Data Science II
(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

Moving Beyond Linearity

- The truth is almost never linear!
- But often the linearity assumption is good enough
- When it's not, consider
  - polynomials
  - piecewise polynomials
  - splines
  - local regression
  - generalized additive model (GAM)
  - multivariate adaptive regression splines (MARS)
  - ...
- More flexibility without losing the ease of linear models
- For now, we assume X is one-dimensional

Polynomial Regression

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$
- Create new variables $X_1 = X, X_2 = X^2$ and then treat as multiple linear regression
- More interested in the fitted values at any value x_0

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \ldots + \hat{\beta}_d x_0^d$$

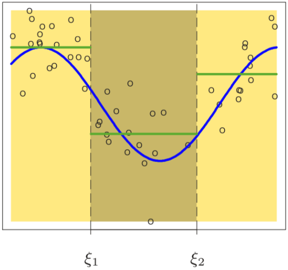
- Choice of d
  - Fix the degree d at some reasonably low value
  - Use cross-validation
- Caveat: polynomials have notorious tail behavior – bad for extrapolation

Step Functions

- Piecewise constant
- Cut the variable into distinct regions and construct new variables, $c_0(X), \ldots, c_K(X)$, where

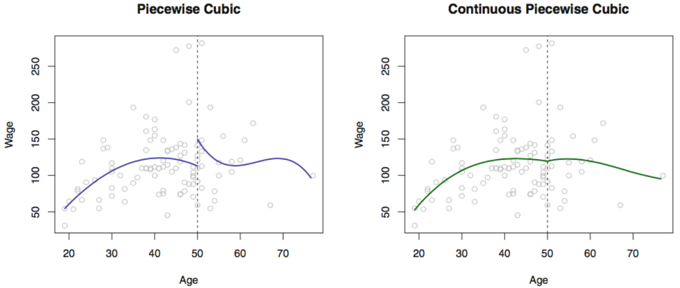
$$c_0(x) = I(x < \xi_1),\ c_1(x) = I(\xi_1 \leq x < \xi_2), \ldots,$$
$$c_{K-1}(x) = I(\xi_{K-1} \leq x < \xi_K),\ c_K(x) = I(x \geq \xi_K)$$

- Create a series of dummy variables representing each group
- Choice of cutpoints or knots can be problematic



Piecewise Polynomials

- Different polynomials in regions defined by knots
- Better to add constraints to the polynomials, e.g. continuity



[ISL] Figure 7.3

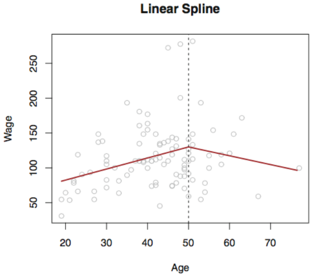
Linear Splines

- A piecewise linear polynomial continuous at each knot ξ_k
- Model: $y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) \ldots \beta_{K+1} b_{K+1}(x_i) + \epsilon_i$
- b_k are basis functions

$$b_0(x) = 1, \; b_1(x) = x, \; b_{k+1}(x) = (x - \xi_k)_+, k = 1, \ldots, K$$

where

$$(x - \xi_k)_+ = \begin{cases} x - \xi_k & \text{if } x > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

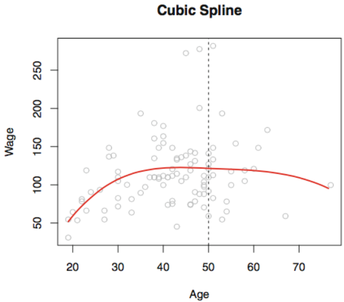


Cubic Splines

- A piecewise cubic polynomial with continuous derivatives up to order 2 at each knot ξ_k
- $y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \ldots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$ where

