스플라인 회귀모형

서울시립대 통계학과 이용희

FALL 2019 학부

1 함수의 기저

반응변수 Y와 설명변수 X가 다음과 같은 관계를 가진다고 하자.

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

이러한 관계를 회귀모형(regression model)이라고 하며 Y의 평균이 X에 따라서 변하는 괸계를 설정하는 모형이다. 만약 $m(x)$의 형태를 회귀계수의 선형식으로 나타낼 수 있다면 우리는 이를 선형회귀모형(linear regression model)이라고 한다.

$$m(x) = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

설명변수는 고정된 값이거나 또는 확률변수일 수도 있다.

비모수 회귀모형(nonparametric regression model)에서는 $m(x)$의 형태에 특별한 제한을 두지 않는다. 따라서 함수 $m(x)$는 무수히 많고 다양한 형태를 가질 수 있다.

이제 설명변수의 수가 1개라고 가정하면 함수 $m(x)$를 가장 단순하게 표현할 수 있는 모형이 1차 회귀모형이다.

$$y = a + bx + e$$

만약 반응변수와 설명변수의 관계가 선형이 아니라 비선형이라면 $m(x)$를 p-차 다항식으로 사용할 수 있다.

$$m(x) = a + b_1x + b_2x^2 + \cdots + b_px^p \tag{1}$$

이렇게 $m(x)$를 다항식으로 표현하는 것은 우리가 알 수 없는 함수를 $p + 1$개의 기저들(basis), 즉 1, x, x^2, ..,x^p의 선형조합으로 근사하는 것이며 다항식의 경우는 각 기저 $N_k(x)$ 는 설명변수의 k-차 항 x^k이다.

$$\beta_0N_0(x) + \beta_1N_1(x) + \beta_2N_2(x) + \cdots + \beta_pN_p(x) \tag{2}$$

일반적으로 임의의 함수는 다양한 형태의 기저들로 표현할 수 있으며 기저들의 형태에 따라 다음과 같은 것들이 있다.

- 다항식(polynimials)
- 퓨리에 함수(Fourier series)
- 웨이블릿 함수(Wavelet series)

2 Regesssogram 의 확장: 직선 스플라인(Linear Spline)

이제 n개의 자료 $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$이 주어졌다고 하자. 편의상 설명변수 x가 구간 $(0, 1)$의 값이라고 가정한다. 설명변수들이 포함되는 구간 $(0, 1)$를 $K+1$개의 구간 $\{B_j = (\xi_{k-1}, \xi_k) | k = 1, 2, .., K+1\}$ 으로 나누어 보자. 즉

$$\xi_0 = 0 < \xi_1 < \xi_2 < \cdots < \xi_{K-1} < \xi_K < 1 = \xi_{K+1}$$

위의 구간에서 K개의 내부점들 $\xi_1 < \xi_2 < \cdots < \xi_{K-1} < \xi_K$은 일반적으로 knots(연결점)이라고 부른다,

이제 반응변수들의 평균 $E(Y|x)$ 을 단순하게 각 구간에서 상수라고 가정하면

$$E(Y|x) = a_k \quad \text{if } x \in B_k$$

각 구간에 속하는 반응변수들의 평균으로 추정할 수 있다. 이러한 추정법을 우리는 전 강의에서 Regressogram이라고 하였다.

$$\hat{E}(Y|x) = \frac{1}{n_k} \sum_{i=1}^{n} y_i I_{B_k}(x_i) \quad \text{if } x \in B_k \tag{3}$$

이제 Regeressogram의 개념을 확장시켜서 각 구간의 회귀모형을 직선식으로 확장하여 보자. 즉

$$E(Y|x) = a_k + b_k x \quad \text{if } x \in B_k \tag{4}$$

모형 (4)는 각 구간마다 회귀직선을 적합하는 것과 같은 모형이다. 이러한 모형에서 추정된 각 회귀식들은 각 구간의 연결점 $\xi_1, \xi_2, \ldots, \xi_K$ 에서 불연속이다. 각 연결점 ξ_k에서 연속인 회귀식들을 구할 수 있을까?

