CH7 Moving Beyong Linearity

이전까지 내용은 linear model에 초점을 맞춰왔다.
linear model은 해석하기 쉽다는 장점이 있지만 예측력이 떨어진다.

이번 장에서는 linearity assumption을 완화해도
해석력은 유지되는 모델들이 대해 알아보자.

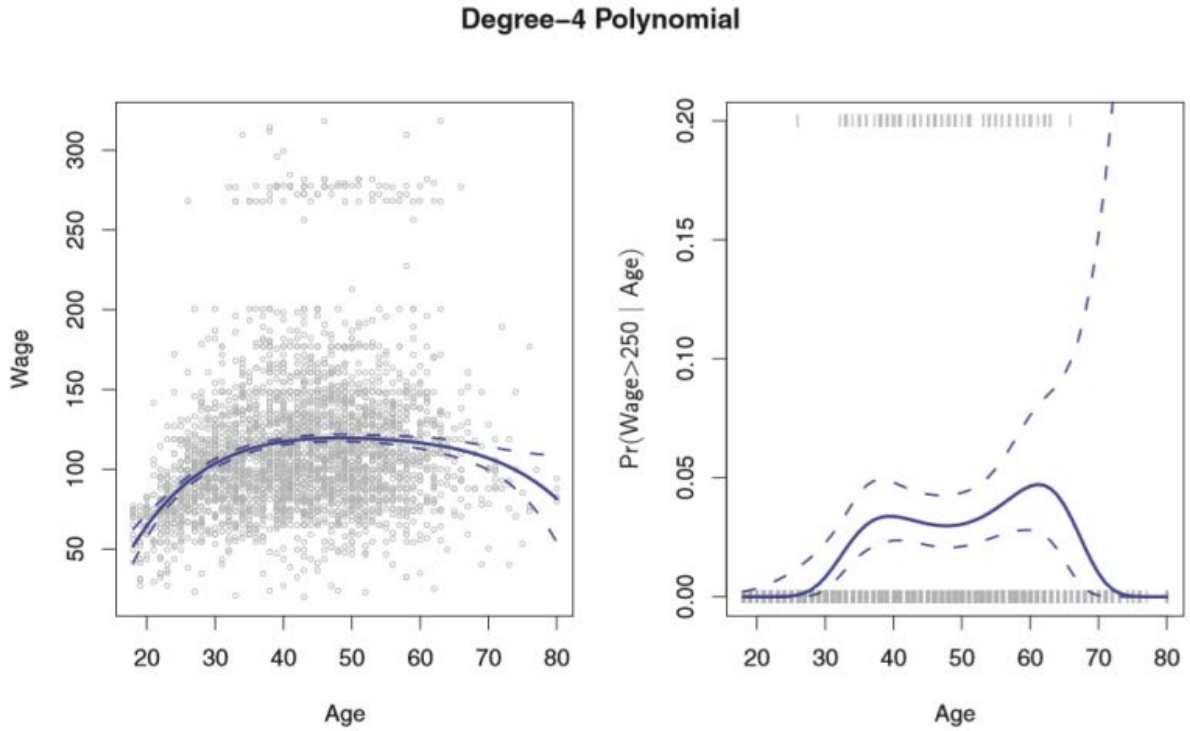
- Polynomial regression
- Step fuction
- Regression splines
- Smoothing splines
- Local regression
- Generalized addtive model(GAM)

1. Polynomial regression

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

위의 수식은 설명변수가 xi, xi^2, xi^3, ... , xi^d의 표준적인 선형모델이기 때문에 그 계수들은 최소제곱 선형회귀를 사용하여 쉽게 추정할 수 있다.

일반적으로는 3또는 4보다 큰 값의 d를 사용하는 경우는 드물다. 왜냐하면, 다항식 곡선이 지나치게 유연해지기 때문에 아주 이상한 형태를 가질 수 있기 때문. 이것은 X변수의 경계 근처에서 특히 심함.



