3 Nonparametric Regression

3.1 Nadaraya-Watson Regression

Let the data be (y_i, X_i) where y_i is real-valued and X_i is a q-vector, and assume that all are continuously distributed with a joint density $f(y, x)$. Let $f(y \mid x) = f(y, x)/f(x)$ be the conditional density of y_i given X_i where $f(x) = \int f(y, x)\, dy$ is the marginal density of X_i. The regression function for y_i on X_i is

$$g(x) = E(y_i \mid X_i = x).$$

We want to estimate this nonparametrically, with minimal assumptions about g.

If we had a large number of observations where X_i exactly equals x, we could take the average value of the $y_i's$ for these observations. But since X_i is continously distributed, we won't observe multiple observations equalling the same value.

The solution is to consider a neighborhood of x, and note that if X_i has a positive density at x, we should observe a number of observations in this neighborhood, and this number is increasing with the sample size. If the regression function $g(x)$ is continuous, it should be reasonable constant over this neighborhood (if it is small enough), so we can take the average the the y_i values for these observations. The obvious trick is is to determine the size of the neighborhood to trade off the variation in $g(x)$ over the neighborhood (estimation bias) against the number of observations in the neighborhood (estimation variance).

we will observe a large number of X_i in any given neighborhood of x_i.

Take the one-regressor case $q = 1$.

Let a neighborhood of x be $x \pm h$ for some bandwidth $h > 0$. Then a simple nonparametric estimator of $g(x)$ is the average value of the $y_i's$ for the observations i such that X_i is in this neighborhood, that is,

$$\begin{aligned}
\hat{g}(x) &= \frac{\sum_{i=1}^{n} 1(|X_i - x| \leq h)\, y_i}{\sum_{i=1}^{n} 1(|X_i - x| \leq h)} \\
&= \frac{\sum_{i=1}^{n} k\left(\dfrac{X_i - x}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\dfrac{X_i - x}{h}\right)}
\end{aligned}$$

where $k(u)$ is the uniform kernel.

In general, the kernel regression estimator takes this form, where $k(u)$ is a kernel function. It is known as the Nadaraya-Watson estimator, or local constant estimator.

When $q > 1$ the estimator is

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) y_i}{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)}$$

where $K(u)$ is a multivariate kernel function.

As an alternative motivation, note that the regression function can be written as

$$g(x) = \frac{\int y f(y,x)\, dy}{f(x)}$$

where $f(x) = \int f(y,x)\, dy$ is the marginal density of X_i. Now consider estimating g by replacing the density functions by the nonparametric estimates we have already studied. That is,

$$\hat{f}(y,x) = \frac{1}{n|H|h_y} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) k\left(\frac{y_i - y}{h_y}\right)$$

where h_y is a bandwidth for smoothing in the y-direction. Then

$$\begin{aligned}
\hat{f}(x) &= \int \hat{f}(y,x)\, dy \\
&= \frac{1}{n|H|h_y} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) \int k\left(\frac{y_i - y}{h_y}\right) dy \\
&= \frac{1}{n|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)
\end{aligned}$$

and

$$\begin{aligned}
\int y\hat{f}(y,x)\, dy &= \frac{1}{n|H|h_y} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) \int y k\left(\frac{y_i - y}{h_y}\right) dy \\
&= \frac{1}{n|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) y_i
\end{aligned}$$

and thus taking the ratio

$$\begin{aligned}
\hat{g}(x) &= \frac{\frac{1}{n|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) y_i}{\frac{1}{n|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)} \\
&= \frac{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) y_i}{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)}
\end{aligned}$$

again obtaining the Nadaraya-Watson estimator. Note that the bandwidth h_y has disappeared.

The estimator is ill-defined for values of x such that $\hat{f}(x) \leq 0$. This can occur in the tails of the distribution of X_i. As higher-order kernels can yield $\hat{f}(x) < 0$, many authors suggest using only second-order kernels for regression. I am unsure if this is a correct recommendation. If a higher-order kernel is used and for some x we find $\hat{f}(x) < 0$, this suggests that the data is so sparse in that neighborhood of x that it is unreasonable to estimate the regression function. It does not

require the abandonment of higher-order kernels. We will follow convention and typically assume that k is second order ($\nu = 2$) for our presentation.