이제 구간 $(0,1)$을 다음과 같이 $K+1=3$개의 구간으로 나누어 보자. 이 경우 연결점은 ξ_1과 ξ_2, 두 개가 있다.

$$B_1 = (0, \xi_1) \quad B_2 = (\xi_1, \xi_2) \quad B_3 = (\xi_2, 1)$$

모형 (4)에서 주어진 회귀계수는 모두 $2(K+1)=6$ 개이며 추정값을 구하는 방법은 다음과 같은 오차제곱합을 구하는 최소제곱법을 사용할 수 있다.

$$SSE = \sum_{x_i \in B_1} (y_i - a_1 - b_1 x_i)^2 + \sum_{x_i \in B_2} (y_i - a_2 - b_2 x_i)^2 + \sum_{x_i \in B_3} (y_i - a_3 - b_3 x_i)^2 \tag{5}$$

이제 모형 (4)에서 구한 직선회귀식이 연결점 ξ_1과 ξ_2에서 연속이 되려면 다음과 같은 두 조건을 만족해야 한다.

$$a_1 + b_1\xi_1 = a_2 + b_2\xi_1, \quad a_2 + b_2\xi_2 = a_3 + b_3\xi_2 \tag{6}$$

각 연결점의 연속을 만족하려면 위의 두 식을 만족해야 하므로 이제 추정해야 하는 모수의 개수는 4개이다. 왜냐하면 원래의 회귀계수의 개수 6 에서 제약식 2개의 개수를 제외해야 하기 때문이다.

제약식 (6)을 다시 쓰면 다음과 같이 쓸수 있으며

$$\begin{aligned} a_2 &= a_1 + (b_1 - b_2)\xi_1 \\ a_3 &= a_2 + (b_2 - b_3)\xi_2 \\ &= a_1 + (b_1 - b_2)\xi_1 + (b_2 - b_3)\xi_2 \end{aligned}$$

이 제약식을 이용하여 오차제곱합 (5)에서 두 번째 항과 세 번째 항을 다음과 같이 전개할 수 있다

$$\begin{aligned}
\sum_{x_i \in B_2} (y_i - a_2 - b_2 x_i)^2 &= \sum_{x_i \in B_2} (y_i - a_1 - (b_1 - b_2)\xi_1 - b_2 x_i)^2 \\
&= \sum_{x_i \in B_2} [y_i - a_1 - b_1 \xi_1 - b_2 (x_i - \xi_1)]^2 \\
&= \sum_{x_i \in B_2} [y_i - a_1 - b_1 x_i - (b_2 - b_1)(x_i - \xi_1)]^2 \\
\sum_{x_i \in B_3} (y_i - a_3 - b_3 x_i)^2 &= \sum_{x_i \in B_3} (y_i - a_1 - (b_1 - b_2)\xi_1 - (b_2 - b_3)\xi_2 - b_3 x_i)^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 \xi_1 + b_2 \xi_1 - b_2 \xi_2 - b_3 (x_i - \xi_2)]^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 x_i + b_1 x_i - b_1 \xi_1 + b_2 \xi_1 - b_2 \xi_2 - b_3 (x_i - \xi_2)]^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 x_i + b_1 (x_i - \xi_1) + b_2 (\xi_1 - x_i + x_i) - b_2 \xi_2 - b_3 (x_i - \xi_2)]^2 \\
&= \sum_{x_i \in B_3} [y_i - a_1 - b_1 x_i - (b_2 - b_1)(x_i - \xi_1) - (b_3 - b_2)(x_i - \xi_2)]^2
\end{aligned}$$

위의 전개식에서 모수를 다음과 같이 다시 정의하면

$$\beta_0 = a_1, \quad \beta_1 = b_1, \quad \beta_2 = b_2 - b_2, \quad \beta_3 = b_3 - b_2$$

오차제곱합 SSE는 다음과 같이 전개할 수 있다.