왼쪽 그림은 Wage자료에서 나이(Age)에 대한 임금(Wage)을 나타낸 그래프
최소제곱을 사용하여 4 차 다항식을 적합한 결과가 파란색 실선
(추정치의 분산은 각 계수의 추정치의 분산과 공분산 이용해서 계산)
점선으로 된 곡선의 쌍은 (2x) 표준오차 곡선
표준오차의 2 배 값을 그래프로 나타내는 이유는 이 값이 정규분포의 오차항들에 대해 95% 신뢰구간의 근사치에 해당하기 때문

위 그림을 보면 25 만 달러보다 더 많은 수입이 있는 고소득자 그룹과 저소득자 그룹이 있는 것 같음. wage 를 이들 2 개 그룹으로 분할하고 예측하는데 age 의 다항식 함수를 설명변수로 이용하는 로지스틱 회귀를 사용

다시 말하면 다음의 모델을 적합

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

이 자료의 표본크기는 상당히 크지만(n = 3,000) 고소득자는 고작 79 명이다. 이 때문에 추정된 계수들의 분산이 높고 그 결과 신뢰구간이 넓다.

2. Step fuction

global 한 구조를 피하기 위해 step function 을 사용

X 의 범위에 c1, c2, ..., ck 의 절단점(cutpoint)을 사용하여 K+1 개의 새로운 변수를 만든다.

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

여기서 I 함수는 조건이 참이면 1, 그렇지 않으면 0 을 반환하는 Indicator fucntion 이다.

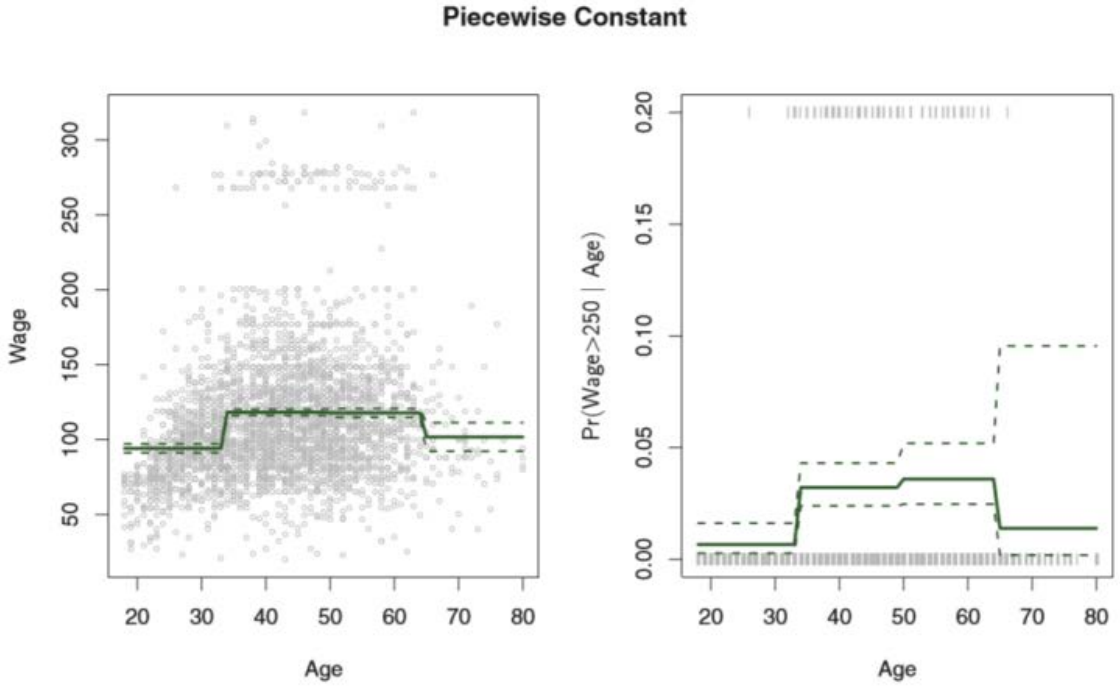
예를 들어, I(ck < X)은 ck < X 이면 1, 그렇지 않으면 0

임의의 X 에 대해, X 는 구간 중 정확히 어느 하나에 속해야 하므로,

C0(X) + C1(X) + ... + CK(X) =1 이다.

그 다음에 최소 제곱법을 사용하여 선형모델을 적합하여 C1(X), C2(X), ..., CK(X)를 설명변수로 사용한다.

