07. 선형성을 넘어서

선형모델 - 설명과 실현 단순, 및 해석과 추론 측면 강점, but **예측능력면에서 상당히 제한적**

–> 선형이란 가정은 항상 근사적이고, 때로는 잘 맞지 않기 때문

해석력은 가능한 한 높게 유지하면서 선형 가정 완화하는 방법에 대해 다룸.

7.1 다항식회귀

- 비선형적 설정으로 선형회귀를 확장하는 표준적인 방법

  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ 에서 $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$로
  - 설명변수가 $x_i, x_i^2, x_i^3, \ldots, x_i^d$인 표준 선형모델, 따라서 최소제곱 적합 가능
  - 일반적으로 4보다 큰 d를 사용하는 경우는 드뭄, 지나치게 유연해질 수 있기 때문
  - 계수 추정치 각각, 추정치 간 공분산행렬로 $\hat{f}(x_0)$의 추정분산 계산 가능
  - 이를 통해 점별 표준오차를 이용해서 confidence band를 만들 수 있다.
- 로지스틱 회귀에도 적용 가능
  - $p = \frac{exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d)}{1 + exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d)}$

7.2 계단함수

- 고차항을 설명변수로 사용하는 것은 X의 비선형 함수에 전역구조를 도입하는 것.
  - 전역구조란? 아마 X가 범위 등에 상관없이 함숫값에 그대로 영향을 미치는 것을 의미하는듯
- 전역구조 도입을 피하기 위해 X의 범위를 여러개의 bin으로 분할, 각 bin에 다른 상수를 적합
  - 연속적인 변수를 순서범주형 변수로 변환하는 것.
  - X의 범위에 $c_1, c_2, \ldots, c_K$의 절단점을 도입하여 K+1개의 새로운 변수를 만드는 것.

$$
\begin{array}{lcl}
C_0(X) & = & I(X < c_1), \\
C_1(X) & = & I(c_1 \leq X < c_2), \\
C_2(X) & = & I(c_2 \leq X < c_3), \\
 & \vdots & \\
C_{K-1}(X) & = & I(c_{K-1} \leq X < c_K), \\
C_K(X) & = & I(c_K \leq X),
\end{array} \tag{7.4}
$$

  - 여기서 I는 조건이 참이면 1, 아니면 0을 반환하는 Indicator Function(=dummy variable)
  - X는 K+1개 구간 어느 하나에 속해야 하므로, $C_0(X) + C_1(X) + \ldots + C_K(X) = 1$
  - $C_0(X), C_1(X), \ldots, C_K(X)$를 설명변수로 사용하며 최소제곱 적합

    $y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$
  - 역시 로지스틱 회귀적합에도 이용 가능

$p = \frac{exp(\beta_0+\beta_1 C_1(x_i)+\beta_2 C_2(x_i)+...+\beta_K C_K(x_i))}{1+exp(\beta_0+\beta_1 C_1(x_i)+\beta_2 C_2(x_i)+...+\beta_K C_K(x_i))}$

- 설명변수에 breakpoint가 없으면 조각별 상수함수들은 상황변화를 놓칠 수 있다.
- 생물통계학, 역학에서 널리 사용

7.3 기저함수

- 다항식회귀, 계단함수의 일반화 버전

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

- 기저함수 $b_1, b_2, \ldots, b_k$는 fixed & known(미리 선택됨)
- 다항식회귀는 $b_j(x_i) = x_i^j$, 계단함수는 $b_j(x_i) = I(c_j <= x_i <= c_{j+1})$
- 회귀계수를 추정하는 데 최소제곱 사용 가능, 따라서 계수 추정치에 대한 표준오차, 모델의 전체 유의성에 대한 overall F test 등 선형모델에 대한 추론도구를 모두 사용 가능
- Wavelets, Fourier series등이 기저함수로 사용가능
- 다음절에서는 기저함수로 매우 자주 선택되는 회귀 스플라인에 대해 알아본다!

