Topic 7. Moving beyond Linearity

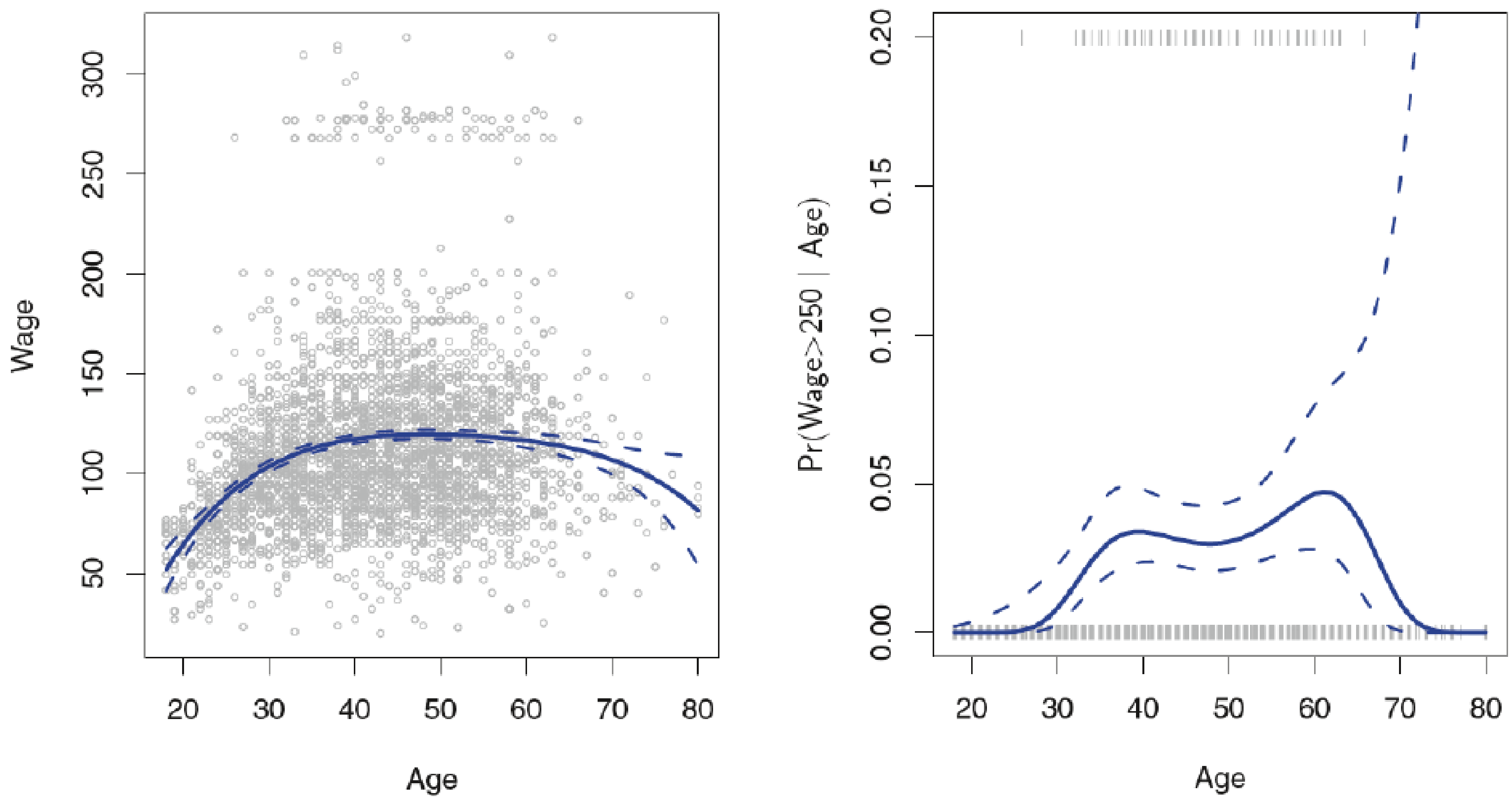
Polynomial Regression

- Polynomial regression extends simple linear regression by replacing the linear regression function with a polynomial function

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

- This model is a special case of multiple linear regression, and thus coefficients can be easily estimated.
- Usually d is set as 2, 3, or 4, and rarely goes beyond 4.
- Usually centered predictor $\tilde{x}_i = x_i - \bar{x}$ is used, to reduce correlation among different-ordered terms.

Examples with Degree-4 Polynomial Regression



Polynomial Logistic Regression

- The left figure suggests that the wages are from two distinct populations.
- Construct a binary response
  - High earners groups earning more than $250,000 per annum.
  - Low earners group otherwise.
- Polynomial logistic regression assumes that

$$\text{logit}\big(Pr(y_i > 250|x_i)\big) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d$$

- Coefficients can be estimated similarly.

Step Functions

- Step function provides a way to approximate nonlinear structure locally.

- It converts continuous variable into ordered binary variable.
  - Let $c_1, \dots, c_k$ be K breaks points in the range of X.
  - Construct $K+1$ new variables:

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

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

$$\vdots$$

$$C_K(X) = I(X \geq c_k)$$

where $I(\cdot)$ is an indicator function.

