

Nonparametric Methods

Michael R. Roberts

Department of Finance
The Wharton School
University of Pennsylvania

July 28, 2009

Overview

- Great for data analysis and robustness tests.
- Also used extensively in program evaluation
 - 1 Estimation of propensity scores
 - 2 Estimation of conditional regression functions
- Goal here is to introduce and operationalize nonparametric
 - 1 density estimation, and
 - 2 regression

Probability Density Functions (PDF)

- Basic characteristics of a random variable X is its PDF, f or CDF, F
- Given a sample of observations $X_i : i = 1, \dots, N$, goal is to estimate the PDF
- Options
 - 1 Parametric: Assume a functional form for f and estimate the parameters of the function. E.g., $N(\mu, \sigma^2)$
 - 2 Nonparametric: Estimate the full function, f , without assuming a particular functional form for f .
- Nonparametric “let the data speak.”
- We’re going to follow Silverman (1986) closely.

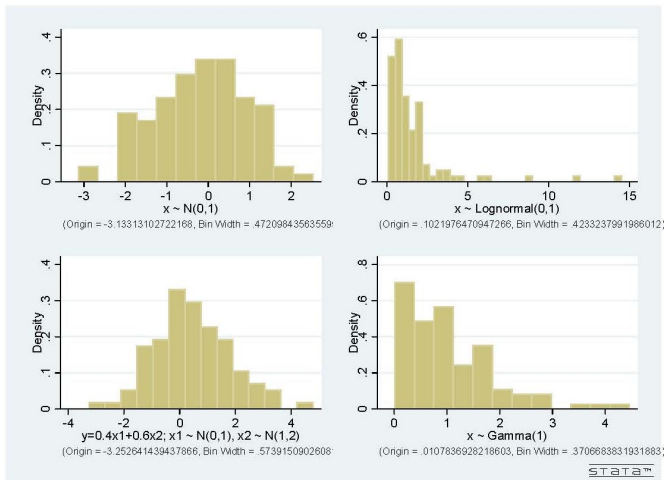
Histogram

- **Origin:** x_0
- **Bin Width:** h (a.k.a. **window width**)
- **Bins:** $[x_0 + mh, x_0 + (m + 1)h)$ for $m \in \mathbb{Z}$
- **Histogram:**

$$\hat{f}(x) = \frac{1}{nh} (\# \text{ of } X_i \text{ in the same bin as } x)$$

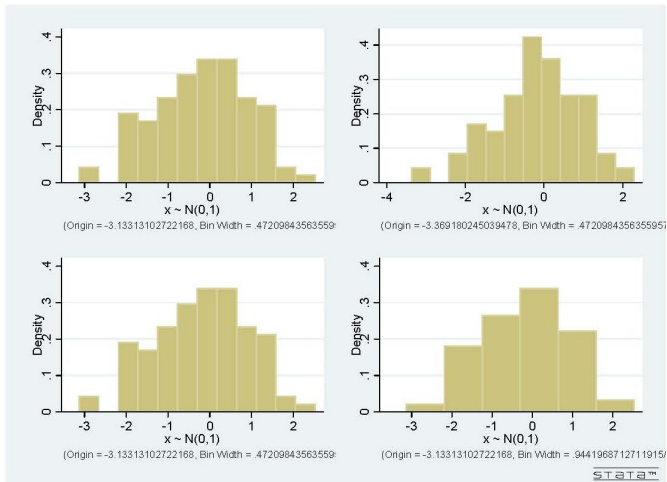
Sample Histograms

- $N = 100$, Origin = $\text{Min}(X_i)$, Bin Width = $0.79 \times \text{IQR} \times N^{1/5}$



Sensitivity of Histograms

- Histogram estimate is sensitive to choice of origin and bin width



Naive Estimator

- The density, f , of rv X can be written

$$f(x) = \lim_{h \rightarrow 0} \frac{1}{2h} \Pr(x - h < X < x + h)$$

- Given h , we can estimate $\Pr(x - h < X < x + h)$ by the proportion of observations falling in the interval (bin)

$$\hat{f}(x) = \frac{1}{2nh} [\# \text{ of } X_i \text{ falling in } (x - h, x + h)]$$

- Mathematically, this is just

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^N \frac{1}{h} W\left(\frac{x - X_i}{h}\right)$$

where

$$W(x) = \begin{cases} 1/2 & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Naive Estimator - An Example

- Consider a sample $\{X_i\}_{i=1}^{10}$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

