

Moving Beyond Linearity

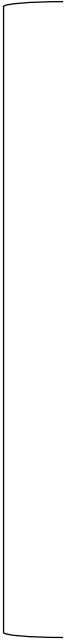
An Introduction to Statistical Learning

황성원

Nonlinear World!

비선형적 접근법 종류

단일 입력 변수 X - 출력 Y

- 
1. 다항회귀
 2. 조각별 상수함수
 3. 회귀 스플라인
 4. 평활 스플라인
 5. 국소회귀

다중 입력 변수 X_1, X_2, \dots, X_p - 출력 Y

6. 일반화가법모델

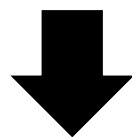
1. 다항회귀

(Polynomial Regression)

메인 아이디어 - 다항회귀

: 적합 모델 변경 : 1차 다항식 (선형) → 고차 다항식 (비선형)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

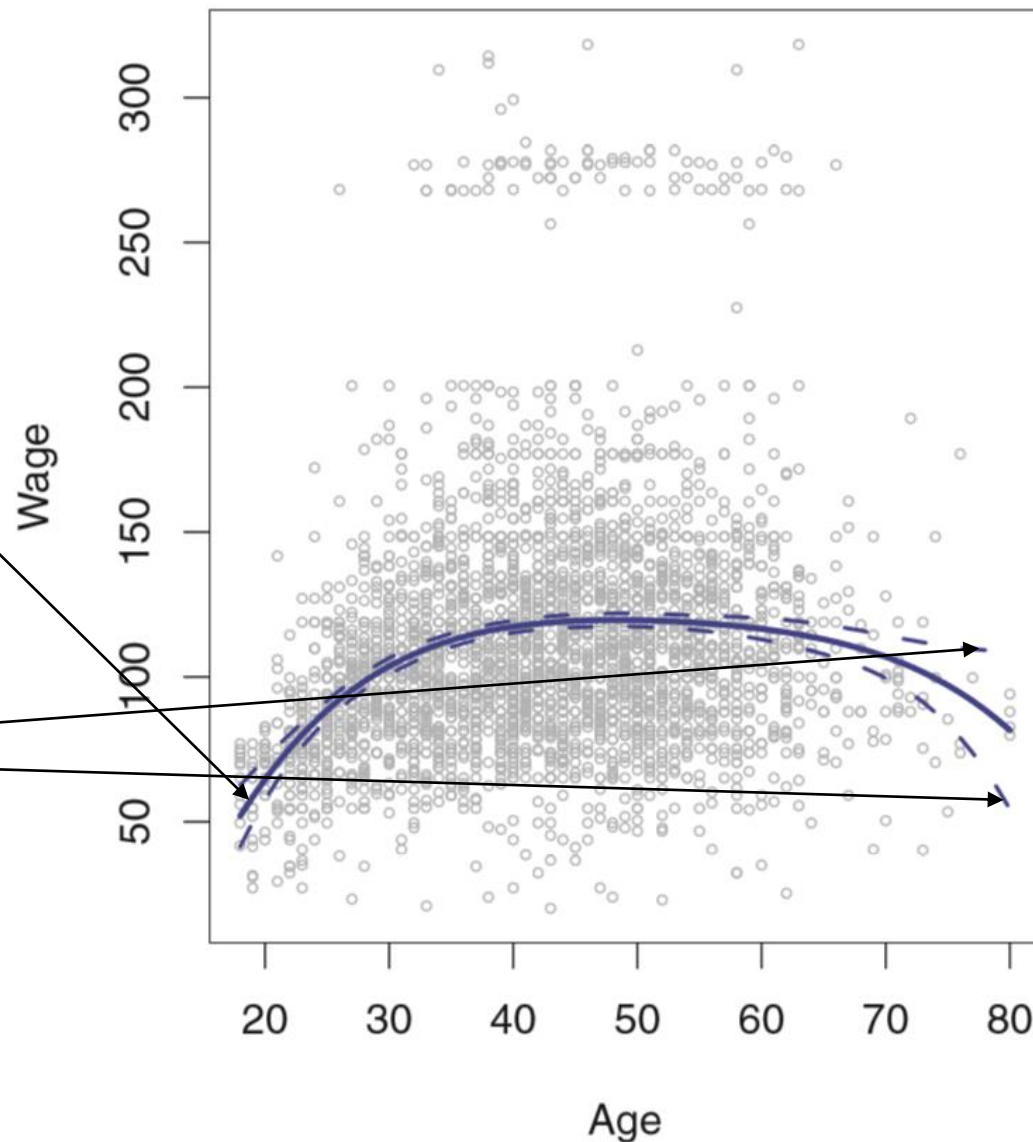


$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \dots + \beta_d x_i^d + \epsilon_i$$

회귀에서...

4차 다항식 Fitting Line

정규분포 95% 신뢰구간



$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \dots + \beta_d x_i^d + \epsilon_i$$

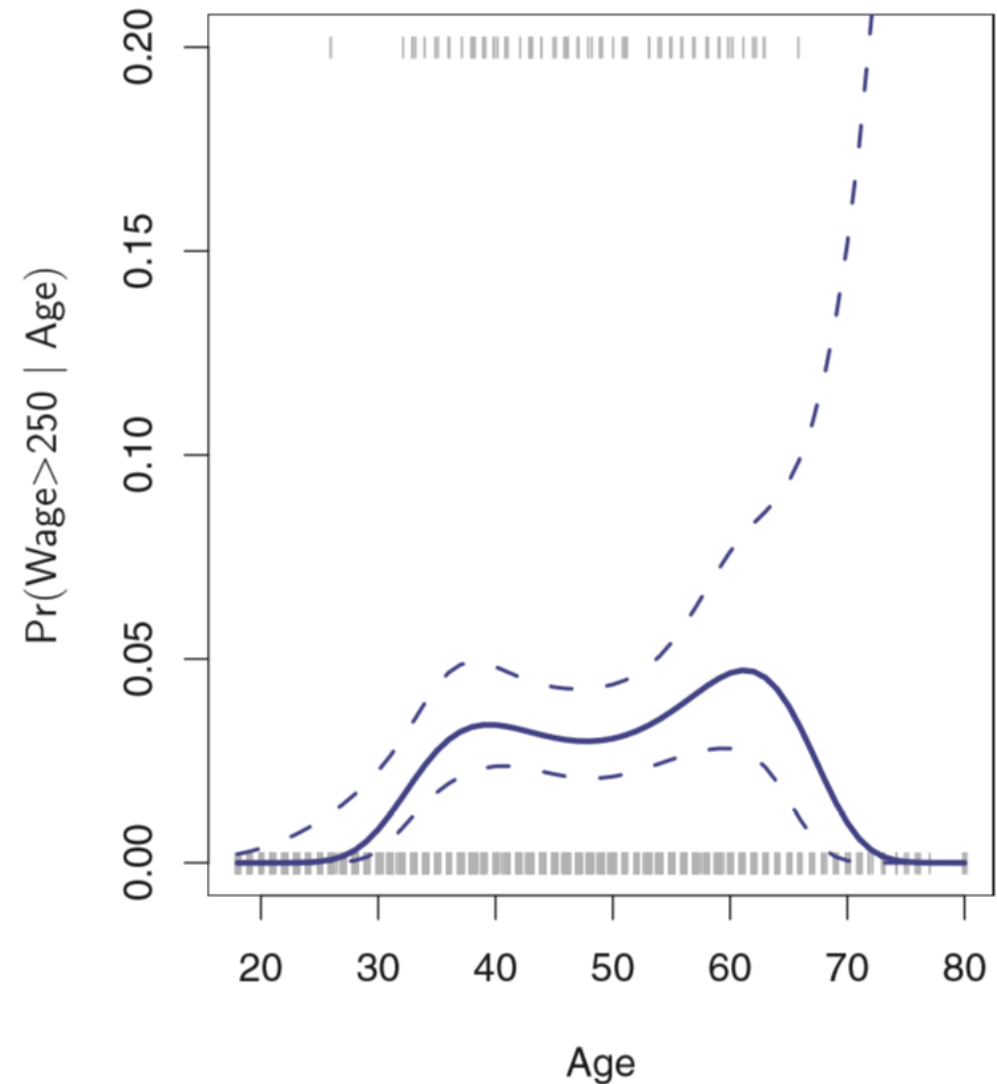
분류에서...

임금이 250 이상(고소득자)일 확률을 가능도로 지정!

전체 표본 수(n) = 3,000명

고소득자 수 = 79명

따라서, 추정된 계수의 분산이 크고, 신뢰구간이 넓다.

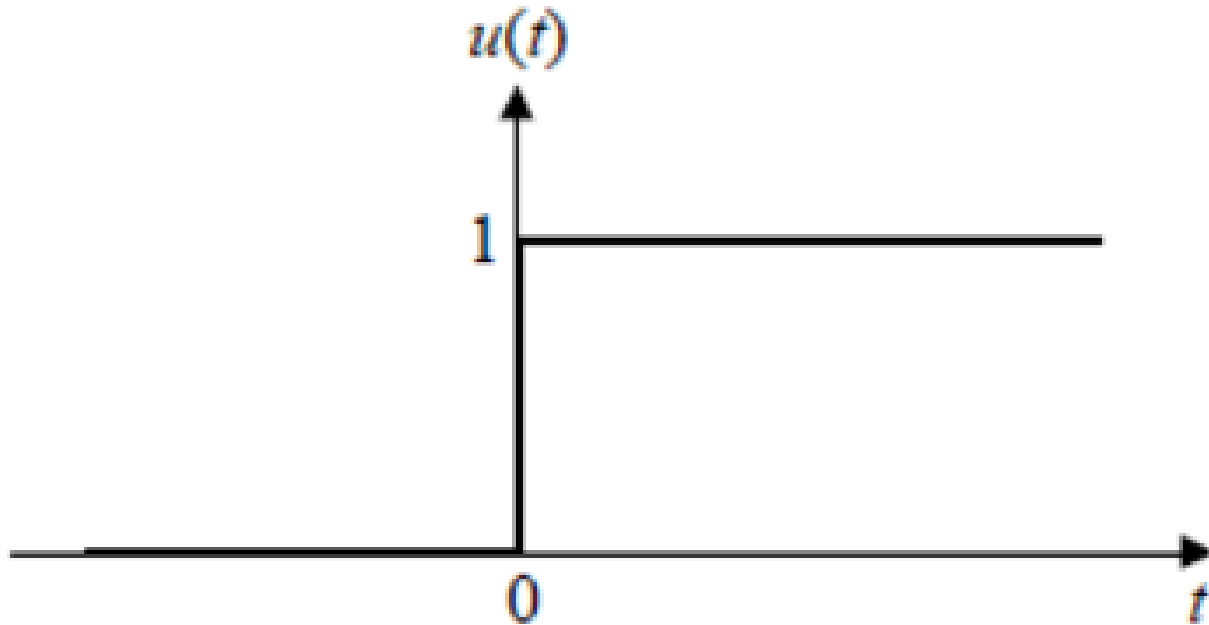


$$\Pr(y_i > 250 | x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d)}$$

2. 조각별 상수함수

(Piecewise Constant)

계단함수 (Step Function)

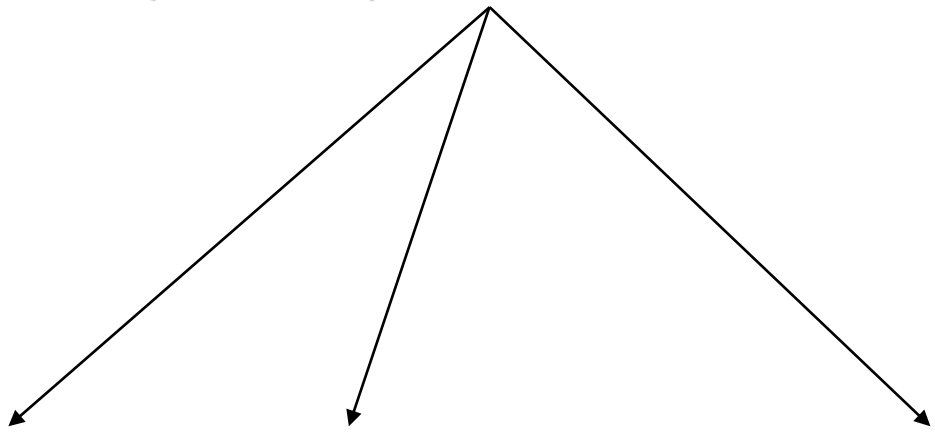


$$u(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases}$$

메인 아이디어 - 조각별 상수함수

: 입력 x 를 K 개의 구간으로 조각 내어 상수 값으로 Fitting 한다!