3.2 Asymptotic Distribution

We analyze the asymptotic distribution of the NW estimator $\hat{g}(x)$ for the case $q = 1$.

Since $E(y_i \mid X_i) = g(X_i)$, we can write the regression equation as $y_i = g(X_i) + e_i$ where $E(e_i \mid X_i) = 0$. We can also write the conditional variance as $E\left(e_i^2 \mid X_i = x\right) = \sigma^2(x)$.

Fix x. Note that

$$\begin{aligned} y_i &= g(X_i) + e_i \\ &= g(x) + (g(X_i) - g(x)) + e_i \end{aligned}$$

and therefore

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

say. It follows that

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

We now analyze the asymptotic distributions of the components $\hat{m}_1(x)$ and $\hat{m}_2(x)$.

First take $\hat{m}_2(x)$. Since $E(e_i \mid X_i) = 0$ it follows that $E\left(k\left(\frac{X_i - x}{h}\right) e_i\right) = 0$ and thus $E(\hat{m}_2(x)) = 0$. Its variance is

$$\begin{aligned} var(\hat{m}_2(x)) &= \frac{1}{nh^2} E\left(k\left(\frac{X_i - x}{h}\right) e_i\right)^2 \\ &= \frac{1}{nh^2} E\left(k\left(\frac{X_i - x}{h}\right)^2 \sigma^2(X_i)\right) \end{aligned}$$

(by conditioning), and this is

$$\frac{1}{nh^2}\int k\left(\frac{z - x}{h}\right)^2 \sigma^2(z) f(z) dz$$

(where $f(z)$ is the density of X_i). Making the change of variables, this equals

$$\begin{aligned}\frac{1}{nh}\int k(u)^2\sigma^2(x+hu)f(x+hu)du &= \frac{1}{nh}\int k(u)^2\sigma^2(x)f(x)du + o\left(\frac{1}{nh}\right)\\ &= \frac{R(k)\sigma^2(x)f(x)}{nh} + o\left(\frac{1}{nh}\right)\end{aligned}$$

if $\sigma^2(x)f(x)$ are smooth in x. We can even apply the CLT to obtain that as $h \to 0$ and $nh \to \infty$,

$$\sqrt{nh}\hat{m}_2(x) \to_d N\left(0, R(k)\sigma^2(x)f(x)\right).$$

Now take $\hat{m}_1(x)$. Its mean is

$$\begin{aligned}E\hat{m}_1(x) &= \frac{1}{h}Ek\left(\frac{X_i - x}{h}\right)(g(X_i) - g(x))\\ &= \frac{1}{h}\int k\left(\frac{z-x}{h}\right)(g(z) - g(x))f(z)dz\\ &= \int k(u)(g(x+hu) - g(x))f(x+hu)du\end{aligned}$$

Now expanding both g and f in Taylor expansions, this equals, up to $o(h^2)$

$$\begin{aligned}&\int k(u)\left(uhg^{(1)}(x) + \frac{u^2h^2}{2}g^{(2)}(x)\right)\left(f(x) + uhf^{(1)}(x)\right)du\\ =\ &\left(\int k(u)udu\right)hg^{(1)}(x)f(x)\\ &+\left(\int k(u)u^2du\right)h^2\left(\frac{1}{2}g^{(2)}(x)f(x) + g^{(1)}(x)f^{(1)}(x)\right)\\ =\ &h^2\kappa_2\left(\frac{1}{2}g^{(2)}(x)f(x) + g^{(1)}(x)f^{(1)}(x)\right)\\ =\ &h^2\kappa_2 B(x)f(x),\end{aligned}$$

where

$$B(x) = \frac{1}{2}g^{(2)}(x) + f(x)^{-1}g^{(1)}(x)f^{(1)}(x)$$

(If k is a higher-order kernel, this is $O(h^\nu)$ instead.) A similar expansion shows that $var(\hat{m}_1(x)) = O\left(\frac{h^2}{nh}\right)$ which is of smaller order than $O\left(\frac{1}{nh}\right)$. Thus

$$\sqrt{nh}\left(\hat{m}_1(x) - h^2\kappa_2 B(x)f(x)\right) \to_p 0$$

and since $\hat{f}(x) \to_p f(x)$,

$$\sqrt{nh}\left(\frac{\hat{m}_1(x)}{\hat{f}(x)} - h^2\kappa_2 B(x)\right) \to_p 0$$

