ApEc 8213: Econometric Analysis III -- Lecture #7

Nonparametric and Semiparametric Regression
Hansen, Chapter 19

I. Introduction

The density estimation methods covered in Lecture 6 can be used to estimate **nonparametric regressions**, which have the conditional expectation function (CEF):

$$\mathrm{E}[Y|\, X = x] = m(x)$$

where ***m*(*x*) can have any nonlinear shape**.

For most of this lecture we will assume that there is only one X variable. The model can be written as:

$$Y = m(X) + e$$

$$\mathrm{E}[e|\, X\,] = 0$$

$$\mathrm{E}[e^2|\, X\,] = \sigma^2(X)$$

Note that we are assuming that X is exogenous, but we are also allowing for heteroscedasticity of unknown form for e. We will **assume that both *f*(*x*)**, the density of X, **and *m*(*x*) are continuous in *x***.

I. Binned Means Estimator + Kernel Regression (19.2, 19.3)

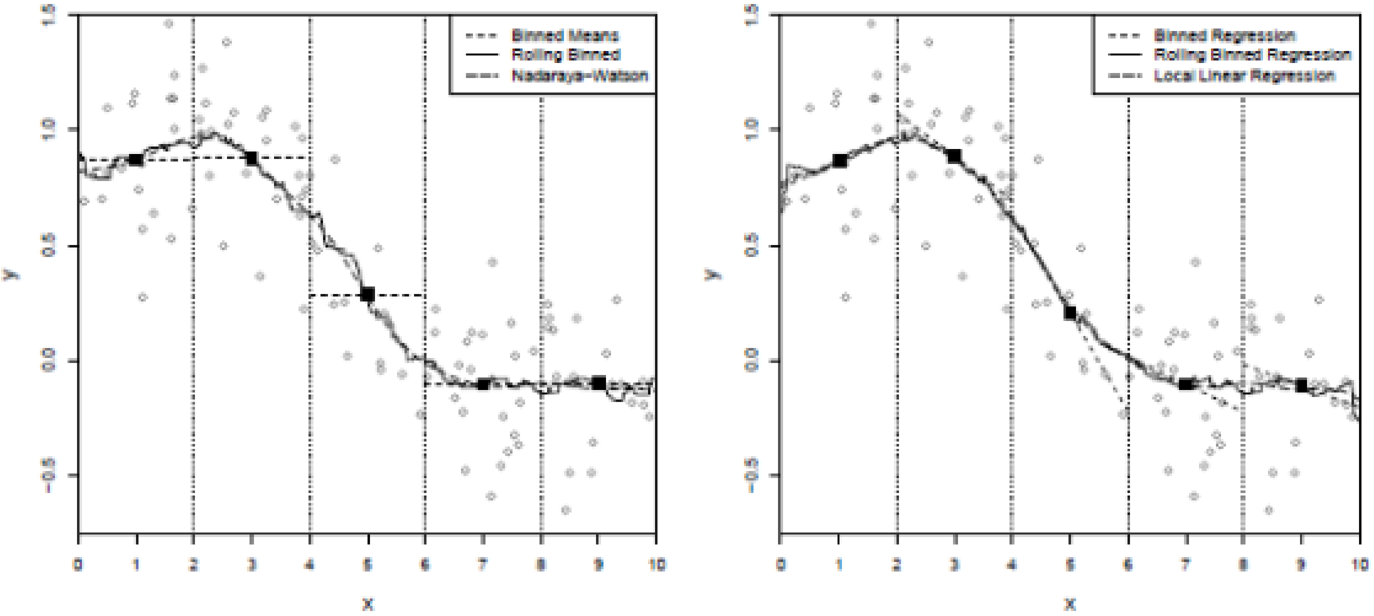
How can one estimate *m*(*x*) for *x*, which is a specific value for the variable X? By definition, $m(x)$ is the mean of Y when $X = x$. But when x is continuous it is **unlikely** that there are **many observations with *X* exactly equal to *x***. **The** "**solution**" here, similar to density estimation, is to **take the mean of *Y* for observations that have *X* "close" to *x***. The simplest case is the "binned means estimator".

