

# 비선형함수추정

서울대학교  
통계학과

2018년 8월

# 다루는 내용 I

이 장에서는 선형회귀보다 일반적인 비선형 회귀모형을 다룬다. 다음과 같은 내용을 다룬다.

1. 다항회귀
2. 계단함수
3. 회귀 스플라인
4. 평활 스플라인
5. 국소회귀
6. 일반화가법모형

# 다항회귀

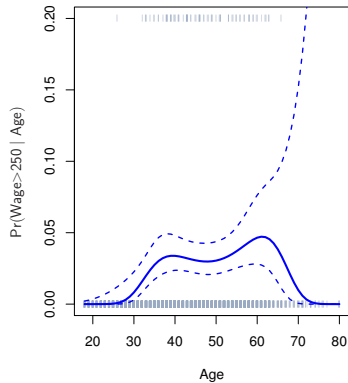
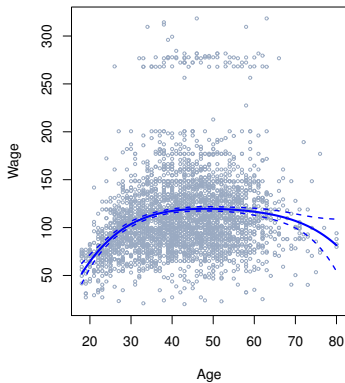
## $d$ 차 다항회귀모형

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), \quad i = 1, 2, \dots, n$$

## 특성들

1. 가장 간단하게  $x$ 의 비선형 회귀 함수를 표현할 수 있는 방법이다.
2. 4차가 넘어가면 함수의 모양이 너무 유연해져서, (특히 설명변수의 경계영역에서) 이상한 모양이 될 수 있다. 4차 이상의 모형은 잘 쓰지 않는다.

### Degree-4 Polynomial



# 노트. 다항회귀

## 그림

- ▶ ISLR 패키지에 있는 Wage 자료이다. 이 자료는 미국 중부아틀란틱 주(mid-atlantic states, 뉴잉글랜드와 남부 아틀란틱 사이의 주들로 New York, New Jersey, Pennsylvania, Delaware, Maryland, Washington D.C., Virginia, and West Virginia를 말한다. 기상에 관한 경우에 코네티컷을 포함하기도 한다.)들의 3000명의 연봉을 기록한 자료이다. 12개의 변수가 있다.

## ▶ 데이터셋의 변수들

1. year 연봉이 기록된 해
2. age 노동자의 나이
3. sex 성별
4. marital 결혼상태를 나타내는 팩터. 레벨의 의미는 다음과 같다.  
1. Never Married 2. Married 3. Widowed 4. Divorced and 5. Separated.
5. race 인종을 나타내는 팩터. 레벨의 의미 1. White 2. Black 3. Asian and 4. Other indicating race
6. education 교육을 나타내는 팩터. 레벨의 의미 1. < HS Grad 2. HS Grad 3. Some College 4. College Grad and 5. Advanced Degree
7. region 지역을 나타내는데, 이 경우 중부 아틀란틱만 있다.
8. jobclass 직종의 형태를 나타내는 팩터. 레벨의 의미 1. Industrial and 2. Information
9. health 건강상태를 나타내는 팩터. 레벨의 의미 1. <=Good and 2. >=Very Good
10. health\_ins 건강보험 가입 여부를 나타내는 팩터. 레벨의 의미 1. Yes and 2. No
11. logwage 연봉에 로그를 취한 값
12. wage 연봉

- ▶ 그림의 왼쪽은 4차 다항회귀모형을 적합한 그림이다. 점선은 추정량  $\pm 2 \times$  표준오차를 표현한다.
- ▶ 그림의 오른쪽은 25만달러 이상의 연봉을 받은 사람들의 확률을 로지스틱 모형으로 추정한 것이다. 즉,

$$\mathbb{P}(y > 250|x) = \frac{e^{\beta_0 + \beta_1 x + \dots + \beta_4 x^4}}{1 + e^{\beta_0 + \beta_1 x + \dots + \beta_4 x^4}}$$

# 다항회귀 R 코드 I

## 다항회귀모형의 적합

```
library(ISLR)
attach(Wage)
```

패키지를 로드한다. 이 예는 임금자료로 분석을 한다.

```
fit=lm(wage~poly(age,4),data=Wage)
coef(summary(fit))
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	111.70361	0.7287409	153.283015	0.000000e+00
## poly(age, 4)1	447.06785	39.9147851	11.200558	1.484604e-28
## poly(age, 4)2	-478.31581	39.9147851	-11.983424	2.355831e-32
## poly(age, 4)3	125.52169	39.9147851	3.144742	1.678622e-03
## poly(age, 4)4	-77.91118	39.9147851	-1.951938	5.103865e-02



## 다항회귀 R 코드 II

poly 함수는 주어진 변수와 차수로 직교다항함수기저를 이용해 계획행렬을 생성한다.