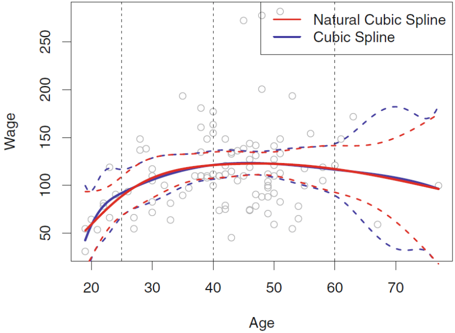
$$b_0(x) = 1,\ b_1(x) = x,\ b_2(x) = x^2,\ b_3(x) = x^3,$$

$$b_{k+3}(x) = (x - \xi_k)_+^3,\ k = 1, \ldots, K$$



Natural Cubic Splines

- Cubic splines can have high variance at the outer range of the predictors



- Natural cubic spline extrapolates linearly beyond the boundary knots
- More stable estimates at the boundaries

Natural cubic spline

Basis function of natural cubic spline with K knots

- $N_1(x) = 1$
- $N_2(x) = x$
- The remaining basis are $N_{k+2}(x) = d_k(x) - d_{K-1}(x)$ for $k = 1, \ldots, K-2$, where

$$d_k(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k}, k = 1, \ldots, K-1.$$

Knots placement

- One strategy is to decide K, the number of knots, and then place them at appropriate quantiles of the observed X
- A cubic spline with K knots has $K + 4$ parameters or degrees of freedom
- A natural cubic spline with K knots has K degrees of freedom
- Method that avoids the knot selection problem?

Smoothing Splines

Consider this criterion for fitting a smooth function $g(x)$ to the training data

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

- The second term is a roughness penalty
  - $\lambda \to 0$?
  - $\lambda \to \infty$?
- The solution is a natural cubic spline, with knots at every unique value of x_i

$$g(x) = \sum_{j=1}^{n} N_j(x)\theta_j$$

Smoothing spline

- $g(x) = \sum_{j=1}^{n} N_j(x)\theta_j$
- $\{\boldsymbol{N}\}_{ij} = N_j(x_i)$, $\{\boldsymbol{\Omega}_N\}_{jk} = \int N_j''(t)N_k''(t)dt$
- The criterion reduces to

$$RSS(\theta, \lambda) = (\boldsymbol{y} - \boldsymbol{N}\theta)^\intercal(\boldsymbol{y} - \boldsymbol{N}\theta) + \lambda\theta^\intercal\boldsymbol{\Omega}_N\theta$$

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

Smoothing spline

Demmler-Reinsch basis (1975)

Choosing λ

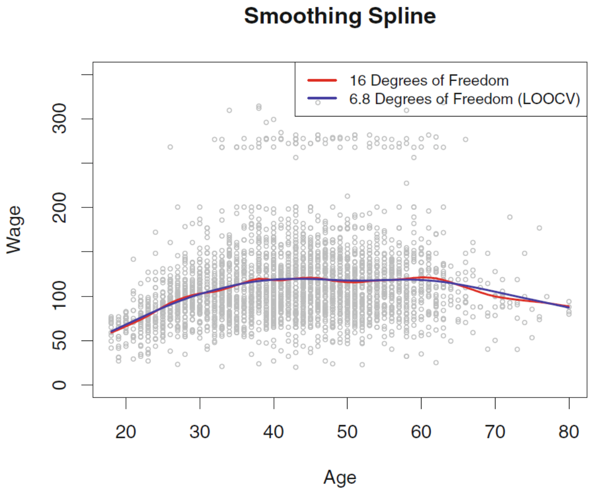
- Smoothing splines avoid the knot-selection issue, leaving a single λ to be chosen
- The vector of n fitted values can be written as $\hat{\boldsymbol{g}}_\lambda = \boldsymbol{S}_\lambda \boldsymbol{y}$
- The effective degrees of freedom are given by

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

- Cross-validation
- The leave-one-out cross-validation (LOOCV) error is given by

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

Wage data



[ISL] Figure 7.8