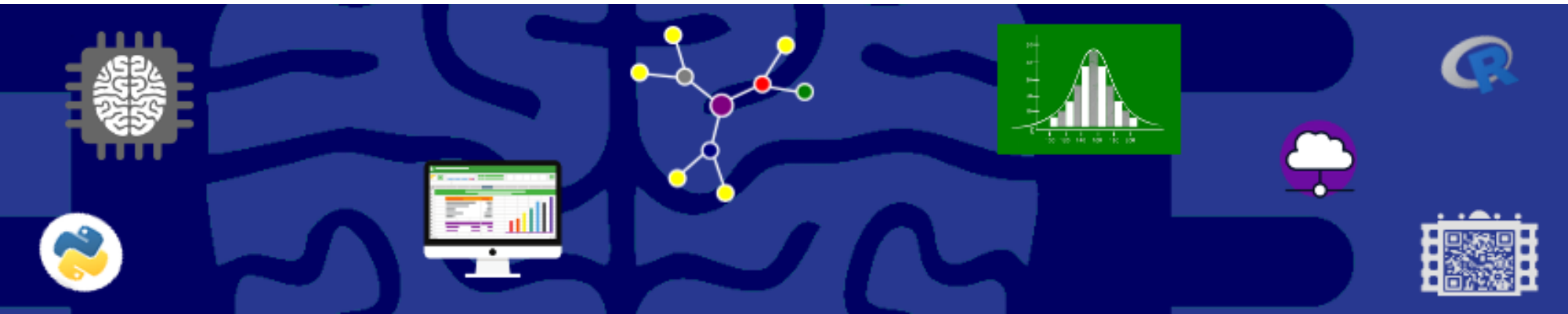
Introduction to Data Science

Winter Semester 2018/19

Oliver Ernst

TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



Contents I

Contents II

Contents III

Contents

Nonlinear Regression Models

Chapter overview

- Despite the benefits of simplicity and interpretability of the standard linear model for regression, it will suffer from large bias if the model generating the data depends nonlinearly on the predictors.
- In this chapter we explore methods which make the linear regression model more flexible by using **linear combinations of nonlinear functions**, specifically
  1. polynomial and piecewise polynomial functions,
  2. piecewise constant functions,
  3. piecewise piecewise polynomial functions with penalty terms and
  4. generalized additive model functions

  of the predictors.

Contents

- For univariate models, **polynomial regression** replaces the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

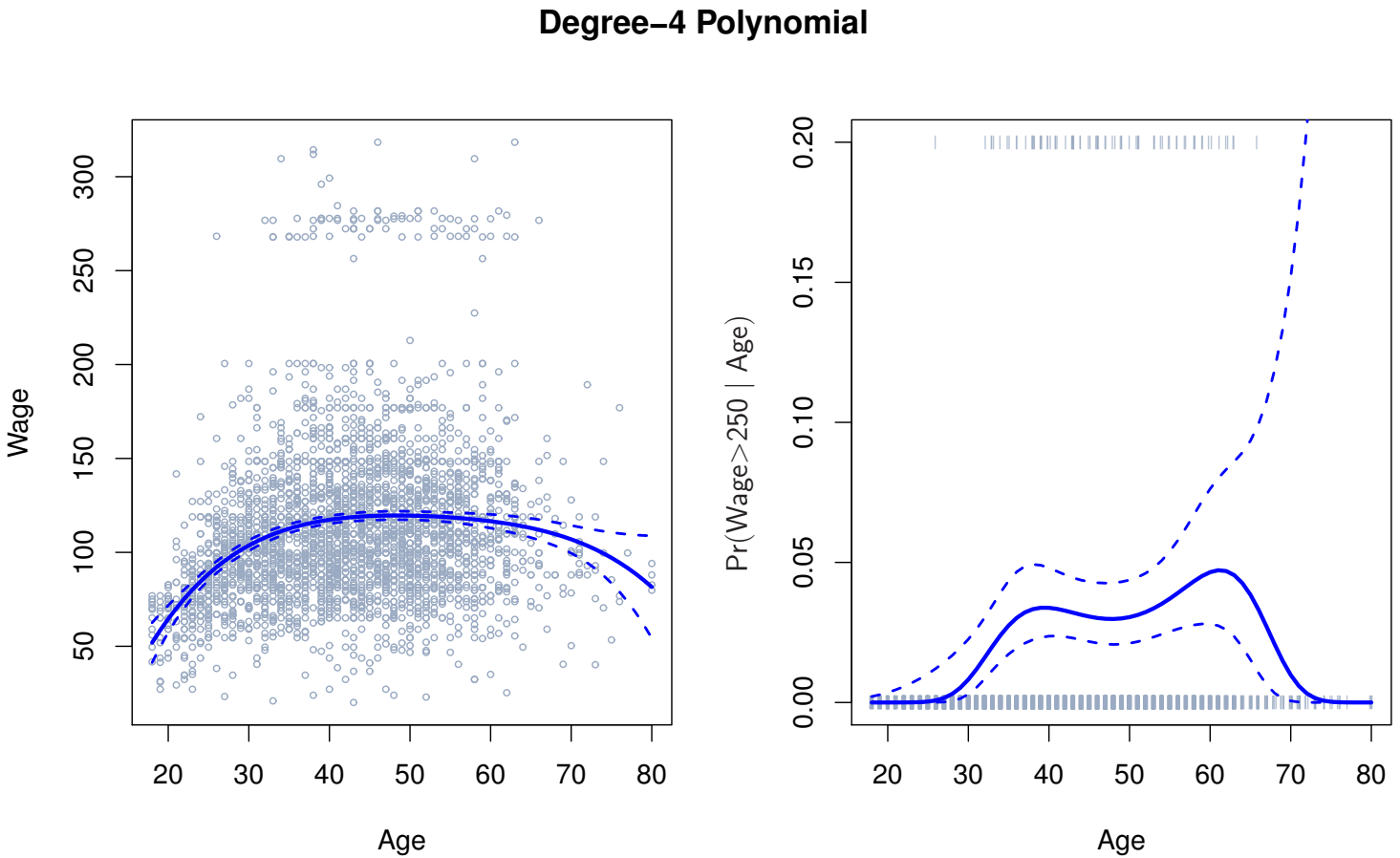
with a polynomial of degree $d > 1$ in the predictor variable

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \varepsilon.$$

- High degree polynomials are often difficult to handle due to their oscillatory behavior and their unboundedness for large arguments, so that degrees higher than 4 can become problematic if employed naively.
- **Example:** `Wage` data set: income and demographic information for males who reside in the central Atlantic region of the United States.
Fit response `wage` [in $ 1000] to predictor `age` by LS using a polynomial of degree $d = 4$.