In summary, we have

$$\begin{aligned}\sqrt{nh}\left(\hat{g}(x)-g(x)-h^2\kappa_2 B(x)\right) &= \sqrt{nh}\left(\frac{\hat{m}_1(x)}{\hat{f}(x)}-h^2\kappa_2 B(x)\right)+\frac{\sqrt{nh}\hat{m}_2(x)}{\hat{f}(x)} \\ &\xrightarrow{d} \frac{N\left(0,R(k)\sigma^2(x)f(x)\right)}{f(x)} \\ &= N\left(0,\frac{R(k)\sigma^2(x)}{f(x)}\right)\end{aligned}$$

When X_i is a q-vector, the result is

$$\sqrt{n\left|H\right|}\left(\hat{g}(x)-g(x)-\kappa_2\sum_{j=1}^{q}h_j^2 B_j(x)\right)\xrightarrow{d}= N\left(0,\frac{R(k)^q\sigma^2(x)}{f(x)}\right)$$

where

$$B_j(x)=\frac{1}{2}\frac{\partial^2}{\partial x_j^2}g(x)+f(x)^{-1}\frac{\partial}{\partial x_j}g(x)\frac{\partial}{\partial x_j}f(x).$$

3.3 Mean Squared Error

The AMSE of the NW estimator $\hat{g}(x)$ is

$$AMSE\left(\hat{g}(x)\right)=\kappa_2^2\left(\sum_{j=1}^{q}h_j^2 B_j(x)\right)^2+\frac{R(k)^q\sigma^2(x)}{n\left|H\right|f(x)}$$

A weighted integrated MSE takes the form

$$\begin{aligned}WIMSE &= \int AMSE\left(\hat{g}(x)\right)f(x)M(x)\left(dx\right) \\ &= \kappa_2^2\int\left(\sum_{j=1}^{q}h_j^2 B_j(x)\right)^2 f(x)M(x)\left(dx\right)+\frac{R(k)^q\int\sigma^2(x)M(x)dx}{nh_1h_2\cdots h_q}\end{aligned}$$

where $M(x)$ is a weight function. Possible choices include $M(x)=f(x)$ and $M(x)=1\left(f(x)\geq\delta\right)$ for some $\delta>0$. The AMSE nees the weighting otherwise the integral will not exist.

3.4 Observations about the Asymptotic Distribution

In univariate regression, the optimal rate for the bandwidth is $h_0=Cn^{-1/5}$ with mean-squared convergence $O(n^{-2/5})$. In the multiple regressor case, the optimal bandwidths are $h_j=Cn^{-1/(q+4)}$ with convergence rate $O\left(n^{-2/(q+4)}\right)$. This is the same as for univariate and q-variate density estimation.

If higher-order kernels are used, the optimal bandwidth and convergence rates are again the same as for density estimation.

The asymptotic distribution depends on the kernel through $R(k)$ and κ_2. The optimal kernel minimizes $R(k)$, the same as for density estimation. Thus the Epanechnikov family is optimal for regression.

As the WIMSE depends on the first and second derivatives of the mean function $g(x)$, the optimal bandwidth will depend on these values. When the derivative functions $B_j(x)$ are larger, the optimal bandwidths are smaller, to capture the fluctuations in the function $g(x)$. When the derivatives are smaller, optimal bandwidths are larger, smoother more, and thus reducing the estimation variance.

For nonparametric regression, reference bandwidths are not natural. This is because there is no natural reference $g(x)$ which dictates the first and second derivative. Many authors use the rule-of-thumb bandwidth for density estimation (for the regressors X_i) but there is absolutely no justification for this choice. The theory shows that the optimal bandwidth depends on the curvature in the conditional mean $g(x)$, and this is independent of the marginal density $f(x)$ for which the rule-of-thumb is designed.

3.5 Limitations of the NW estimator

Suppose that $q = 1$ and the true conditional mean is linear $g(x) = \alpha + x\beta$. As this is a very simple situation, we might expect that a nonparametric estimator will work reasonably well. This is not necessarily the case with the NW estimator.

