Nonparametric Statistics

Ch.9 Regression function estimation

Motivation

- Very often we are interested in investigating relationship between variables. The simple linear regression model is to model the relationship between two variables as a linear function. However, what if the linear relationship is not true?

- A simple remedy is to use a nonlinear function like a higher order polynomial function. However, it might be still restrictive or the choice of a function class might be very subjective.

- In nonparametric function estimation, no assumptions are made on function structures. Estimation procedures purely depend on the characteristic of data itself.

Parametric vs Nonparametric

- Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from a bivariate distribution. Consider the following regression model.

$$Y = m(X) + \epsilon,$$

where $E(\epsilon|X) = 0$. We want to find the conditional mean function

$$m(x) = E(Y|X = x).$$

Parametric We may assume that the regression function is actually linear and the error follows $N(0, \sigma^2)$.

$$Y = \beta_0 + \beta_1 X + \epsilon \ , \qquad \epsilon \sim N(0, \sigma^2).$$

The parameter space is $F = \{(\beta_0, \beta_1, \sigma^2) : \beta_0, \beta_1 \in R, \sigma^2 \in R^+\}$ and $\dim(F) = 3$.

Nonparametric We do not give any structure on $m(\cdot)$. Instead, we only impose some smoothness conditions. $F = \{m : m' \text{ is continuous}\}$ and $\dim(F) = \infty$.

Parametric regression model

- Constant regression : $m(x) = \beta_0 \quad \forall\, x$

$$Y = \beta_0 + \epsilon \ , \qquad \epsilon \sim N(0, \sigma^2).$$

$$\widehat{\beta_0} = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0)^2 = \bar{Y}$$

- Linear regression : $m(x) = \beta_0 + \beta_1 x \quad \forall\, x$

$$Y = \beta_0 + \beta_1 X + \epsilon \ , \qquad \epsilon \sim N(0, \sigma^2).$$

$$(\widehat{\beta_0}, \widehat{\beta_1}) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = (\bar{Y} - \widehat{\beta_1}\,\bar{X}\ , \frac{\sum_i (X_i - \bar{X}) Y_i}{\sum_i (X_i - \bar{X})^2})$$

- Polynomial regression : $m(x) = \sum_{j=0}^{p} \beta_j x^j \quad \forall\, x$

$$Y = (x) = \sum_{j=0}^{p} \beta_j X^j + \epsilon \ , \qquad \epsilon \sim N(0, \sigma^2).$$

$$(\widehat{\beta_0}, \dots, \widehat{\beta_p}) = \underset{\beta_j}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j X^j \right)^2$$

Parametric regression model

- Nonlinear regression : $m(x) = c/(1 + \exp(ax + b)) \quad \forall\, x$

$$Y = c/(1 + \exp(aX + b)) + \epsilon \ , \qquad \epsilon \sim N(0, \sigma^2).$$

$$(\hat{a}, \hat{b}, \hat{c}) = \underset{a,b,c}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - c/(1 + \exp(aX_i + b)))^2$$

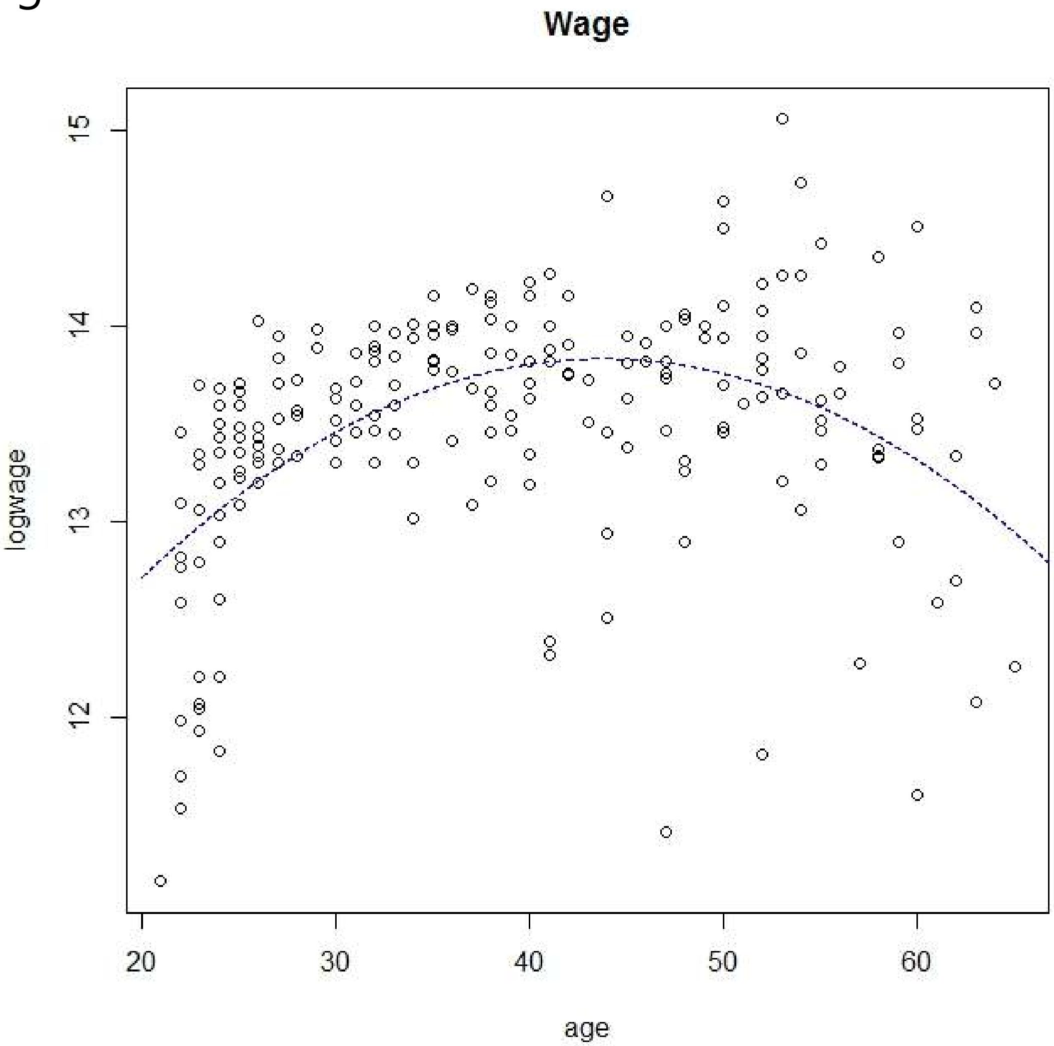
- Generalized linear model : $g\big(m(x)\big) = \beta_0 + \beta_1 x \quad \forall x$

Parametric regression models are sometimes too restrict or subjective.
Nonparametric regression models have no pre-determined form.

