Chapter 7: Moving Beyond Linearity

Yonghyun Kwon

Department of Mathematics, Korea Military Academy

Moving Beyond Linearity

The truth is never linear!

Or almost never!

But often the linearity assumption is good enough.

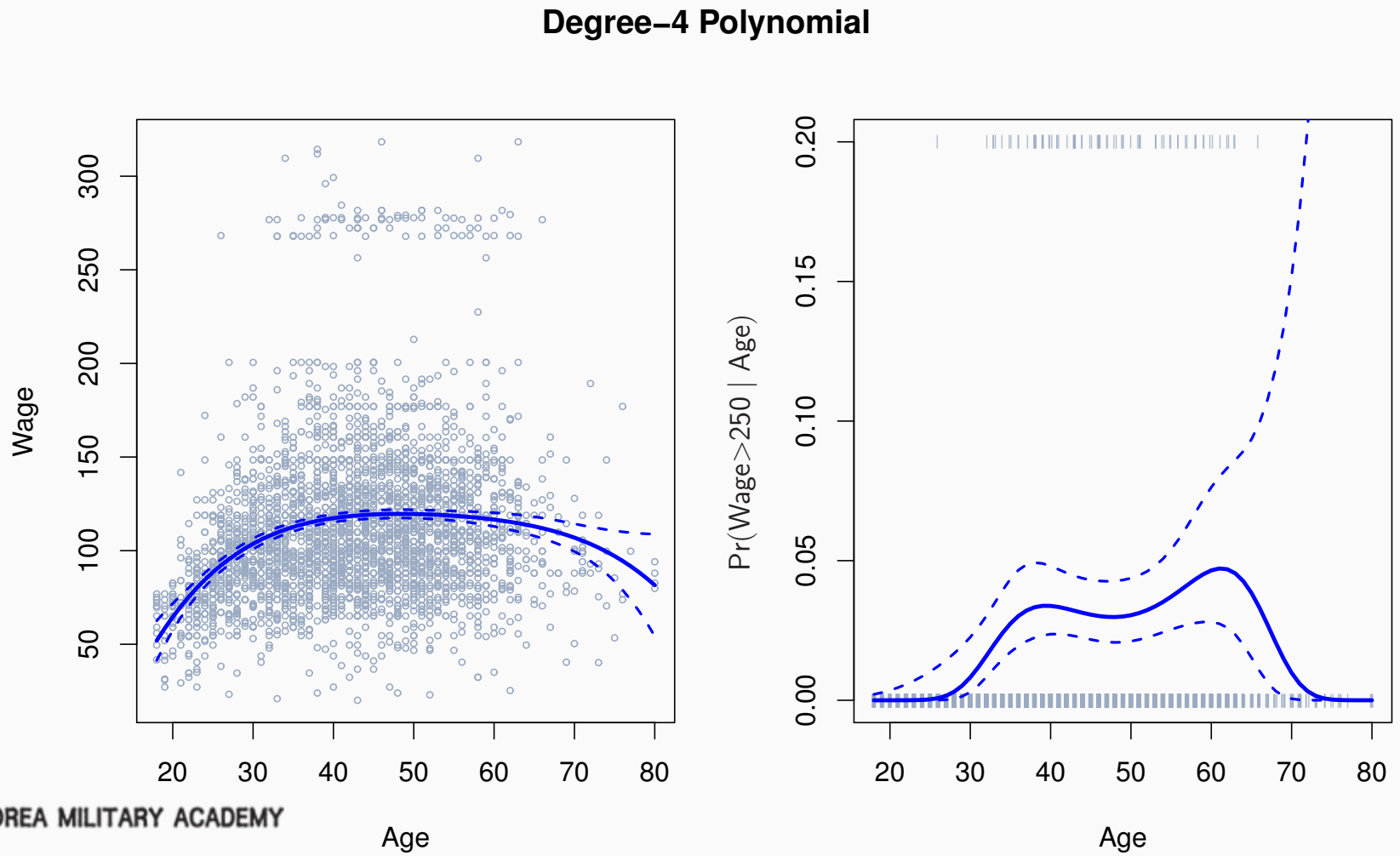
When it's not ...

- *polynomials*,
- *step functions*,
- *splines*,
- *local regression*, and
- *generalized additive models (GAMs)*

offer a lot of flexibility, without losing the ease and interpretability of linear models.

Polynomial Regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \varepsilon_i$$



Polynomial Regression: Details

- Create new variables $X_1 = X$, $X_2 = X^2$, ... and then treat as multiple linear regression.
- Not really interested in the coefficients; more interested in the *fitted function values* at any value x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4.$$

- Since $\hat{f}(x_0)$ is a linear function of the $\hat{\beta}_\ell$, we can obtain a simple expression for the *pointwise variance* $\mathrm{Var}[\hat{f}(x_0)]$ at any x_0.
  We display $\hat{f}(x_0) \pm 2 \cdot \mathrm{se}[\hat{f}(x_0)]$.
- We either fix the degree d at some reasonably low value, or use *cross-validation* to choose d.

Polynomial Regression: Details (cont.)

