ECON 721: Lecture Notes on Nonparametric Density and Regression Estimation

Petra E. Todd

Fall, 2014

Contents

Chapter 1

Review of Stochastic Order Symbols

Let X_n be a sequence of random variables. Then

$$X_n = o_P(1) \text{ if } X_n \to_P 0 \text{ as } n \to \infty$$

and

$$X_n = O_P(a_n) \text{ if } X_n/a_n \text{ is stochastically bounded.}$$

Recall that X_n/a_n is stochastically bounded if

$$\forall \epsilon > 0,\ \exists M < \infty : P(|X_n/a_n| > M) < \epsilon,\ \forall n$$

If a sequence is $o_p(1)$ then it is $O_p(1)$. Convergence in probability implies stochastic boundedness but not vice versa.

Chapter 2

Nonparametric Density Estimation

Nonparametric density estimators allow estimation of densities without having to specify the functional form of the density. Instead, weaker assumptions such as continuity and differentiability can be assumed.

2.1 Histogram Estimator

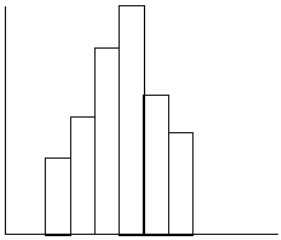
The most commonly used nonparametric density estimator is the histogram estimator. Suppose we estimate density by a histogram with bin width equal to h

Proportion of data in bin:

$$\frac{1}{n}\sum_{i=1}^{n} 1(x_i \in [x_0, x_0 + h]) \approx \hat{f}(x_0)h$$

This suggests a density estimator:

$$\hat{f}(x_0) = \frac{1}{nh}\sum_{i=1}^{n} 1(x_i \in [x_0, x_0 + h])$$



histogram

2.1.1 What is the variance and bias of this estimator?

Bias

$$\begin{aligned}
E\hat{f}(x_0) &= \frac{1}{h}E(1(x_i \in [x_0, x_0+h])) \\
&= \frac{1}{h}\int_{-\infty}^{\infty} 1(x_i \in [x_0, x_0+h]) f(x_i)dx_i \\
&= \frac{1}{h}\int_{x_0}^{x_0+h} \{f(x_0) + f'(\bar{x})(x_i - x_0)\}dx_i \qquad\qquad \bar{x} \in [x_i, x_0] \\
&= f(x_0)\underbrace{\frac{1}{h}\int_{x_0}^{x_0+h} dx_i}_{1} + \underbrace{\left[\int_{x_0}^{x_0+h} f'(\bar{x})(x_i - x_0)\right]\frac{1}{h}}_{R(x_0,h)} dx_i
\end{aligned}$$

Assume the first derivative is bounded:

$$|\ f'(x)| \leq c < \infty \qquad\qquad \forall x$$

In that case, the remainder term can be bounded from above by

$$R(x_0, h) \leq \frac{\int_{x_0}^{x_0+h} |f'(\bar{x})|\,|x_i - x_0|\,dx_i}{h} \leq C \cdot h^2 \cdot \frac{1}{h} = O_p(h)$$

Thus,

$$\left|E\hat{f}(x_0) - f(x_0)\right| \leq Ch$$

Consistency of the histogram requires that $h \to 0$ as $n \to \infty$. Note that these derivations assumed f is differentiable and derivative is bounded.

Variance

Mean-square consistency requires that $var \to 0$ as well. If observations are iid, we can write

$$var\ \hat{f}(x_0) = \frac{1}{n^2h^2}\sum_{i=1}^{n} var\{1(x_i \in [x_0, x_0 + h])\}$$

We will show that the variance is bounded above by something that goes to zero (below, it is bounded by 0).

$$\begin{aligned} var\{1(x_i \in [x_0, x_0 + h])\} &= E\left(1(x_i \in [x_0, x_0 + h])^2\right) - E(1(\cdot))^2 \\ &\leq E\left(1(x_i \in [x_0, x_0 + h])^2\right) \\ &\leq E\left(1(x_i \in [x_0, x_0 + h])\right) \end{aligned}$$

We already analyzed this term above in studying the bias. We showed that

$$E(1(x_i \in [x_0, x_0 + h])) = hf(x_0) + O_p(h^2)$$

Assume that $f(x_0) \leq C'$, so,

$$var\hat{f}(x_0) \leq \frac{C' \cdot n \cdot h}{n^2h^2} = \frac{C'}{nh} = O_p(\frac{1}{nh})$$

Consistency therefore requires $nh \to \infty$

Remarks

(i) The estimator is not root-n consistent, because the variance gets large as $h \to 0$

(ii) For consistency, we require both $h \to 0$ and $n \cdot h \to \infty$

2.1.2 Selecting the bin size

How might you choose h? One way is to minimize the mean-squared error (MSE), focussing on the lowest order terms (the terms that converge the slowest):

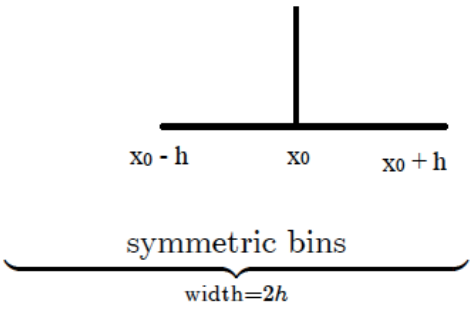
$$MSE = C^2h^2 + \frac{C'}{nh}$$

$$\frac{\partial}{\partial h} = 2 \cdot C^2 h + -\frac{C'}{nh^2} = 0$$
$$2 \cdot C^2 h^3 = +\frac{C'}{n}$$

$$h = \underbrace{[\frac{C'}{2C^2}]^{\frac{1}{3}} n^{-\frac{1}{3}}}_{optimal\ choice\ of\ bin\ width}$$

2.2 Symmetric Binning Histogram Estimator

Instead of constructing the bins as before, consider constructing the bins symmetrically around point of evaluation:



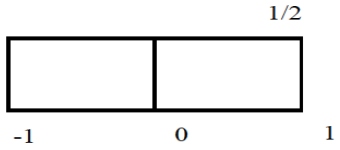
$$\hat{f}(x_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} 1(|x_0 - x_i| < h)$$

Can rewrite as

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^{n} \underbrace{\frac{1}{2} \cdot 1(\left|\frac{x_0 - x_i}{h}\right| < 1)}_{\text{"uniform kernel"}}$$

Note that the area under the kernel function integrates to 1 and the kernel function is symmetric.

Figure 2.1: Symmetric uniform kernel function



2.3 Standard kernel density estimator

Instead of uniform kernel, use kernel with weights that give more weight to closer observations and also go smoothly to zero. We also require that it integrates to 1 and is symmetric.

Let

$$s = \frac{x_0 - x_i}{h}$$

Examples:

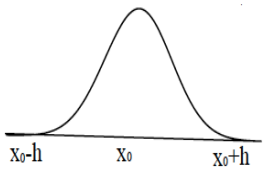
$k(s) = \frac{15}{16}(s^2-1)^2$	if $\lvert s\rvert \leq 1$	"biweight" or "quartic" kernel
$= 0$	$else$	
$k(s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}}$		"normal kernel"

With the normal kernel, the weights are always positive.