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



$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}$$

Step fuction 그 자체로는 잘 사용되지 않고, 이것의 응용형태가 사용된다고 함

Basis Functions

사실 위의 polynomial regression 과 step-function regression 은 basis function 의 특별한 경우

변수 X 에 적용될 수 있는 함수 또는 변환들 b1(X), b2(X), ..., bK(X) 을 가지는 것

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

기저함수들 b1(X), b2(X), ..., bK(X)은 정해져 있음(선택)

Polynomial regression 의 경우 기저함수 bj(xi) = xi^j

step-function regression 의 경우 기저함수 bj(xi) = I

위 모델을 기저함수 bk(X)을 predictor 로 하는 standard linear model 로 볼 수 있다.

Basis function 에는 많은 방법들이 있고, 기저함수로 매우 자주 선택되는 spline 이 있다.

3. Regression splines

3.1 Piecewise Polynomials

X 의 전체 범위에 걸쳐 고차원 다항식을 적합하는 방법 대신,

X 의 범위를 구분하여 각 범위에 저차원의 다항식을 적합하는 방식

cubic regression model of the form

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

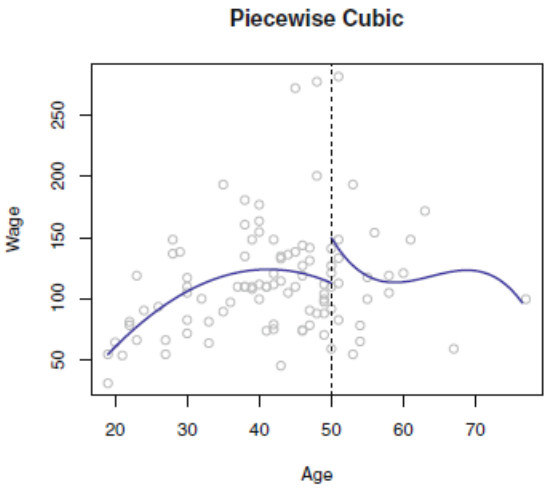
여기서 계수 β는 x 의 범위에 따라 값이 달라지게 된다.

이렇게 계수의 값이 변하게 되는 지점을 knots 라고 함.

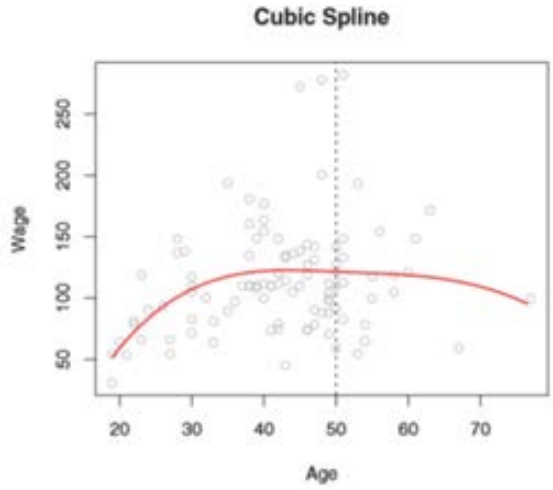
Ex) 점 c 에서 single knot 를 갖는 cubic polynomial

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

마찬가지로 각 계수들은 최소 제곱이용해서 fitting 가능



위 그림은 age = 50 에서 knot 를 가지는 cubic polynomial 로 fitting 한 것
함수가 연속적이지 않아 이상하게 보인다는 문제점 발생



Smooth 한 그래프를 그리기 위해
다항식들의 1 차 및 2 차 도함수는 age = 50 에서 연속해야 한다는 제약 조건이 필요.

(다항식의 연속성, 1 차 도함수의 연속성, 2 차 도함수의 연속성)

3.2 The Spline Basis Representation

어떻게 d 차 적합 하면서 d-1 차까지의 미분 가능하다는 조건하에서 d 차원 다항식을 적합할 수 있을까?

기저모델을 사용하여 회귀 스플라인을 표현할 수 있다. K 개 매듭을 가지는 삼차 스플라인은 적절한 기저함수 b1, b2, ..., bK+3 에 대해 다음과 같은 모델