Step Functions

- X must be in exactly one of the $K + 1$ intervals, and

$$C_0(X) + C_1(X) + C_2(X) + \cdots + C_K(X) = 1$$

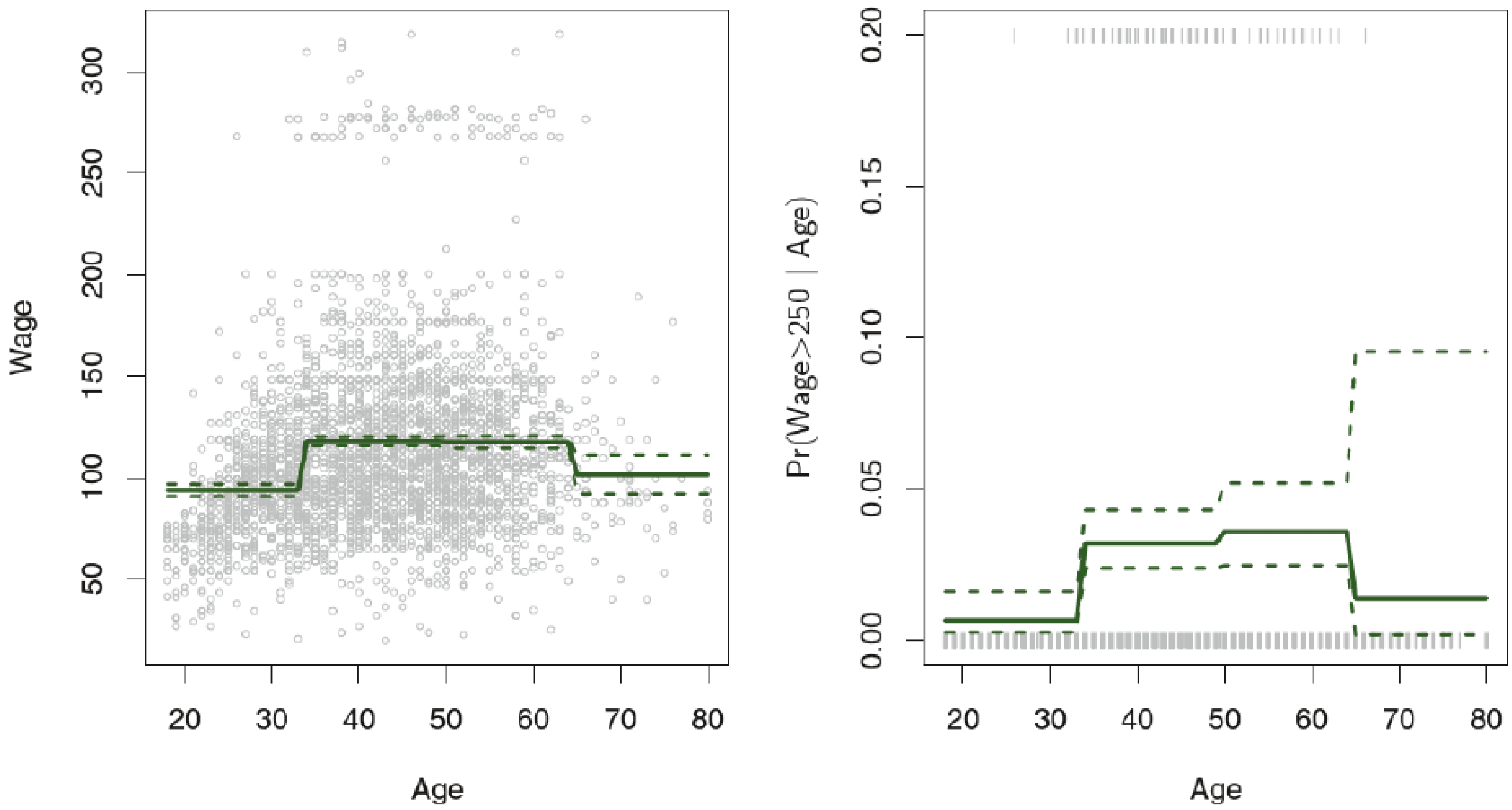
- The linear regression function can be replaced by

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

- The fitted response is $\beta_0 + \beta_k$ if $c_k \leq X < c_{k+1}$, and β_0 if $X < c_1$.

- Thus the regression function is a step function, piecewise constant function.

Example of Step Functions



Interaction Regression Model

- When there are more than one predictors in the model, we can consider their interactions, e.g.,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i$$

- The mean response change per unit change in each variable now depends on the other variables.

Piecewise Polynomials

- Piecewise polynomial regression fits separate low-degree polynomials over different regions of X.

