Yonsei University, Department of Applied Statistics

Chapter 7 Moving Beyond Linearity

12기 YBIGTA 신보현

February 19, 2019

여태까지, 선형 모델에 중점을 두고 논의를 진행해왔다. 선형 모델은 선형성이라는 강력한 가정이 충족된다면 좋은 성능을 보이지만, 그렇지 않다면 꽤나 나쁜 결과를 보일 때가 많다. 6단원에서는 ridge, lasso, pcr, pls 등 선형모델을 향상시킨 여러 기법에 대해서 배웠지만 이러한 모델도 여전히 선형성의 가정에 기반한다. 따라서 7단원에서는 선형 가정을 완화함과 동시에 해석력 또한 최대한 유지하는 기법들에 대해서 배울 것이다. 앞으로 배울 모델들에 대해서 간략하게 알아보자.

- Polynomial regression
  다항 회귀는 원래 변수의 제곱, 세 제곱 등의 예측변수를 추가하는 방법이다. 이렇게 함으로써 데이터의 비선형성을 모델에 반영할 수 있다.
- Step functions
  범주형 변수를 생성하기 위해 변수를 K개의 서로 분리된 지역으로 자른다. 이것은 piecewise constant function을 적합하는 효과와 동일하다.
- Regression splines
  앞선 모델보다 더 flexible하고 사실 이들의 확장이기도 하다. X 변수를 K개의 지역으로 나눈 후, 각 지역에서 다항회귀를 적합한다. 하지만 이러한 다항식은 바운더리에서 부드럽게 (smoothly) 연결되도록 제한된다. 만약 변수가 충분한 지역으로 나뉜다면, 이것은 매우 flexible한 적합을 만들 것이다.
- Smoothing splines
  regression splines과 비슷하지만 smoothness penalty하에 SSE를 최소화한다는 점에서 다른 점을 보인다.
- Local regression
  splines과 비슷하지만, 지역들이 겹치는 것이 허용되고 사실 매우 부드러운 방법으로 겹쳐진다.
- Generalized additive models
  여러 예측 변수를 다루기 위해 위의 방법들을 확장한다.

7.1에서 7.6까지는 하나의 예측변수와 반응변수간에 flexible한 관계를 적합할 것이며 7.7에서는 여러 예측변수의 함수로 반응변수를 적합할 것이다.

7.1 Polynomial Regression

다항 회귀는 반응변수와 예측변수간의 관계를 아래와 같이 설정한다.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \cdots + \beta_d x_i^d + \epsilon_i$$

기존의 다중 회귀와 비교해볼 때, 다중 회귀는 예측변수가 모두 1차였지만 다항 회귀는 그러한 제한을 두지 않는다는 점이다. 하지만 계수들은 기존의 다중 회귀와 마찬가지로 최소자승추정법으로 추정이 된다. $x_1, x_2^2, \cdots, x_d^d$들을 $k_1, k_2, \cdots, k_d$로 보면 되기 때문이다. 보통, 차수는 3차, 4차를 넘어가지 않고 차수가 높아진다면 특히 X의 가장자리에서 이상한 모양이 형성될 수 있다.

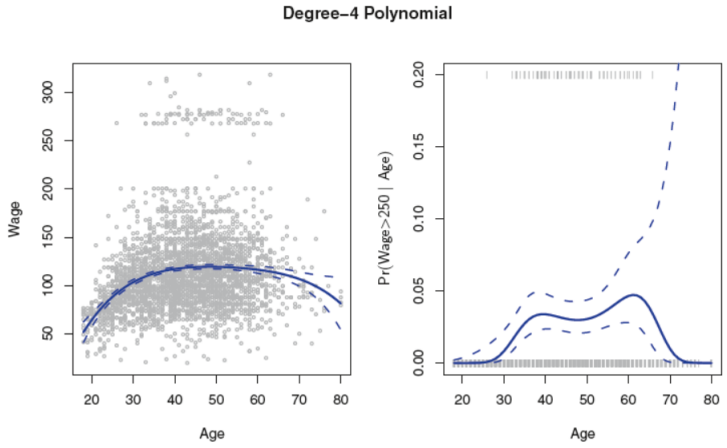

FIGURE 7.1. *The* `Wage` *data.* Left: *The solid blue curve is a degree-4 polynomial of* `wage` *(in thousands of dollars) as a function of* `age`*, fit by least squares. The dotted curves indicate an estimated 95 % confidence interval.* Right: *We model the binary event* `wage>250` *using logistic regression, again with a degree-4 polynomial. The fitted posterior probability of* `wage` *exceeding* $250,000 *is shown in blue, along with an estimated 95 % confidence interval.*

figure 7.1은 wage vs age 데이터에서 4차 다항 회귀를 적합한 결과이다. 즉, 아래와 같이 식을 적합한 것이다.

$$\hat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_4 x_4^4$$

어떤 age 값에서 fitted value의 분산은 어떻게 계산할까?($Var(\widehat{f}(x_0))$) 다중 회귀를 통해 추정된 계수의 분산-공분산 행렬을 알 수 있다. σ가 알려져있다고 가정해보자. $Cov(\widehat{\beta}) = \sigma^2(X'X)^{-1}$이고 $\widehat{f}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 + \cdots + \widehat{\beta}_4 x_0^4 = \begin{bmatrix} 1 & x_0 & \cdots & x_0^4 \end{bmatrix}^T \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \cdots \\ \widehat{\beta}_4 \end{bmatrix} = \ell' \widehat{\beta}$

이므로 $Var(\widehat{f}(x_0)) = \ell' Cov(\widehat{\beta}) \ell$로 계산하면 된다. 이런 방식으로 각 점에서 fitted value의 표준 오차를 계산하고 두 배의 표준오차를 더하고 뺀 그래프를 그리는데, 정규분포를 따르는 오차항에 대해서, 이것이 95% 신뢰구간에 해당하기 때문이다.

figure 7.1을 보면, 고소득 집단(250,000 달러 이상 버는 집단)과 저소득 집단이 있는 것으로 파악된다. 따라서 두 클래스에 분류를 하는 로지스틱 다항 회귀를 시행하는 것도 나쁘지

않은 선택일 것이다.