```
fit2=lm(wage~poly(age,4,raw=T),data=Wage)
coef(summary(fit2))
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	-1.841542e+02	6.004038e+01	-3.067172	0.0021802539
## poly(age, 4, raw = T)1	2.124552e+01	5.886748e+00	3.609042	0.0003123618
## poly(age, 4, raw = T)2	-5.638593e-01	2.061083e-01	-2.735743	0.0062606446
## poly(age, 4, raw = T)3	6.810688e-03	3.065931e-03	2.221409	0.0263977518
## poly(age, 4, raw = T)4	-3.203830e-05	1.641359e-05	-1.951938	0.0510386498

직교다항함수기저가 아니라 다항함수 그 자체로 계획행렬을 생성하려면 아래와 같은 옵션을 쓰면 된다.

```
fit2a=lm(wage~age+I(age^2)+I(age^3)+I(age^4),data=Wage)
coef(fit2a)
fit2b=lm(wage~cbind(age,age^2,age^3,age^4),data=Wage)
```

위와 동일한 결과를 준다.

# 다항회귀 R 코드 III

## 예측과 그림 그리기

```
agelims=range(age)
age.grid=seq(from=agelims[1],to=agelims[2])
```

변수 age의 최대값과 최소값을 구하고, 그 것을 이용해 격자값을 구한다. 이 격자값들에서 예측값을 생성하려고 한다.

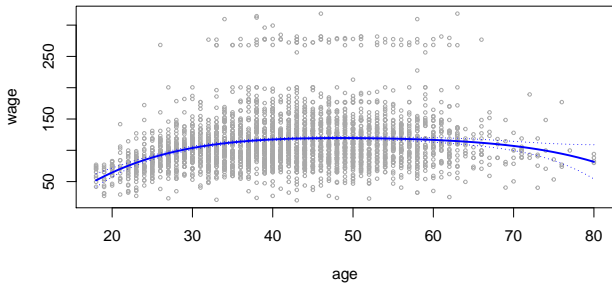
```
preds=predict(fit,newdata=list(age=age.grid),se=TRUE)
se.bands=cbind(preds$fit+2*preds$se.fit,preds$fit-2*preds$se.fit)
```

격자값에서 예측값을 구하고, 함수추정량의 표준오차를 이용해서 신뢰구간의 밴드를 구했다.

```
plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
title("Degree-4 Polynomial",outer=T)
lines(age.grid,preds$fit,lwd=2,col="blue")
matlines(age.grid,se.bands,lwd=1,col="blue",lty=3)
```

# 다항회귀 R 코드 IV

Degree-4 Polynomial



# 다항회귀 R 코드 V

## 노트: 그림설명

1. `cex`는 글자와 심볼의 척도를 나타낸다. 1은 디폴트 크기이고 0.5는 디폴트 크기의 절반, 1.5는 1.5배인 것이다.
2. `lwd`는 선의 굵기를 나타내는 척도로 1은 디폴트 굵기, 1.5는 디폴트의 1.5배를 의미한다.

# 계단함수를 이용한 회귀모형 I

## $x$ 를 범주형변수들로 변환

$c_1 < c_2 < \dots < c_K$ 를  $x$ 의 범위의 구분점들이라 하자. 다음과 같이 범주형 변수들을 정의한다.

$$C_0(x) = I(x < c_1)$$

$$C_1(x) = I(c_1 \leq x < c_2)$$

$\vdots$

$$C_{K-1}(x) = I(c_{K-1} \leq x < c_K)$$

$$C_K(x) = I(c_K \leq x).$$

## 계단함수를 이용한 회귀모형

## 계단함수를 이용한 회귀모형 II

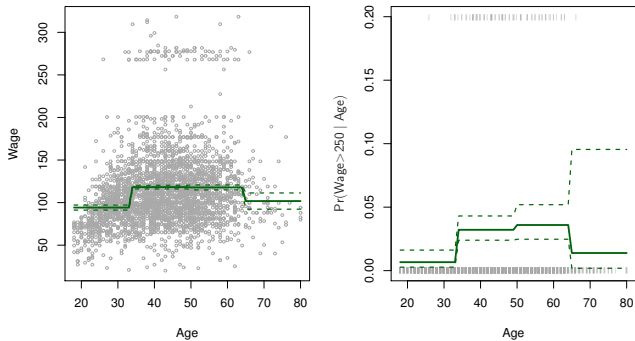
$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \dots + \beta_K C_K(x_i) + \epsilon_i, \\ \epsilon_i \sim (0, \sigma^2), i = 1, 2, \dots, n$$

### 참고사항

1.  $C_0(x) + C_1(x) + \dots + C_K(x) = 1, \forall x$  이어서, 모형에서  $C_0(x)$ 는 포함하지 않았다.

