Moving Beyond Linearity

David J. Hessen

Utrecht University

March 7, 2019

Statistical learning

Supervised learning

- a single response y
- multiple predictors $x_1, \ldots, x_p$

Multiple regression (interval response): $y = f(x_1, \ldots, x_p) + \varepsilon$

Binary logistic regression: $\pi = \dfrac{\exp\{f(x_1, \ldots, x_p)\}}{1 + \exp\{f(x_1, \ldots, x_p)\}}$

The assumption of linearity: $f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

Statistical learning

Why this inflexible (very restrictive) approach?

- If linearity is true, then there is no bias and no more flexible method competes $\rightarrow$ the variance of the estimator of $f(x_1, \ldots, x_p)$ will be smaller
- Often the linearity assumption is good enough
- Very interpretable

Statistical learning

What can be done when linearity is not good enough?

1. Polynomial regression
2. Piecewise polynomials
3. Regression splines
4. Smoothing splines
5. Local regression
6. Generalized additive models

Modeling approaches 1 to 5 are presented for the relationship between response y and a single predictor x

Polynomial regression

$$
\begin{aligned}
&\text{linear function} && : f(x) = \beta_0 + \beta_1 x \\
&\text{quadratic function} && : f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \\
&\text{cubic function} && : f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\
&\qquad \vdots && \qquad\qquad \vdots \\
&\text{degree-}d\text{ polynomial:} && f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d
\end{aligned}
$$

It's just the standard linear model

$$f(x_1, \ldots, x_d) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_d x_d$$

where

$$x_1 = x,\ x_2 = x^2,\ \ldots,\ x_d = x^d$$

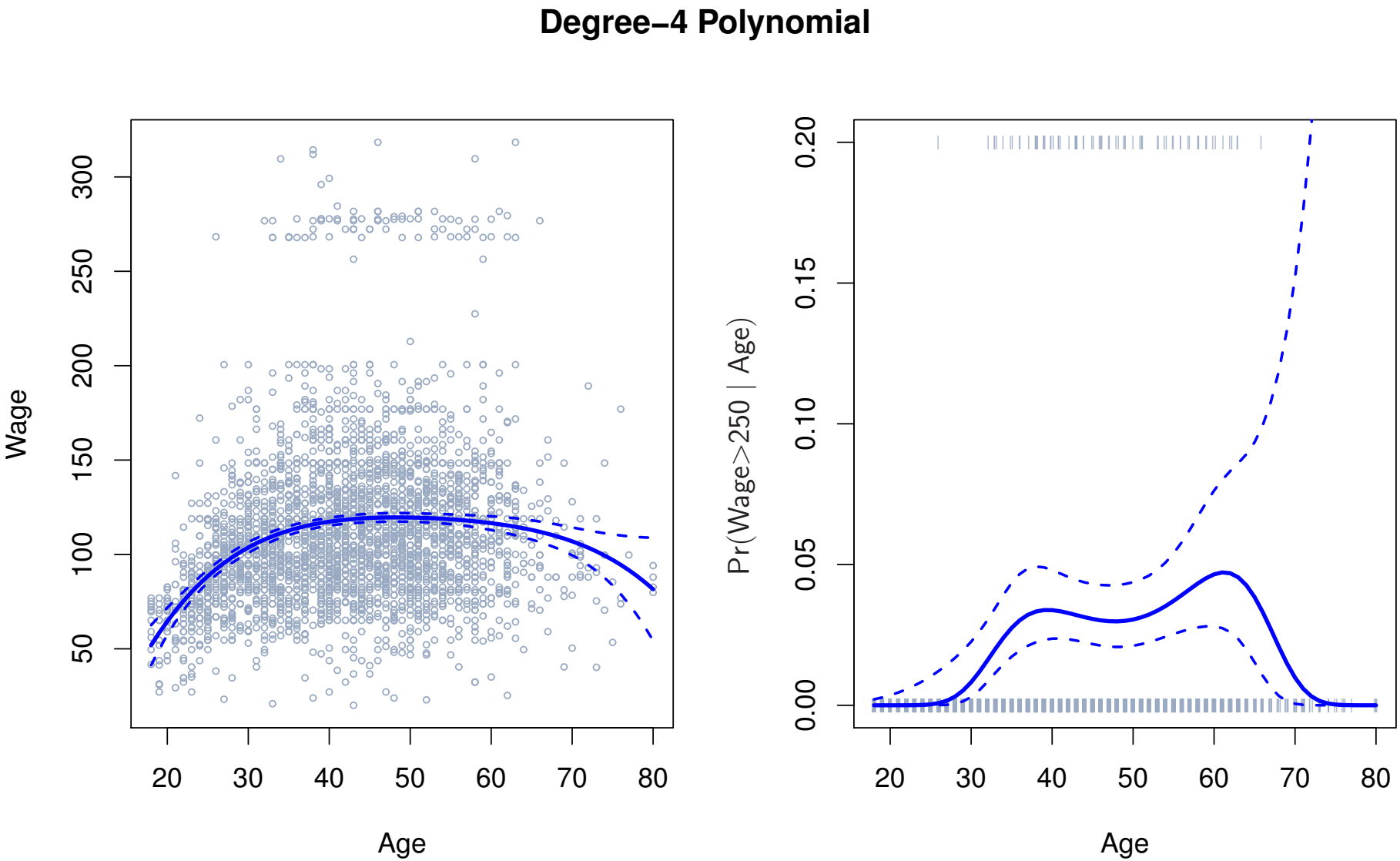
Polynomial regression

- The coefficients $\beta_0, \beta_1, \ldots, \beta_d$ can be easily estimated using least squares
- The interest is more in the fitted value

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

  than in the coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d$
- Usually, either d is fixed to 3 or 4, or cross-validation is used to choose d
- Especially near the boundary of x the polynomial curve can become overly flexible (bad for extrapolation)

