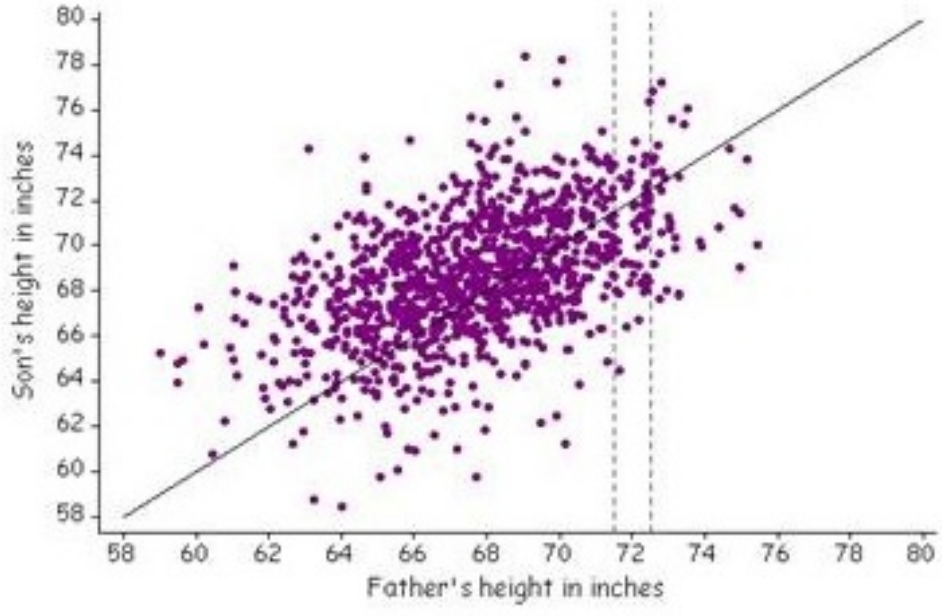
비모수 커널 추정법

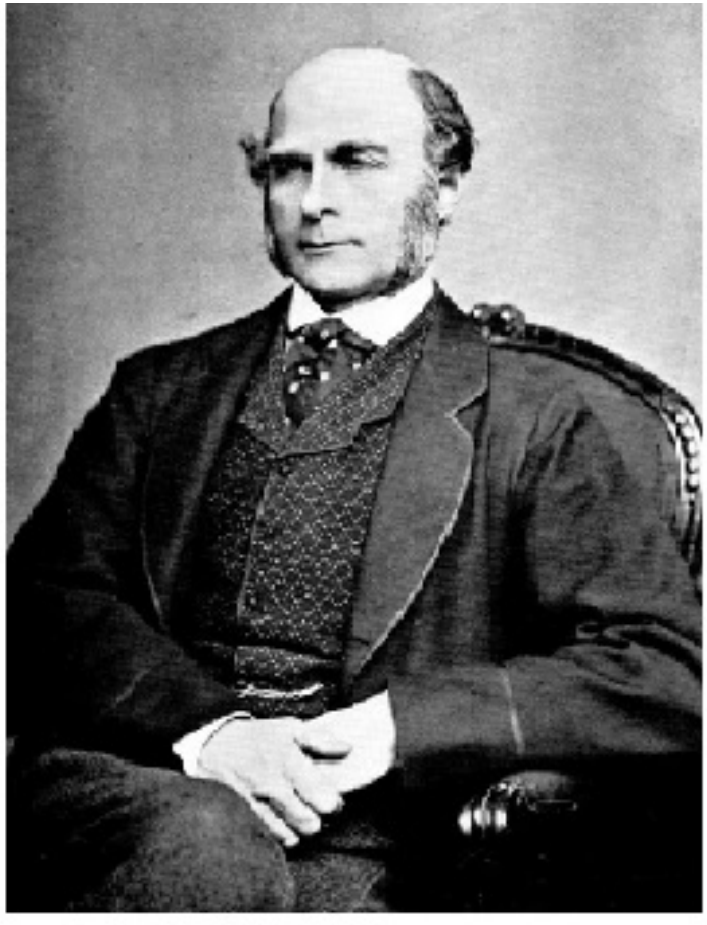
양성준
전북대학교 통계학과

2022-06-02

회귀분석(Regression analysis)이란?

- 역사 : Francis Galton (1822-1911)





- 평균으로의 회귀 : 아들의 키는 아버지보다 평균치로 회귀하는 경향이 있다.

회귀분석(Regression analysis)이란?

- 변수들 사이의 연관성

$$Y \sim g(X)$$

X : 입력/설명/독립/예측변수, Y : 출력/반응/종속변수

- 변수들 간의 연관성을 나타내는 함수 g를 추정하기 위한 일련의 과정

회귀분석(Regression analysis)이란?

- (X, Y)에 대해 다음 순서쌍의 관측을 가정

$$(X_j, Y_j),\ j = 1, 2, \ldots, n$$

- 자료생성과정은 (관측불가능한) 오차항의 영향을 받음

$$Y_j = g(X_j) + u_j$$

u_j : 평균이 0인 오차변수

- 예] X : 연령, Y : 임금
  X : 체온, Y : 심장박동수
  X : GDP, Y : 코로나치명률
  X : 연령, Y : 퇴직률

Q. 어떻게 g를 추정할 것인가?

모형설정 : 모수적(Parametric) 방법

- 선형모형 : $g(x) = a + bx$

$$Y_j = a + bX_j + u_j$$

- 이차모형

$$Y_j = a + bX_j + cX_j^2 + u_j$$

- 지수모형

$$Y_j = ae^{bX_j} + u_j$$

- 이외에도 g에 대한 수없이 많은 선택이 가능

모수적 방법의 특징

- 장점 : 적용 및 추정이 상대적으로 간단하다. 실제 연관성의 체계에 대한 믿음이나 정보가 있는 경우 가장 효율적이고, 명확한 해석이 가능하다.

- 단점 : 모형 설정이 주관적일 수 있다. 설정된 모형의 틀을 벗어나는 자료의 특징은 설명할 수 없다. (모형 설정의 오류)

비모수적 방법

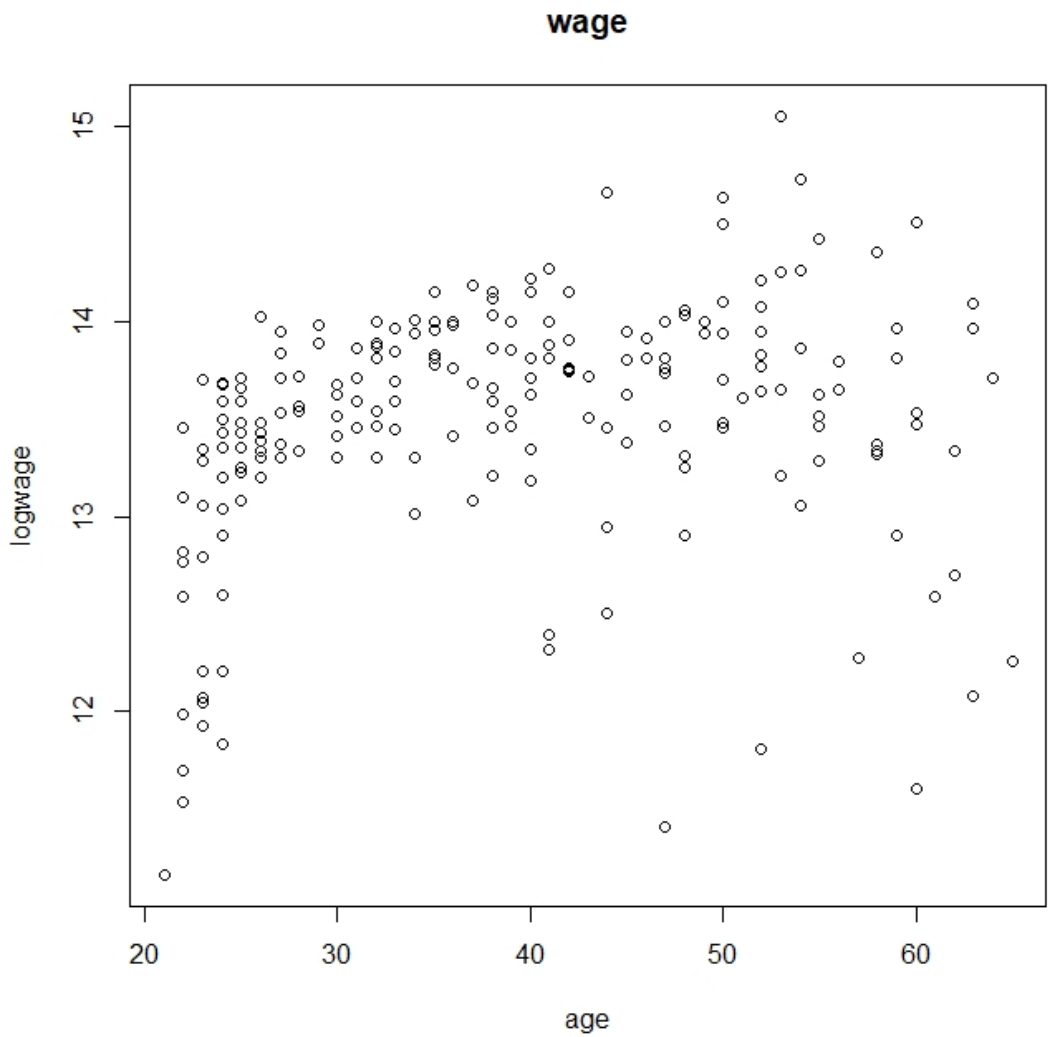
- 모형에 대한 사전 설정/가정을 최소화
- 데이터 자체의 특성만을 이용하여 연관성을 추정하고자 함