# 계단함수를 이용한 회귀모형 III

Piecewise Constant



# 노트. 계단함수 I

## 그림

1. 그림의 왼쪽은 계단함수를 이용한 회귀모형을 적합한 그림이다. 점선은 추정량  $\pm 2 \times$  표준오차를 표현한다.
2. 그림의 오른쪽은 25만달러 이상의 연봉을 받은 사람들의 확률을 계단함수를 이용한 로지스틱 모형으로 추정한 것이다. 즉,

$$\mathbb{P}(y > 250|x) = \frac{e^{\beta_0 + \beta_1 C_1(x) + \dots + \beta_4 C_4(x)}}{1 + e^{\beta_0 + \beta_1 C_1(x) + \dots + \beta_4 C_4(x)}}$$



# 계단함수를 이용한 회귀모형 R 코드 I

```
table(cut(Wage$age,4))
```

```
##  
## (17.9,33.5] (33.5,49] (49,64.5] (64.5,80.1]  
##          750      1399        779        72
```

```
fit=lm(wage~cut(age,4),data=Wage)  
coef(summary(fit))
```

```
##              Estimate Std. Error  t value    Pr(>|t|)  
## (Intercept)    94.158392    1.476069  63.789970 0.000000e+00  
## cut(age, 4) (33.5,49]    24.053491    1.829431  13.148074 1.982315e-38  
## cut(age, 4) (49,64.5]    23.664559    2.067958  11.443444 1.040750e-29  
## cut(age, 4) (64.5,80.1]    7.640592    4.987424   1.531972 1.256350e-01
```

## 노트.

cut(age, 4)는 변수 age를 4개의 영역으로 나눠준다. 격자의 분리점을 정하고 싶으면 breaks 옵션을 쓰면 된다.

# 기저함수(basis function)를 이용한 회귀모형 I

## 기저함수를 이용한 회귀모형

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \dots + \beta_K b_K(x_i) + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), \quad i = 1, 2$$

여기서,  $b_1, \dots, b_K$  는 기저함수이다.

## 참고사항

1. 다항회귀모형과 계단함수를 이용한 회귀모형도 기저함수를 이용한 회귀모형으로 볼 수 있다.
2. 이 외에도 푸리에 기저(Fourier basis), 웨이블릿 기저(wavelet basis) 등을 사용할 수 있다.

# 회귀스플라인(regression spline) I

## 동기

1. 다항회귀모형의 문제점 : 다항회귀모형을 이용해서 모형의 유연성을 높이려면 다항식의 차수를 높여야 한다. 그런데 차수를 높이면 원치 않는 모양의 이상한 회귀함수 모양이 나타날 수 있다. 이는 어떤 관측치가 멀리 떨어진 곳의 함수 추정량에도 영향을 미치기 때문이다.
2. 계단함수의 문제점 : 이를 극복하기 위해서는 계단함수 같은 기저함수를 이용하면 된다. 그런데, 계단함수는 연속함수에 적용할 수 없다.
3. 스플라인 모형: 위의 문제점을 해결하기 위해 매듭(knot)으로 나누어진 구간에 낮은 차원의 다항식 모형을 적합하는 것을 고려한다. 조각별 다항식(piecewise polynomial)을 스플라인이라 한다.

# 회귀스플라인(regression spline) II

## 회귀스플라인

1. 차수가  $d$ 이고 매듭이  $\xi_1, \dots, \xi_K$ 인 스플라인이란 매듭으로 이루어진 각 구간에서 차수가  $d$ 인 다항식이고  $d - 1$ 차 도함수가 연속인 함수를 말한다. 이것들을 회귀스플라인이라고 한다.
2. 3차 스플라인(cubic spline) : 차수가  $d = 3$ 인 스플라인. 각 구간이 3차 다항식이고, 2차 도함수가 연속. 매듭의 불연속성이 사람의 눈에 보이지 않은 가장 차수가 낮은 스플라인. 이 이상 높은 차수는 사용할 필요가 없다.
3. 1차 스플라인(linear spline) :  $d = 1$ 인 스플라인.
4. 보통  $d = 0, 1, 3$ 이 가장 많이 사용된다.