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



The diagram consists of three arrows originating from a single point below the first equation and pointing to the corresponding terms in the second equation: β_0 , $\beta_1 C_1(x_i)$, and $\beta_K C_K(x_i)$.

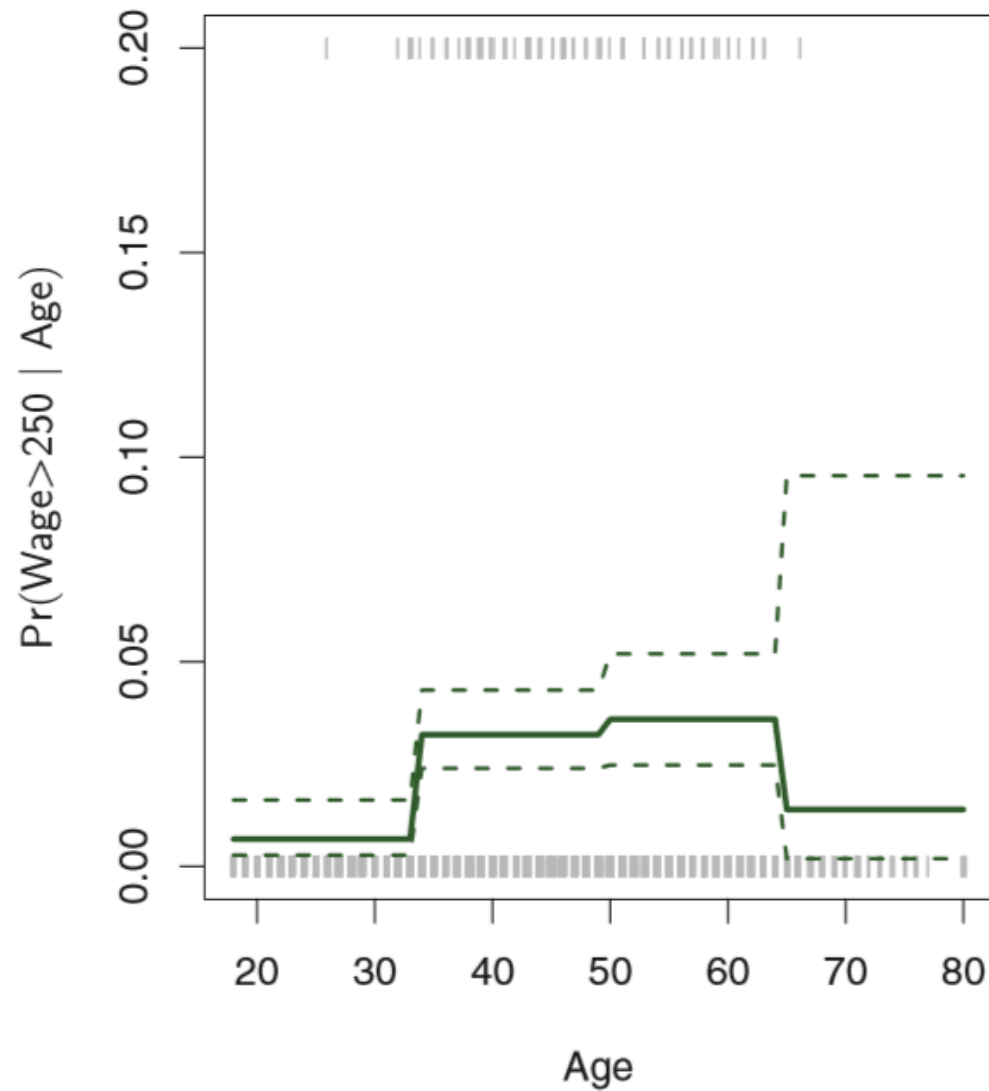
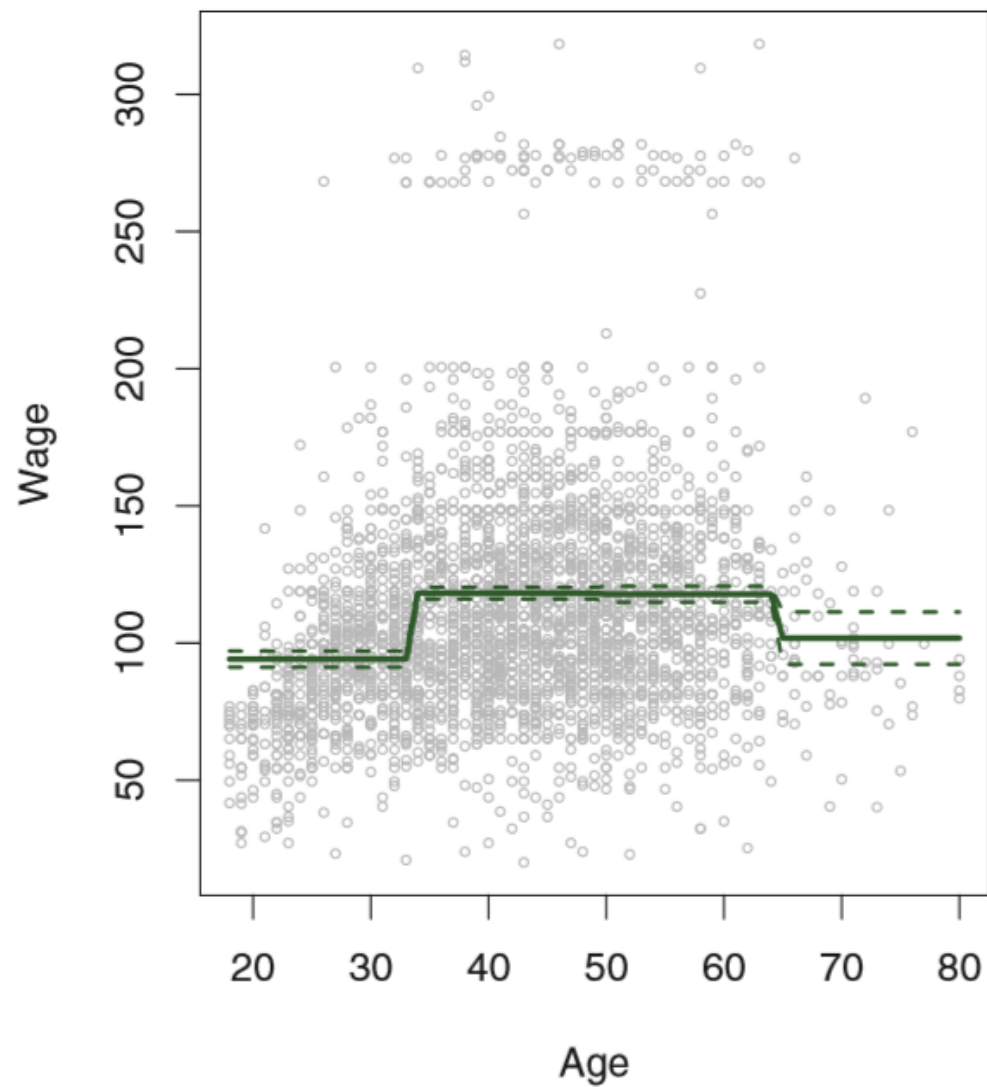
$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \dots + \beta_K C_K(x_i) + \epsilon_i$$

조각 내는 방법 (계단함수 사용)

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \dots + \beta_K C_K(x_i) + \epsilon_i$$

$$\begin{aligned} C_0(X) &= I(X < c_1), \\ C_1(X) &= I(c_1 \leq X < c_2), \\ C_2(X) &= I(c_2 \leq X < c_3), \\ &\vdots \\ C_{K-1}(X) &= I(c_{K-1} \leq X < c_K), \\ C_K(X) &= I(c_K \leq X), \end{aligned}$$

회귀 / 분류 결과



3. 회귀 스플라인

(Regression Spline)

기저함수 (Basis Function) 방법론

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \dots + \beta_K b_K(x_i) + \epsilon_i$$

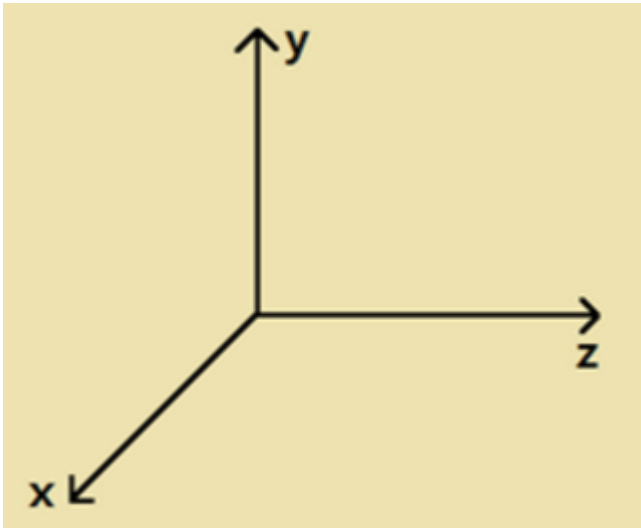
일반적인 x 를 입력 변수로 쓰는 것이 아닌,

x 에 어떠한 함수를 취한 것을 x 로 쓴다!

기저벡터 (Basis Vector)