Take the absolutely simplest case that there is not regression error, i.e. $y_i = \alpha + X_i\beta$ identically. A simple scatter plot would reveal the deterministic relationship. How will NW perform?

The answer depends on the marginal distribution of the x_i. If they are not spaced at uniform distances, then $\hat{g}(x) \neq g(x)$. The NW estimator applied to purely linear data yields a nonlinear output!

One way to see the source of the problem is to consider the problem of nonparametrically estimating $E(X_i - x \mid X_i = x) = 0$. The numerator of the NW estimator of the expectation is

$$\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)(X_i - x)$$

but this is (generally) non-zero.

Can the problem by resolved by choice of bandwidth? Actually, it can make things worse. As the bandwidth increases (to increase smoothing) then $\hat{g}(x)$ collapses to a flat function. Recall that the NW estimator is also called the local constant estimator. It is approximating the regression function by a (local) constant. As smoothing increases, the estimator simplifies to a constant, not to a linear function.

Another limitation of the NW estimator occurs at the edges of the support. Again consider the case $q = 1$. For a value of $x \leq \min(X_i)$, then the NW estimator $\hat{g}(x)$ is an average only of y_i values for obsevations to the right of x. If $g(x)$ is positively sloped, the NW estimator will be upward biased. In fact, the estimator is inconsistent at the boundary. This effectively restricts application

of the NW estimator to values of x in the interior of the support of the regressors, and this may too limiting.

3.6 Local Linear Estimator

We started this chapter by motivating the NW estimator at x by taking an average of the y_i values for observations such that X_i are in a neighborhood of x. This is a local constant approximation. Instead, we could fit a linear regression line through the observations in the same neighborhood. If we use a weighting function, this is called the local linear (LL) estimator, and it is quite popular in the recent nonparametric regression literature.

The idea is to fit the local model

$$y_i = \alpha + \beta' (X_i - x) + e_i$$

The reason for using the regressor $X_i - x$ rather than X_i is so that the intercept equals $g(x) = E(y_i \mid X_i = x)$. Once we get the estimates $\hat{\alpha}(x)$, $\hat{\beta}(x)$, we then set $\hat{g}(x) = \hat{\alpha}(x)$. Furthermore, we can use $\hat{\beta}(x)$ to estimate of $\dfrac{\partial}{\partial x} g(x)$.

If we simply fit a linear regression through observations such that $|X_i - x| \leq h$, this can be written as

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left(y_i - \alpha - \beta' (X_i - x)\right)^2 1(|X_i - x| \leq h)$$

or setting

$$Z_i = \begin{pmatrix} 1 \\ X_i - x \end{pmatrix}$$

we have the explicit expression

$$\begin{aligned} \begin{pmatrix} \hat{\alpha}(x) \\ \hat{\beta}(x) \end{pmatrix} &= \left(\sum_{i=1}^{n} 1(|X_i - x| \leq h) Z_i Z_i'\right)^{-1} \left(\sum_{i=1}^{n} 1(|X_i - x| \leq h) Z_i y_i\right) \\ &= \left(\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) Z_i Z_i'\right)^{-1} \left(\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) Z_i y_i\right) \end{aligned}$$

where the second line is valid for any (multivariate) kernel funtion. This is a (locally) weighted regression of y_i on X_i. Algebraically, this equals a WLS estimator.

In contrast to the NW estimator, the LL estimator preserves linear data. That is, if the true data lie on a line $y_i = \alpha + X_i'\beta$, then for any sub-sample, a local linear regression fits exactly, so $\hat{g}(x) = g(x)$. In fact, we will see that the distribution of the LL estimator is invariant to the first derivative of g. It has zero bias when the true regression is linear.

As $h \to \infty$ (smoothing is increased), the LL estimator collapses to the OLS regression of y_i on X_i. In this sense LL is a natural nonparametric generalization of least-squares regression.

The LL estimator also has much better properties at the boundard than the NW estimator. Intuitively, even if x is at the boundard of the regression support, as the local linear estimator fits a (weighted) least-squares line through data near the boundary, if the true relationship is linear this estimator will be unbiased.