Recall the histograms discussion in Lecture 6. The kernel density estimator with a rectangular kernel was a histogram "centered" around x. Similarly, the **binned means estimator** of $m(x)$ for any x within a bin that covers the range from $x - h$ to $x + h$ is the **mean of *Y* for all observations with *x* in that range**. This can be written as:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} 1[|X_i - x| \leq h] Y_i}{\sum_{i=1}^{n} 1[|X_i - x| \leq h]} \qquad (19.1)$$

where 1[ ] is the "indicator function". The **simplest application** of this is to calculate these means for **pre-specified bins**. For example, set $x = 1$ and $h = 1$, and for all x between 0 and 2 use (19.1) to calculate a single $\widehat{m}(x)$ that applies to all x in this range. For all x between 2 and 4, set $x = 3$ (and keep $h = 1$) and do the same. This is shown by the dashed horizontal lines in the left side of Figure 19.1.

Figure 19.1



(a) Nadaraya-Watson (b) Local Linear

This is **not a very good way to estimate *m*(*x*)**, and in particular this is **not at all continuous** and $m(x)$ is assumed to be continuous. It would make **much more sense to have "rolling bins"**. That is, choose about 100 values of x equally spaced in the range of x and calculate (19.1) for each of these 100 points, and then "connect the dots". This is shown by the dark "squiggly" line in the left of Figure 19.1. This is much better but it is not very smooth.
To **smooth this**, let's try using **kernel functions**.

Kernel Regression

Recall the kernel functions $K(u)$ from Lecture 6. **Replace the indicator function** in (19.1) **with a kernel function:**

$$\widehat{m}_{\text{nw}}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \qquad (19.2)$$

This is a **kernel regression** estimator of $m(x)$, and more specifically it is called the **Nadaraya-Watson** estimator (hence the "nw" subscript). As will be seen below, this could also be called the "local constant" estimator. Hansen recommends using the Gaussian kernel function for $K(u)$, which will ensure that $\widehat{m}_{nw}(x)$ is differentiable. However, this is also true of the biweight kernel.

We will discuss "optimal" bandwidth (h) below, but for now it is interesting to point out that **as $\boldsymbol{h} \to \infty$ then $\boldsymbol{\widehat{m}_{\text{nw}}(x)} \to \boldsymbol{\overline{Y}}$**, the sample mean of Y. The dashed curve on the left-hand side of Figure 19.1 shows $\widehat{m}_{\text{nw}}(x)$ for a Gaussian kernel and $h = 1/\sqrt{3}$.

II. Local Linear + Local Polynomial Estimators (19.4, 19.5)

The **shape of $\boldsymbol{m(x)}$** for the **Nadaraya-Watson estimator** is essentially **a constant**: it **calculates a weighted mean of $\boldsymbol{Y}$** with $K(u)$ as the weights. To see this is to note that $\widehat{m}_{\text{nw}}(x)$ solves the following minimization problem:

$$\widehat{m}_{\text{nw}}(x) = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(Y_\text{i} - m)^2$$

That is, differentiating this summation with respect to *m* and setting this derivative to zero yields an *m* equal to $\widehat{m}_{\mathrm{nw}}(x)$. This is a weighted regression with only a constant term.

But regression functions are usually sloped, and it **may be better** to **allow $\boldsymbol{\widehat{m}(x)}$** to be **linear in $\boldsymbol{x}$**. [Draw a picture!]