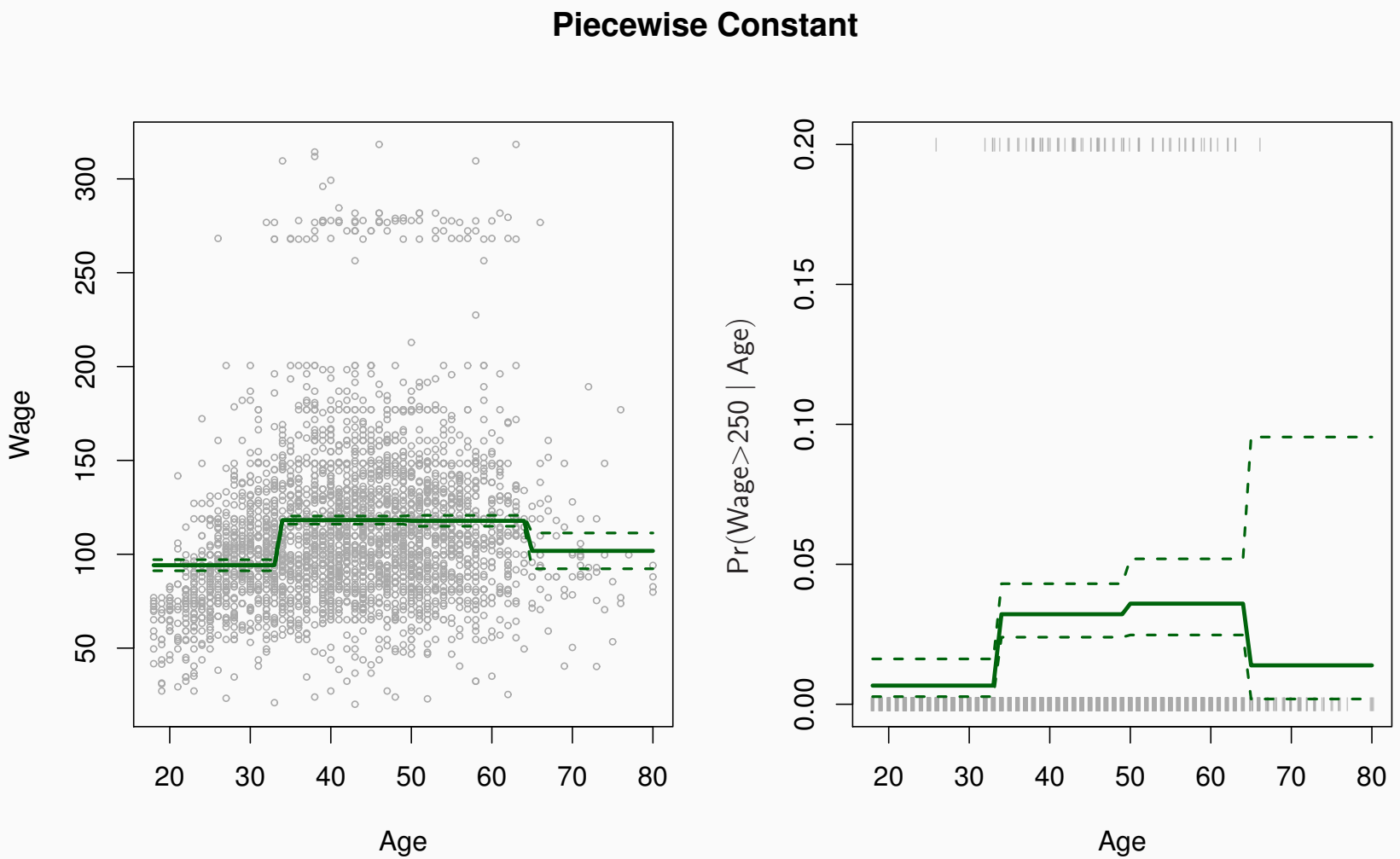
- *Logistic regression* follows naturally. For example, we model

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

- To get confidence intervals, compute upper and lower bounds on the *logit scale*, then invert to the probability scale.
- Can apply on several variables — stack the variables into one matrix, and separate out the pieces afterwards (see GAMs later).
- *Caveat:* polynomials have notorious *tail behavior* — very bad for extrapolation.
- In `R`: `y ~ poly(x, degree = 3)`

Step Functions

Another way of creating transformations of a variable: *cut* the variable into distinct regions.



Step Functions (cont.)

- Easy to work with. Creates a series of *dummy variables* representing each group.
- Useful way of creating interactions that are easy to interpret. For example, the interaction effect of `Year` and `Age`:

$$I(\texttt{Year} < 2005) \cdot \texttt{Age}, \qquad I(\texttt{Year} \geq 2005) \cdot \texttt{Age}$$

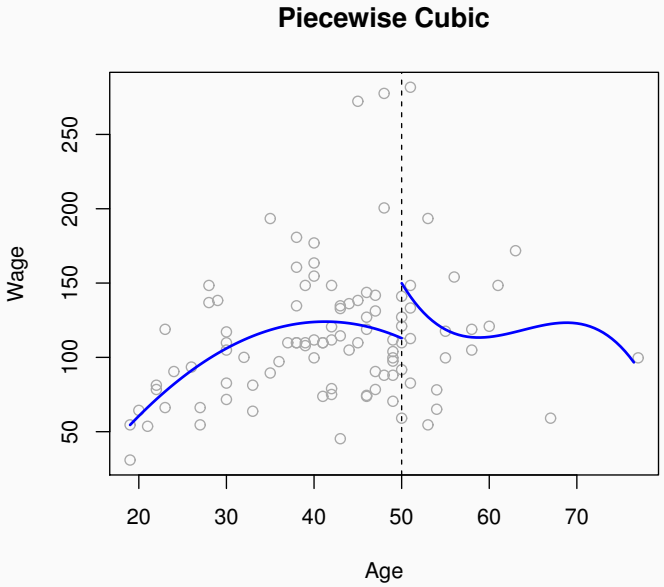
  allows for different linear functions in each time period.
- In `R`: `I(year < 2005)` or `cut(age, c(18, 25, 40, 65, 90))`.
- Choice of *cutpoints (knots)* can be problematic. For creating nonlinearities, smoother alternatives such as *splines* are available.