- Let the bin width = 2, then

$$\begin{aligned}\hat{f}(4) &= \frac{1}{10} \left\{ \frac{1}{2} W\left(\frac{4-1}{2}\right) + \frac{1}{2} W\left(\frac{4-2}{2}\right) + \dots + \frac{1}{2} W\left(\frac{4-10}{2}\right) \right\} \\ &= \frac{1}{10} \left\{ 0 + 0 + \left(\frac{1}{2} \frac{1}{2}\right) + \left(\frac{1}{2} \frac{1}{2}\right) + \left(\frac{1}{2} \frac{1}{2}\right) + 0 + \dots + 0 \right\} \\ &= \frac{3}{40}\end{aligned}$$

Naive Estimator - An Example from Silverman

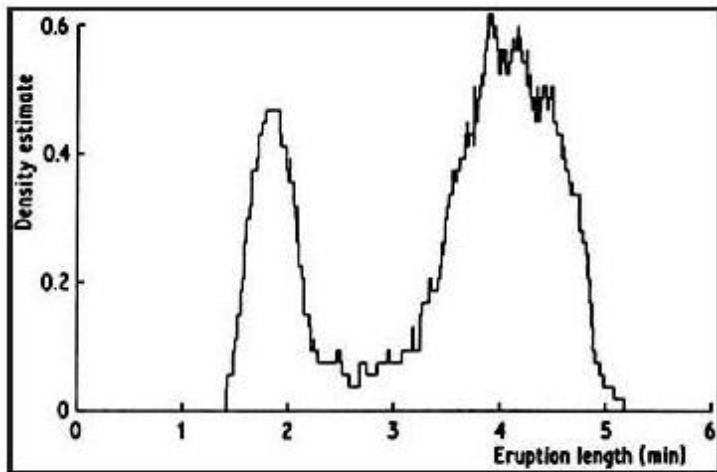


Fig. 2.3 Naive estimate constructed from Old Faithful geyser data, $h = 0.25$.

Naive Estimator - Discussion

- From def of $W(x)$, estimate of f is constructed by placing box of width $2h$ and height $(2nh)^{-1}$ on each observation and summing.
- Attempt to construct histogram where every point, x , is the center of a sampling interval $(x + h, x - h)$
- We don't need a choice of origin, x_0 , anymore
- Choice of bin width, h , remains and is crucial for controlling degree of smoothing
 - Large h produce smoother estimates
 - Small h produce more jagged estimates
- Drawbacks: \hat{f} is discontinuous, jumps at points $X + i \pm h$ and zero derivative everywhere else

Definition & Intuition

- Replace weight fcn W in naive estimator by a **Kernel Function** K :

$$\int_{-\infty}^{\infty} K(x) dx = 1$$

- Kernel estimator is:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right)$$

where h is **window width** or **smoothing parameter** or **bandwidth**

- Intuition:
 - Naive estimator is a sum of boxes centered at observations
 - Kernel estimator is a sum of bumps centered at observations

Kernel choice determines shape of bumps

Kernel Estimator - Example

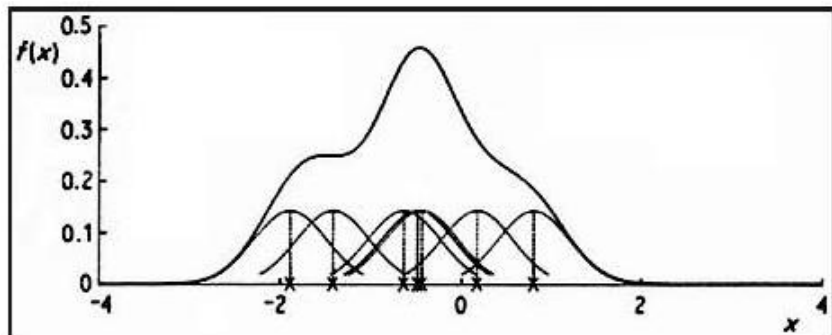


Fig. 2.4 Kernel estimate showing individual kernels. Window width 0.4.

Varying the Window Width

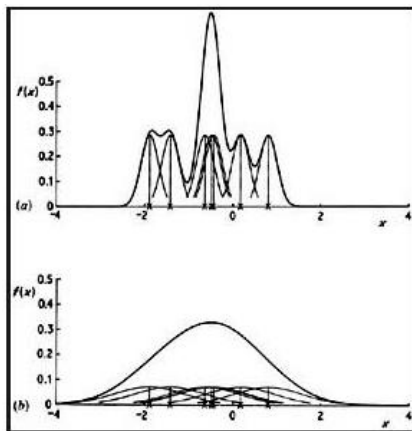


Fig. 2.5 Kernel estimates showing individual kernels. Window widths: (a) 0.2; (b) 0.8.