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

figure 7.1의 오른쪽이 로지스틱 다항 회귀를 적합한 결과이다. 굵은 선으로 추정된 고소득 집단일 확률이, 점선으로 신뢰구간이 표시되었다. 오른쪽으로 갈수록 신뢰구간이 매우 넓어짐을 확인할 수 있는데, 이는 분산이 커진 결과로 추정된다. 전체 데이터 3000개 중, 고소득 집단인 사람은 79명 뿐이었으므로 이렇게 분산이 높게 나온 것으로 보인다.

7.2 Step Functions

다항 함수를 예측변수로 사용하는 것은 모델에 비선형성을 전역적으로(global) 적용한다. 만약 전체적으로 비선형성이 있는 것이 아닌, 부분적으로 비선형성이 존재하고 그 이외에는 선형성이 존재한다면 이러한 방법은 좋지 않을 것이다. 이러한 상황에 step functions을 사용함으로써 국소적(locally)으로 비선형성을 가정할 수 있다. X의 범위를 bins으로 나누고 각 bin마다 다른 상수(different constant)을 적합한다. 이는 연속형 변수를 ordered categorical 변수로 바꾸는 것과 결과적으로 동일하다.
좀 더 상세하게 말하면, X의 범위에서 cutpoints인 $c_1, c_2, \cdots, c_K$을 만들고 $K+1$개의 새로운 변수를 만든다.

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

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

$$C_2(X) = I(c_2 \leq X < c_3)$$

$$\vdots$$

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

$$C_K(X) = I(c_K \leq X)$$

여기서 어떤 X 값에 대해서도 $C_0(X) + C_1(X) + \cdots + C_K(X) = 1$인데, 그 이유는 X의 범위를 $K+1$개의 범위로 나누었기 때문에 어떤 X 값도 $K+1$개의 범위 중 하나에 속해야 하고, 이 중으로 속할 수는 없기 때문이다(안 겹치게 범위를 나누었기 때문에) 마치 더미 변수와 같은 느낌이라고 생각하면 된다. 그리고 이들을 이용해서 최소자승추정법을 실행한다.

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

여기서 $C_0(X)$은 빠졌는데, 그 이유는, 이를 포함시키면 design matrix의 열이 선형 종속이 되어 OLS가 유일하게 존재하지 않기 때문이다. 만약 $X < c_1$이라면 모든 변수들이 0이

되어 남아있는 β_0가 $X < c_1$일 때의 Y의 mean value라고 해석될 수 있다. 또한 위 식은 $c_j \leq X < c_{j+1}$에 대해서 반응 변수를 $\beta_0 + \beta_j$라고 예측을 하므로β_j는 average increase라고 해석될 수 있다.

step functions을 적합한 예는 figure 7.2에 나와있다.

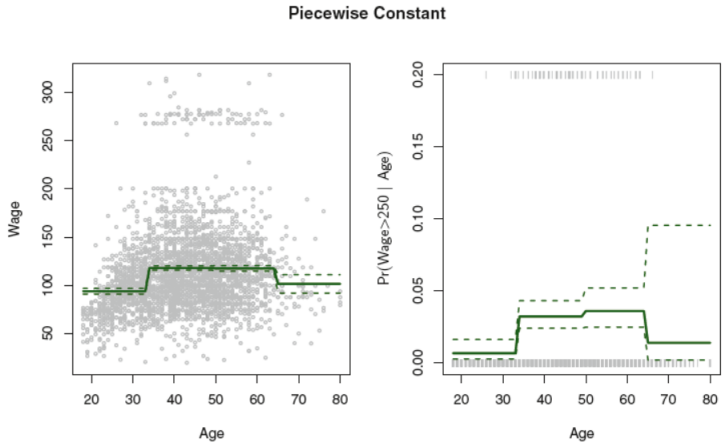

FIGURE 7.2. *The* `Wage` *data.* Left: *The solid curve displays the fitted value from a least squares regression of* `wage` *(in thousands of dollars) using step functions of* `age`*. The dotted curves indicate an estimated 95 % confidence interval.* Right: *We model the binary event* `wage>250` *using logistic regression, again using step functions of* `age`*. The fitted posterior probability of* `wage` *exceeding* $250,000 *is shown, along with an estimated 95 % confidence interval.*

오른쪽 그림은 로지스틱 다항 회귀를 적합한 결과이다.

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

안타깝게도, 예측변수에 자연스러운 절단점이 없다면 piecewise-constant 함수는 의미를 상실할 수 있다. 예를 들어, 첫번째 bin에서 적합된 결과는 당연히 상수일텐데, 실제 데이터를 보면 상승하는 추세를 확인할 수 있다. 그럼에도, step functions는 바이오통계 등 다른 분야에서 인기가 있는 방법이다.

7.3 Basis Functions

다항 회귀와 step functions은 사실, basis function 접근법의 특별한 예이다. 아이디어는, 변수 X에 대해서 적용할 수 있는 함수나 변형을 생각하는 것이다. 즉, 아래의 모델을 적합한다.