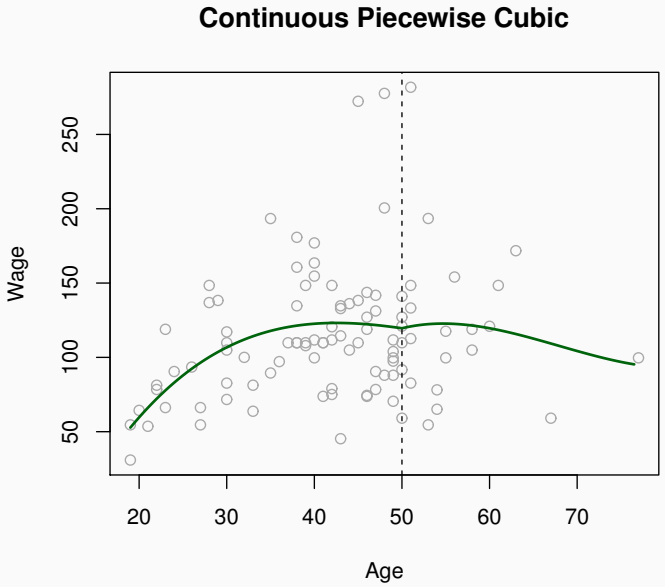
Piecewise Polynomials

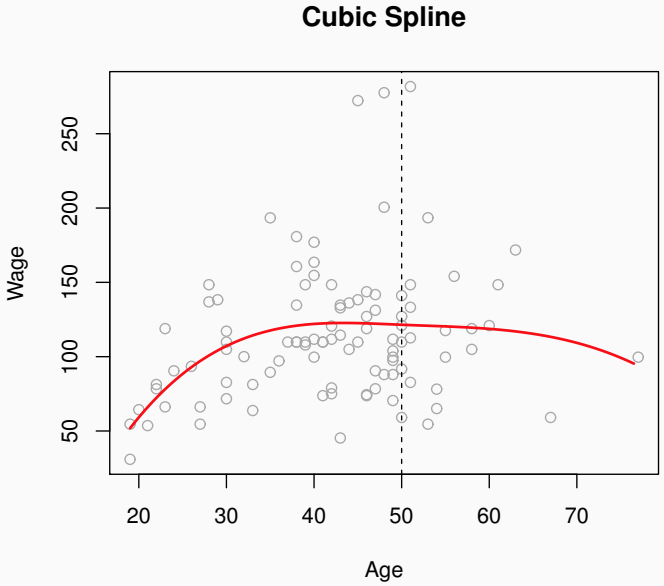
- Instead of a single polynomial in X over its whole domain, use *different polynomials* in regions defined by knots. E.g.:

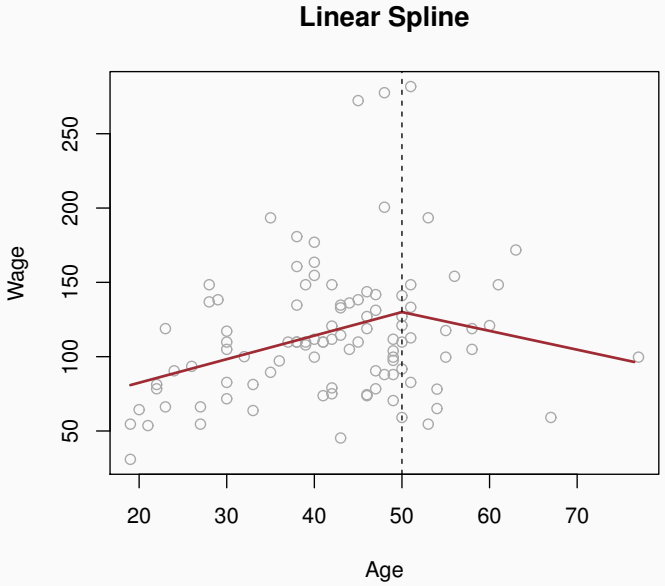
$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i & \text{if } x_i < c, \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i & \text{if } x_i \geq c. \end{cases}$$

- Better to add *constraints* (e.g. continuity, smoothness) at the knots.
- *Splines* impose the maximum amount of continuity.


Piecewise Cubic
Wage
Age
50
100
150
200
250
20
30
40
50
60
70


Continuous Piecewise Cubic
Wage
Age
50
100
150
200
250
20
30
40
50
60
70


Cubic Spline
Wage
Age
50
100
150
200
250
20
30
40
50
60
70


Linear Spline
Wage
Age
50
100
150
200
250
20
30
40
50
60
70

Linear Splines

Definition

A *linear spline* with knots at ξ_k, $k = 1, \ldots, K$, is a piecewise linear polynomial continuous at each knot.