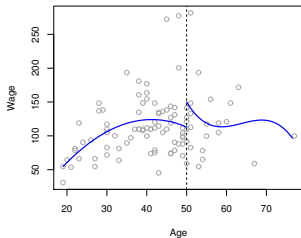
# 회귀스플라인(regression spline) III

## 노트. 참고.

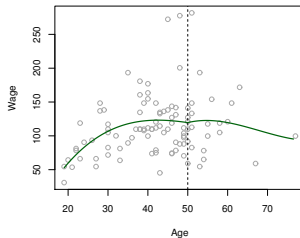
$K$  개의 매듭을 가진 3차 스플라인은  $K + 4$  개의 기저가 필요하다.  $\xi_1$  보다 작은 구간에서 3차 다항식을 나타내는 4개의 기저가 필요하다. 매듭이 하나 추가될 때 마다 기저가 하나씩 필요하다. 도함수 조건을 고려하지 않으면, 매듭이 하나 추가 될 때, 새로운 3차 다항식이 필요하므로 기저가 4개가 필요하다. 그런데, 2차 도함수까지 연속이어야 하므로, 세 가지 제한조건이 생긴다. 따라서 매듭이 한 개 생길 때마다, 한 개의 기저가 필요하다.

# 회귀스플라인(regression spline) IV

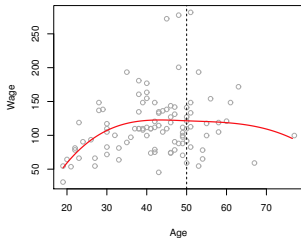
**Piecewise Cubic**



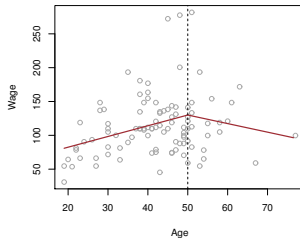
**Continuous Piecewise Cubic**



**Cubic Spline**



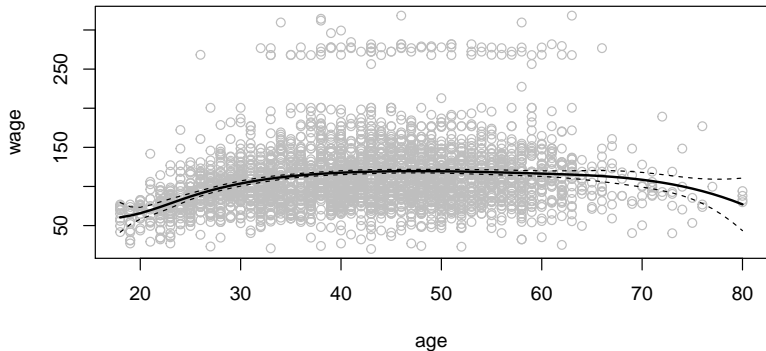
**Linear Spline**



# 회귀 스플라인 R 코드 I

```
library(splines)
fit=lm(wage~bs(age,knots=c(25,40,60)),data=Wage)
pred=predict(fit,newdata=list(age=age.grid),se=T)
plot(age,wage,col="gray")
lines(age.grid,pred$fit,lwd=2)
lines(age.grid,pred$fit+2*pred$se,lty="dashed")
lines(age.grid,pred$fit-2*pred$se,lty="dashed")
```

## 회귀 스플라인 R 코드 II



bs는 회귀스플라인 기저를 예측변수로 갖는 계획행렬을 생성한다. 이를 이용해 선형회귀모형을 적용해서 회귀스플라인 모형을 적합한다.



# 회귀 스플라인 R 코드 III

```
dim(bs(age,knots=c(25,40,60)))
```

```
## [1] 3000    6
```

```
dim(bs(age,df=6))
```

```
## [1] 3000    6
```

```
attr(bs(age,df=6),"knots")
```

```
##    25%    50%    75%
```

```
## 33.75 42.00 51.00
```

bs 함수에서 자유도는 매듭의 개수를 결정한다. 자유도만 주어져도 계획행렬을 구할 수 있다.

# 자연 스플라인(natural spline) I

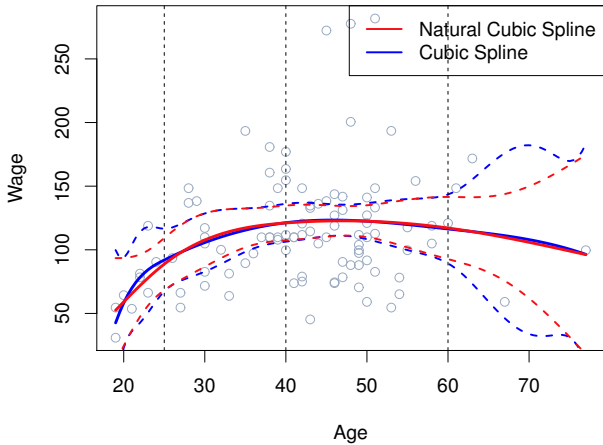
## 자연 스플라인(natural spline)

1. 자연스플라인은 3차 스플라인에서  $x < \xi_1$  혹은  $x > \xi_k$  일 때, 1차 다항식으로 대치한 것을 말한다.
2. 회귀스플라인은 경계 부근에서 큰 분산을 갖는다.  
자연스플라인은 경계부근에서 안정된 분산을 갖게 한다.