The so-called *Nadaraya-Watson Kernel Estimator* is given by

$$\hat{f}(x_0) = \frac{1}{nh_n} \sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n})$$

Figure 2.2: Kernel function with weights that go smoothly to zero



quartic kernel function

2.3.1 Bias

$$
\begin{aligned}
E\hat{f}(x_0) &= \int_{-\infty}^{\infty} \frac{1}{h_n} k(\frac{x_0 - x_i}{h_n}) f(x_i) dx_i \\
let\ s &= \frac{x_0 - x_i}{h_n} \quad ds = -\frac{1}{h_n} \quad x_i = x_0 - sh_n \\
when\ x_i &= -\infty, s_i = \infty \\
&= -\int_{-\infty}^{\infty} k(s) f(x_0 - sh_n) ds \\
&= \int_{-\infty}^{\infty} k(s) f(x_0 - sh_n) ds
\end{aligned}
$$

Assume f is r times differentiable with bounded r^{th} derivative

$$
\begin{aligned}
f(x_0 - sh_n) = f(x_0) + f'(x_0)(-h_n s) + \frac{1}{2} f''(x_0) h_n^2 s^2 + ... \\
... + \frac{1}{r!} f^r(x_0)(-h_n s)^r + \underbrace{\frac{1}{r!}[f^r(\bar{x}) - f^r(x_0)](-h_n s)^r}_{\text{Remainder term R}(x_0,s)}
\end{aligned}
$$

If

$$\int_{-\infty}^{\infty} k(s) s^R = 0 \quad R = 1..r$$

then bias is

$$\int_{-\infty}^{\infty} k(s)R(X_0, s)ds \leq h_n^r C_2 \int_{-\infty}^{\infty} |s|^r k(s)ds$$
(since we assumed bounded rth derivative)

Thus, this estimator improves on histogram if some moments equal 0. Typically, k satisfies

$$\int_{-\infty}^{\infty} k(s)ds = 1$$
$$\int_{-\infty}^{\infty} k(s)sds = 0$$

If we require

$$\int_{-\infty}^{\infty} k(s)s^2ds = 0$$

Then $k(s)$ must be negative in places, as in the figure below.

Remark If f^r satisfies a lipschitz condition

$$i.e. \, |f^r(\bar{x}) - f^r(x_0)| \leq m \, |\bar{x} - x_0|$$

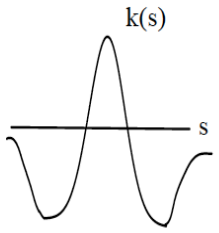
then

$$Bias \leq C_0 \cdot h_n^{r+1} \int_{-\infty}^{\infty} |s|^r k(s)ds = \underbrace{O_p(h_n^{r+1})}_{higher \; order}$$

2.3.2 Variance

$Var \, \hat{f}(x_0) = \frac{1}{n^2 h_n^2} \sum_{i=1}^{n} Vark(\frac{x_0 - x_i}{h_n})$
$= \frac{1}{nh_n^2}[Ek^2(\frac{x_0 - x_i}{h_n}) - Ek(\frac{x_0 - x_i}{h_n})^2]$ can show using same techniques
$\leq \frac{C_1}{nh}$
$= O_p(\frac{1}{nh_n})$ $\Rightarrow$ does not improve on histogram

Figure 2.3: Kernel function that is sometimes negative



2.3.3 Choice of Bandwidth

Choose h to minimize the asymptotic mean-squared error (AMSE):

$$\min_h AMSE(h) = C_2^2 h_n^{2r} + \frac{C_1}{nh_n}$$

$$\frac{\partial}{\partial h} = C_2^2 \cdot 2r \cdot h_n^{2r-1} - \frac{C_1}{nh_n^2} = 0$$

$$h_n^{2r+1} = \frac{C_1}{C_2^2 2rn}$$

$$\Rightarrow h_n = (\frac{C_1}{2rC_2})^{\frac{1}{2r+1}} \cdot (n)^{-\frac{1}{2r+1}}$$

h_n shrinks to zero as n grows large

Now, let's look at order of AMSE (without making the Lipschitz assumption). Plug in the optimal value for h_n that was obtained above. Let $C_3 = \frac{C_1}{2rC_2}^{\frac{1}{2r+1}}$.

$$\begin{aligned} MSE(h^k) &= C_2 C_3^{2r} n^{-\frac{2r}{2r+1}} + \frac{C_1}{n \cdot C_3 n^{-\frac{1}{2r+1}}} \\ &= C_2 C_3^{2r} n^{-\frac{2r}{2r+1}} + C_1 C_3^{-1} n^{-\frac{2r}{2r+1}} \\ &= O_p(n^{-\frac{2r}{2r+1}}) \end{aligned}$$

$$\text{so } \sqrt{MSE} = O_p(n^{-\frac{r}{2r+1}})$$

$$\frac{r}{2r+1} < \frac{1}{2}$$

(parametric rate is $n^{-\frac{1}{2}}$)

Recall

$$(\hat{\beta}_{ols} - \beta_0) \sim (0, \underbrace{\sigma^2 (\frac{X'X}{N})^{-1}}_{Var} \cdot N^{-1})$$

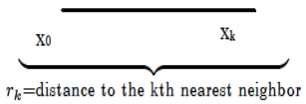
$RMSE = \sqrt{MSE}$ is $O_p(N^{-\frac{1}{2}})$

How do you generalize kernel density estimator to more dimensions?

$$\hat{f}(x_0, z_0) = \frac{1}{n h_{n,x} h_{n,z}} \sum_{i=1}^{n} k_1(\frac{x_0 - x_i}{h_{n,x}}) k_2(\frac{z_0 - z_i}{h_{n,z}})$$

2.4 Alternative Estimator- k nearest neighbor

Alternatively, the bandwidth can be chosen so as to include a certain number of "neighbors" in each point estimation. In this case, the bandwidth will be variable:



$$f(x_0) \cdot 2r_k = \frac{\#data \text{ in bin}=k-1}{n}$$
$$\hat{f}(x_0) = \frac{k-1}{2r_k \cdot n}$$

For m-dimensional data,

$$\hat{f}(x_0) = \frac{k-1}{(2r_k)^m \cdot n}$$

2.5 Asymptotic Distribution of Kernel Density

We already did consistency (we showed conditions required for conv. in mean square)

$$\hat{f}(x_0) = \frac{1}{nh_n}\sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n}) \underset{\text{by LLN}}{\overset{p}{\to}} E\ \hat{f}(x_0) \to f(x_0) \qquad as\ h \to 0$$
$$VAR\ \hat{f}(x_0) \overset{p}{\to} 0 \quad as\ nh_n \to \infty$$

Now, asymptotic distribution

$$\sqrt{nh_n}(\frac{1}{nh_n}\sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n}) - f(x_0)) = \sqrt{nh_n}(\frac{1}{nh_n}\sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n}) - E\ k(\frac{x_0 - x_i}{h_n})) \quad (1)$$
$$+ \sqrt{nh_n}(\frac{1}{nh_n}\sum_{i=1}^{n} E\ k(\frac{x_0 - x_i}{h_n}) - f(x_0)) \quad (2)$$

Analyze term (1):