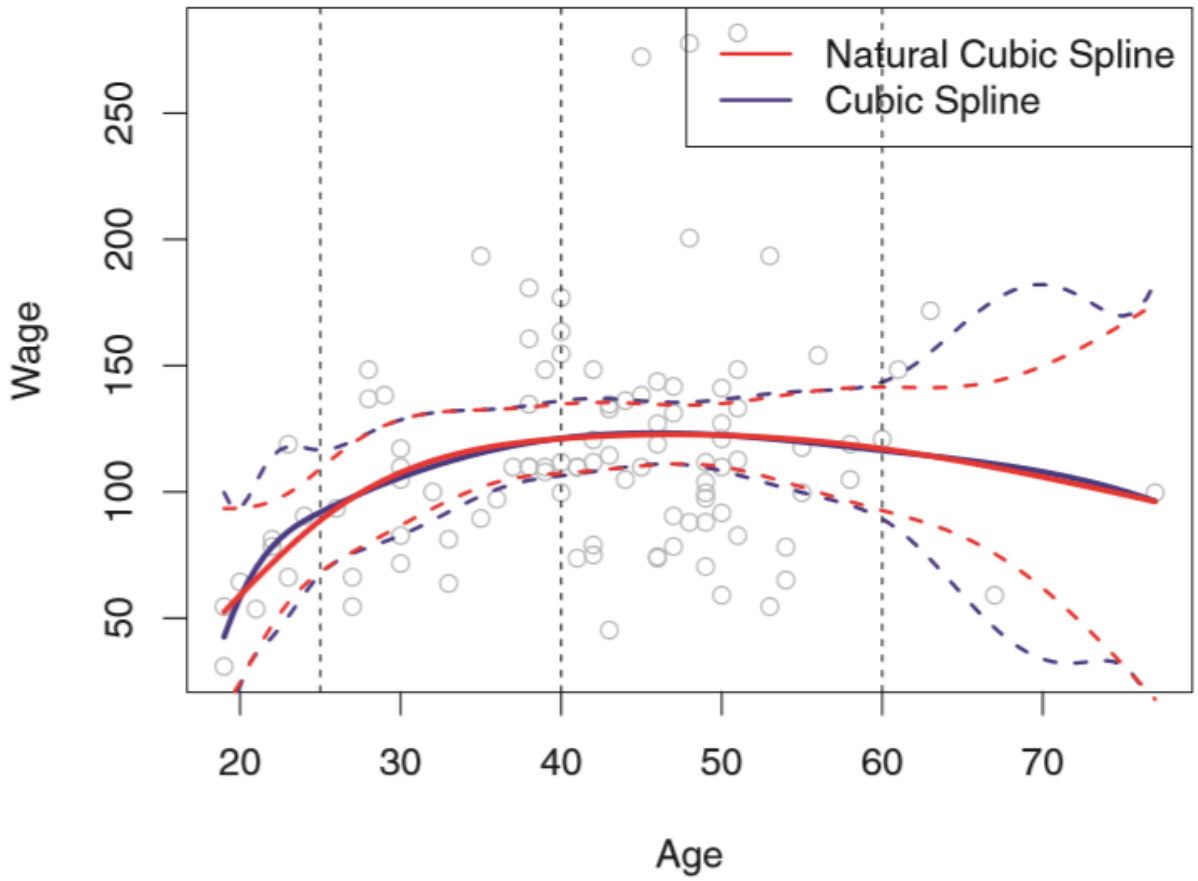
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

여러 방법이 있지만, cubic spline 을 나타내는 대표적인 방법은
Knot 당 하나의 truncated power basis function 을 추가하는 것이다.

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

미분 계산을 해보면 d 차 적합 하면서 d-1 차까지의 미분 가능하다는 사실을 알 수 있음.

하지만 spline 은 설명변수들의 외측 범위에서 - 즉, X 가 매우 작거나 매우 큰 값을 취할 때 높은 분산을 가질 수 있다.