"Let the data speak for themselves"

Example : Wage data

Ex] X : age, Y : logarithm of wage



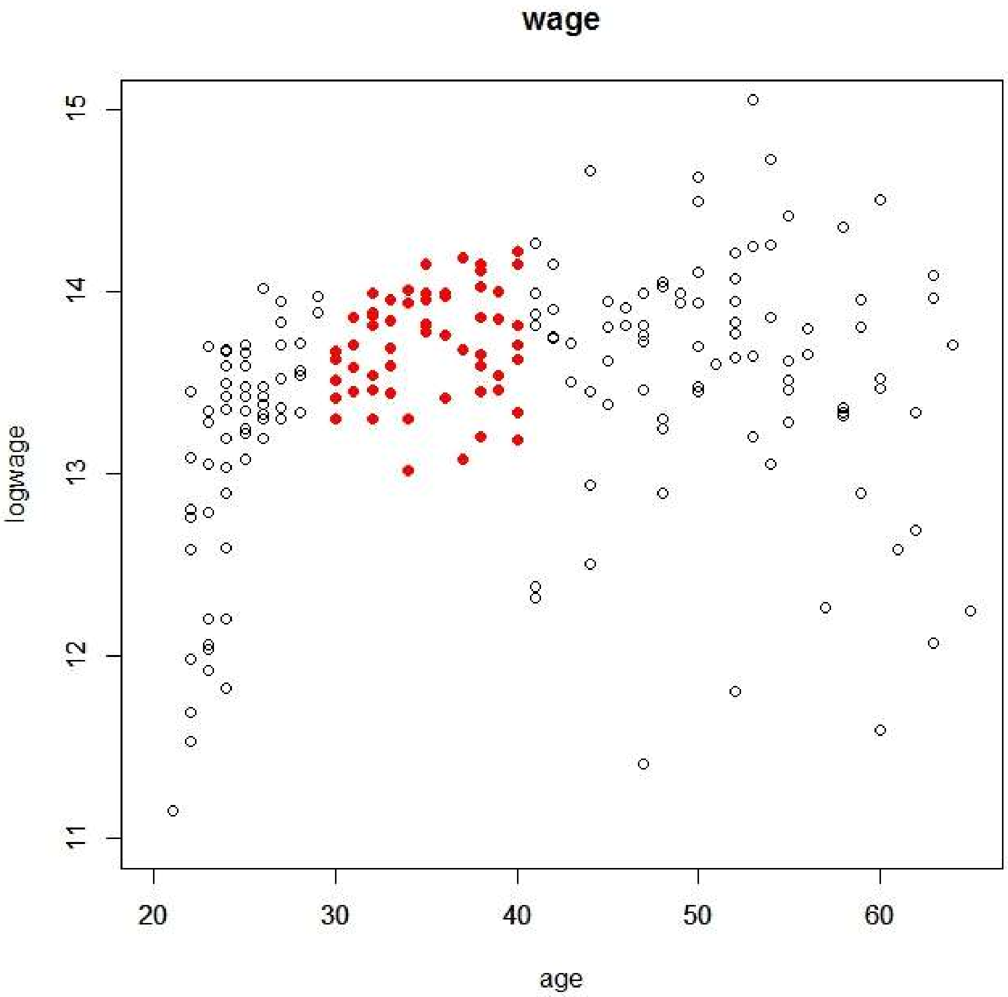
Quadratic function -> (parametric)

Is it enough to capture the characteristic of the data?

IDEA : Localizing

- **Idea** of the local constant regression : $m(x) \approx \beta_0 \quad near \quad x$

Suppose that we are interested in the mean wage at age=35. Which points are more important to make a guess? 60-year-old people or 30-year-old people? A natural answer would be "points close to age=35".



Nonparametric regression model

❖ Local constant regression : $m(x) \approx \beta_0 \quad near \quad x$

Moving average

If we take the points from which the distance is less than or equal to 5.

$$\hat{m}(35) = \frac{1}{N_x}\sum_{i=1}^{n} I(|X_i - 35| \le 5) \cdot Y_i$$
$$= average \quad of \ \ Y_i \ near \ \ (age \ = 35),$$

where $N_x = \sum_{i=1}^{n} I(|X_i - 35| \le 5)$.

For every x,

$$\hat{m}(x) = \frac{1}{N_x}\sum_{i=1}^{n} I(|X_i - x| \le h) \cdot Y_i$$

where $N_x = \sum_{i=1}^{n} I(|X_i - x| \le h)$.

Nonparametric regression model

❖ Local constant regression : $m(x) \approx \beta_0 \quad near \quad x$

Nadaraya-Watson estimator

Moving-average estimation assigns uniform weights to the neighborhood of $x.$ Why not consider different weights?

For every x,

$$\begin{aligned}\widehat{m}(x) &= \frac{1}{N_x}\sum_{i=1}^{n} W_i(x) \cdot Y_i \\ &= weighted \quad average \quad of \ Y_i \ near \ \ x,\end{aligned}$$

where $N_x = \sum_{i=1}^{n} W_i(x)$.

If we set $W_i(x) = K(\frac{X_i - x}{h})$, $\widehat{m}(x)$ is called 'Nadaraya-Watson estimator'. Here K is called 'kernel function' and a positive constant h 'bandwidth' or 'smoothing parameter'.

Nonparametric regression model

- Local constant regression : $m(x) \approx \beta_0 \quad near \quad x$

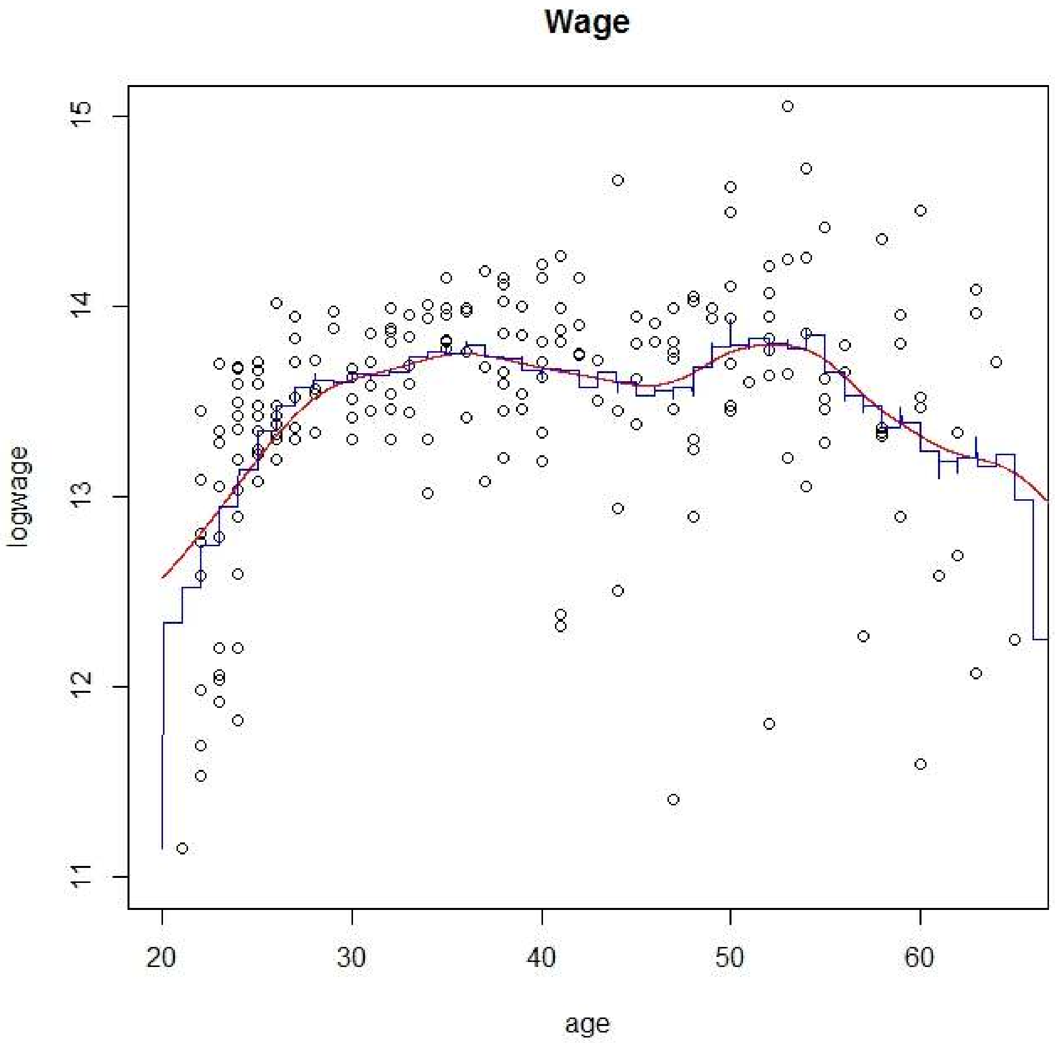
Nadaraya-Watson estimator

(smooth red line)

vs

Moving average

(rigid blue line)