"Let the data speak for themselves"

- 커널 추정법은 비모수적 방법 중 하나로 g에 대해서는 적당히 부드러운(smooth) 성질을 가진다는 가정만 상정
- 자료 특성의 국소적인 변화를 모수적 방법에 비해 민감하게 추정할 수 있음

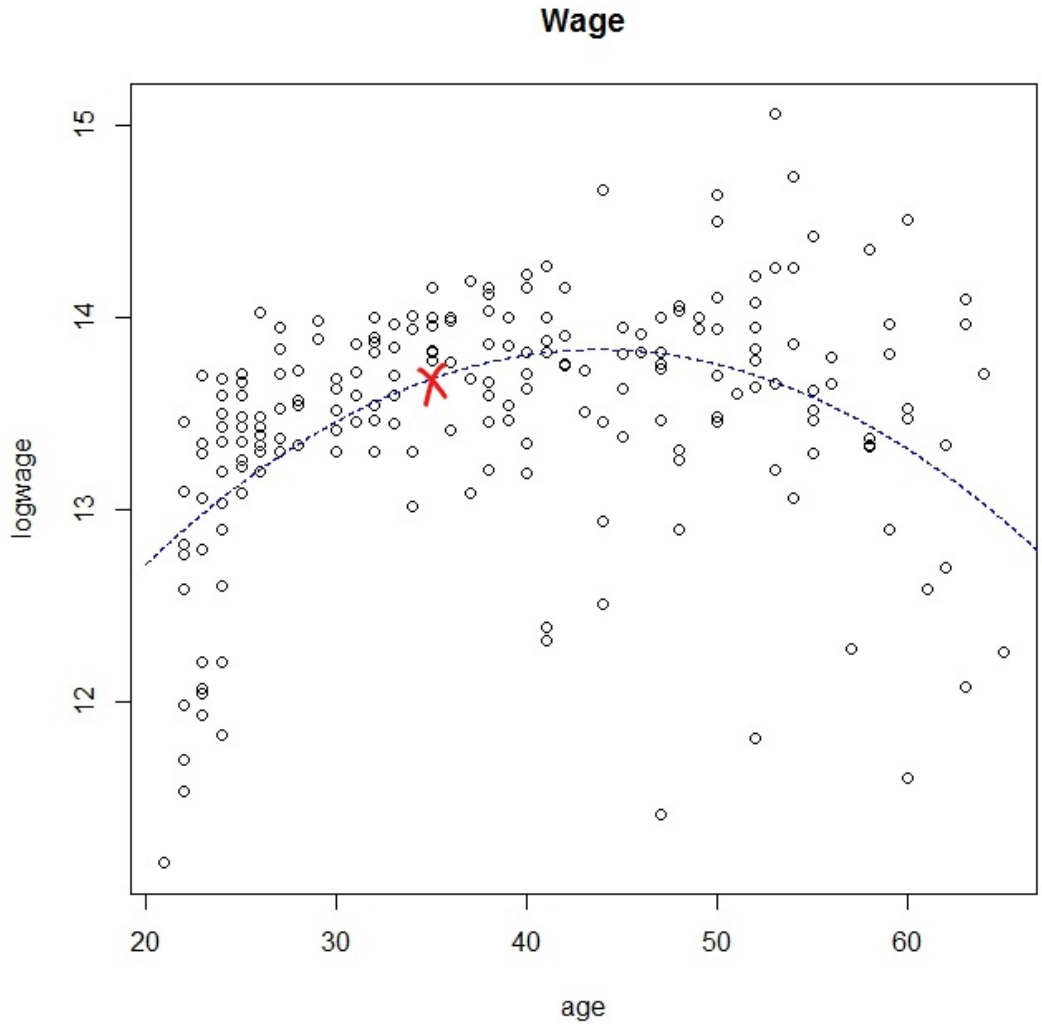
예시 : 임금데이터

- X : 연령, Y : 임금
- 35세의 평균임금은? : $g(35)$



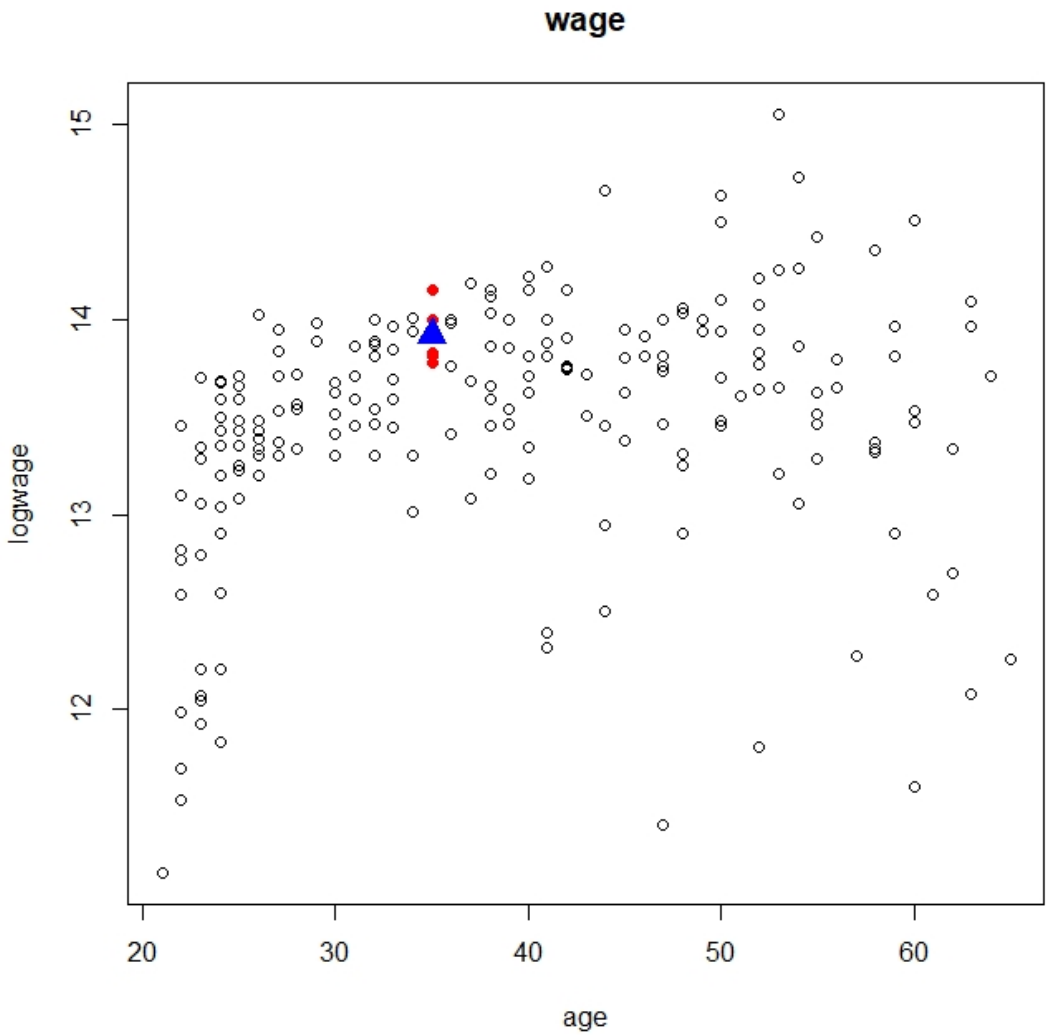
이차모형 적합

- 데이터의 특성을 충분히 반영하지 못함