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

여기서, $b_1(\cdot), \cdots, b_K(\cdot)$은 fixed unknown임에 주목하자. 즉, 미리 이 함수를 선택하는 것이다. 다항 회귀에 대해서는, $b_j(x_i) = x_i^j$이고 step functions(piecewise constant functions)에 대해서는 $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$이다. 동일하게, 위 모델에 대해서도 최소자승법을 실행하여 계수를 추정한다. basis functions는 매우 다양한 형태가 있고 정하기 나름이다. 이제 가장 빈번하게 사용되는 basis function인 regression splines을 알아보자.

7.4 Regression Splines

7.4.1 Piecewise Polynomials

X의 전 범위에 높은 차수의 다항식을 적합하는 대신, piecewise polynomial regression은 X의 다른 범위에 별도의 낮은 차수의 다항식을 적합한다. 예를 들어, piecewise cubic polynomial은 $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$를 X 범위에서 각 파트에 다른 추정된 계수로 적합한다. 추정된 계수가 바뀌는 지점을 knots라고 부른다.
예를 들어, knots가 없는 3차 다항 회귀는 그저 이전과 동일한 3차 다항 회귀이고 점 c에서 하나의 knot를 가지는 3차 다항 회귀 모델은 아래와 같다.

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

다시 말해, 데이터에 두 개의 다른 다항 함술ㄹ 적합하는 것이다.
더 많은 knots를 사용하면 더 flexible한 piecewise polynomial이 형성된다. 일반적으로, K개의 다른 knots를 사용하면, $K+1$개의 서로 다른 다항 함수를 적합하게 된다. 다항 함수로는 삼차 다항식뿐만 아니라 앞서 보았던 step functions도 사용할 수 있다.

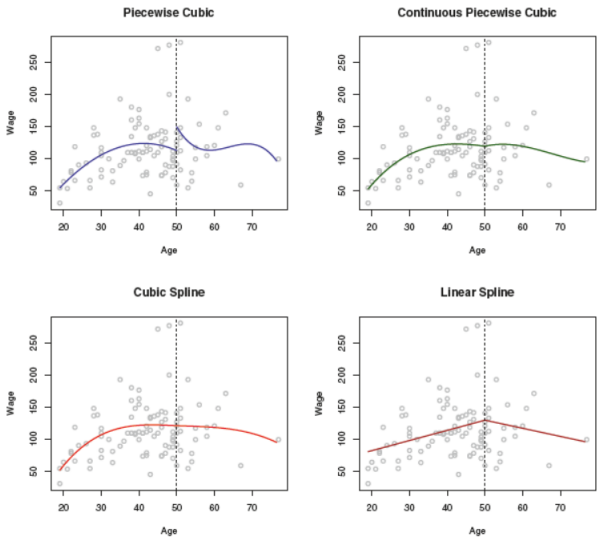

FIGURE 7.3. *Various piecewise polynomials are fit to a subset of the* Wage *data, with a knot at* age=50. Top Left: *The cubic polynomials are unconstrained.* Top Right: *The cubic polynomials are constrained to be continuous at* age=50. Bottom Left: *The cubic polynomials are constrained to be continuous, and to have continuous first and second derivatives.* Bottom Right: *A linear spline is shown, which is constrained to be continuous.*

figure 7.3 왼쪽 위에 Wage 데이터에 piecewise cubic polynomial을 적합한 결과를 보면, knots에서 불연속일뿐만 아니라 모양도 매우 이상하다.

7.4.2 Constraints and Splines

figure 7.3 왼쪽 위 그림에 나타난 문제를 해결하기 위해서 적합된 곡선을 연속적이게 적합을 해야 한다. 다르게 말해서, knots에서 점프를 하면 안 된다. 오른쪽 위 그림이 이를 적용한 것인데, V자 모양이 조금 부자연스럽다.

왼쪽 밑 그림은 1차 도함수, 2차 도함수가 knot에서 연속적이게 제한을 한 그림이다. 다르게 말해서 knot에서 연속할뿐만 아니라 매우 부드럽게 연결되야 한다는 것이다. 이러한 방법을 cubic spline이라고 부른다. piecewise cubic polynomials에 부여하는 각 제한은 결과적으로 나오는 모델의 복잡도를 줄임으로써 효과적으로 자유도 한 개를 frees up 한다(책의 내용을 그대로 번역한 것이라 조금 어색함) 따라서, 왼쪽 위 그림에서 8개의 계수가 추정되었으므로 자유도가 8인데 왼쪽 아래 그림에서는 세 개의 제한(연속성, 1차 & 2차 도함수에서의 연속성)으로 인해서 자유도가 5가 되었다. 일반적으로 K개의 knots을 가지는 cubic spline은 $4 + K$의 자유도를 가진다.

오른쪽 밑 그림은 knot에서 연속적인 linear spline이다.

d차 spline(regression spline)의 일반적인 정의는, piecewise degree-d polynomial이고 d-1차까지 각 knot에서 연속적이어야 한다. 이러한 정의에 따르면, linear spline은 knot으로 분리된 각 지역에서 line을 적합하고 knot에서 연속적이게 만든 것이다.

7.4.3 The Spline Basis Representation

방금 전 살펴본 d차 spline(regression spline)의 일반적인 정의는 꽤나 복잡하게 보인다. 어떻게 d차 piecewise polynomial을 적합하는데 knot에서 d-1차까지 연속적이게 만들까? 사실, cubic spline은 basis model을 이용해서 나타낼 수 있다. K개의 knots을 가진 cubic spline은 아래와 같이 표현할 수 있다.