MA vs NW

❖ Local constant regression : $m(x) \approx \beta_0 \quad near \quad x$

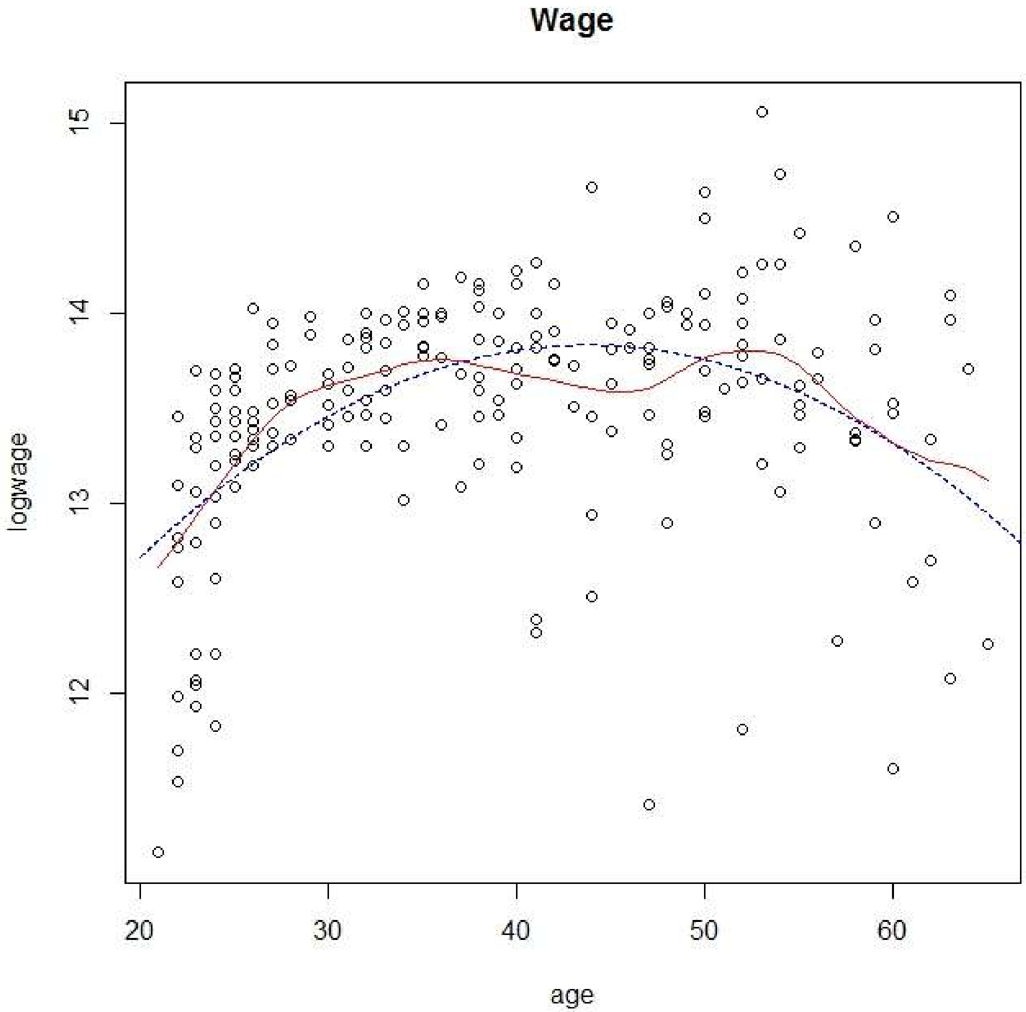
Nadaraya-Watson estimator

(red solid line)

vs

Moving average

(blue dashed line)



Nonparametric regression model

- h controls the smoothness of the estimated line. Too large h results in losing the advantage of localizing, and too small h makes the estimated regression line very rough. The choice of smoothing parameter is very important, we will learn basic principles to select a optimal one in some sense.
- We also need to choose a kernel function K. Normally, a symmetric density function like Gaussian density is employed. It is known that the choice of K is not so important.
- Sometimes, nonparametric regression models are used as primary estimators to check parametric assumptions or to find proper parametric forms.
- We will mainly learn about kernel-based regression, but other methods such as splines or wavelets can be also mentioned if we have time to deal with them.