Deriving the asymptotic distribution of the LL estimator is similar to that of the NW estimator, but much more involved, so I will not present the argument here. It has the following asymptotic distribution. Let $\hat{g}(x) = \hat{\alpha}(x)$. Then

$$\sqrt{n\,|H|}\left(\hat{g}(x) - g(x) - \kappa_2 \sum_{j=1}^{q} h_j^2 \frac{1}{2}\frac{\partial^2}{\partial x_j^2} g(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q \sigma^2(x)}{f(x)}\right)$$

This is quite similar to the distribution for the NW estimator, with one important difference that the bias term has been simplified. The term involving $f(x)^{-1}\dfrac{\partial}{\partial x_j} g(x) \dfrac{\partial}{\partial x_j} f(x)$ has been eliminated. The asymptotic variance is unchanged.

Strictly speaking, we cannot rank the AMSE of the NW versus the LL estimator. While a bias term has been eliminated, it is possible that the two terms have opposite signs and thereby cancel somewhat. However, the standard intuition is that a simplified bias term suggests reduced bias in practice. The AMSE of the LL estimator only depends on the second derivative of $g(x)$, while that of the NW estimator also depends on the first derivative. We expect this to translate into reduced bias.

Magically, this does not come as a cost in the asymptotic variance. These facts have led the statistics literature to focus on the LL estimator as the preferred approach.

While I agree with this general view, a side not of caution is warrented. Simple simulation experiments show that the LL estimator does not always beat the NW estimator. When the regression function $g(x)$ is quite flat, the NW estimator does better. When the regression function is steeper and curvier, the LL estimator tends to do better. The explanation is that while the two have identical asymptotic variance formulae, in finite samples the NW estimator tends to have a smaller variance. This gives it an advantage in contexts where estimation bias is low (such as when the regression function is flat). The reason why I mention this is that in many economic contexts, it is believed that the regression function may be quite flat with respect to many regressors. In this context it may be better to use NW rather than LL.

3.7 Local Polynomial Estimation

If LL improves on NW, why not local polynomial? The intuition is quite straightforward. Rather than fitting a local linear equation, we can fit a local quadratic, cubic, or polynomial of arbitrary order.

Let p denote the order of the local polynomial. Thus $p = 0$ is the NW estimator, $p = 1$ is the LL estimator, and $p = 2$ is a local quadratic.

Interestingly, the asymptotic behavior differs depending on whether p is even or odd.

When p is odd (e.g. LL), then the bias is of order $O(h^{p+1})$ and is proportional to $g^{(p+1)}(x)$

When p is even (e.g. NW or local quadratic), then the bias is of order $O(h^{p+2})$ but is proportional to $g^{(p+2)}(x)$ and $g^{(p+1)}(x)f^{(1)}(x)/f(x)$.

In either case, the variance is $O\left(\frac{1}{n|H|}\right)$

What happens is that by increasing the polynomial order from even to the next odd number, the order of the bias does change, but the bias simplifies.

By increasing the polynomial order from odd to the next even number, the bias order decreases. This effect is analogous to the bias reduction achieved by higher-order kernels.

While local linear estimation is gaining popularity in econometric practice, local polynomial methods are not typically used. I believe this is mostly because typical econometric applications have $q > 1$, and it is difficult to apply polynomial methods in this context.

3.8 Weighted Nadaraya-Watson Estimator

In the context of conditional distribution estimation, Hall et. al. (1999, JASA) and Cai (2002, ET) proposed a weighted NW estimator with the same asymptotic distribution as the LL estimator. This is discussed on pp. 187-188 of Li-Racine.

The estimator takes the form

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} p_i(x) K\left(H^{-1}\left(X_i - x\right)\right) y_i}{\sum_{i=1}^{n} p_i(x) K\left(H^{-1}\left(X_i - x\right)\right)}$$

where $p_i(x)$ are weights. The weights satisfy

$$\begin{aligned} p_i(x) &\geq 0 \\ \sum_{i=1}^{n} p_i(x) &= 1 \\ \sum_{i=1}^{n} p_i(x) K\left(H^{-1}\left(X_i - x\right)\right)\left(X_i - x\right) &= 0 \end{aligned}$$

The first two requirements set up the $p_i(x)$ as weights. The third equality requires the weights to force the kernel function to satisfy local linearity.