Nonlinear Regression Models

Polynomial Regression



Left: Polynomial ($d = 4$) LS fit of `wage` against `age` (solid blue) with 95% confidence interval (blue dashed). Right: Model of event $\{$`wage` $> 250\}$ using logistic regression with $d = 4$, fitted posterior probability (solid blue) with 95% confidence interval (blue dashed).

Left panel in previous figure:

- Given fit of particular age value x_0

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

  use variance/covariance estimates of $\hat{\beta}_j$ to estimate variance of $\hat{f}(x_0)$.
- If $\hat{\boldsymbol{C}} \in \mathbb{R}^{5\times 5}$ is the estimated covariance matrix of the β_j, then

$$\mathbf{Var}\, \hat{f}(x_0) = \boldsymbol{\ell}_0^\top \hat{\boldsymbol{C}} \boldsymbol{\ell}_0, \qquad \text{where} \qquad \boldsymbol{\ell}_0 = (1, x_0, x_0^2, \ldots, x_0^4)^\top.$$

- Estimated **pointwise standard error** of $\hat{f}(x_0)$ is the square root of this variance.
- Repeat calculation for all x_0, plotting $\pm 2\times$ standard error (corresponds to $\approx$ 95% confidence interval for normally distributed errors) yields dashed lines.

Right panel in previous figure:

- Observations seem to fall into 2 classes: *high earners* ($>$ \$250K) and *low earners*; treat `wage` as binary response variable with these two groups.
- Using logistic regression, can predict this binary response using polynomial functions of predictor `age`.
- This corresponds to fitting

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

- Gray marks in figure denote ages of high and low earners.
- Solid blue: fitted probabilities of being high/low earner given `age`, dashed blue gives 95% confidence interval (very wide).
- Only 79 high earners of $n = 3000$ observations, results in high variance of coefficients and therefore wide confidence intervals.

Contents

Idea:

- Polynomials are globally defined on the domain of the predictor(s) X.
- To model more locally varied response behavior, divide domain of X into subdomains and use different response model on each.
- Simplest case: different constant function on each subinterval.
- Amounts to converting a continuous variable into an **unordered categorical variable**.

- Introduce "cut points" $c_1 < c_2 < \cdots < c_K$ in range of X, construct $K+1$ new (dummy) variables with indicator function $\mathbb{1}(\cdot)$

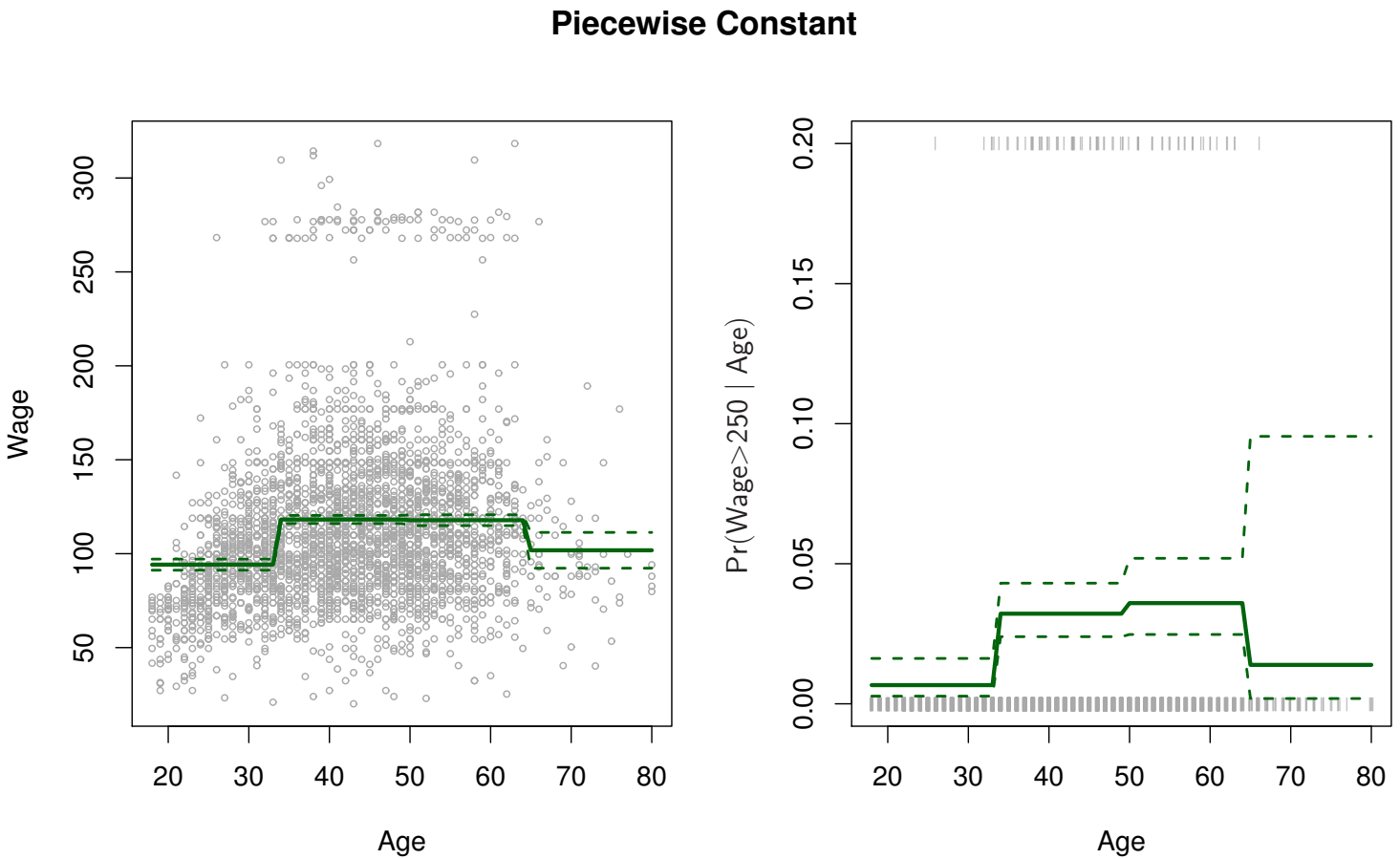
$$\begin{aligned} C_0(X) &= \mathbb{1}(X < c_1), \\ C_1(X) &= \mathbb{1}(c_1 \le X < c_2), \\ &\vdots \\ C_{K-1}(X) &= \mathbb{1}(c_{K-2} \le X < c_{K-1}), \\ C_K(X) &= \mathbb{1}(c_K \le X). \end{aligned} \tag{7.1}$$