Several kernel functions

1. Uniform kernel (produces moving average)

$$K(x) = \frac{1}{2}\, I(|x| \leq 1)$$

2. Triangular kernel

$$K(x) = (1 - |x|)\, I(|x| \leq 1)$$

3. Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$$

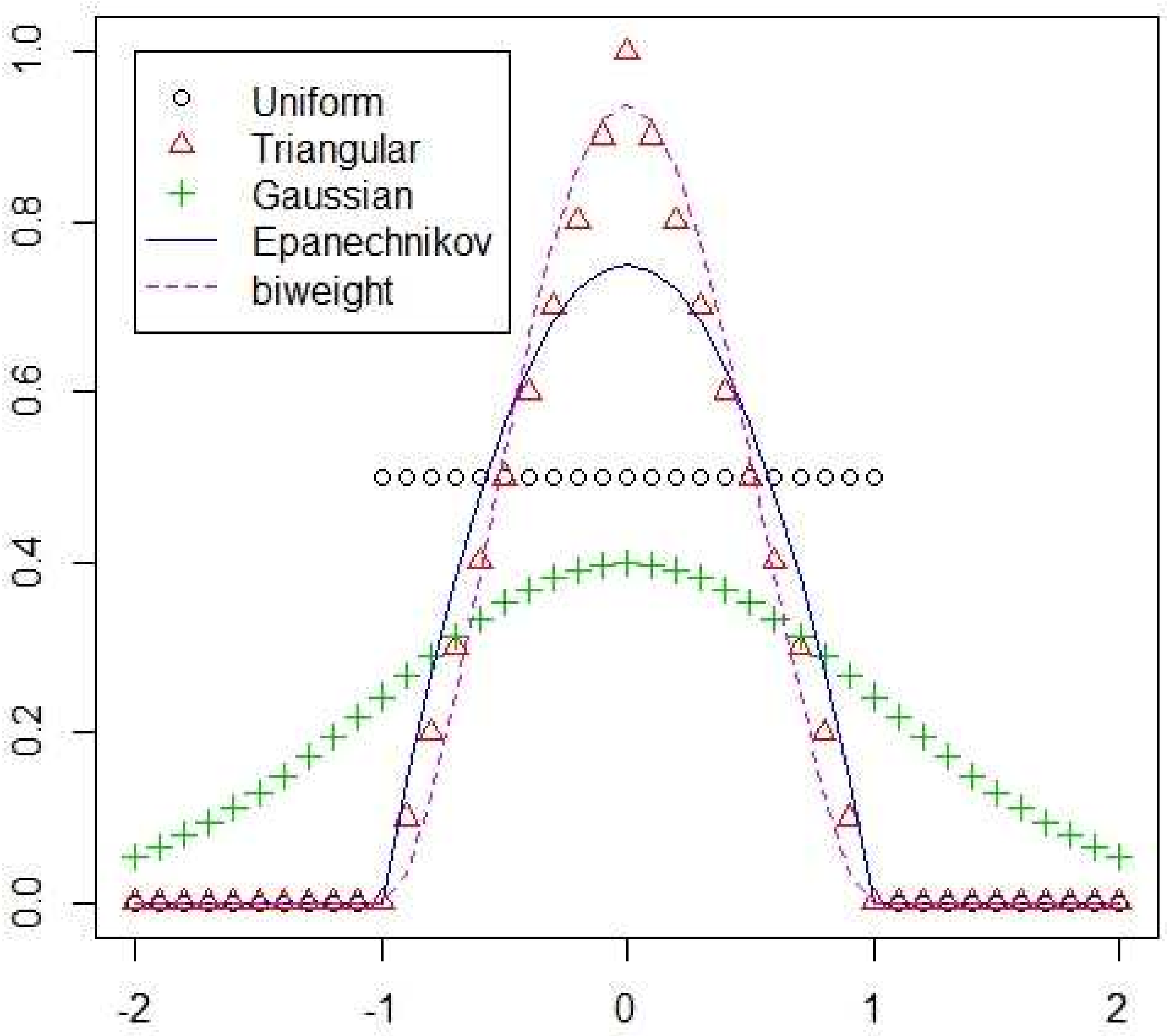
4. Epanechnikov kernel

$$K(x) = \frac{3}{4}(1 - x^2)\, I(|x| \leq 1)$$

5. biweight kernel

$$K(x) = \frac{15}{16}(1 - x^2)^2\, I(|x| \leq 1)$$

Common kernel functions



Local constant regression

- (X_i, Y_i) : iid copies of (X, Y)

  We want to estimate $m(x)$ based on (X_i, Y_i) where

$$Y = m(X) + \epsilon \ , \qquad E(\epsilon | X) = 0$$

- (Parametric or Global) Constant regression : $m(u) = \beta_0 \quad \forall\, u$

$$Y = \beta_0 + \epsilon \ , \qquad \epsilon \sim N(0, \sigma^2).$$

$$\widehat{\beta_0} = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0)^2 = \bar{Y}$$

Local constant regression

- (X_i, Y_i) : iid copies of (X, Y)

  We want to estimate $m(x)$ based on (X_i, Y_i) where

  $$Y = m(X) + \epsilon \ , \qquad E(\epsilon|X) = 0$$

- Local constant regression : $m(u) \approx \beta_0 \quad near \quad x$

  Use only (X_i, Y_i) such that $X_i \approx x$ to approximate $m(X_i)$.

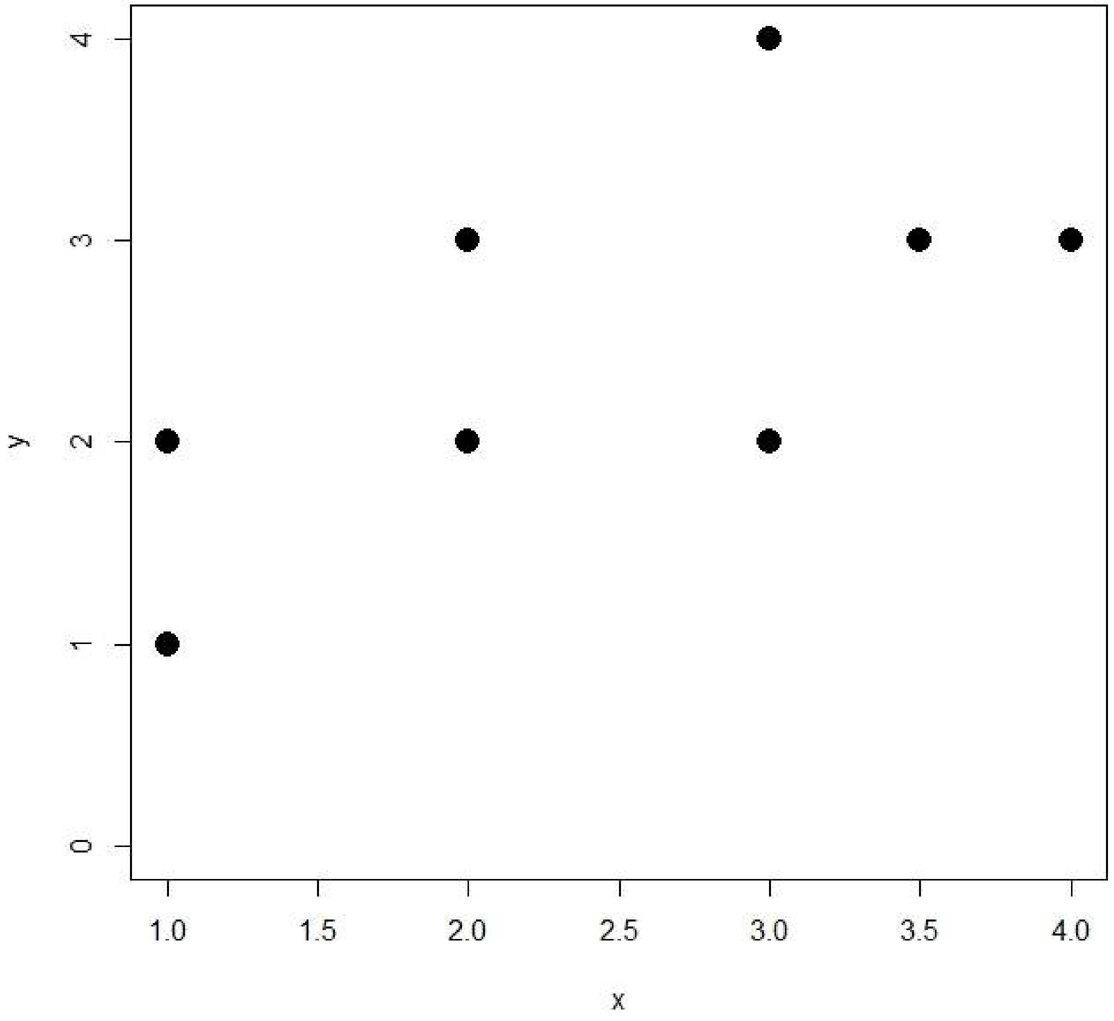
  $$\hat{m}(x) = \widehat{\beta_0} = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0)^2 K_h(X_i - x) = \frac{\sum_i K_h(X_i - x) Y_i}{\sum_i K_h(X_i - x)} = \frac{\sum_i K((X_i - x)/h) Y_i}{\sum_i K((X_i - x)/h)}$$

  where K is a kernel function which is symmetric. The local constant regression estimator is called the Nadaraya-Watson estimator.

  - Note that $\widehat{\beta_0}$ depends on x, the point of interest.
  - We need to repeat solving infinitely many LS problems to obtain $\hat{m}$.
  - It can be shown that $\hat{m}(x)$ is a consistent estimator of $m(x)$, that is, $\hat{m}(x) \to m(x)$ in probability under some suitable conditions.