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\sqrt{h_n}}k(\frac{x_0-x_i}{h_n}) - E(\frac{1}{\sqrt{h_n}}k(\frac{x_0-x_i}{h_n}))\right]$$

can find a CLT for this part

$$\sim N\ (0, VAR\frac{1}{\sqrt{h_n}}k(\frac{x_0-x_i}{h_n}))$$

$$VAR\frac{1}{\sqrt{h_n}}k(\frac{x_0-x_i}{h_n}) = \underbrace{\frac{1}{h_n}E\ (k(\frac{x_0-x_i}{h_n})^2)}_{f(x_0)\int_{-\infty}^{\infty}k^2(s)ds} - \frac{1}{h_n}E\ (k(\frac{x_0-x_i}{h_n}))^2$$

$$\text{Assume } \int_{-\infty}^{\infty}k(s)ds = 1, \int_{-\infty}^{\infty}k(s)sds = 0$$

$$[\frac{1}{h_n}\cdot O(h_n^2)] \to 0 \quad as\ h_n \to 0$$

$$\text{Term (1)} \sim N(0, f(x_0)\int_{-\infty}^{\infty}k^2(s)ds)$$

Now, show Term (2)$\xrightarrow{p}$ 0 (will require more stringent conditions on the bandwidth)

$$\sqrt{nh_n}(\frac{1}{nh_n}\sum_{i=1}^{n}E\ k(\frac{x_0-x_i}{h_n}) - f(x_0)) \ = \sqrt{nh_n}\{E\frac{1}{h_n}k(\frac{x_0-x_i}{h_n}) - f(x_0)\}$$

$$= \frac{1}{h_n}\int_{-\infty}^{\infty}k(\frac{x_0-x_i}{h_n})f(x_i)dx_i \qquad s = \frac{x_0-x_i}{h_n} \qquad h_nds = -dx_i$$

$$= \int_{-\infty}^{\infty}k(s)f(x_0-sh_n)ds$$

$$= \int_{-\infty}^{\infty}k(s)[f(x_0)+f'(x_0)(-sh_n)+\frac{f''(\bar{x})}{2}s^2h_n^2]ds$$

$$= f(x_0)\int k(s)ds + -f'(x_0)\cdot h_n\int_{-\infty}^{\infty}k(s)sds + h_n^2\int_{-\infty}^{\infty}\frac{f''(\bar{x})}{2}k(s)s^2ds$$

$\bar{x}$ between x_0 and $x_0 + sh$.

Assume $|f''(\bar{x})| \le M$ and $\int_{-\infty}^{\infty}k(s)sds = 0$ (satisfied if k is symmetric).

then

$$E\ \frac{1}{h_n}k\left(\frac{x_0 - x_i}{h_n}\right) - f(x_0)\ is\ O_p(h_n^2)$$

which implies that

$$\sqrt{nh_n}E(\frac{1}{h_n}k(\frac{x_0 - x_i}{h_n}) - f(x_0))\ \text{ is } O_p(\sqrt{nh_n}\cdot h_n^2) = O_p(\sqrt{nh_n^5})$$
$$\sqrt{nh_n}E(\frac{1}{h_n}k(\frac{x_0 - x_i}{h_n}) - f(x_0)) = O_p(\sqrt{nh_n^5})$$

Hence, we require $nh_n^5 \to 0$ in addition to $h_n \to 0$, $nh_n \to \infty$. The required condition for asymptotic normality is stronger than for consistency.

Chapter 3

Nonparametric Regression

3.1 Overview

Consider the model:

$$y_i = g(x_i) + \epsilon_i$$

where we assume

$$E(\epsilon_i|x_i) = 0$$
$$E(\epsilon_i^2|x_i) = c < \infty$$

and we would like to estimated $g(x)$ without having to impose functional form assumptions on g, except assumptions such as continuity and differentiability.

There are two types of nonparametric estimation methods: local and global. These two approaches reflect two different ways to reduce the problem of estimating a function into estimation of real numbers.

Local approaches consider a real valued function $g(x)$ at a single point $x = x_0$. The problem of estimating a function becomes estimating a real number $h(x_0)$. If we are interested in evaluating the function in the neighborhood of the point x_0, we can approximate the function by $g(x_0)$ or, if $g(x)$ is continuously differentiable at x_0, then a better approximation might be $g(x_0) + g'(x_0)(x - x_0)$. Thus, the problem of estimating a function at a

point may be thought of as estimating two real numbers $g(x_0)$ and $g'(x_0)$, making use of observations in the neighborhood. Either way, if we want to estimate the function over a wider range of x values, the same, pointwise problem can be solved at the different points of evaluation.

Global approaches introduce a coordinate system in a space of functions, which reduces the problem of estimating a function into that of estimating a set of real numbers. Any element v in a d-dimensional vector space can be uniquely expressed using a system of independent vectors $\{b_j\}_{j=1}^d$ as $v = \sum_{j=1}^d \theta_j \cdot b_j$, where one can think of $\{b_j\}_{j=1}^d$ as a system of coordinates and $(\theta_1, ..., \theta_d)'$ as the representation of v using the coordinate system. Likewise, using an appropriate set of linearly independent functions $\{\phi_j(x)\}_{j=1}^\infty$ as coordinates any square integrable real valued function can be uniquely expressed by a set of coefficients. That is, given an appropriate set of linearly independent functions $\{\phi_j(x)\}_{j=1}^\infty$, any square integrable function $g(x)$ has unique coefficients $\{\theta_j\}_{j=1}^\infty$ such that

$$g(x) = \sum_{j=1}^{\infty} \theta_j \cdot \phi_j(x).$$

One can think of $\{\phi_j(x)\}_{j=1}^\infty$ as a system of coordinates and $(\theta_1, \theta_2, ...)'$ as the representation of $g(x)$ using the coordinate system. This observation allows us to translate the problem of estimating a function into a problem of estimating a sequence of real numbers $\{\theta_j\}_{j=1}^\infty$.

Well known bases are polynomial series and Fourier series. These bases are infinitely differentiable everywhere. Other well known bases are polynomial spline bases and wavelet bases.

The literature on nonparametric estimation methods is vast. These notes will focus on local methods, in particular, kernel-based methods, which are among the most commonly used (at least by economists!).

3.2 The Local approach

Examples: kernel regression, local polynomial regression

Early smoothing methods date back to the 1860's in actuarial literature (see Cleveland (1979, JASA)), where they were used to study life expectancies at different ages.

One possibility is to estimate the conditional mean by:

$$E(y|x) = \int yf(y|x)dy = \int y\frac{f(x,y)}{f(x)}dy$$
$$\hat{E}(y|x) = \int y\frac{\hat{f}(x,y)}{\hat{f}(x)}dy.$$

Or, with repeated data at each point, could estimate by

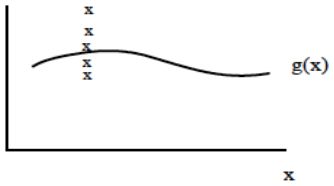
$$g(x_0) \text{ by } \hat{g}(x_0) = \frac{\sum_i y_i 1(x_i = x_0)}{\sum_i 1(x_i = x_0)}$$

A drawback is that we can only estimate this way at points where we have data. Need a way of interpolating or extrapolating between and outside data observations. We could create cells as in the histogram estimator:

$$\hat{g}(x_0) = \frac{\sum_{i=1}^{n} y_i 1(x_i \in [x_0 - h_n, x_0 + h_n])}{\sum_{i=1}^{n} 1(x_i \in [x_0 - h_n, x_0 + h_n])}$$
$$= \frac{\sum_{i=1}^{n} y_i 1(\left|\frac{x_0 - x_i}{h_n}\right| < 1)}{\sum_{i=1}^{n} 1(\left|\frac{x_0 - x_i}{h_n}\right| < 1)}$$

If we want $\hat{g}(x)$ to be smooth, we need to choose a kernel function that goes to 0 smoothly

Figure 3.1: Repeated data estimator



$$\hat{g}(x_0) = \frac{\sum_{i=1}^{n} y_i k(\frac{x_0 - x_i}{h_n})}{\sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n})}$$

$$= \frac{\frac{1}{nh_n}\sum_{i=1}^{n} y_i k(\frac{x_0 - x_i}{h_n}) \qquad (\to g(x_0)f(x_0))}{\frac{1}{nh_n}\sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n}) \qquad (\to f(x_0))}$$

$$= \sum_{i=1}^{n} y_i w_i(x_0)$$

where

$$w_i(x_0) = \frac{k(\frac{x_0 - x_i}{h_n})}{\sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n})} \qquad note : \sum_{i=1}^{n} w_i(x_0) = 1$$

We don't need $\int k(s)ds = 1$, but we do require that kernel used in numerator and denominator integrate to the same value.