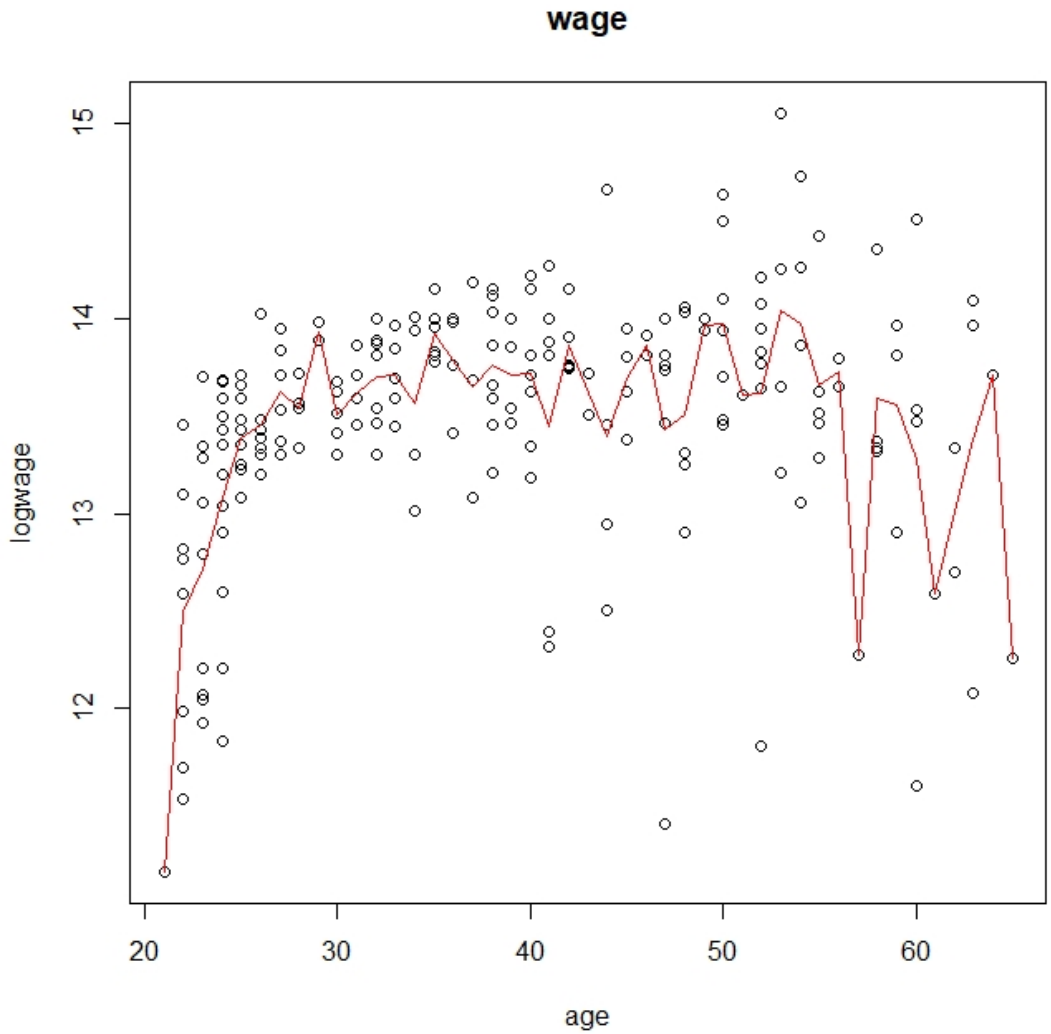
특성을 만족하는 관측치만 활용

- 연령이 35세에 해당하는 관측치의 평균
- 정보의 제한적 활용



특성을 만족하는 관측치만 활용

- 연령별 임금 평균 추정치
- 연령에 따른 변화가 지나치게 심하게 나타남



국소평균 (local average)

- 연령별 평균에 의한 추정은 다음과 같은 문제를 가짐

  - 추정된 g가 연속이라는 보장이 없음
  - 특정 연령에서의 관측치가 없거나 부족한 경우 추정이 불안정

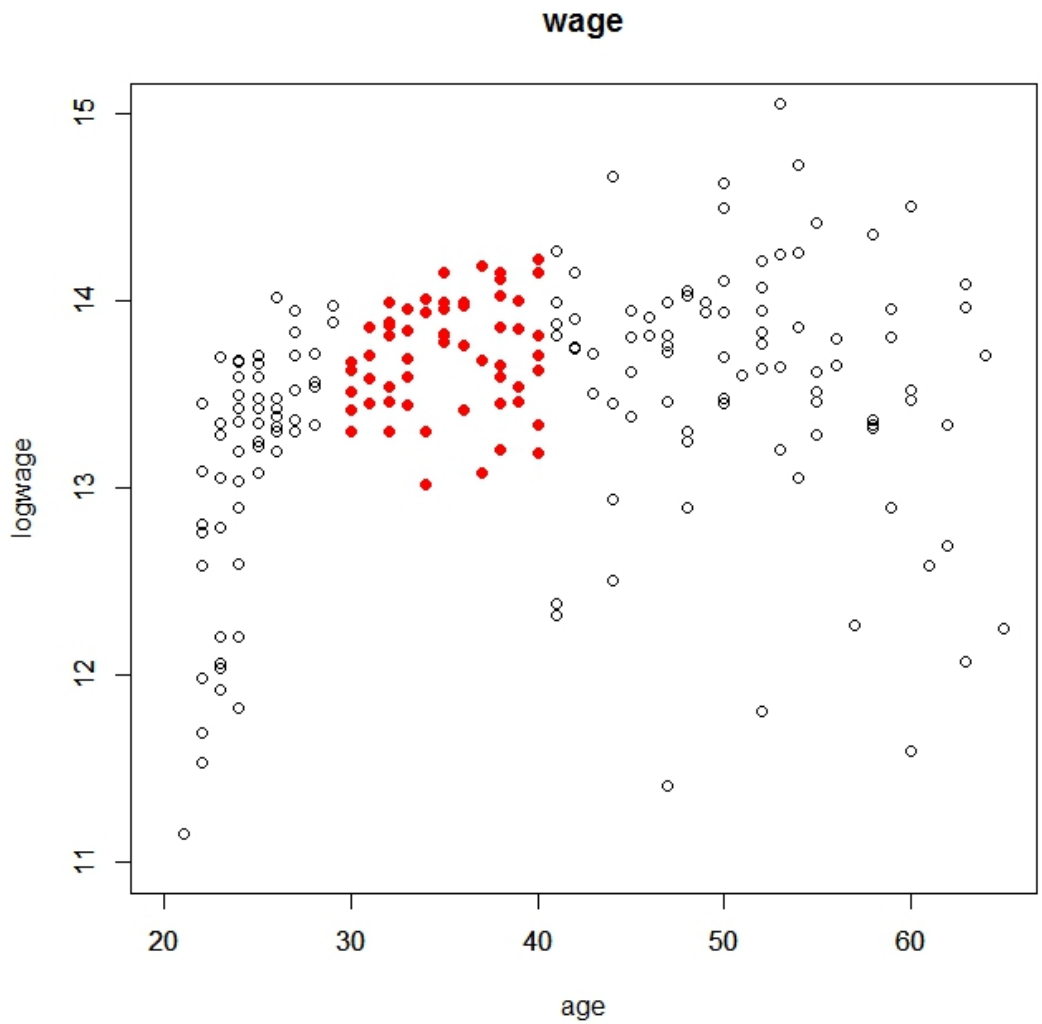
- 연령이 비슷하면 평균임금도 비슷하지 않을까?

