Econ 582
Nonparametric Regression

Eric Zivot

May 28, 2013

Nonparametric Regression

So far we have only considered linear regression models

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ E[\varepsilon_i|\mathbf{x}_i] = 0 \\ E[y_i|\mathbf{x}_i &= \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta},\ \varepsilon_i = y_i - E[y_i|\mathbf{x}_i = \mathbf{x}] \\ \frac{\partial E[y_i|\mathbf{x}_i = \mathbf{x}]}{\partial \mathbf{x}} &= \boldsymbol{\beta} \end{aligned}$$

The assume that $E[y_i|\mathbf{x}_i = \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ is a linear function of x is often done for convenience.

In general, when the components of x are continuously distributed then

$$E[y_i|\mathbf{x}_i = \mathbf{x}] = m(\mathbf{x})$$

can take on any nonlinear shape.

Two cases to consider

- If $E[y_i|\mathbf{x}_i = \mathbf{x}] = m(\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \mathbb{R}^p$ then we have a parametric nonlinear regression model

  $$y_i = m(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon_i$$

  and the parameters $\boldsymbol{\theta}$ and be estimated using nonlinear regression techniques

- If $E[y_i|\mathbf{x}_i = \mathbf{x}] = m(\mathbf{x})$ cannot be modeled parametrically or the parametric form $m(\mathbf{x}, \boldsymbol{\theta})$ is unknown then we have a non-parametric regression

  $$y_i = m(\mathbf{x}) + \varepsilon_i$$

  and we can estimate the function $m(\mathbf{x})$ at each point $\mathbf{x}$ using nonparametric regression techniques.

Binned Estimation of $m(\mathbf{x})$

Consider a nonparametric regression with a single covariate x

$$y_i = m(x_i) + \varepsilon_i$$

Fix the point $x_i = x$ and consider estimating $m(x)$ using a local average of y_i values associated x_i values near x such that $|x_i - x| \leq h$

$$\begin{aligned}
\hat{m}(x) &= \frac{\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right) y_i}{\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right)} = \sum_{i=1}^{n} w_i(x_i) y_i \\
1\left(|x_i - x| \leq h\right) &= 1 \text{ if } |x_i - x| \leq h; 0 \text{ otherwise} \\
w_i(x_i) &= \frac{1\left(|x_i - x| \leq h\right)}{\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right)}
\end{aligned}$$

Note, $\sum_{i=1}^{n} w(x_i) = 1$.

Example: Nonparametric regression (Hansen)

The true model is

$$\begin{aligned} y_i &= m(x_i) + \varepsilon_i,\ i = 1, \ldots, 100 \\ m(x) &= 10 \cdot \log(x) \\ & x_i \sim iid\ N(4, 1) \\ & \varepsilon_i \sim N(0, 16) \\ & y_i | x_i \sim iid\ N(m(x_i), 16) \end{aligned}$$

For binned estimation let $x = 2, 3, 4, 5, 6$ and $h = 0.5$.

Remarks:

- Binned estimator is a step-function (discontinuous estimate of $m(x)$)
- For coarse grid of x the steps (squares in figure) are large
- For a fine grid of x the steps (solid line in figure) are smaller
- The bandwidth h determines the smoothness of the estimate: Small h gives small bins and less smoothness