3.2.1 Two types of potential problems

(i) When density at x_0 is low, the denominator tends to be close to 0 and we get a division by zero problem.

(ii)At boundary points, the estimator has higher order bias, even with appropriately chosen $k()$. This problem is known as *boundary bias*.

3.2.2 Interpretation of Local Polynomial Estimators as a local regression

Solve this problem at each point of evaluation, x_0 (which may or may not correspond to points in the data):

$$\hat{a} : \underset{\{a,b_1,..,b_k\}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - a - b_1(x_i - x_0) - b_2(x_i - x_0)^2 - ... - b_k(x_i - x_0)^k)^2 k_i$$

$$k_i = k\left(\frac{x_0 - x_i}{h_n}\right)$$

$\hat{b}_1$ provides an estimator for $g'(x_0)$
$\hat{b}_2$ provides an estimator for $\frac{g''(x_0)}{2}$
if $k = 0$,

$$\hat{a} = \frac{\sum_{i=1}^{n} y_i k(\frac{x_0 - x_i}{h_n})}{\sum_{i=1}^{n} k(\frac{x_0 - x_i}{h_n})} \leftarrow \text{standard Nadaraya-Watson kernel regression estimator}$$

$$= \sum_{i=1}^{n} y_i w_i(x_0) \qquad \text{where } \sum_{i=1}^{n} w_i(x_0) = 1$$

if $k = 1$

$$\hat{a} = \frac{\sum_{i=1}^{n} y_i k_i \sum_{j=1}^{n} (x_j - x_0)^2 - \sum_{k=1}^{n} y_k k_k (x_k - x_0) \sum_{l=1}^{n} k_l (x_l - x_0)}{\sum_{i=1}^{n} k_i \sum_{j=1}^{n} k_j \left(x_j - x_0\right)^2 - \{\sum_{k=1}^{n} k_k (x_k - x_0)\}^2}$$

where

$$k_i = k(\frac{x_0 - x_i}{h_n})$$

We can again write the expression as $\sum_{i=1}^{n} y_i w_i(x_0)$

where the weights satisfy

$$\sum_{i=1}^{n} w_i(x_0) = 1$$

and

$$\sum_{i=1}^{n} w_i(x_0)(x_i - x_0) = 0$$

$k = 0$ weights for the standard Nadaraya-Watson kernel estimator do not satisfy this second property.

The LLR estimator (k=1) improves on kernel estimator in two ways

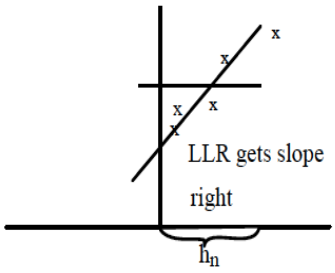
(a) the asymptotic bias does not depend on design density of the data (i.e. f(x) does not appear in the bias expression.)

(b) the bias has a higher order of convergence at boundary points

Intuitively, the standard kernel estimator fits a local constant, so when there is no data on the other side, the estimate will tend to be too high or too low. It does not capture the slope. LLR gets the slope.

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

Figure 3.2: Boundary bias - How LLR improves on kernel regression



$$
\begin{aligned}
\hat{g}_{LLR}(x_0) - g(x_0) &= \sum_{i=1}^{n} w_i(x_0) y_i - g(x_0) \\
&= \sum_{i=1}^{n} w_i(x_0)\varepsilon_i + \sum_{i=1}^{n} w_i(x_0)(g(x_i) - g(x_0)) \\
&= \sum_{i=1}^{n} w_i(x_0)\varepsilon_i + g'(x_0)\,(x_i - x_0) + [g'(\bar{x}) - g'(x_0)]\,(x_i - x_0) \\
&= \underbrace{\sum_{i=1}^{n} w_i(x_0)\varepsilon_i}_{variance\ part} + g'(x_0)\underbrace{\sum_{i=1}^{n} w_i(x_0)\,(x_i - x_0)}_{=0\ with\ LLR} + \\
&\quad + \sum_{i=1}^{n} w_i(x_0)\underbrace{(g'(\bar{x}) - g'(x_0))}\,(x_i - x_0)
\end{aligned}
$$

If g' satisfies a Lipschitz condition, the last term can be bounded.

3.3 Properties of Nonparametric Regression Estimators

1. Consistency (will show convergence in mean square)
2. Asymptotic Normality

In showing the asymptotic properties, we will work with the standard kernel regression estimator, because it is simpler to analyze than LLR and the basic techniques are the same.

$$\hat{m}(x_0) = \frac{\frac{1}{nh_n}\sum_i y_i k(\frac{x_0 - x_i}{h_n})}{\frac{1}{nh_n}\sum_i k(\frac{x_0 - x_i}{h_n})}$$

Assume

(i) $h_n \to 0,\ nh_n \to \infty$
(ii) $\int_{-\infty}^{\infty} sk(s)ds = 0 \quad \int_{-\infty}^{\infty} k(s)ds \neq 0 \quad \int_{-\infty}^{\infty} k(s)s^2 ds < C$
(iii) $f(x)$ *and* $g(x)$ *are* c^2 with bounded second derivatives

3.3.1 Consistency

We already showed

$$\frac{1}{nh_n}\sum_i k(\frac{x_0 - x_i}{h_n}) \overset{MS}{\to} f(x_0) \int_{-\infty}^{\infty} k(s)ds \qquad \text{(denominator)}$$

will now show

$$\frac{1}{nh_n}\sum_i y_i k(\frac{x_0 - x_i}{h_n}) \overset{MS}{\to} f(x_0)g(x_0) \int_{-\infty}^{\infty} k(s)ds \text{ (numerator)}$$

Then apply Mann.Wald theorem (plim of a continuous function is continuous function of plim) to conclude ratio$\overset{P}{\to} g(x_0)$

Denominator:

$$VAR(\frac{1}{nh_n}\sum_{i=1}^{n}k(\frac{x_0-x_i}{h_n})) \leq \frac{1}{n^2h^2}\cdot n\ E(k^2(\frac{x_0-x_i}{h_n}))$$
$$= \frac{1}{nh^2}\cdot O_p(h_n)$$
$$= O_p(\frac{1}{nh})$$

Thus,

$$\frac{1}{nh_n}\sum_i k(\frac{x_0-x_i}{h_n}) \overset{MS}{\to} E(\frac{1}{nh_n}\sum_i k(\frac{x_0-x_i}{h_n}))$$
$$= f(x_0)\int k(s)ds + O_p(h_n)$$