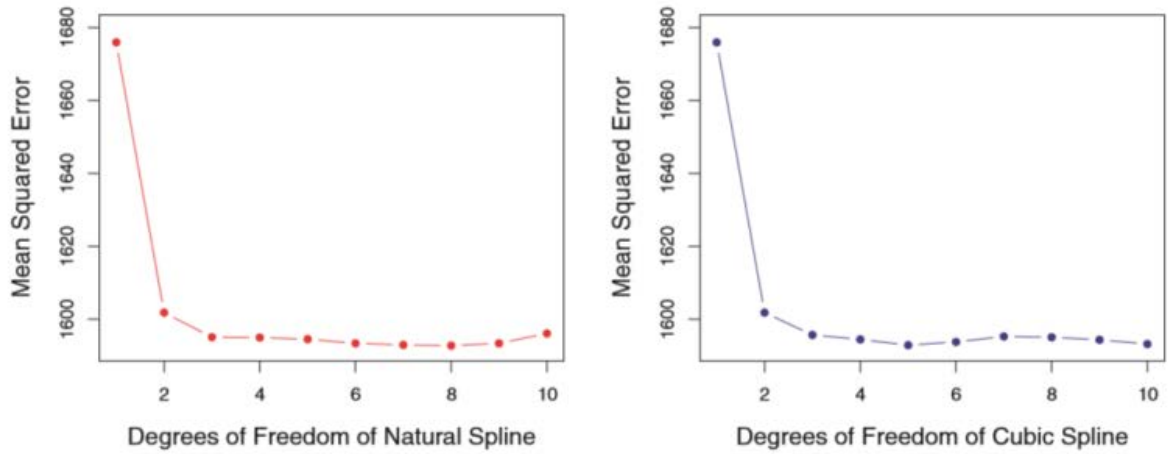
natural spline 은 함수가 경계에서(X 가 가장 작은 knot 보다 작거나 가장 큰 knot 보다 큰 영역에서) 선형이어야 한다는 추가적인 경계 제한조건이 있는 회귀 스플라인이다.

이 추가적인 제한조건은 natural spline 은 일반적으로 경계부근에서 더 안정적인 추정치를 제공한다는 것을 의미한다. 위 그림에서 natural spline 은 붉은색 선으로 표시되며 대응하는 신뢰구간은 더 좁은 것을 알 수 있다.

3.3 Choosing the Number and Locations of the Knots

spline 을 적합할 때 매듭들은 어디에 위치시켜야 하는가?
-> 보통 데이터에 균일하게 knots 를 위치시킴

Knots 를 몇 개나 사용해야 하는가?
->Cross-validation RSS 를 계산한 후, 가장 작은 RSS 를 주는 K 값을 선택



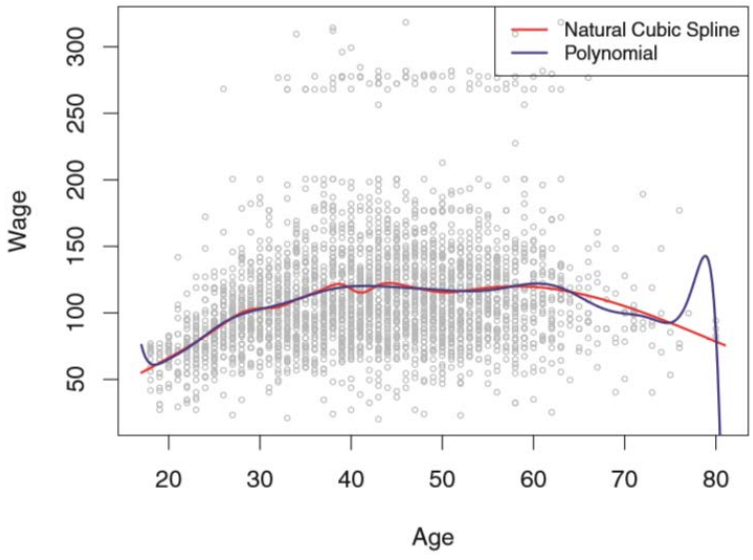
위 그림은 Wage 자료에 적합된 다양한 자유도의 spline 들에 대한 10-fold cross-validation MSE 를 계산한 것이다.

두 곡선은 자유도가 증가함에 따라 곧 평평해진다는 것을 알 수 있다. 자유도는 natural spline 의 경우 3, cubic spline 의경우도 3, 4 이면 충분해 보인다는 것을 알 수 있다.

3.4 Comparison to Polynomial Regression

회귀 스플라인들은 종종 다항식 회귀에 비해 월등히 좋은 결과를 준다. 왜냐하면 다항식은 유연한 적합을 위해 높은 차수를 사용해야 하지만, 스플라인은 차수는 고정시키고 knots 수를 증가시켜 유연성을 높일 수 있기 때문

Spline 은 함수 f 가 빠르게 변하는 영역에는 더 많은 매듭을 위치시켜 유연성을 높이고 그렇지 않은 영역에는 매듭 수를 줄일 수 있다.