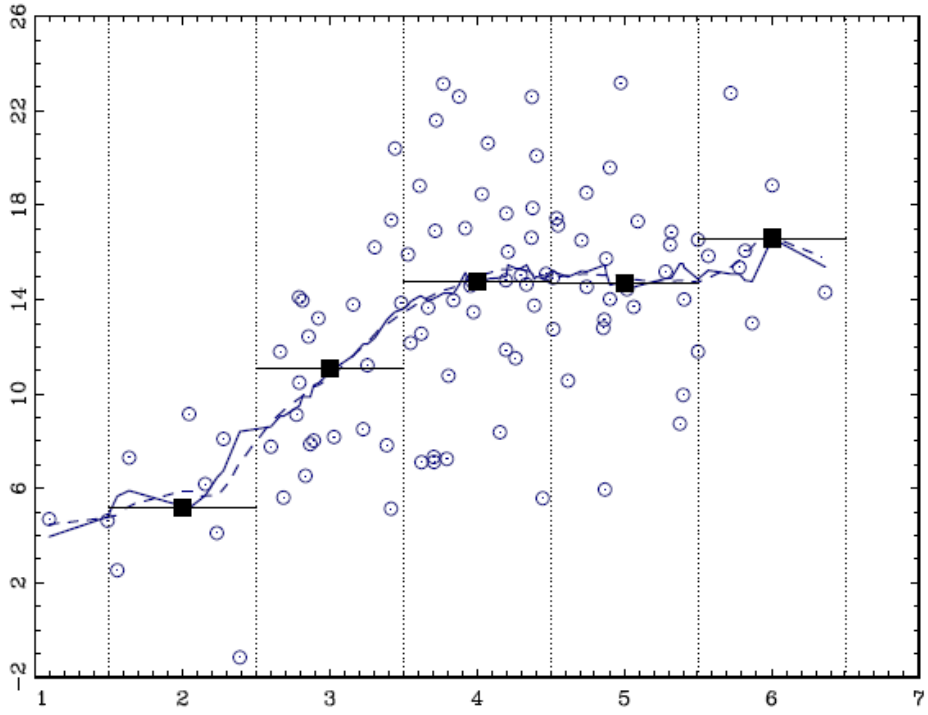

Figure 1: Binned estimator at $x = 2, 3, 4, 5, 6$ with $h = 1/2$ and NW estimator with Epanechnikov kernel

Kernel Regression

- Binned estimator is discontinuous because $w_i(x_i)$ is constructed from indicator functions which are discontinuous

- If $w_i(x_i)$ is constructed from a continuous function then $\hat{m}(x_i)$ will also be continuous.

- Kernel estimators of $m(x)$ are continuous estimators based on continuous kernel weight functions

Example: Kernel weight function based on uniform distribution

Define the weights $1(|x_i - x| \leq h)$ in terms of the uniform density on $[-1, 1]$

$$k_0(u) = \frac{1}{2} 1(|u| \leq 1) = \text{rectangular kernel}$$

Then

$$1(|x_i - x| \leq h) = 1\left(\left(\frac{x_i - x}{h}\right) \leq 1\right) = 2k_0\left(\frac{x_i - x}{h}\right)$$

and

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} k_0\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} k_0\left(\frac{x_i - x}{h}\right)}$$

Definition: A second-order kernel function $k(u)$ satisfies

- $0 \leq k(u) \leq \infty$
- $k(u) = k(-u)$
- $\int k(u)du = 1$
- $\sigma_k^2 = \int u^2 k(u)du < \infty$

Kernel Estimator

Given a kernel weight function $k(u)$, a kernel estimator of $m(x)$ has the form

$$\begin{aligned}\hat{m}(x) &= \frac{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)} \\ &= \sum_{i=1}^{n} w_i(x_i) y_i\end{aligned}$$

where

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

Note: The kernel estimator is also known as the *Nadaraya-Watson* estimator, the *kernel regression* estimator or the *local constant* estimator.

Role of Bandwidth Parameter h

- Bandwidth determines smoothness of estimator: large h gives smoother $\hat{m}(x)$; smaller h gives rougher (more erratic) $\hat{m}(x)$

- $h \to 0 \Rightarrow \hat{m}(x_i) \to y_i$

- $h \to \infty \Rightarrow \hat{m}(x_i) \to \bar{y}$

Commonly used Kernels

1. Epanechnikov kernel

$$k_1(u) = \frac{3}{4}(1 - u^2)1\left(|u| \leq 1\right)$$

2. Gaussian kernel

$$k_4(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

Two important properties of kernels

$$\begin{aligned} \sigma_k^2 &= \int_{-\infty}^{\infty} u^2 k(u) du \\ R_k &= \int_{-\infty}^{\infty} k(u)^2 du \end{aligned}$$