Numerator:

$$E(\frac{1}{nh_n}\sum_{i=1}^{n}y_ik(\frac{x_0-x_i}{h_n})) = \frac{1}{h_n}E(E(y_i|x_i)k(\frac{x_0-x_i}{h_n}))$$
$$= \frac{1}{h_n}E(g\,(x_i)\,k(\frac{x_0-x_i}{h_n}))$$
$$= \frac{1}{h_n}\int_{-\infty}^{\infty} g\,(x_i)\,k(\frac{x_0-x_i}{h_n})f(x_i)dx_i$$
$$\text{Let } \phi(x_i) = f\,(x_i)\,g(x_i) \text{ and Taylor expand (as before)}$$
$$= \phi\,(x_0)\int_{-\infty}^{\infty} k(s)ds + O_p(h_n)$$
$$\text{where } \phi(x_0) = g(x_0)f(x_0)$$

Also, can show

$$VAR(\frac{1}{nh_n}\sum_{i=1}^{n}y_ik\left(\frac{x_0-x_i}{h_n}\right) \leq \frac{c}{nh_n} \tag{3.1}$$
$$\frac{c}{nh_n} \to 0\ as\ nh_n \to \infty \tag{3.2}$$

Hence,

$$ratio \to \frac{g\left(x_0\right) f\left(x_0\right) \int k(s)ds}{f\left(x_0\right) \int k(s)ds} = g\left(x_0\right)$$

provided that
$h_n \to 0\ nh_n \to \infty$

3.3.2 Asymptotic Normality

Show

$$\sqrt{nh_n}(\hat{g}(x_0) - g(x_0))\tilde{}$$

$$N\left\{\left[\frac{1}{2}g''\left(x_0\right) + \frac{g'\left(x_0\right) f'\left(x_0\right)}{f\left(x_0\right)}\right] h_n^2 \sqrt{nh_n} \int_{-\infty}^{\infty} u^2 k(u)du,\, f(x_0)^{-1}\sigma^2(x_0) \int_{-\infty}^{\infty} k^2\left(u\right) du\right\}$$

where
$\sigma^2(x_0) = E\left(\varepsilon_i^2 | x_i = x_0\right) \leftarrow$ conditional variance

$$\sqrt{nh_n}\left[\frac{\sum y_i k_i}{\sum k_i} - g\left(x_0\right)\right] = \sqrt{nh_n}\left[\frac{\sum (y_i - g(x_0))k_i}{\sum k_i}\right] \tag{3.3}$$

$$\begin{aligned} y_i &= g(x_i) + \varepsilon_i \qquad\qquad E(\varepsilon_i | x_i) = 0 \\ &= \underbrace{\sqrt{nh_n}\frac{\sum \varepsilon_i k_i}{\sum k_i}}_{variance\ part} + \underbrace{\sqrt{nh_n}\frac{\sum (g(x_i) - g(x_0))k_i}{\sum k_i}}_{bias\ part} \\ &= \sqrt{nh_n}(\frac{\frac{1}{nh_n}\sum \varepsilon_i k_i}{\frac{1}{nh_n}\sum k_i}) + \sqrt{nh_n}(\frac{\frac{1}{nh_n}\sum (g(x_i) - g(x_0))k_i}{\frac{1}{nh_n}\sum k_i}) \end{aligned}$$

Assuming $\int_{-\infty}^{\infty} k(s)ds = 1$, we showed before that

$$\frac{1}{nh_n}\sum k_i \overset{MS}{\to} f(x_0) + O(h_n)$$

We will now analyze the distribution of the numerator

$$\sqrt{nh_n}\frac{1}{nh_n}\sum \varepsilon_i k_i = \frac{1}{\sqrt{nh_n}}\sum_i \varepsilon_i k(\frac{x_0 - x_i}{h_n}) = \frac{1}{\sqrt{n}}\sum_i \frac{1}{\sqrt{h_n}}\varepsilon_i k(\frac{x_0 - x_i}{h_n})$$

and apply CLT to get

$$\sim N\left(0,\ VAR\left(\frac{1}{\sqrt{h_n}}\varepsilon_i k(\frac{x_0 - x_i}{h_n})\right)\right) \tag{3.4}$$

$$\begin{aligned} VAR\left(\frac{1}{\sqrt{h_n}}\varepsilon_i k(\frac{x_0 - x_i}{h_n})\right) &= \frac{1}{h_n}E\left(\varepsilon_i^2 k^2(\frac{x_0 - x_i}{h_n})\right) + \text{higher order terms} \\ &= \frac{1}{h_n}E\big(\underbrace{E(\varepsilon_i^2|x_i)}_{=\sigma^2(x_i)}k^2(\frac{x_0 - x_i}{h_n})\big) \\ &= \frac{1}{h_n}\int_{-\infty}^{\infty}\sigma^2(x_i)k^2(\frac{x_0 - x_i}{h_n})f(x_i)dx_i \\ &= \sigma^2(x_0)f(x_0)\int k^2(s)ds + O(h_n) \quad \text{by change of var} \end{aligned}$$

Now, analyze the bias term

$$\sqrt{nh_n}\frac{1}{nh_n}\sum(g(x_i)-g(x_0))k_i = \frac{1}{\sqrt{nh_n}}\sum_i(g(x_i)-g(x_0))k(\frac{x_0-x_i}{h_n})$$

$$\longrightarrow \frac{1}{\sqrt{nh_n}}\cdot n\int_{-\infty}^{\infty}(g(x_i)-g(x_0))k(\frac{x_0-x_i}{h_n})f(x_i)dx_i$$

$$= \sqrt{nh_n}\int_{-\infty}^{\infty}(g(x_0-sh_n)-g(x_0))\cdot k(s)\cdot f(x_0-sh_n)ds$$

$$= \sqrt{nh_n}\int_{-\infty}^{\infty}\{g'(x_0)\,(-sh_n)+\frac{1}{2}g''(x_0)s^2h_n^2+\frac{1}{2}[g''(\bar{x})-g''(x_0)]s^2h_n^2\}$$

$$\cdot\{f(x_0)-sh_nf'(x_0)+\frac{s^2h_n^2}{2}f''(\bar{x})\}k(s)ds$$

+higher order (will ignore)

–if g'' satisfies lipschitz condition

$$|g''(\bar{x})-g''(x_0)| \le C\,|\bar{x}-x_0|$$

then $\frac{[g''(\bar{x})-g''(x_0)]}{2}$term will be higher order -other terms

$$= -g'f\overbrace{\int sk(s)ds}^{=0}+g'f'h_n^2\int s^2k(s)ds-\frac{g'f''(\bar{x})}{2}\underbrace{h_n^3}_{higher\ order}\int k(s)s^3ds$$

$$+\frac{1}{2}g''(x_0)f(x_0)\cdot h_n^2\int_{-\infty}^{\infty}k(s)s^2ds+higher\ order\ terms$$

get

$$\sqrt{nh_n}\sum(g(x_i)-g(x_0))k_i \to \sqrt{nh_n}\,[g'(x_0)f'(x_0)+\frac{1}{2}g''(x_0)f(x_0)]\cdot h_n^2\int_{-\infty}^{\infty}k(s)s^2ds$$

require

$$\sqrt{nh_n}h_n^2 \to 0 \quad i.e.\ nh_n^5 \to 0$$

Putting these results together, we get

$$\sqrt{nh_n}(\hat{g}(x_0)-g(x_0)) \sim N\left([\frac{g'(x_0)f^{'}(x_0)}{f(x)}+\frac{1}{2}g''(x_0)]\sqrt{nh_n}h_n^2\int k(s)s^2ds, f(x_0)^{-1}\sigma^2(x_0)\int k(s)^2ds\right)$$