  - 특정 연령의 평균연령 추정을 위해 이웃한 관측치를 활용

- 국소화 (localizing)

국소평균 (local average)

- 35세 근처의 관측치들의 평균으로 $g(35)$를 추정



국소평균 (local average)

- 국소평균

$$\begin{aligned} \hat{g}(35) &= \frac{1}{N_{35}} \sum_{j=1}^{n} logwage_j \cdot I(|age_j - 35| \leq 5) \\ &= \text{35세와 차이가 5세 이하인 관측치들 임금의 평균} \end{aligned}$$

$N_{35} = \sum_{j=1}^{n} I(|age_j - 35| \leq 5)$: 이웃 관측치 수

$I(\cdot)$: 지시함수로 조건이 만족되면 1, 아니면 0의 값을 가짐

- 35 대신 원하는 연령을 넣으면 모든 연령에 대한 $\hat{g}(\cdot)$ 추정 가능

커널평활 (Kernel smoothing)

- 국소평균은 관심지점 근처 모든 관측치에 동등한 가중치를 부여
- 더 가까운 관측치가 더 중요하지 않을까?
- 국소 평균의 일반화

$$\begin{aligned}\hat{g}(35) &= \frac{1}{N_{35}} \sum_{j=1}^{n} logwage_j \cdot W_j(35) \\ &= \text{35세와 차이가 5세 이하인 관측치들 임금의 가중평균}\end{aligned}$$

$N_{35} = \sum_{j=1}^{n} W_j(35)$: 가중치 총합

$W_j(\cdot)$: 가중치 함수

- 35 대신 원하는 연령을 넣으면 모든 연령에 대한 $\hat{g}(\cdot)$ 추정 가능

커널평활 (Kernel smoothing)

- W_j의 결정이 추정 전체를 결정하는 핵심임
- 커널(Kernel) 함수 K를 도입하여

$$W_j(35) = K\left(\frac{age_j - 35}{h}\right), \ h : \text{평활모수}$$

와 같이 설정하는 것을 국소상수 추정이라 한다.

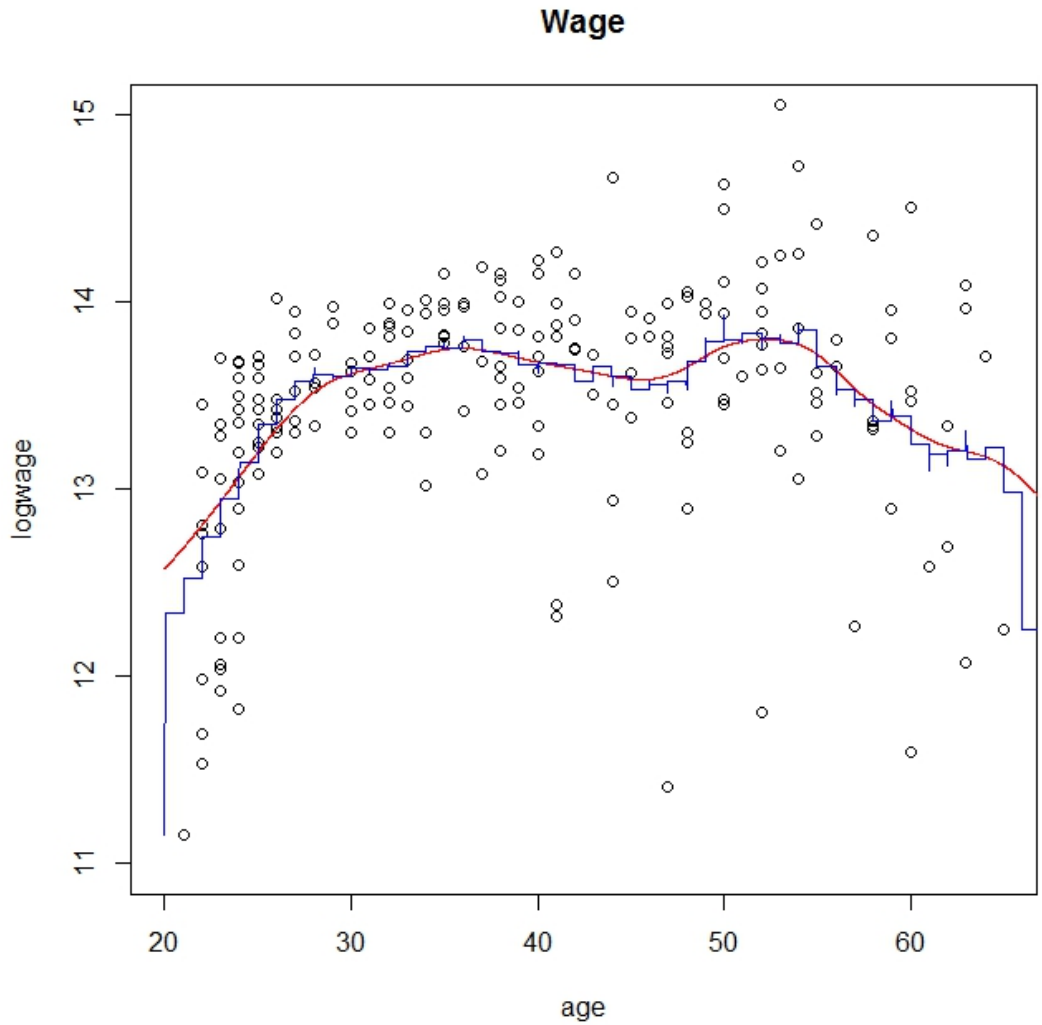
- 국소상수 (local constant) 추정 : Nadaraya-Watson (NW)

$$\hat{g}(35) = \frac{\sum_{j=1}^{n} logwage_j \cdot K\left(\frac{age_j - 35}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{age_j - 35}{h}\right)}$$

- 35 대신 원하는 연령을 넣으면 모든 연령에 대한 $\hat{g}(\cdot)$ 추정 가능

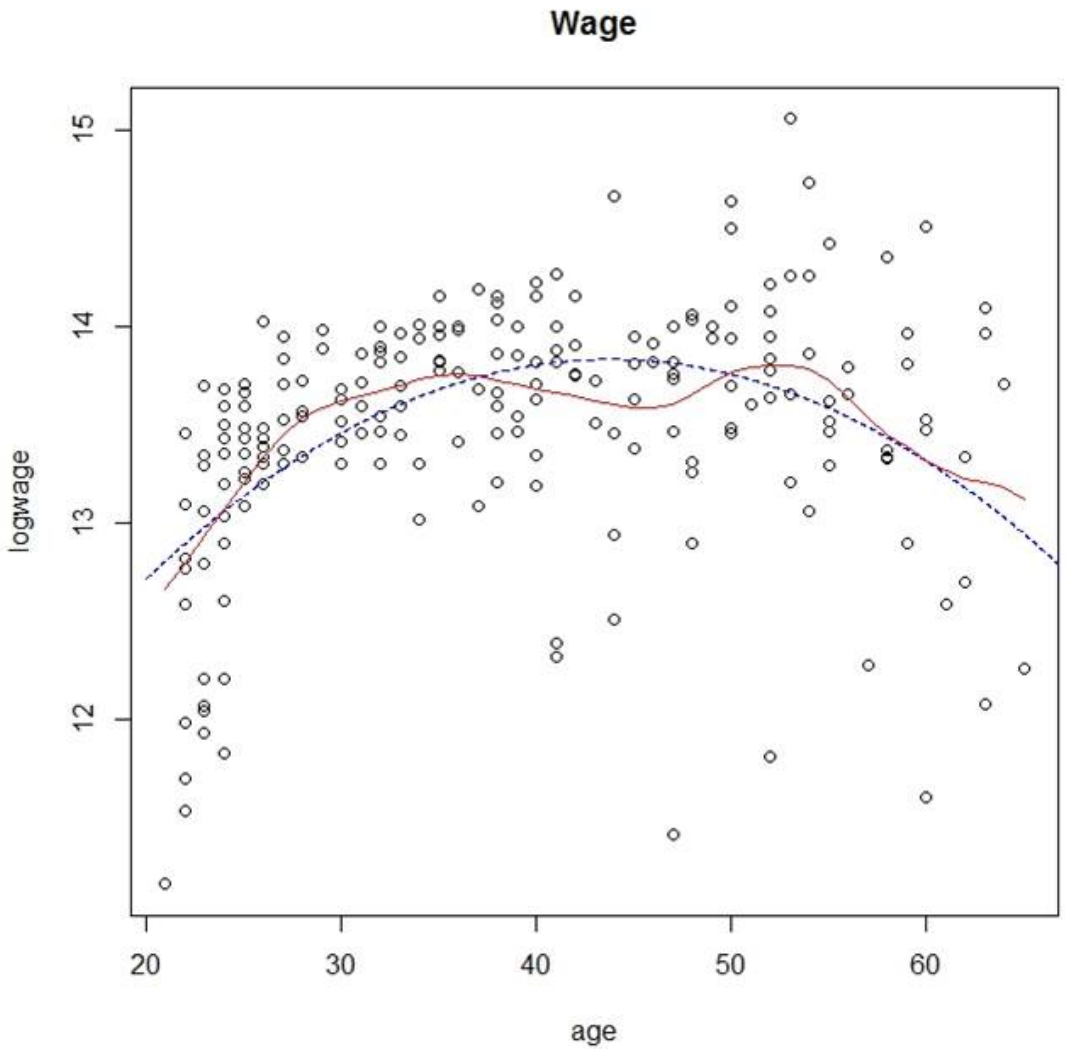
국소평균 vs 국소상수

- 국소평균(blue)은 계단형, 국소상수(red)는 부드러운 함수로 나타남



이차모형 vs 국소상수

- 이차 (blue) vs 국소상수 (red)