Example

- We have the next data.
  $(1,1)$ $(1,2)$ $(2,2)$ $(2,3)$
  $(3,2)$ $(3,4)$ $(3.5,3)$ $(4,3)$
  Calculate
  $\hat{f}_{h=1}(2.5), \hat{f}_{h=1.5}(2.5)$ and
  $\hat{f}_{h=1}(3)$ and $\hat{f}_{h=1.5}(3)$



Example

$$\hat{f}_{h=1}(2.5) = \frac{2K\left(\frac{1}{2}\right) + 3K\left(\frac{1}{2}\right) + 2K\left(\frac{1}{2}\right) + 4K\left(\frac{1}{2}\right)}{K\left(\frac{1}{2}\right) + K\left(\frac{1}{2}\right) + K\left(\frac{1}{2}\right) + K\left(\frac{1}{2}\right)} = \frac{11}{4}$$

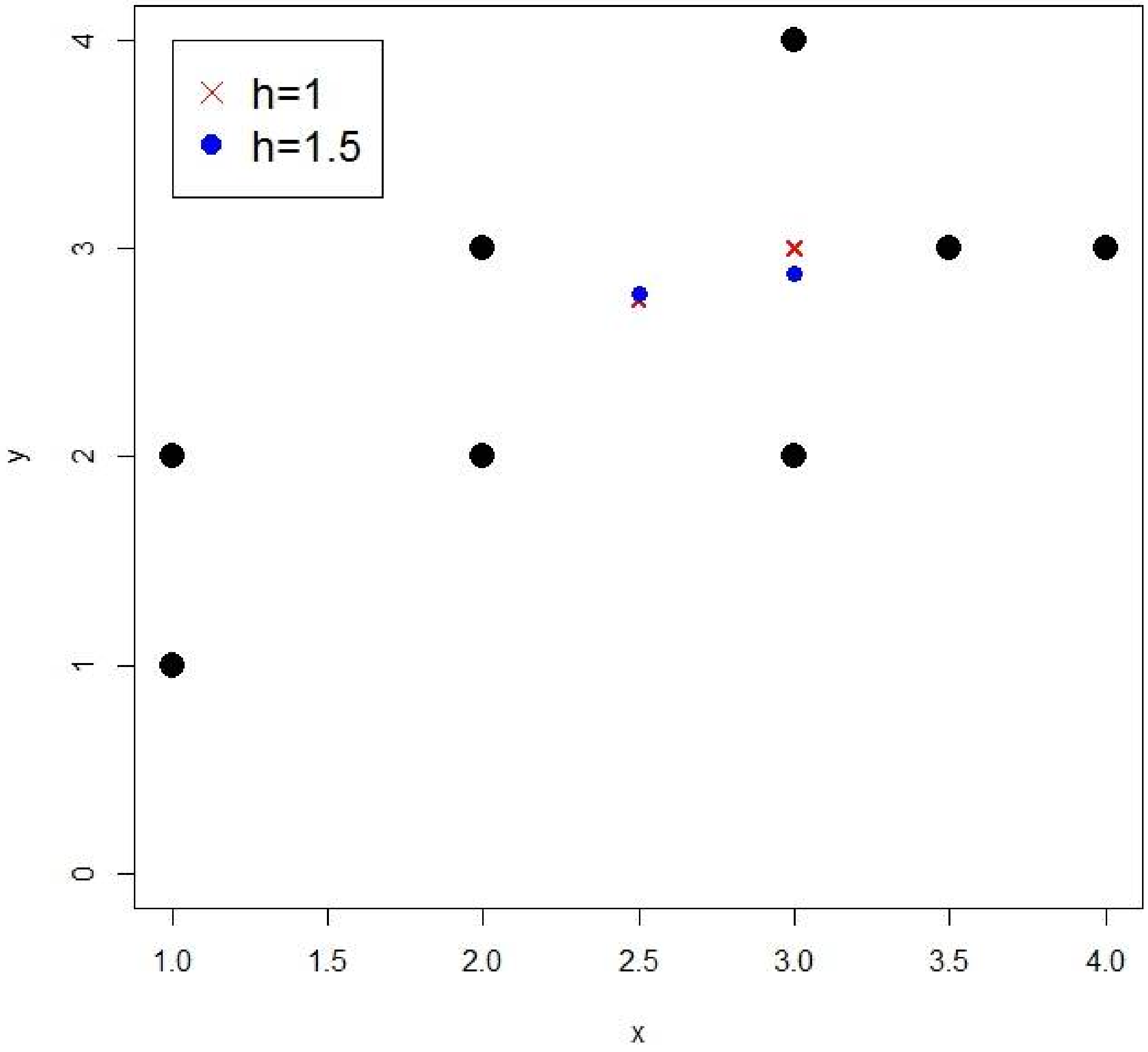
$$\hat{f}_{h=1.5}(2.5) = \frac{2K\left(\frac{1}{3}\right) + 3K\left(\frac{1}{3}\right) + 2K\left(\frac{1}{3}\right) + 4K\left(\frac{1}{3}\right) + 3K\left(\frac{2}{3}\right)}{K\left(\frac{1}{3}\right) + K\left(\frac{1}{3}\right) + K\left(\frac{1}{3}\right) + K\left(\frac{1}{3}\right) + K\left(\frac{2}{3}\right)} = \frac{103}{37}$$

$$\hat{f}_{h=1}(3) = \frac{2K(0) + 4K(0) + 3K\left(\frac{1}{2}\right)}{K(0) + K(0) + K\left(\frac{1}{2}\right)} = 3$$

$$\hat{f}_{h=1.5}(3) = \frac{2K\left(\frac{2}{3}\right) + 3K\left(\frac{2}{3}\right) + 2K(0) + 4K(0) + 3K\left(\frac{1}{3}\right) + 3K\left(\frac{2}{3}\right)}{K\left(\frac{2}{3}\right) + K\left(\frac{2}{3}\right) + K(0) + K(0) + K\left(\frac{1}{3}\right) + K\left(\frac{2}{3}\right)} = \frac{118}{41}$$

$$\because K\left(\frac{1}{2}\right) = \frac{9}{16}\ , K\left(\frac{1}{3}\right) = \frac{2}{3}\ , K\left(\frac{2}{3}\right) = \frac{5}{12}\ , K(0) = \frac{3}{4}$$