Properties of Commonly Used Kernels

Kernel	**Equation**	R_k	σ_k^2
Uniform	$k_0(u) = \frac{1}{2}1\left(\lvert u\rvert \leq 1\right)$	$1/2$	$1/3$
Epanechnikov	$k_1(u) = \frac{3}{4}(1-u^2)1\left(\lvert u\rvert \leq 1\right)$	$3/5$	$1/5$
Biweight	$k_2(u) = \frac{15}{16}(1-u^2)^2 1\left(\lvert u\rvert \leq 1\right)$	$5/7$	$1/7$
Triweight	$k_3(u) = \frac{35}{32}(1-u^2)^3 1\left(\lvert u\rvert \leq 1\right)$	$350/429$	$1/9$
Gaussian	$k_4(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$	$1/(2\sqrt{\pi})$	1

Local Linear Estimator

The Nadaraya-Watson (NW) kernel estimator is often called a local constant estimator as it locally (about x) approximates $m(x)$ as a constant function. In fact, the NW estimator solves the minimization problem

$$\hat{m}(x) = \arg\min_{\alpha} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)(y_i - \alpha)^2$$

This is a weighted regression of y_i on an intercept only.

A local linear approximation solves the minimization problem

$$\left\{\hat{\alpha}(x), \hat{\beta}(x)\right\} = \arg\min_{\alpha} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)(y_i - \alpha - \beta(x_i - x))^2$$

The local linear estimator of $m(x)$ is the estimated intercept

$$\hat{m}(x) = \hat{\alpha}(x)$$

The local linear estimator of the regression derivative $\nabla m(x)$ is the estimated slope coefficient

$$\widehat{\nabla m}(x) = \hat{\beta}(x)$$

Matrix notation

Define

$$\mathbf{z}_i = \begin{pmatrix} 1 \\ x_i - x \end{pmatrix}, \; k_i = k\left(\frac{x_i - x}{h}\right)$$

Then the LL estimator is the weighted LS estimator

$$\begin{aligned} \begin{pmatrix} \hat{\alpha}(x) \\ \hat{\beta}(x) \end{pmatrix} &= \left(\sum_{i=1}^{n} k_i(x)\mathbf{z}_i(x)\mathbf{z}_i(x)'\right)^{-1} \sum_{i=1}^{n} k_i(x)\mathbf{z}_i(x)y_i \\ &= (\mathbf{Z}'\mathbf{K}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{K}\mathbf{y} \end{aligned}$$

where

$$\underset{(n\times n)}{\mathbf{K}} = \begin{pmatrix} k_1(x) & & \\ & \ddots & \\ & & k_n(x) \end{pmatrix}$$

Remarks

- $h \to \infty \Rightarrow m(x) \to \hat{\alpha}_{OLS} + \hat{\beta}_{OLS} x$ because $k_i(x) \to \frac{1}{n}$
- NW does better than LL when $m(x)$ is close to a flat line
- LL does better than NW when $m(x)$ is meaningfully nonconstant
- LL does better than NW for x values near the boundary of support of x_i