$$\begin{aligned}
SSE = &\sum_{x_i \in B_1} (y_i - \beta_0 - \beta_1 x_i)^2 + \sum_{x_i \in B_2} [y_i - \beta_0 - \beta_1 x_i - \beta_3 (x_i - \xi_1)]^2 \\
&+ \sum_{x_i \in B_3} [y_i - \beta_0 - \beta_1 x_i - \beta_2 (x_i - \xi_1) - \beta_3 (x_i - \xi_2)]^2
\end{aligned} \tag{7}$$

식 (7)에서 얻은 오차제곱합은 다음과 같은 회귀모형을 적합한 경우에 얻을 수 있는 오차제곱합이다.

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 (x - \xi_1)_+ + \beta_3 (x - \xi_2)_+ \tag{8}$$

위의 식에서 함수 $(x)_+$ 는 다음과 같이 정의된 함수이다.

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

이제 연결점(knots)에서 연속인 조건을 만족하는 직선들을 추정하는 문제는 식 (8)에 주어진 회귀식을 추정하는 문제와 같음을 보였다. 즉, 주어진 구간에서 서로 연결되는 최적의 직선식을 구하는 문제는 함수를 다음과 같은 기저로 조합된 것으로 보고 최적의 계수를 구하는 것과 동일한 문제이다.

$$N_0(x) = 1, \quad N_1(x) = x, \quad N_2(x) = (x - \xi_1)_+, \quad N_3(x) = (x - \xi_2)_+$$

위의 기저의 특징은 일부 기저함수의 값이 주어진 구간의 전 ξ_1과 ξ_2에 의존한다는 것이다. 아래 그림은 연결점이 2개인 경우 4개의 기저를 나타내는 그림이다.

이제 위의 문제에서 만약 $K+1$개의 구간이 있다면 원래 회귀계수의 개수 $2(K+1)$에서 제약식의 수 K를 제외한 총 $K+2$ 개의 기저가 필요하며 다음과 같다.

$$N_0(x) = 1, \quad N_1(x) = x, \quad N_k(x) = (x - \xi_k)_+, k = 1, 2, \ldots, K$$

3 스플라인 회귀 (Spline regression)

이제 주어진 구간에서 직선식이 아닌 p-차 다항식 (1) 을 고려하자. $K+1$개의 구간에서 p-차 다항식이 매유 부드럽게 연결되기 위한 조건은 연결점들에서 연속이며 더 나아가 $1, 2, .., p-1$차의 미분값이 동일한 것이다. 이러한 조건을 만족하는 최적의 p-차 다항식을 구하는 문제는 다음과 같은 $K+p+1$ 개의 기저로 표현된 함수식에서 최적의 함수를 구하는 문제와 같다. 아래의 기저들을 절단된 다항함수 (truncated power) 기저라고 부른다.

$$\begin{aligned} N_j(x) &= x^j, \quad j = 0, 1, 2\, dots, p \\ N_{p+k}(x) &= (x - \xi_k)_+^p, \quad k = 1, 2, \ldots, K \end{aligned} \tag{9}$$

위와 같은 스플라인 회귀에서 가장 자주 사용되는 것은 3차 스플라인 회귀(Cubic spline, $p = 3$)이다. 각 연결점에서 연속이고 1차와 2차 마분값이 같은 조건을 주고 각 구간에서 다항식을 구하는 것이다. 이때 구간의 하한과 상한, 즉 경계점(boundary)에서 직선의 성질을 가지는 조건, 즉

