Spline & B-Spline

12191885 김재겸
12221917 김한별

in An Introduction to Statistical Learning with applications in R

1 What is a Spline?

2. Spline – 1) Truncated Polynomial

3. Spline – 2) Cubic Spline

4. Spline – 3) Natural Spline

5. B-Spline

What is a **Spline**?

- Linear model의 문제점 보완

-> Non-linear model !

1) Polynomial Regression

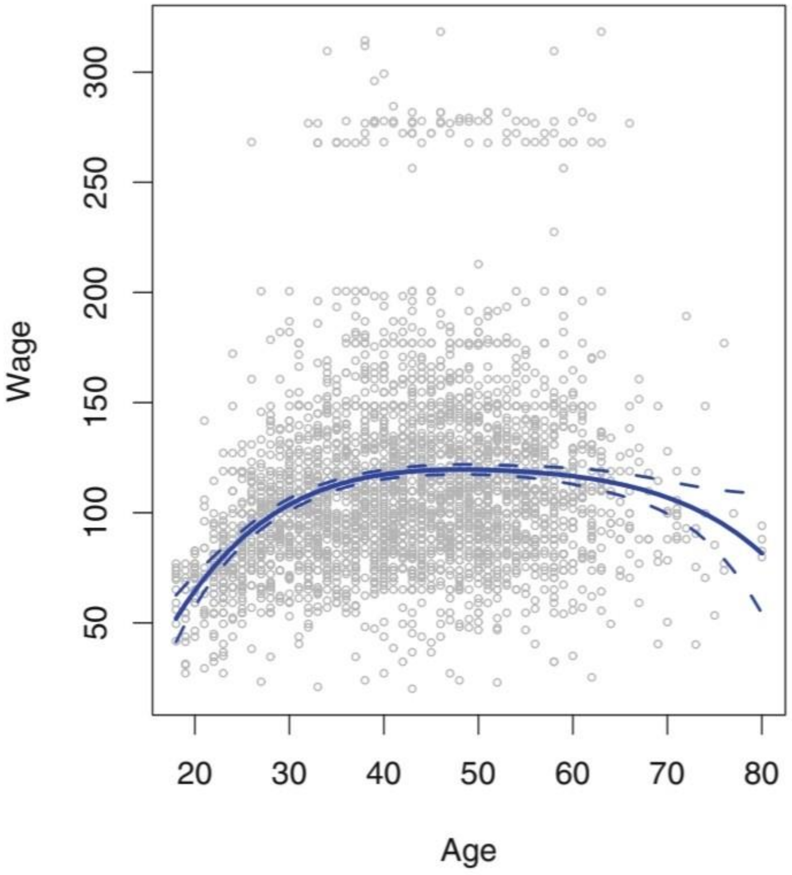
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 + \varepsilon_i$$