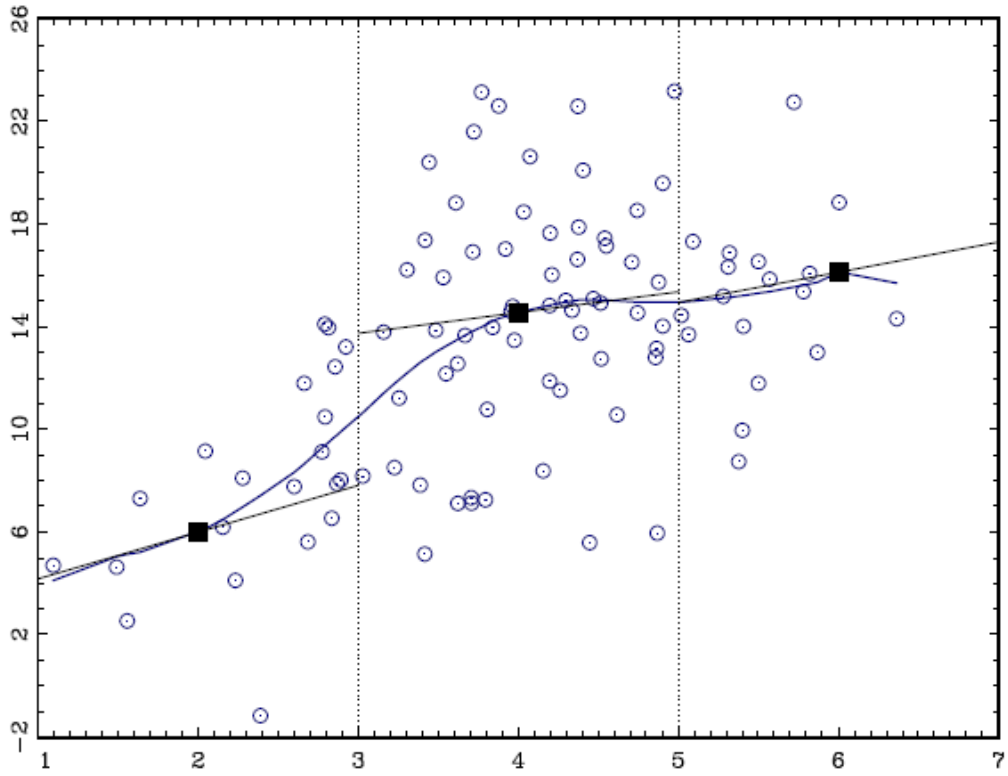

Figure 2: Local linear estimator.

Nonparametric Residuals and Regression Fit

Define the nonparametric residual as

$$\hat{e}_i = y_i - \hat{m}(x_i)$$

Problem: $\hat{e}_i$ is not a good error measure for small h because $\hat{m}(x_i) \longrightarrow y_i$ as $h \longrightarrow 0$ and so

$$\hat{e}_i \longrightarrow 0 \text{ as } h \longrightarrow 0$$

Need a residual that does not suffer from this overfitting problem

Leave-one-out (Jacknife) Residuals (NW Estimator)

Idea: For the NW estimator, we can prevent $\hat{m}(x_i) \longrightarrow y_i$ as $h \longrightarrow 0$ by leaving out x_i and y_i from the non-parametric fit

$$\hat{m}_{-i}(x) = \frac{\sum_{j \neq i} k\left(\frac{x_j - x}{h}\right) y_j}{\sum_{j \neq i}^{n} k\left(\frac{x_j - x}{h}\right)}$$

The leave-one-out (Jacknife) NW predictor and residual for observation i are

$$\begin{aligned} \tilde{y}_i &= \hat{m}_{-i}(x) \\ \tilde{e}_i &= y_i - \tilde{y}_i \end{aligned}$$

Leave-one-out (Jacknife) Residuals (LL Estimator)

The Jacknife LL estimator has the form

$$\begin{aligned}
\begin{pmatrix} \tilde{\alpha}_i \\ \tilde{\beta}_i \end{pmatrix} &= \left(\sum_{j \neq i} k_{ij} \mathbf{z}_{ij} \mathbf{z}_{ij}' \right)^{-1} \sum_{j \neq i} k_{ij} \mathbf{z}_{ij} y_j \\
\mathbf{z}_{ij} &= \begin{pmatrix} 1 \\ x_j - x_i \end{pmatrix} \\
k_{ij} &= k\left(\frac{x_j - x_i}{h} \right)
\end{aligned}$$

and the LL residual is

$$\tilde{e}_i = y_i - \tilde{\alpha}_i$$

Cross Validation and Bandwidth Selection

$$\begin{aligned} y_i &= m(x_i) + e_i,\ var(e_i) = \sigma^2 \\ & \quad x_i \perp e_i \text{ for all } i \\ \hat{m}(x_i, h) &= \text{nonparametric estimate of } m(x_i) \end{aligned}$$

Problem: How to choose h?