$$m''(0) = m'''(0) = 0, \quad m''(1) = m'''(1) = 0$$

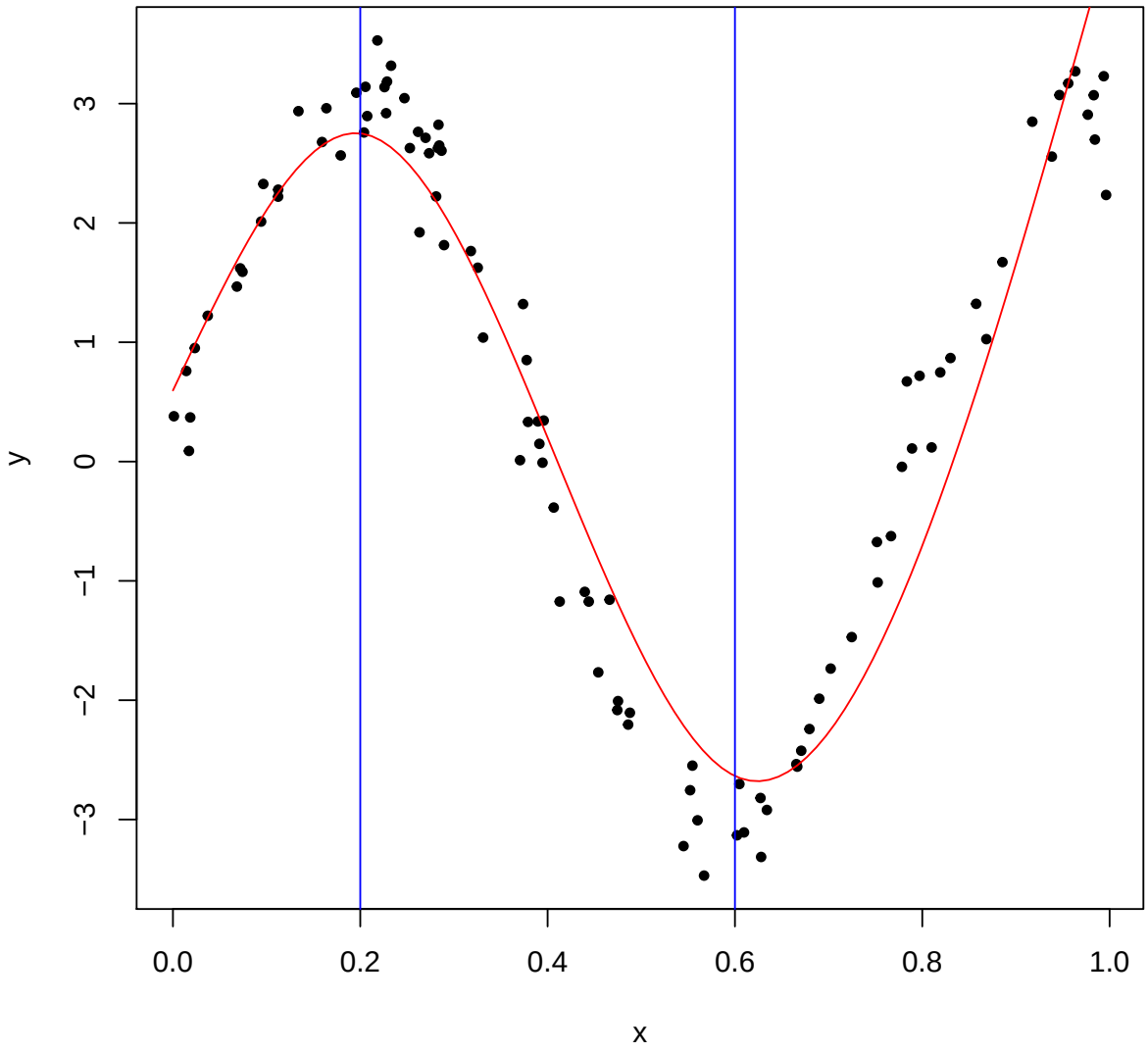
을 만족하는 스플라인을 자연 스플라인(natural spline)이리고 부르며 가장 자주 사용된다.