Example



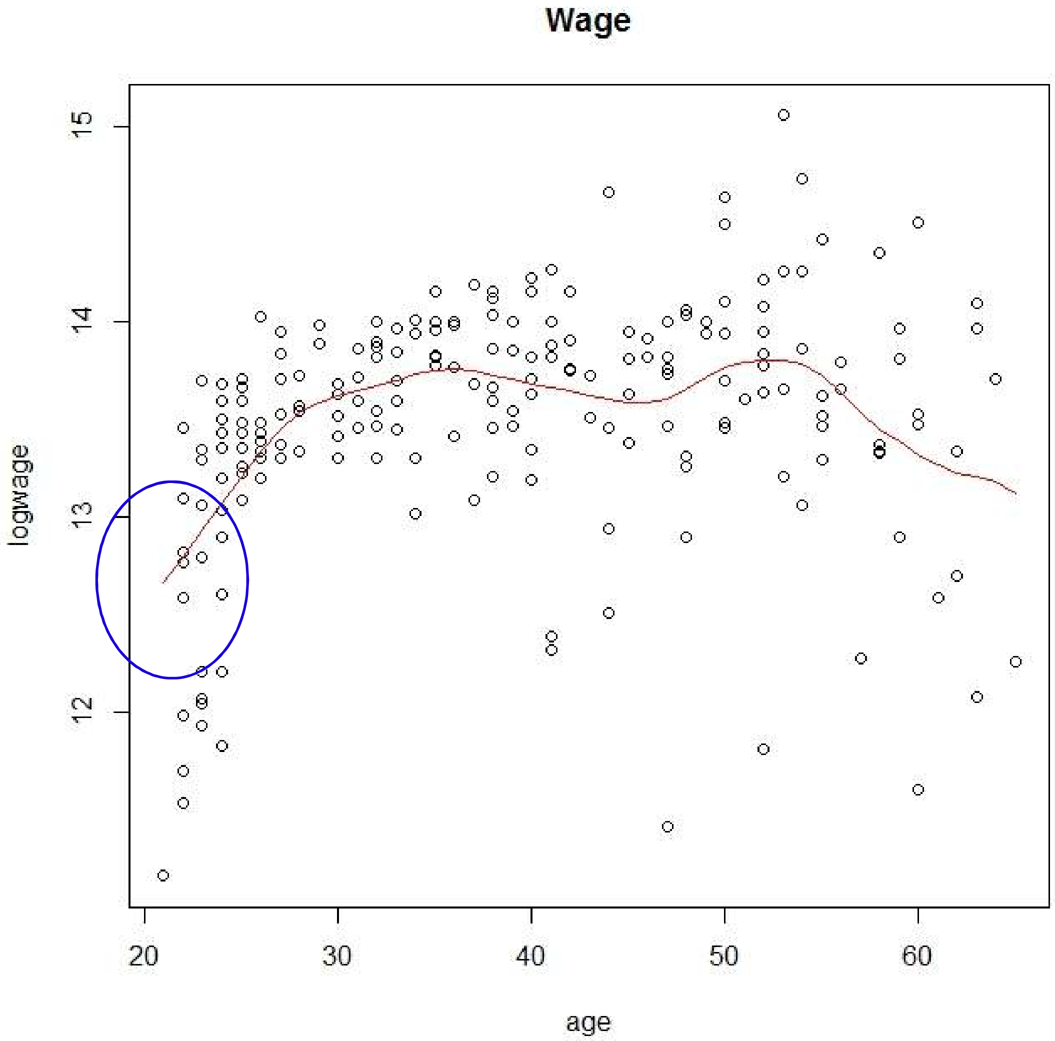
NW estimator

- The local constant estimator is simple, easy to understand and naturally extended from the moving average idea.

- The choice of the kernel also affects the asymptotic distribution. However, it is known that the amount is not so large.

- It is well known that the NW estimator suffers from the boundary problem. Consider the case that we want to estimate the regression function at $x \leq \min(X_i)$ and the true regression function is positively sloped near x. Then, the NW estimator will be upward biased. This means that the NW estimator cannot effectively estimate the regression function near the boundary of the support of X. This is the limitation of the NW estimator.

Boundary effect



Local linear estimator

❖ The boundary problem can be overcome via the local linear fitting.

❖ Linear regression : $m(u) = \beta_0 + \beta_1 u \quad \forall\, x \;\;,\;\; \epsilon \sim N(0, \sigma^2)$.

$$(\widehat{\beta_0}, \widehat{\beta_1}) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = (\bar{Y} - \widehat{\beta_1}\;\bar{X}\;, \frac{\sum_i (X_i - \bar{X}) Y_i}{\sum_i (X_i - \bar{X})^2})$$

❖ Local linear regression : $m(u) \approx \beta_0 + \beta_1 (u - x) \quad \mathit{near} \quad x$

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 (X_i - x))^2\, K_h(X_i - x)$$

- $\widehat{m}^l(x) = \hat{\beta}_0$
- We can also estimate the derivative of m at x. Note that $m'(x) \approx \beta_1$. Therefore, $\widehat{m'}^{\,l}(x) = \hat{\beta}_1$
- This gives a better approximate than local constant fitting.

Local linear estimator

❖ The above minimizing equation can be rewritten with a matrix form.

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1(X_i - x))^2 \, K_h(X_i - x) = (Y - X_x\beta)^\top W_x(Y - X_x\beta)$$

Where, $Y = (Y_1, \dots, Y_n)^\top$, $W_x = diag\ \ (K_h(X_1 - x), \dots, K_h(X_n - x))^\top$, $X_x = (X_{x0}, X_{x1})$, $X_{x0} = (1, \dots, 1)^\top$, $X_{x1} = (X_1 - x, \dots, X_n - x)^\top$, $\beta = (\beta_0, \beta_1)^\top$.

Therefore,

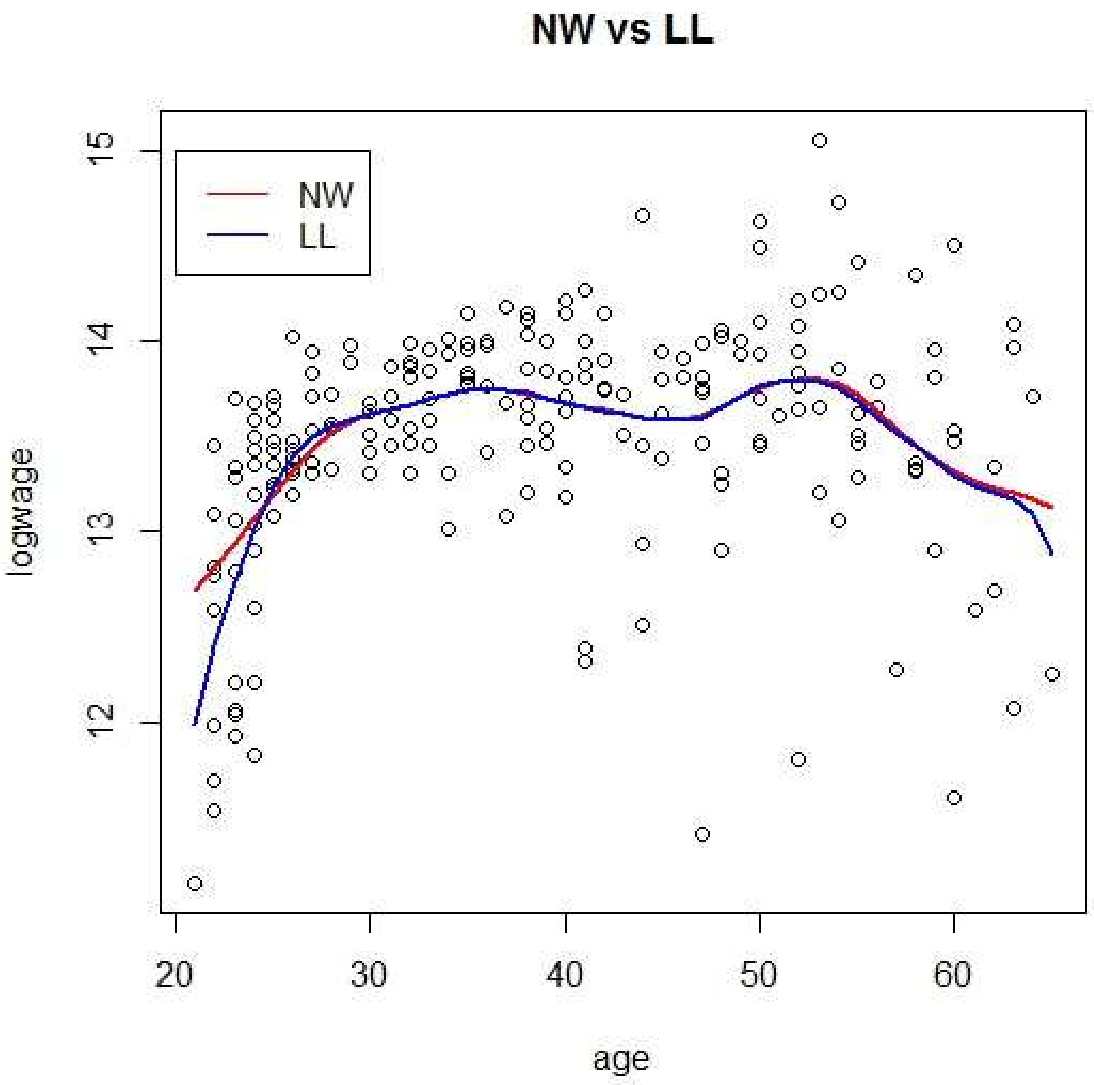
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(Y - X_x\beta)^\top W_x(Y - X_x\beta) = (X_x^\top W_x X_x)^{-1} X_x^\top W_x Y$$

- $\widehat{m}^l(x) = e_1^\top (X_x^\top W_x X_x)^{-1} X_x^\top W_x Y = \hat{\beta}_0$, $e_1 = (1,0)^\top$.
- $\widehat{m'}^{l}(x) = e_2^\top (X_x^\top W_x X_x)^{-1} X_x^\top W_x Y = \hat{\beta}_1$, $e_2 = (0,1)^\top$.
- This is nothing but solving a weighted least squares problem.