7.4 회귀 스플라인

7.4.1 조각별 다항식

- X의 범위를 구분하여 각 범위에 저차원 다항식을 적합, 역시 최소제곱 적합 사용
- 매듭이란 : 계수들이 변하는 점, 즉 X의 범위들의 경계
- e.g. 삼차회귀모델 적합에서, 매듭이 없을 경우

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

- 매듭 1개(점 c에서)일때,

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

- 매듭이 많아질수록, 조각별 다항식은 유연해진다.
- 문제점 : 함수의 불연속 유발 가능

7.4.2 제약조건과 스플라인

- 차수가 d인 스플라인은 (d-1)차까지의 도함수가 연속적인 제약조건을 만족한다.
- 삼차스플라인의 예시
  - 매듭이 K개일때, 제약조건 없는 경우 4+K*4의 자유도를 가짐
  - 3가지 제한조건(원함수, 1차도함수, 2차도함수의 연속)을 만족함으로 4+K(4-3), 4+K의 자유도를 가짐.
  - 따라서 매듭이 K개인 d차 스플라인은 d+1+K의 자유도를 가진다.

7.4.3 스플라인 기저표현

- 조각별 다항식 + 연속조건은 다소 복잡
- 적절한 기저함수로 회귀 스플라인 표현이 가능
- K개의 매듭을 가지는 삼차 스플라인에 대해, 기저함수 $b_1, b_2, \ldots, b_{K+3}$으로 다음과 같은 모델링이 가능

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

- 위 식을 사용하여 스플라인을 나타내는 가장 직접적인 방법은, 삼차다항식에 대한 기저(x, x^2, x^3)을 가지고 시작하여 매듭당 하나의 절단 멱 기저함수(truncated power basis function)을 추가하는 것

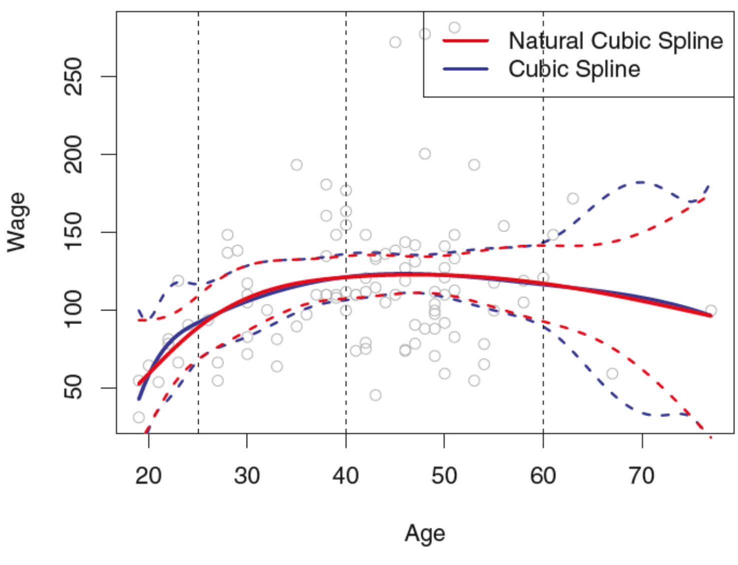
$$h(x, \xi) = (x - \xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise,} \end{cases} \tag{7.10}$$

- ξ는 매듭, $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$에 $\beta_4 h(x, \xi)$ 형태의 항을 추가하면, 3차 도함수만 ξ에서 불연속(즉, 2차 도함수까지 연속으로 제약조건 만족)
- K개의 매듭으로 일반화 하면, 다음과 같은 식에 대해 최소제곱 적합을 진행하는 것

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d + \beta_{d+1} h(X, \xi_1) + \beta_{d+2} h(X, \xi_2) + \ldots + \beta_{d+K} h(X, \xi_K) + \epsilon_i$$

$$\text{단}, h(X, \xi) = (x - \xi)_+^d$$

- 추정해야할 계수가 d+K+1개이므로 d+K+1의 자유도 사용
- 스플라인은 경계(설명변수의 양 극)에서 높은 분산을 취할 수 있는 단점
  - 이를 보완하기 위한 방법은 함수가 경계에서 선형이라는 추가적인 제한조건을 걸어준다.