커널의 선택

- 커널함수 $K(\cdot)$는 보통 다음 성질을 가지도록 선택

  - $\int_{-\infty}^{\infty} K(u)du = 1$

  - $K(u) \geq 0$ (non-negative)

  - $K(u) = K(-u)$ (symmetric)

  - $\int_{-\infty}^{\infty} uK(u)du = 0$

  - 적당한 유계조건, 미분가능조건, 적률조건, 꼬리부분에 대한 조건이 추가되기도 함

여러 커널 함수

- Uniform kernel : 국소평균추정

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

$K(\frac{age_j - 35}{5}) = \frac{1}{2} I(|age_j - 35| \leq 5)$

- Triangular kernel

$$K(u) = (1 - |u|) I(|u| \leq 1)$$

- Gaussian kernel

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

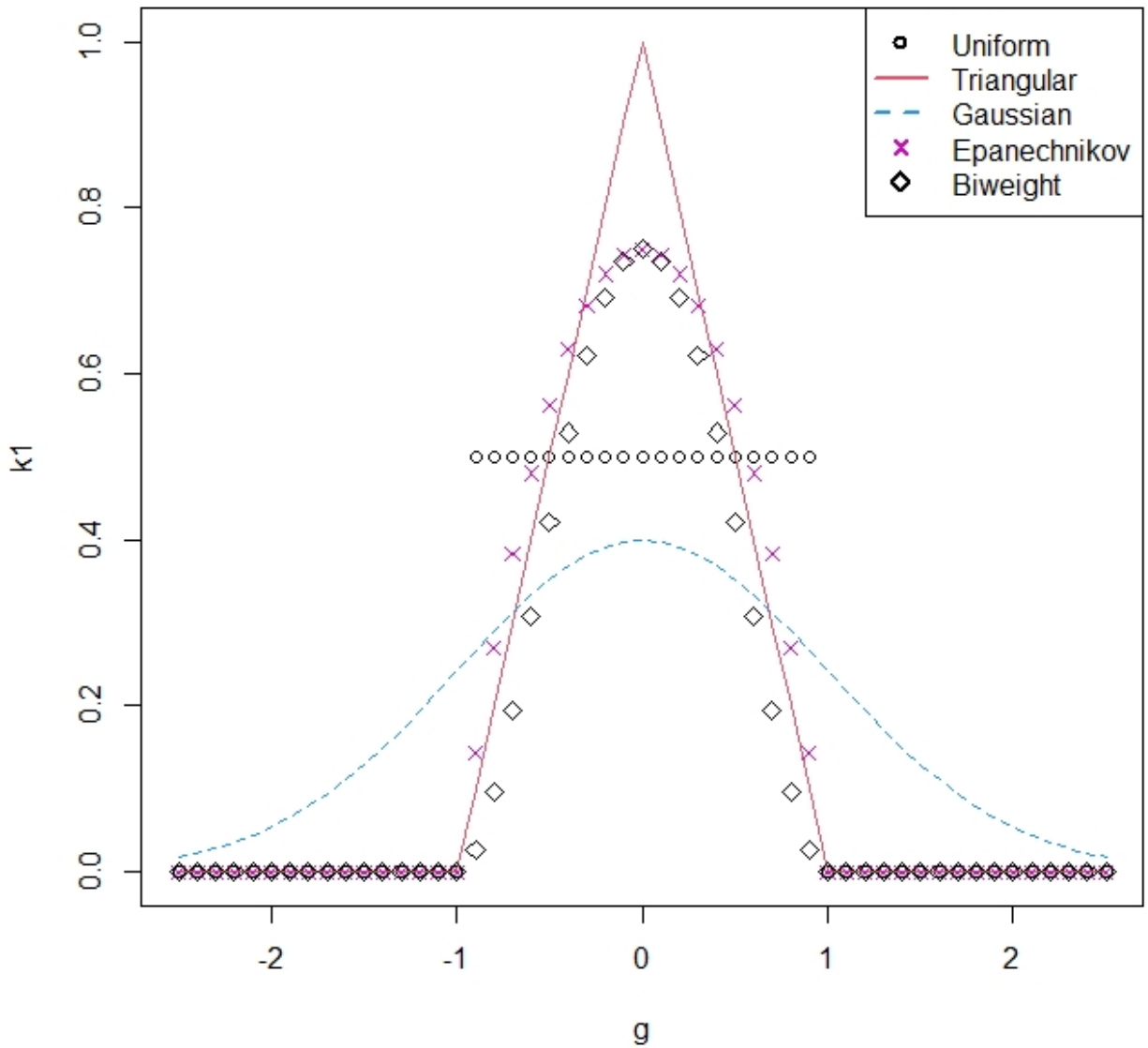
- Epanechnikov kernel

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

- Biweight kernel

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

여러 커널 함수


Uniform
Triangular
Gaussian
Epanechnikov
Biweight
k1
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
g

커널 함수의 선택

- 커널 함수의 선택에 따라 추정값이 달라지지만 전체 추정의 질을 좌우하는 정도는 아닌 것으로 알려져 있음
- Gaussian과 Epanechnikov 커널이 주로 쓰임
- Gaussian은 모든 관측치에 조금이라도 가중치를 부여하는 커널로 자료가 희소하거나 한 경우에 사용하기에 더 편리함
- 추정의 질을 향상시키기 위한 목적으로 특별한 형태의 커널을 사용하기도 함

평활모수

- 평활모수 h는 bandwidth라 하며 추정의 질을 좌우하는 매우 중요한 요소로 세심한 선택이 필요

- Uniform kernel에서의 예

  - $h = 5$: 나이차가 5세 이하인 관측치의 평균으로 추정

  - $h = 2$: 나이차가 2세 이하인 관측치의 평균으로 추정

- h를 크게 하면 멀리 있는 관측치에도 가중치를 크게 부여. $h \to \infty$면 모든 관측치에 거의 동일한 가중치를 부여하게 됨.

- h를 작게 하면 매우 가까운 관측치에만 의미 있는 가중치를 부여.