Example Discussion

- X 's correspond to data points (the sample: $N = 7$)
- Centered over each data point, is a little curve — bump — $1/(nh)K[(x - X_i)/h]$
- The estimated density, \hat{f} , constructed by adding up each bump at each data point is also shown
- As $h \rightarrow 0$ we get a sum of Dirac delta function spikes at the observations
- If K is a PDF, then so is \hat{f}
- \hat{f} inherits the continuity and differentiability properties of K
- For data with long-tails, get spurious noise to appear in the tails since window width is fixed across entire sample
 - If window width widened to smooth away tail detail, detail in main part of dist is lost
 - adaptive methods address this problem

Long Tail Data

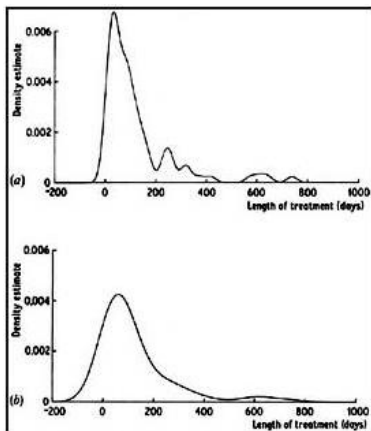


Fig. 2.9 Kernel estimates for suicide study data. Window widths: (a) 20; (b) 60.

Sample Kernels: Definitions

$$\text{Rectangular (Uniform)} : K(t) = \begin{cases} \frac{1}{2} & |t| < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Triangular} : K(t) = \begin{cases} 1 - |t| & |t| < 1 \\ 0 & \text{otherwise} \end{cases}$$

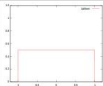
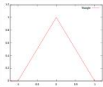

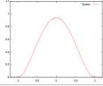
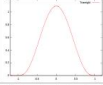
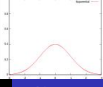
$$\text{Epanechnikov} : K(t) = \begin{cases} \frac{3}{4} (1 - \frac{1}{5}t^2) / \sqrt{5} & |t| < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Biweight (Quartic)} : K(t) = \begin{cases} \frac{15}{16} (1 - t^2)^2 & |t| < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Triweight} : K(t) = \begin{cases} \frac{35}{32} (1 - t^2)^3 & |t| < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Gaussian} : K(t) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)t^2}$$

Sample Kernels - Figures

Kernel Functions, $K(u)$			$\int u^2 K(u) du$	$\int K^2(u) du$
Uniform	$K(u) = \frac{1}{2} \mathbf{1}_{\{ u \leq 1\}}$		$\frac{1}{3}$	$\frac{1}{2}$
Triangular	$K(u) = (1 - u) \mathbf{1}_{\{ u \leq 1\}}$		$\frac{1}{6}$	$\frac{2}{3}$
Epanechnikov	$K(u) = \frac{3}{4} (1 - u^2) \mathbf{1}_{\{ u \leq 1\}}$		$\frac{1}{5}$	$\frac{3}{5}$
Quartic	$K(u) = \frac{15}{16} (1 - u^2)^2 \mathbf{1}_{\{ u \leq 1\}}$		$\frac{1}{7}$	$\frac{5}{7}$
Triweight	$K(u) = \frac{35}{32} (1 - u^2)^3 \mathbf{1}_{\{ u \leq 1\}}$		$\frac{1}{9}$	$\frac{350}{429}$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$		1	$\frac{1}{2\sqrt{\pi}}$

Measures of Discrepancy

- Mean Square Error (Pointwise Accuracy)

$$\begin{aligned}MSE_x(\hat{f}) &= E[\hat{f}(x) - f(x)]^2 \\ &= \underbrace{[E\hat{f}(x) - f(x)]^2}_{\text{Bias}} + \underbrace{\text{Var}\hat{f}(x)}_{\text{Variance}}\end{aligned}$$