예제로서 다음과 같은 모형에서 생성된 자료를 가지고 자연 스플라인을 적합해 보자.

$$y_i = 3\sin(8x_i) + e_i, \quad e_i \sim N(0, (0.3)^2)$$

다음은 100개의 자료를 가지고 2개의 연결점 $0, 2$와 0.6을 이용한 3차 스플라인 함수를 적합하고 그리는 프로그램이다.

```
library(splines)
n = 100
x = runif(n)
x.grid = seq(0,1,length.out = 100)
y = 3*sin(8*x) + rnorm(n,0,.3)
fit = lm(y ~ ns(x,knots = c(0.2,0.6)) )
plot(x,y,pch=20)
pred  = predict(fit,newdata=list(x=x.grid))
lines(x.grid, pred,col="red")
abline(v=c(0.2,0.6), col = 'blue')
```



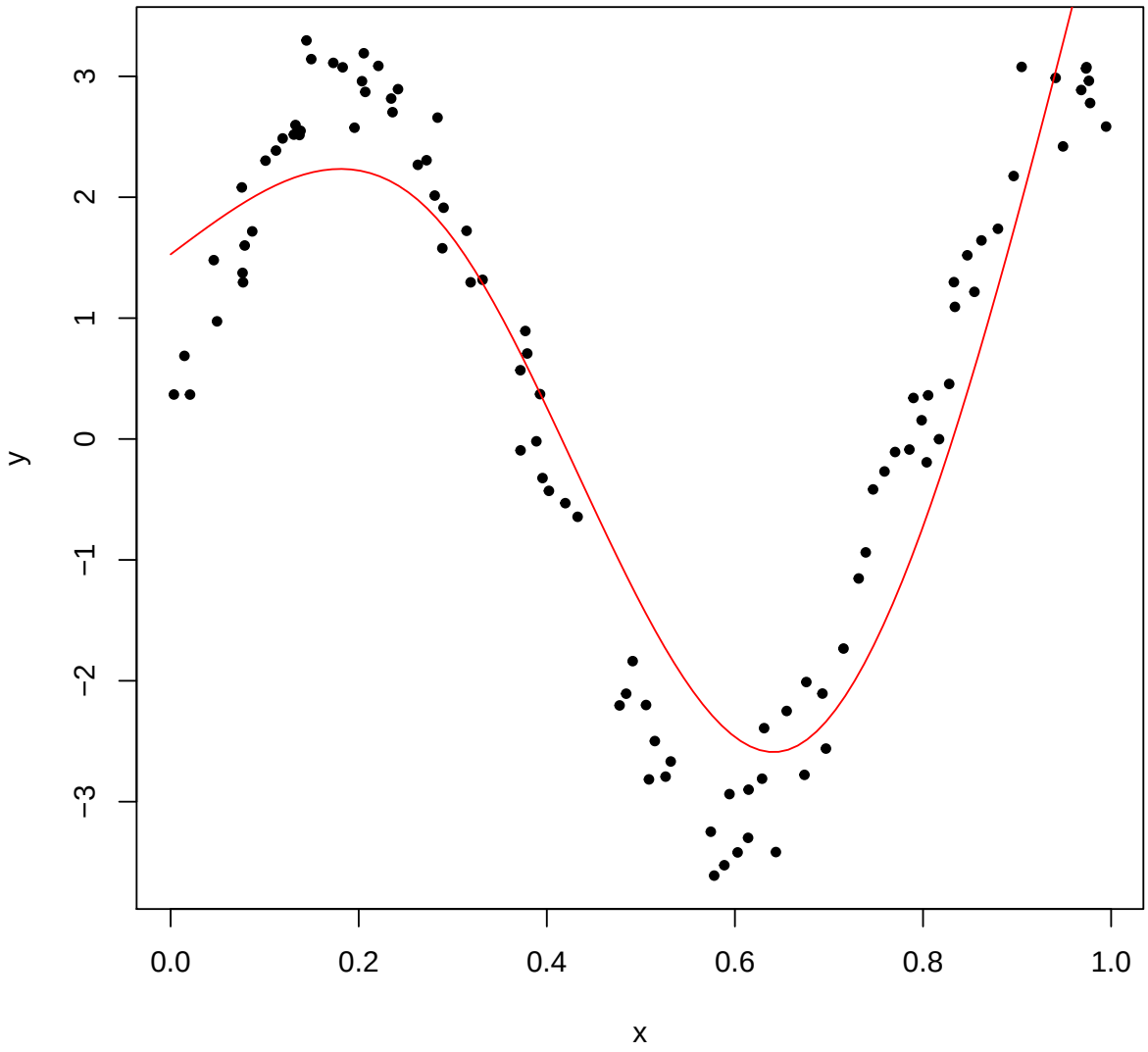
다음 프로그램에서는 연결점을 자동으로 선택하고 차수를 3차(자유도=3)로 사용하는 방법이다.

```
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
fit = lm(y ~ ns(x,df=3) )
summary(fit)

##
```

```
## Call:
## lm(formula = y ~ ns(x, df = 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1022 -0.6825  0.1623  0.6409  1.3008
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.5478     0.2809   5.510 3.01e-07 ***
## ns(x, df = 3)1  -8.1325     0.3378 -24.073  < 2e-16 ***
## ns(x, df = 3)2   1.1552     0.7088   1.630    0.106
## ns(x, df = 3)3   2.1136     0.2747   7.695 1.25e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8021 on 96 degrees of freedom
## Multiple R-squared:  0.8624,Adjusted R-squared:  0.8581
## F-statistic: 200.6 on 3 and 96 DF,  p-value: < 2.2e-16

plot(x,y,pch=20)
pred  = predict(fit,newdata=list(x=x.grid))
lines(x.grid, pred,col="red")
```