- 기저벡터 -

$$\hat{j} = (0, 1, 0)$$



$$\hat{i} = (1, 0, 0)$$

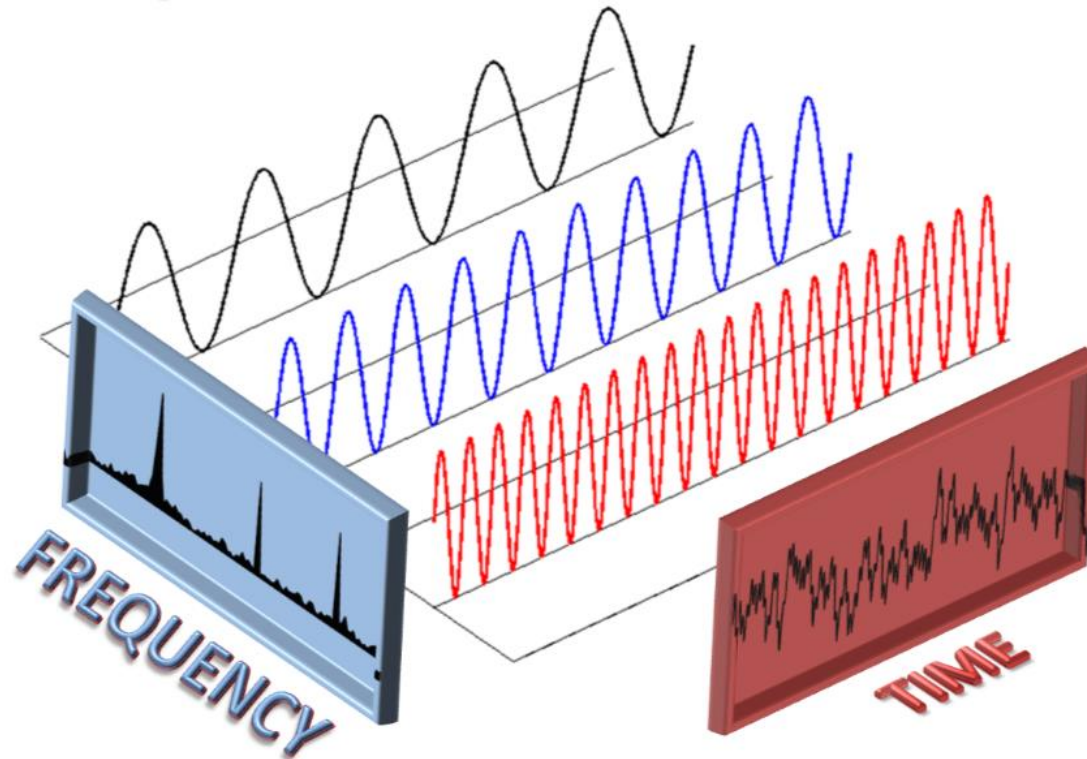
$$\hat{k} = (0, 0, 1)$$

$$V = 4\hat{i} + 6\hat{j} - 2\hat{k}$$

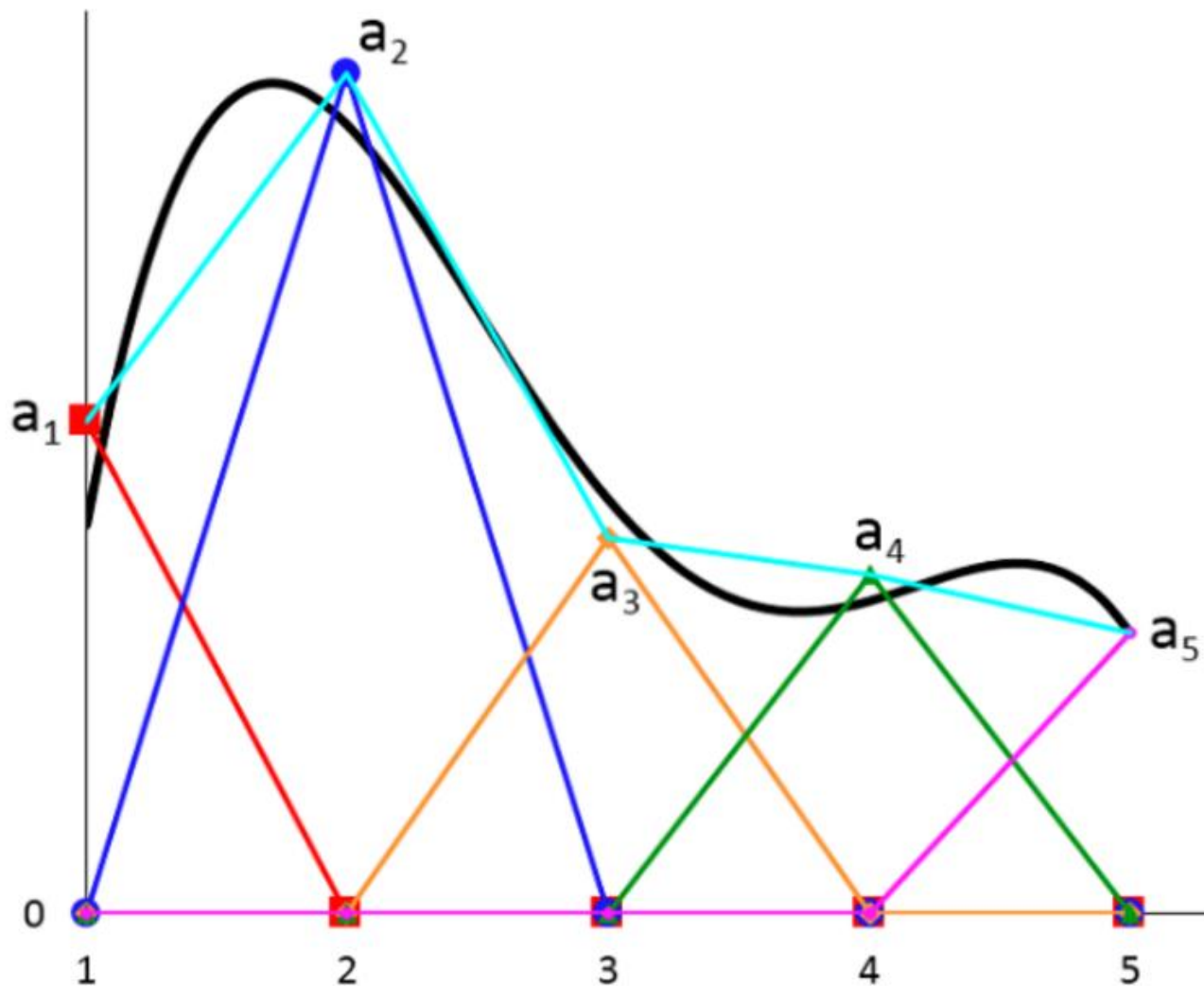
기저함수 (Basis Function) 예: 푸리에급수

$$f(x) = A_0 + A_1 \cos x + B_1 \sin x + A_2 \cos 2x + B_2 \sin 2x + \dots$$

$$= \frac{A_0}{2} + \sum_{n=1}^{\infty} \left[A_n \cos \frac{n\pi}{L} x + B_n \sin \frac{n\pi}{L} x \right]$$



여러 가지 기저 함수 적용 형태



다항회귀의 경우

$$b_j(x_i) = x_i^j$$

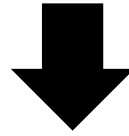
조각별 상수함수의 경우

$$b_j(x_i) = I(c_j \leq x_i < c_{j+1})$$

조각별 상수가 아니라 조각별 다항식!

일반 3차 다항식

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$$



c를 기준으로 조각난 3차 다항식

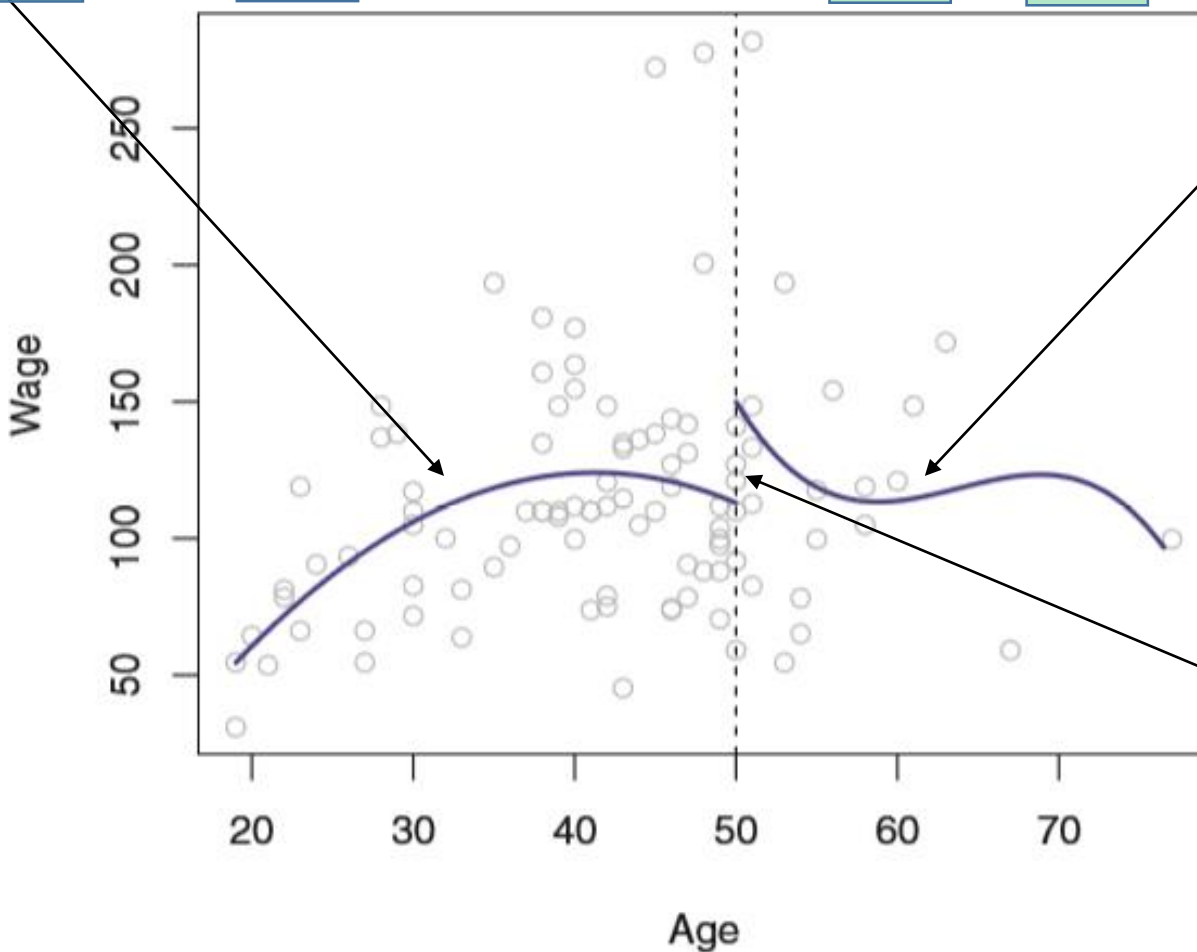
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

조각별 상수가 아니라 조각별 다항식!

$$\beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i$$

$$\beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i$$

자유도: 8



매듭(knots)

연속 조각별 다항식

불연속

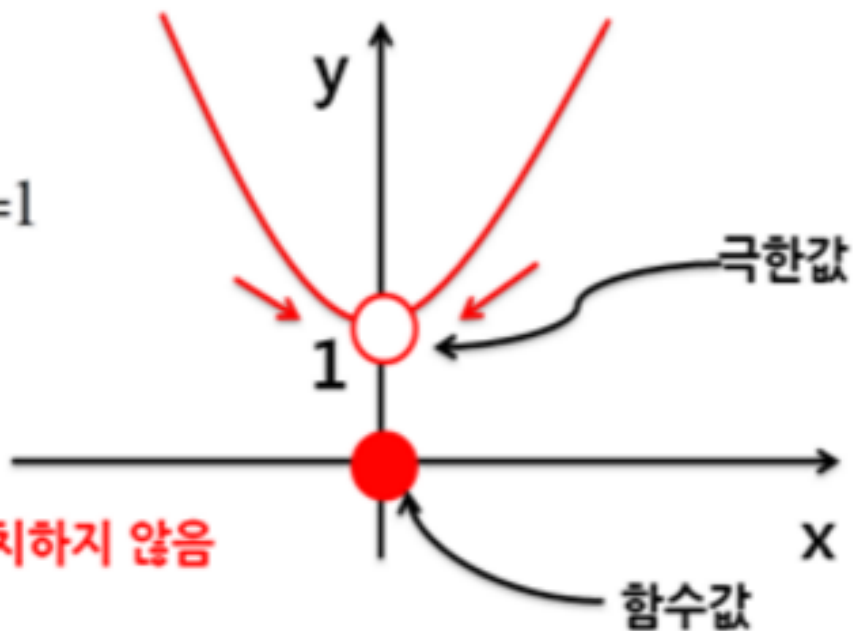
$f(x)$ 는 $x=a$ 에서 불연속이 된다.

(1) $f(0)=0$

(2) $\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} f(x) = 1$

(3) $f(0) \neq \lim_{x \rightarrow 0} f(x)$

함수값과 극한값이 일치하지 않음

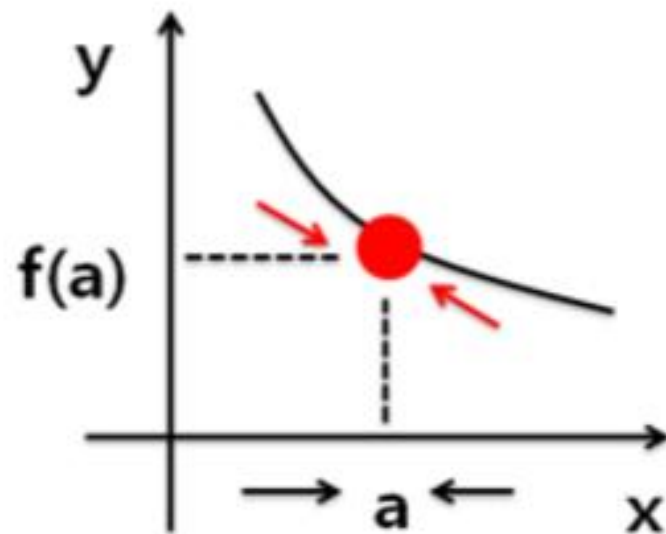


(1) $f(a)$ 값이 존재

(2) $\lim_{x \rightarrow a} f(x)$ 존재

$$\Leftrightarrow \lim_{x \rightarrow a+0} f(x) = \lim_{x \rightarrow a-0} f(x)$$