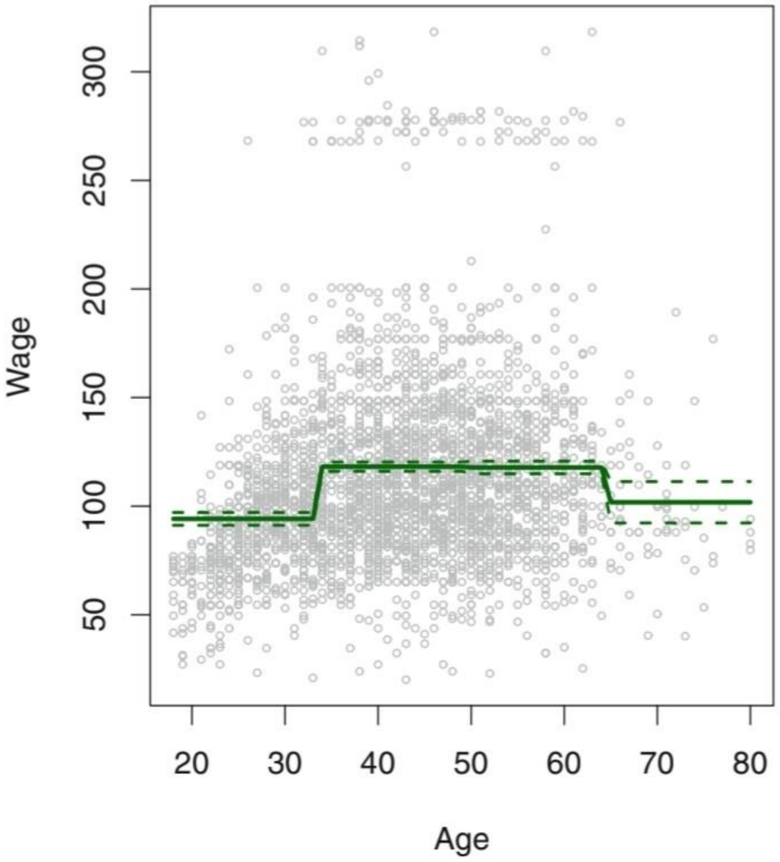
2) Step functions

x를 cutpoint $(c_1, c_2, \ldots, c_k)$ K+1구간으로 나눠서 각 구간별로 fitting

3) Spline

Polynomial Regression & Step functions





Spline

- 어떤 m개의 cutpoint(knot)를 기준으로 m+1개의 구간이 있을 때

  연속하는 k차 “piecewise” 다항식

  -> 구간별로 정의된 매끈한 다항식

- 각 knot에서 1차, 2차, k-1차 도함수까지 연속인 선: kth spline

Formally, a function $f : \mathbb{R} \to \mathbb{R}$ is a kth order spline with knot points at $t_1 < \ldots < t_m$, if

- f is a polynomial of degree k on each of the intervals $(-\infty, t_1], [t_1, t_2], \ldots [t_m, \infty)$, and
- $f^{(j)}$, the jth derivative of f, is continuous at $t_1, \ldots t_m$, for each $j = 0, 1, \ldots k-1$.