To allow $m(x)$ to have a slope **for *X* close to *x***, write *Y* as:

$$Y = m(X) + e \approx m(x) + m'(x)(X - x) + e$$

This is the **local linear (LL)** approximation. In effect, it is defined as choosing α and β to minimize the following:

$$\{\widehat{m}_{\mathrm{LL}}(x), \widehat{m}'_{\mathrm{LL}}(x)\} = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(Y_{\mathrm{i}} - \alpha - \beta(X_{\mathrm{i}} - x))^2$$

where $\widehat{m}_{\mathrm{LL}}(x)$ is the **estimate of $\boldsymbol{\alpha}$** and $\widehat{m}'_{\mathrm{LL}}(x)$ **estimates $\boldsymbol{\beta}$**. Both estimates are **different for different values of *x***.

To estimate this, regress Y_i on a constant term and $X_i - x$. More specifically, define:

$$Z_i = \begin{pmatrix} 1 \\ X_i - x \end{pmatrix}, \quad \hat{\beta}_{\mathrm{LL}}(x) = \begin{pmatrix} \widehat{m}_{LL}(x) \\ \widehat{m}'_{\mathrm{LL}}(x) \end{pmatrix}$$

Then the LL estimator is:

$$\hat{\beta}_{\mathrm{LL}}(x) = \left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Z_i(x) Z_i(x)'\right)^{-1} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Z_i(x) Y_i$$

This is essentially weighted least squares with $\boldsymbol{K\left(\frac{X_i - x}{h}\right)}$ as the weights. This is for a particular value of x, and as in density estimation you **do this for about 100 values of $\boldsymbol{x}$**, evenly spaced, and "connect the dots". The right side of Fig. 19.1 shows how this looks, separately for binned regression, rolling binned regression, and local linear regression.

Local Polynomial Estimator

The LL approach can be extended to higher order polynomials in $(X - x)$ using a Taylor expansion:

$$Y \approx m(x) + m'(x)(X - x) + m''(x)(X - x)/2 + \ldots + m^{(p)}(x)(X - x)/p! + e$$

For more details, see Section 19.5 in Hansen. He points out that there is a **trade-off** between the order of the polynomial (p) and the bandwidth (h). **Increasing $\boldsymbol{p}$ improves** the model **fit**, which allows for a larger h for a given level of bias, **but** increasing p **increases** the **estimation variance**.

III. Asymptotic Bias and Asymptotic Variance (19.6, 19.7)

Given that $\mathrm{E}[Y|\, X = x] = m(x)$, and that after conditioning on X the only variable in the **Nadaraya-Watson estimator** is Y, the **conditional expectation** (mean) of that estimator is:

$$\mathrm{E}[\widehat{m}_{\mathrm{nw}}(x)|\,\boldsymbol{X}] = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\mathrm{E}[Y_i|X_i]}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)m(X_i)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \quad (19.3)$$

To calculate the asymptotic bias of $\widehat{m}_{\mathrm{nw}}(x)$ as $n \to \infty$, we **need the following assumptions**:

Assumption 19.1

1. $h \to 0$
2. $nh \to \infty$
3. $m(x)$, $f(x)$ and $\sigma^2(x)$ are continuous in some neighborhood $\mathcal{N}$ around x, where $\sigma^2(x) = \mathrm{E}[e^2|\, X = x]$
4. $f(x) > 0$

The assumptions that ***h* → 0 and *nh* → ∞** imply that the number of **"effective" observations goes to ∞ as *h* → 0**.

Theorem 19.1. Suppose that Assumption 19.1 holds and that $m''(x)$ and $f'(x)$ are continuous in some neighborhood $\mathcal{N}$ around x, Then:

1. $\mathrm{E}[\widehat{m}_{\mathrm{nw}}(x)|\,\boldsymbol{X}] = m(x) + h^2 B_{\mathrm{nw}}(x) + o_p(h^2) + O_p\left(\sqrt{h/n}\right)$

   where $B_{\mathrm{nw}}(x) = (1/2)m''(x) + (f(x))^{-1}f'(x)m'(x)$.