(3) $f(a) = \lim_{x \rightarrow a} f(x)$



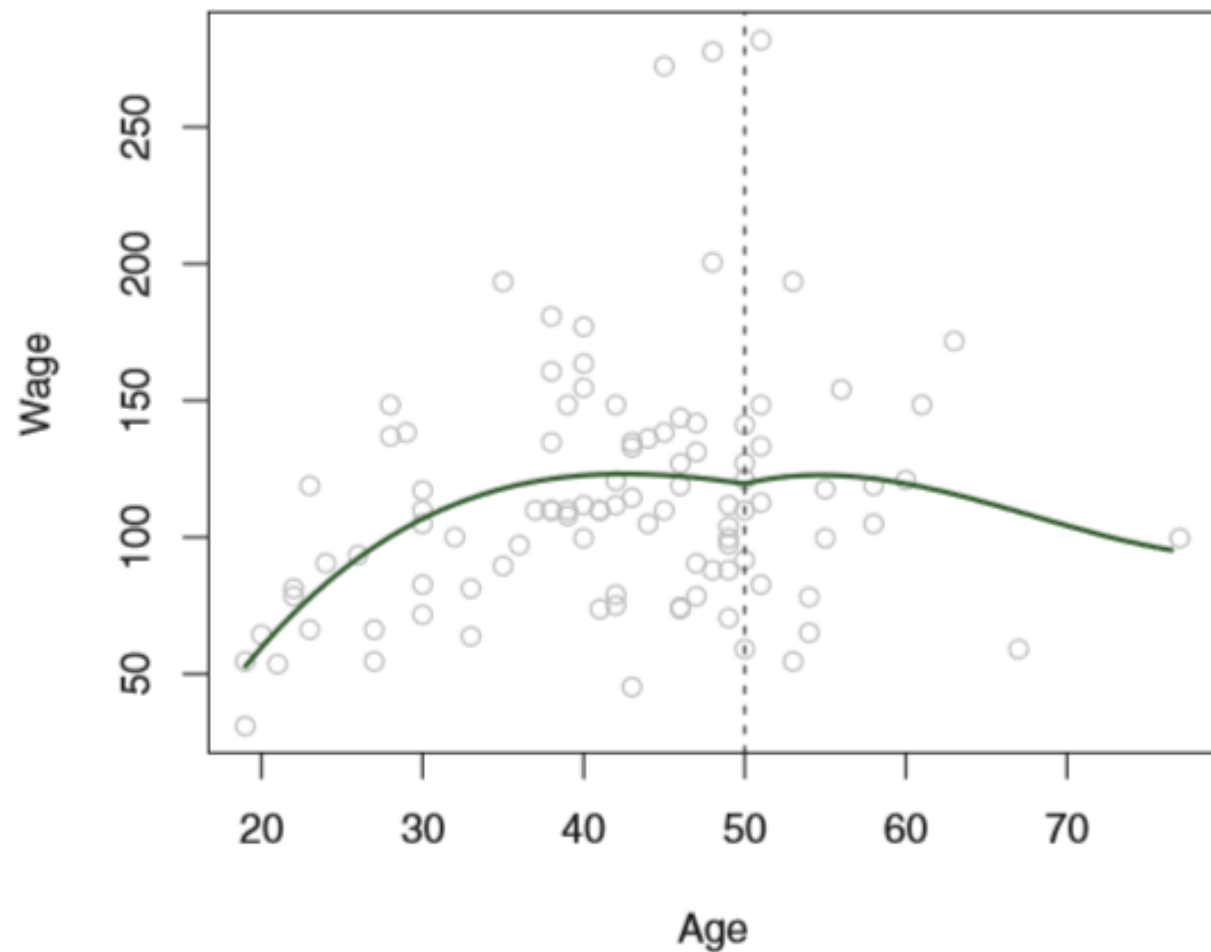
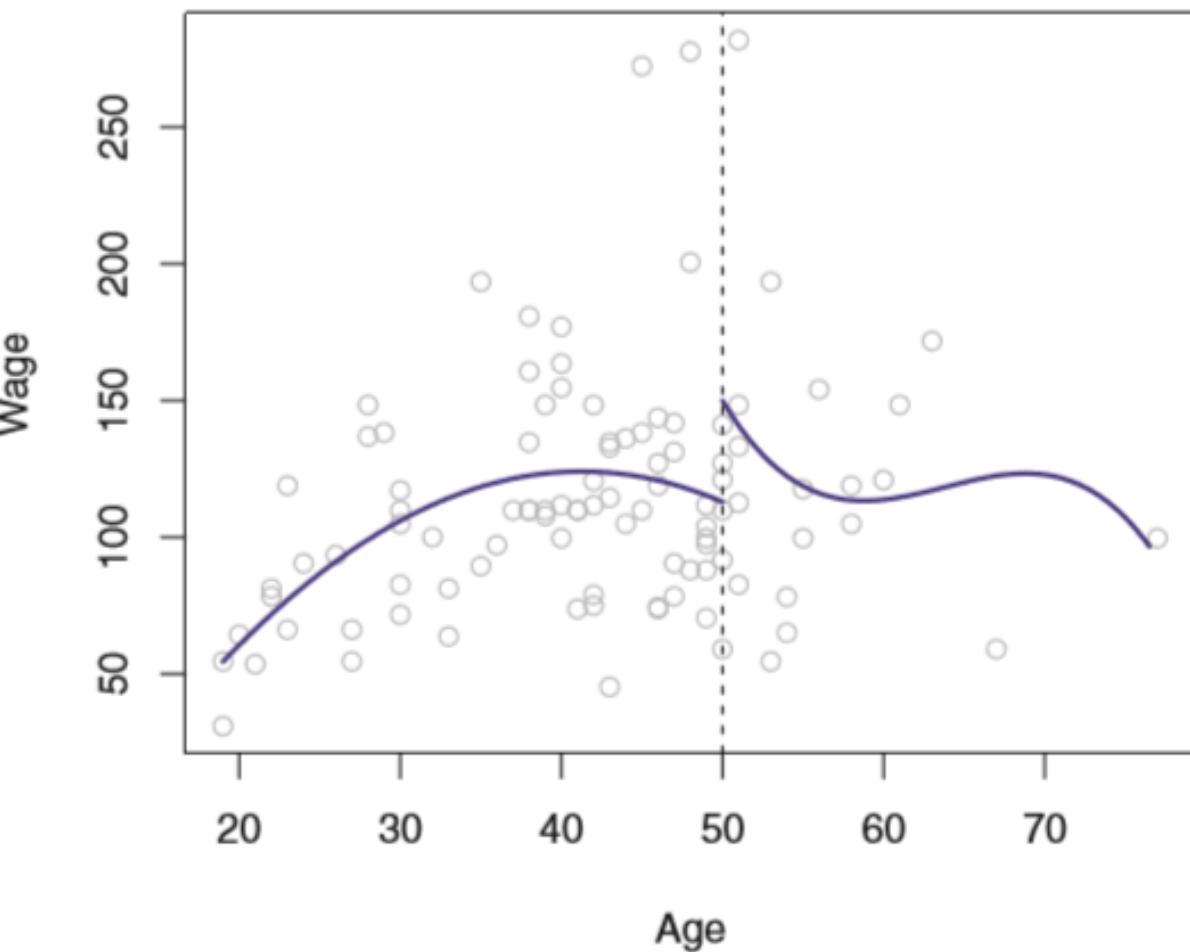
연속

연속 조각별 다항식

Piecewise Cubic

+ 연속조건 =

Continuous Piecewise Cubic

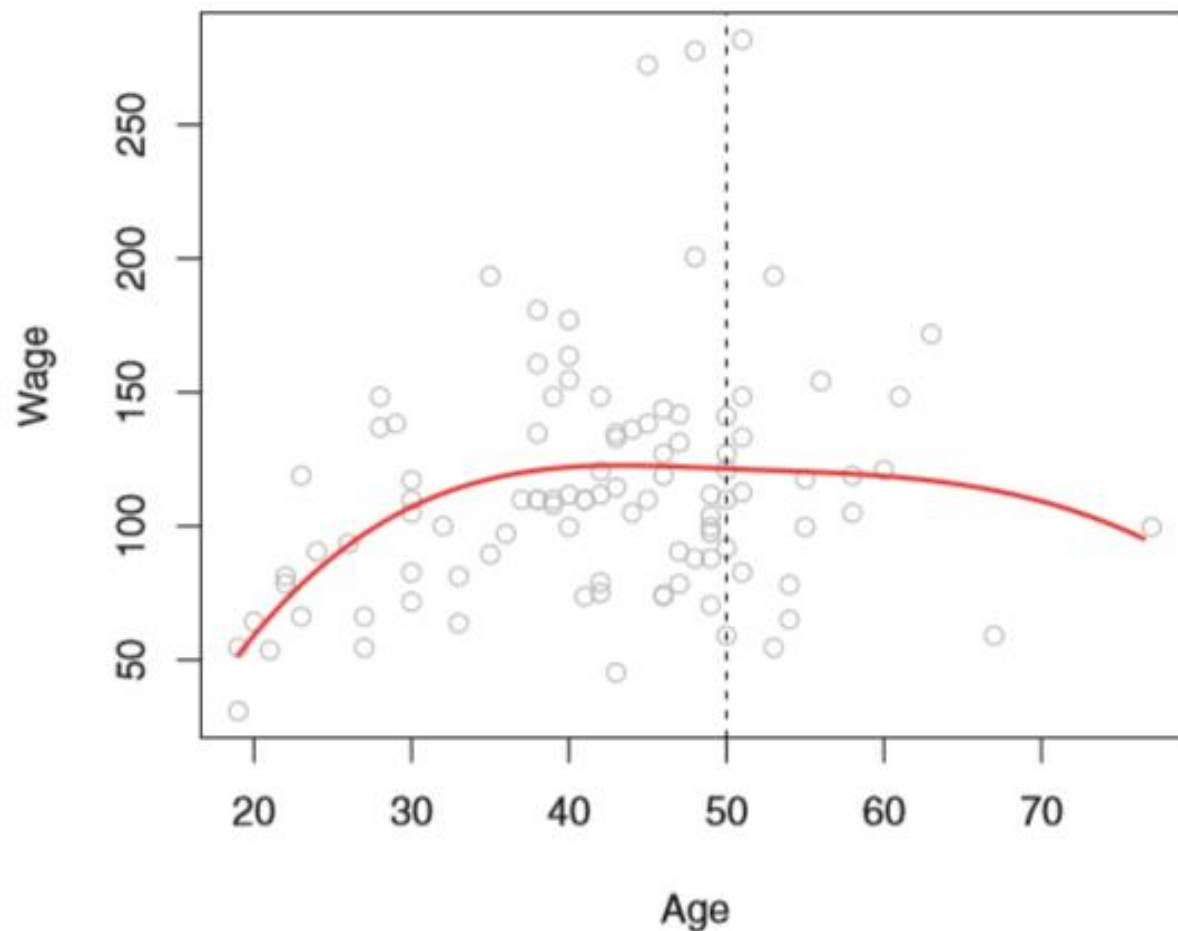
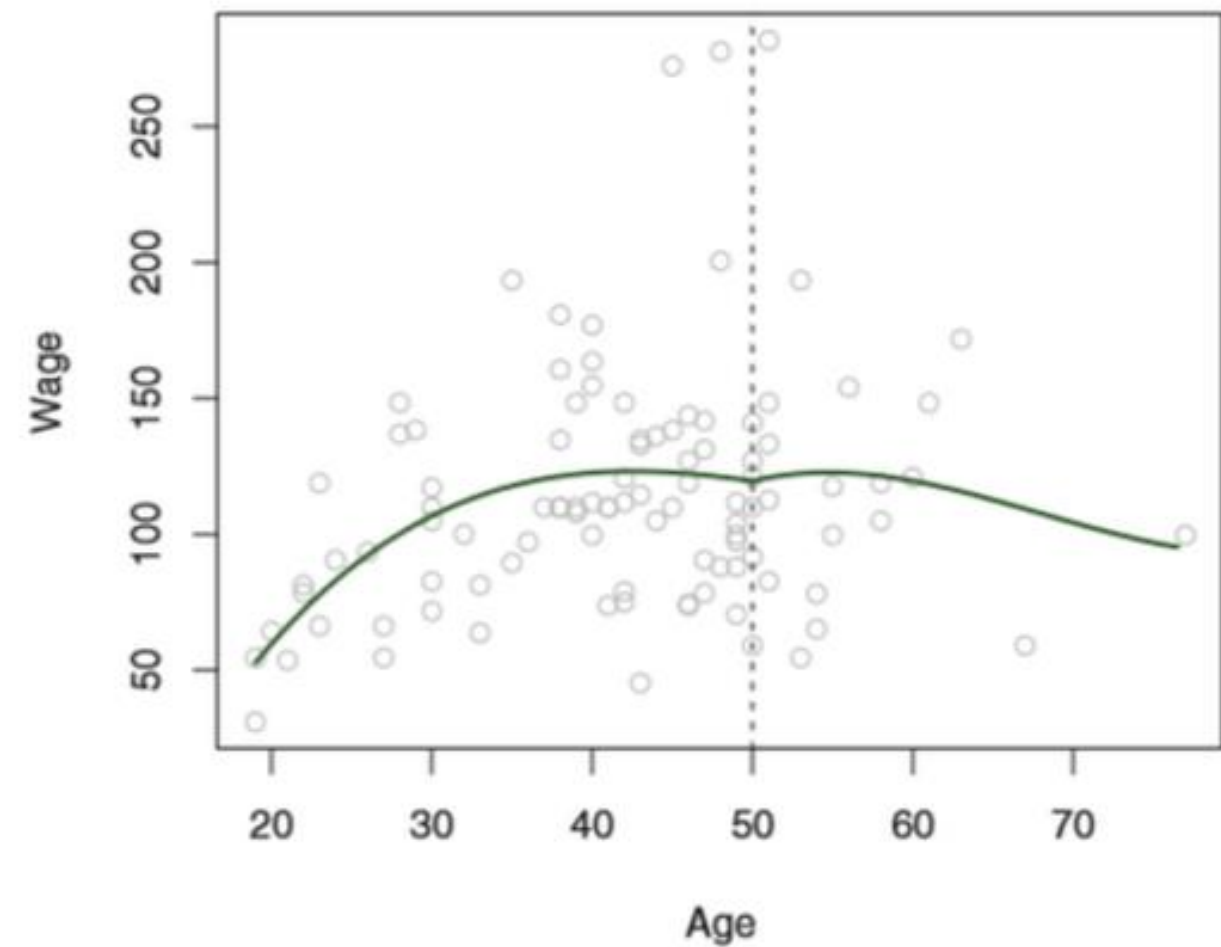