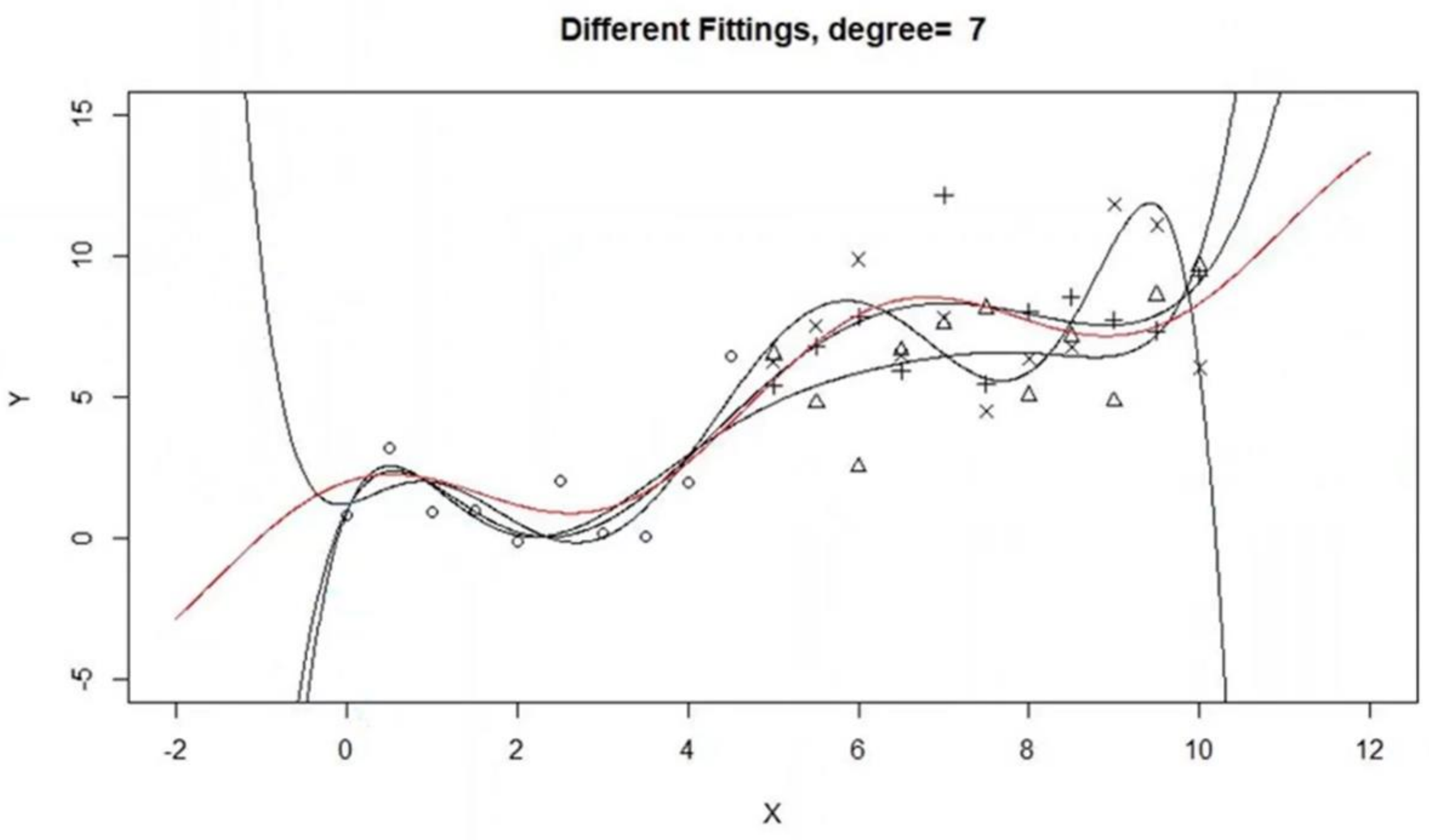
Spline

- k = 1,   piecewise constant function
- k = 2,   piecewise linear function
- k = 3,   piecewise quadratic function