- h large $\Rightarrow$ smoother estimator (smaller variance of $\hat{m}(x_i, h)$) but higher bias at each x_i

- h small $\Rightarrow$ noiser estimator (higher variance of $\hat{m}(x_i, h)$) but lower bias at each x_i (recall, $\hat{m}(x_i, h) \longrightarrow y_i$ as $h \longrightarrow 0$)

Key point: Desirable to choose h to minimize the bias-variance tradeoff

MSE, IMSE and MSFE

The mean-squared error (MSE) at x is defined as

$$MSE_n(x,h) = E\left[(\hat{m}(x,h) - m(x))^2\right] = bias(\hat{m}(x,h), m(x))^2 + var(\hat{m}(x,h))$$

and is a function of both x and h

The integrated MSE, a weighted average MSE over all x, is

$$\begin{aligned} IMSE_n(h) &= \int MSE_n(x,h) f_x(x) dx = E\left[MSE_n(x,h)\right] \\ f_x(x) &= \text{pdf of } x \end{aligned}$$

and is only a function of h.

Goal: Find h to minimize $IMSE_n(h)$

Problem: $IMSE_n(h)$ depends on $f_x(x)$ which is unknown

Result: $IMSE_n(h)$ can be estimated using the sample mean-squared forecast error (MSFE)

Let (y_{n+1}, x_{n+1}) be out-of-sample observations independent of the sample. The prediction of y_{n+1} given x_{n+1} is

$$\hat{y}_{n+1} = \hat{m}(x_{n+1}, h)$$

The MSFE is defined as

$$MSFE_n(h) = E\left[(y_{n+1} - \hat{y}_{n+1})^2\right] = E\left[(y_{n+1} - \hat{m}(x_{n+1}, h))^2\right]$$

Using the trivial identity

$$\begin{aligned} y_{n+1} - \hat{m}(x_{n+1}, h) &= y_{n+1} - m(x_{n+1}) + m(x_{n+1}) - \hat{m}(x_{n+1}, h) \\ &= e_{n+1} + m(x_{n+1}) - \hat{m}(x_{n+1}, h) \end{aligned}$$

It can be shown that

$$\begin{aligned} MSFE_n(h) &= E\left[(e_{n+1} + m(x_{n+1}) - \hat{m}(x_{n+1}, h))^2\right] \\ &= \sigma^2 + \int MSE_n(x, h) f_x(x) dx \\ &= \sigma^2 + IMSE_n(h) \end{aligned}$$

Hence, minimizing $MSFE_n(h)$ is equivalent to minimizing $IMSE_n(h)$

Estimating $MSFE_n(h)$

Using the Jacknife nonparametric residuals

$$\tilde{e}_i(h) = y_i - \tilde{m}_{-i}(x_i, h)$$

an estimate of $MSFE_n(h)$ is

$$\widehat{MSFE}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_i(h)^2$$

Treated as a function of h, $\widehat{MSFE}_n(h)$ is called the *cross-validation criterion*

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_i(h)^2$$

Optimal Bandwidth Estimation

The bandwidth h that minimizes an estimate of the IMSE solves

$$\begin{aligned} \hat{h} &= \arg\min_{h \geq h_l} CV(h) \\ h_l &> 0 \end{aligned}$$

Notes:

- Typically, the univariate minimization is done by evaluating $CV(h)$ over a grid $[h_l < h_1 < h_2, \ldots, h_J]$ and choosing $\hat{h}$ as the value that gives the smallest $CV(h)$ over the grid.