삼차 스플라인 (Cubic Spline)

Continuous Piecewise Cubic

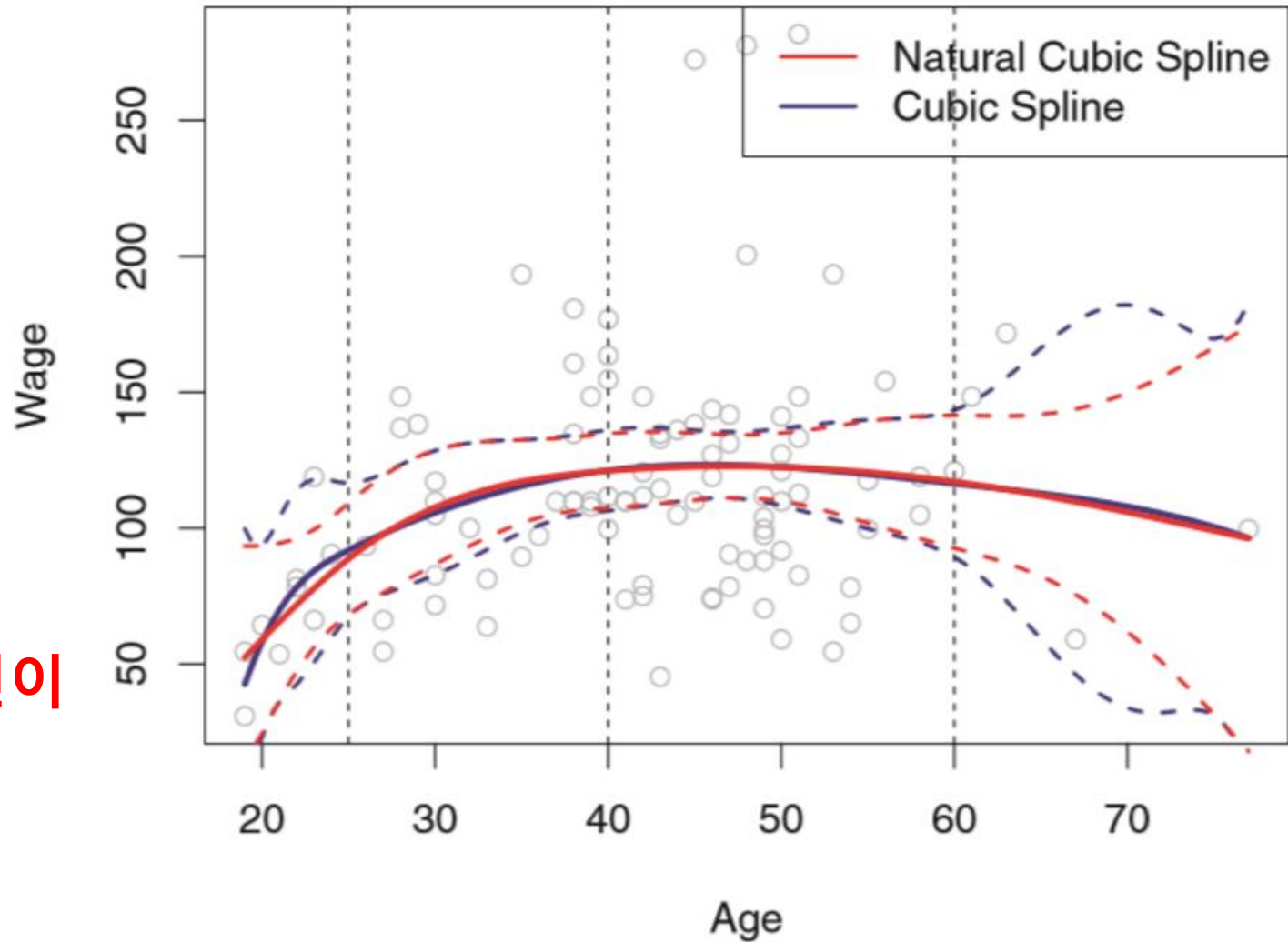
+ 1차 미분/2차 미분 연속 =

Cubic Spline



자연 삼차 스플라인 (Natural Cubic Spline)

제한 조건이 많아지면,
자유도를 낮추고,
이는 분산 감소로
신뢰구간의 안정성 확보
따라서, 자연 삼차 스플라인이
좋다!

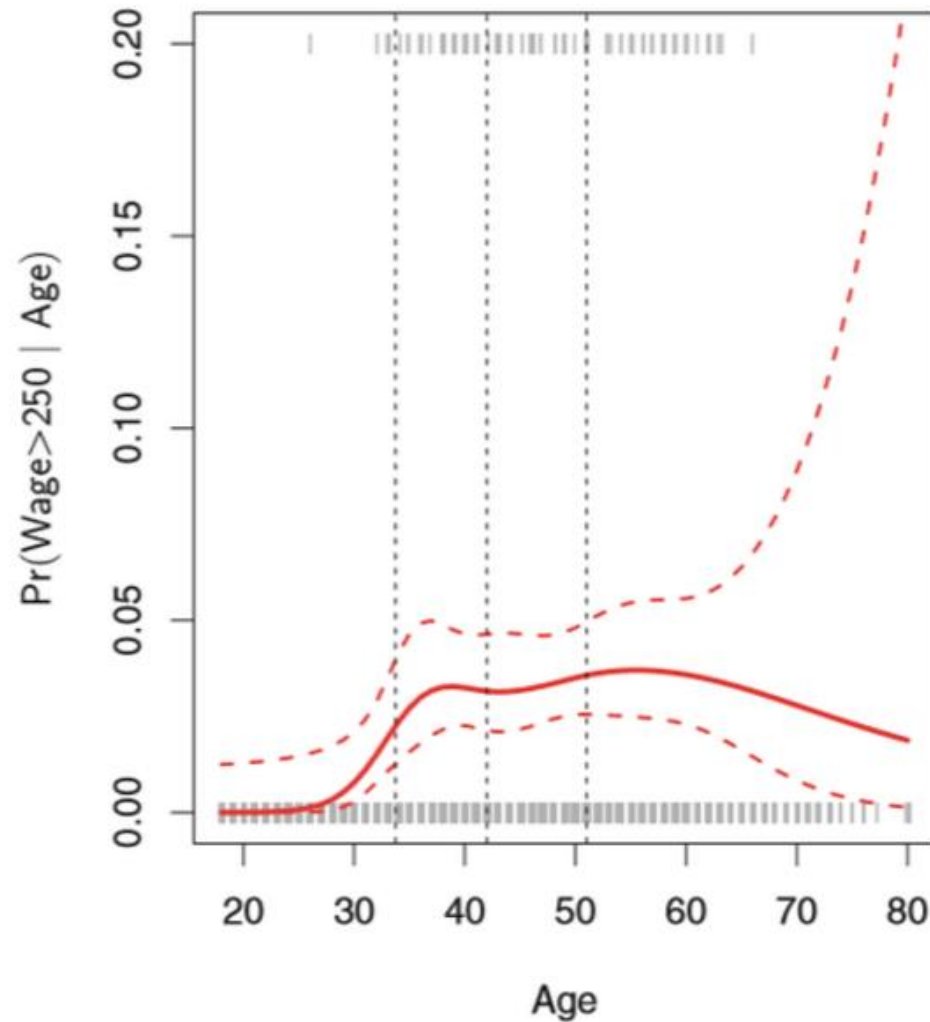
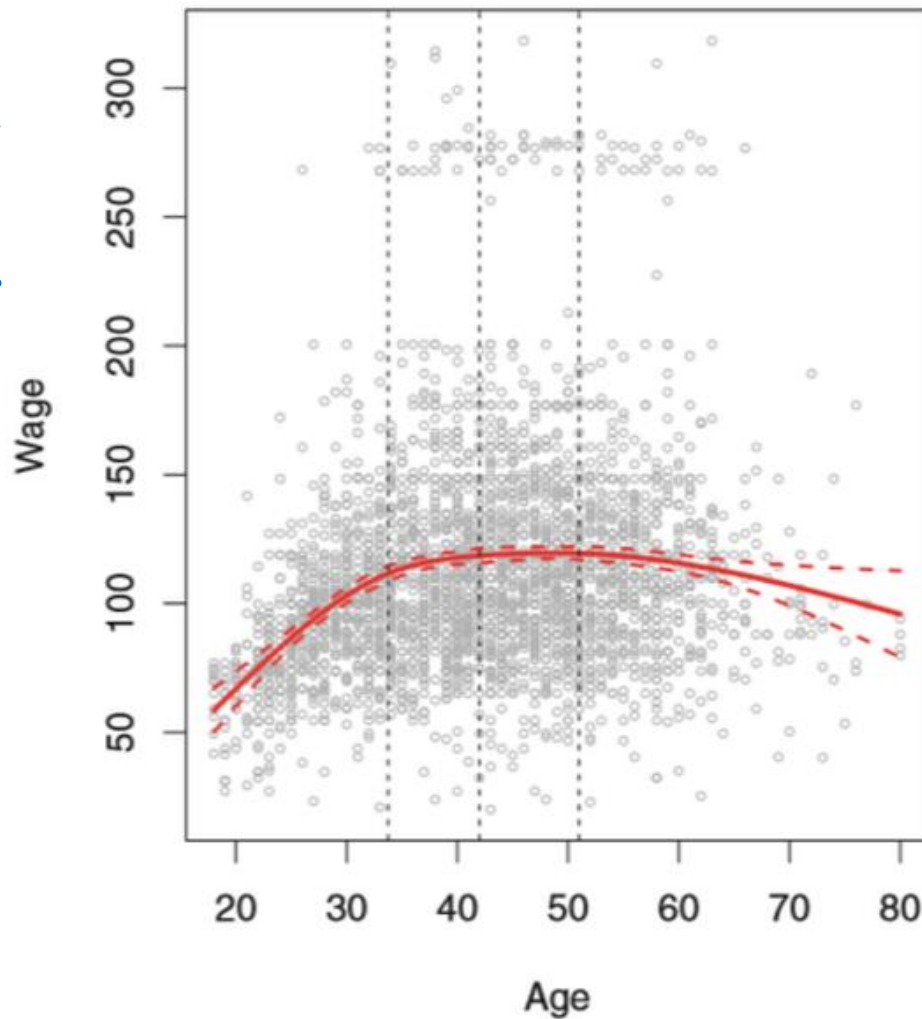


삼차 스플라인 + 끝자락 선형 가정 = 자연 삼차 스플라인

매듭 수와 위치 선택

매듭이 많으면
유연성이 커지고,
계산량도 커진다.