Polynomial regression



Polynomial regression

- In the left-hand panel of the figure, the solid blue curve is given by

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

and the pair of dotted blue curves indicate an estimated 95% confidence interval given by

$$\hat{f}(x) \pm 2 \cdot se\{\hat{f}(x)\}$$

- In the right-hand panel, the solid blue curve is given by

$$\hat{\pi}(y > 250 \,|\, x) = \exp\{\hat{f}(x)\}/[1 + \exp\{\hat{f}(x)\}] = \text{sigm}\{\hat{f}(x)\}$$

and the pair of dotted blue curves indicate an estimated 95% confidence interval given by

$$\text{sigm}[\hat{f}(x) \pm 2 \cdot se\{\hat{f}(x)\}]$$

Piecewise polynomials

A step function (a piecewise polynomial of order zero) can be used to avoid imposing a global structure

Cutpoints or knots $c_1, c_2, \ldots, c_K$ in the range of x are chosen and are used to create $K + 1$ dummy variables

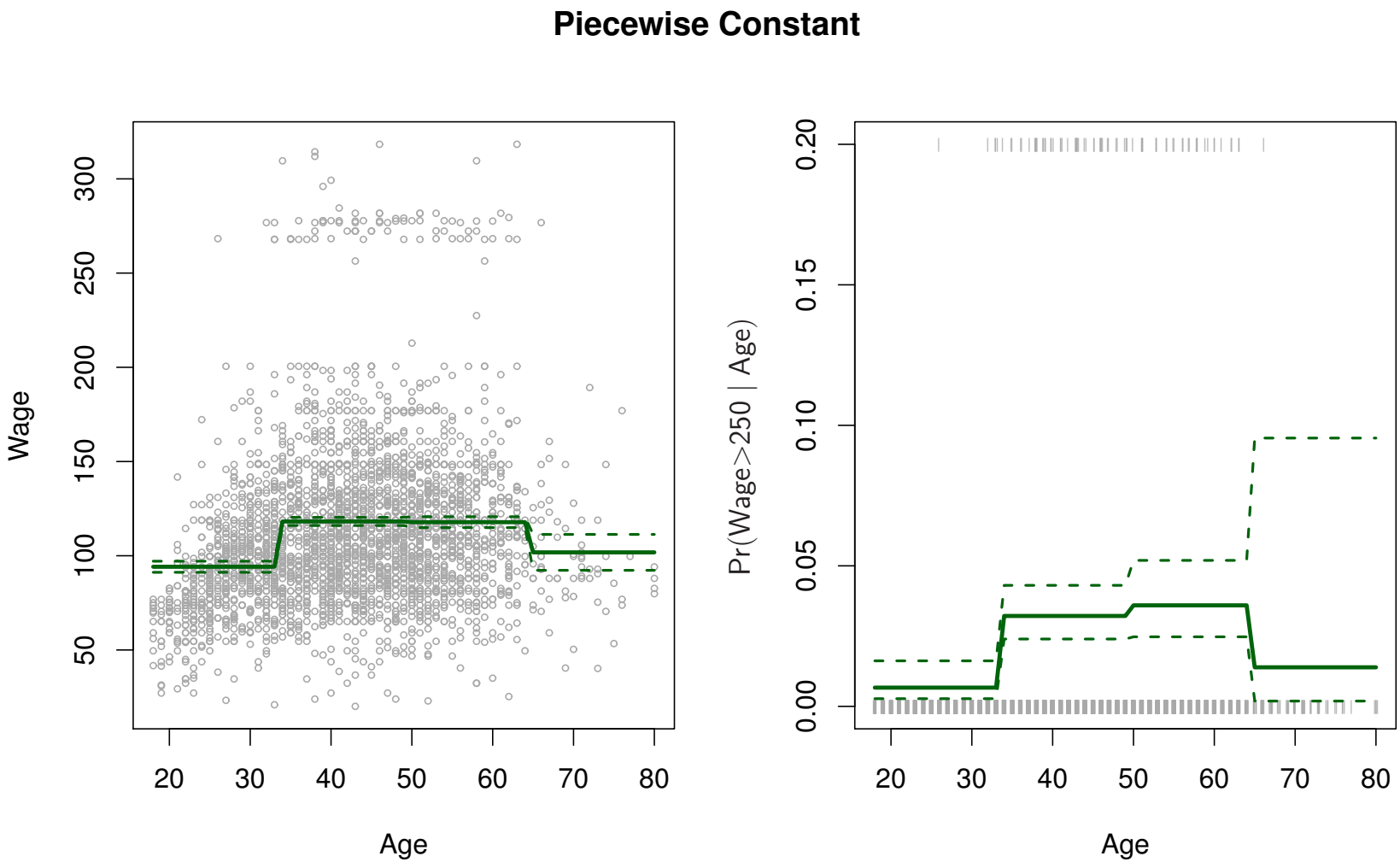
$$\begin{aligned} C_0(x) &= I(x < c_1) \\ C_1(x) &= I(c_1 \leq x < c_2) \\ &\vdots \qquad\qquad \vdots \\ C_{K-1}(x) &= I(c_{K-1} \leq x < c_K) \\ C_K(x) &= I(x \geq c_K) \end{aligned}$$