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

cubic spline을 basis function을 이용해서 어떻게 나타낼까? 다양한 basis functions을 이용하여 cubic spline을 나타낼 수 있다. cubic spline을 basis function을 이용하여 나타낼 수 있는 가장 직접적인 방법은 3차 다항식(x, x^2, x^3)부터 시작을 해서 ***truncated power basis function***을 knot당 하나씩 추가하는 것이다. truncated power basis function은 아래와 같이 정의된다.

$$h(x,\xi) = (x-\xi)^3_+ = \begin{cases} (x-\xi)^3 & if\ x > \xi \\ 0 & o.w \end{cases} \quad where\ \xi\ is\ the\ knot$$

knot 하나를 추가 함으로써, $\beta_4 h(x,\xi)$항을 추가하는 것인데, 이는 3차 도함수에서 불연속이고 1차, 2차 도함수에서는 각 knot에서 연속성을 유지할 것이다. 이 말이 무슨 뜻인지 구체적으로 살펴보자. $\beta_4 h(x,\xi)$을 추가하여 3차 regression spline 모델을 구성하면 아래와 같다.

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \beta_{41}(x-\xi)^3 + \epsilon_i & if\ x_i > \xi \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & o.w \end{cases}$$

이를 최소자승 추정법으로 계수를 추정을 하고 1차, 2차, 3차 도함수를 구하면 아래와 같다.

$$\widehat{f}(x)^{(1)} = \begin{cases} \widehat{\beta}_{11} + 2\widehat{\beta}_{21} x_i + 3\widehat{\beta}_{31} x_i^2 + 3\widehat{\beta}_{41}(x-\xi)^2 & if\ x_i > \xi \\ \widehat{\beta}_{11} + 2\widehat{\beta}_{21} x_i + 3\widehat{\beta}_{31} x_i^2 & o.w \end{cases}$$

$$\widehat{f}(x)^{(2)} = \begin{cases} 2\widehat{\beta}_{21} + 6\widehat{\beta}_{31} x_i + 6\widehat{\beta}_{41}(x-\xi) & if\ x_i > \xi \\ 2\widehat{\beta}_{21} + 6\widehat{\beta}_{31} x_i & o.w \end{cases}$$

$$\widehat{f}(x)^{(3)} = \begin{cases} 6\widehat{\beta}_{31} + 6\widehat{\beta}_{41} & if\ x_i > \xi \\ 6\widehat{\beta}_{31} & o.w \end{cases}$$

이를 살펴보면, knot 주변을 살펴볼 때, 원래 모델에서, 1차, 2차 도함수에서 모두 연속이지만

3차 도함수에서만 불연속인 것을 확인할 수 있다. 즉, cubic spline을 만들기 위해서 basis function으로 $\beta_4 h(x, \xi)$을 추가하면, regression spline의 조건을 모두 충족하는 것이다.
일반적으로, cubic spline을 만들 때, K개의 knots을 설정하고자 하면, K개의 truncated power function을 추가한다. 즉, $X, X^2, X^3, h(X, \xi_1), \cdots, h(X, \xi_K)$ *where* $\xi_1, \cdots, \xi_K$ *are knots*을 예측변수로 하고 최소자승추정법을 시행하는 것이다. 이는 총, $K+4$개의 회귀 계수를 추정하는 것과 결과적으로 동일하므로, K개의 knots을 가진 cubic spline을 적합하는 것은 $K+4$의 자유도를 가진다.
하지만 splines는 예측변수의 바운더리 부분에서 높은 분산을 가진다. 즉, X가 매우 적거나 높은 값을 가지는 것이다.

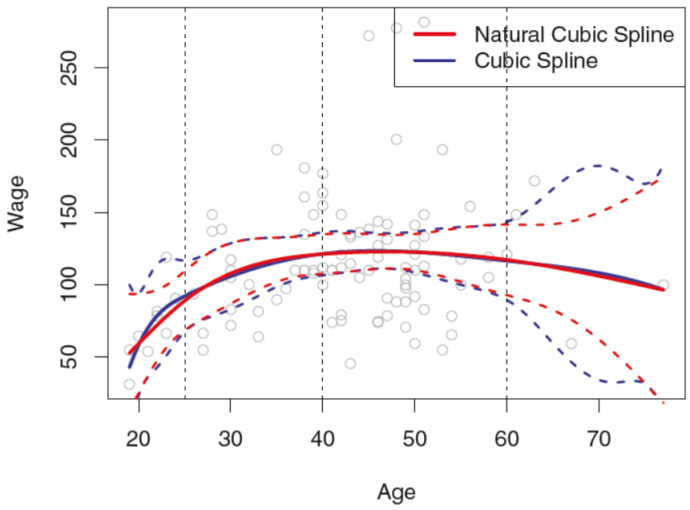

FIGURE 7.4. *A cubic spline and a natural cubic spline, with three knots, fit to a subset of the* `Wage` *data.*

figure 7.4는 cubic spline 적합 결과를 보여주는데, confidence band가 매우 넓은 것을 확인할 수 있다. 이에 대한 대안으로, natural spline는 추가적인 바운더리 제약이 있는 regression spline이다. 적합된 함수는 바운더리에서 (X가 가장 작은 knot보다 작거나, 가장 큰 knot보다 큰 구간) 선형이라는 추가 조건이 있다. 이러한 추가 제약을 통해 natural splines는 바운더리에서 좀 더 안정적인 추정치를 낸다.