We represent this model as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \varepsilon_i,$$

where the b_k are *basis functions*:

$$b_1(x_i) = x_i, \qquad b_{k+1}(x_i) = (x_i - \xi_k)_+, \quad k = 1, \ldots, K.$$

Here the $(\cdot)_+$ means *positive part*:

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

Cubic Splines

Definition

A *cubic spline* with knots at ξ_k, $k = 1, \ldots, K$, is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot.

Using the *truncated power basis*:

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

$$b_1(x_i) = x_i, \quad b_2(x_i) = x_i^2, \quad b_3(x_i) = x_i^3, \quad b_{k+3}(x_i) = (x_i - \xi_k)_+^3$$

where

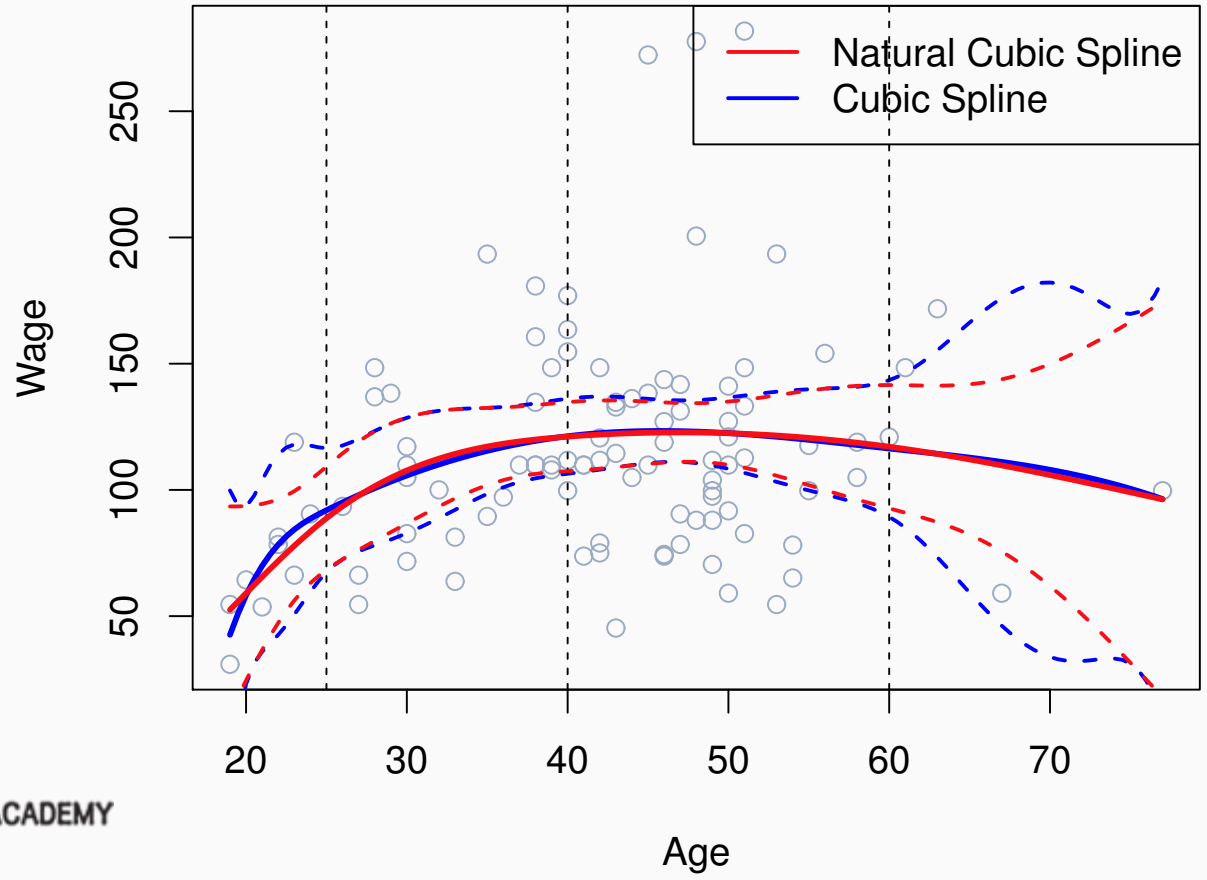
$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k, \\ 0 & \text{otherwise.} \end{cases}$$

Note: A cubic spline with K knots has $K + 4$ parameters (degrees of freedom).