2. $\mathrm{E}[\hat{m}_{\mathrm{LL}}(x)|\,\boldsymbol{X}] = m(x) + h^2B_{\mathrm{LL}}(x) + o_p(h^2) + O_p\left(\sqrt{h/n}\right)$

where $B_{\mathrm{LL}}(x) = (1/2)m''(x)$

The $o_p(h^2) + O_p\left(\sqrt{h/n}\right)$ terms go to 0 as $n \rightarrow \infty$. $\boldsymbol{B_{\mathrm{nw}}(x)}$ **and** $\boldsymbol{B_{\mathrm{LL}}(x)}$ are the **asymptotic bias** of these estimators. For both estimators, this asymptotic bias is **proportional to** $\boldsymbol{h^2}$ (greater smoothing increases the bias), and it also **increases with the "curvature", i.e.** $\boldsymbol{m''(x)}$, of the $m(x)$ function: if $m''(x) < 0$ they are downward biased and if $m''(x) > 0$ then they are upward biased. This is called **smoothing bias**.

The **Nadaraya-Watson estimator** adds an **additional source of bias**: $(f(x))^{-1}f'(x)m'(x)$. **If** $\boldsymbol{f'(x) > 0}$, there are **more points to the right of** $\boldsymbol{x}$ **than to the left**, and averaging over these points will be biased if $m(x)$ has a non-zero slope.

Question: What is the direction of this bias?

The **local linear estimator does not have this problem** because it allows for a non-zero slope of $m(x)$. Hansen concludes (and I agree) that the local linear estimator $\hat{m}_{\mathrm{LL}}(x)$ is preferred to the Nadaraya-Watson estimator.

Asymptotic Variance

Subtracting equation (19.3) from equation (19.2) yields:

$$\widehat{m}_{\text{nw}}(x) - \text{E}[\widehat{m}_{\text{nw}}(x) | \boldsymbol{X}] = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) e_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}, \quad \text{with } e_i = Y_i - m(X_i).$$

After conditioning on X (and x, which is the fixed point that we are considering), **the only variable in this is $\boldsymbol{e_i}$.**

Assuming that the observations are i.i.d, $\text{Var}(\widehat{m}_{\text{nw}}(x) | \boldsymbol{X})$ is:

$$\text{Var}(\widehat{m}_{\text{nw}}(x) | \boldsymbol{X}) = \frac{\sum_{i=1}^{n} \left[K\left(\frac{X_i - x}{h}\right)\right]^2 \sigma^2(X_i)}{\left[\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right]^2} \qquad (19.4)$$

where $\sigma^2(X_i) = \text{E}[e_i^2 | X = X_i]$.

This can be simplified as $n \to \infty$:

Theorem 19.2. Under Assumption 19.1:

1. $\text{Var}(\widehat{m}_{\text{nw}}(x) | \boldsymbol{X}) = \dfrac{R_K \sigma^2(x)}{f(x) nh} + o_p\left(\dfrac{1}{nh}\right)$

2. $\text{Var}(\widehat{m}_{\text{LL}}(x) | \boldsymbol{X}) = \dfrac{R_K \sigma^2(x)}{f(x) nh} + o_p\left(\dfrac{1}{nh}\right)$

where $\sigma^2(x) = \text{E}[e^2 | X = x]$ and $R_K = \int_{-\infty}^{\infty} (K(u))^2 du$ is the "roughness" of kernel $K(u)$.

Note that the **asymptotic variance of these two estimators is identical**. It **decreases when *nh*** (the number of "effective" observations) **increases**, and it **increases when $f(x)$ is smaller** (when there are relatively few observations near this value of x). It is also larger when the (conditional) variance of e, $\sigma^2(x)$, is larger.