## 노트.

자연스플라인이 경계부근에서 안정된 분산을 갖는 이유.  
경계에서 3차회귀모형 대신에 1차모형을 적합해서 모형의 유연성을 줄였다. 따라서, 편향을 크게하는 대신, 분산을 줄이게 된다.

## 자연 스플라인(natural spline) II



# 자연 스플라인(natural spline) III

## 자연 3차 스플라인의 기저표현

$K$ 개의 매듭이 있을 때,  $K$ 개의 기저가 필요하다. 3차 스플라인은  $K + 4$ 개의 기저가 필요한데, 가장 작은 구간과 가장 큰 구간에서 2개씩의 제한 조건이 있어서 4를 빼줘야 한다.  $K$ 개의 기저는 다음과 같다.

$$N_1(x) = 1, \quad N_2(x) = x, \quad N_{k+2}(x) = d_k(x) - d_{K-1}(x), \quad k = 1, 2, \dots, K-$$

여기서,

$$d_k(x) := \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k}$$

이다.  $N_k$ 의 2차와 3차 도함수는  $x > \xi_K$ 에서 0이다.

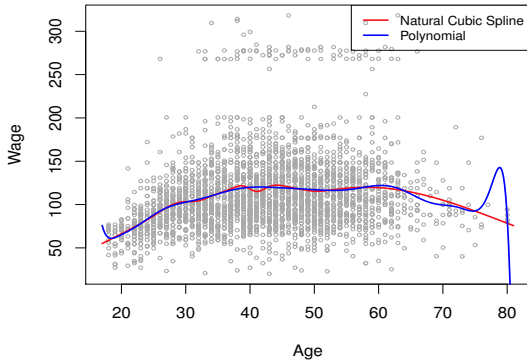
# 자연 스플라인(natural spline) IV

## 매듭의 위치와 개수를 정하는 법

1. 위치. 매듭이 촘촘하게 있는 구간에서 스플라인은 더 유연하다. 대개의 경우 설명변수의 분위수를 이용해 정한다.  $K$  개의 매듭의 위치를 정하는 경우,  $x_{(n\frac{i}{K+1})}, i = 1, 2, \dots, K$  에 정한다.
2. 매듭의 개수  $K$  는 교차검증방법을 이용해서 정한다.

# 자연 스플라인(natural spline) V

## 다항회귀모형과 스플라인모형의 비교



## 자연 스플라인(natural spline) VI

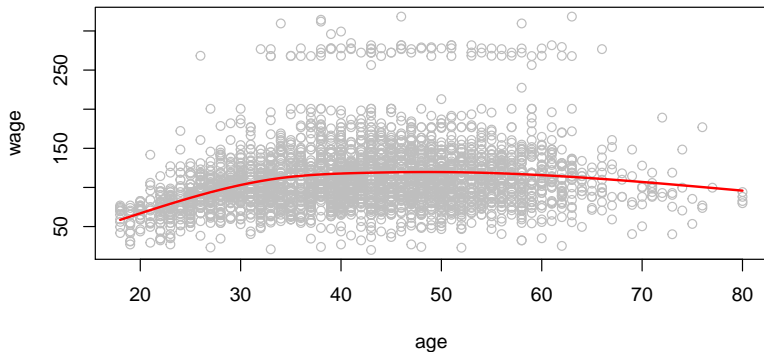
자유도가 15인 다항회귀모형과 자연스플라인 모형이 비교되었다. 다항회귀인 경우 경계부근에서 원치않는 효과가 있다.

# 자연 스플라인 R 코드 I

```
fit2=lm(wage~ns(age,df=4),data=Wage)
pred2=predict(fit2,newdata=list(age=age.grid),se=T)
plot(age,wage,col="gray")
lines(age.grid, pred2$fit,col="red",lwd=2)
```



## 자연 스플라인 R 코드 II



함수 `ns`는 자연스플라인기저로 생성되는 계획행렬을 생성한다. 자연스플라인모형은 이 기저에 `lm` 함수를 적용한다.

# 평활스플라인(smoothing spline) I

## 정의

$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(x)^2 dx$$

를 최소화 하는 함수  $g$ 를 평활스플라인이라고 한다.

# 평활스플라인(smoothing spline) II

## 설명

1.  $\lambda$  : 조율파라미터(tuning parameter)
2.  $\sum_{i=1}^n (y_i - g(x_i))^2$  : 손실함수(loss function)
3.  $\int g''(x)^2 dx$  : 벌점(penalty). 이차도함수는 함수의 구불구불한 정도를 나타낸다.