The weights are determined by empirical likelihood. Specifically, for each x, you maximize $\sum_{i=1}^{n} \ln p_i(x)$ subject to the above constraints. The solutions take the form

$$p_i(x) = \frac{1}{n\left(1 + \lambda'\left(X_i - x\right) K\left(H^{-1}\left(X_i - x\right)\right)\right)}$$

where λ is a Lagrange multiplier and is found by numerical optimization. For details about empirical likelihood, see my *Econometrics* lecture notes.

The above authors show that the estimator $\hat{g}(x)$ has the same asymptotic distribution as LL.

When the dependent variable is non-negative, $y_i \geq 0$, the standard and weighted NW estimators

also satisfy $\hat{g}(x) \geq 0$. This is an advantage since it is obvious in this case that $g(x) \geq 0$. In contrast, the LL estimator is not necessarily non-negative.

An important disadvantage of the weighted NW estimator is that it is considerably more computationally cumbersome than the LL estimator. The EL weights must be found separately for each x at which $\hat{g}(x)$ is calculated.

3.9 Residual and Fit

Given any nonparametric estimator $\hat{g}(x)$ we can define the residual $\hat{e}_i = y_i - \hat{g}(X_i)$. Numerically, this requires computing the regression estimate at each observation. For example, in the case of NW estimation,

$$\hat{e}_i = y_i - \frac{\sum_{j=1}^{n} K\left(H^{-1}(X_j - X_i)\right) y_j}{\sum_{j=1}^{n} K\left(H^{-1}(X_j - X_i)\right)}$$

From $\hat{e}_i$ we can compute many conventional regression statistics. For example, the residual variance estimate is $n^{-1}\sum_{i=1}^{n} \hat{e}_i^2$, and R^2 has the standard formula.

One cautionary remark is that since the convergence rate for $\hat{g}$ is slower than $n^{-1/2}$, the same is true for many statistics computed from $\hat{e}_i$.

We can also compute the leave-one-out residuals

$$\begin{aligned} \hat{e}_{i,i-1} &= y_i - \hat{g}_{-i}(X_i) \\ &= y_i - \frac{\sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right) y_j}{\sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right)} \end{aligned}$$

3.10 Cross-Validation

For NW, LL and local polynomial regression, it is critical to have a reliable data-dependent rule for bandwidth selection. One popular and practical approach is cross-validation. The motivation starts by considering the sum-of-squared errors $\sum_{i=1}^{n} \hat{e}_i^2$. One could think about picking h to minimize this quantity. But this is analogous to picking the number of regressors in least-squares by minimizing the sum-of-squared errors. In that context the solution is to pick all possible regressors, as the sum-of-squared errors is monotonically decreasing in the number of regressors. The same is true in nonparametric regression. As the bandwidth h decreases, the in-sample "fit" of the model improves and $\sum_{i=1}^{n} \hat{e}_i^2$ decreases. As h shrinks to zero, $\hat{g}(X_i)$ collapses on y_i to obtain perfect fit, $\hat{e}_i$ shrinks to zero and so does $\sum_{i=1}^{n} \hat{e}_i^2$. It is clearly a poor choice to pick h based on this criterion.

Instead, we can consider the sum-of-squared leave-one-out residuals $\sum_{i=1}^{n} \hat{e}_{i,i-1}^2$. This is a reasonable criterion. Because the quality of $\hat{g}(X_i)$ can be quite poor for tail values of X_i, it may be more sensible to use a trimmed verion of the sum of squared residuals, and this is called the cross-validation criterion

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{i,i-1}^2 M(X_i)$$

(We have also divided by sample size for convenience.) The funtion $M(x)$ is a trimming function, the same as introduced in the definition of WIMSE earlier.

The cross-validation bandwidth h is that which minimizes $CV(h)$. As in the case of density estimation, this needs to be done numerically.

To see that the CV criterion is sensible, let us calculate its expectation. Since $y_i = g(X_i) + e_i$,

$$\begin{aligned} E(CV(h)) &= E\left(\left(e_i + g(X_i) - \hat{g}_{-i}(X_i)\right)^2 M(X_i)\right) \\ &= E\left(\left(g(X_i) - \hat{g}_{-i}(X_i)\right)^2 M(X_i)\right) - 2E\left(e_i\left(g(X_i) - \hat{g}_{-i}(X_i)\right) M(X_i)\right) + E\left(e_i^2 M(X_i)\right). \end{aligned}$$