- Plots of $CV(h)$ against h provide a visual guide to choosing h

Asymptotic Distribution Theory

Theorem. Let $\hat{m}(x,h)$ denote either the NW or LL estimator of $m(x)$. If x is interior to the support of x_i and $f(x_i) > 0$, then as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$,

$$\sqrt{nh}(\hat{m}(x,h) - m(x) - h^2\sigma_k^2 B(x)) \xrightarrow{d} N\left(0, \frac{R_k \sigma^2(x)}{f_x(x)}\right)$$

$$\hat{m}(x,h) \overset{A}{\sim} N\left(m(x) + h^2\sigma_k^2 B(x), \frac{R_k \sigma^2(x)}{nh \cdot f_x(x)}\right)$$

where

$$\begin{aligned} \sigma^2(x) &= E[e_i^2 | x_i = x] \\ \sigma_k^2 &= \int_{-\infty}^{\infty} u^2 k(u) du, \; R_k = \int_{-\infty}^{\infty} k(u)^2 du \end{aligned}$$

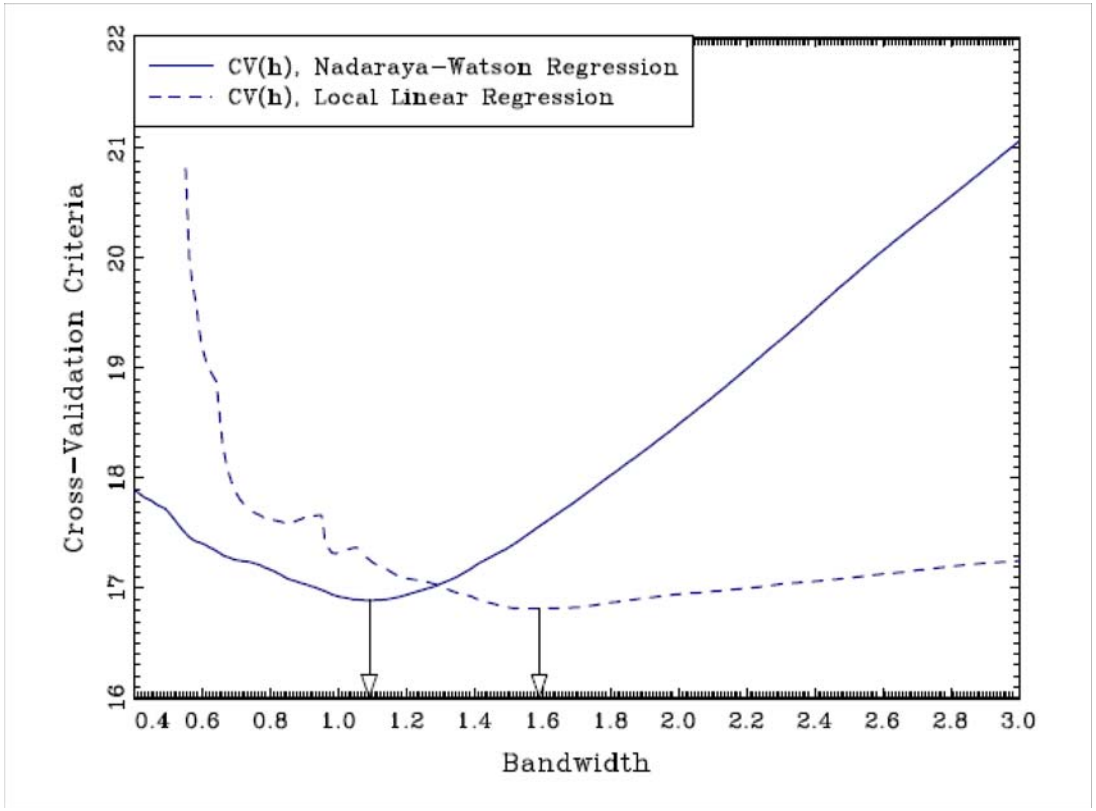

Figure 3: Cross-validation criteria, NW and LL estimators.