평활모수

- $h \uparrow$

  - 모형이 단순해짐

  - 상대적으로 부드러운 추정곡선 생성

  - 추정량의 편의 증가, 분산 감소

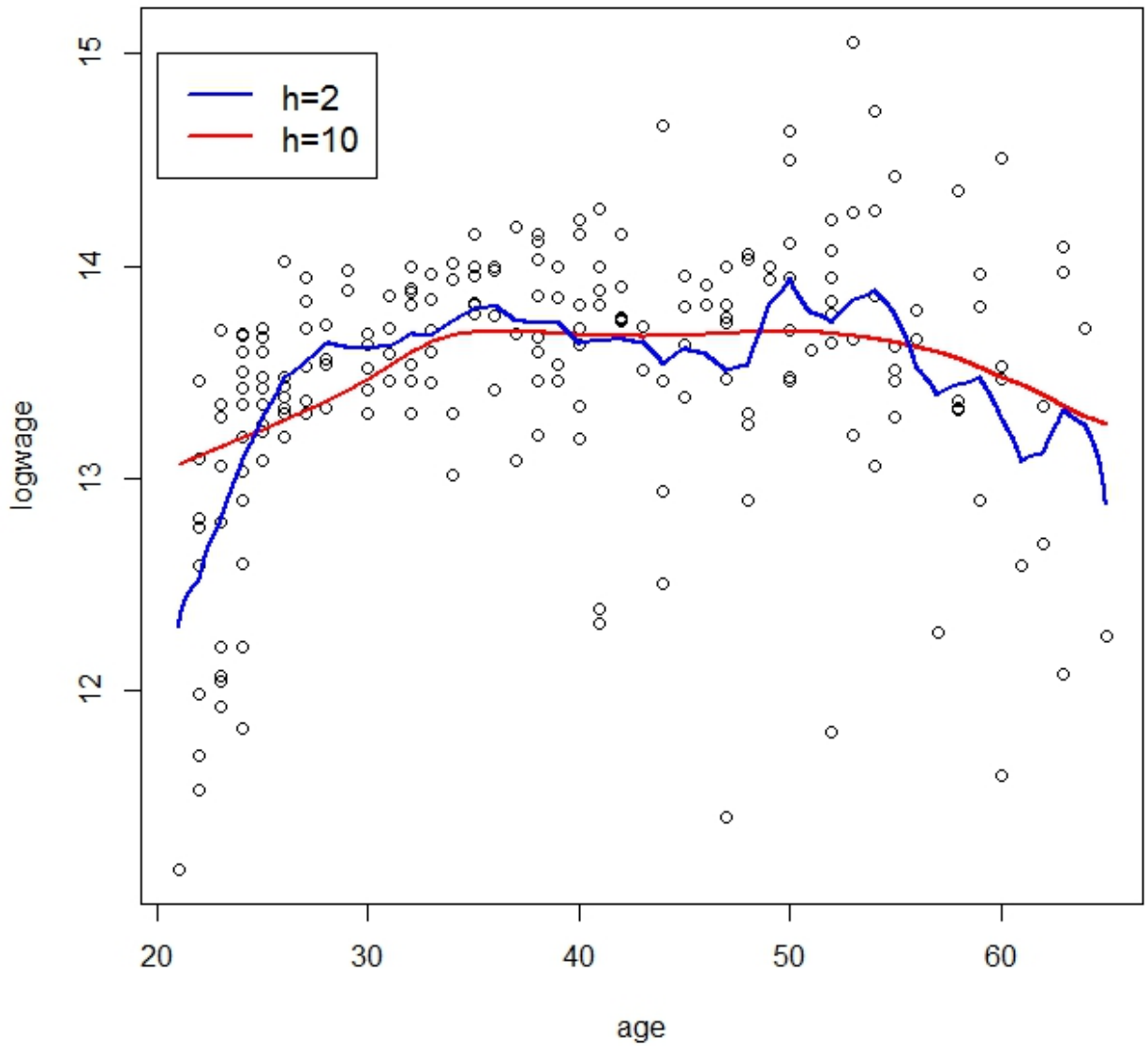
- $h \downarrow$

  - 모형이 복잡해짐

  - 자료의 특성에 민감하게 반응하는 변화가 심한 추정곡선 생성

  - 추정량의 편의 감소, 분산 증가

평활모수에 따른 추정량의 변화



평활모수의 선택

- h의 선택을 위한 다양한 방식이 제안되어 있는데 대부분 다음과 같은 MISE(Mean Integrated Squared Error)의 최소화를 목적으로 함

$$E\left[\int (\hat{g}(x) - g(x))^2 dx\right], \ g : \text{true function}$$

- 국소상수추정의 경우 이론적으로 평활모수 h는 다음을 만족하는 것이 좋음이 알려져 있음

$$h = cn^{-1/5}, \ c > 0$$

c에 대한 적절한 추정이 필요함

평활모수의 선택

- 계산능력의 향상으로 인해 이론적인 결과 대신 교차타당검증법을 이용한 평활모수의 선택도 빈번히 사용됨

- 교차타당검증 (Cross-Validation)

  - 모형의 추정을 위한 관측치와 평가를 위한 관측치를 분리하여 사용하는 방법
  - leave-one-out, K-fold 방법 등이 있음

Leave-one-out 방법

- 전체 자료를 추정을 위한 $n-1$개와 평가를 위한 1개로 분리하는 방법

  - j번째를 제외한 $n-1$개의 관측치를 이용하여 모형을 추정

  - 추정치 $\hat{Y}_j = \hat{g}(X_j)$ 생성

  - j번째 관측치의 실제 관측된 값 Y_j의 차이를 계산

$$Err_j = Y_j - \hat{Y}_j$$

  - 모든 관측치에 대하여 동일한 과정을 수행한 후 예측오차를 얻음

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n} Err_j$$

- 다양한 h에 대하여 $CV(h)$를 계산하여 최소값을 주는 h를 최종 선택

Leave-one-out 방법

- 평가를 위해 n번의 추정이 필요하여 계산량이 많음
- 어떤 경우 한 번의 추정으로 $CV(h)$ 계산이 가능하도록 한 연구결과가 있음

An Introduction to Statistical Learning, 2nd ed.

K-fold 방법

- 전체 자료를 K개로 분할하여 각 fold를 한번씩 평가에 이용하는 방법
  - k번째 fold를 제외한 나머지 관측치를 이용하여 모형을 추정
  - k번째 fold에 속한 관측치들에 대하여 추정치 생성
  - k번째 fold에서 실제 관측값과 추정치 사이의 차이를 계산하여 합함

$$Err_k$$

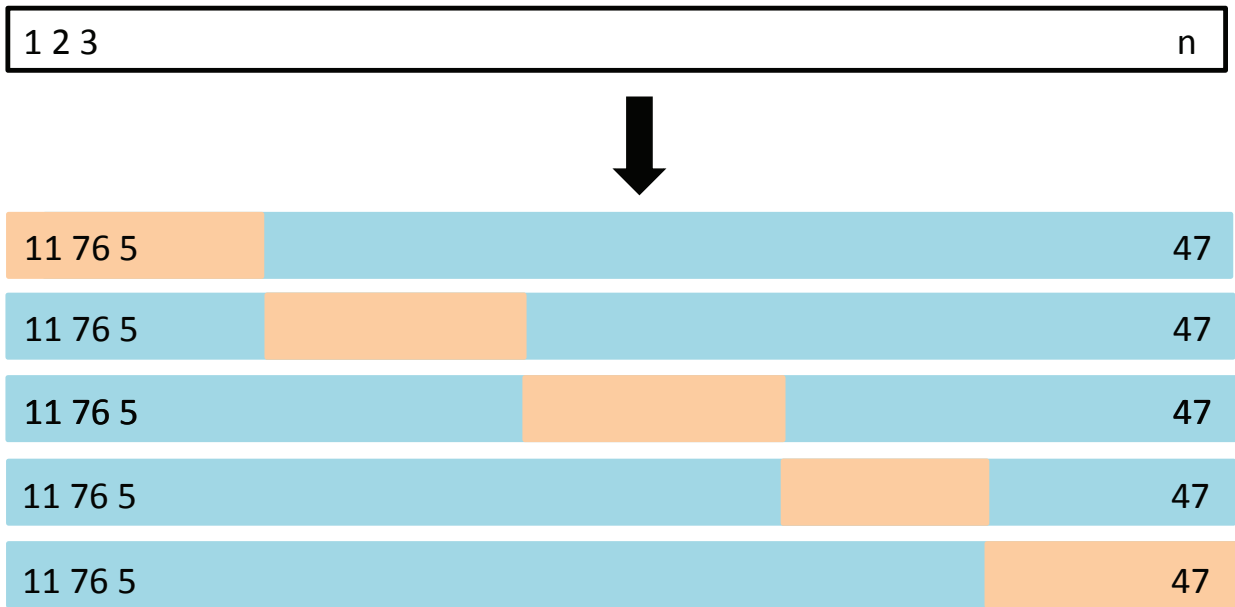
  - 모든 fold에 대하여 동일한 과정을 수행한 후 예측오차를 얻음

$$CV(h) = \frac{1}{K}\sum_{k=1}^{K} Err_k$$

- 다양한 h에 대하여 $CV(h)$를 계산하여 최소값을 주는 h를 최종 선택

K-fold 방법

- 평가를 위해 K번의 추정이 필요
- $K = n$이면 leave-one-out과 동일



An Introduction to Statistical Learning, 2nd ed.

h, g에 대한 조건

- h, g는 이론적으로 다음과 같은 조건을 보통 가정

  - g는 2차미분이 존재하며 연속

  - h는 $nh \to \infty,\ nh^8 \to 0$을 만족

- (X_j, Y_j)의 독립/동일분포조건, $\sigma^2(x) = E(u_j^2|x)$의 미분가능조건 등이 추가되기도 함

- 전술한 것과 같은 적절한 조건 하에서 커널추정량은 일치추정량이고 점근적으로 정규분포가 됨이 알려져 있음

- 조건들은 자료의 형태나 추정방식 등에 따라 조금씩 다르게 주어짐

국소상수추정의 한계

- 국소상수추정량

$$\hat{g}(x) = \frac{\sum_{j=1}^{n} Y_j \cdot K\left(\frac{X_j - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}$$