- Tradeoff: Bias can be reduced at expense of increased variance by adjusting the amount of smoothing
- Mean Integrated Square Error (Global Accuracy)

$$\begin{aligned}MISE_x(\hat{f}) &= E \int [\hat{f}(x) - f(x)]^2 dx \\ &= \underbrace{\int [E\hat{f}(x) - f(x)]^2 dx}_{\text{Integrated Bias}} + \underbrace{\int \text{Var}\hat{f}(x) dx}_{\text{Integrated Variance}}\end{aligned}$$

Useful Facts

- The bias is *not* a fxn of sample size
 - ⇒ Increasing sample size will not reduce bias
 - ∴ Need to adjust the weight fxn (i.e., Kernel)
- Bias is a fxn of window width (and Kernel)
 - ⇒ Decreasing window width reduces bias
 - If window width fxn of sample size, then bias

Choosing the Smoothing Parameter

- Optimal window width derived as minimizer of (approximate) MISE is a fxn of the unknown density f
- Appropriate choice of smooth parameter depends on the goal of the density estimation
 - 1 If goal is data exploration to guide models and hypotheses, subjective criteria probably ok (see below)
 - 2 When drawing conclusions from estimated density, undersmoothing is probably good idea (easier to smooth than unsmooth a picture)

Reference to a Standard Distribution

- Use a standard family of distributions to assign a value to unknown density in optimal window width computation.
- E.g., assume f normal with $Var = \sigma^2$ and Gaussian kernel \implies

$$h^* = 1.06\sigma n^{-1/5}$$

- Can estimate σ from the data using SD
- If pop dist is multimodal or heavily skewed, h^* will oversmooth

Robust Measures of Spread

- Can use robust measure of spread ($R = \text{IQR}$) to get different optimal smoothing parameter

$$h^* = 0.79Rn^{-1/5}$$

but this exacerbates problems from multimodality/skew because it oversmooths

- Can try

$$h^* = 1.06An^{-1/5} \text{ or } h^* = 0.9An^{-1/5} \text{ or}$$

where

$$A = \min(SD, IQR/1.34)$$

Setup

- The basic problem is to estimate a function m :

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

where x_i is scalar rv (for ease), $E(\varepsilon_i|x) = 0$

- This is just a generalization of the linear model:

$$m(x_i) = x_i' \beta$$

- The goal is to estimate m

First Stab

-

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

where x_i is k -vector of rv's, $E(\varepsilon_i|x) = 0$

- This is just a generalization of the linear model:

$$m(x_i) = x_i' \beta$$

- The goal is to estimate m

Local Regression

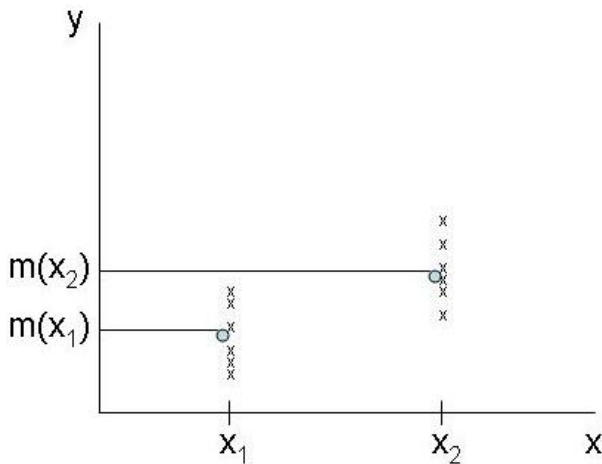
- Imagine x_i is a discrete rv.
- For each value that x_i can take, such as x , we can just average all of the y_i at that point to estimate m .

$$\hat{m} = \frac{1}{N_x} \sum_{i: x_i = x} y_i$$

where N_x is the number of observations where $x_i = x$

- This estimator is consistent (and a lot like OLS)

Local Regression - An Illustration



More Generally

- This local averaging procedure can be defined by

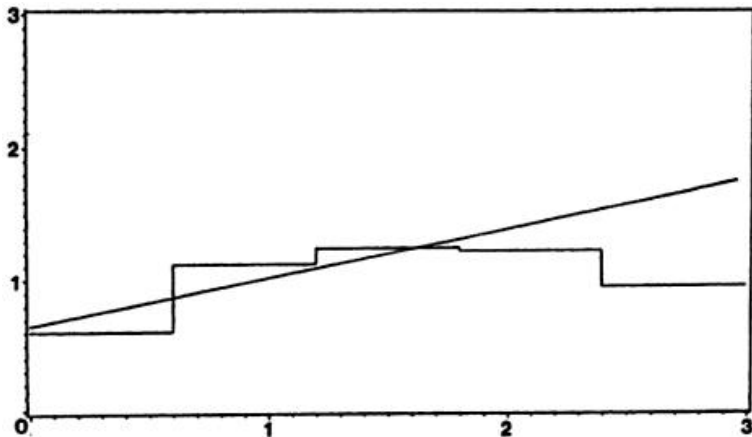
$$\hat{m} = \frac{1}{N} \sum_{i=1}^N W_{ni}(x) Y_i \quad (1)$$

where $\{W_{ni}(x)\}_{i=1}^N$ is a sequence of **weights** which may depend on the whole vector $\{X_i\}_{i=1}^N$