- Since events exhaustive and mutually exclusive we have $\sum_{k=0}^{K} C_k(X) \equiv 1$.
- Now fit LS model using $C_1(X), \ldots, C_K(X)$ as predictors[9]:

$$y_i = \beta_0 + \beta_1 C_1(X) + \beta_2 C_2(X) + \cdots + \beta_K C_K(X) + \varepsilon_i.$$

$\beta_j (j > 0)$: average increase in response for $X \in [c_j, c_{j+1})$ relative to $X < c_1$.

[9] Omit C_0 as this is redundant with the intercept.



Left: piecewise constant fit of `wage` against `age` (solid) with 95% confidence band (dashed). Right: modeling event $\{$`wage` $> 250\}$ using logistic regression (solid) with 95% confidence band (dashed).

Previous figure:

- Left: capturing response behavior requires choosing the cut points appropriately. Increasing trend of `wage` with `age` clearly missed in first bin.
- Right: logistic regression fits

$$\mathbf{P}(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(X) + \cdots + \beta_K C_K(X)}{1 + \exp(\beta_0 + \beta_1 C_1(X) + \cdots + \beta_K C_K(X)}$$

to predict probability of being high earner given `age`.

Piecewise constant approximation popular in biostatistics and epidemiology, where bins often correspond to 5-year age groups.

Nonlinear Regression Models

General regression functions

- Polynomial, piecewise constant regression examples of **basis function approach**, where linear combination of transformations $\{b_k(X)\}_{k=1}^K$ of predictor variables used for fitting:

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

- Basis functions b_k chosen a priori. Examples:

$$b_k(x_i) = \begin{cases} x_i^k & \text{polynomial regression,} \\ \mathbb{1}(c_k \leq x_i < c_{k+1}) & \text{piecewise constant regression.} \end{cases}$$

- Model still linear in the coefficients, hence all inferential methods of linear LS still applicable (standard errors for coefficient estimates, F-statistics for model significance etc.).
- Many possible choices: wavelets, Fourier modes, splines, etc.

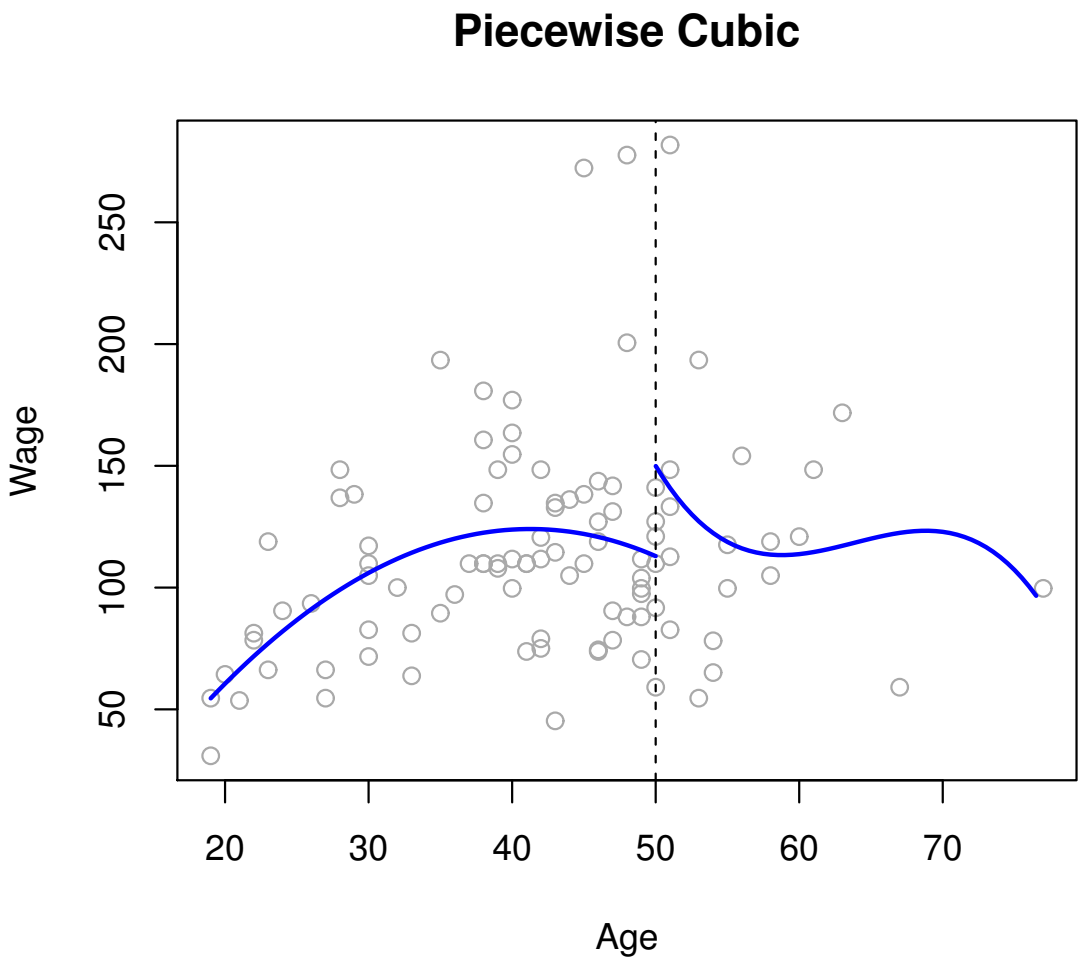
Contents

- As in piecewise constant modes, introduce partition of X domain into sub-intervals.
- Fit a different low-degree polynomial in each subinterval.
- E.g. cubic:

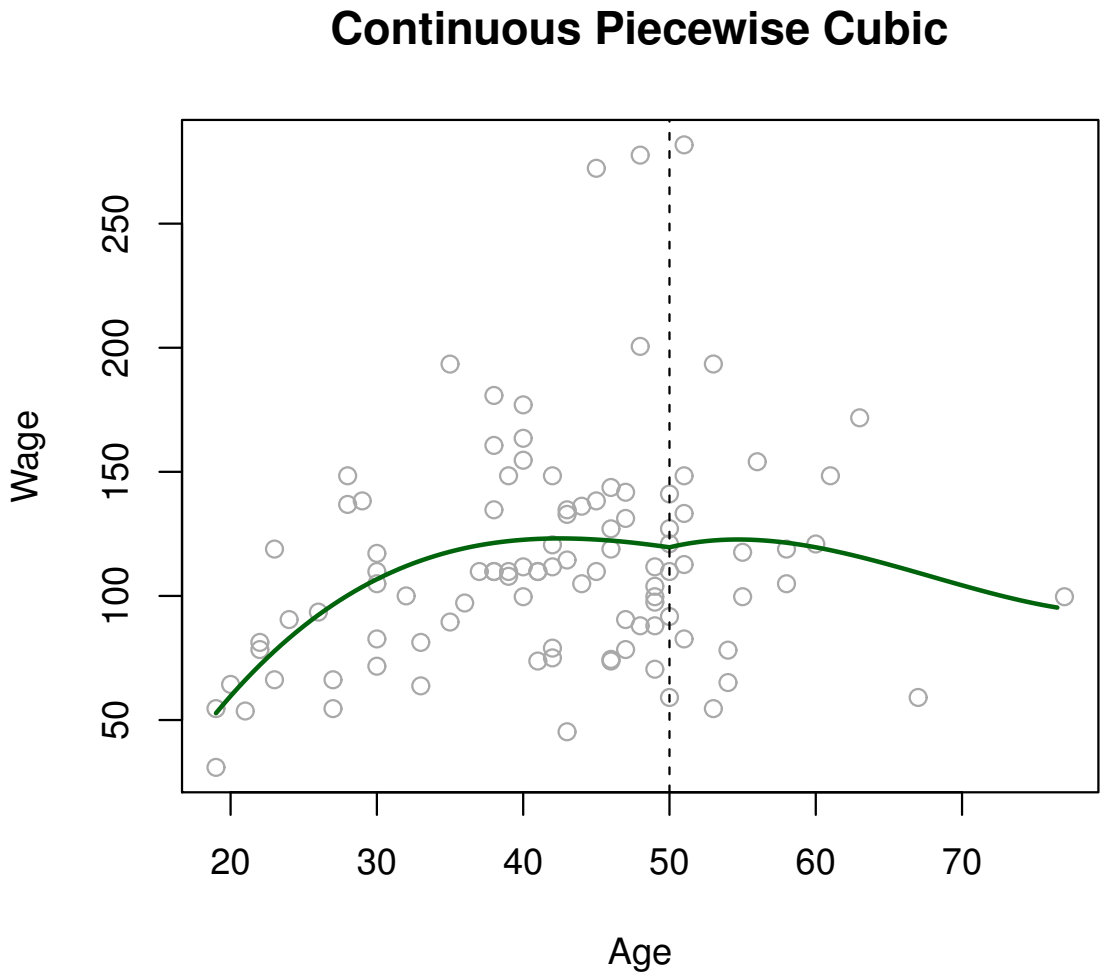
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i, \tag{7.2}$$

  with separate coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ in each subinterval.
- Spline termionology: cut points called **knots**.
- Piecewise cubic with single knot at $X = c$:

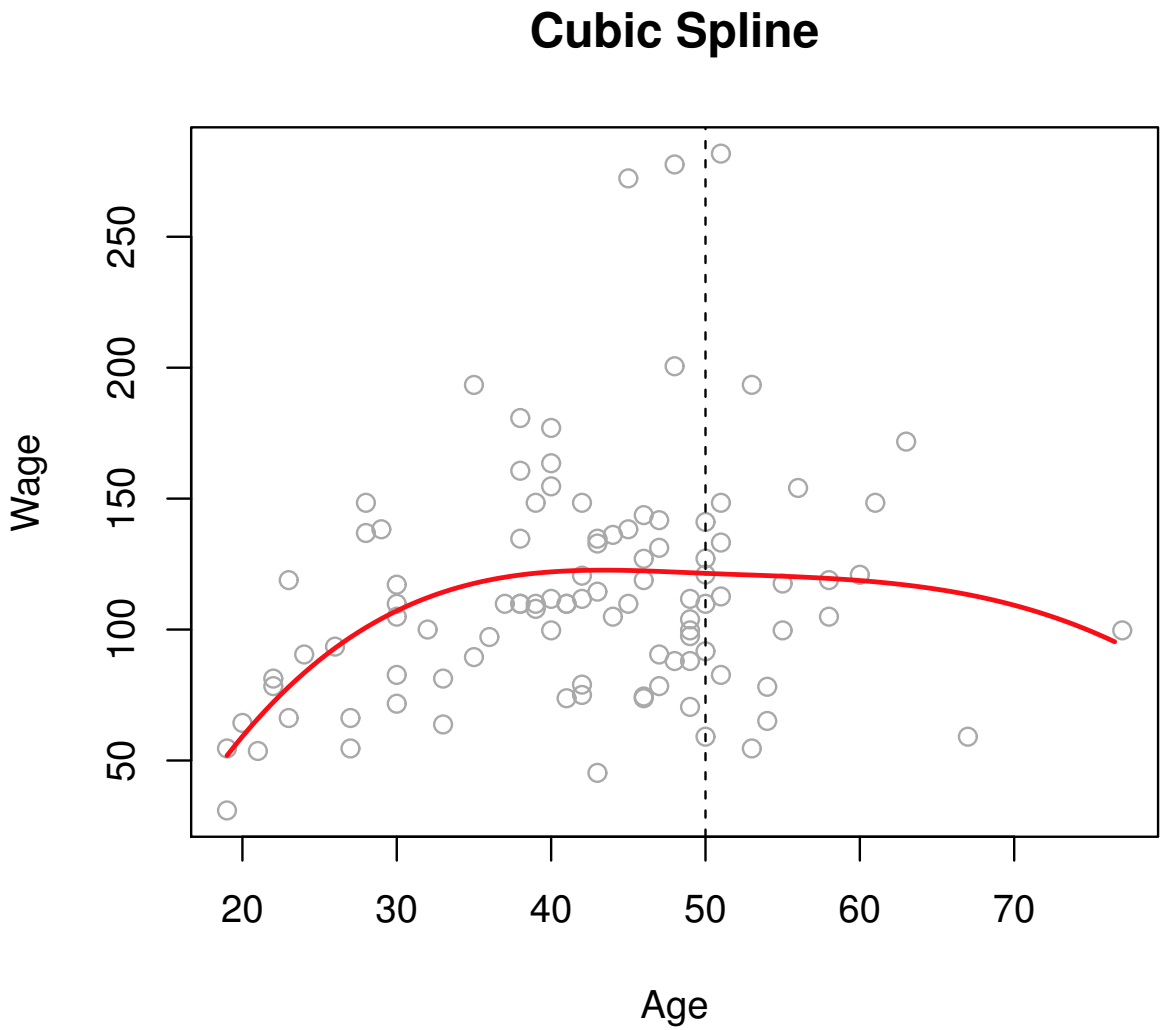
$$y_i = \begin{cases} \beta_{0,1} + \beta_{1,1} x_i + \beta_{2,1} x_i^2 + \beta_{3,1} x_i^3 + \varepsilon_i, & \text{if } x_i < c, \\ \beta_{0,2} + \beta_{1,2} x_i + \beta_{2,2} x_i^2 + \beta_{3,2} x_i^3 + \varepsilon_i, & \text{if } x_i \geq c. \end{cases}$$