4. Smoothing splines

곡선을 적합하는 데 있어서 데이터에 잘 맞는 함수를 찾을 때, RSS 을 가장 작게 하는 함수를 원한다. 하지만 아무런 제한조건을 두지 않으면 overfitting 하는 문제 발생.

우리가 원하는 것은 RSS 가 작지만 적당히 smoothing 한 함수
방법은 다음의 식을 최소로 하는 함수 g 를 찾는 것

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

여기서 λ는 조율 파라미터
Ridge regression 과 Lasso 와 관련하여 보았던 "손실(Loss) + 패널티(Penalty)" 와 비슷한 형태

앞에 식은 g 가 데이터에 잘 적합되게 하는 손실
λ가 곱해져 있는 오른쪽 식은 g 의 변동성에 제한을 주는 패널티.

1 차 도함수 g'(t)는 t 에서 함수의 기울기이고
,2 차 도함수 g''(t)는 기울기가 변하는 정도(양)에 해당한다.
그러므로 함수의 2 차 도함수는 거침(roughness)의 정도이다.

g(t)가 t 근방에서 아주 꾸불꾸불(wiggly)하면 2 차 도함수의 절대값은 크고, 그렇지 않으면 0 에 가까워진다. 적분기호는 t 의 전 범위에 걸친 합으로 생각할 수 있다. 즉, g'(t)의 총 변화에 대한 측도라고 생각할 수 있다.

g 가 아주 smooth 하면 g'(t)는 상수에 가까울 것이고 적분항은 작은 값을 가질 것이다.
반대로, g 가 변동이 심하면 g'(t)는 변화가 아주 클 것이고 적분항은 큰 값을 가질 것이다.
그러므로, 위 식에서 적분항은 g 가 smooth 하게한다. 값이 클수록 g 는 더 smooth 해질 것이다.

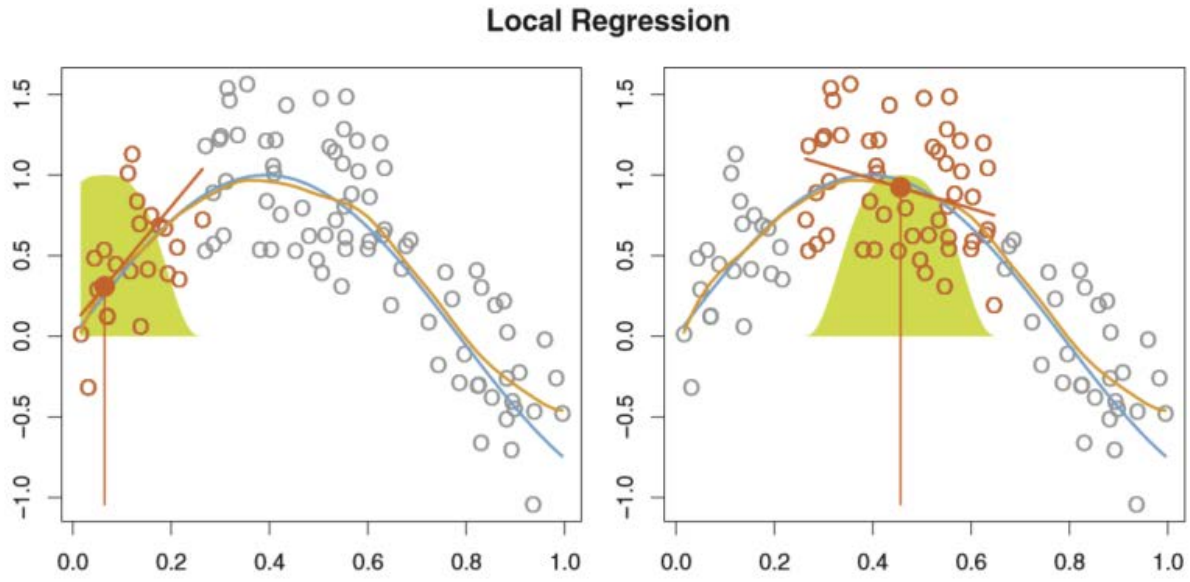
그런데 위 식을 최소로 하는 함수 g(x)는 어떤 특별한 성질을 가진다. 즉, 이 함수는 x1, ..., xn 에서 knot 를 갖는 삼차 다항식이고 이 함수의 1 차 도함수와 2 차 도함수는 각 knot 에서 연속이다. 또한 이 함수는 극단적인 knot 이외의 영역에서는 선형이다.