다음 프로그램에서는 25개의 연결점을 선택하고 차수를 3차(자유도=3)로 사용하는 방법이다. 과적합(overfitting)이 발생하였다.

```
kknots=seq(0+0.1,1-0.1,len=25)
x = runif(n)
y = 3*sin(8*x) + rnorm(n,0,.3)
fit = lm(y ~ ns(x,df=3,knots = kknots ))
```

```r
plot(x,y,pch=20)
pred  = predict(fit,newdata=list(x=x.grid))
lines(x.grid, pred,col="red")
```

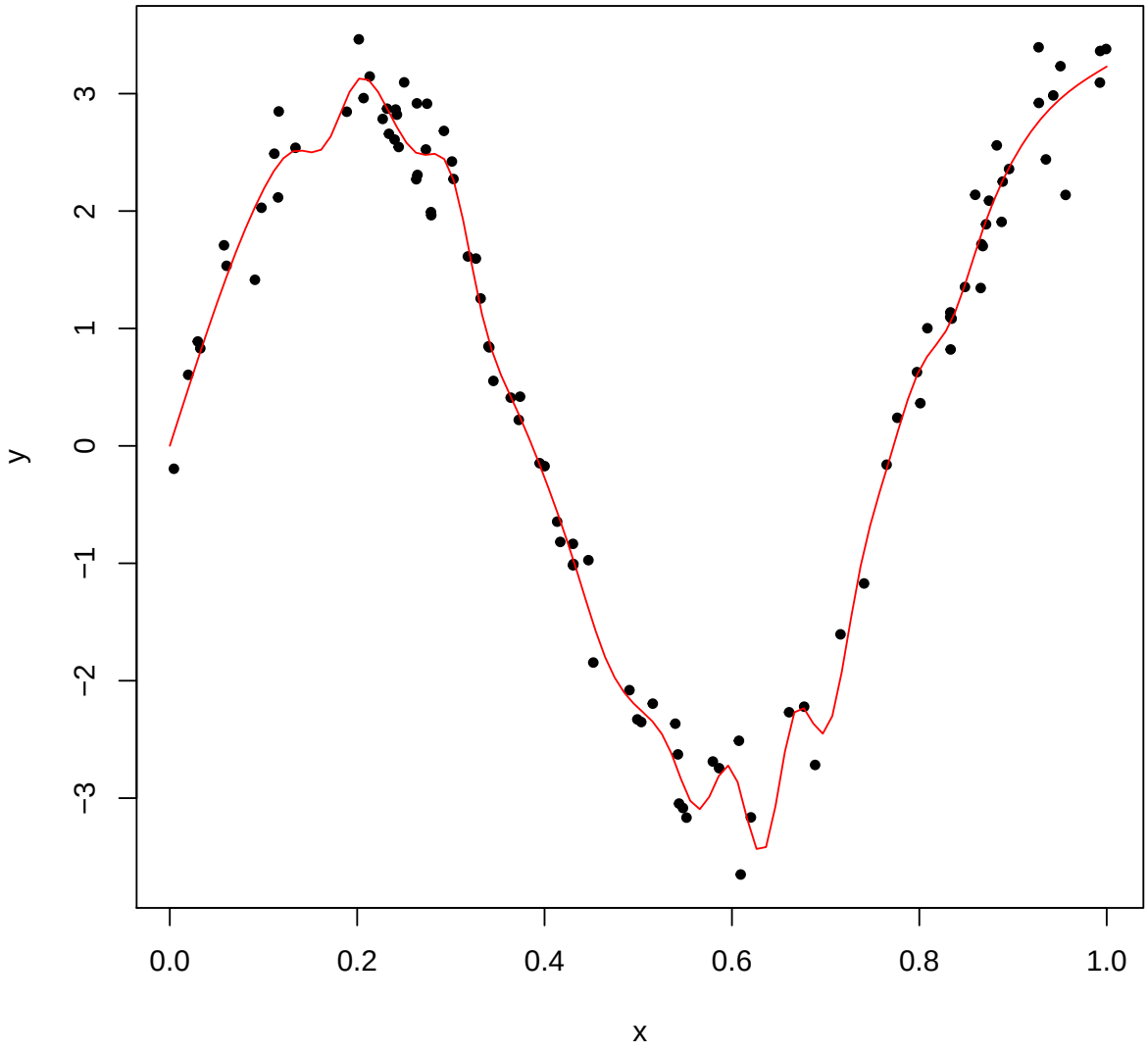


Figure 1: 연결점에서의 연속과 절단다항함수