7.4.4 Choosing the Number and Locations of the Knots

spline을 적합할 때, knots을 어디에 두어야 할까? knots가 많으면 많을수록 추정된 함수는 flexible하게 될 것이다. 이렇게 생각을 하면, 변동이 심한 곳에서 knots을 많이 두고, 변동이 적은 곳에서는 knots을 많이 두는 것도 하나의 방법일 것이다. 이러한 방법도 좋지만, 실제

로는 knots을 균등하게(uniform fasion) 두는 것이 일반적이다. 적절한 자유도를 채택하고, 컴퓨터로 하여금 자동적으로 데이터에 균등하게 knots을 배열하게 하는 것이 하나의 방법이다.

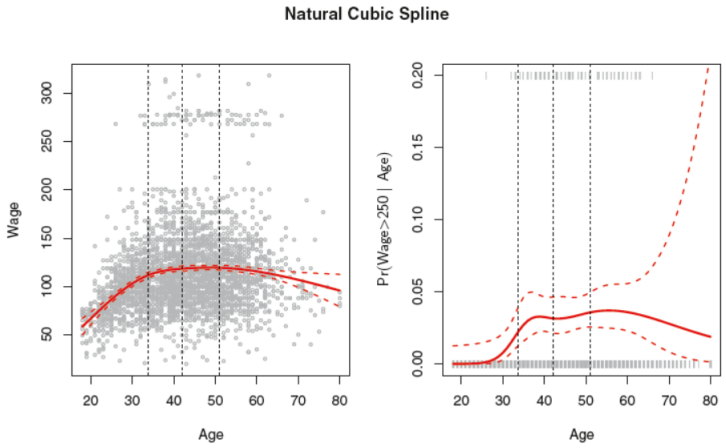

FIGURE 7.5. *A natural cubic spline function with four degrees of freedom is fit to the* `Wage` *data.* Left: *A spline is fit to* `wage` *(in thousands of dollars) as a function of* `age`. Right: *Logistic regression is used to model the binary event* `wage>250` *as a function of* `age`. *The fitted posterior probability of* `wage` *exceeding* $250,000 *is shown.*

figure 7.5는 figure 7.4와 같이 natural cubic spline을 3개의 knots을 적합한 결과인데, 자유도 4를 명시함으로써 자동적으로 age의 1분위, 2분위 3분위에 knots을 배열하도록 설정했다. 사실 여기에는 바운더리 knots까지 포함하여 5개의 knots이 있다. 5개의 knots을 가진 cubic splines는 총 5+4=9의 자유도를 가진다. 하지만 natural cubic splines는 각 바운더리에서 선형이어야 하는 조건으로 인해서 각각 자유도를 2씩 잃어, 자유도 5가 된다. 이것이 상수를 포함하기 때문에, 자유도를 4라고 하는 것이다(책의 내용을 우선 번역 했는데 의미는 생각해보기)

얼마나 많은 knots을 사용해야할까? cross-validation을 사용하는 것이 객관적인 방법이다.

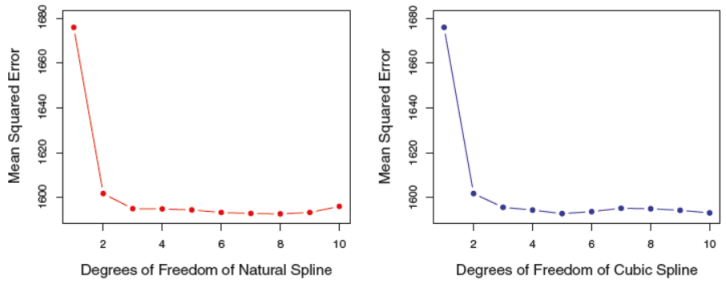

FIGURE 7.6. *Ten-fold cross-validated mean squared errors for selecting the degrees of freedom when fitting splines to the* `Wage` *data. The response is* `wage` *and the predictor* `age`. Left: *A natural cubic spline.* Right: *A cubic spline.*

figure 7.6은 Wage 데이터에 10 fold CV를 적용한 결과이다. 왼쪽 그림은 natural spline, 오른쪽 그림은 cubic spline이다. 두 모델 모두 비슷한 결과인데, 1차 spline(linear spline)이 적절하지 않음을 동일하게 보여준다. 3~4차 정도가 가장 적절해 보인다.

7.4.5 Comparison to Polynomial Regression

regression splines은 종종 polynomial regression보다 더 우수한 결과를 낸다. 그 이유는 높은 차수를 사용하는 다항 회귀와는 다르게, splines은 차수를 높이지 않고 knots을 통해 flexibility을 늘리기 때문이다. 일반적으로 이러한 접근법이 더 안정적인 추정치를 낸다.

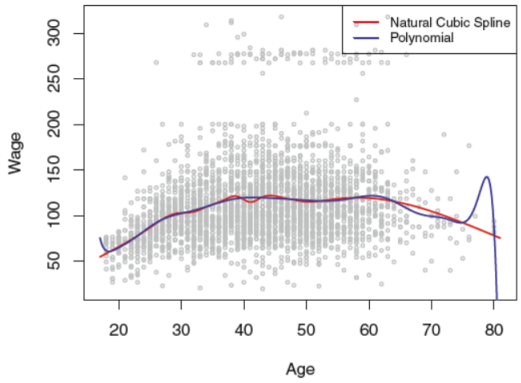

FIGURE 7.7. *On the* Wage *data set, a natural cubic spline with 15 degrees of freedom is compared to a degree-15 polynomial. Polynomials can show wild behavior, especially near the tails.*

figure 7.7을 보면, 다항 회귀가 바운더리에서 매우 이상한 형태의 함수를 추정했고 반면 natural cubic spline은 안정적인 추정치를 냈음을 확인할 수 있다.