사실 위 식을 최소로 하는 함수 g(x)는 x1, ..., xn 에 knot 가 있는 natural cubic spline 이 수축된 모양이다. (λ가 수축 정도를 결정)

그렇다면 λ에 어떠한 값을 넣어야 할까? 훈러닝 혹은 교재 참고 ..!
https://www.youtube.com/watch?v=GFi8jeOkXaU&list=PLTGzWF3DajHQZ7zXesjid0zxmGdaNS4-K&index=29

5. Local regression

목표점 x0 에서 그 주변의 관측치들 만을 사용하여 비선형함수를 fitting 하는 방법



위 그림에서 한 목표점은 4 부근에 있고, 또 다른 목표점은 경계부근의 0.05 에 있다.
파란선은 데이터를 생성한 함수 f(x)를, 밝은 오렌지색 선은 local regression 을 통한 추정치
알고리즘은 아래와 같다.

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

1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.

2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.

3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

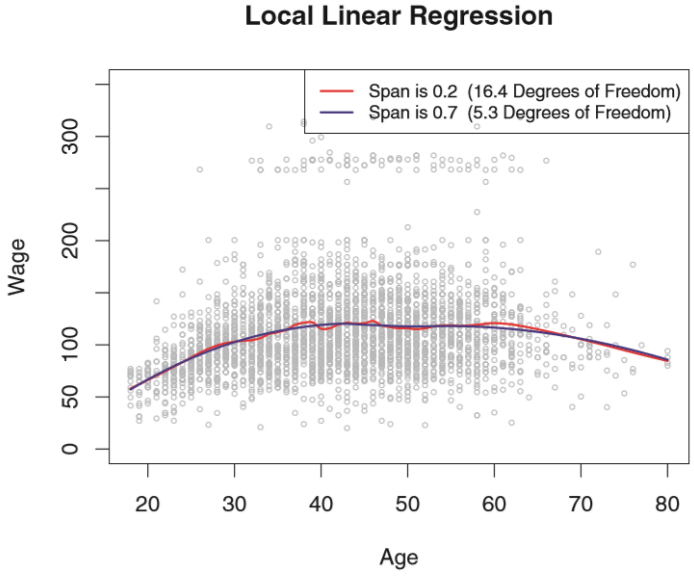
$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{7.14}$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

국소회귀를 수행하기 위해 선택해야 할 것이 세가지 있다.

1. Step 2 에서가중치 함수 K 를 어떻게 정의할 지 선택
2. Step3 에서 선형, 상수 또는 이차 회귀가 적합할 지 선택
3. 가장 중요한 것은 Step1 에서 s(span)의 값 설정

s 값이 작을수록 local 하게 fitting 되고, s 값이 global 하게 fitting 될 것이다.
참고로 s 값은 cross-validation 을 이용하거나 직접 설정.



6. Generalized addtive model(GAM)

여러 개의 설명변수 X1, ..., Xp 를 기반으로 Y 를 예측하는 문제
-> 다중선형회귀의 확장

GAMs 는 가산성은 유지하면서 각 변수의 비선형함수들을 허용하여 표준선형모델을 확장한 것
질적 및 양적 반응변수 모두에 적용될 수 있다!