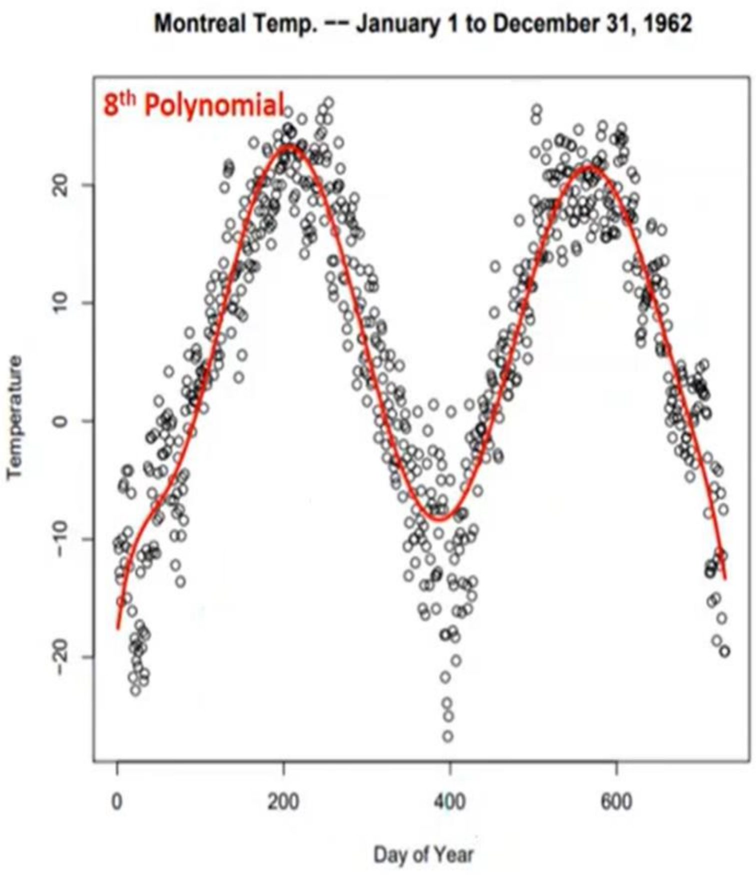
- 궁금한 점…

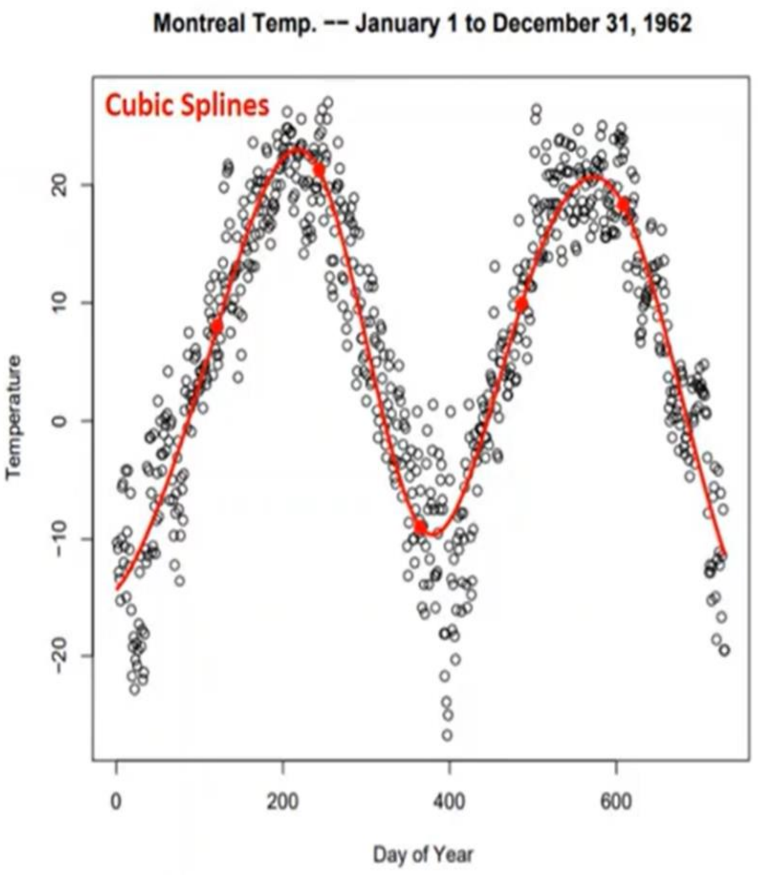
  다항식으로 fitting 하면 될 것 같은데 왜 굳이 spline?

Spline 사용하는 이유?



Spline 사용하는 이유?

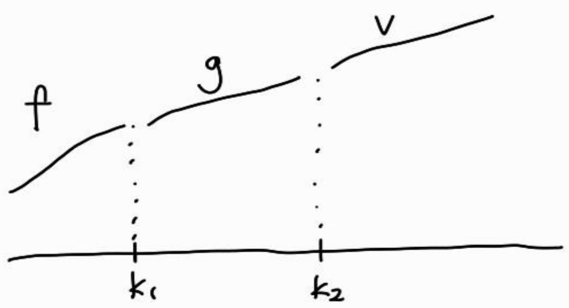




Spline – 1) Truncated Polynomial

$k=2$, $D=3$, cubic spline

(구간) (차수)



i) $f(k_1) = g(k_1)$

ii) $f'(k_1) = g'(k_1)$

iii) $f''(k_1) = g''(k_1)$

이렇게

+가 붙은 다항식을

Truncated Polynomial

$$f(x) + \underbrace{\beta_1 (x-k_1)_+^3}_{\substack{0 & x<k_1 \\ (x-k_1)^3 & x \geq k_1}} + \beta_2 (x-k_2)_+^3 \leftarrow$$

$f(x)$ 가 있으면

$g(x) = \underbrace{f(x) + (x-k_1)^3}$

$V(x) = (\quad) + (x-k_2)^3$

$= f(x) + (x-k_1)^3 + (x-k_2)^3$

$g'(x) = f'(x) + 3(x-k_1)^2$

$g''(x) = f''(x) + 6(x-k_1)$

Spline – 1) Truncated Polynomial

- truncated 다항식:

$$h(x,\xi) = (x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}$$

- M개의 knot에 대해 K차 spline을 fitting할 경우 회귀식:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_i^k + \sum_{j=1}^{m} \hat{\beta}_j (x_i - \xi_j)_+^K = \sum_{j=1}^{K+1+m} \hat{\beta}_{j-1}\, g_j(x_i)$$