- For example, a piecewise cubic regression model is

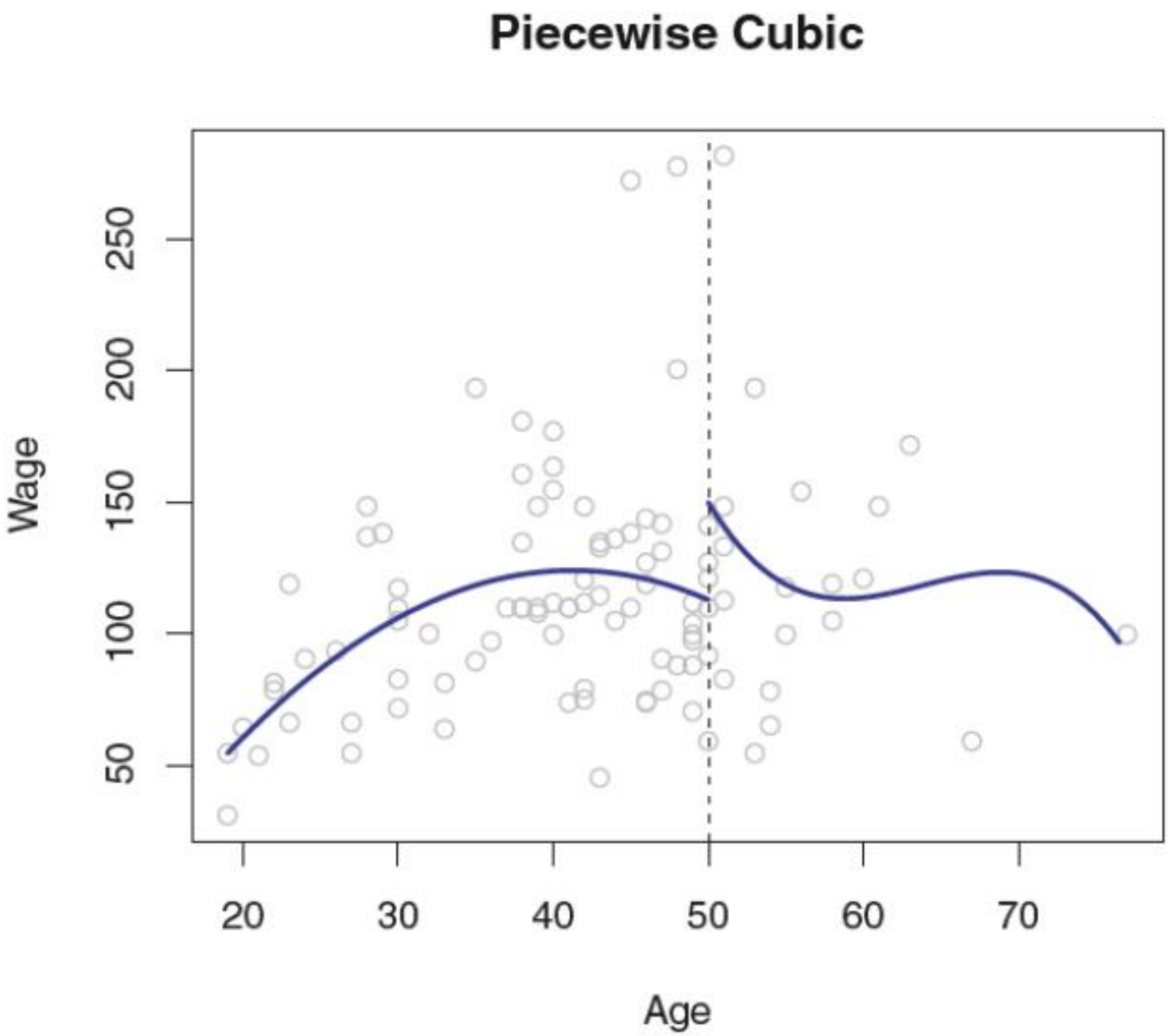
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & if\, x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}{x_i}^2 + \beta_{32}{x_i}^3 + \epsilon_i & if\, x_i \geq c \end{cases}$$

  c is a **knot**.

- This model gives two different cubic models for observations with $x_i < c$ and $x_i \geq c$.

- Using more knots leads to a more flexible piecewise polynomial.

Example of Piecewise Cubic



Continuous Piecewise Polynomials

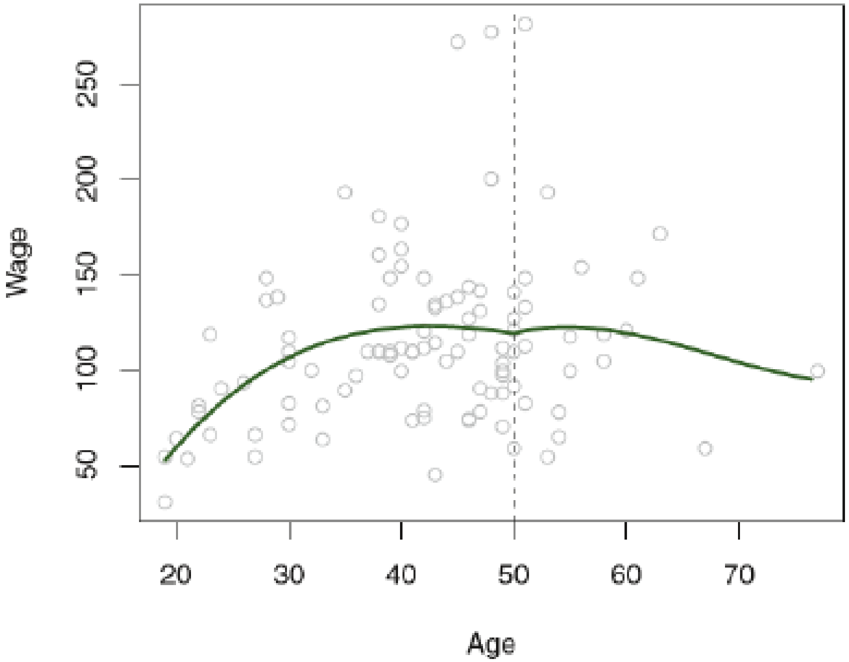
- The fitted curve in last figure is “problematic”: the predicted wages jump at age 50!