# 평활스플라인(smoothing spline) III

## 평활스플라인은 자연3차스플라인

평활스플라인은 매듭이 중복되지 않는(distinct)  $x_1, \dots, x_n$  인 자연3차스플라인이다.

## 유효자유도(effective degree of freedom)

$g$ 가 회귀함수일 때,  $(g(x_1), \dots, g(x_n))'$ 의 추정값을

$$\hat{g}_\lambda = S_\lambda y$$

와 같이 나타낼 수 있다. 이 때, 유효자유도는

$$df_\lambda := \text{tr}(S_\lambda)$$

로 정의된다.

# 평활스플라인(smoothing spline) IV

## 노트. 참고.

1. 평활스플라인이 자연스플라인이기 때문에, 평활스플라인의 자유도는 중복되지않는  $x_i$ 들의 개수( $n$ )라고 생각할 수 있지만 그렇지 않다.  $\lambda$ 가 0에서  $\infty$ 로 움직이면서 자유도는 0에서  $n$ 까지 움직인다.
2.  $S_\lambda$ 는 모자행렬로 선형회귀모형에서  $H = X'(X'X)^{-1}X$ 이다. 이 때,  $tr(H) = p + 1$ 로 회귀계수의 개수와 같다.

# 평활스플라인(smoothing spline) V

## 조율파라미터의 결정

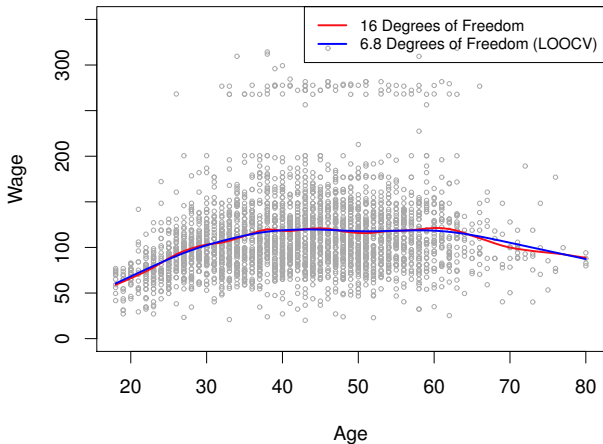
$\lambda$ 의 값은 교차검증을 이용해 결정할 수 있다. 그런데, 평활스플라인에서 하나빼기교차검정오차는 다음과 같이 수식이 알려져 있어, 빠르게 계산할 수 있다.

$$RSS_{cv}(\lambda) := \sum_{i=1}^n (y_i - \hat{g}_{\lambda}^{(-i)}(x_i))^2 = \sum_{i=1}^n \left[ \frac{y_i - \hat{g}_{\lambda}(x_i)}{1 - (S_{\lambda})_{ii}} \right]^2$$

여기서,  $\hat{g}_{\lambda}^{(-i)}(x)$ 는  $i$ 번째 관측치를 제외하고 구축한 함수의 추정치이고,  $(S_{\lambda})_{ii}$ 는  $S_{\lambda}$ 의  $i$ 번째 대각행렬이다.

# 평활스플라인(smoothing spline) VI

Smoothing Spline



# 평활 스플라인 R 코드 I

```
plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
title("Smoothing Spline")
fit=smooth.spline(age,wage,df=16)
fit2=smooth.spline(age,wage,cv=TRUE)
```

```
## Warning in smooth.spline(age, wage, cv = TRUE): cross-validation with
non-unique 'x' values seems doubtful
```

```
fit2$df
```

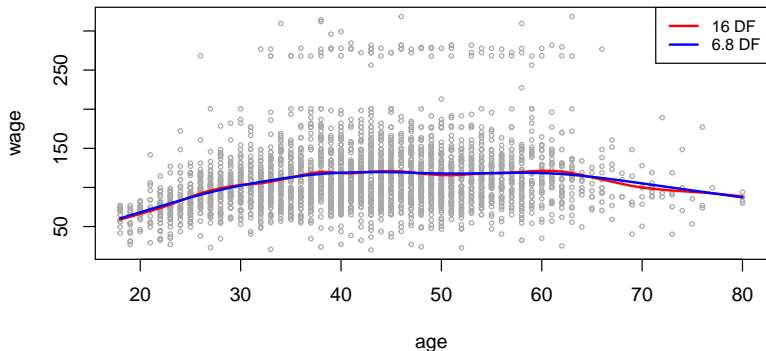
```
## [1] 6.794596
```

```
lines(fit,col="red",lwd=2)
lines(fit2,col="blue",lwd=2)
legend("topright",legend=c("16 DF","6.8 DF"),col=c("red","blue"),lty=1,lwd=2,cex=
```



## 평활 스플라인 R 코드 II

Smoothing Spline



평활스플라인은 `smooth.spline` 함수를 이용해서 적합한다.  
fit은 `df=16`이 주어져서 조율파라미터가 가 결정이 되었다.  
fit2는 교차검정을 이용해서 조율파라미터가 결정되었다.