Natural Cubic Splines

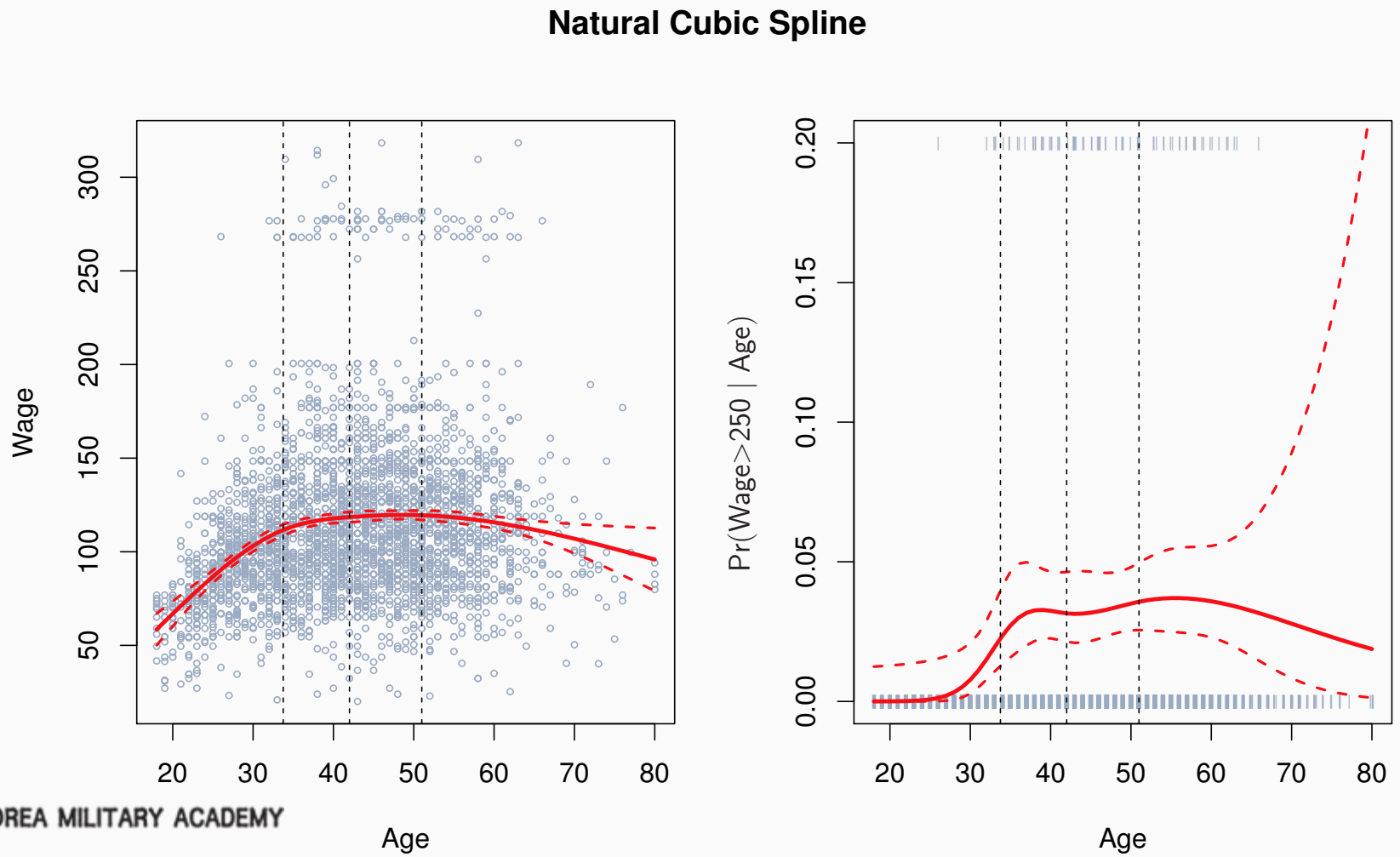
Definition

A *natural cubic spline* extrapolates linearly beyond the boundary knots. This adds $4 = 2 \times 2$ extra constraints, allowing more internal knots for the same degrees of freedom.



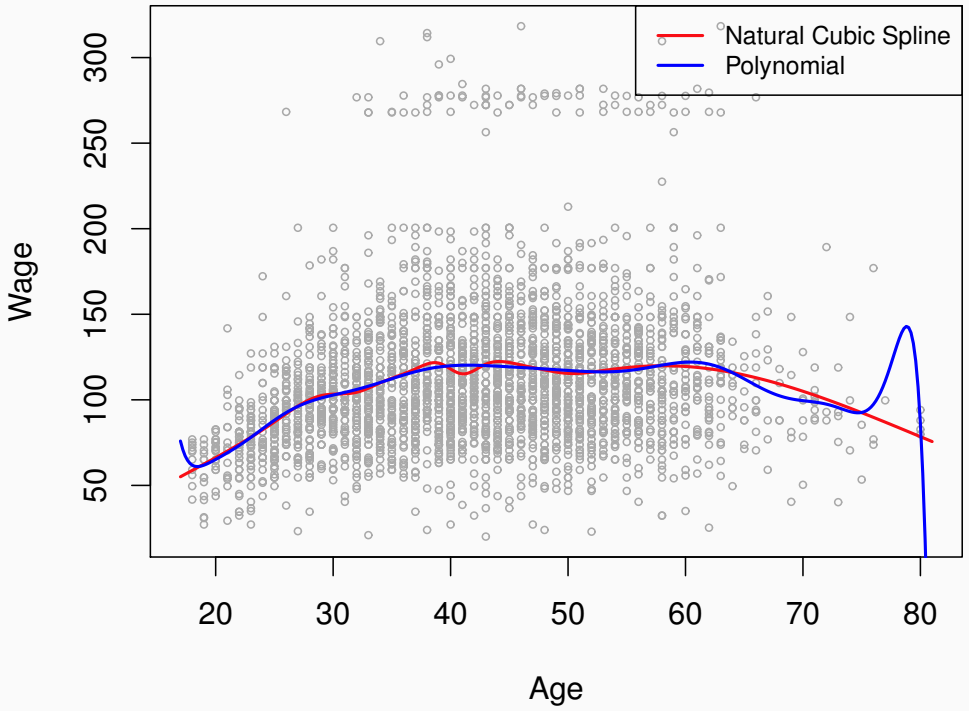
Fitting Splines in R

- `bs(x, ...)` in package `splines`: fits splines of any degree.
- `ns(x, ...)` in package `splines`: fits natural cubic splines.