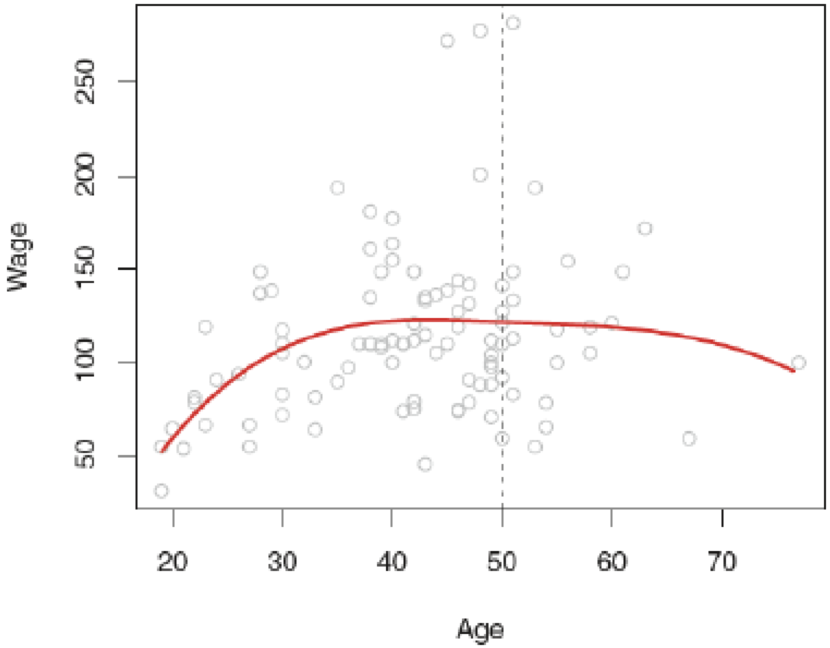
- One can fit a piecewise polynomial with the constraint that the fitted curve must be continuous.

- In addition, one also require the derivatives of the piecewise polynomials are continuous.

Continuous Piecewise Polynomials (Cont.)

- **Left:** Continuous piecewise polynomials; **Right:** Piecewise polynomials with continuous first and second order derivatives (cubic spline)





Basis Functions

- In general, we can fit a regression model

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

where $b_1(\cdot), \ldots, b_K(\cdot)$ are fixed and known basis functions.

- For polynomial regression: $b_k(x) = x^k$

- For step functions: $b_k(x) = I(c_k \leq x < c_{k+1})$

- Many other possible basis functions can be employed, leading to various nonlinear models.

Cubic Spline

- **Cubic spline** is a continuous piecewise polynomials with continuous first and second order derivatives.

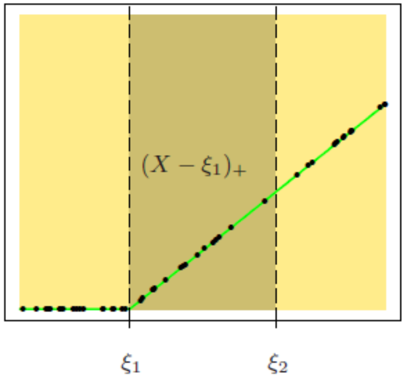
- A cubic spline with K knots can be modeled as

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

where the basis functions are x, x^2, x^3, and the truncated power basis function

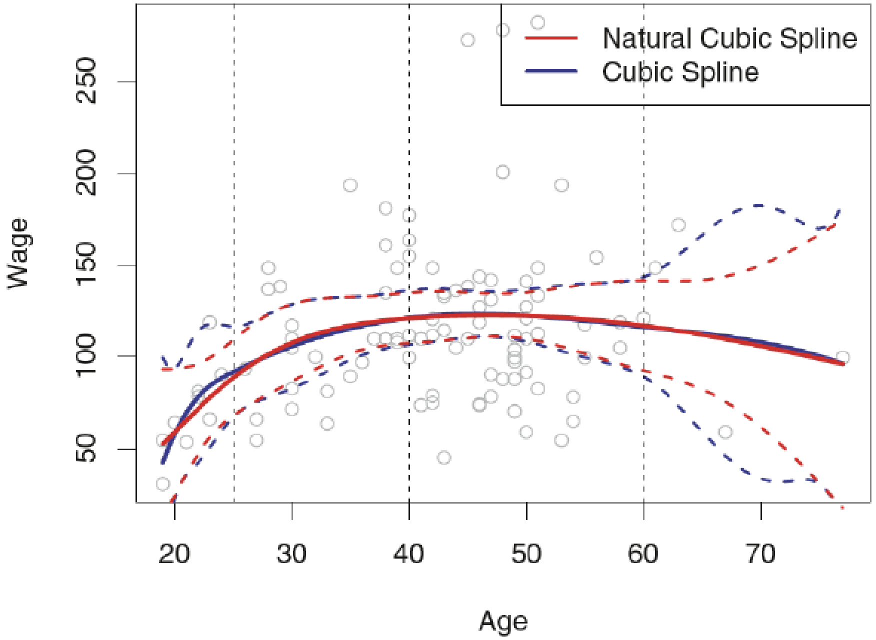
$$h(x, \xi_k) = (x - \xi_k)_+^3 = \max\big((x - \xi_k)^3, 0\big)$$

at each knot ξ_k, $k = 1, \ldots, K$.