The third term does not depend on the bandwidth so can be ignored. For the second term we use the law of iterated expectations, conditioning on X_i and I_{-i} (the sample excluding the i'th observation) to obtain

$$E\left(e_i\left(g(X_i) - \hat{g}_{-i}(X_i)\right) M(X_i) \mid I_{-i}, X_i\right) = E(e_i \mid X_i)\left(g(X_i) - \hat{g}_{-i}(X_i)\right) M(X_i) = 0$$

so the unconditional expecation is zero. For the first term we take expectations conditional on I_{-i} to obtain

$$E\left(\left(g(X_i) - \hat{g}_{-i}(X_i)\right)^2 M(X_i) \mid I_{-i}\right) = \int \left(g(x) - \hat{g}_{-i}(x)\right)^2 M(x) f(x)\,(dx)$$

and thus the unconditional expectation is

$$\begin{aligned} E\left(\left(g(X_i) - \hat{g}_{-i}(X_i)\right)^2 M(X_i)\right) &= \int E\left(g(x) - \hat{g}_{-i}(x)\right)^2 M(x) f(x)\,(dx) \\ &= \int E\left(g(x) - \hat{g}(x)\right)^2 M(x) f(x)\,(dx) \\ &= \int MSE\left(\hat{g}(x)\right) M(x) f(x)\,(dx) \end{aligned}$$

which is $WIMSE(h)$.

We have shown that

$$E(CV(h)) = WIMSE(h) + E\left(e_i^2 M(X_i)\right)$$

Thus CV is an estimator of the weighted integrated squared error.

As in the case of density estimation, it can be shown that it is a good estimator of $WIMSE(h)$, in the sense that the minimizer of $CV(h)$ is consistent for the minimizer of $WIMSE(h)$. This holds true for NW, LL and other nonparametric methods. In this sense, cross-validation is a general, practical method for bandwidth selection.

3.11 Displaying Estimates and Pointwise Confidence Bands

When $q = 1$ it is simple to display $\hat{g}(x)$ as a function of x, by calculating the estimator on a grid of values.

When $q > 1$ it is less simple. Writing the estimator as $\hat{g}(x_1, x_2, ..., x_q)$, you can display it as a function of one variable, holding the others fixed. The variables held fixed can be set at their sample means, or varied across a few representative values.

When displaying an estimated regression function, it is good to include confidence bands. Typically this are pointwise confidence intervals, and can be computed using the $\hat{g}(x) \pm 2s(x)$ method, where $s(x)$ is a standard error. Recall that the asymptotic distribution of the NW and LL estimators take the form

$$\sqrt{nh_1 \cdots h_q}\left(\hat{g}(x) - g(x) - Bias(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q \sigma^2(x)}{f(x)}\right).$$

Ignoring the bias (as it cannot be estimated well), this suggests the standard error formula

$$s(x) = \sqrt{\frac{R(k)^q \hat{\sigma}^2(x)}{nh_1 \cdots h_q \hat{f}(x)}}$$

where $\hat{f}(x)$ is an estimate of $f(x)$ and $\hat{\sigma}^2(x)$ is an estimate of $\sigma^2(x) = E\left(e_i^2 \mid X_i = x\right)$.

A simple choice for $\hat{\sigma}^2(x)$ is the sample mean of the residuals $\hat{\sigma}^2$. But this is valid only under conditional homoskedasticity. We discuss nonparametric estimates for $\sigma^2(x)$ shortly.

3.12 Uniform Convergence

For theoretical purposes we often need nonparametric estimators such as $\hat{f}(x)$ or $\hat{g}(x)$ to converge uniformly. The primary applications are two-step and semiparametric estimators which depend on the first step nonparametric estimator. For example, if a two-step estimator depends on the residual $\hat{e}_i$, we note that

$$\hat{e}_i - e_i = g\left(X_i\right) - \hat{g}\left(X_i\right)$$

is hard to handle (in terms of stochastically bounding), as it is an estimated function evaluated at a random variable. If $\hat{g}(x)$ converges to $g(x)$ pointwise in x, but not uniformly in x, then we don't know if the difference $g\left(X_i\right) - \hat{g}\left(X_i\right)$ is converging to zero or not. One solution is to apply a uniform convergence result. That is, the above expression is bounded in absolute value, for $|X_i| \leq C$ for some $C < \infty$ by

$$\sup_{|x| \leq C} \left|g\left(x\right) - \hat{g}\left(x\right)\right|$$

and this is the object of study for uniform convergence.