IV. Asymptotic IMSE + Bandwidth Choice (19.8 - 19.12)

We saw in Lecture 6 that the **mean squared error** of an estimator can be expressed as the **sum of** its **squared bias and** its **variance**. This also **holds for $\hat{m}(x)$ asymptotically**, so we have (AMSE = asymptotic MSE):

$$\text{AMSE}(\hat{m}(x)) = h^4(B(x))^2 + \frac{R_K\sigma^2(x)}{nhf(x)}$$

where $B(x) = B_{\text{nw}}(x)$ for the Nayarada-Watson estimator and $B(x) = B_{\text{LL}}(x)$ for the local linear estimator.

Of course, this is just for a single value of x, so **we should integrate this over all values of x**:

$$\text{AIMSE}(\hat{m}(x)) = \int_S [h^4(B(x))^2 + \frac{R_K\sigma^2(x)}{nhf(x)}] f(x)w(x)dx \quad (19.5)$$

$$= h^4\bar{B} + \frac{R_K}{nh}\bar{\sigma}^2$$

where S = support of X (region where $f(x) > 0$), $w(x)$ is a weighting function (explained more below), and:

$$\bar{B} = \int_S (B(x))^2 f(x) w(x) dx$$

$$\bar{\sigma}^2 = \int_S \sigma^2(x) w(x) dx$$

If *S* is bounded (range of X is bounded) then there is **no need for a weighting function**. If S is *not* bounded then $\bar{\sigma}^2$ may go to infinity. To avoid this, one can impose upper and lower bounds, denoted by ξ_1 and ξ_2, respectively, and define the **weighting function** as giving equal weight to all x within these bounds: $w(x) = 1[\xi_1 \leq x \leq \xi_2]$.

As with density estimation, we should **choose the bandwidth that minimizes AIMSE**:

Theorem 19.3. The bandwidth that minimizes AIMSE is:

$$h_0 = \left(\frac{R_K \bar{\sigma}^2}{4\bar{B}}\right)^{1/5} n^{-1/5} \qquad (19.6)$$

When h is proportional to $n^{-1/5}$, then AIMSE($\widehat{m}(x)$) = $O(n^{-4/5})$, which means that $n^{4/5}$AIMSE($\widehat{m}(x)$) is bounded (a technical property used later).

Inserting h_0 into the expression for AIMSE($\widehat{m}(x)$), i.e. equation (19.5), which can be denoted as AIMSE$_0$, **yields**:

$$\text{AIMSE}_0 \approx 1.65(R_K{}^4\bar{B}\bar{\sigma}^8)^{1/5}n^{-4/5}$$

Notice that this depends on the kernel function $K(u)$ only through R_K. Since the **Epanechnikov kernel has the smallest value of *R_K*** (see Lecture 8), we have:

Theorem 19.4. The AIMSE (19.5) of the Nadaraya-Watson and local linear regression estimators is minimized by the Epanechnikov kernel function.

Hansen says that **the efficiency benefit** of the Epanechnikov kernel **is small** (1-3%) and recommends the Gaussian kernel.

Reference Bandwidth

A little manipulation of equation (19.6) yields:

$$h_0 = \left(\frac{R_K}{4}\right)^{1/5}\left(\frac{\bar{\sigma}^2}{n\bar{B}}\right)^{1/5} \approx 0.58\left(\frac{\bar{\sigma}^2}{n\bar{B}}\right)^{1/5} \qquad (19.7)$$

Of course, you **still need to calculate $\bar{B}$ and $\bar{\sigma}^2$**, which depend on $f(x)$. Fan and Gijbels (1996) proposed a **"rule of thumb" (ROT) approach** to calculate $\bar{B}$ and $\bar{\sigma}^2$ for the local linear (LL) estimator using regression methods.

First, select bounds (ξ_1, ξ_2) for x and set $w(x) = 1[\xi_1 \le x \le \xi_2]$. In practice, you do not use observations outside of ξ_1 and ξ_2.