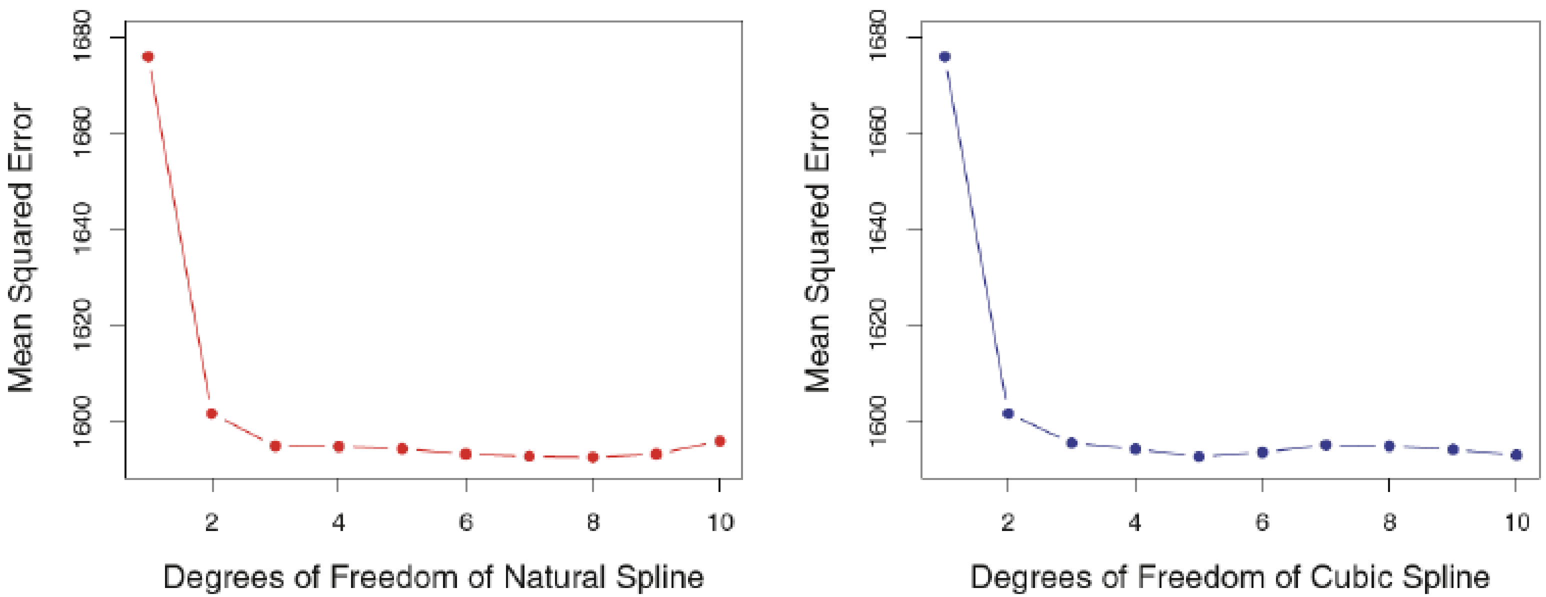
Cubic Spline (Cont.)

- Splines can have high variance at the outer range of the predictors.
- Natural spline is a regression spline requiring the model to be linear at the boundary. With this additional constraint, it generally produces more stable estimates at the boundaries.



Determining Knots

- Knots can be placed in a uniform fashion, say the knots in last figure are the 25th, 50th and 75th percentiles of Age.
- Sophisticated ways of placing knots are also available.
- The number of knots affects the complexity (degree of freedom) of the fitted model, and can be chosen via cross validation.



Smoothing Spline

- Smoothing spline is to find g that minimizes

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

- $\sum_{i=1}^{n}(y_i - g(x_i))^2$ is a **loss function**, encouraging g to fit the data well.
- $\int (g''(t))^2 dt$ is a **penalty** term that penalizes the complexity of g, where $g''(t)$ measures the roughness of g.
- $\lambda \geq 0$ is a tuning parameter.
- Other loss functions and penalty terms can also be used.

Smoothing Spline (Cont.)

- The minimizer $\hat{g}(x)$ can be shown to have the properties:
  - It is a piecewise cubic polynomial with knots at $x_1, \ldots, x_n$.
  - It has continuous first and second derivatives at each knot.
  - It is linear in the region outside of the extreme knots.

- It is a natural cubic spline with knots at $x_1, \ldots, x_n$.