Natural Cubic Spline

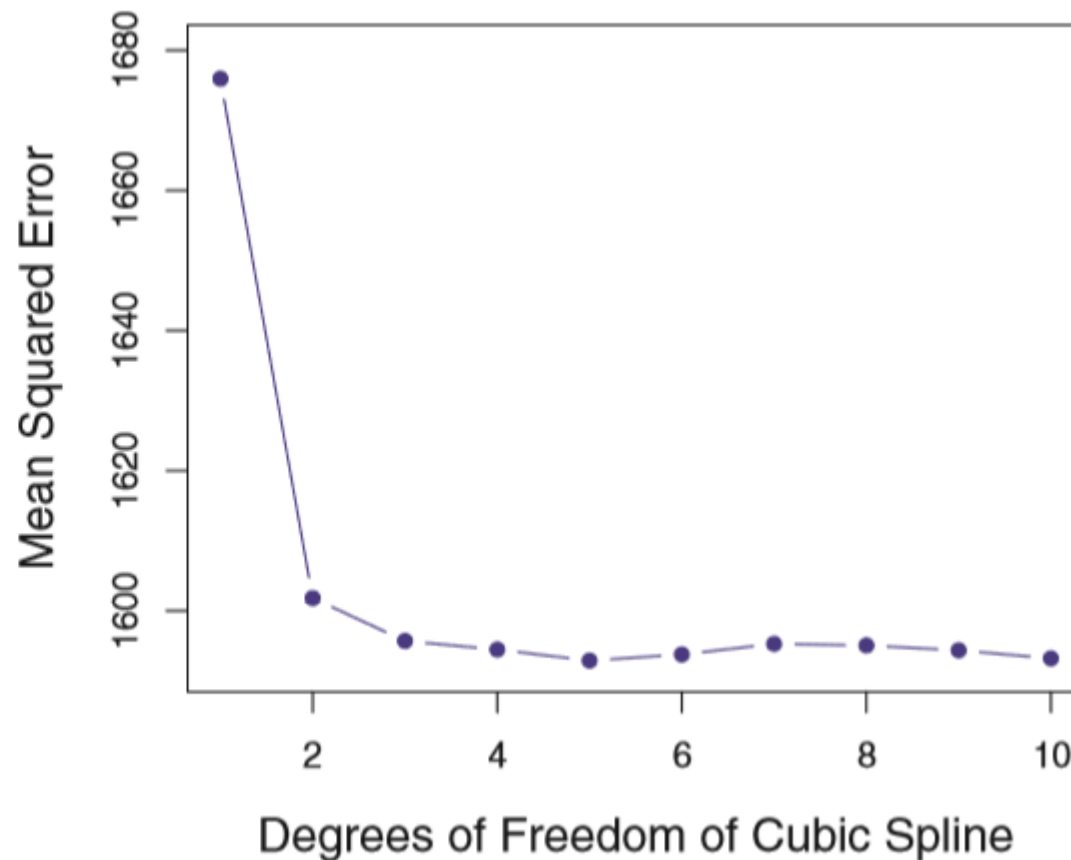
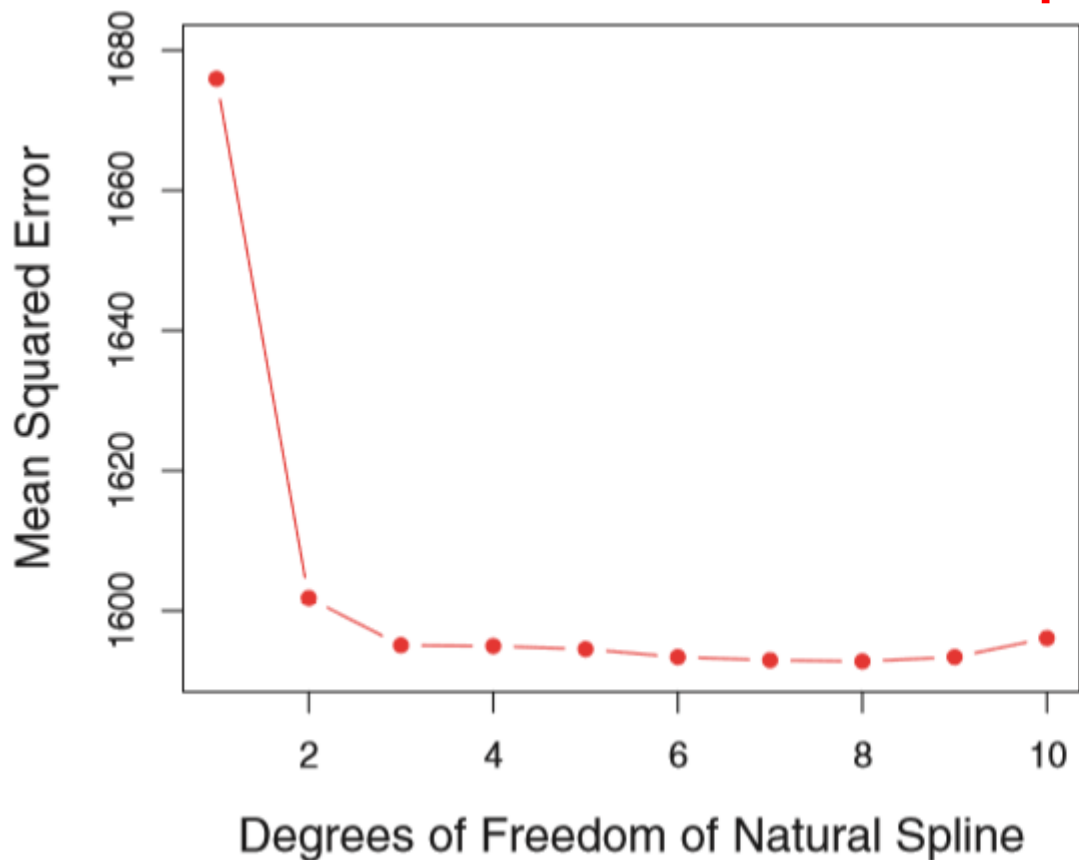


매듭 수와 자유도

보통 매듭은 일정한 간격으로 하고, 개수를 설정하는 문제 → 교차검증(CV)으로!



자유도

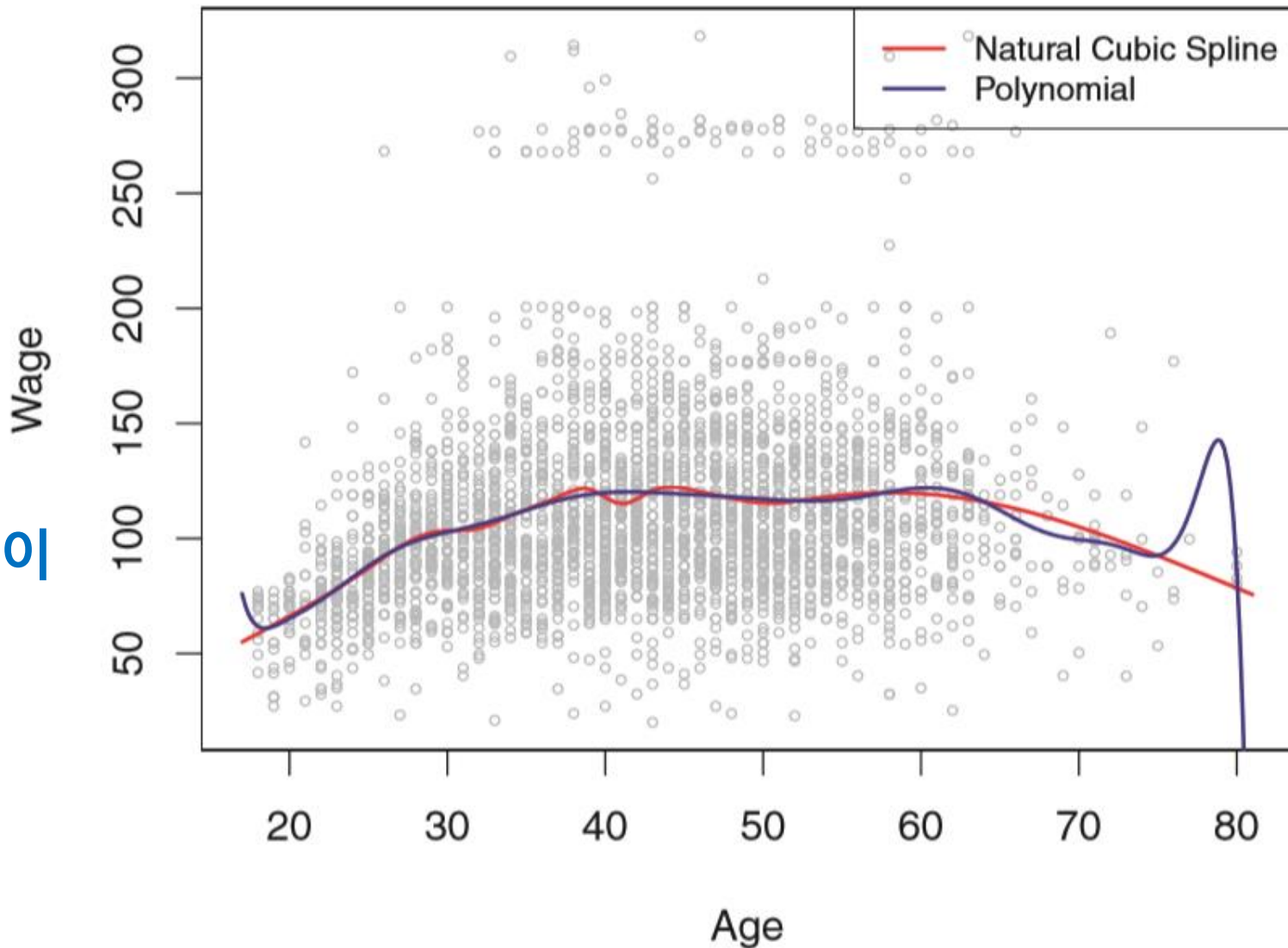


다항회귀와 비교

경계부분에서

삼차 스플라인과는 신뢰구간 차이

다항회귀와는 값 자체의 차이

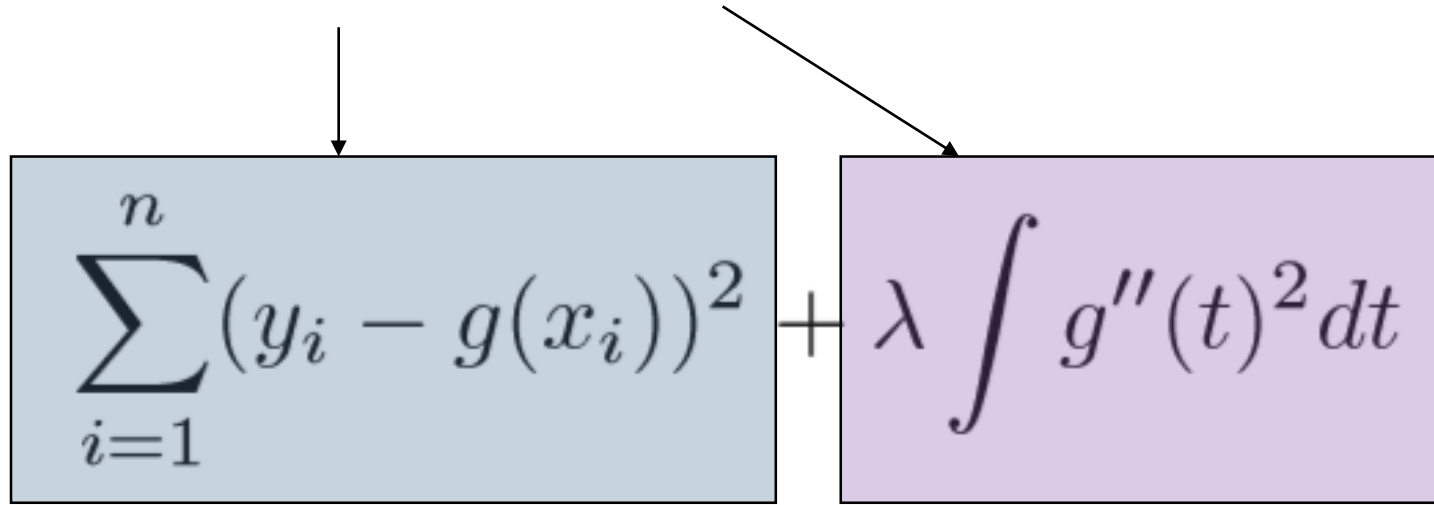


(15차 다항 VS. 15자유도 자연 삼차 스플라인)

4. 평활 스플라인

(Smoothing Spline)

목적 함수 = RSS + Penalty



The diagram illustrates the components of the objective function. A vertical arrow points from the 'RSS' part of the title to a light blue box containing the formula for the Residual Sum of Squares: $\sum_{i=1}^n (y_i - g(x_i))^2$. A diagonal arrow points from the 'Penalty' part of the title to a light purple box containing the formula for the penalty term: $\lambda \int g''(t)^2 dt$. The two boxes are separated by a plus sign, representing the total objective function.