Least squares is used to fit

$$f(x) = \beta_0 + \beta_1 C_1(x) + \beta_2 C_2(x) + \ldots + \beta_K C_K(x)$$

Piecewise polynomials

Knots: $c_1 = 35$, $c_2 = 50$, $c_3 = 65$



Piecewise polynomials

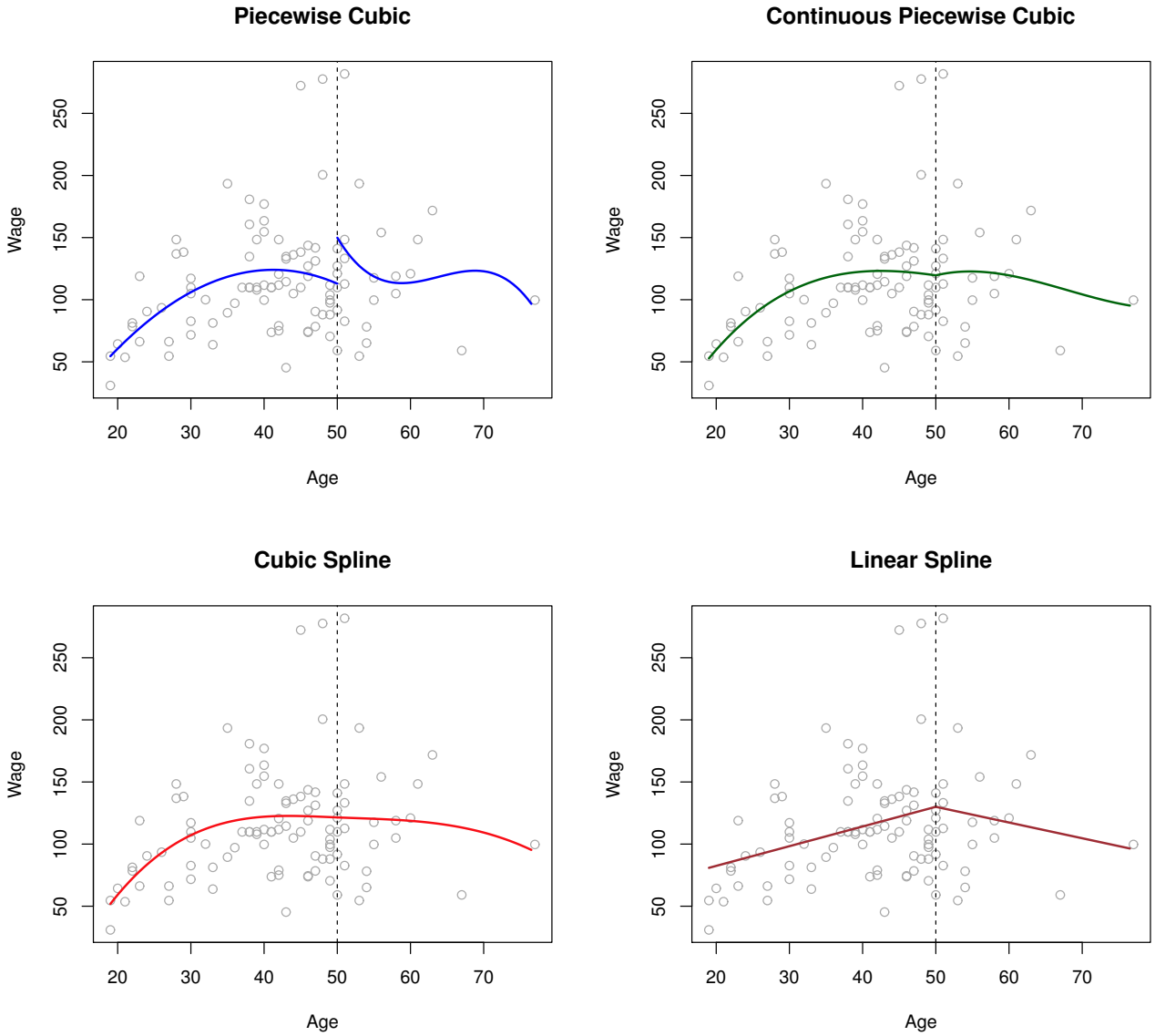
- Step functions are easy to work with
- Interactions can easily be created and are easy to interpret
- The choice of the knots can be problematic
- Piecewise constant functions can miss the action

A piecewise polynomial of degree-d

$$f(x) = \begin{cases} \beta_{00} + \beta_{10}x + \beta_{20}x^2 + \ldots + \beta_{d0}x^d & \text{if } x < c_1 \\ \beta_{01} + \beta_{11}x + \beta_{21}x^2 + \ldots + \beta_{d1}x^d & \text{if } c_1 \leq x < c_2 \\ \vdots & \vdots \\ \beta_{0K} + \beta_{1K}x + \beta_{2K}x^2 + \ldots + \beta_{dK}x^d & \text{if } x \geq c_K \end{cases}$$

Usually, separate low-degree polynomials are fitted over different regions of x

Piecewise polynomials



Regression splines

A degree-d spline is a piecewise degree-d polynomial with continuity in derivatives up to degree $d-1$ at each knot (d constraints per knot)

A great reduction in complexity compared to a piecewise polynomial

The number of parameters of a degree-d piecewise polynomial is

$$(K+1)(d+1) = Kd + K + d + 1 \quad (\#\text{ intervals} \times \#\text{ coefficients})$$

The number of constraints is Kd (# knots $\times$ # constraints per knot)

The number of parameters of a degree-d spline is therefore $K + d + 1$

Consequently, a smaller $var\{\hat{f}(x)\}$ for a degree-d spline

Regression splines

A degree-d spline can be represented by the linear model

$$f(x) = \beta_0 + \beta_1 b_1(x) + \ldots + \beta_d b_d(x) + \beta_{d+1} b_{d+1}(x) + \ldots + \beta_{d+K} b_{d+K}(x)$$

where

$$\begin{aligned} b_1(x) &= x \\ b_2(x) &= x^2 \\ &\vdots \quad\quad \vdots \\ b_d(x) &= x^d \\ b_{d+k}(x) &= \left\{ \begin{array}{ll} (x - c_k)^d & \text{if } x > c_k \\ 0 & \text{otherwise} \end{array} \right\} \text{ for } k = 1, \ldots, K \end{aligned}$$

are basis functions

Regression splines

A linear spline

$$f(x) = \beta_0 + \beta_1 b_1(x) + \beta_2 b_2(x) + \ldots + \beta_{1+K} b_{1+K}(x)$$

where

$$b_1(x) = x$$

$$b_{1+k}(x) = \left\{ \begin{array}{ll} x - c_k & \text{if } x > c_k \\ 0 & \text{otherwise} \end{array} \right\} \text{ for } k = 1, \ldots, K$$