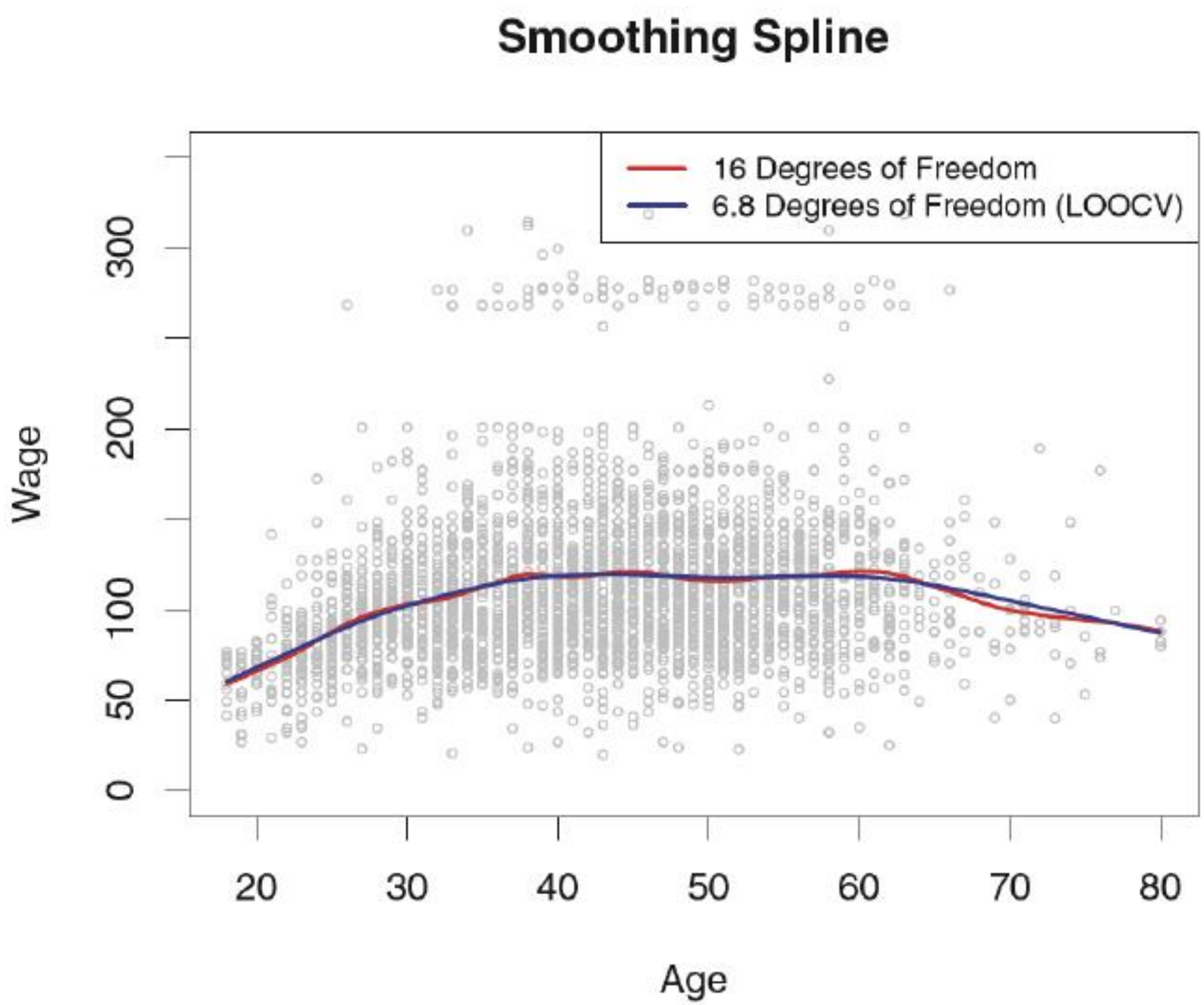
- But it is NOT the same natural cubic spline from the basis function approach.

- It is a shrunken version of such a natural cubic spline, where the level of shrinkage is controlled by λ.

Tuning Parameter

- When $\lambda = 0$, the penalty has no effect and g will interpolate the training observations.
- When $\lambda \to \infty$, smoothing spline degenerates to simple linear regression.
- Clearly, λ controls the bias-variance trade-off of the smoothing spline, and different λ leads to different $\hat{g}_\lambda$.
- The optimal λ can be determined by cross validation.