# 국소회귀(local regression) I

$x = x_0$ 에서  $\hat{g}$ 의 계산

$x_0$ 에 가까운  $s = k/n$ 개의 관측치에 가중치  $K(x_i, x_0)$ 를 주고 1차 회귀모형을 적합한다. 즉,

$$\sum_{i=1}^n K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i)^2$$

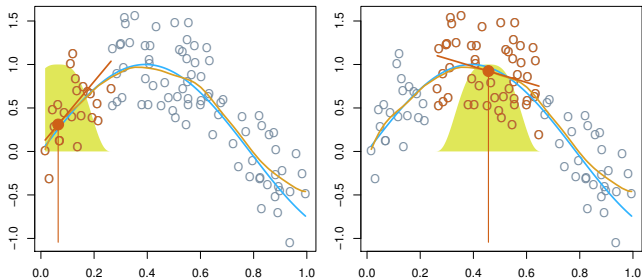
을 최소화하여  $\beta_0$ 와  $\beta_1$ 을 얻는다. 예측치는

$$\hat{f}(x_0) := \hat{\beta}_0 + \hat{\beta}_1 x_0$$

가 된다. 위에서 선형회귀 대신  $p$ 차 회귀모형을 쓸 수 있다. 위에서  $s$ 를 펼침(span)이라 부른다.

# 국소회귀(local regression) II

Local Regression



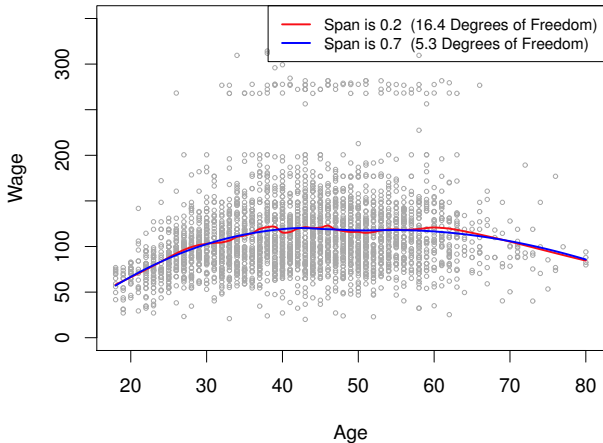
# 국소회귀(local regression) III

## 펼침 $s$ 의 결정

국소회귀를 적합할 때, 가중치  $K$ , 회귀모형의 차수, 펼침  $s$ 를 결정해야 한다. 이 중  $s$ 의 결정이 제일 중요하다.  $s$ 는 모형의 유연성을 결정하고, 교차검증으로 결정할 수 있다.

# 국소회귀(local regression) IV

## Local Linear Regression



# 국소회귀(local regression) V

## 국소회귀모형의 확장

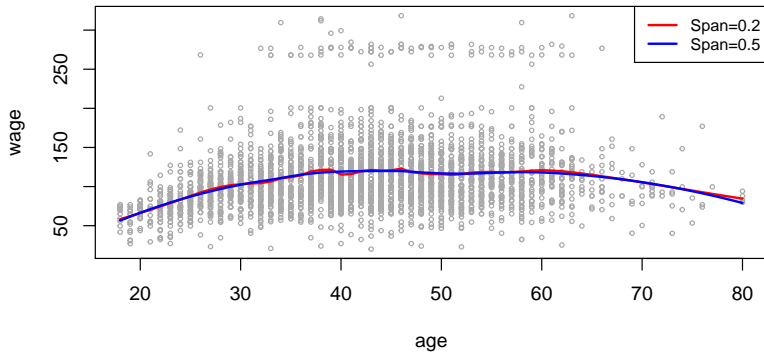
1. 일부 설명변수에는 국소회귀모형을 쓰고, 일부 설명변수에는 전역회귀모형을 쓸 수 있다. 시간이 설명변수일 때, 시간을 국소회귀모형으로 쓰면 시간변동계수모형(time varying coefficient model)이 된다.
2. 국소회귀모형은 고차원 자료에 확장이 잘 된다.
3. 회귀모형의 차수  $p$ 가 3, 4보다 크면 성능이 나빠질 수 있다.