Regression splines

A quadratic spline

$$f(x) = \beta_0 + \beta_1 b_1(x) + \beta_2 b_2(x) + \beta_3 b_3(x) + \ldots + \beta_{2+K} b_{2+K}(x)$$

where

$$\begin{aligned} b_1(x) &= x \\ b_2(x) &= x^2 \\ b_{2+k}(x) &= \left\{ \begin{array}{ll} (x - c_k)^2 & \text{if } x > c_k \\ 0 & \text{otherwise} \end{array} \right\} \text{ for } k = 1, \ldots, K \end{aligned}$$

Regression splines

A cubic spline

$$f(x) = \beta_0 + \beta_1 b_1(x) + \beta_2 b_2(x) + \beta_3 b_3(x) + \beta_4 b_4(x) + \ldots + \beta_{3+K} b_{3+K}(x)$$

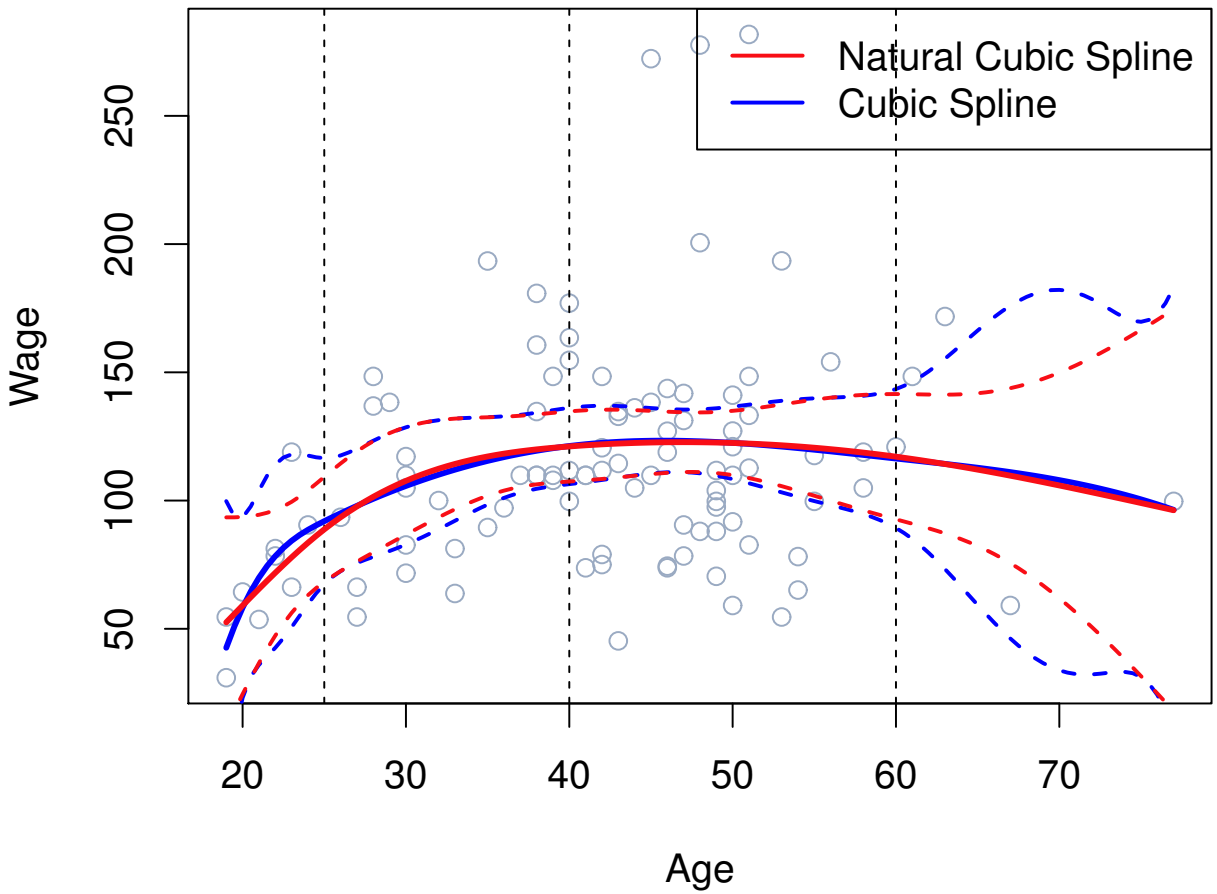
where

$$\begin{aligned} b_1(x) &= x \\ b_2(x) &= x^2 \\ b_3(x) &= x^3 \\ b_{3+k}(x) &= \left\{ \begin{array}{ll} (x - c_k)^3 & \text{if } x > c_k \\ 0 & \text{otherwise} \end{array} \right\} \text{ for } k = 1, \ldots, K \end{aligned}$$