  - 경계에서 조금 더 안정적인 추정치(대응되는 신뢰대역이 더 좁다)

7.4.4 매듭의 수와 위치 선택

- 함수가 가장 빠르게 변할 것 같은 곳에 많은 곳을 위치시키는 방법이 있을 수 있지만

- 보통은 균일하게 매듭을 위치
    - 자유도를 지정한 후, 소프트웨어가 데이터의 균등 분위수(uniform quantiles)에 각각 매듭을 위치
    - 예를 들어, 자연 삼차 스플라인에 대해 자유도 4를 명시하면, 3개의 매듭이 만들어짐.
        - 경계매듭을 포함하여 5개 매듭의 삼차 스플라인은 9의 자유도를 갖지만, 자연 스플라인은 각 경계에서 선형성을 강제하는 2개의 추가적 제한조건이 있음. 따라서 9-2*2=5에서,
        - 위 추가적 제약조건에 상수가 포함되는데, 이것이 절편(이미 자유도에 포함이 된)에 흡수되므로 자유도는 1을 더 빼줘서 4로 간주.
    - 이를 좀 더 일반화하여 자연 d차 스플라인에 대해 f의 자유도를 명시하면,
    - $f-d+2$개의 매듭을 생성

$$(d+(k+2)+1)-2*(d-1)-1=f \Leftrightarrow k=f+d-4$$

- 적절한 자유도는 CV로 결정할 수 있다.

7.4.5 다항식회귀와 비교

- 다항식회귀는 유연성을 위해 높은 차수를 사용
- 스플라인은 차수를 고정하고 매듭수로 유연성을 조절할 수 있다.
- 일반적으로 스플라인이 더 안정적인 추정치를 제공

7.5 평활 스플라인

7.5.1 평활 스플라인의 개요

- 7.4의 회귀 스플라인은, 1)매듭을 지정 2)기저함수 도출 3)최소제곱 적합을 이용
- 평활 스플라인은 다음 식을 최소로 하는 함수 g를 찾는 것.

$$\sum_{i=1}^{n}(y_i-g(x_i))^2+\lambda\int g''(t)^2dt \tag{7.11}$$

- Loss(손실, 여기서는 RSS) + Penalty의 형태
- 페널티 항이 이차도함수(roughness의 척도)에 대한 적분값이므로, 유연성에 대한 페널티
- λ =0 일때는 모든 관측치들을 보간하는 지나치게 유연한 함수
- $\lambda\to\infty$ 일때, g는 선형 최고제곱선
- $g(x)$는 $x_1,x_2,\ldots,x_n$에 매듭이 있는 조각별 삼차 다항식이고, 함수의 1,2차 도함수 모두 매듭에서 연속
    - 즉, $g(x)$는 모든 훈련 관측치($x_1,x_2,\ldots,x_n$)에 매듭이 있는 자연 삼차 스플라인이다.
    - 하지만 앞에서 다뤘듯 기저함수를 이용한 자연 삼차 스플라인은 아니다.

7.5.2 평활 파라미터 λ의 선택

- λ를 제어하여 유효자유도(*effective degree of freedom*)를 제어

- λ가 0에서 ∞로 증가함에 따라 유효자유도 df_λ는 n에서 2로 줄어듬
- 왜 자유도 대신 실효자유도를 다루는가?
  - 평활 스플라인은 n개의 파라미터를 가지므로 명목상 n의 자유도를 갖지만, 이 파라미터들은 실제로는 심하게 수축, 따라서 어떠한 측도가 될 수 없다.
  - df_λ는 유연성의 측도로의 대안이 될 수 있다. 따라서 df_λ가 높아질수록 유연하다.(낮은 편향과 높은 분산)
  - $\hat{g}_\lambda$를 특정 λ에 대한 위 (7.11)식의 해($x_1, x_2, ..., x_n$에 대한 fitted value, 따라서 n-벡터)라고 하고,
  - 이를 만약 어떤 n X n행렬인 S_λ와 반응벡터 y로 다음과 같이 표현한다면,