Local linear estimator

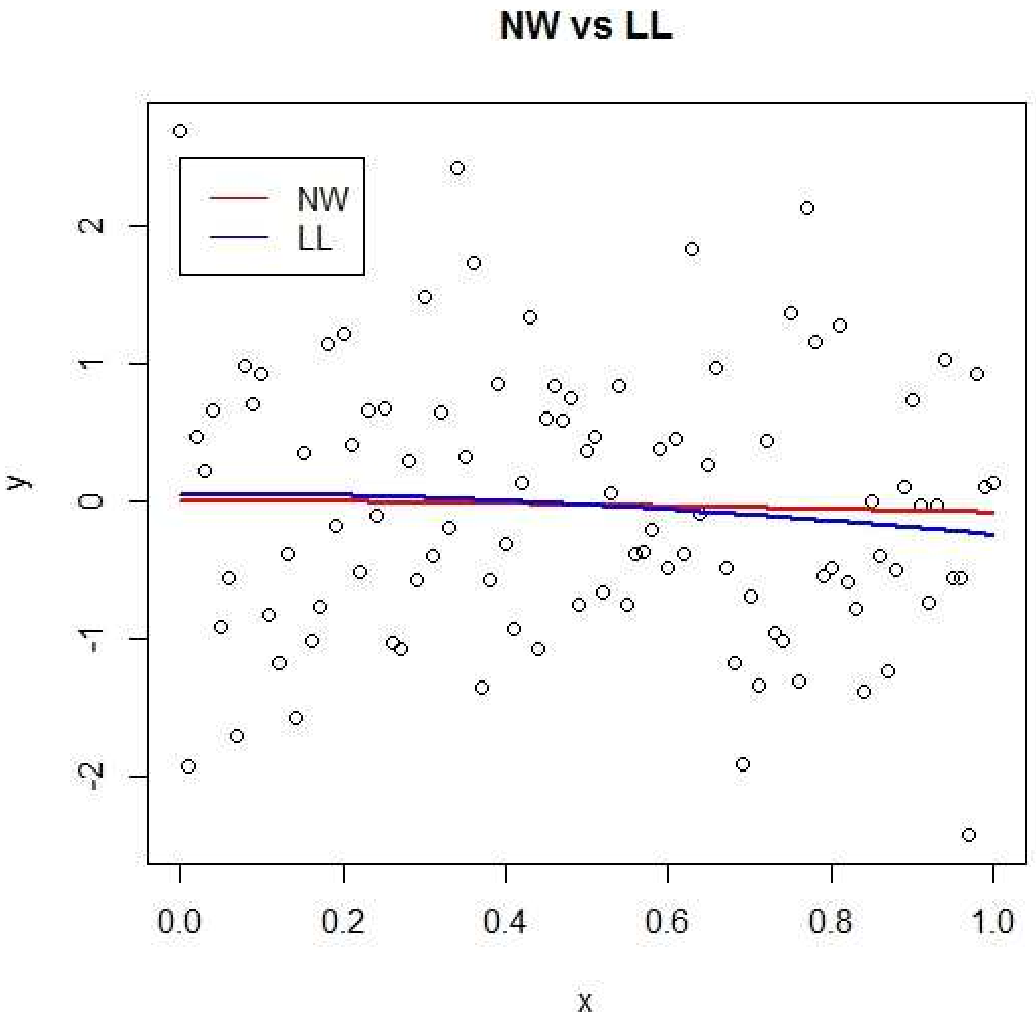
- The local linear estimator improves the NW estimator.
- The major advantage of the local linear estimator compared to the NW estimator is that it has much better properties at the boundary. The local linear estimator is preferable to the NW estimator in a general sense.
- However, the local constant estimator can beat the local linear estimator if the true regression line is quite flat.
- Both estimators are consistent.

Local constant vs Local linear



Local constant vs Local linear

True regression line : $m(x) = 0$



Local polynomial estimator

- ❖ Local polynomial regression : $m(u) \approx \sum_{j=0}^{p} \beta_j (u-x)^j \quad near \quad x$

$$(\hat{\beta}_0, \dots, \hat{\beta}_p) = \underset{\beta_0, \dots, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right)^2 K_h(X_i - x)$$

  - $\widehat{m}^p(x) = \hat{\beta}_0$
  - We can also estimate the derivatives of m at x. $\widehat{m^{(r)}}^p(x) = r!\,\hat{\beta}_r$, $r \le p$.

- ❖ The above minimizing equation can be also rewritten with a matrix form, which will be omitted.

Choice of Kernel

- As mentioned earlier, the choice of K does not have much impact on estimation in practice.

- In local constant or local linear fitting, it is known that the Epanechnikov kernel is optimal in the sense that it minimizes the leading term of

$$E\left[\int \left(\widehat{m}(x) - m(x)\right)^2 dx\right]$$

when h is fixed.

- When a kernel function is compactly supported, regression estimates in a sparse region may not be available. This is not the case with a kernel having non-compact support such as the Gaussian kernel. So, from a practical point of view, using the Gaussian kernel is sometimes convenient. In fact, $\sum_i K_h(X_i - x) = 0$ if there is no data in $\{t: |t - x| \le h\}$ so that the local constant estimator is not defined at x.

Bandwidth selection

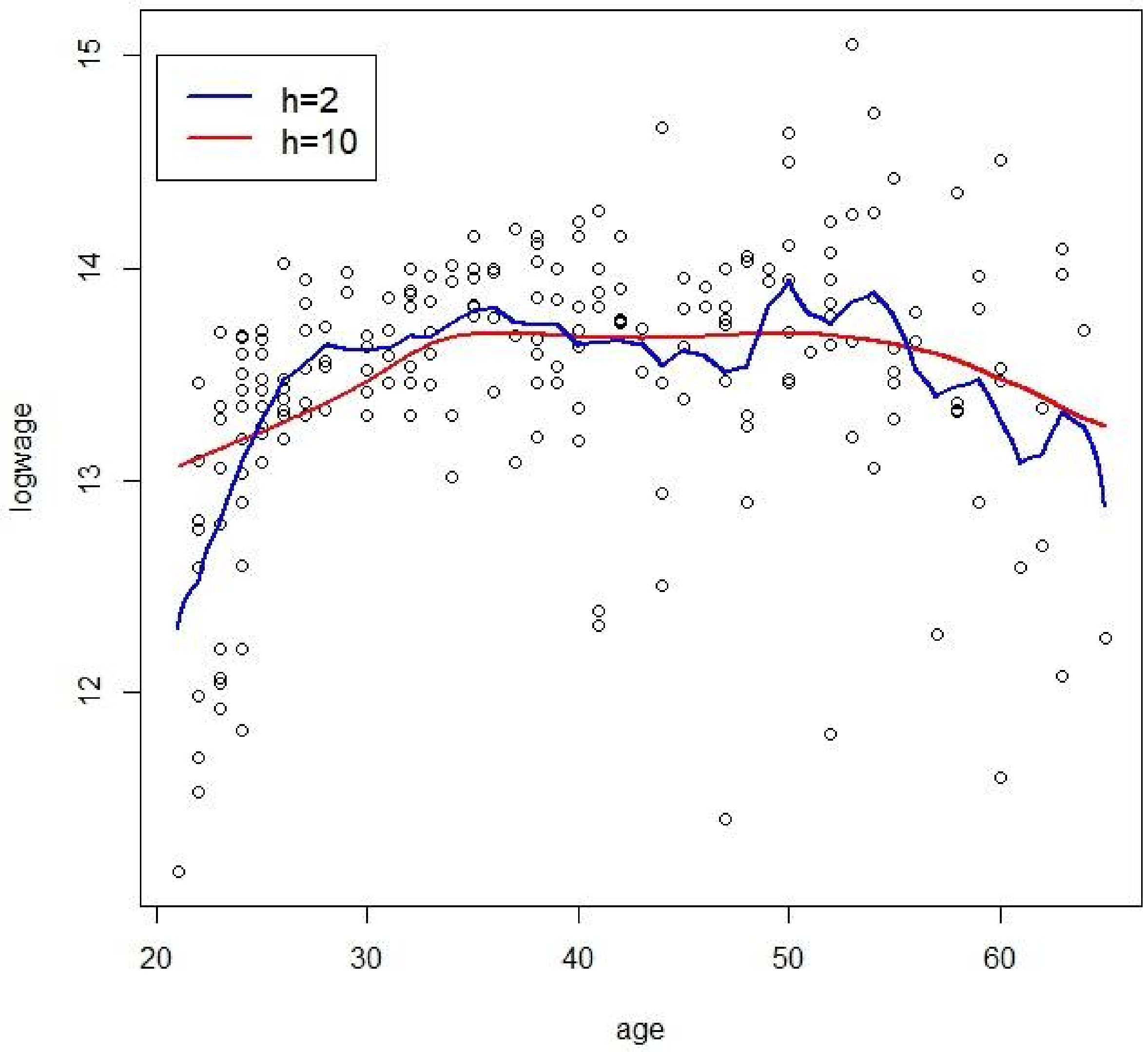
- As in the density estimation, selecting the bandwidth h is very important.