# 국소회귀 R 코드 I

```
plot(age, wage, xlim=agelims, cex=.5, col="darkgrey")
title("Local Regression")
fit=loess(wage~age, span=.2, data=Wage)
fit2=loess(wage~age, span=.5, data=Wage)
lines(age.grid, predict(fit, data.frame(age=age.grid)), col="red", lwd=2)
lines(age.grid, predict(fit2, data.frame(age=age.grid)), col="blue", lwd=2)
legend("topright", legend=c("Span=0.2", "Span=0.5"), col=c("red", "blue"), lty=1, lwd=2)
```

# 국소회귀 R 코드 II

Local Regression





# 일반화가법모형(generalized additive models) I

## 모형

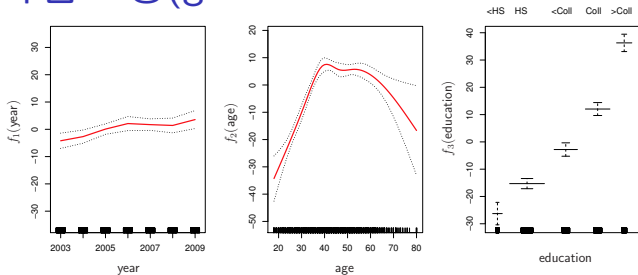
여러 개의 설명변수가 있을 때, 각 변수에 1변수 회귀함수를 적용하는 것을 말한다.

$$y_i = \beta_0 + f_1(x_{i1}) + \dots + f_p(x_{ip}) + \epsilon_i, \quad i = 1, 2, \dots, n.$$

## 연봉자료의 예

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

# 일반화가법모형(generalized additive models) II



# 일반화가법모형(generalized additive models) III

## 역적합(backfitting)

설명변수 하나씩 돌아가면서 나머지 설명변수와 그의 적합한 함수를 고정한 후, 한 개의 변수에 관해서만 회귀함수를 구하는 방법

## 노트. 참고.

평활스플라인의 경우 최소제곱법을 사용할 수 없기 때문에, 일반화가법모형 전체에 최소제곱법을 적용하는 것이 어렵다.

# 일반화가법모형(generalized additive models) IV

## 일반화가법모형의 장점과 단점

1. 각 설명변수에 비선형 함수  $f_j$ 를 적용하기 때문에, 설명변수를 변환할 필요가 없다.
2. 설명변수의 비선형 함수는 예측 성능을 향상시킬 수 있다.
3. 모형이 가법이기 때문에, 다른 설명변수들을 고정시켰을 때, 한 설명변수의 효과가  $f_j$ 이다.
4.  $f_j$ 의 유연성은 자유도로 요약된다.
5. (단점) 가법모형이기 때문에 교호작용을 표현하지 못한다. 하지만, 교호작용을 표현하고 싶으면  $f(x_i, x_j)$ 를 적합하면 된다.

# 일반화가법모형(generalized additive models) V

## 이산형 반응변수와 가법모형

반응변수가 이산형일 때도 가법모형을 적용할 수 있다.

$$\log \left( \frac{p(x)}{1 - p(x)} \right) = \beta_0 + f_1(x_1) + \dots + f_p(x_p)$$

# 일반화 가법모형 R 코드 I

```
gam1=lm(wage~ns(year,4)+ns(age,5)+education,data=Wage)
```

예측변수 모두가 자연스플라인 기저로 구성된 가법모형을 적용하려면 lm을 이용하면 된다. 가법모형이 큰 선형모형일 뿐이기 때문이다.

```
library(gam)
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.12
```

```
gam.m3=gam(wage~s(year,4)+s(age,5)+education,data=Wage)
```

하나 이상의 예측변수에 평활스플라인을 적용하려면 gam 패키지의 함수 gam이 필요하다. 함수 gam 안의 함수 s는 평활스플라인의 기저를 생성하는 함수로 gam 패키지 안에 있는 함수이다.

# 일반화 가법모형 R 코드 II

```
summary(gam.m3)

##
## Call: gam(formula = wage ~ s(year, 4) + s(age, 5) + education, data = Wage)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -119.43  -19.70   -3.33   14.17   213.48
##
## (Dispersion Parameter for gaussian family taken to be 1235.69)
##
##      Null Deviance: 5222086 on 2999 degrees of freedom
## Residual Deviance: 3689770 on 2986 degrees of freedom
## AIC: 29887.75
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## s(year, 4)    1   27162   27162   21.981 2.877e-06 ***
## s(age, 5)     1  195338  195338  158.081 < 2.2e-16 ***
## education    4 1069726  267432  216.423 < 2.2e-16 ***
## Residuals   2986 3689770    1236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##              Npar Df Npar F    Pr(F)
## (Intercept)
## s(year, 4)    3   1.086 0.3537
## s(age, 5)     4  32.380 <2e-16 ***
## education
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 참고문헌

아래의 책에서 제공하는 그림들을 사용하였다.

1. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. [An introduction to statistical learning.](#) Springer, 2013.