Regression splines

Unfortunately, a spline can have high variance at the outer range of x

A natural spline is forced to be linear in the extreme regions of x



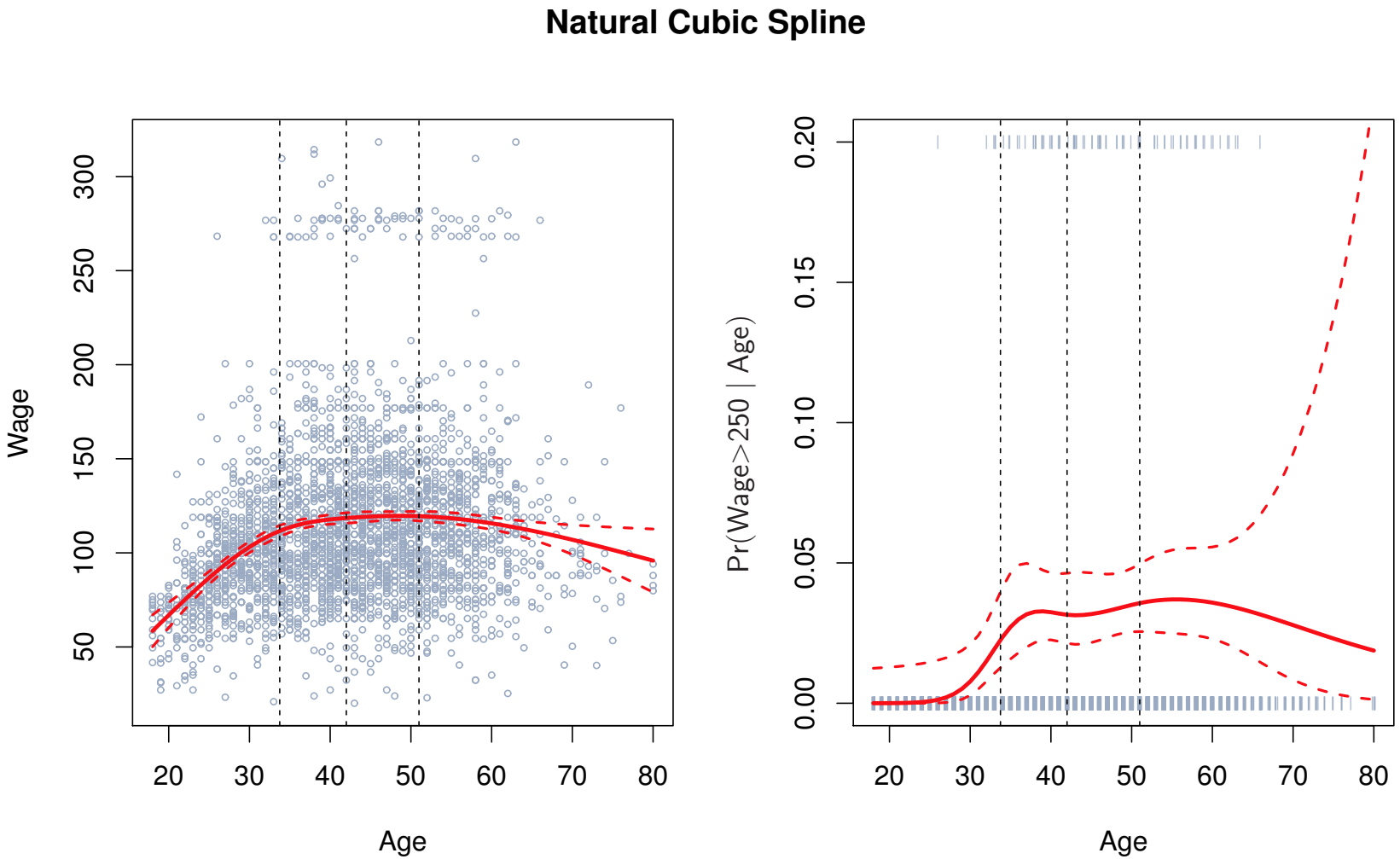
Regression splines

Where should the K knots be placed?

- More knots in regions where $f(x)$ is expected to change more rapidly and fewer knots in regions where $f(x)$ is expected to be more stable
- At uniform quantiles of the data

Regression splines

Knot placement at the 25th, 50th, and 75th percentiles of x



Regression splines

How many knots should be used?

- Try out which number produces the best looking curve
- Cross-validation for different numbers of knots $\rightarrow$ the value of K that gives the smallest overall cross-validated RSS is chosen

Smoothing splines

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the observed sample data

A smoothing spline is a function $f(x)$ that minimizes

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx$$

for a fixed nonnegative tuning or smoothing parameter λ

The first term is the RSS and tries to make $f(x)$ match the data

The second term is a roughness penalty and controls how wiggly $f(x)$ is

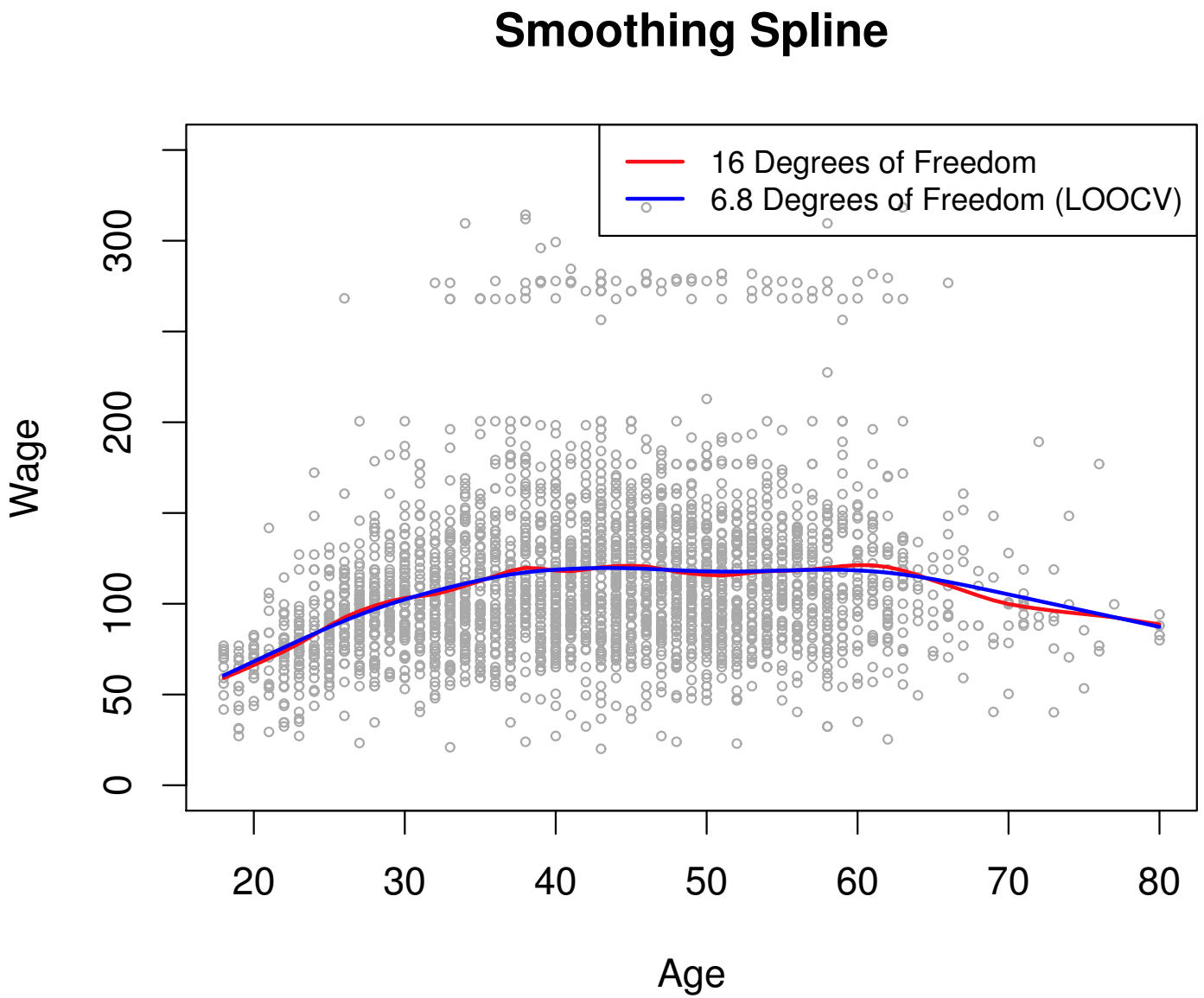
- The smaller λ, the more wiggly $f(x)$, eventually interpolating y_i when $\lambda = 0$
- As $\lambda \to \infty$, the function $f(x)$ becomes linear