7.5 Smoothing Splines

7.5.1 An overview of Smoothing Splines

이제 knot 선택 문제를 피하고 최대의 knots을 사용하는 방법에 대해 알아본다. 데이터에 대해서, 부드러운 곡선을 적합하고자 할 때, 어떤 함수 $g(x)$을 찾고자 하는데, SSE($\sum_{i=1}^{n}(y_i - g(x_i))^2$을 가장 최소로 하는 $g(x)$을 찾고자 한다. 하지만 $g(x_i)$에 어떤 제한을 두지 않으면 SSE을 0으로 만드는, 즉 y_i을 덮어 버리는(interpolate) g을 선택할 수 있는 위험이 있다. 이러한 함수 g는 훈련 데이터에 매우 과적합 되어서 분산이 매우 높아, 적합되지 않은 새로운 데이터인 테스트 세트에 대해서는 성능이 매우 나쁠 것이다. g가 SSE를 최소로 할뿐만 아니라, 매우 부드러운 형태를 원한다.

어떻게 함수 g를 부드럽게 만들까? 많은 방법이 있지만, 마치 6단원에서 살펴본 regularization처럼 최소화하고자 하는 quantitiy에 어떤 penalty을 추가하는 것이다. 즉, 아래의 식을 최소화 하는 함수 g를 찾는다.

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g^{''}(t)^2 dt \; where \; \lambda \; is \; a \; nonnegative \; tuning \; parameter \qquad (1)$$

위의 식을 최소화하는 함수 g를 smoothing splines라고 부른다. $\sum_{i=1}^{n}(y_i - g(x_i))^2$은 loss function으로써, g가 데이터에 잘 적합되게 하는 역할을 하고 $\lambda \int g^{''}(t)^2 dt$은 penalty 항으로써, g의 변동성을 제한하는 역할을 한다. $g^{''}(t)$은 g의 2차 도함수를 의미한다. 1차 도함수인 $g^{'}(t)$은 함수 t에서의 순간 기울기, 2차 도함수인 $g^{''}(t)$는 기울기가 얼마나 변화하는지를 나타낸다. 따라서 $g(t)$가 t 근방에서 구불구불한 모양이면 $g^{''}(t)$의 절대값은 클 것이고 그렇지 않다면 0에 가까울 것이다. $\int$은 적분을 의미하는데, t범위 에서의 summation이라고 생각하자(적분의 정의를 살펴보면 summation의 극한이기 때문이다) 이를 종합하면 $\int g^{''}(t)^2 dt$는 전체 범위에서 $g^{'}(t)$에서의 전체 변화를 나타낸다. 만약 g가 매우 부드러우면 $g^{'}(t)$는 상수에 가까울 것이고 $g^{''}(t)$가 0에 가까워져 $\int g^{''}(t)^2 dt$가 작은 값을 가지게 될 것이다. 반대로 g가 구불구불하고 변동이 심하다면 $g^{'}(t)$도 변동이 심할 것이고, $g^{''}(t)$도 큰 값을 가지게 되어 $\int g^{''}(t)^2 dt$ 또한 큰 값을 가질 것이다. 따라서 $\lambda \int g^{''}(t)^2 dt$는 g가 부드러운 모양이 되게 하는 역할을 한다. λ가 커질수록, $\int g^{''}(t)^2 dt$을 작게 해야, 전체적인 quantity도 최소화할 수 있고 그에 따라서 부드러운 모양이 나오게 되는 것이다. 극단적으로, $\lambda \longrightarrow \infty$라면 $\int g^{''}(t)^2 dt$을 0과 가깝게 만들어야 하며 이는 곧 g가 일직선이 됨을 의미한다. 6단원에서 regularization에서와 마찬가지로 spline의 λ 또한 bias-variance trade-off를 조정한다.

사실, 위 식을 최소화하는 $g(x)$는 $x_1, \cdots, x_n$에서 knots을 가지고 각 knot에서 1차, 2차 도함수가 연속인 piecewise cubic polynomial이다(책에는 증명이 없이 이렇다고만 나와있다) 게다가, 각 바운더리에서 선형 형태를 유지한다. 다시 말해서, 위 식을 최소화하는 $g(x)$는 $x_1, \cdots, x_n$에서 knots을 가지는 natural cubic spline인 것이다! 하지만 앞서 본 natural cubic spline의 shrunken 버전인데, λ가 shrinkage의 정도를 조절하기 때문이다.

7.5.2 Choosing the Smoothing Parameter λ

방금 살펴본 smoothing spline은 데이터의 개수만큼 knots이 있기 때문에 너무 많은 자유도를 가지고 있는 것이 아닌가 의아할 수 있다. 하지만 tuning parameter인 λ가 smoothing spline의 roughness, 즉 ***effective degrees of freedom***을 조절해준다. λ가 0에서 ∞로 증가할수록, effective degrees of freedom(이제 df_λ라고 적는다)는 n에서 2로 줄어듦을 보일 수 있다.(고만 나와있고 증면은 없다)

smoothing spline에서는 degrees of freedom 대신에 왜 effective degrees of freedom을 말할까? 보통 degrees of freedom은 자유로운 모수의 개수, 예를 들어 다항 회귀에서 적합된

계수의 개수를 의미한다. 비록 smoothing spline이 n개의 모수를 가지고 있어 degrees of freedom이 n일지라도, 이러한 n개의 모수는 constrained되거나 shrunk된다. 따라서 df_λ는 smoothing spline의 유연성의 척도인데, 더 높을 수록 더 flexible하다는 뜻이다. effective degrees of freedom의 정의는 다소 기술적이다.

