z)dz \\ &= \int K(u)g(x-hu)f(x-hu)du, \end{aligned}$$

which is the same formula as for the expectation of $\hat{f}(x)$ with "$g(x)f(x)$" replacing "$f(x)$" throughout. Assuming the product $g(x)f(x)$ is twice continously differentiable, etc., the same Taylor's series expansion as for the bias of $\hat{f}(x)$ yields the bias of $\hat{t}(x)$ as

$$\begin{aligned} E[\hat{t}(x)] - g(x)f(x) &= \frac{h^2}{2} tr\left(\frac{\partial^2 g(x)f(x)}{\partial x \partial x'} \cdot \int uu'K(u)du\right) + o(h^2) \\ &= O(h^2). \end{aligned}$$

And the variance of $\hat{t}(x)$ is

$$\begin{aligned}
Var(\hat{t}(x)) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^p}K\left(\frac{x-x_i}{h}\right)y_i\right) \\
&= \frac{1}{n}E\left(\frac{1}{h^p}K\left(\frac{x-x_i}{h}\right)y_i\right)^2 - \frac{1}{n}(E[\hat{t}(x)])^2 \\
&= \frac{1}{n}\int\frac{1}{h^{2p}}\left[K\left(\frac{x-z}{h}\right)\right]^2[\sigma^2(z)+g(z)^2]f(z)dz - \frac{1}{n}(E[\hat{t}(x)])^2 \\
&= \frac{1}{nh^p}\int[K(u)]^2[\sigma^2(x-hu)+g(x-hu)^2]f(x-hu)du - \frac{1}{n}(E[\hat{f}(x)])^2 \\
&= \frac{[\sigma^2(x)+g(x)^2]f(x)}{nh^p}\int[K(u)]^2du + o\left(\frac{1}{nh^p}\right),
\end{aligned}$$

where $\sigma^2(x) \equiv Var[y_i|x_i = x]$. So, as for the kernel density estimator, the MSE of the numerator of $\hat{g}(x)$ is of order $[O(h^2)]^2 + O(1/nh^p)$, and the optimal bandwidth h^* has

$$h^* = O\left(\left(\frac{1}{n}\right)^{1/(p+4)}\right),$$

just like $\hat{f}(x)$. A "delta method" argument then implies that this yields the best rate of convergence of the ratio $\hat{g}(x) = \hat{t}(x)/\hat{f}(x)$ to the true value $g(x)$.

Derivation of the asymptotic distribution of $\hat{g}(x)$ uses that "delta method" argument. First, the Liapunov condition can be verified for the triangular array

$$z_{in} \equiv \frac{1}{h^p}K\left(\frac{x-x_i}{h}\right)(\lambda_1+\lambda_2 y_i),$$

where λ_1 and λ_2 are arbitrary constants, leading to the same requirement as for $\hat{f}(x)$ (namely, $nh^p \to \infty$ as $h \to 0$ and $n \to \infty$) for $\bar{z}_n$ to be asymptotically normal, with

$$\begin{aligned}
\sqrt{nh^p}(\bar{z}_n - E[\bar{z}_n]) &= \sqrt{nh^p}\left(\lambda_1(\hat{f}(x)-E[\hat{f}(x)]) - \lambda_2(\hat{t}(x)-E\left[\hat{t}(x)\right])\right) \\
&\to^d \mathcal{N}(0, \left[\lambda_1^2 + 2\lambda_1\lambda_2 g(x) + \lambda_2^2\left(g(x)^2+\sigma^2(x)\right)\right]f(x)\int[K(u)]^2du). \qquad (**)
\end{aligned}$$

The Cramér-Wald device then implies that the numerator $\hat{t}(x)$ and denominator $\hat{f}(x)$ are jointly asymptotically normal, and the usual delta method approximation

$$\begin{aligned}
\sqrt{nh^p}(\hat{g}(x) - E[\hat{t}(x)]/E[\hat{f}(x)]) &= \frac{\sqrt{nh^p}\left(E[\hat{f}(x)](\hat{t}(x)-E[\hat{t}(x)]) - E[\hat{t}(x)](\hat{f}(x)-E[\hat{f}(x)])\right)}{\hat{f}(x)E[\hat{f}(x)]} \\
&= \frac{\sqrt{nh^p}\left((\hat{t}(x)-E[\hat{t}(x)]) - g(x)(\hat{f}(x)-E[\hat{f}(x)])\right)}{f(x)} \\
&\quad + o_p\left(\sqrt{nh^p}\left(\hat{t}(x)-E[\hat{t}(x)]\right)\right) + o_p\left(\sqrt{nh^p}(\hat{f}(x)-E[\hat{f}(x)])\right)
\end{aligned}$$

yields

$$\sqrt{nh^p}(\hat{g}(x) - E[\hat{t}(x)]/E[\hat{f}(x)]) \to^d \mathcal{N}(0, \frac{\sigma^2(x)}{f(x)}\int[K(u)]^2du)$$

after (1) is applied with $\lambda_1 = -g(x)/f(x)$ and $\lambda_2 = 1/f(x)$.

When the bandwidth tends to zero at the optimal rate,

$$h_n = c\left(\frac{1}{n}\right)^{1/(p+4)},$$

then the asymptotic distribution of $\hat{g}(x)$ is biased when centered at the true value $g(x)$,

$$\sqrt{nh^p}(\hat{g}(x) - g(x)) \rightarrow^d \mathcal{N}(\delta(x), \frac{\sigma^2(x)}{f(x)}\int [K(u)]^2\, du),$$

where now

$$\begin{aligned}
\delta(x) &\equiv \lim \frac{\sqrt{nh^p}\left[(E[\hat{t}(x)] - t(x)) - g(x)(E[\hat{f}(x)] - f(x))\right]}{f(x)} \\
&= \frac{c^{(p+4)/2}}{2f(x)} tr\left[\left(\frac{\partial^2 g(x)f(x)}{\partial x \partial x'} - g(x)\frac{\partial^2 f(x)}{\partial x \partial x'}\right) \cdot \int uu'K(u)du\right].
\end{aligned}$$

And if the bandwidth tends to zero *faster* than the optimal rate, i.e., "undersmoothing" is assumed, so that

$$h^* = o\left(\frac{1}{n}\right)^{1/(p+4)},$$

then

$$\lim \frac{\sqrt{nh^p}\left[(E[\hat{t}(x)] - t(x)) - g(x)(E[\hat{f}(x)] - f(x))\right]}{f(x)} = 0,$$

and the bias term vanishes from the asymptotic distribution,

$$\sqrt{nh^p}(\hat{g}(x) - g(x)) \rightarrow^d \mathcal{N}(0, \frac{\sigma^2(x)}{f(x)}\int [K(u)]^2\, du),$$

as for the kernel density estimator $\hat{f}(x)$.

Discrete Regressors

Some Other Nonparametric Regression Methods

Cross-Validation