Second, estimate a q^{th} order polynomial for $m(x)$:

$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^q$$

Fan and Gijbels suggest $q = 4$. To calculate $\bar{B}$ we need $m''(x)$, which we can approximate by $2\hat{\beta}_2 + 6\hat{\beta}_3 x + 12\hat{\beta}_4 x^2 + \ldots + q(q-1)\hat{\beta}_q x^{q-2}$. **Call this $\boldsymbol{\widehat{m}''(x)}$**.

Third, $\bar{B}$ can then be estimated as:

$$\hat{B} = (1/n) \sum_{i=1}^{n} \left(\frac{1}{2}\widehat{m}''(X_i)\right)^2 1[\xi_1 \le x \le \xi_2] \qquad (19.8)$$

Fourth, assume that the regression error is homoscedastic and thus use $\hat{\sigma}^2$ from the polynomial regression to calculate $\bar{\sigma}^2 = \hat{\sigma}^2(\xi_2 - \xi_1)$. **Putting this all together gives**:

$$h_{ROT} = 0.58\left(\frac{\hat{\sigma}^2(\xi_2 - \xi_1)}{n\hat{B}}\right)^{1/5} \qquad (19.9)$$

Strictly speaking, this is for the local linear estimator only, but see Hansen (p.697) for an argument that this can also be used for the Nadaraya-Watson estimator.

Hansen briefly discusses (p.698) the **choice of $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$**. When there is no upper or lower bound for x he suggests setting these to select a range of x that is "equal to the region of interest for $\widehat{m}(x)$", which is a bit vague. He also suggests what amounts to "dropping the tails" of x, e.g. throwing out the top 5% and bottom 5% of the values of x.

Hansen discusses in detail (Section 19.10) how **“estimation at a boundary”** can lead to bias if you use the Nadaraya-Watson estimator instead of the local linear estimator. I’m not sure how common it is to “estimate at a boundary”, but since there are no disadvantages to using the **local linear estimator**, and it has less bias than the **Nadaraya-Watson estimator**, you should **use the former** rather than the latter, unless you have a clear reason for using the latter.

It is useful to **calculate errors for particular observations**. In general, **using $Y_i - \hat{m}(X_i)$ is not recommended** because as $h \rightarrow 0$ you put more weight on Y_i as an estimate for $\hat{m}(X_i)$, which implies that your estimate of the error will also → 0. Most people recommend a **“leave-one-out” approach** where you calculate $\hat{m}(X_i)$ without using observation i. For example, for the i^{th} observation the **leave-one-out estimate of Y_i**, which can be **denoted by $\tilde{Y}_i$ or $\tilde{m}_{-i}(x)$**, is:

$$\tilde{Y}_i = \tilde{m}_{-i}(x) = \frac{\sum_{j \neq i} K\left(\frac{X_j - x}{h}\right) Y_j}{\sum_{j \neq i} K\left(\frac{X_j - x}{h}\right)}$$

where the “$-i$” subscript indicates that the i^{th} observation is not used. **The leave-one-out prediction error is**:

$$\tilde{e}_i = Y_i - \tilde{Y}_i \qquad (19.10)$$

See Section 19.11 for more details.

Cross-Validation Bandwidth Selection

Applied statisticians prefer to select bandwidths using cross-validation, which is a "leave-one-out" method. The **notation is changed slightly** to show that the estimator of $m(x)$ also depends on the bandwidth: $\boldsymbol{\widehat{m}(x, h)}$**.**

Recall that we want to **select *h* to minimize** the integrated mean-squared error (**IMSE**) of $\widehat{m}(x, h)$, which is:

$$\mathrm{IMSE}_n(h) = \int_S \mathrm{E}[(\widehat{m}(x, h) - m(x))^2]\, f(x)w(x)dx$$