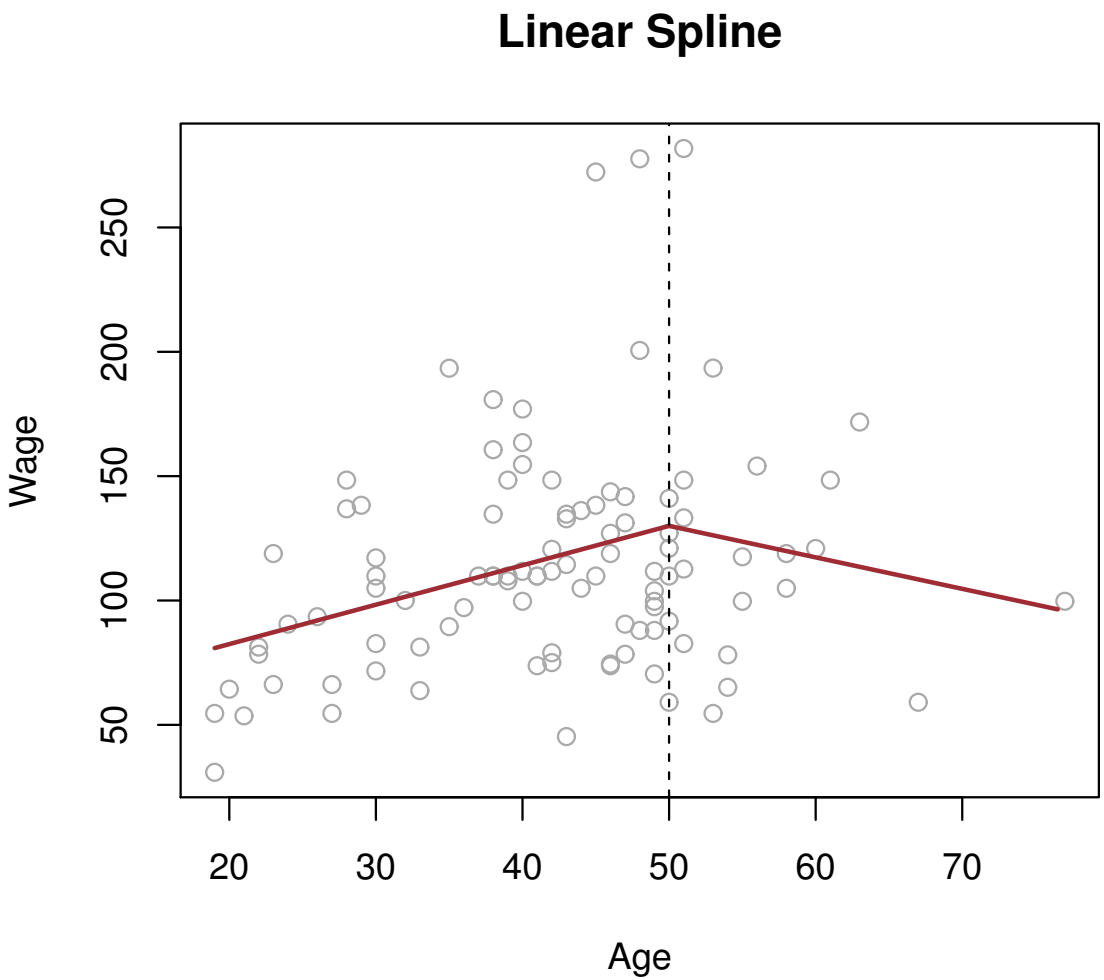
A piecewise cubic fit of `wage` against `age` for the `Wage` data set. Note the discontinuity at the (single) knot $c = 50$. Model has $8 = 2 \times 4$ degrees of freedom.



A piecewise cubic fit of the same data, now with the added constraint that the two polynomials should agree at the knot. This still leaves a 'kink' at the knot, i.e., a discontinuity of the first derivative.



A piecewise cubic fit of the same data, now with the added constraint that the two polynomials as well as their first derivatives should agree at the knot.



A piecewise linear fit of the same data with continuity constraint.