Spline – 1) Truncated Polynomial

이 경우 Design Matrix $\mathbb{G} \in \mathcal{R}^{n\times(K+1+m)}$

$$\mathbb{G}_{ij} = g_j(x_i),\ \ for\ i \in [n],\ j \in [K+1+m]$$

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^D & (x_1-\xi_1)_+^D & \cdots & (x_1-\xi_K)_+^D \\ 1 & x_2 & x_2^2 & \cdots & x_2^D & (x_2-\xi_1)_+^D & \cdots & (x_2-\xi_K)_+^D \\ 1 & x_3 & x_3^2 & \cdots & x_3^D & (x_3-\xi_1)_+^D & \cdots & (x_3-\xi_K)_+^D \\ & & & \vdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^D & (x_n-\xi_1)_+^D & \cdots & (x_n-\xi_K)_+^D \end{bmatrix}$$

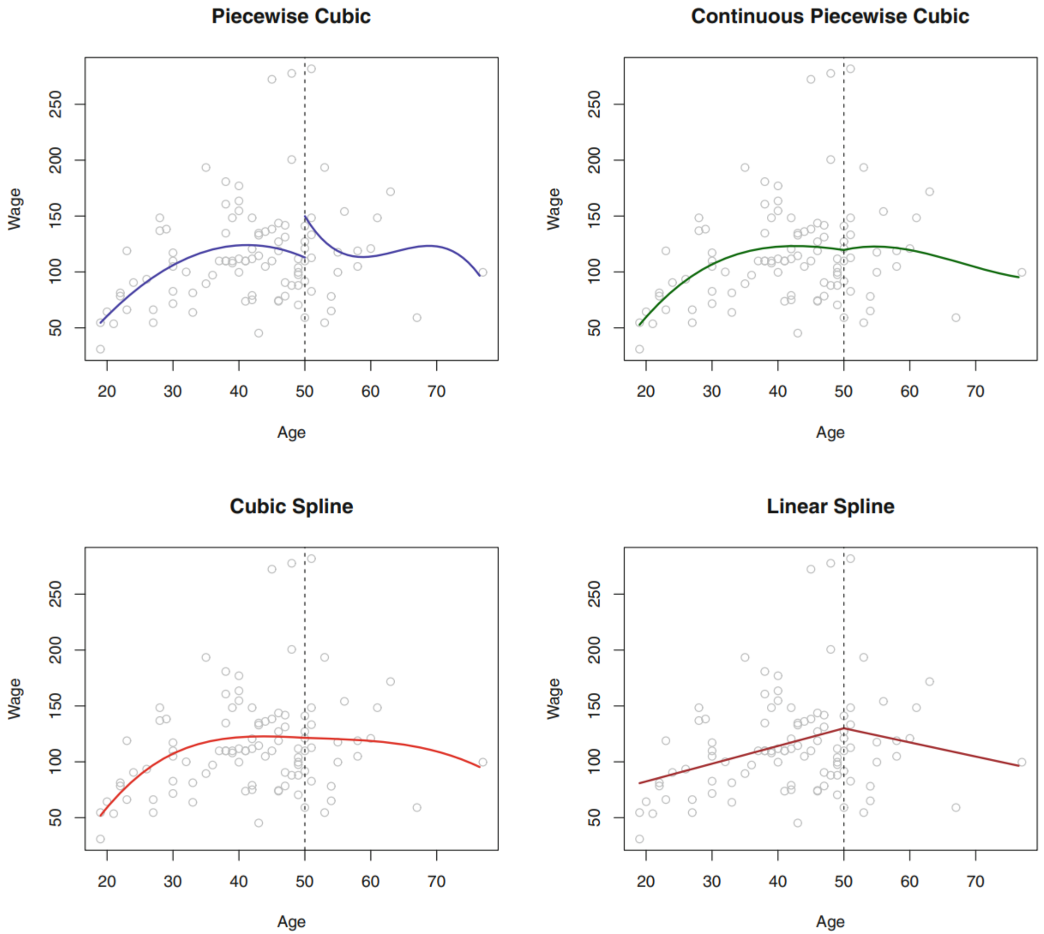
Design Matrix의 각 열 $g_j(x)$은 결국 **K개의 knot로 이뤄진 D차 piecewise 다항식이라는 연속함수공간(벡터스페이스)를 "truncated polynomial"로 span하는 기저**로 볼 수 있다.