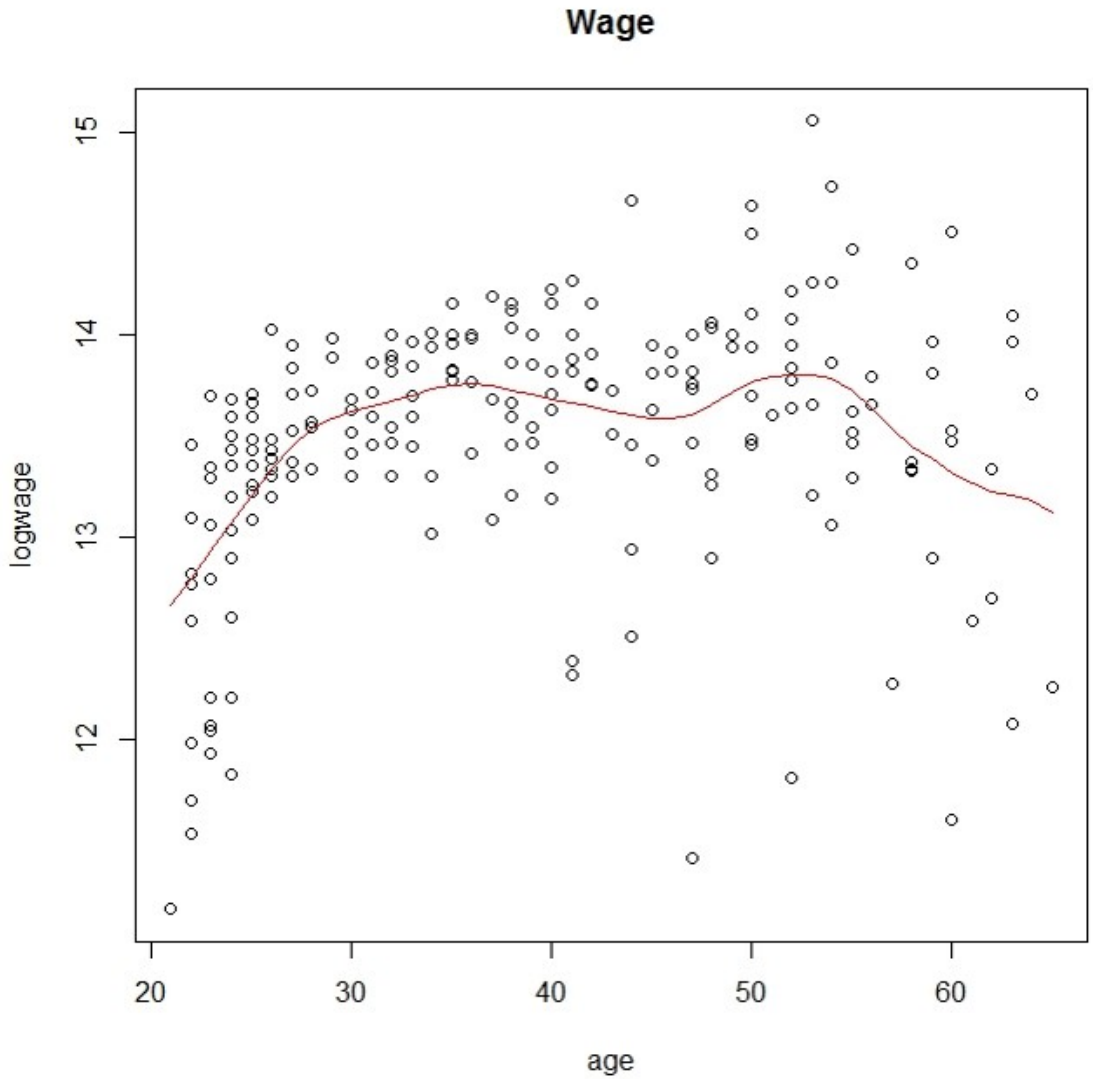
- 위 추정량은 다음 식을 최소화시키는 해임이 알려져 있음

$$\hat{g}(x) = \arg\min_{a} \sum_{j=1}^{n} (Y_j - a)^2 K\left(\frac{X_j - x}{h}\right)$$

- 즉, x 근처에서 국소적으로 $g(u) \approx a$로 모형화
- 가중평균의 틀에서 쉽게 이해할 수 있고 적용이 간단하나 함수의 경계 근처에서 편의가 발생할 수 있음이 알려져 있음

Boundary effect

- 다음과 같이 함수의 경계에서 추세가 존재하는 경우 편의 발생



국소선형 (local linear) 추정

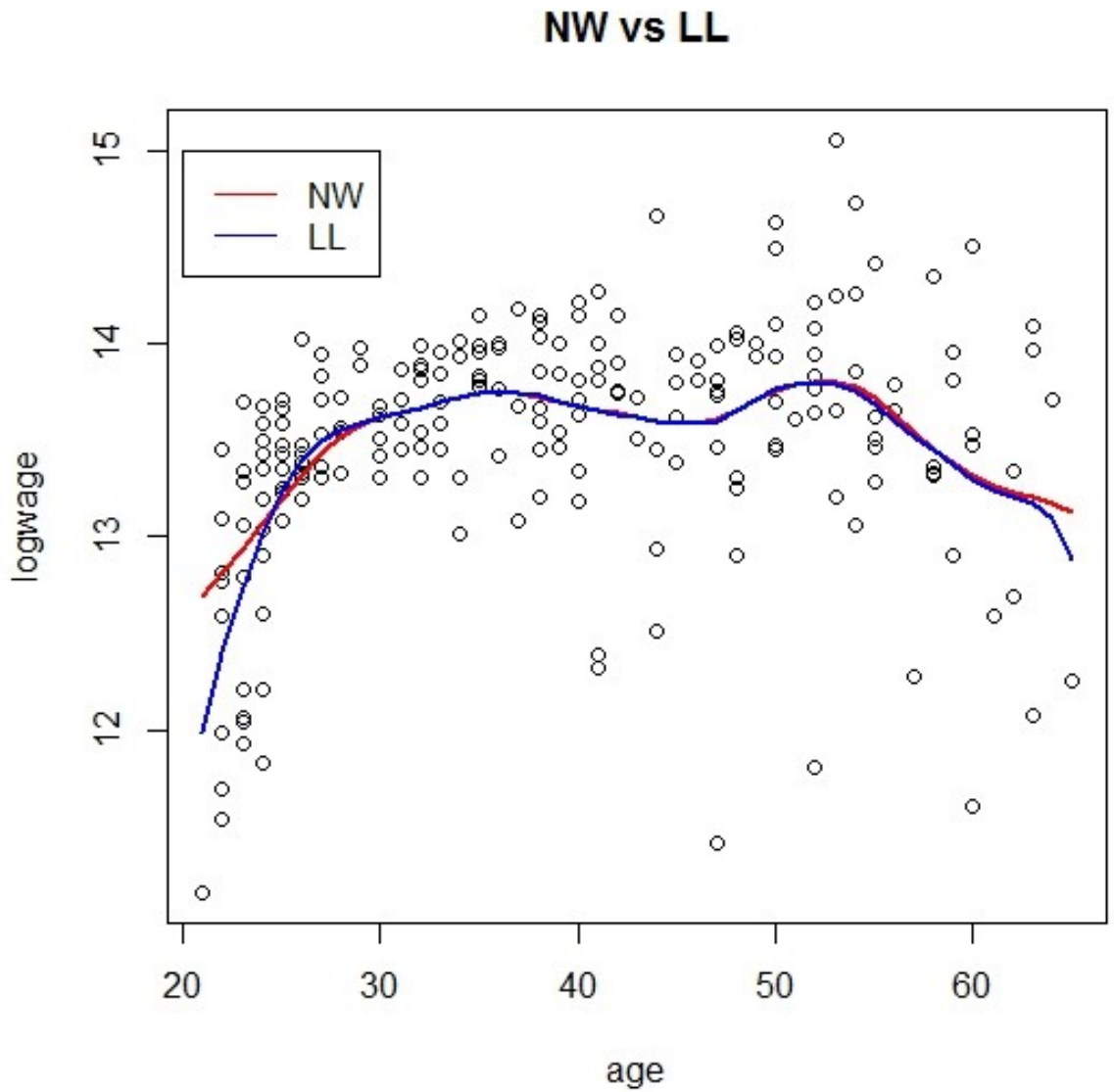
- 국소선형추정량 : 다음 식을 최소화시키는 해로 정의

$$\hat{g}(x) = \arg\min_{a} \sum_{j=1}^{n} (Y_j - a - b(X_j - x))^2 K\left(\frac{X_j - x}{h}\right)$$