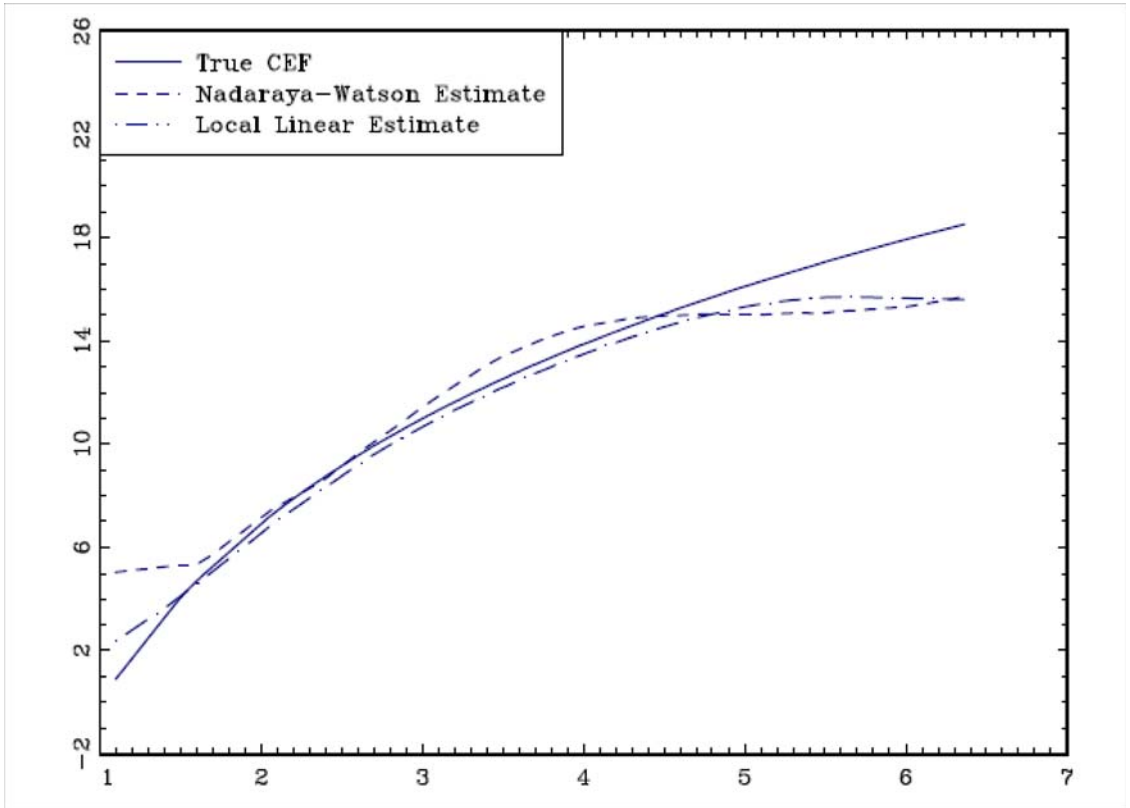

Figure 4: NW and LL estimates using data-dependent CV bandwidths.

The asymptotic bias terms for the NW and LL estimators are

$$\begin{aligned} B_{NW}(x) &= \frac{1}{2}m''(x) + f_x(x)^{-1} f_x'(x) m'(x) \\ B_{LL}(x) &= \frac{1}{2} f_x(x) m''(x) \end{aligned}$$

Remarks:

- Asymptotic variances of NW and LL estimators are the same but biases differ

- $\hat{m}(x, h)$ converges at rate $\sqrt{nh}$ instead of the usual CLT rate of $\sqrt{n}$

- Because $h \to 0$ as $n \to \infty$, $\sqrt{nh}$ diverges slower than $\sqrt{n}$. Hence, nonparametric estimators converge more slowly to their asymptotic distributions than parametric estimators

- $\hat{m}(x, h)$ has an asymptotic bias term $h^2 \sigma_k^2 B(x)$ which depends on h, σ_k^2, $m'(x)$, $m''(x)$ and $f_x(x)$ and $f'_x(x)$

- Asymptotic bias decreases in h and asymptotic variance increases in h

- $B_{NW}(x)$ depends on both $m'(x)$ and $m'(x)$ whereas $B_{LL}(x)$ only depends on $m''(x)$