6.1 GAMs for Regression Problems

각 설명변수와 반응변수 사이의 비선형적 관계를 고려하기 위해 아래 다중선형회귀모델

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

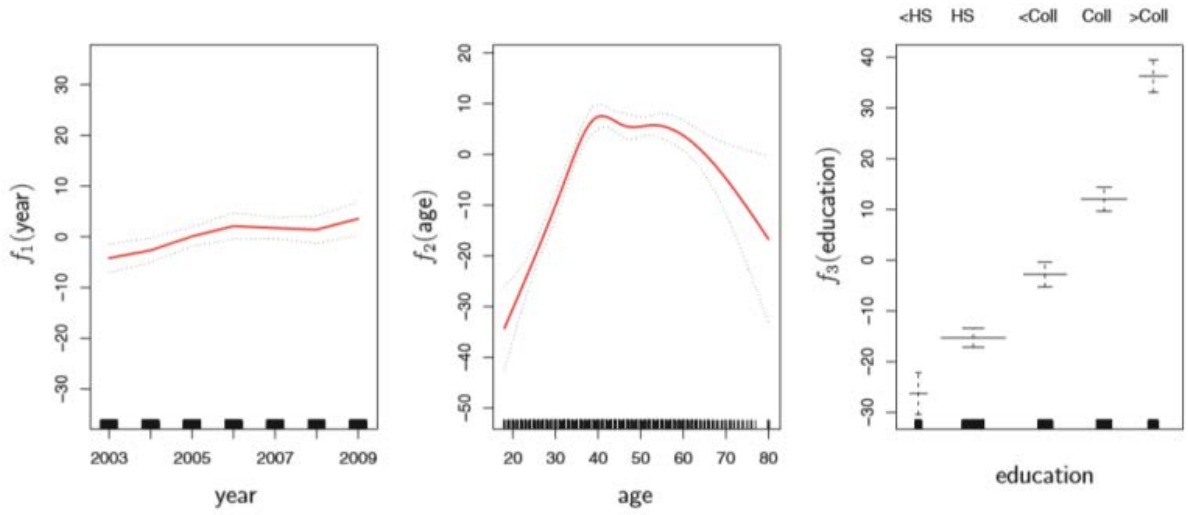
각 β를 smooth 한 비선형함수 f 로 대체한 모델

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

EX)

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

연도(year)와 나이(age)는 양적 변수
교육(education)은 5 개의 수준(<HS, HS, <Coll, Coll, >Coll)을 가지는 질적변수
첫 두 함수는 natural spline 사용하여 적합.
세 번째 함수는 각 수준에 대해 별도 상수를 사용하여 더미변수 이용해 적합.



6.2 GAMs for Classification Problems

GAMs 은 Y 가 질적 변수 경우에도 사용될 수 있다.

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

비선형적인 관계를 다루기 위해 아래와 같이 확장한 모델 사용

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

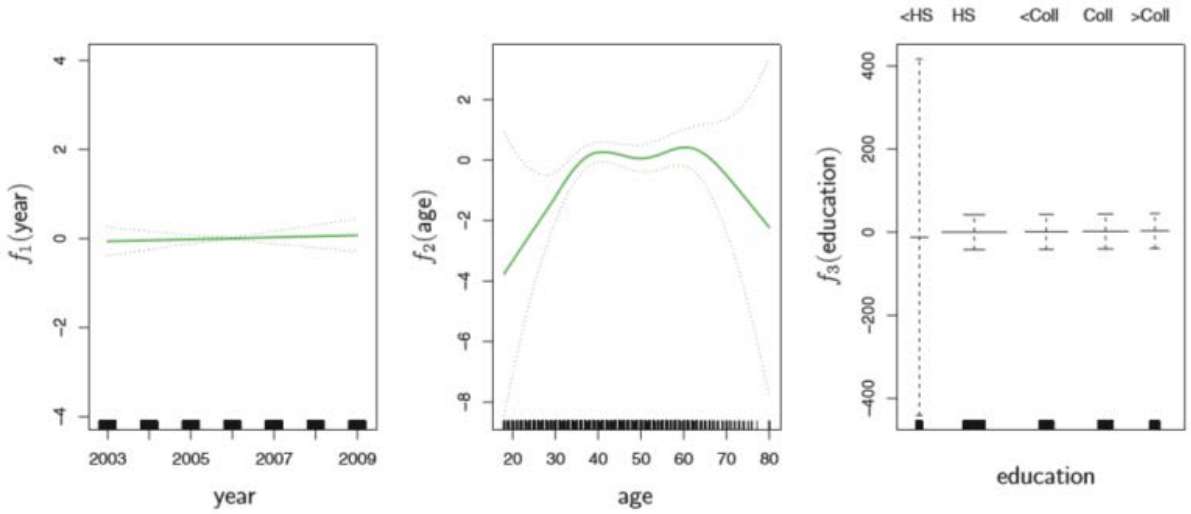
EX)

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times \texttt{year} + f_2(\texttt{age}) + f_3(\texttt{education})$$

여기서, p(X) = P(wage > 250 | year, age, education)

f2 는 smoothing spline 을 사용하여 적합

f3 은 각 수준에 대해 별도 상수를 사용하여 더미변수 이용해 적합



고등학교 교육을 받지 못한 사람은 제외하고 GAM 을 다시 적합