Example of Smoothing Spline



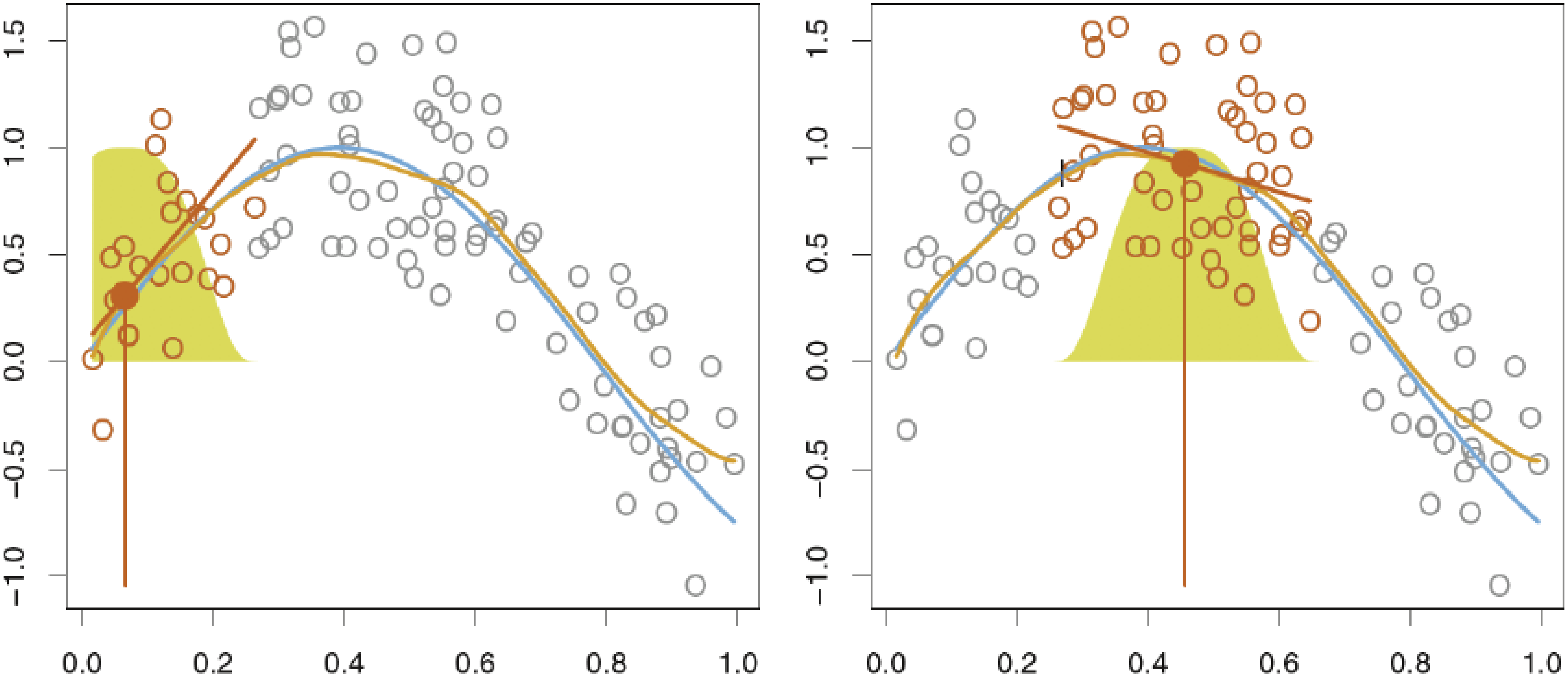
Local Linear Regression

- **Local linear regression** computes the fit at x_0 by fitting a linear model only to its nearby training observations.
  - Gather s training observations whose x_i are closest to x_0.
  - Assign weight $K_{i0} = K(x_i, x_0)$ to each point, which is smaller if x_i is further away from x_0, and is 0 if it is not one of the s closest observations.
  - Fit a weighted linear regression by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