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

- 유효자유도는 $df_\lambda = \sum_{i=1}^{n} S_{\lambda ii}$, 즉 S_λ의 대각원소의 합으로 구할 수 있다.
- 평활 스플라인은 매듭의 수 결정이 필요없다, 왜냐하면 이미 매듭이 $x_1, x_2, ..., x_n$으로 정해져있기 때문
- 적절한 λ를 선택하는 것이 중요
  - 이는 평활 스플라인에서 LOOCV를 효과적으로 계산하는 다음 식에 의해 선택될 수 있다.

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}}\right]^2.$$

7.6 국소회귀

- 국소회귀는 목표점 x_0에서 그 주변의 훈련관측치만을 사용하여 적합을 계산
- 다음 알고리즘을 통해 국소회귀 진행
  - 1. 훈련 포인트들의 x_i가 x_0에 가장 가까운 $s = k/n$만큼을 모은다.
    2. 이 이웃에 각 점에 가중치 $K_{i0} = K(x_i, x_0)$를 할당. 가까울 수록 가중치는 높고, k개의 이웃 외의 모든 점은 가중치가 0이다.
    3. 앞의 가중치를 사용하여 아래 식을 최소로 하는 $\hat{\beta}_0$ 와 $\hat{\beta}_1$을 찾으므로써 가중 최소제곱회귀 적합

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

4. x_0에서 적합된 값은 $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$로 주어진다.

- 국소회귀는 최근접이웃 방법처럼 예측할때마다 모든 훈련 데이터를 필요로 하는 기억 기반 절차
- 가중치 함수 K를 정의하고, 위의 세 번째 단계에서 선형, 상수, 이차회귀를 적합할지를 선택해야 한다.(위의 예는 선형)
- 가장 중요한건 s, 생성(span)을 정하는 것
  - s는 유연성을 제어, s값이 작을수록 유연, 클수록 평활&전역적 적합
  - CV를 통해 선택하거나 직접 지정할 수 있다.
- 다수의 설명변수가 있는 설정에서 다음과 같이 일반화 할 수 있다.
  - 특정 변수에 대해는 전역적이지만 다른 변수에 대해서는 국소적인(예를 들면 시간) **가변 계수 모델**

        - 시간에 대해 국소적인 모델은 최근에 수집된 데이터에 모델을 적응시키는 데 유용하게 쓰임
    - 하나의 변수가 아니라 p개 변수에 대해 p차원 이웃을 통해 적합하는 회귀
        - 3또는 4를 초과하는 p는 사용하지 않음. 왜냐하면 가까운 훈련 관측치가 매우 적어 성능 나쁨

7.7 일반화가법모델(GAM)