- 즉, x 근처에서 국소적으로 $g(u) \approx a + b(u - x)$로 선형 모형화
- $\hat{a} = \hat{g}(x)$이고 $\hat{b} = \hat{g}'(x)$
- 국소선형추정의 경계에서의 편의 문제를 해결
- 커널함수와 평활모수의 선택은 국소선형추정과 동일한 방식

추정량 비교

- 국소상수(NW) vs 국소선형(LL)



추정량 비교

- 경계 부분이 아닌 곳에서는 비슷한 양상을 보임
- 추정 대상이 거의 평평한 함수일 경우 국소상수추정이 더 우수
- 추정량의 분산은 동일함

퇴직률 추정

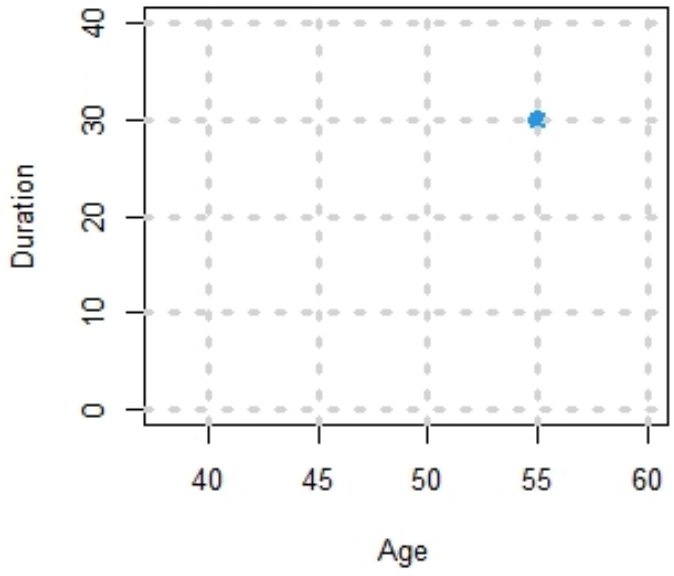
- 과거에는 영향을 미치는 요인들을 통제 후 조합별 자료를 이용하여 추정하였음
- 조합별 자료가 충분하지 않은 경우 분산이 크게 나타남
- 성별/직종별로 구분 후 퇴직률을 연령 및 재직기간의 함수로 가정 후 커널추정법 적용

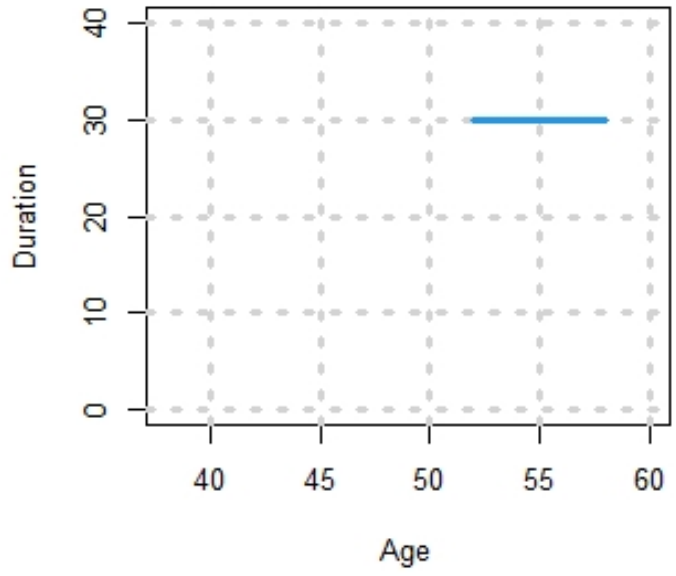
$$\text{퇴직률} = g(\text{연령},\text{재직기간}) + u$$

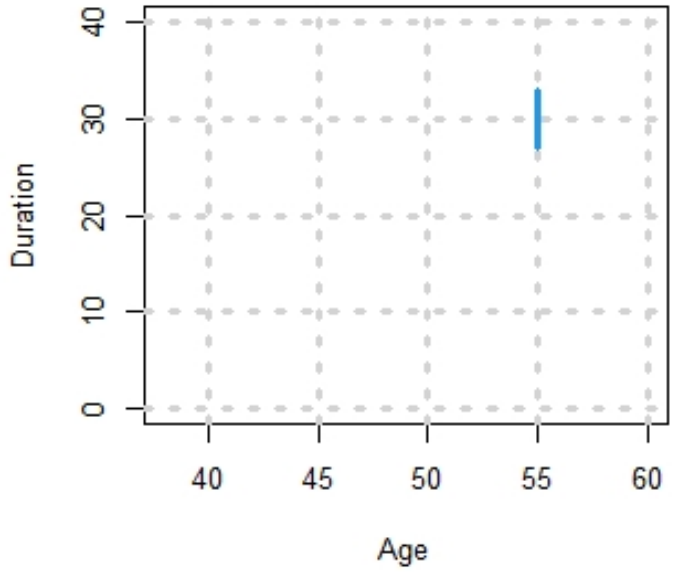
- 연령 혹은 재직기간 중 하나를 고정 후 다른 요인의 함수로 추정하는 것도 가능

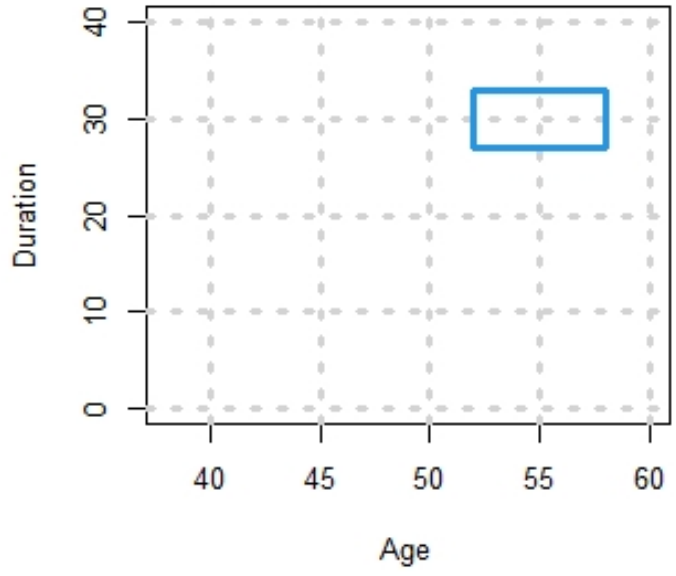
퇴직률 추정

- 55세
- 30년 재직









다변량으로의 확장

- 설명변수가 복수인 경우 회귀함수는 다변량함수가 된다.

$$Y_j = g(X_{1j}, X_{2j}, \ldots, X_{pj}) + u_j$$

- 다변량 국소상수/국소선형 추정 ($p = 2$)

  - 국소상수 : 다음 함수의 최소화 문제의 해 $\hat{a} = \hat{g}(x_1, x_2)$

$$\sum_{j=1}^{n} (Y_j - a)^2 K\left(\frac{X_{1j} - x_1}{h_1}\right) K\left(\frac{X_{2j} - x_2}{h_2}\right)$$

  - 국소선형 : 다음 함수의 최소화 문제의 해 중 $\hat{a}$

$$\sum_{j=1}^{n} (Y_j - a - b_1(X_{1j} - x_1) - b_2(X_{2j} - x_2))^2 K\left(\frac{X_{1j} - x_1}{h_1}\right) K\left(\frac{X_{2j} - x_2}{h_2}\right)$$

다변량으로의 확장

- 설명변수의 개수 p에 따라 평활모수의 선택을 위한 계산량이 늘어남
- 위에서 소개한 커널은 product 형태로 사각형 모양의 관측치 영역을 선택하게 된다. 다른 형태의 커널을 도입하여 영역을 수정할 수 있다.
- p가 커짐에 따라 자료의 부족현상이 심해진다. 이를 해결하기 위해 평활모수를 크게 잡거나 가법모형의 이용 등을 고려할 수 있다.

$$\text{가법모형} : g(x_1, x_2) = g_1(x_1) + g_2(x_2)$$