이후는 OLS처럼 $\min_\beta \|Y - G\beta\|_2^2$, 이를 만족하는 $\hat\beta$는 $\hat\beta = (G^TG)^{-1}G^TY$로 구할 수 있다.

Spline – 2) Cubic Spline

- k = 4일 때의 spline
- 일반적으로 사용됨
  - 1, 2차함수: 곡선의 복잡도를 충분히 표현 x
  - 4차 이상: 복잡한 곡선 표현 가능하지만, 곡선이 불안정해질 수 있다.
- ex) knot: 2개일 때 cubic spline

Spline – 2) Cubic Spline

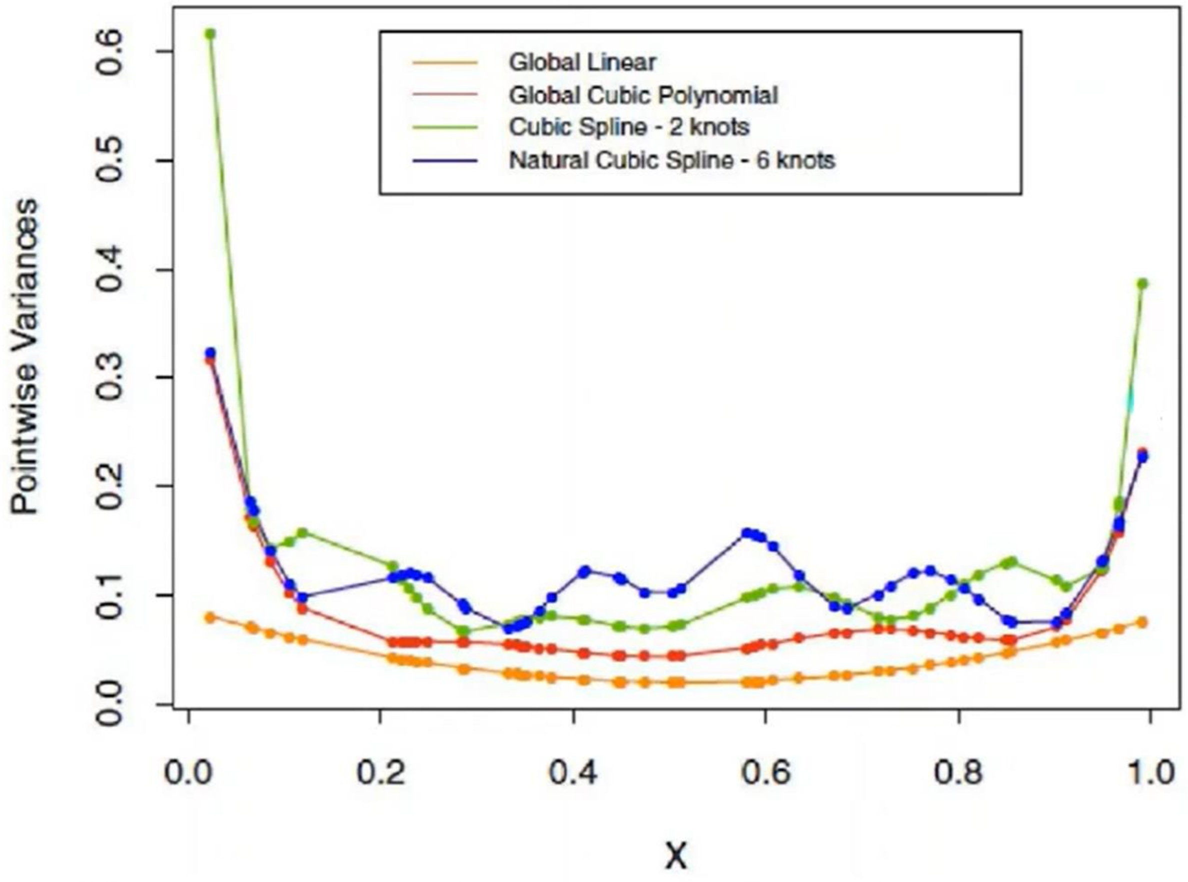


Spline – 3) Natural Spline

- D차 spline의 문제점
    - Boundary(처음과 마지막 knot의 바깥 구간)에서 분산이 크다.
    - knot의 양 끝구간엔 데이터가 거의 존재하지 않기 때문이다.
    - 그 구간에서는 제한된 샘플로 많은 계수를 추청해야 하므로 추정치의 분포가 자유도가
    - 더 낮은 t분포를 따르니 CI가 더 클 수 밖에 없다.

