Rozszerzenia bazy

Piecewise polynomials and splines

Karolina Sopata

21 marca 2022 r.

Model liniowy

$$E(Y|X) = f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

- prosty w interpretacji
- nie powoduje overfittingu przy małej liczbie N i/lub dużym p
- nie zawsze adekwatny

Przykład kwadratowego modelu regresji:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

Rozszerzenia bazy

X - wektor zmiennych objaśniających o długości p
$h_m(X) : \mathbb{R}^p \to \mathbb{R}$
m = 1,...,M

$$E(Y|X) = f(X) = \sum_{m=1}^{M} \beta_m h_m(X) = \beta_1 h_1(X) + ... + \beta_M h_M(X)$$

$h_m(X)$ - funkcje bazowe rozpinające przestrzeń funkcji f

Przykłady funkcji bazowych

- $h_m(X) = X_m$, gdzie m=1,...,p
- $h_m(X) = X_j^2$
  $h_m(X) = X_j X_k$
- $h_m(X) = log(X_j)$
  $h_m(X) = \sqrt{X_j}$
- $h_m(X) = ||X||$
- $h_m(X) = \mathbb{1}(L_m \leqslant X_k < U_m)$

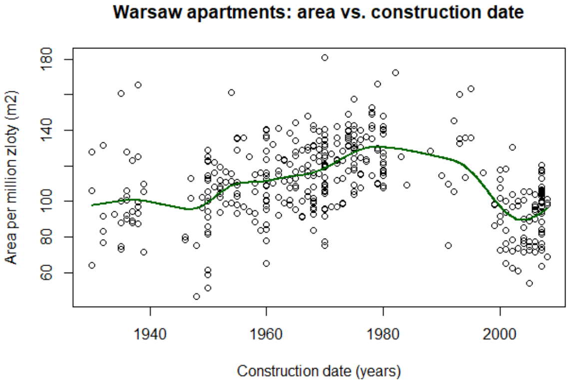
Piecewise polynomials and splines

- przyjmujemy, że wektor X jest jednowymiarowy

- **piecewise polynomial function** - wielomian odcinkowy - dzielimy dziedzinę X na rozłączne przedziały, na każdym z przedziałów definiujemy osobną funkcję wielomianową

- **spline** - funkcja sklejana - rzeczywista funkcja gładka, dla której istnieje rodzina podprzedziałów dziedziny taka, że funkcja ta jest wielomianem na każdym z tych podprzedziałów

Przykład funkcji gładkiej



Rysunek: Model powierzchni jaką można nabyć w m^2 za milion złotych w zależności od daty budowy budynku mieszkalnego, 30 funkcji bazowych

Różne drogi wyboru funkcji bazowych

1. **Restriction methods** - stopień złożoności modelu jest ograniczony poprzez z góry określony limit klas funkcji jakie mogą zostać użyte

$$f(X) = \sum_{j=1}^{p} f_j(X_j) = \sum_{j=1}^{p} \sum_{m=1}^{M_j} \beta_{jm} h_{jm}(X_j)$$

2. **Selection methods** - wybór spośród wszystkich funkcji, sprawdzenie które funkcje istotnie zmieniają dopasowanie modelu do danych, np. metoda CART, MARS
3. **Regularization methods** - użycie wszystkich dostępnych funkcji z zastosowaniem pewnych restrykcji, np. ridge regression, LASSO

Piecewise polynomials - przykład 1

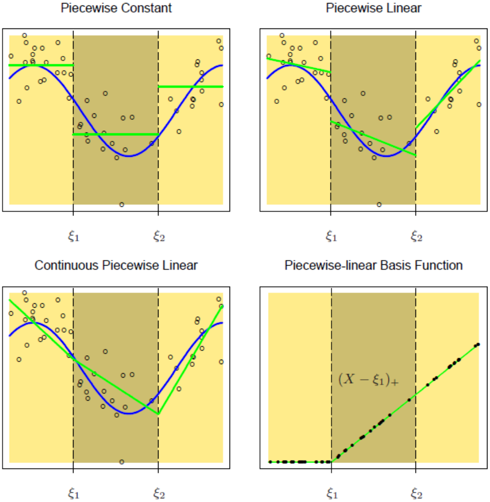
$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 \leq X)$$

- powyższe funkcje przyjmują dodatnie wartości na rozłącznych przedziałach
- estymator najmniejszych kwadratów dla modelu

$$f(X) = \sum_{m=1}^{3} \beta_m h_m(X)$$

sprowadza się do $\hat{\beta}_m = \bar{Y}_m$ (średnia zmiennych objaśnianych na m-tym przedziale)

Piecewise polynomials - przykład 1 c.d.