- $B_{NW}(x) = B_{LL}(x) = 0$ if $m(x)$ is constant (i.e., $m'(x) = m''(x) = 0$)

- $B_{LL}(x)$ is typically lower than $B_{NW}(x)$

Estimating Asymptotic Standard Errors

The asymptotic distribution theory gives the result

$$AVAR(\hat{m}(x,h)) = \frac{R_k \sigma^2(x)}{nh \cdot f_x(x)}$$

The known quantities are R_k, n and h. The unknown quantities are $\sigma^2(x) = E[e_i^2 | x_i = x]$ and $f_x(x)$. An estimate of $AVAR(\hat{m}(x,h))$ uses estimates for $\sigma^2(x)$ and $f_x(x)$

$$\begin{aligned} \widehat{AVAR}(\hat{m}(x,h)) &= \frac{R_k \hat{\sigma}^2(x)}{nh \cdot \hat{f}_x(x)} \\ \widehat{ASE}(\hat{m}(x,h)) &= \sqrt{\frac{R_k \hat{\sigma}^2(x)}{nh \cdot \hat{f}_x(x)}} \end{aligned}$$

Question: How to estimate $\sigma^2(x)$ and $f_x(x)$?

Nonparametric Estimation of $\sigma^2(x) = E[e_i^2|x_i = x]$ and $f_x(x)$

A nonparametric estimate of $\sigma^2(x)$ has the form

$$\tilde{\sigma}^2(x) = \frac{\sum_{i=1}^n k_i(x)\tilde{e}_i^2}{\sum_{i=1}^n k_i(x)}$$

where $\tilde{e}_i$ is the Jackknife residual.

A nonparametric estimate of $f_x(x)$ has the form

$$\hat{f}_x(x) = \frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)$$

Extension to Multiple Regression

$$\begin{aligned} y_i &= E[y_i|\mathbf{x}_i = \mathbf{x}] + y_i - E[y_i|\mathbf{x}_i = \mathbf{x}] \\ &= m(\mathbf{x}) + e_i \\ \mathbf{x}_i &= (x_{1i}, \ldots, x_{di})' \end{aligned}$$

For any vector $\mathbf{x}$ and observation i, define the kernel weights and bandwidth vector

$$\begin{aligned} k_i(\mathbf{x}) &= k\left(\frac{x_{1i} - x_1}{h_1}\right) k\left(\frac{x_{2i} - x_2}{h_2}\right) \cdots k\left(\frac{x_{di} - x_d}{h_d}\right) \\ \mathbf{h} &= (h_1, \ldots, h_d)' \end{aligned}$$

Nonparametric Estimators

Multivariate NW estimator:

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} k_i(\mathbf{x}) y_i}{\sum_{i=1}^{n} k_i(\mathbf{x})}$$

Multivariate LL estimator:

$$\begin{aligned}
\mathbf{z}_i &= \begin{pmatrix} 1 \\ \mathbf{x}_i - \mathbf{x} \end{pmatrix} \\
\begin{pmatrix} \hat{\alpha}(\mathbf{x}) \\ \hat{\beta}(\mathbf{x}) \end{pmatrix} &= \left(\sum_{i=1}^{n} k_i(\mathbf{x}) \mathbf{z}_i(\mathbf{x}) \mathbf{z}_i(\mathbf{x})' \right)^{-1} \sum_{i=1}^{n} k_i(\mathbf{x}) \mathbf{z}_i(\mathbf{x}) y_i \\
&= (\mathbf{Z}'\mathbf{K}\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{K}\mathbf{y}
\end{aligned}$$

Remarks

- Finding the cross-validation bandwidth vector

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \; CV(\mathbf{h})$$

  is a cumbersome numerical problem if d is large

- Asymptotic distribution theory is similar to univariate case with one important difference: convergence rate to asymptotic normal distribution depends on the dimension of $\mathbf{x}$, d. The higher is d, the slower is the convergence rate. This is called the "curse of dimensionality" and is a major problem in nonparametric regression.