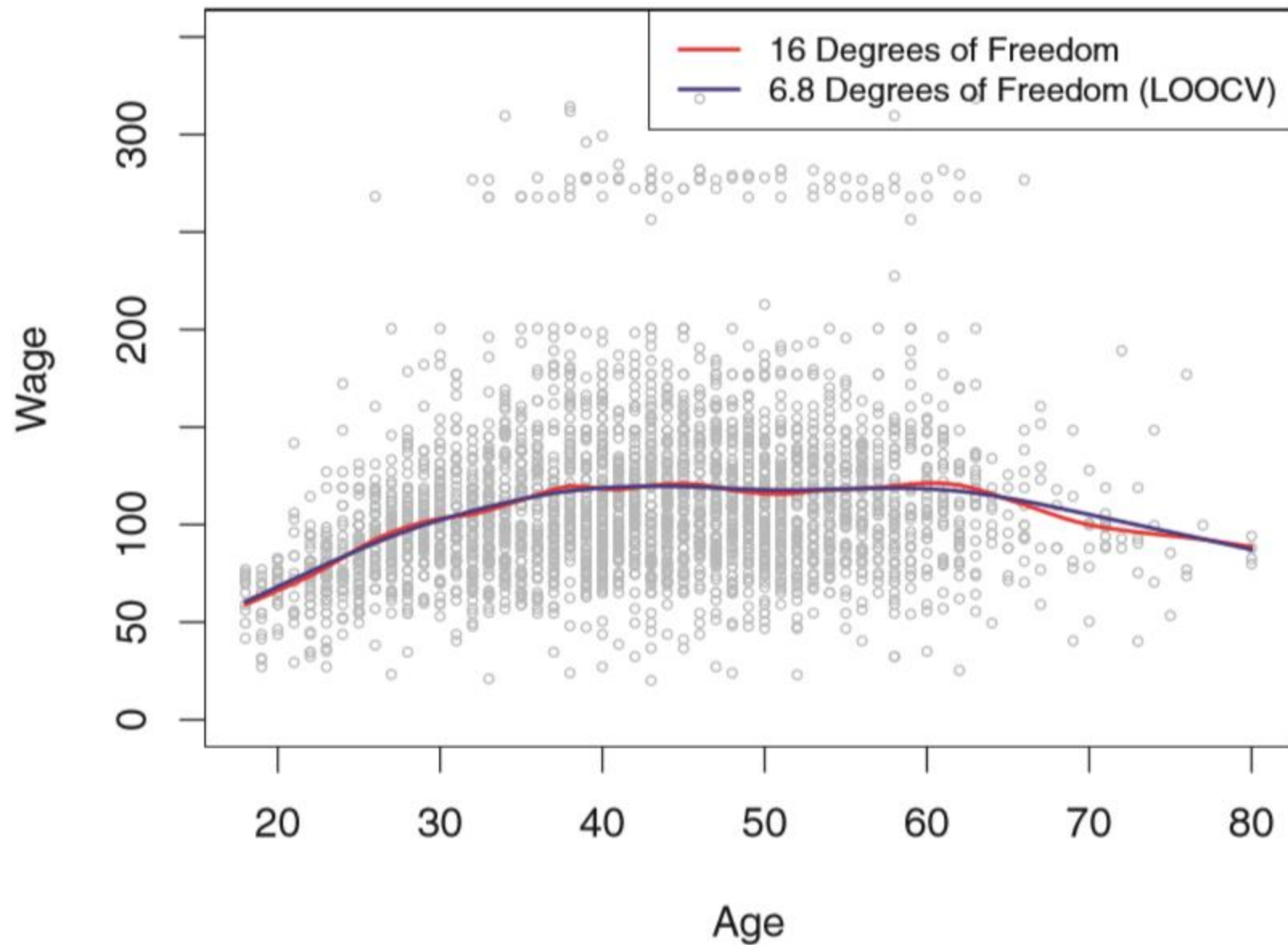
$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

2차 미분 = 1차 미분의 변동 = 기울기의 변동

2차 미분 값이 적으면 → 전 구간에 대한 기울기의 변동이 크지 않게 →

결론적으로, Smoothing (평활)하게 Fitting 한다!

평활 스플라인 적합 모습

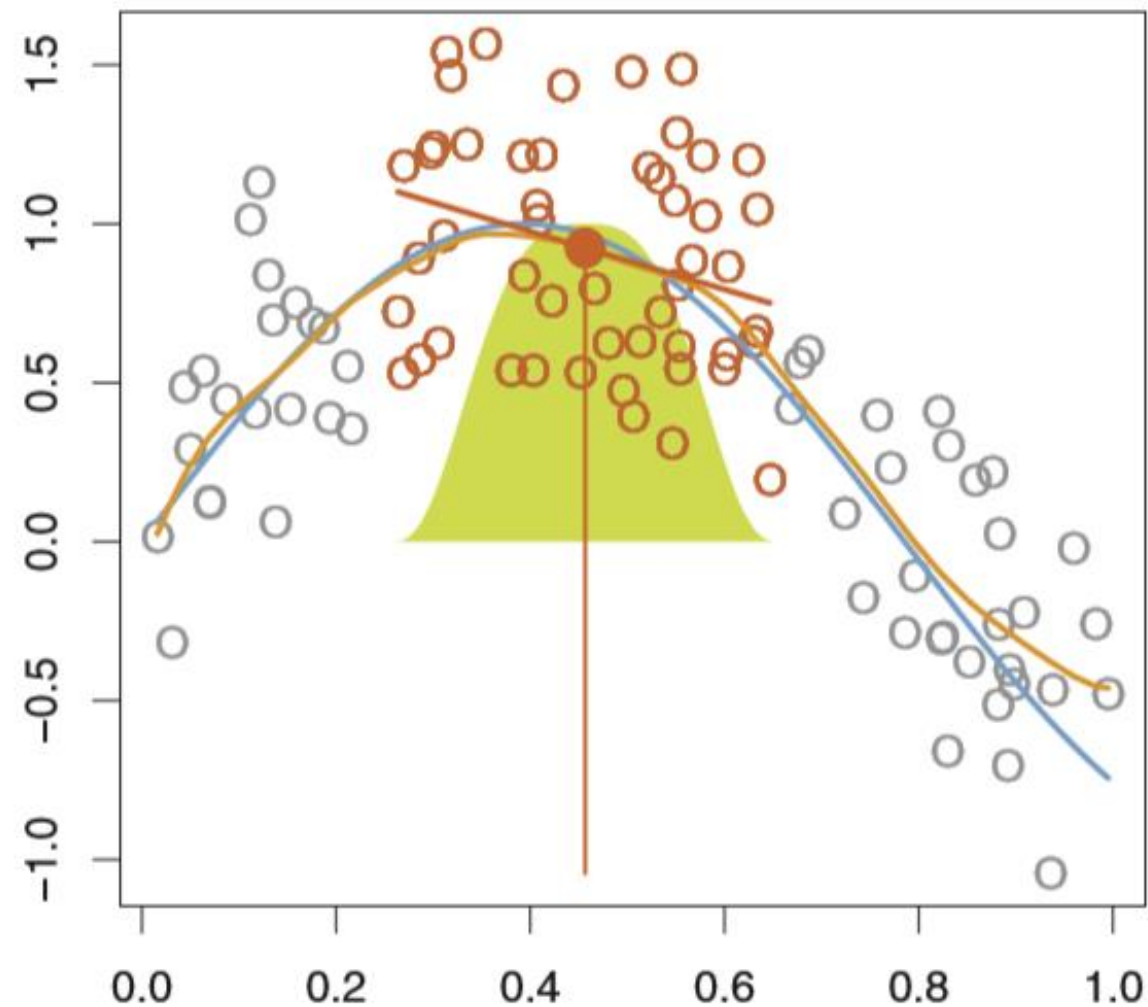
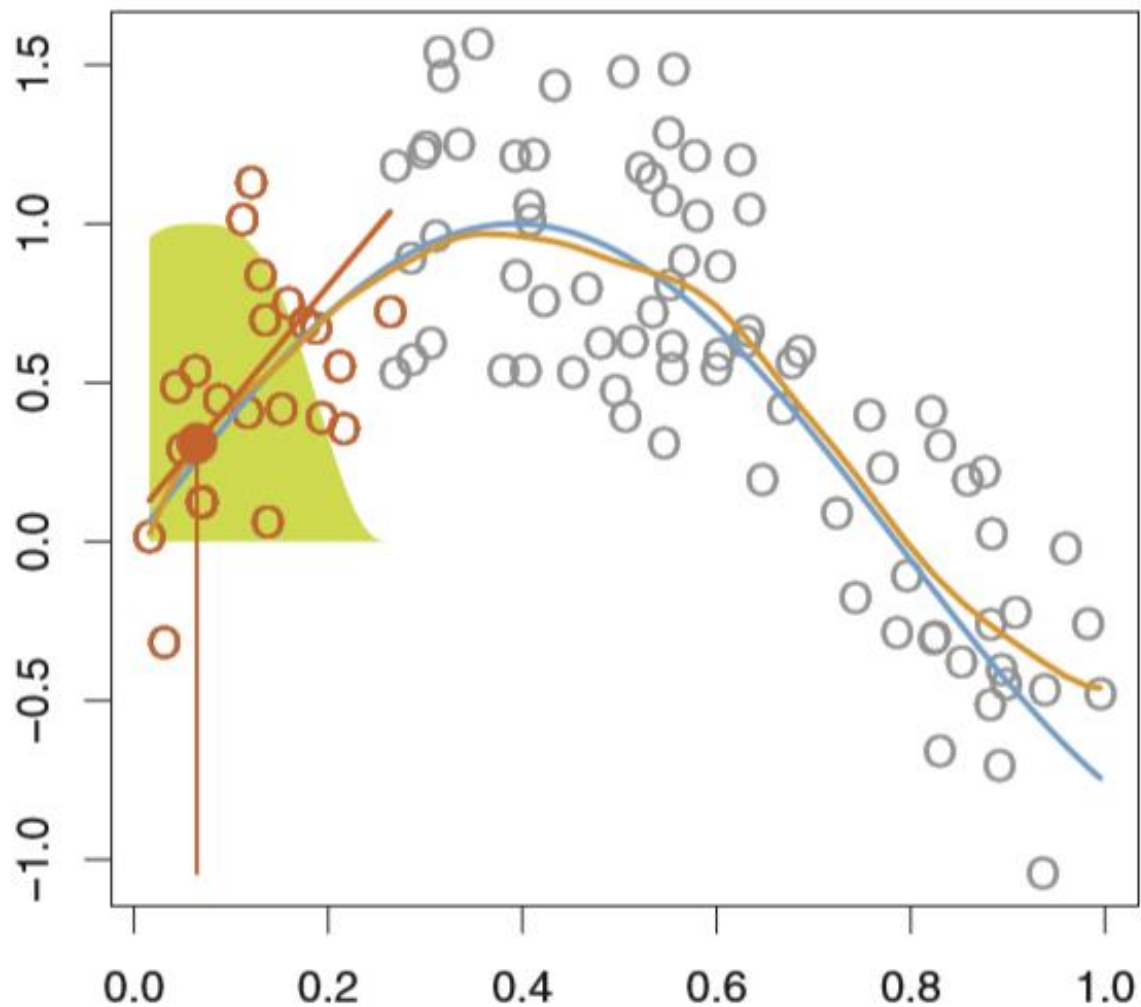


5. 국소회귀

(Local Regression)

메인 아이디어: 부분적으로 적합!

Local Regression



알고리즘

Algorithm 7.1 *Local Regression At $X = x_0$*

s값에 따라 유연성 결정!