앞서, (1)을 최소화하는 모델은 natural cubic spline임을 배웠다. 즉, $g(x) = \sum_{j=1}^{n} N_j(x)\theta_j$로 표현할 수 있다(ESL exercise 5.4, 해답은 https://stats.stackexchange.com/questions/172217/why-are-the-basis-functions-for-natural-cubic-splines-expressed-as-they-are-es 참조) 즉, (1)을 다시 써보면,

$$\begin{aligned} Q(\theta, \lambda) &= \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g^{''}(t)^2 dt \\ &= (\mathbf{y} - \mathbf{N}\theta)^T((\mathbf{y} - \mathbf{N}\theta) + \lambda\theta^T \mathbf{\Omega}_N \theta \end{aligned}$$

$$where\ \{\mathbf{N}\}_{ij} = N_j(x_i)\ and\ \{\mathbf{\Omega}_N\}_{jk} = \int N_j^{''}(t) N_k^{''}(t) dt$$

최소자승법에 의해서 $Q(\theta, \lambda)$을 최소화하는 $\widehat{\theta}$을 쉽게 찾을 수 있다.

$$\widehat{\theta} = (\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y}$$

fitted smoothing spline은 아래와 같다.

$$\hat{f}(x) = \sum_{i=1}^{n} N_j(x)\widehat{\theta}_j$$

벡터 노테이션으로는

$$\hat{\mathbf{f}} = \mathbf{N}\widehat{\theta} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y} = \mathbf{S}_\lambda\mathbf{y}$$

여기서 $\mathbf{S}_\lambda$는 smoother matrix이다. $\mathbf{S}_\lambda$은 $\mathbf{y}$에 의존하지 않고 오로지 x_i, λ에만 의존한다. effective degrees of freedom은 smoother matrix의 diagonal elements이다.

$$df_\lambda = \sum_{i=1}^{n}\{\mathbf{S}_\lambda\}_{ii}$$

즉, effective degrees of freedom은 다중 선형 회귀에서 Hat matrix의 trace와 비슷한 개념이다. 다중 선형 회귀에서 fitted value는 $\hat{\mathbf{f}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$로 표현한다. 즉, $trace(\mathbf{H})$와 df_λ는 비슷한 역할을 할 것임을 짐작할 수 있다.

smoothing spline을 적합할 때, knots의 개수나 위치에 관해 걱정할 필요가 없다. 훈련

데이터를 knots로 정의하기 때문이다. 하지만 λ을 정해야 하는 문제가 있다. 물론 이를 cross-validation을 통해서 해결한다. LOOCV가 아래 공식을 이용하여 거의 동일한 계산 시간으로 smoothing spline을 효과적으로 계산할 수 있음이 밝혀졌다.

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

$\hat{g}_{\lambda}^{(-i)}(x_i)$은 i번째 관측치를 제외하고 적합한 smoothing spline에서 x_i의 fitted value에 대한 노테이션이다.

7.6 Local Regression

local regression은 비선형 함수를 적합하는 또 다른 방법이다. 이는 어떤 점 근방의 몇 개의 관측치만을 사용하여 근방의 함수를 적합한다.

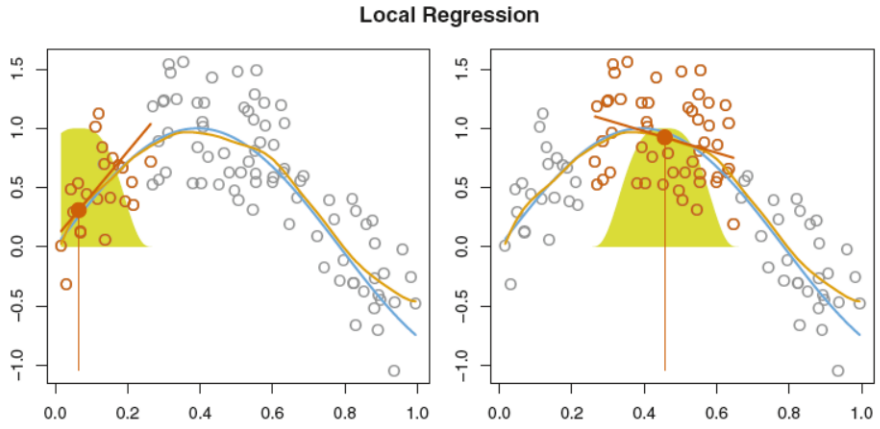

FIGURE 7.9. *Local regression illustrated on some simulated data, where the blue curve represents $f(x)$ from which the data were generated, and the light orange curve corresponds to the local regression estimate $\hat{f}(x)$. The orange colored points are local to the target point x_0, represented by the orange vertical line. The yellow bell-shape superimposed on the plot indicates weights assigned to each point, decreasing to zero with distance from the target point. The fit $\hat{f}(x_0)$ at x_0 is obtained by fitting a weighted linear regression (orange line segment), and using the fitted value at x_0 (orange solid dot) as the estimate $\hat{f}(x_0)$.*

figure 7.9의 왼쪽 그림을 보면 0.05 근방의 몇 개의 점을 사용하여 근방의 선을 적합하고 오른쪽 그림또한 0.4 근방의 몇 개의 점을 사용하여 근방의 선을 적합하였다. 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 $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