Smoothing splines

Remarkly, a smoothing spline is a natural cubic spline with knots at the unique values of $x_1, \ldots, x_n$

- Smoothing splines avoid the knot-selection issue, leaving a single λ to be chosen
- Instead of λ, the effective degrees of freedom (the number of parameters minus the number of constraints) can be specified (as λ increases from 0 to ∞, df_λ decreases from n to 2)
- The value of either λ or df_λ that minimizes the cross-validated RSS can be chosen

Smoothing splines



Local regression

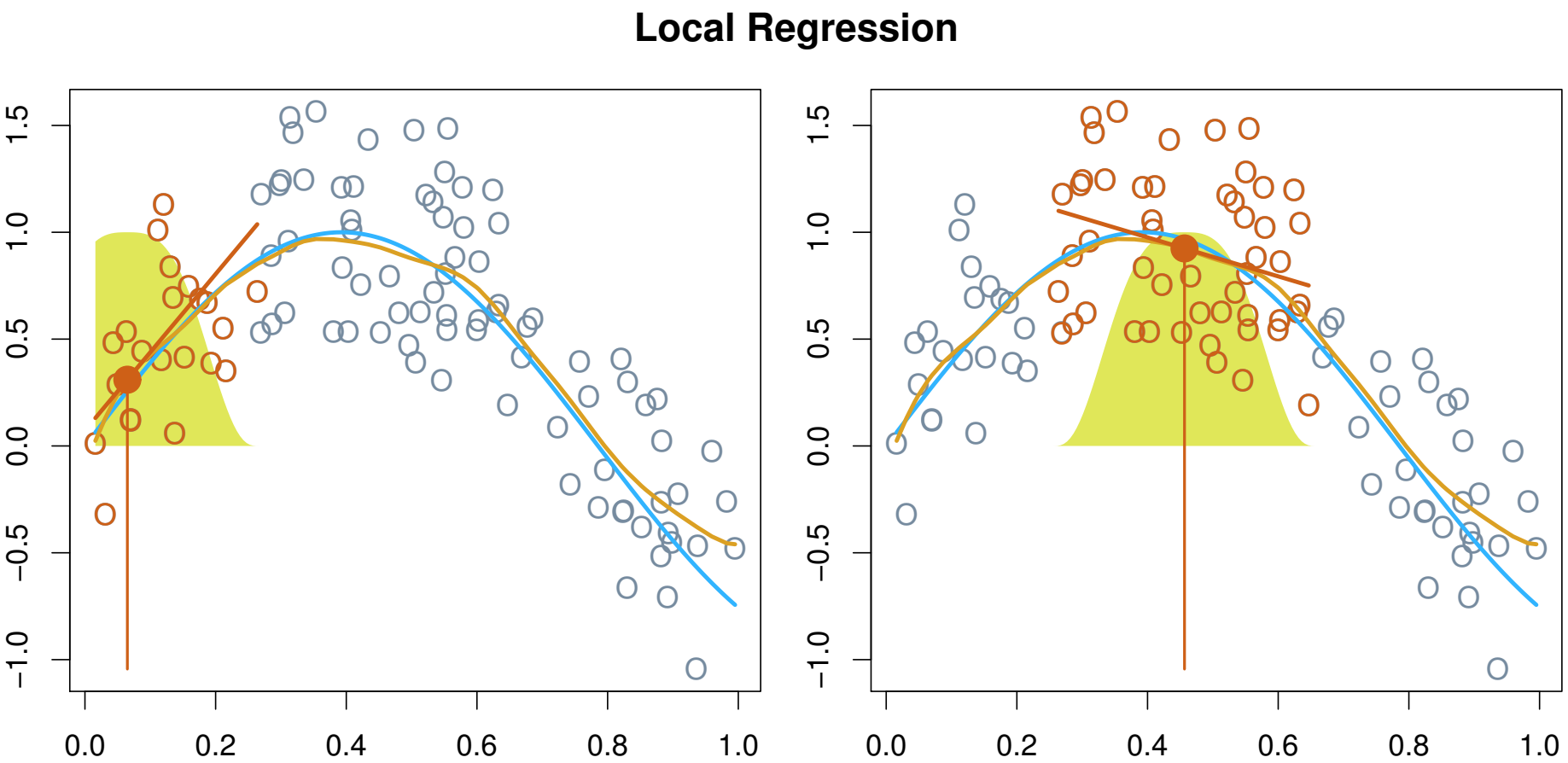
Algorithm for local regression at x_0

1. Gather the fraction $s = k/n$ of cases whose x-values are closest to x_0
2. Assign a weight w_i to each of the k cases
3. Minimize the weighted sum of squares

$$\sum_{i=1}^{k} w_i \{y_i - f(x_i)\}^2$$

   with respect to the parameters of $f(x)$
4. The fitted value at x_0 is given by $\hat{f}(x_0)$