- Same bias versus variance tradeoff:
 - Large window width \implies a lot of smoothing \implies a lot of bias but small variance
 - Small window width \implies a lot of smoothing \implies little bias but a lot of variance

Nonparametric Regression Example

- Assume constant weights \implies jagged discontinuous function



Least Squares

- Local averaging formula (1) is a least squares estimator
- Assume weights $\{W_{ni}(x)\}_{i=1}^N$ are > 0 & sum to 1 $\forall x$

$$N^{-1} \sum_{i=1}^N W_{ni}(x) = 1$$

- Then \hat{m} is a least squares estimate at x since \hat{m} is the solution to

$$\begin{aligned} & \min_{\theta} N^{-1} \sum_{i=1}^N W_{ni}(x) (Y_i - \theta)^2 \\ &= N^{-1} \sum_{i=1}^N W_{ni}(x) (Y_i - \hat{m}(x))^2 \end{aligned}$$

- Local avg is like finding a local WLS estimate

The Kernel

- Kernel regression defines the weight fcn W by a continuous, bounded (often symmetric) real function — the kernel K — that integrates to one.
- The weight sequence is:

$$W_{Ni}(x) = K_{h_N}(x - X_i) / \hat{f}_{h_N}$$

where

$$\hat{f}_{h_N} = N^{-1} \sum_{i=1}^N K_{h_N}(x - X_i)$$

$$K_{h_N}(u) = h_N^{-1} K(u/h_N)$$

is the kernel with scale factor h_N and N is still the sample size

- \hat{f}_{h_N} is the *Rosenblatt-Parzen* kernel density estimator of the marginal density of X

Nadaraya-Watson Estimator

- The complete weighting sequence is:

$$W_{Ni}(x) = h_N^{-1}K(x - X_i/h_N)/N^{-1} \sum_{i=1}^N h_N^{-1}K(x - X_i/h_N)$$

- This form of weights was proposed by Nadaraya and Watson.
- Hence, the Nadaraya-Watson estimator is

$$\begin{aligned}\hat{m}_h(x) &= N^{-1} \sum_{i=1}^N W_{Ni}(x) Y_i \\ &= \frac{N^{-1} \sum_{i=1}^N K_{h_N}(x - X_i) Y_i}{N^{-1} \sum_{i=1}^N K_{h_N}(x - X_i)}\end{aligned}$$

- Shape of kernel weights determined by choice of K
- Size of the weights determined by h_N (bandwidth)
- For choice of Kernel, see earlier slide

Example

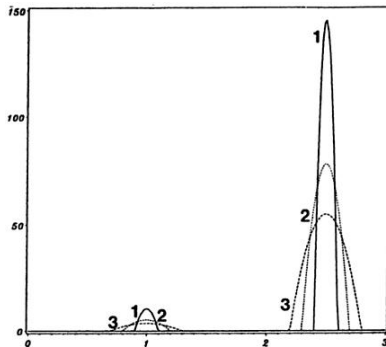


Figure 3.2. The effective kernel weights for the food versus net income data set. $K_h(x \cdot) / \hat{f}_h(x)$ at $x = 1$ and $x = 2.5$ for $h = 0.1$ (label 1), $h = 0.2$ (label 2), $h = 0.3$ (label 3) with Epanechnikov kernel $K(u) = 0.75(1 - u^2)I(|u| \leq 1)$ and density estimate as in Figure 1.5, year = 1973, $n = 7125$. Family Expenditure Survey (1968–1983).