세 번째 스텝에서 각 관측치마다 가중치가 다름에 주목하자. 다시 말해서 OLS가 아닌 WLS(weighted least squares)을 통해서 local regression의 계수 추정치를 구하는 것이다. local regression을 실행하기 위해서는 가중치 함수 K, 세 번째 스텝에서 선형, 상수, 비선형 회귀를 시행할 것인지 등의 논의가 필요하다. 하지만 가장 중요한 것은 *span* s을 정하는 것이다. 이 span은 smoothing spline의 tuning parameter λ와 같은 역할을 한다. 비선형 적합의 flexibility을 결정한다. 매우 큰 s 값은 모든 훈련 데이터를 사용함으로써 전역적인 적합(global fit)이 될 것이다.

local regression 아이디어는 많은 방법으로 일반화될 수 있따. p개의 변수에 대해서 다중 회귀를 시행할 때, 어떤 변수에 대해서는 전역적으로, 다른 변수에 대해서, 예를 들어 시간에 대해서는 국소적으로 적합을 하는 것이 하나의 예이다. 이렇게 변화하는 계수(varying coefficients) 모델은 가장 최근에 모아진 데이터에 대해서 모델을 적합할 때 효과적인 방법이다. local regression은 변수가 2개 이상일 때에도 일반화될 수 있지만, 일반적으로 변수가 3개나 4개 이상이 된다면, Nearest-neighbors 접근법이 차원이 높아질수록 성능이 안 좋아지는 것과 동일하게 바람직하지 않은 결과를 낼 수 있다.

7.7 Generalized Additive Models

GAM은 표준적인 각 변수의 비선형성을 허용하고 additivity는 유지하면서 선형 모델을 확장한다. 연속형 변수, 범주형 변수 모두에 대해서 쓰일 수 있고 먼저 연속형 변수에 대해서 살펴보자.

7.7.1 GAMs for Regression Problems

예측 변수에 대한 비 선형 함수 $f_j(x_{ij})$을 통해서 비선형 관계를 모델링한다.

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

GAM은 additive 모델이라고 불리는데 그 이유는, 각 함수 f_j가 서로 더해져 있기 때문이다. 7.1에서 7.6까지 단 변수에 대해서 비선형 관계를 적합하는 다양한 방법을 배웠는데, GAMs은 이러한 방법을 additive 모델로 모두 사용할 수 있다는 장점이 있다.

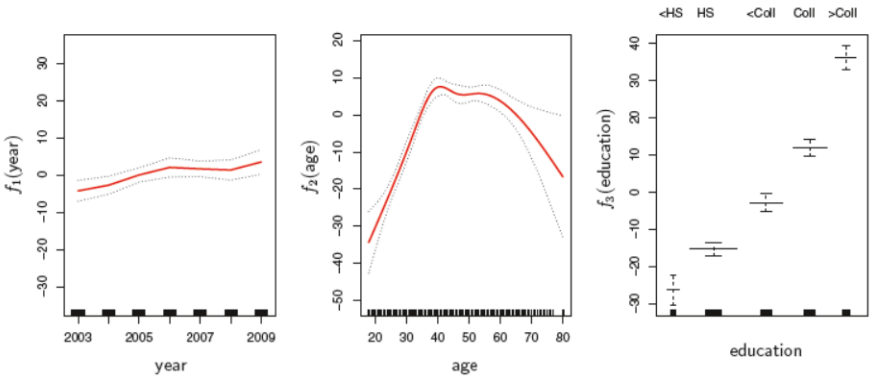

FIGURE 7.11. *For the* **Wage** *data, plots of the relationship between each feature and the response,* **wage**, *in the fitted model (7.16). Each plot displays the fitted function and pointwise standard errors. The first two functions are natural splines in* **year** *and* **age**, *with four and five degrees of freedom, respectively. The third function is a step function, fit to the qualitative variable* **education**.

figure 7.11은 year, age 변수에 대해서는 natural spline을, education 변수에 대해서는 더미 변수를 사용하여 최소자승법을 실행한 결과이다. 이와 같이 각 변수마다 사용하고 싶은 방법을 골라서 적용하는 것이다.

Pros and Cons of GAMs

- GAMs은 각 변수 X_j에 대해 비선형 함수 f_j을 적합하여 표준 선형 회귀가 놓치는 비선형 관계를 모델링할 수 있다. 이는 각 변수에 대해서 개인적으로 여러 다양한 변환을 시도해볼 필요가 없음을 의미한다.
- 비선형 적합은 반응변수에 대해 더 정확한 예측을 만든다.
- 모델이 additive하기 때문에 X_j가 반응변수 Y에 미치는 영향을 다른 변수를 고정한 채, 알 수 있다. 따라서 추론이 목적이라면 GAMs는 유용할 것이다.

- GAMs의 가장 큰 단점은 additive하게 제한된다는 것이다. 많은 변수들에 대해서 중요한 상호 관계를 놓칠 수 있다. 하지만 선형 회귀와 마찬가지로 교호작용 항을 추가함으로써 GAMs에도 $X_j \times X_k$의 상호작용을 파악할 수 있다.

7.7.2 GAMs for Classification Problems

GAMs은 분류문제에도 사용될 수 있다. 간단하게 살펴보기 위해, Y가 0또는 1 값을 가진다고 하고 $p(X) = Pr(Y = 1 \mid X)$를 반응변수가 1이 될 조건부 확률이라고 하자. 로지스틱 회귀에 GAMs을 적용하면 아래와 같다.

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 f_1(X_1) + \beta_2 f_2(X_2) + \cdots + \beta_p f_p(X_p)$$

위에서 언급한 장, 단점을 분류 문제에서의 GAMs도 모두 동일하게 가진다.