Local regression



Local regression

Choices to be made

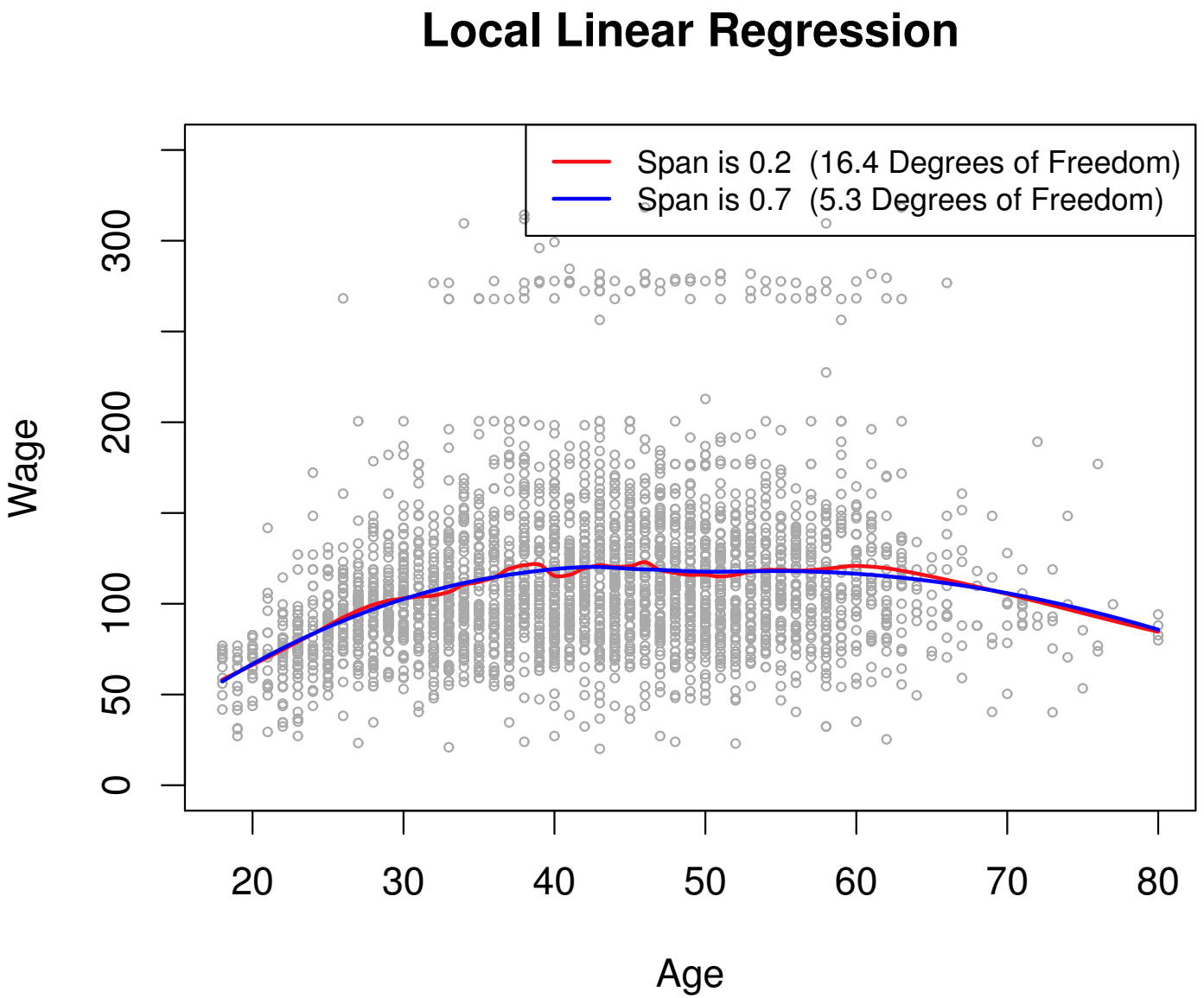
- The weights $w_1, \ldots, w_k$
- The form of $f(x)$
- The span s

The span s plays a role like λ in a smoothing spline

The smaller s, the more wiggly

Cross-validation can be used to choose s, or s can be specified directly

Local regression



Generalized additive models

A flexible way to predict response y from multiple predictors $x_1, \ldots, x_p$

A natural way to extend the standard multiple linear model

$$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

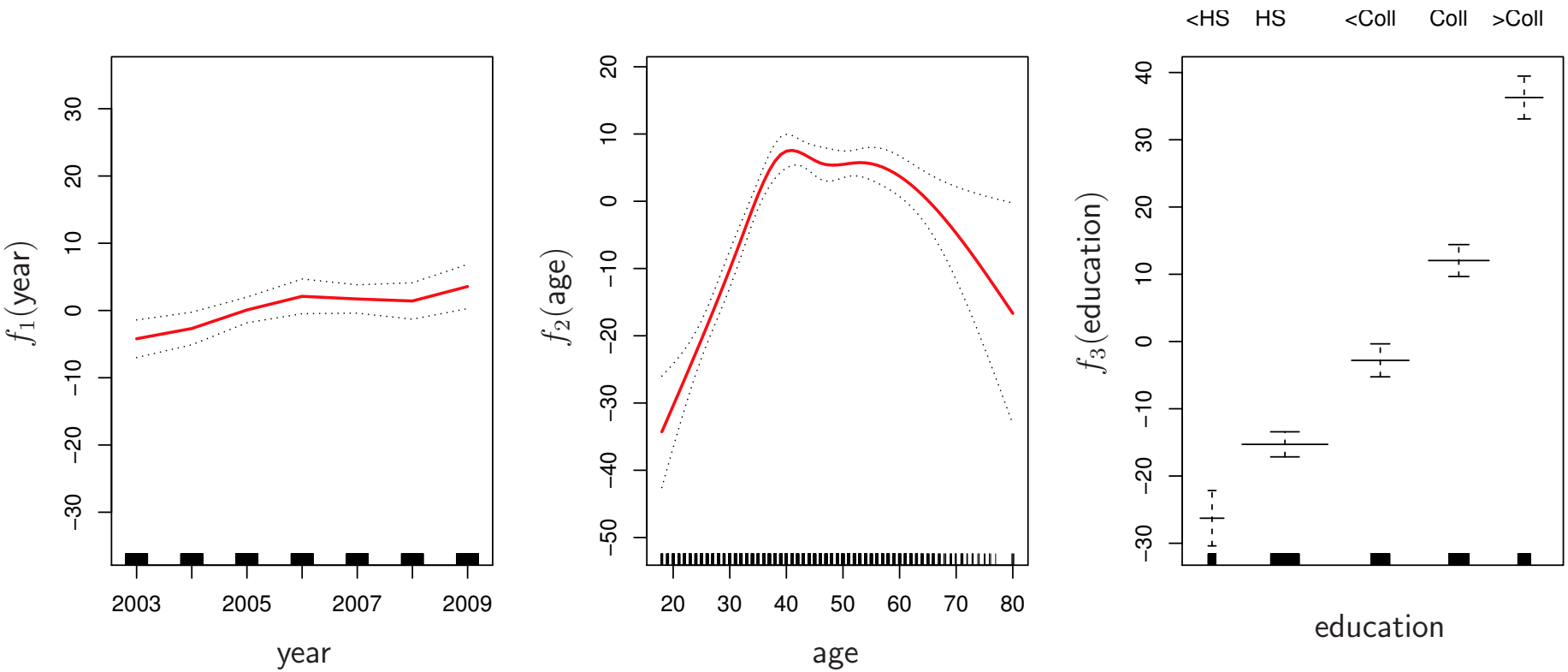
in order to allow for non-linear relationships, is to write