Choice of Kernel

- 1 Smaller bandwidth \implies greater concentration of weights around x
- 2 In regions with sparse data where marginal density estimate \hat{f}_h is small, sequence $\{W_{ni}(x)\}_{i=1}^N$ gives more weight to obs around x
 - There are a lot of X_i 's concentrated around the value $X = 1$, not so many around $X = 2.5$ \implies the density of X , estimated by \hat{f}_h is very large around $X = 1$ and very small around $X = 2.5$ \implies the weights, W_{Ni} , are very small around $X = 1$ and very large around $X = 2.5$ since \hat{f}_h is in the denominator of the weight fn

Univariate Regression 1

- Same model

$$Y_i = m(X_i) + \varepsilon_i$$

- We want to fit this model at a particular x -value, say x_0
- Ultimately, we fit the model at either a representative range of x -values or the N sample points, $x_i : i = 1, \dots, N$
- Run a p th-order regression of Y on X around x_0

$$Y_i = \alpha + \beta_1(X_i - x_0) + \beta_2(X_i - x_0)^2 + \dots + \beta_p(X_i - x_0)^p + \varepsilon_i$$

- Weight the observations according to proximity to x_0 . E.g.,

$$\text{Tricube : } K(t) = \begin{cases} (1 - |t|^3)^3 & |t| < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $t = (X_i - x_0)/h$, h is window width

Univariate Regression 2

- Fitted value at x_0 (i.e., height of estimated regression curve) is $\hat{y}_0 = \alpha$
- It's just the intercept because we centered the predictor x at x_0
- Sometimes we adjust h so that each local regression includes a fixed proportion s of the data
- s is the **span** of the local regression smoother
 - Larger span s , smoother the result
 - Larger the order of the local polynomial, more flexible the smooth

Example

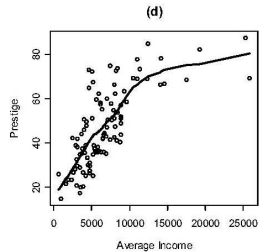
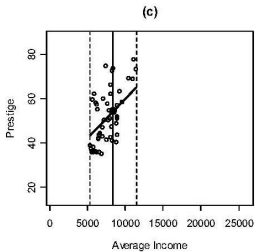
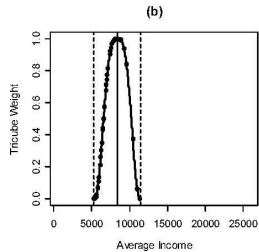
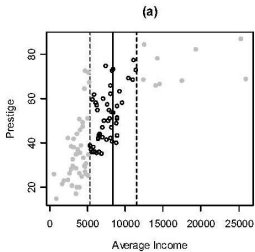


Fig (a): Window Width & Span

- Focus on one point, $x_0 = x_{(80)}$ (i.e., the 80th largest x value)
- This point is denoted by the solid vertical line
- Fig (a) shows the window that includes the 50 nearest x -neighbors of $x_{(80)}$
 - This implies a span s of $\approx 50\%$ ($50/102$)

Fig (b): Kernel

- The tricube kernel provides the weights for all of the observations in the window
- Note the weights are declining in the distance from the reference point $x_{(80)}$
- Note that the tricube $K(t)$ is strictly positive only for $|t| < 1$
- But, the raw distances as measured along the x-axis are much greater than 1
- This is because the argument t is $(X_i - x_0)/h$. So, big h shrinks the argument t

Fig (c): Local Weighted Linear Regression

- The line is a:
 - locally (Just the 50 obs around $x_{(80)}$,
 - weighted (each observation is weighted by the Kernel $K((X_i - x_{(80)})/h)$,
 - linear (assume the polynomial is of order $p = 1$),
 - regression.
- The fitted value of y at $x_{(80)}$, $\hat{y}|_{x_{(80)}}$ is presented as a large solid dot

Fig (d): The Curve

- Local regressions are estimated for a range of x -values (e.g., all the sample points)
- The fitted values are connected to form the curve
 - How are the points connected?

Other Smoothers

- Alternatives to kernel regression include:
 - 1 k Nearest Neighborhood smoothers
 - 2 Orthogonal series smoothers
 - 3 Spline smoothers
 - 4 Recursive smoothers
 - 5 Convolution smoothers
 - 6 Median smoothers

References

- Fox, John, 2002, *Nonparametric Regression Appendix to An R and S-Plus Companion to Applied Regression*
- Silverman, B. W. 1986, *Density Estimation for Statistics and Data Analysis* Chapman & Hall, London, U.K.
- Pagan, Adrian and Aman Ullah, 2006, *Nonparametric Econometrics* Cambridge University Press, Cambridge, U.K.
- Hardle, Wolfgang 1990, *Applied Nonparametric Regression* Cambridge University Press, Cambridge, U.K.