- Cubic spline with K knots: $4 + K$ degrees of freedom.
- General definition of (univariate) spline: piecewise polynomial of degree d with continuity of derivatives of orders $0, 1, 2 \ldots, d-1$.
- Cubic spline model with K knots can be modeled as

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

  using appropriate basis functions.
- One possible basis (cubic case): start off with monomials x, x^2, x^3, then add for each knot ξ one **truncated monomial**

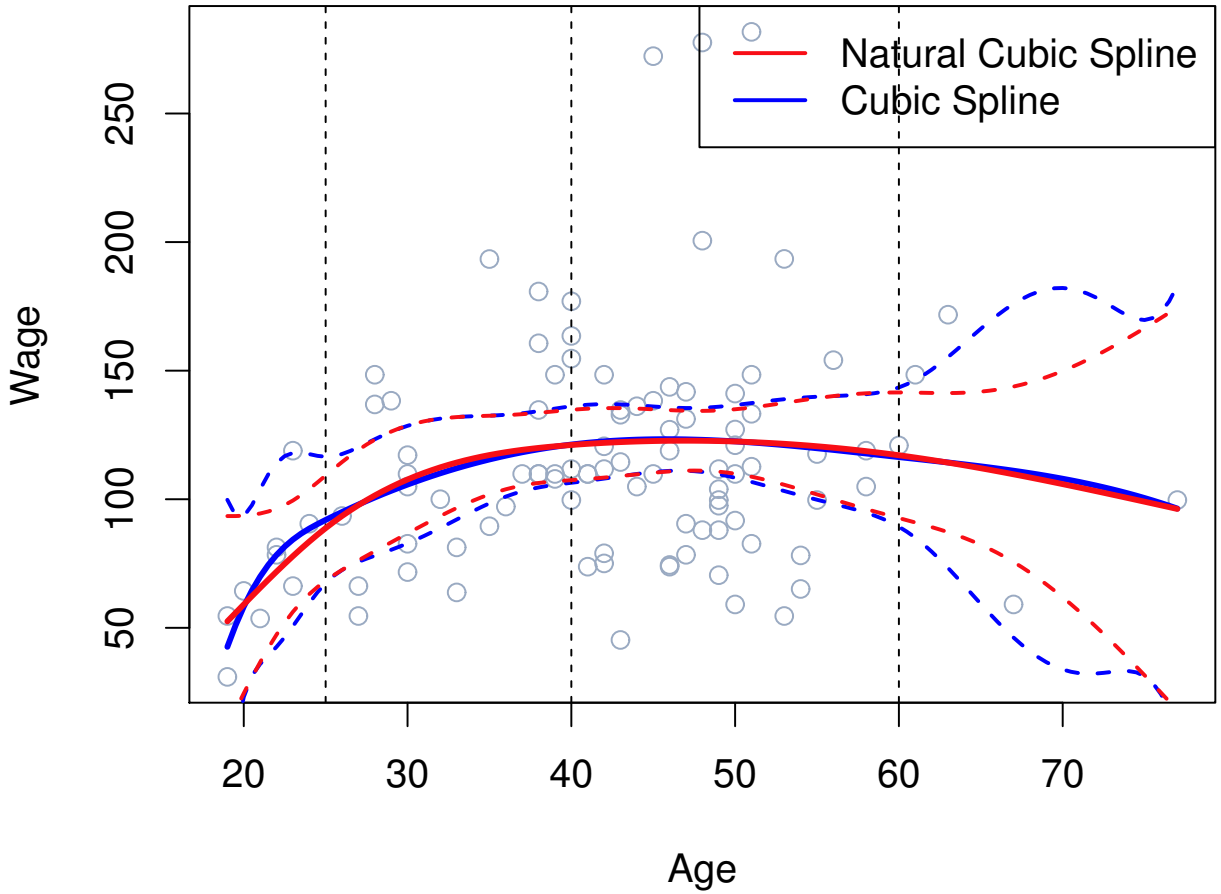
$$h(x, \xi) := (x - \xi)_+^3 := \begin{cases} (x-\xi)^3 & \text{if } x > \xi, \\ 0 & \text{otherwise.} \end{cases}$$

- Adding single basis function $h(x, \xi)$ to model (7.2) will introduce discontinuity only in third derivative at $x = \xi$.

Nonlinear Regression Models

LS regression with splines

To fit LS regression model with cubic splines using K knots $\{\xi_k\}_{k=1}^K$, use $K+3$ predictor variables $X, X^2, X^3, h(X,\xi_1), \ldots, h(X,\xi_K)$.