(which is what we set out to show)

Remarks:

Getting asymptotic distribution requires plug-in estimator of $g', f', f, g'', \sigma^2(x_0)$
(all the ingredients of the bias and variance)

g'-obtain by local linear or local quadratic regression
f'-derivative of estimator for f gives estimator of f'

$$\hat{f}'(x_0) = \frac{1}{nh_n}\sum_{i=1}^{n}k'(\frac{x_0-x_i}{h_n})$$

f–obtain by standard density estimator
$\int_{-\infty}^{\infty}k^2(s)ds,\ \int_{-\infty}^{\infty}k(s)sds$–constants that depend on kernel function $k(s)$
$\sigma^2(x_0)$–conditional variance. Estimate by

$$\hat{E}(\varepsilon_i^2|x_0) = \frac{\sum w_i(x_0)\hat{\varepsilon}_i^2}{\sum w_i(x_0)} \leftarrow fitted\ residuals\ \hat{y}-\hat{m}(x_i)=\hat{\varepsilon}_i$$

We showed distribution of kernel estimator. What about distribution of LLR? Fan (JASA, 1992) showed

$$\sqrt{nh_n}(\hat{g}_{LLR}(x_0)-g(x_0)) = N\left[\underbrace{\frac{1}{2}g''(x_0)\int k(s)s^2ds\cdot\sqrt{nh_n}h_n^2}_{\text{one less term in bias expression}},\ \underbrace{f(x_0)^{-1}\sigma^2(x_0)\int k^2(s)ds}_{\text{same variance}}\right]$$

Remarks:

–Bias of LLR does not depend on $f(x_0)$ (the design density of the data). Because of this feature, Fan refers to the estimator as being "Design-Adaptive"

–Fan showed that in general, better to use odd-order local polynomial. There is no cost in variance, bias of odd-order does not depend on $f(x)$.

–LLR estimator does not suffer from boundary bias problem

3.3.3 How might we choose the bandwidth?

Could choose the bandwidth to minimize the pointwise asymptotic MSE(AMSE)

$$AMSE_{LLR} = [h_n^2 \tfrac{1}{2} g''(x_0) \int k(s)s^2 ds]^2 + \tfrac{f(x_0)^{-1}\sigma^2(x_0)}{nh_n} \int k^2(s)ds$$

$$\tfrac{\partial}{h_n} = 4h_n^3 \tfrac{1}{4} g''(x_0)^2 [\int k(s)s^2 ds]^2 + -\tfrac{f(x_0)^{-1}\sigma^2(x_0)}{nh_n^2} \int k^2(s)ds = 0$$

$$\Rightarrow h_n = \left[\tfrac{f(x_0)^{-1}\sigma^2(x_0)}{g''(x_0)^2} \cdot \underbrace{\frac{\int k^2(s)ds}{[\int k(s)s^2 ds]^2}}_{\text{constant}} \right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

Remarks:

(1) higher variance $\sigma^2(x_0)$ $\Rightarrow$wider bandwidth

(2) more variability in function ($g''(x_0)$) $\Rightarrow$wider bandwidth

(3) more data$\Rightarrow$narrower bandwidth (with kernel regression (i.e. local polynomial of degree 0), bias term also depends on $f(x)$ and therefore optimal bandwidth choice will depend on $f(x)$)

(4) difficulty–obtaining optimal bandwidth requires estimates of $\sigma^2(x_0)$, $g''(x_0)$ (and in the case of the usual regression estimator $f(x_0)$. This is the problem of how to choose "pilot bandwidth." One could assume normality in choosing a pilot for $f(x_0)$ (this is Silverman's rule-of-thumb method). Could also use fixed bw or nearest neighbor for $\sigma^2(x_0)$, $g''(x_0)$

Figure 3.3: Functions of different smoothness

Figure 4: functions of different smoothness

(5)Could minimize $AMISE = E\int(\hat{g}(x_0) - g(x_0))dx_0$ and pick a global bandwidth

For further information on choosing the bandwidth, see below and see book by Fan and Gijbels (1996).

Chapter 4

Bandwidth Selection for Density Estimation

4.1 First Generation Methods

4.1.1 Plug-in method (local version)

One could derive the optimal bandwidth at each point of evaluation, x_0, which would lead to a pointwise localized bandwidth estimator

$$\begin{aligned}\hat{h}(x_0) &= \arg\min\ AMSE\ \hat{f}(x_0) = E\left(\hat{f}\left(x_0\right) - f\left(x_0\right)\right)^2 \\ &= \frac{h^4}{4} f''(x_0)^2 \left[\int k(s) s^2 ds\right]^2 + \frac{1}{nh} f(x_0) \int k^2(s) ds \\ \Rightarrow \hat{h}(x_0) &= \left[\frac{\int k^2(s) ds}{\left[\int k(s) s^2 ds\right]^2} \cdot \frac{f(x_0)}{\left(f''(x_0)\right)^2}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}\end{aligned}$$

We need to estimate $f\left(x_0\right), f''\left(x_0\right)$ to get optimal plug-in bandwidth, and the optimal bandwidth would be different for different points of estimation.

4.1.2 Global plug-in method

Because the above method is computationally intensive, one could instead derive a optimal bandwidth for all the points of evaluation by minimizing a different criterion - the mean integrate squared error:

$$MISE\ \hat{f}(x_0) = E\int \left(\hat{f}(x_0) - f(x_0)\right)^2 dx_0 \quad \leftarrow \text{removes } x_0 \text{ by integration}$$
$$= \frac{1}{nh}\underbrace{\int f(x_0)\,dx_0}_{=1}\underbrace{\int k^2(s)ds}_{A} + \frac{h^4}{4}(\underbrace{\int k(s)s^2ds)^2(\int f''(x_0)^2dx_0)}_{B}$$

$$h^4_{MISE} = (\frac{A}{B^2})^{\frac{1}{5}}n^{-\frac{1}{5}}$$

note: $\int f''(x_0)^2dx_0$ is a measure of the variation in $f(x_0)$. If f is highly variable then h will be small. Here, we still need estimate of $f''(x_0)$ but no longer require an estimate of $f(x_0)$ (it was eliminated by integrating)

4.1.3 Rule-of-thumb Method (Silverman)

If

$$x \sim\ Normal$$

then Silverman shows (in his book *Density Estimation*)

$$h^4_{MISE} \approx n^{-\frac{1}{5}}SD(x) \qquad\qquad SD = \text{standard deviation}$$

This method if often applied to nonnormal data, but it has a tendency to "oversmooth", particularly when data density is multimodel. Sometimes, the standard deviation is replaced by the 75%-25% interquartile range.

This is an older method, now available in some software packages, but the performance is not so good. Other methods are better.