Knot Placement

- One strategy: decide K (number of knots), then place them at appropriate *quantiles* of the observed X.
- A cubic spline with K knots has $K + 4$ parameters (d.f.).
- A natural spline with K knots has K degrees of freedom.



Comparison of a degree-14 polynomial and a natural cubic spline, each with 15 d.f.

```
ns(age, df=14)
poly(age, deg=14)
```

Smoothing Splines

A bit mathematical . . .

Consider fitting a smooth function $g(x)$ by minimizing:

$$\min_{g \in \mathcal{S}} \sum_{i=1}^{n} \left(y_i - g(x_i)\right)^2 + \lambda \int g''(t)^2 \, dt$$

- The first term is *RSS* — makes $g(x)$ match the data.
- The second term is the *roughness penalty* controlling wiggliness, modulated by $\lambda \geq 0$:
  - Smaller $\lambda \Rightarrow$ more wiggly (interpolates y_i when $\lambda = 0$).
  - $\lambda \to \infty \Rightarrow g(x)$ becomes linear.
- The solution is a *natural cubic spline* with a knot at every unique x_i.

Smoothing Splines: Details

- Smoothing splines avoid the knot-selection issue — only a single λ needs to be chosen.
- In R: `smooth.spline()` fits a smoothing spline.
- The vector of n fitted values is $\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$, where $\mathbf{S}_\lambda$ is an $n \times n$ smoother matrix.
- The *effective degrees of freedom* are:

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

Smoothing Splines: Choosing λ

- Specify *df* rather than λ directly:

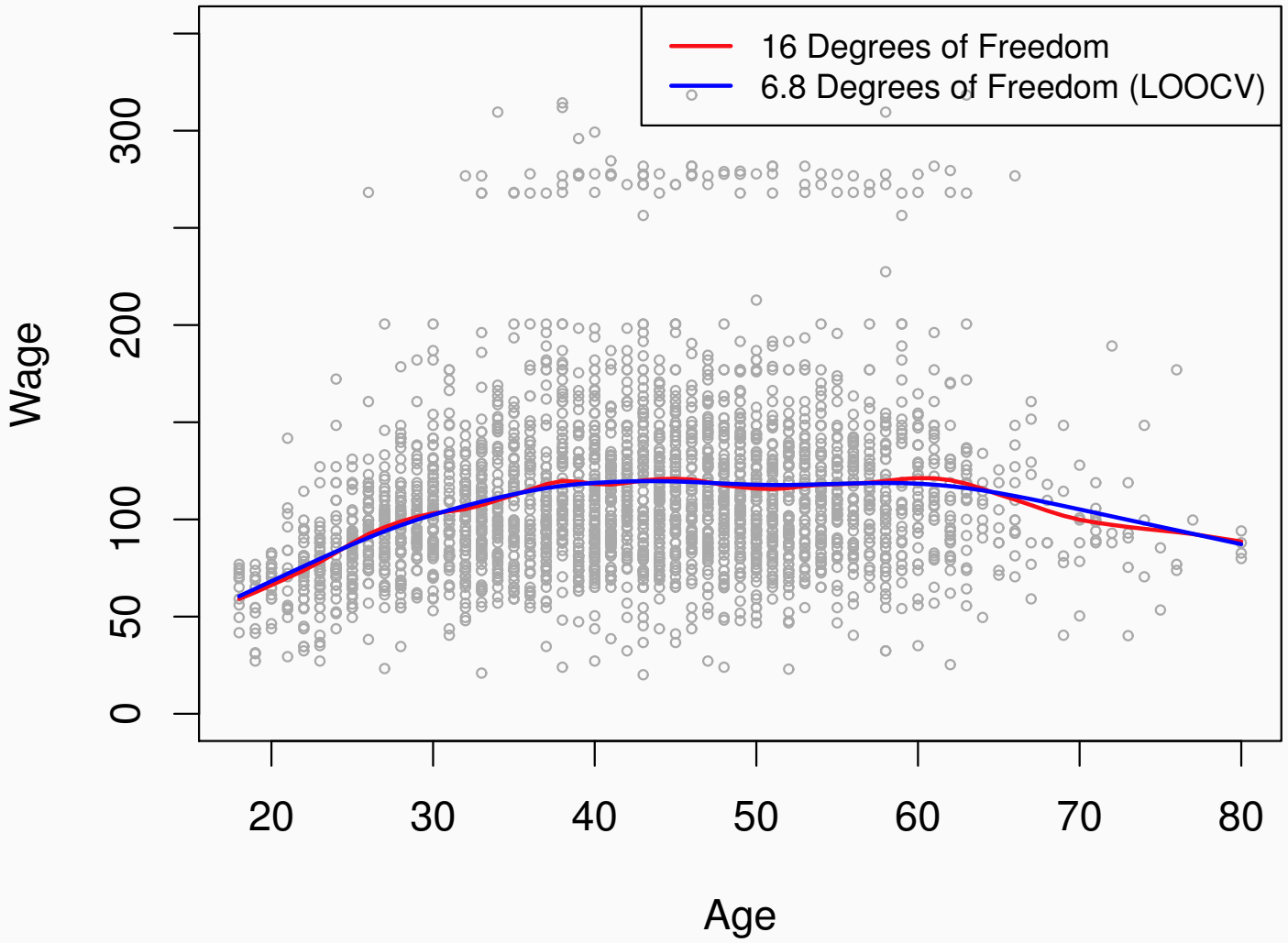
  `smooth.spline(age, wage, df = 10)`

- The *leave-one-out cross-validated* (LOOCV) error has a closed form:

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

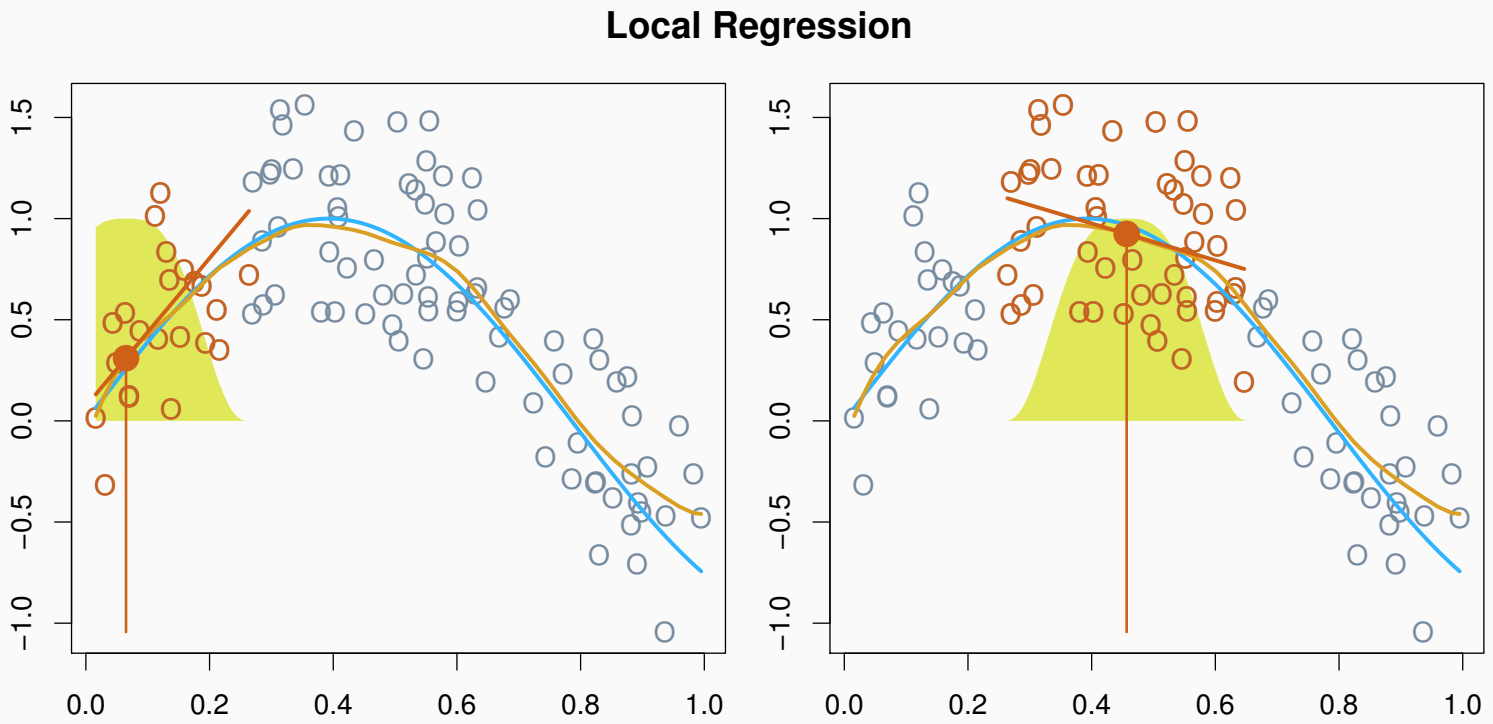
  In R: `smooth.spline(age, wage)` (uses LOOCV automatically).

Smoothing Spline



Local Regression

- With a sliding weight function, fit *separate linear fits* over the range of X by weighted least squares.