Use the leave-one-out prediction error from equation (19.10), which can be expressed as:

$$\tilde{e}_i(h) = Y_i - \widetilde{m}_{-i}(X_i, h)$$

One can use this to estimate $\mathrm{IMSE}_n(h)$ as follows

$$\mathrm{CV}(h) = (1/n) \sum_{i=1}^{n} (\tilde{e}_i(h))^2 w(X_i) \qquad (19.11)$$

CV(*h*) is called the **cross-validation criterion**. It turns out to be **an unbiased predictor of the IMSE *plus a constant*** for a sample with $n - 1$ observations:

Theorem 19.7.

$$\mathrm{E}[\mathrm{CV}(h)] = \bar{\sigma}^2 + \mathrm{IMSE}_{n-1}(h) \qquad (19.12)$$

where $\bar{\sigma}^2 = \mathrm{E}[e^2 w(X)]$ and $\mathrm{IMSE}_{n\text{-}1}(h)$ is defined as $\mathrm{IMSE}_n(h)$ where $\widetilde{m}_{-i}(X_i, h)$ replaces $(\widehat{m}(x, h)$.

Since $\bar{\sigma}^2$ is a constant that does not depend on h, **the h that minimizes E[CV(h)] is the same h that minimizes $\mathbf{IMSE}_{n\text{-}1}(h)$**, and when n is large the h that minimizes $\mathrm{IMSE}_{n\text{-}1}(h)$ is very close to the h that minimizes $\mathrm{IMSE}_n(h)$. So, select an h than minimizes E[CV(h)].

Thus the **cross-validation bandwidth** $\hat{h}$ is defined as:

$$h_{\mathrm{cv}} = \underset{h \geq h_{min}}{\operatorname{argmin}} \mathrm{CV}(h)$$

where h_{min} is a restriction that "can be" imposed to avoid "unreasonably small bandwidths" (not explained in Hansen, so maybe do not impose it). There is no explicit derivation for h_{cv}, you need to search for it. See Hansen (p.702) for some suggestions on how to do this.

V. Asymptotic Distribution, Conditional Variance and Confidence Bands (19.13 – 19.17).

Before discussing the asymptotic distribution of $\widehat{m}_{\mathrm{nw}}(x)$ and $\widehat{m}_{\mathrm{LL}}(x)$, note that they consistently estimate $m(x)$:

Theorem 19.8. Under Assumption 19.1:

$$\widehat{m}_{\mathrm{nw}}(x) \underset{p}{\rightarrow} m(x) \quad \text{and} \quad \widehat{m}_{\mathrm{LL}}(x) \underset{p}{\rightarrow} m(x)$$

Turn next to the asymptotic distribution of $\widehat{m}_{\mathrm{nw}}(x)$ and $\widehat{m}_{\mathrm{LL}}(x)$:

Theorem 19.9. Suppose that Assumption 19.1 holds, and that $m''(x)$ and $f'(x)$ are continuous in some neighborhood $\mathcal{N}$ around x, and that for some $r > 2$ and $x \in \mathcal{N}$ we have:

$$\mathrm{E}[|e|^r \mid X = x] \leq \bar{\sigma} < \infty \quad \text{and } nh^5 = O(1)$$

Then:

$$\sqrt{nh}(\widehat{m}_{\mathrm{nw}}(x) - m(x) - h^2 B_{\mathrm{nw}}(x)) \underset{d}{\rightarrow} \mathrm{N}\Big(0, \frac{R_K \sigma^2(x)}{f(x)}\Big) \quad (19.16)$$

$$\sqrt{nh}(\widehat{m}_{\mathrm{LL}}(x) - m(x) - h^2 B_{\mathrm{LL}}(x)) \underset{d}{\rightarrow} \mathrm{N}\Big(0, \frac{R_K \sigma^2(x)}{f(x)}\Big)$$

The assumption that $nh^5 = O(1)$ implies that the bandwidth must go to 0 at least at the rate $n^{-1/5}$. Also, because $h \rightarrow 0$ then $\sqrt{nh}$ diverges more slowly than $\sqrt{n}$, and so these nonparametric estimators **converge more slowly than most parametric estimators**. Convergence at the rate $\sqrt{nh}$ implies that nh is the "effective sample size".