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

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

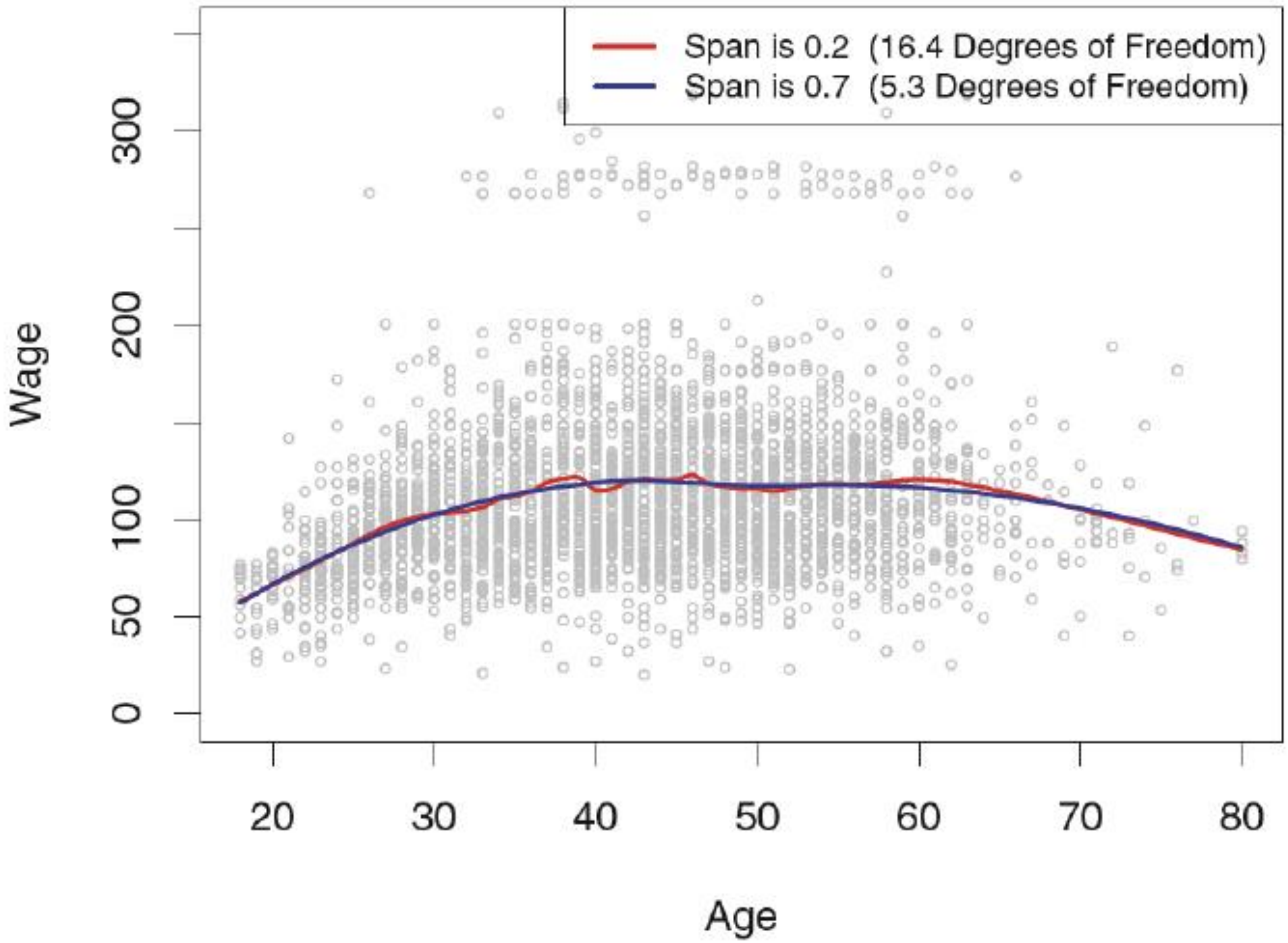
Local Linear Regression



Some Remarks

- Choice of weighting function K.
- Choice of local constant, linear or quadratic regression functions
- Choice of **span** s
  - It controls the flexibility of the non-linear fit.
  - Smaller s leads to more local and wiggly fit, whereas a very large s leads to a global fit by using all of the training observations.
  - Can be determined by cross validation.
- The idea can generalize to the **varying coefficient model**, which is a global model in some variables and local in others.

Example with Local Regression



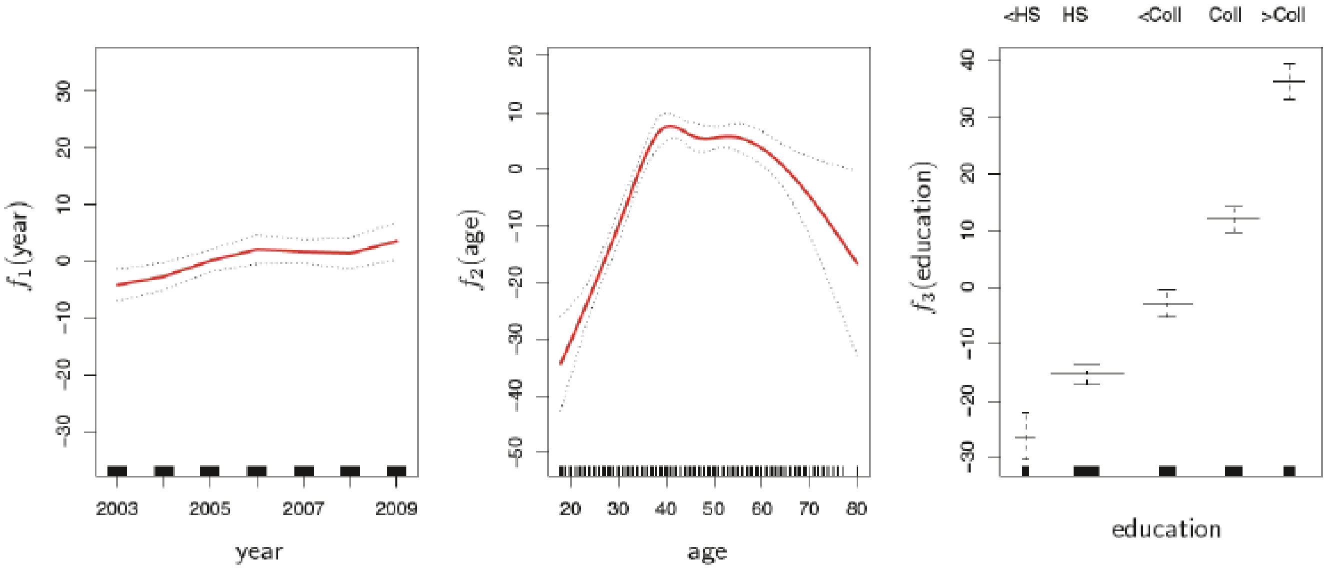
Generalized Additive Model (GAMs)

- **Generalized additive model** provides a general framework for modeling nonlinear function with multiple variables.
- It assumes that

$$y_i = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

where each f_j is a nonlinear function for X_j.

- f_i can be estimated by any nonlinear model.



Pros and Cons

- It fits nonlinear f_j to each X_j, so as to automatically model non-linear relationships to multiple variables.

- The model is additive in nature, so we examine the effect of each individual X_j on Y while holding all of the other variables fixed.

- The additive form can also be restrictive as it rules out possible interaction terms.

- If interactions are needed, one may consider to include $f_{jk}(X_j, X_k)$ in the additive model.