Rysunek: W lewym górnym rogu: odcinkowa funkcja stała, w prawym górnym rogu: odcinkowa funkcja liniowa; źródło: *T.Hastie, The elements of statistical learning*

Piecewise polynomials - przykład 1 c.d.

$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 \leq X)$$

$h_4(X) = \mathbb{1}(X < \xi_1)X$
$h_5(X) = \mathbb{1}(\xi_1 \leqslant X < \xi_2)X$
$h_6(X) = \mathbb{1}(X \leqslant \xi_2)X$

Warunek ciągłości w węzłach:

$$f(\xi_i^-) = f(\xi_i^+)$$

Stąd:

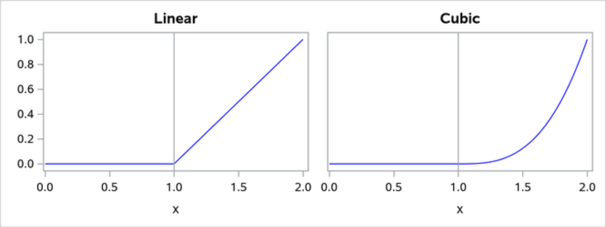
$$\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$$

$$\beta_2 + \xi_2\beta_5 = \beta_3 + \xi_1\beta_6$$

Truncated power function

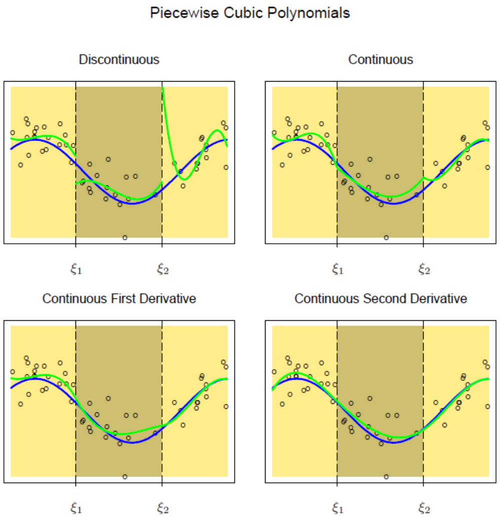
Truncated power function dla węzła ξ_i:

$$h_i(X) = (X - \xi_i)_+^d = \begin{cases} 0, & X < \xi_i \\ (X - \xi_i)^d, & X \geqslant \xi_i \end{cases}$$



Rysunek: Truncated power function dla węzła $\xi_i = 1$ dla d=1 oraz d=3; *źródło: https://documentation.sas.com*

Piecewise polynomials - przykład 2



Rysunek: Przykłady sześciennych wielomianów odcinkowych dopasowanych do tych samych danych; źródło: *T.Hastie, The elements of statistical learning*

Cubic spline

Generalnie:
$h_j(X) = X^{j-1}, j = 1, ..., M$
$h_{M+l}(X) = (X - \xi_l)_+^{M-1}$, l = 1,...,K

Funkcje bazowe dla sześciennej funkcji składanej z dwoma węzłami:

K = 2
M = 4
$h_1(X) = 1$
$h_2(X) = X$
$h_3(X) = X^2$
$h_4(X) = X^3$
$h_5(X) = (X - \xi_1)_+^3$
$h_6(X) = (X - \xi_2)_+^3$

Model

$$E(Y|X) = f(X) = \beta_1 h_1(X) + ... + \beta_6 h_6(X)$$

Splines

- ważnym zagadnieniem jest odpowiedni dobór stopnia funkcji sklejanej, liczby węzłów i ich położenia