The proofs for Theorems 19.8 and 19.9 are in Section 19.26 of Hansen for $\widehat{m}_{\mathrm{nw}}(x)$ and Fan and Gijbels (1996) for $\widehat{m}_{\mathrm{LL}}(x)$.

On pages 704-05, Hansen discusses "**under-smoothing**". The idea is that bias can be "technically eliminated" by selecting a bandwidth h that converges to 0 faster than the optimal rate of $n^{-1/5}$. However, doing so increases the variance of the estimators, and has other problems, so **Hansen does not recommend it**.

Conditional Variance Estimation

The conditional variance is:

$$\sigma^2(x) = \mathrm{Var}[Y | X = x] = \mathrm{E}[e^2 | X = x]$$

An estimate of $\sigma^2(x)$ is **useful for constructing confidence intervals** (and for other tasks). Since $\sigma^2(x)$ is the CEF (conditional expectation function) of e^2 given X, it can be estimated using a nonparametric regression. Consider the **Nadaraya-Watson estimator**. Suppose e is observed, then:

$$\bar{\sigma}^2(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) e_i^2}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \quad (\bar{\sigma}^2 \text{ denotes “ideal”})$$

We need to estimate e_i. One option is $\hat{e}_i(h) = Y_i - \widehat{m}(X_i)$. In fact, a leave-one-out estimate, $\tilde{e}_i(h) = Y_i - \widetilde{m}_{-i}(X_i)$, is better because it avoids “overfitting”. Thus we have:

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tilde{e}_i^2}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \qquad (19.17)$$

This depends on the bandwidth, but **this bandwidth does not have to be the same as the one used to estimate *m*(*x*).** Hansen suggests using ROT or cross-validation using $\tilde{e}_i^2$ as the dependent variable to estimate $\hat{\sigma}^2(x)$. See p.706 for an explanation of why NW is better than LL estimation.

Hansen shows in sections 19.16 and 19.17 how to calculate estimate the variance of $\hat{\beta}(x)$ when using Nadaraya-Watson, local linear or polynomial estimation. He shows how to extend this to clustered errors in section 19.20.

VI. Partially Linear Regression (19.24)

In some cases you may be OK with a regression that has some variables (Z) specified as linear while X is specified as nonparametric. This is a **partially linear regression**:

$$Y = m(X) + Z'\beta + e \quad (19.26)$$

Usually there is only one *X* variable, but this can be extended to several X variables. Robinson (1988) worked this out. The conditional expectation of equation (19.26) is:

$$\mathrm{E}[Y|\,X] = m(X) + \mathrm{E}[Z|\,X]'\beta$$

Subtract this from (19.26):

$$Y - \mathrm{E}[Y|\,X] = (Z - \mathrm{E}[Z|\,X])'\beta + e$$

This is a linear regression of the nonparametric regression error Y – E[Y| X] on the vector of nonparametric regression errors Z – E[Z| X]. **To estimate this**:

1. Nonparametrically regress Y and all Z's on X, and save the fitted values.
2. Regress Y – fitted values on Z's – fitted values to get $\hat{\beta}$.
3. Nonparametrically regress $Y_i - Z_i'\hat{\beta}$ on X_i to estimate $m(x)$.

VII. Multivariate Regression and Specification Tests

It is possible to have more than one X variable, but it is hard to draw a picture of it if you have three or more X variables. See Section 19.22 in Hanson for some details. Deaton's 1997 book is also a useful reference.

Finally, there are specification tests to compare parametric and semiparametric regressions. See pp.119-124 of Yatchew's book *Semiparametric Regression for the Applied Econometrician* (2003).

Happy Valentine's Day!

I ***LOVE***

ECONOMETRICS!