It turns out that there is some cost to obtain uniformity. While the NW and LL estimators pointwise converge at the rates $n^{-2/(q+4)}$ (the square root of the MSE convergence rate), the uniform

convergence rate is

$$\sup_{|x|\leq C} |g(x) - \hat{g}(x)| = O_p\left(\left(\frac{\ln n}{n}\right)^{2/(q+4)}\right)$$

The $O_p(\cdot)$ symbol means "bounded in probability", meaning that the LHS is bounded beneath a constant this the rate, with probability arbitrarily close to one. Alternatively, the same rate holds almost surely. The difference with the pointwise case is the addition of the extra $\ln n$ term. This is a very slow penalty, but it is a penalty none-the-less.

This rate was shown by Stein to be the best possible rate, so the penalty is not an artifact of the proof technique.

A recent paper of mine provides some generalizations of this result, allowing for dependent data (time series). B. Hansen, Econometric Theory, 2008.

One important feature of this type of bound is the restriction of x to the compact set $|x| \leq C$. This is a bit unfortunate as in applications we often want to apply uniform convergence over the entire support of the regressors, and the latter can be unbounded. One solution is to ignore this technicality, and just "assume" that the regressors are bounded. Another solution is to apply the result using "trimming", a technique which we will probably discuss later, when we do semiparametrics. Finally, as shown in my 2008 paper, it is also possible to allow the constant $C = C_n$ to diverge with n, but at the cost of slowing down the rate of convergence on the RHS.

3.13 NonParametric Variance Estimation

Let $\sigma^2(x) = var(y_i \mid X_i = x)$. It is sometimes of direct economic interest to estimate $\sigma^2(x)$. In other cases we just want to estimate it to get a confidence interval for $g(x)$.

The following method is recommended. Write the model as

$$\begin{aligned} y_i &= g(X_i) + e_i \\ E(e_i \mid X_i) &= 0 \\ e_i^2 &= \sigma^2(X_i) + \eta_i \\ E(\eta_i \mid X_i) &= 0 \end{aligned}$$

Then $\sigma^2(x)$ is the regression function of e_i^2 on X_i.

If e_i^2 were observed, this could be done using NW, weighted NW, or LL regression. While e_i^2 is not observed, it can be replaced by $\hat{e}_i^2$ where $\hat{e}_i = y_i - \hat{g}(X)$ are the nonparametric regression residuals. Using a NW estimator

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^n K\left(H^{-1}(X_i - x)\right)\hat{e}_i^2}{\sum_{i=1}^n K\left(H^{-1}(X_i - x)\right)}$$

and similarly using weighted NW or LL. The bandwidths H are not the same as for estimation of $\hat{g}(x)$, although we use the same notation.

As discussed earlier, the LL estimator $\hat{\sigma}^2(x)$ is not guarenteed to be non-negative, while the NW and weighted NW estimators are always non-negative (if non-negative kernels are used).

Fan and Yao (1998, Biometrika) analyze the asymptotic distribution of this estimator. They obtain the surprising result that the asymptotic distribution of this two-step estimator is identical to that of the one-step idealized estimator

$$\tilde{\sigma}^2(x) = \frac{\sum_{i=1}^n K\left(H^{-1}\left(X_i - x\right)\right) e_i^2}{\sum_{i=1}^n K\left(H^{-1}\left(X_i - x\right)\right)}.$$

That is, the nonparametric regression of $\hat{e}_i^2$ on x_i is asymptotically equivalent to the nonparametric regression of e_i^2 on x_i.

Technically, they demonstrated this result when $\hat{g}$ and $\hat{\sigma}^2$ are computed using LL, but from the nature of the argument it appears that the same holds for the NW estimator. They also only demonstrated the result for $q = 1$, but the extends to the $q > 1$ case.

This is a neat result, and is not typical in two-step estimation. One convenient implication is that we can pick bandwidths in each step based on conventional one-step regression methods, ignoring the two-step nature of the problem. Additionally, we do not have to worry about the first-step estimation of $g(x)$ when computing confidence intervals for $\sigma^2(x)$.