1. 특정 입력 X 에 가까운 일부 $s = k/n$ 을 모은다.

2. 이웃의 각 점에 가중치 K 를 할당한다.

여러 방법으로 가중치 배치 가능!

(예: 특정 입력 X 에 가까우면 큰 값, 멀수록 작아지다가 0)

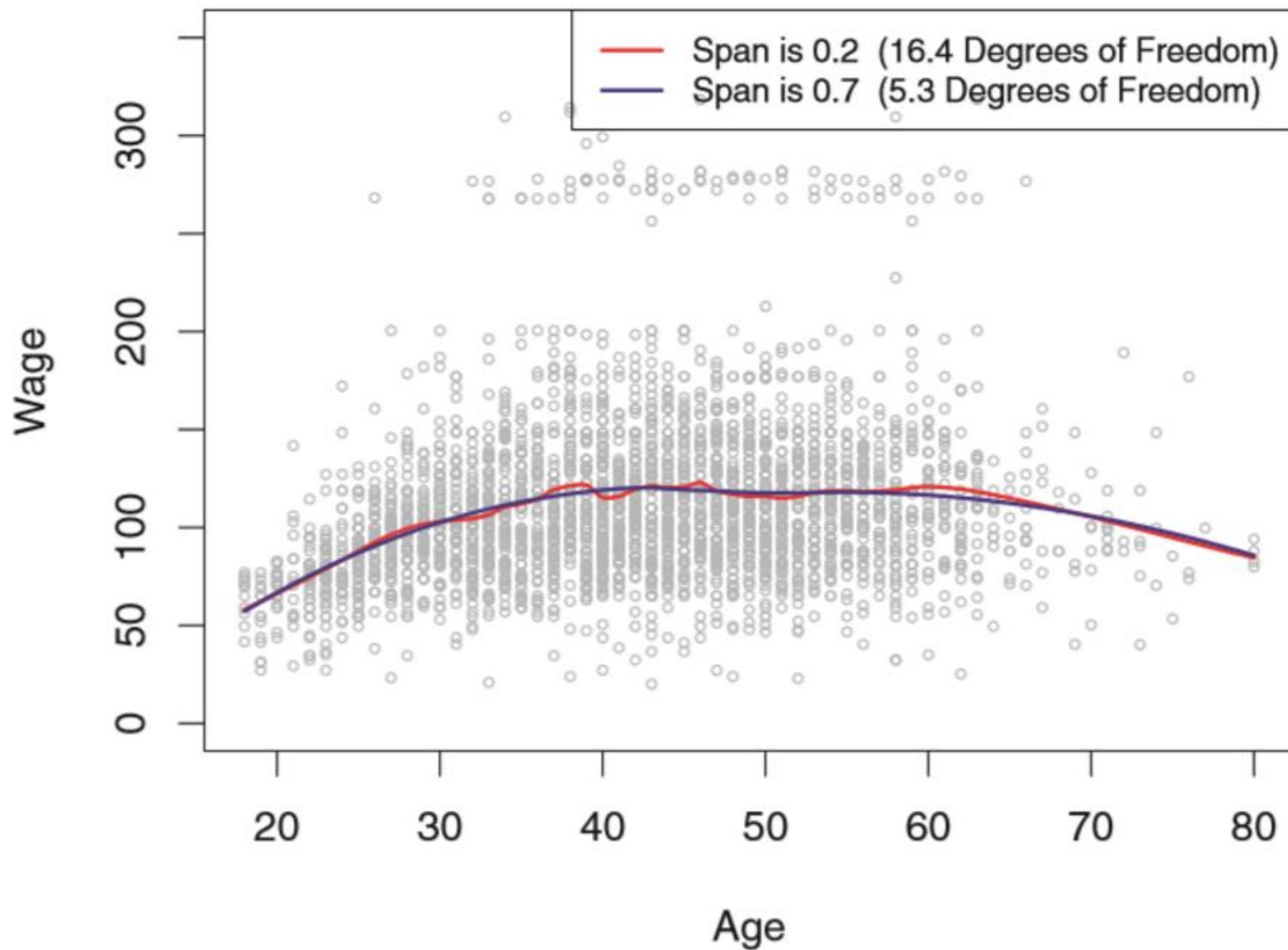
3. 앞의 가중치 적용하여 오른쪽 계수를 구한다.

선형 이외도 가능!

$$\sum_{i=1}^n K_{i0} (y_i - \beta_0 - \beta_1 x_i)^2$$

4. 각 지점에서의 적합된 모델이 다르므로, 추정값은 입력에 맞는 함수를 사용한 결과

s값에 따른 국소회귀 결과 비교



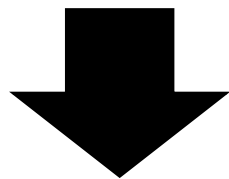
6. 일반화가법모델

(Generalized Additive Models)

메인 아이디어 (회귀)

: 각각의 입력에 비선형 함수를 씌우고 그냥 합한다(가법)!

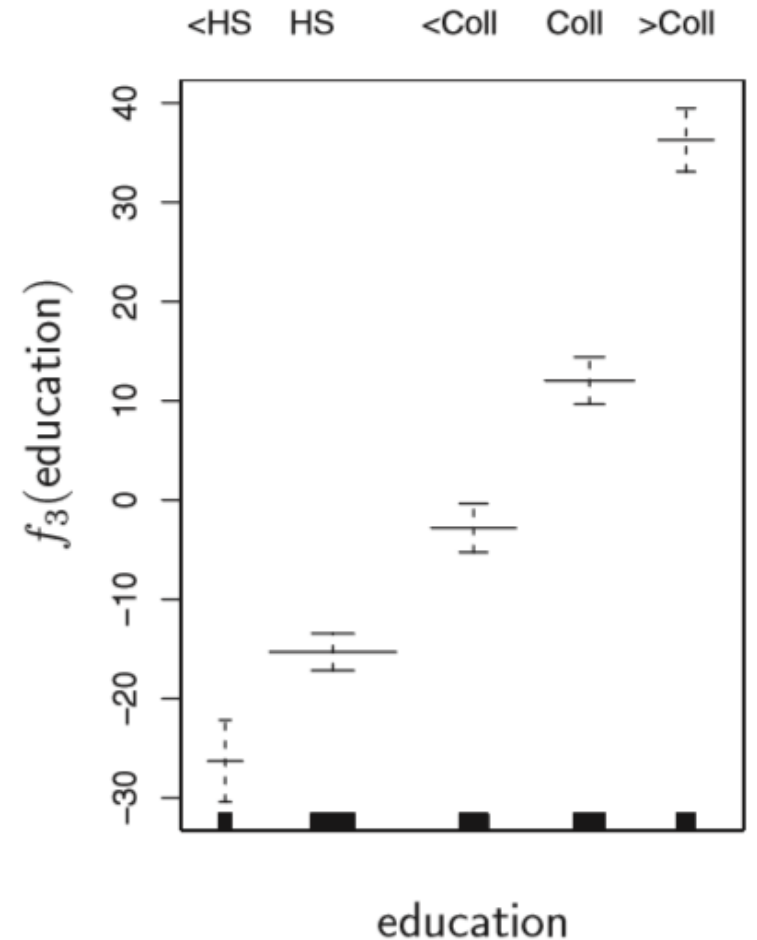
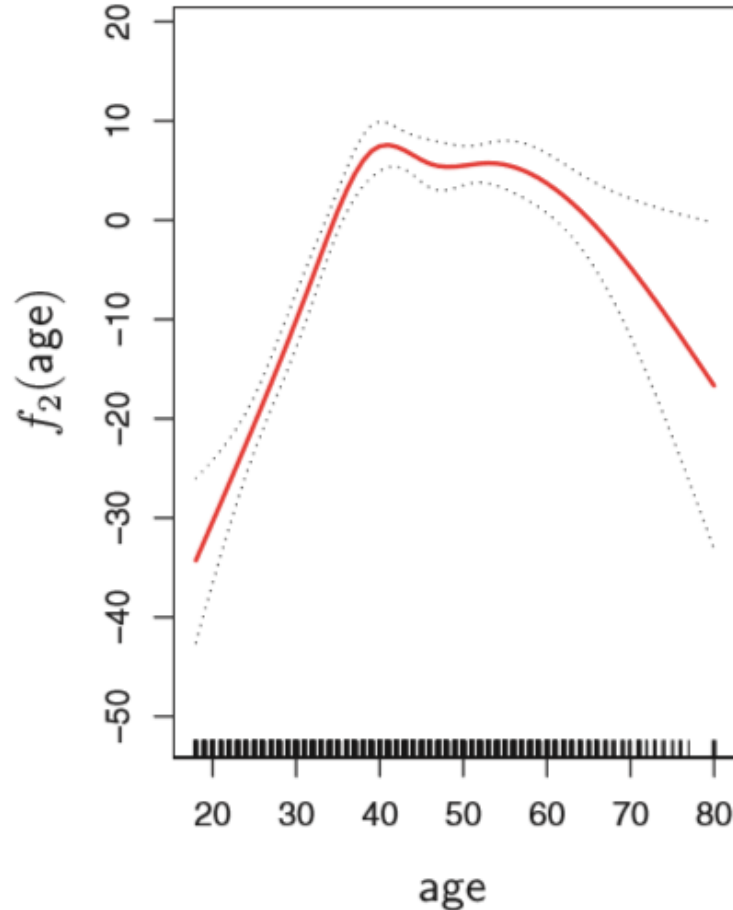
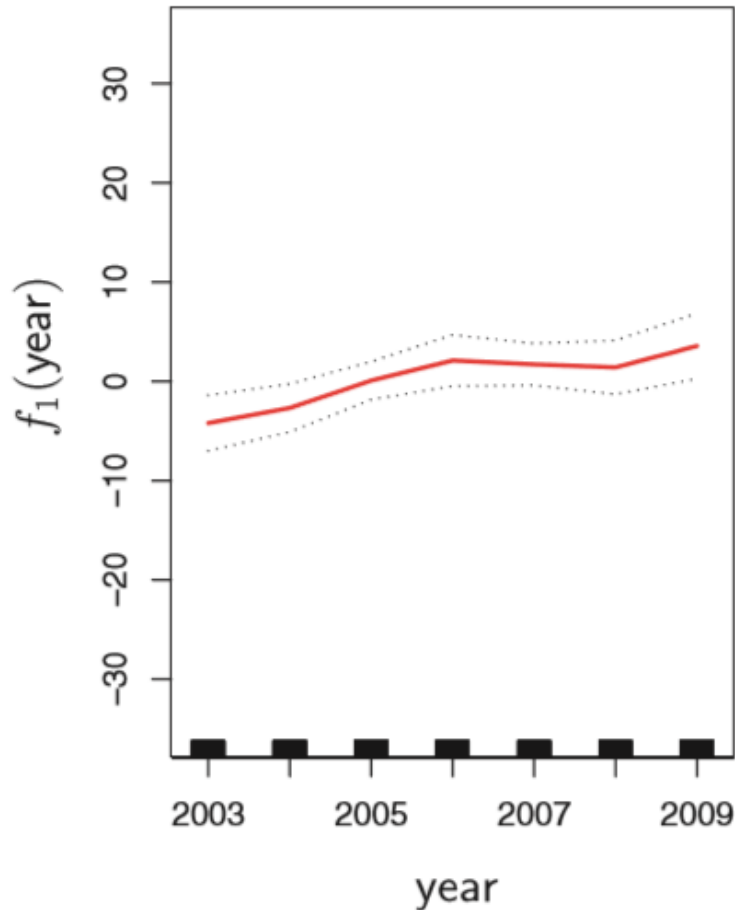
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$



$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^p f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i \end{aligned}$$

예시: 경력, 나이, 교육 정도에 따른 임금

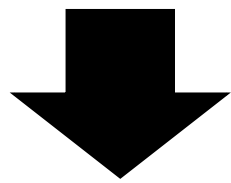
$$\text{wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \epsilon$$



메인 아이디어 (분류)

: 각각의 입력에 비선형 함수를 씌우고 그냥 합한다(가법)!

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$



$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

정리

비선형적 접근법 종류

단일 입력 변수 X - 출력 Y

1. 다항회귀
2. 조각별 상수함수
3. 회귀 스플라인
4. 평활 스플라인
5. 국소회귀

다중 입력 변수 X_1, X_2, \dots, X_p - 출력 Y

6. 일반화가법모델

Thank you!