- h is called "smoothing parameter". It directly determines the level of smoothness of an estimated function.

- The bias of the kernel regression estimators is a increasing function of h, whereas the opposite is true for the variance. Therefore, we need to balance them. This is called the bias-variance tradeoff.

- There are a number of existing research papers on bandwidth selection. Asymptotically, it is known that the choice $h \sim n^{-1/5}$ is optimal in the local linear fitting.

Bandwidth selection



Multivariate kernel regression estimator

- What if there is multiple independent variables? For example, the local constant estimator (NW estimator) can be naturally extended for a general case.

- (X_i, Y_i) : iid copies of $(X, Y) \in R^p \times R$

  We want to estimate $m(x)$ based on (X_i, Y_i) where

  $$Y = \boldsymbol{m}(X) + \epsilon \ , \qquad E(\epsilon | X) = 0$$

  where $\boldsymbol{m}(\cdot): R^d \to R$. Then,

  $$\widehat{\boldsymbol{m}}(x) = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0)^2 \boldsymbol{K}_h(X_i - x) = \frac{\sum_i \boldsymbol{K}_h(X_i - x) Y_i}{\sum_i \boldsymbol{K}_h(X_i - x)}$$

  Where $\boldsymbol{K}_h(u) = K_h(u_1) \times \cdots \times K_h(u_p)$, $u = \left(u_1, \dots, u_p\right)^{\top}$.

Dimensionality problem

- The multivariate kernel regression estimator is a natural extension of the univariate one. However, we are to face a serious problem: the so-called

CurSe of dimensionality

- This problem refers to the fact that a local neighborhood in higher dimensions is no longer local. A neighborhood with a fixed percentage of data points can be very big and far from what is understood by the term “local”.

- For the moment, we assume that the kernel function is compactly supported for convenience. A similar problem arises with a non-compact support kernel.

Dimensionality problem

- Univariate case : Let X be uniformly distributed on $[0,1]$. What is your choice for h if you want to include approximately 25% of data to estimate a regression function at 0.5?

  Answer => 0.125

- Bivariate case : Let (X_1, X_2) be uniformly distributed on $[0,1]^2$. What is your choice for h? if you want to include approximately 25% of data to estimate a regression function at $(0.5,0.5)$?

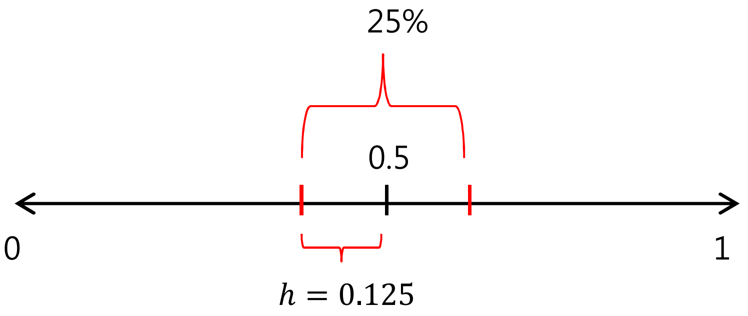
  Answer => 0.25

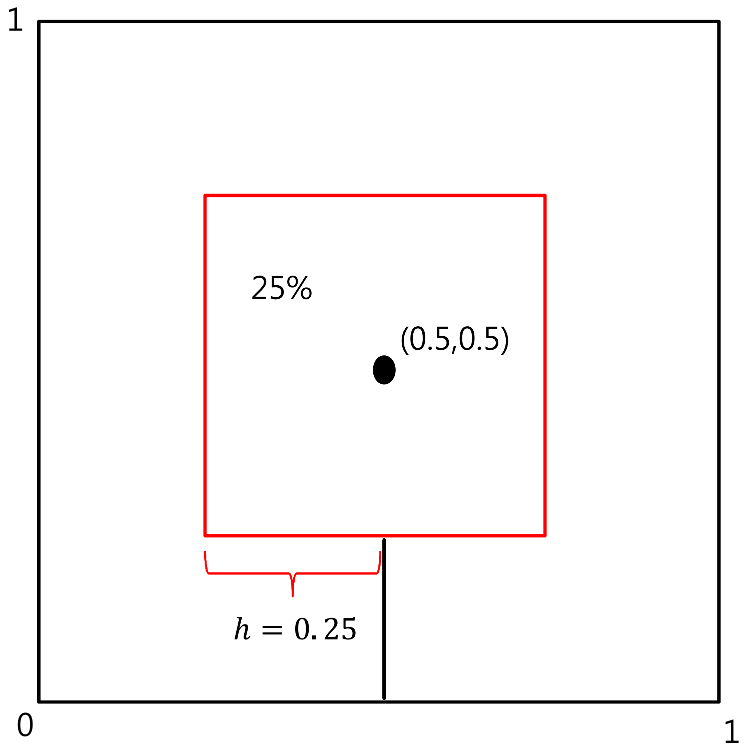
- General case : Let X be uniformly distributed on. What is your choice for h? if you want to include approximately 25% of data to estimate a regression function at $(0.5, \dots, 0.5)$?

  Answer => $2^{-(\frac{2}{p}+1)}$,

  This quantity converges to 0.5 as p increases, which means that one need to cover almost all region in a coordinatewise manner.

Example





Dimensionality problem

- ❖ Therefore, if p increases, we need to broaden our window to include a similar portion of data, which results in making our estimator no longer “local”.

- ❖ Reversely speaking, if we keep the same size of bandwidths, the number of data included in our window decreases geometrically. For example, suppose that $n = 1000$ and $h = 0.25$ then, we have 500 observations in our window when d=1. However, if d=10, only one or less observations in average are available because $1000 * 0.5^{10} = 0.98$.

- ❖ This is a crucial problem to be resolved in the kernel regression context because including multiple variables in models is very common these days. It is unavoidable to analyze complex data.

- ❖ Note that parametric models are not affected by this kind of problem since they normally employ some global criterion to estimate functions.