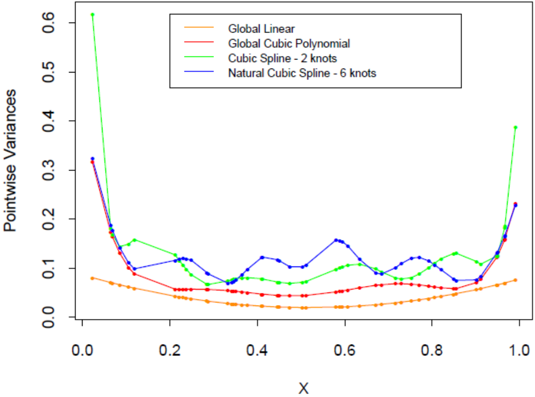
- położenie węzłów może być dobrane metodą parametryzacji rodziny funkcji sklejanych według liczby funkcji bazowych lub stopni swobody i wyznaczone na podstawie rozmieszczenia zmiennych

Polecenie w R:

$$bs(x, df=7)$$

generuje macierz funkcji bazowych na podstawie N obserwacji dla funkcji sklejanej stopnia 3 z 7-3=4 węzłami.

Natural cubic splines



Rysunek: Krzywe wariancji punktowej dla różnych modeli dopasowanych do danych pochodzących z rozkładu jednostajnego z dodanym szumem o stałej wariancji; źródło: *T.Hastie, The elements of statistical learning.*

Natural cubic spline

- warunek: funkcja poza węzłami zewnętrznymi jest liniowa
- wprowadzenie takiego warunku powoduje zmniejszenie się wariancji punktowej poza węzłami zewnętrznymi, natomiast dzieje się to kosztem obciążenia

Przykład

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

gdzie:

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

Smoothing splines

- złożoność dopasowania jest kontrolowana przez parametr regularyzacji λ
- λ - parametr wygładzający

$$RSS(f, \lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt$$

Skrajne przypadki

- $\lambda = 0$ - f może być dowolną funkcją, która przybliża rzeczywistą zależność
- $\lambda = \infty$ - prosta dopasowana metodą najmniejszych kwadratów

Smoothing splines

Zdefiniowane na poprzednim slajdzie kryterium posiada skończenie wymiarowe unikalne minimum będące naturalnym sześciennym splajnem z węzłami w wartościach x_i, gdzie i = 1,...,N.

Rozwiązanie - naturalny splajn sześcienny, stąd funkcję f możemy przedstawić w następującej postaci:

$$f(x) = \sum_{j=1}^{N} N_j(x)\theta_j$$

gdzie $N_j(X)$ - N-wymiarowy zbiór funkcji bazowych reprezentujący daną rodzinę splajnów naturalnych.

Smoothing splines

Kryterium:

$$RSS(\theta, \lambda) = (y - N\theta)^T (y - N\theta) + \lambda \theta^T \Omega_N \theta$$

gdzie $\{N\}_{ij} = N_j(x_i)$,
$\{\Omega_N\}_{jk} = \int N_j''(t) N_k''(t) dt$

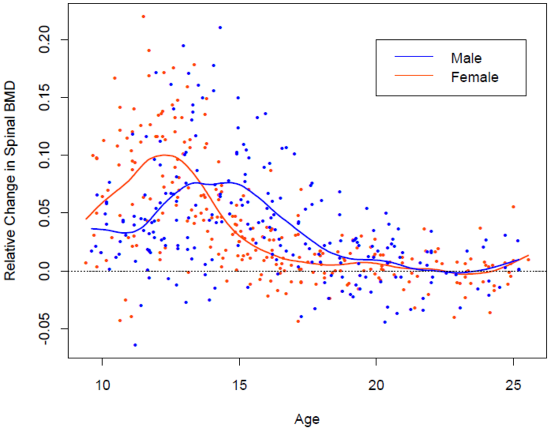
Rozwiązanie (uogólniona regresja grzbietowa):

$$\hat{\theta} = (N^T N + \lambda \Omega_N)^{-1} N^T y$$

Stąd:

$$\hat{f}(X) = \sum_{j=1}^{N} N_j(X) \hat{\theta}_j$$

Przykład dopasowania funkcji gładkiej



Rysunek: Dane: względna zmiana w mineralizacji kości kręgosłupa w ciągu kolejnych dwóch badań w odstępie roku w zależności od wieku badana u młodzieży; *źródło: T.Hastie, The elements of statistical learning.*

Dziękuję za uwagę!