4.1.4 Global Method #2: Least Squares Cross-validation

The idea is to minimize an estimate of the integrated squared error (ISE)

$$\min_h ISE = \int \{\hat{f}(x_0) - f(x_0)\}^2 dx_0$$
$$= \underbrace{\int \hat{f}(x_0)^2dx_0}_{term\ \#1} - \underbrace{2\int \hat{f}(x_0)f(x_0)dx_0}_{term\ \#2} + \underbrace{\int f(x_0)^2dx_0}_{\text{term doesnt depend on h}}$$

$$
\begin{aligned}
Term\ \#1 &= \int \frac{1}{n^2h^2}\sum_i\sum_j k(\frac{x_i - x_0}{h_n})k(\frac{x_j - x_0}{h_n}) \\
let\ -s &= \frac{x_i - x_0}{h_n} \Rightarrow x_0 = x_j + sh \quad \frac{x_0}{h} = \frac{x_j}{h} + s \\
&= \frac{1}{n^2h^2}\sum_i\sum_j \int k(\underbrace{\frac{x_i}{h} - \frac{x_j}{h}}_{\frac{x_i - x_0}{h_n}} - s)k(s)ds \\
&= \frac{1}{n^2h}\sum_i\sum_j \int k(\frac{x_i - x_j}{h} - s)k(s)ds
\end{aligned}
$$

Note: $\int k(a-s)k(s)ds$ is the density of the sum of two random variables with density k (also known as the "convolution")

$$
\begin{aligned}
y &= a + s \\
\Pr(y \le y_0) &= \Pr(a + s \le y_0) \\
&= \Pr(a \le y_0 - s) \\
&= \int_{-\infty}^{\infty} F(y_0 - s)k(s)ds \\
\frac{\partial}{\partial y} &= \int_{-\infty}^{\infty} k(y_0 - s)k(s)ds
\end{aligned}
$$

Now examine Term #2:

$-2\int \hat{f}(x_0)f(x_0)dx_0$

An unbiased estimator for term (2) is

$$
\frac{-2}{N(N-1)}\sum_i\sum_{j\neq i} k(\frac{x_i - x_j}{h}) \approx \frac{-2}{N^2h}\sum_i\sum_{j\neq i} k(\frac{x_i - x_j}{h})
$$

Note that we need to use "leave-one-out" estimator to get independence.

$$
\hat{h}_{CV} = \arg\min \underbrace{\frac{1}{n^2h}\sum_i\sum_j \int k(\frac{x_i - x_j}{h} - s)k(s)ds}_{\text{need to calculate convolution for this part}} - \frac{2}{n^2h} \underbrace{\sum_i\sum_{j\neq i} k(\frac{x_i - x_j}{h})}_{\text{ith point does not get used, } = \hat{f}_{-i}(x_i)}
$$

$$
\underbrace{\hat{f}_{-i}(x_i)}_{\text{leave one out estimator}} = \hat{f}(x_i) - K(o)\frac{1}{nh}
$$

(1) turns out that $\hat{h}_{CV}$ can only be estimated at rate $n^{-\frac{1}{10}}$(result due to Stone) which is *very* slow. However, in Monte Carlo studies LSCV often does better that Rule-of-thumb method.

(2) Marson (1990) found a tendency for local minima near small h areas, which lead to a tendency to undersmooth.

(3) LSCV is usually implemented by searching over a grid of bandwidth values

4.1.5 Biased Cross-validation

$$AMISE = \frac{1}{nh}\int k^2(s)ds + \frac{h_n^4}{4}\int \Gamma^2 k(s)ds \int f''(x)^2 dx + o(\frac{1}{nh} + h^4)$$

what if we plug in $\hat{f}''(x)$?
note:

$$\begin{aligned}
\hat{f}(x) &= \frac{1}{nh}\sum_i k(\frac{x_i - x}{h}) \\
\hat{f}'(x) &= \frac{1}{nh^2}\sum_i k'(\frac{x_i - x}{h}) \\
\hat{f}''(x) &= \frac{1}{nh^3}\sum_i k''(\frac{x_i - x}{h}) \\
\hat{f}''(x)^2 &= \frac{1}{n^2h^6}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\frac{x_i - x}{h})k(\frac{x_j - x}{h})
\end{aligned}$$

What is bias in estimation $b''(x)^2$?

$$\begin{aligned}
E\{\hat{f}''(x)^2\} &= \int E(\frac{1}{n^2h^6}\sum_i k''(\frac{x_i - x}{h})\sum_j k''(\frac{x_j - x}{h_n}))dx \\
&= \frac{1}{nh^6}\int E''(\frac{x_i - x}{h})^2 dx \\
\text{can show } &= f''(x) + \frac{1}{nh^5}\int k''(s)^2 d + \text{ higher order}
\end{aligned}$$

$$h_{BCV} = \arg\min \frac{1}{nh}\int k^2(s)ds + \frac{h^4}{4}\int s^2k(s)ds\Big[\int \hat{f}''(x)^2\,dx - \underbrace{\frac{1}{nh^5}\int k''(s)ds}_{\text{subtract this bias here}}\Big]$$

4.1.6 Second Generation Methods

Sheather and Jones(1991) "Solve -the -equation" Plug-in Method

$$h_{SIPI} = \left[\frac{\int k^2(s)ds}{\int f''_{g(h)}(x)^2dx\left[\int s^2k(s)ds\right]^2}\right]^{\frac{1}{5}} n^{-\frac{1}{5}} \qquad \leftarrow \text{find } h \text{ that solves this}$$

We need to plug in an estimate for $f''(x)$. The bandwidth that is optimal for $f(x)$ is not same as optimal bandwidth for f''

We can find an analogue to $AMSE$ for $R(f'') = \int f''(x)^2dx$

get

$$g = \left[C_1(k)\{R(f''')\}^{\frac{1}{7}}C_2(k)\right] n^{-\frac{1}{7}}$$

$R(f'') = \int f''(x)dx$ enters into the formula for the optimal bandwidth for g

$C_1(k)$ and $C_2(k)$ are functions of the kernel. $R(f''')$ is $\int f'''(x)^2dx$ Now solve for g as a function of h

$$g(h) = \left[C_3(k)\{\frac{R(f'')}{R(f''')}\}^{\frac{1}{7}}C_4(k)\right] h^{\frac{5}{7}}$$

–Now we need an estimate of f'''. At this point, SJ suggest using a rule of thumb method, based on normal density assumption What is main advantage

of this method over other methods?
For most methods

$$\frac{\hat{h} - h_{MISE}}{h_{MISE}} \sim n^{-p}$$

Recall, for CV, $p = \frac{1}{10}$. for SJPI,

$$\rho = \frac{5}{14} \qquad \text{(close to } \frac{1}{2}\text{)}$$

so rate of convergence of the bandwidth is faster. This method is available in some software packages.

Chapter 5

Bandwidth selection for Nonparametric Regression

Recall that the distribution for the LLR estimator was

$$\sqrt{nh_n}(\hat{g}_{LLR}(x_0)-g(x_0)) \sim N\left(\frac{1}{2}g''(x_0)\int k^2(s)sds \cdot \sqrt{nh_n}\cdot h_n^2,\ f(x_0)^{-1}\sigma^2(x_0)\int k^2(s)\,ds\right)$$

5.1 Plug-in Estimator (local method)

–choose bandwidth to minimize the MSE "data-dependent"

$$\min_{h_n} \underbrace{\left[\frac{1}{2}g''(x_0)\right]^2\left[\int k(s)s^2ds\right]^2 h_n^4}_{Bias^2=C_0h_n^4} + \underbrace{\frac{f(x_0)^{-1}\sigma^2(x_0)}{nh_n}\int k^2(S)ds}_{\text{Variance}=C_1n^{-1}h_n^{-1}}$$

$$h_n^* = \left[\frac{f(x_0)^{-1}\sigma^2(x_0)\int k^2(s)ds}{g''(x_0)^2\int k(s)s^2ds}\right]^{\frac{1}{5}} n^{-\frac{1}{5}} \qquad h_n^* : h_n^*(x_0) \text{ variable bandwidth}$$

note:

$$\min C_0h_n^4 + C_1n^{-1}h_n^{-1}$$

$$4C_0h_n^3 - C_1n^{-1}h_n^{-2} = 0$$

$$h_n^5 = \frac{C_1}{4C_0}n^{-1}$$

$$h_n = \left(\frac{C_1}{4C_0}\right)^{-\frac{1}{5}} n^{-\frac{1}{5}}$$

Remarks: g''high $\Rightarrow$smaller bandwidth
want smaller bandwidth in regions where function is highly variable
$\sigma^2(x_0)$ high$\Rightarrow$use a larger bandwidth