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 5

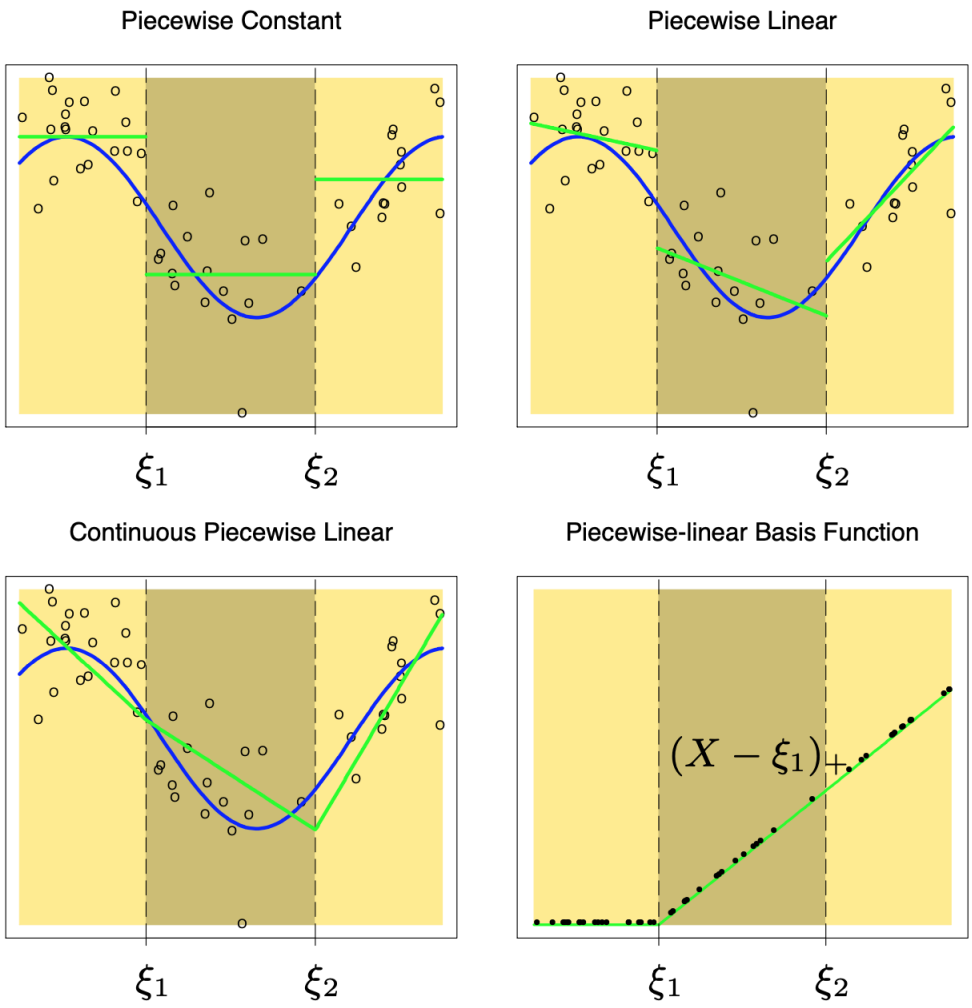


Figure 2: 직선스플라인 회귀에서 4개의 기저

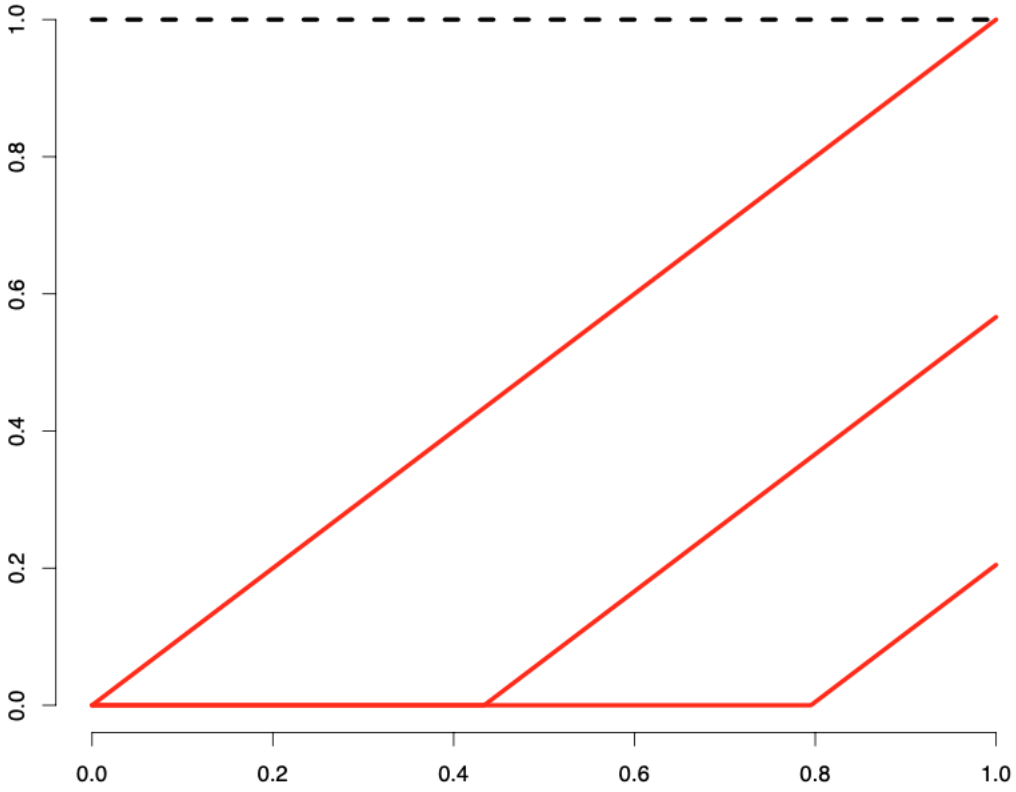


Figure 3: 3차 스플라인