- **Natural Spline :** Boundray에서 인위적으로 1차식이 되도록 조건을 가해
  추정 계수의 개수를 줄인 것

Spline – 3) Natural Spline



B-Spline

- Truncated polynomial

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^D & (x_1-\xi_1)_+^D & \cdots & (x_1-\xi_K)_+^D \\ 1 & x_2 & x_2^2 & \cdots & x_2^D & (x_2-\xi_1)_+^D & \cdots & (x_2-\xi_K)_+^D \\ 1 & x_3 & x_3^2 & \cdots & x_3^D & (x_3-\xi_1)_+^D & \cdots & (x_3-\xi_K)_+^D \\ & & & \vdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^D & (x_n-\xi_1)_+^D & \cdots & (x_n-\xi_K)_+^D \end{bmatrix}$$

$$\hat{\beta} = (G^T G)^{-1} G^T Y$$

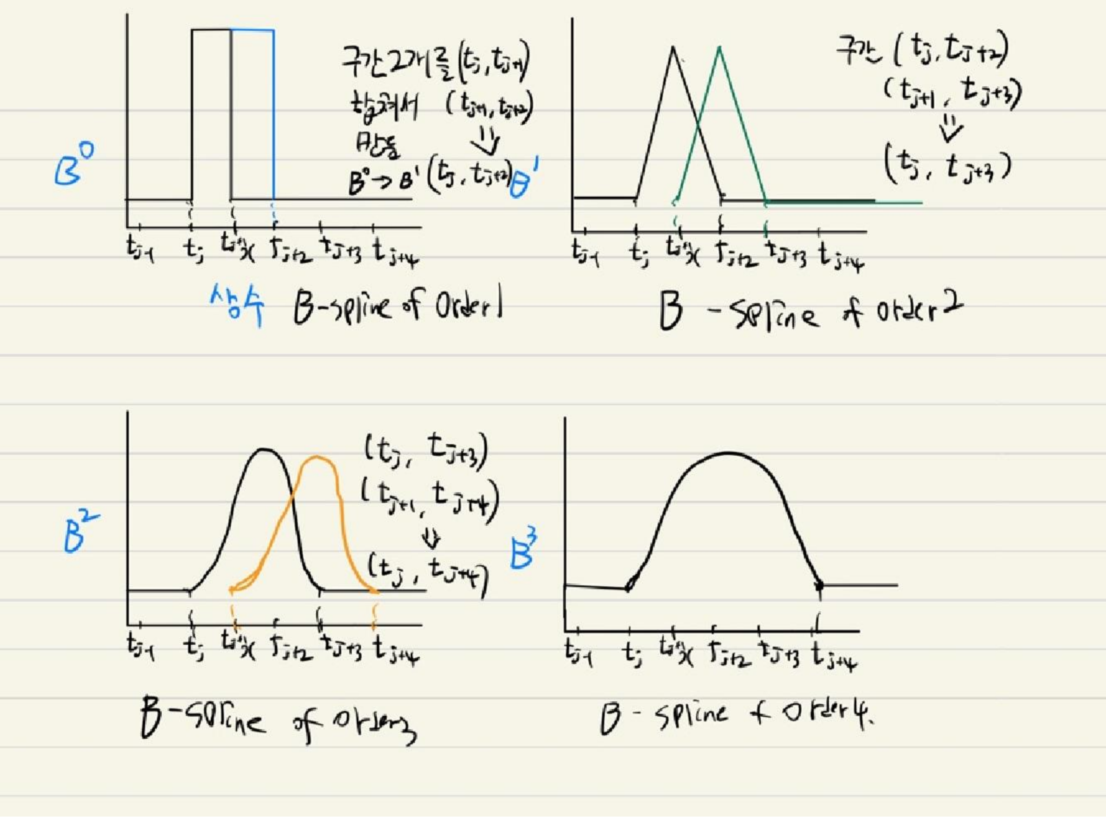
- 문제점
  - 다중공선성의 문제
  - 차수가 올라갈 수록 rounding으로 인한 오차문제가 발생한다는 점에서 수치적으로 불안정함