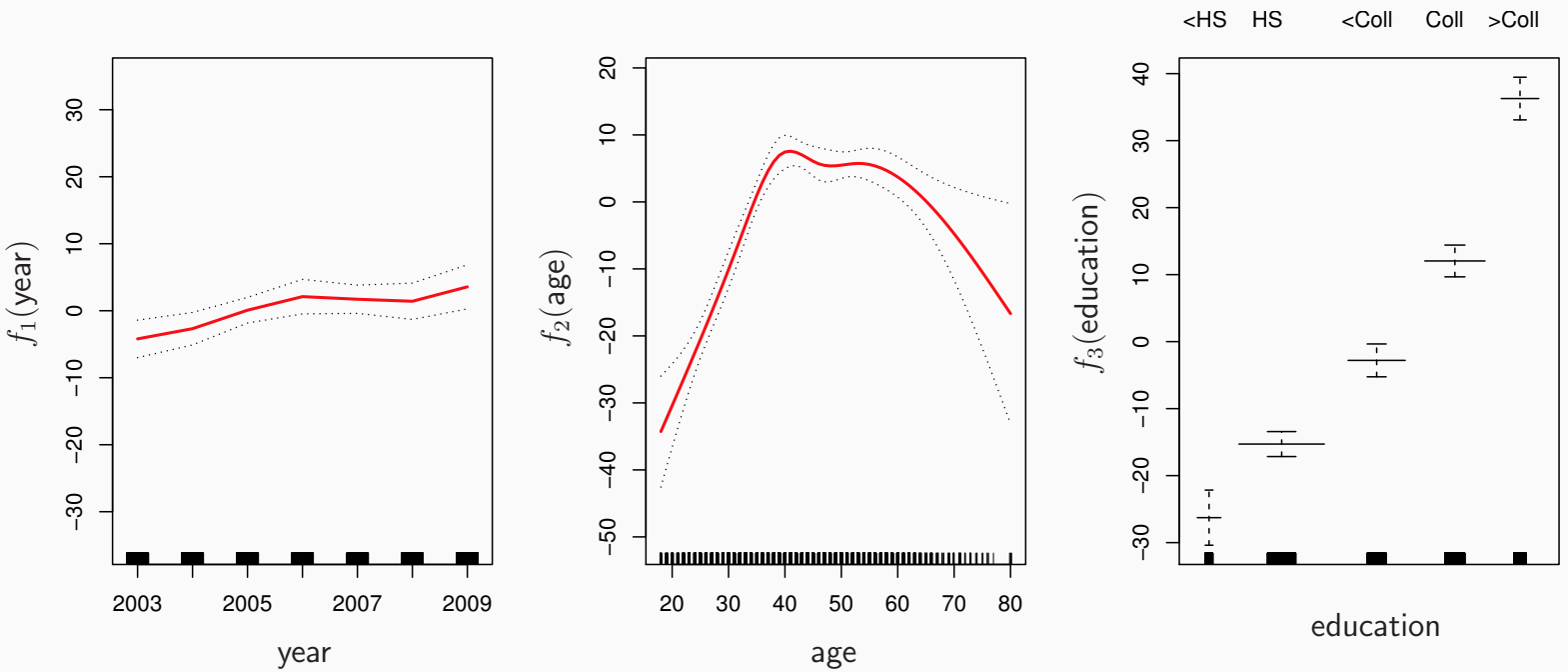
Note: In R: `loess()` function. The span s controls the degree of smoothness.

Generalized Additive Models (GAMs)

GAMs allow flexible nonlinearities in several variables, while retaining the *additive structure* of linear models:

GAM

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i.$$



GAM Details

- Fit a GAM with natural splines using `lm`:

```
lm(wage ~ ns(year, df=5) + ns(age, df=5) +
                                education)
```

- Coefficients are not that interesting; *fitted functions* $\hat{f}_j$ are. Visualized with `plot.gam`.
- Can *mix* terms — some linear, some nonlinear — and use `anova()` to compare models.
- Can also use smoothing splines or local regression:

```
gam(wage ~ s(year, df=5) + lo(age, span=.5) +
                                education)
```

- GAMs are *additive*, although low-order interactions can be included naturally using bivariate smoothers or terms like `ns(age, df=5):ns(year, df=5)`.

GAMs for Classification

For a binary response, use the *logistic GAM*:

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

```
gam(I(wage > 250) ~ year + s(age, df=5) + education,
                    family = binomial)
```

Pros and Cons of GAMs

Pros

- Flexible nonlinear fits in each X_j
- Additive structure aids interpretation
- Inference is possible for each f_j

Cons

- Additivity may miss interaction effects
- More complex models need specialized software

Exercises

1. Consider the piecewise linear regression with a single knot at ξ:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \xi)_+ + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ independently.

A. Write out $\mathrm{E}(y_i \mid x_i)$ for $x_i < \xi$ and for $x_i \geq \xi$. Interpret β_1 and $\beta_1 + \beta_2$.

B. Explain why this model is continuous at $x = \xi$.

C. Suppose $\hat{\beta}_0 = 2$, $\hat{\beta}_1 = 1.5$, $\hat{\beta}_2 = -0.8$, and $\xi = 5$. Sketch the fitted piecewise linear function.

D. Propose a test statistic for $H_0 : \beta_2 = 0$ using the usual t-test output from `lm`. What does H_0 say about the regression function?

Exercises (cont.)

2. Let $g(x)$ be the natural cubic spline minimizing

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

A. Explain in words the role of λ. What happens as $\lambda \to 0$ and as $\lambda \to \infty$?

B. The fitted values satisfy $\hat{\mathbf{g}} = \mathbf{S}_\lambda \mathbf{y}$. The effective degrees of freedom are $df_\lambda = \mathrm{tr}(\mathbf{S}_\lambda)$. Explain why $df_\lambda \in [2, n]$.

C. The LOOCV error is

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

Compare the computational cost of this formula to naive LOO cross-validation (re-fitting n times).