5.2 global Plug-in Estimator

–minimize global criterion s.a.MISE

$$h_n^r = \left\{ \frac{\int \sigma^2(x_0)dx_0 \int k^2(s)ds}{\int g''(x_0)^2 dx_0 \int k(s)s^2 ds} \right\}^{\frac{1}{5}} n^{-\frac{1}{5}}$$

Mf: Fan &Gijbels (1996) Local Polynomial Reg

5.3 Bandwidth Choice in Nonparametric Regression

$y_i = m(x_i) + \varepsilon_i$
$\min \sum_{i=1}^{n} \{y_i - \alpha - ... - \beta_p(x_i - x)^p\} k(\frac{x_i - x}{h_n})$ $\leftarrow$ how to choose h_n , given p

5.4 Global MISE criterion

$$h_{MISE} = \left[\frac{(p+1)p!^2 R(k_p) \overbrace{\int \sigma^2(x)dx}^{VAR(y|x)}}{\partial \mu_{p+1}(k_p)^2 \int m^{p+1}(x)^2 f(x)dx} \right]^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}$$

for LLR, $p = 1$ and get $n^{-\frac{1}{5}}$, for p odd

Fan & Gijbels (1996) Local Polynomial Regression

$k_p =$ function of kernel
$R(k_p) = \int k_p(s)^2 ds$
$\mu_l(k_p) = \int \mu^l k_p(\mu) d\mu$
$= \left[\frac{R(k) \int \sigma^2(x)dx}{\mu_2(k)^2 \int m''(x)^2 f(x)dx} \right]^{\frac{1}{5}} n^{-\frac{1}{5}}$

unknowns are $\sigma^2(x), m''(x), f(x)$

Apply same plug-in method to estimate $\hat{m}''(x)$ then take average
Still need $\int \sigma^2(x)dx$

5.5 blocking method

(1) divide x into N blocks
(2) Estimate $y = m(x) + \varepsilon$ by Q-degree poly within each block

$$\sigma^2(N) = (n - sN)^{-1} \sum_i \sum_j \left\{ y_i - \hat{m}_j^Q(x_i) \right\} 1(x_i \in X_j)$$

5.6 Bandwidth Selection for Nonparametric Regression- Cross-Validation

Recall LLR estimator defined as

$$\hat{a} = \arg\min \sum_{i=1}^{n} (y_i - a - b(x_i - x_0))^2 k(\frac{x_i - x_0}{h_n})$$

Let

$$y = \begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix} \quad w = \begin{pmatrix} k\left(\frac{x_1 - x_0}{h_n}\right) & & 0 \\ & ... & \\ 0 & & k\left(\frac{x_N - x_0}{h_n}\right) \end{pmatrix} \leftarrow N \times N$$

$$x = \begin{pmatrix} 1 & (x_1 - x_0) \\ ... & \\ 1 & (x_N - x_0) \end{pmatrix} \leftarrow N \times Z$$

$\hat{g}_{LLR}(x_0) = [1 \quad 0](x'wx)^{-1}(x'wy)$ $\leftarrow$ coefficient from a weighted least squares problem.
weights depend on x_0, h

Let

$$H(h) = \underset{1\times 2}{[1 \quad 0]}\underset{2\times 2}{(x'wx)^{-1}}\underset{2\times N}{x'}\underset{N\times N}{w} \quad 1\times N$$
$$\hat{g}_{LLR}(x_0) = H(h)\cdot y$$

choose h to $\min E((\hat{g}_{LLR}(x_0) - g(x))^2|x)$
minimize conditional variance at each point of evaluation

$$\begin{aligned}
&= E((H(h)\,y - g)'(H(h)y - g)|x) \\
&= E((H(h)\,g + H(h)\,\varepsilon - g)'((H(h)\,g + H(h)\,\varepsilon - g)|x) \\
&= E(((H(h) - I)g + H(h)\,\varepsilon)'((H(h) - I)g + H(h)\,\varepsilon)|x) \\
&= g'(H(h) - I)'(H(h) - I)g + E(\underset{1\times N}{\varepsilon'}\underset{N\times 1}{H(h)'}\underset{1\times N}{H(h)}\underset{N\times 1}{\epsilon}|x) && \text{recall } trAB = trBA \\
&= g'(H(h) - I)'(H(h) - I)g + E(tr(\varepsilon'H(h)'H(h)\epsilon|x) && tr(A+B) = tr(A) + tr(B) \\
&= g'(H(h) - I)'(H(h) - I)g + E(tr(\varepsilon\varepsilon')\cdot tr(H(h)'H(h))|x && \text{trace is a linear mapping} \\
&= g'(H(h) - I)'(H(h) - I)g + n\sigma_\varepsilon^2 tr(H(h)'H(h)) && \text{need estimator for this}
\end{aligned}$$

Consider what is estimated by

$$\begin{aligned}
(y - H(h)y)'(y - H(h)y) &= y'(I - H(h))'(I - H(h))y \quad \leftarrow \text{sum of fitted residuals} \\
&= (g+\epsilon)'(I - H(h))'(I - H(h))(g+\varepsilon) \\
&= \left\{ \begin{array}{l} +g'(I - H(h))'(I - H(h))g + g'(I - H(h))'(I - H(h))\varepsilon \\ +\varepsilon'(I - H(h))'(I - H(h))g + \varepsilon'(I - H(h))'(I - H(h))\varepsilon \end{array} \right\}
\end{aligned}$$

$$\begin{aligned}
E(g'(I - H(h))'(I - H(h))g|x) &= g'(I - H(h))'(I - H(h))g \\
E(g'(I - H(h))'(I - H(h))\varepsilon|x) &= 0 \\
E(\varepsilon'(I - H(h))'(I - H(h))\varepsilon|x) &= E(tr(\varepsilon'(I - H(h))'(I - H(h))\varepsilon)|x) \\
&= E(tr(\varepsilon\varepsilon')\; tr\left((I - H(h))'(I - H(h))\right)|x) \\
&= n\sigma^2 tr\left((I - H(h))'(I - H(h))\right) \\
&= n\sigma^2\left[trI - 2tr(H(h))trI + trH(h)'H(h)\right]
\end{aligned}$$

together,

$$\begin{aligned}
&(y - H(h)y)'(y - H(h)y) \\
&= g'(I - H(h))'(I - H(h))g + n\sigma^2 trI - \underbrace{2n\sigma^2 trH(h)\,thI}_{\text{don't want this term}} + n\sigma^2 trH(h)'H(h)
\end{aligned}$$

if we use $\hat{g}_{-i}(x_i)$ instead of $\hat{g}(x_i)$ (leave-one-out estimator)

then

$$trH(h) = 0$$
$$\hat{h}_{CV} = \underset{h \leftarrow H}{\arg\min} \sum_{i=1}^{n} (y_i - \hat{g}_{-i}(x_i))^2$$