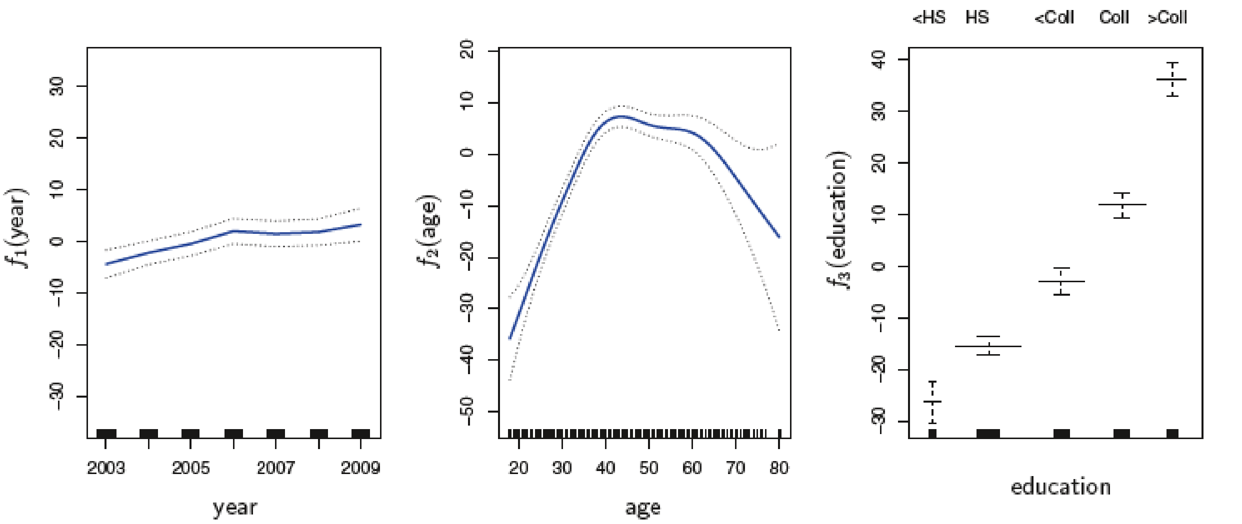
- 각 변수에 대해 가산성은 유지하면서 각 변수마다 비선형함수를 허용하여 표준 선형모델 확장

7.7.1 회귀문제에 대한 GAMs

- 다중 선형모델($y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$) 에 대한 확장

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \end{aligned} \qquad (7.15)$$

- 다중 선형모델의 각 선형요소인 $\beta_i x_{ij}$를 비선형함수 $f_j(x_{ij})$로 대체
- GAM의 장점은, 우리가 앞의 7.1~7.6에서 배운 방법들을 각 비선형함수를 이루는 building block으로 사용할 수 있다는 것!
- 다음의 예는 $wage = \beta_0 + f_1(year) + f_2(age) + f_3(education) + \epsilon$의 모델에서, 범주형 변수 education에는 계단함수를, 수치형 변수 year, age에는 평활 스플라인을 적용한 그림



- 평활 스플라인은 최소제곱이 사용되지 않기 때문에 후방적합의 기법 사용
    - 각 설명변수에 대한 적합을 다른 변수들을 고정한채 교대로 반복하여 업데이트
    - 각 변수에 대한 적합 방법에 부분잔차를 이용하는데,
        - 부분잔차란 예를들어 X_3에 대한 부분 잔차는, $r_i = y_i - f_1(x_{i1}) - f_2(x_{i2})$ 의 형태를 가지고, 이 잔차를 반응변수로 취급하여 f_3을 적합
- 자연 스플라인 등 기저함수를 이용하는 방법은 단순하게 기저변수+가변수(범주형변수에 대한)에 대한 최소제곱 적합

GAM의 장점과 단점

> 장점

- 표준적 선형회귀로 놓칠 수 있는 비선형 관계를 자동적으로 모델링, 변수 변환이 필요없다.
- 표준 선형회귀에 비해 예측 power를 높일 수 있다.
- 가산적이기 때문에 추론에도 역시 유용하다.
- 변수 X_j에 대한 f_j의 평활도를 자유도로 요약 가능하다.

> 단점

- 모델이 가산적이어야 한다는 제한. 따라서 상호작용을 놓칠수 있지만,
- $X_j \times X_k$ 형태의 추가적인 설명변수나
- $f_{jk}(X_j, X_k)$ 형태의 저차원 상호작용 함수를 모델에 추가할 수도 있다.
    - 이 함수들은 2차원 평활기(e.g. 저차원 스플라인)를 사용하여 적합 가능

GAM은 선형모델과 완전 비모수적 모델 사이에서 절충에 유용

7.7.2 분류모델에 대한 GAMs

- 로지스틱 회귀를 로지스틱 회귀 GAM으로 다음과 같이 확장

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p). \qquad (7.18)$$