B-Spline

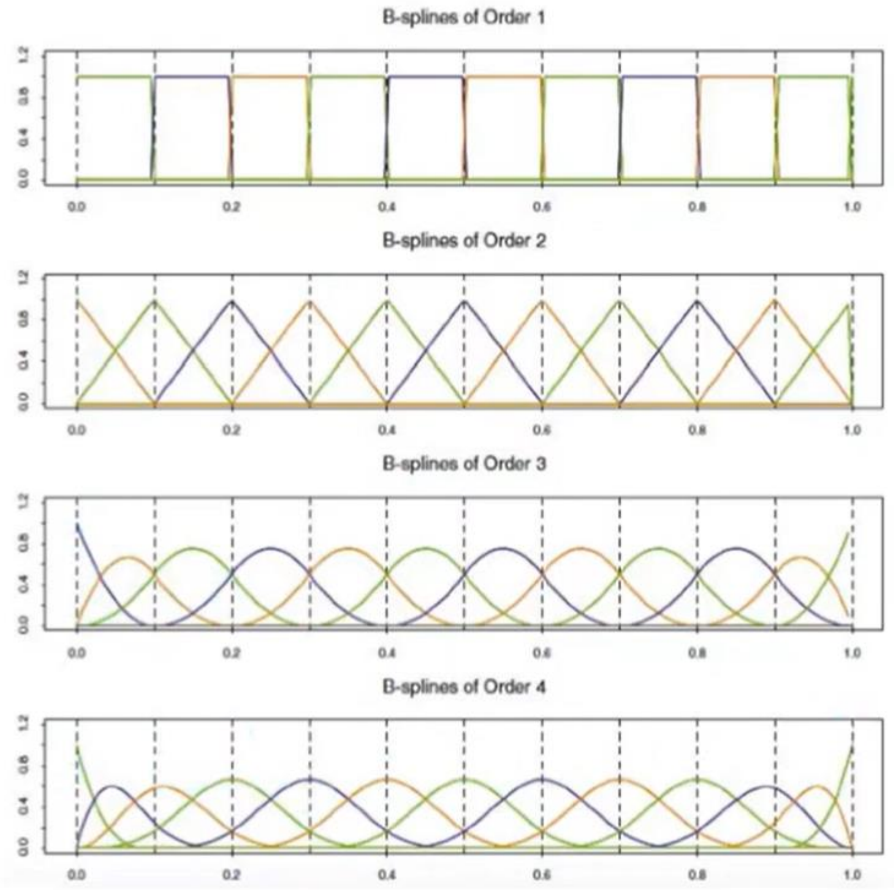
$$B_j^0(x) = \begin{cases} 1, & \text{if } t_j \le x < t_{j+1} \\ 0, & \text{otherwise,} \end{cases}$$

$$B_j^k(x) = \frac{x - t_j}{t_{j+k} - t_j} B_j^{k-1}(x) + \frac{t_{j+k+1} - x}{t_{j+k+1} - t_{j+1}} B_{j+1}^{k-1}(x)$$

B-Spline



B^0
구간 2개를 (t_j, t_{j+1}) 합쳐서 (t_{j+1}, t_{j+2}) 만듦
$B^0 \to B^1$ (t_j, t_{j+2}) B^1
t_{j-1} t_j t_{j+1} x t_{j+2} t_{j+3} t_{j+4}
상수 B-spline of Order1
구간 (t_j, t_{j+2}) (t_{j+1}, t_{j+3}) (t_j, t_{j+3})
B-spline of order2
B^2
(t_j, t_{j+3}) (t_{j+1}, t_{j+4}) (t_j, t_{j+4})
B-spline of order3
B^3
B-spline of order4.



B-splines of Order 1
B-splines of Order 2
B-splines of Order 3
B-splines of Order 4
0.0
0.2
0.4
0.6
0.8
1.0