$$f(x_1, \ldots, x_p) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$$

where $f_1(x_1), \ldots, f_p(x_p)$ are (smooth) non-linear functions

Generalized additive models

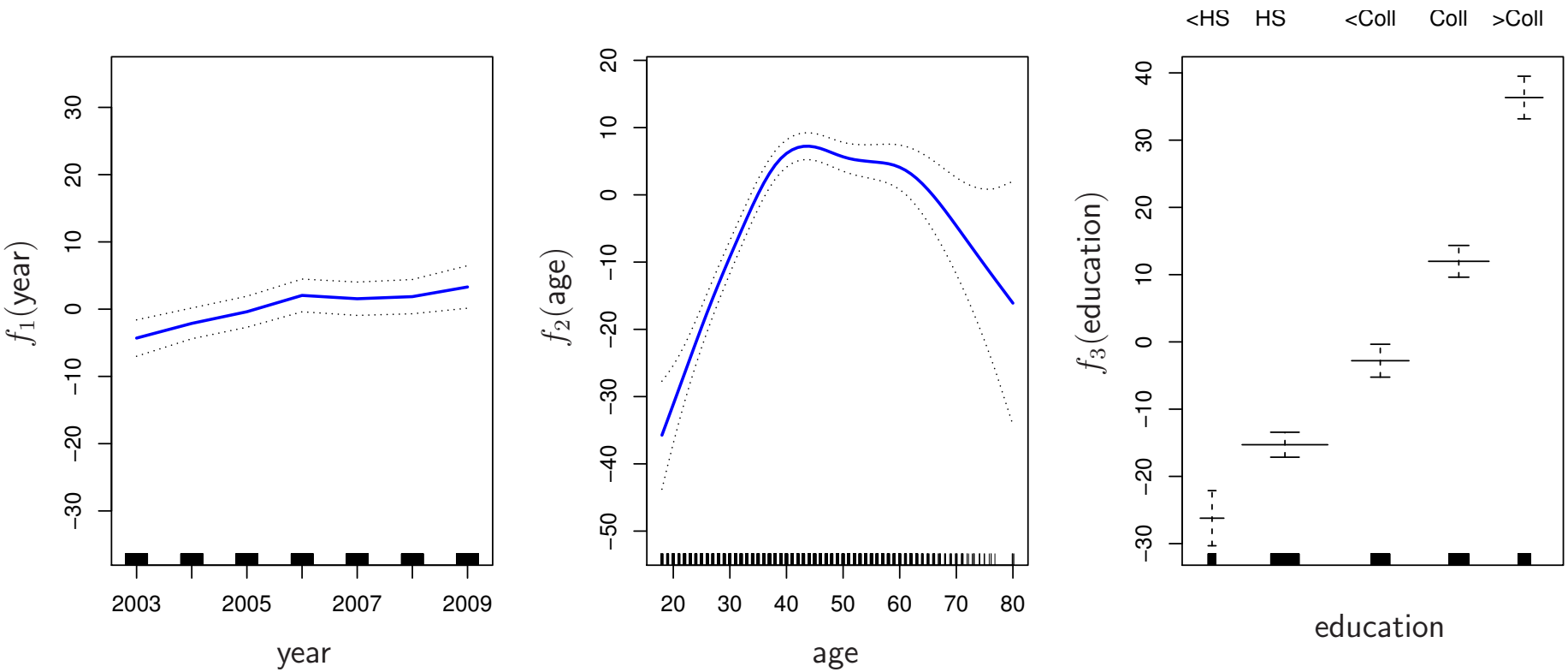
Prediction of $y = \text{wage}$ from $x_1 = \text{year}$, $x_2 = \text{age}$, and $x_3 = \text{education}$



$f(x_1, x_2, x_3) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3)$, where $f_1(x_1)$ and $f_2(x_2)$ are natural splines, and $f_3(x_3)$ is a piecewise linear function

Generalized additive models

Prediction of $y = \text{wage}$ from $x_1 = \text{year}$, $x_2 = \text{age}$, and $x_3 = \text{education}$



$f(x_1, x_2, x_3) = \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3)$, where $f_1(x_1)$ and $f_2(x_2)$ are smoothing splines, and $f_3(x_3)$ is a piecewise linear function

Conclusion

- Many approaches to account for non-linearities
- Cross-validation can help in making different choices
- All approaches discussed can be applied using R
- Book: An Introduction to Statistical Learning by James, Witten, Hastie, and Tibshirani