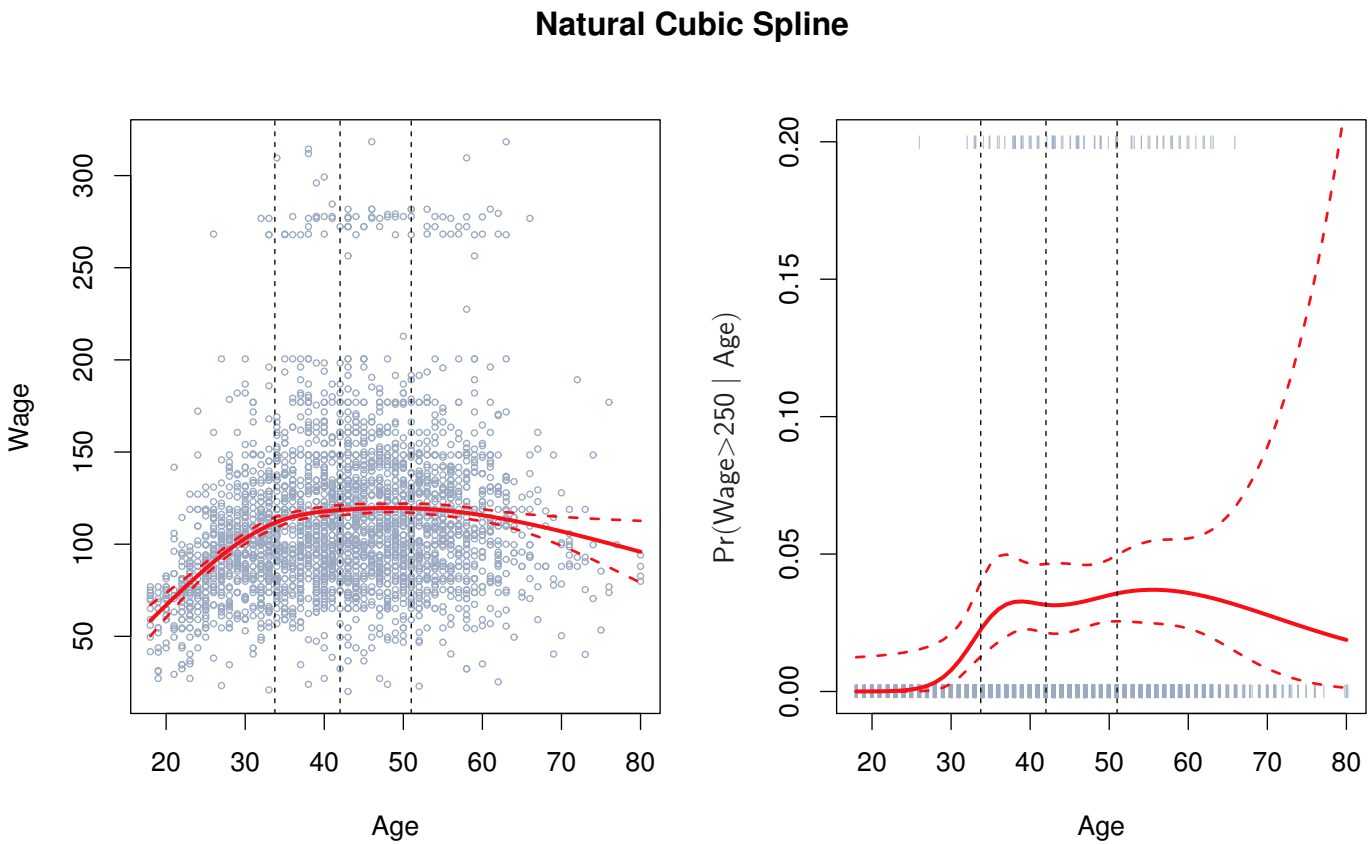


Cubic and natural (linear beyond boundary knots) spline fit using 3 knots to fit a subset of the `Wage` data. Note the large variance near the endpoints.

Nonlinear Regression Models

Choice of spline knots

- Spline most flexible near knots, place these where most variability expected.
- Common practice: space knots uniformly, choose # degrees of freedom, have software place knots at uniform quantiles.



Nonlinear Regression Models

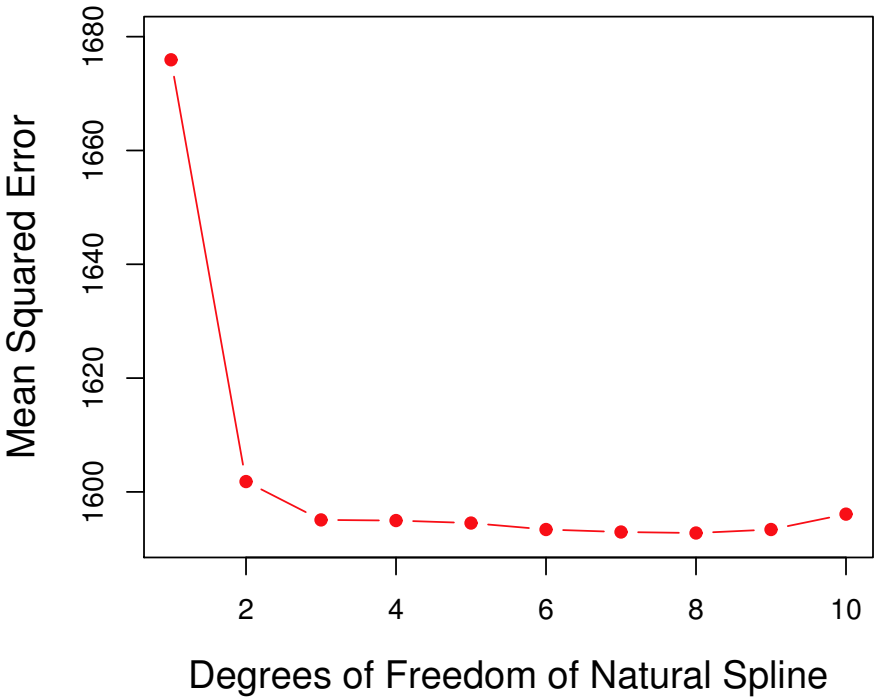
Choice of spline knots

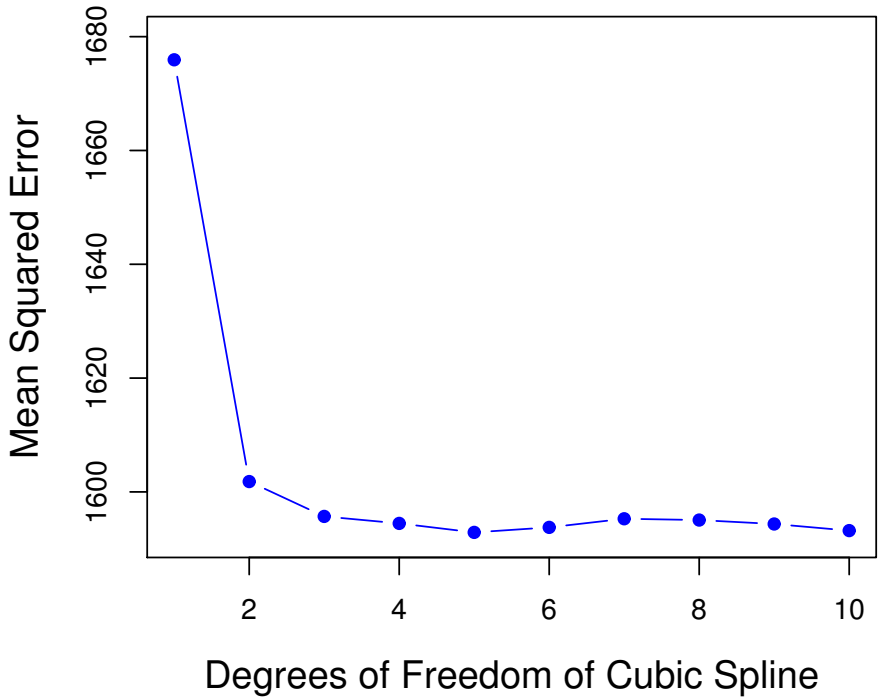
Previous figure:

- Fit natural cubic spline to `Wage` data. Three knots, chosen automatically at 25th, 50th and 75th percentiles of `age`.
- Requested 4 DOF, leading to 3 interior knots. Actually: 5 knots including 2 boundary knots. Corresponds to $9 = 5 + 4$ DOF for cubic spline. Two natural constraints at boundary knots to enforce linearity, leaving $5 = 9 - 4$ DOF. One DOF absorbed in intercept, leaves 4 DOF.
- Right panel: Logistic regression modeling binary event $\{\texttt{wage} > 250\}$. Shown: fitted posterior probability.
- Choosing # knots: trial and error or cross-validation.

Nonlinear Regression Models

Choice of spline knots





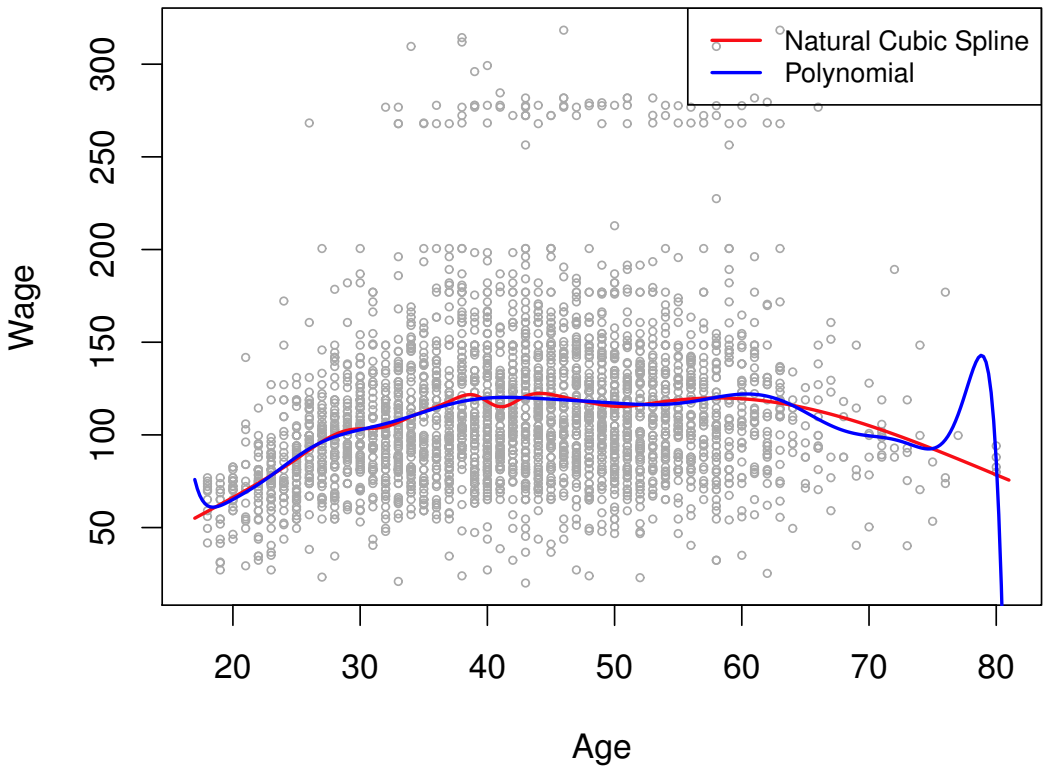
Ten-fold CV MSE for selecting DOF when fitting splines to `Wage` data.

Clear result: 1-DOF not adequate.

Nonlinear Regression Models

Comparison with polynomials

- Spline regression often superior to polynomial.
- More stable as flexibility comes from variation of coefficients of low-degree polynomials and knot placement.



For `Wage` data: comparison of natural cubic spline with 15 DOF to polynomial of degree 15. Latter shows spurious variation near endpoints.

Contents

- Fitting data with smooth function g: want small RSS $= \sum_{i=1}^{n}(y_i - g(x_i))^2$.
- With no constraints on g, can always attain RSS $= 0$ by interpolating the data, leading to overfitting.
- Ensure smoothness by adding penalty term: minimize

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 \,\mathrm{d}t \tag{7.3}$$

  with tuning parameter λ controlling weight assigned to smoothness.
- Limiting values: $\lambda = 0$ corresponds to no smoothing, leading to interpolation for sufficiently many DOF; $\lambda \to \infty$ tends to linear LS fit.
- λ controls bias-variance tradeoff of smoothing spline.
- Can show: minimizer of (7.3) is natural cubic spline with knots at $x_1, \ldots, x_n$. Not the natural cubic spline of the basis function approach, but a **shrunken** version, degree of shrinkage controlled by λ.

- Smoothing spline: natural cubic spline with knots at $x_1, \ldots, x_n$, i.e., n DOF.
- Can show: as $\lambda \to \infty$, **effective degrees of freedom** df_λ decrease from n to 2.
- Smoothing spline has nominally n DOF, these are heavily constrained, i.e., they are "shrunk" by higher weighting of the penalty term.
- Measure of flexibility of smoothing splines: df_λ.
- Mapping from observation vector $\boldsymbol{y} \in \mathbb{R}^n$ to vector $\hat{\boldsymbol{g}}_\lambda$ of n coefficients defining the smoothing spline with penalty parameter λ is linear, i.e.,

$$\boldsymbol{S}_\lambda \hat{\boldsymbol{g}}_\lambda = \boldsymbol{y}, \qquad \boldsymbol{S}_\lambda \in \mathbb{R}^{n\times n}.$$

Effective DOF defined by

$$df_\lambda := \operatorname{tr} \boldsymbol{S}_\lambda = \sum_{i=1}^{n} [\boldsymbol{S}_\lambda]_{i,i}.$$

Nonlinear Regression Models

Smoothing splines: choosing λ

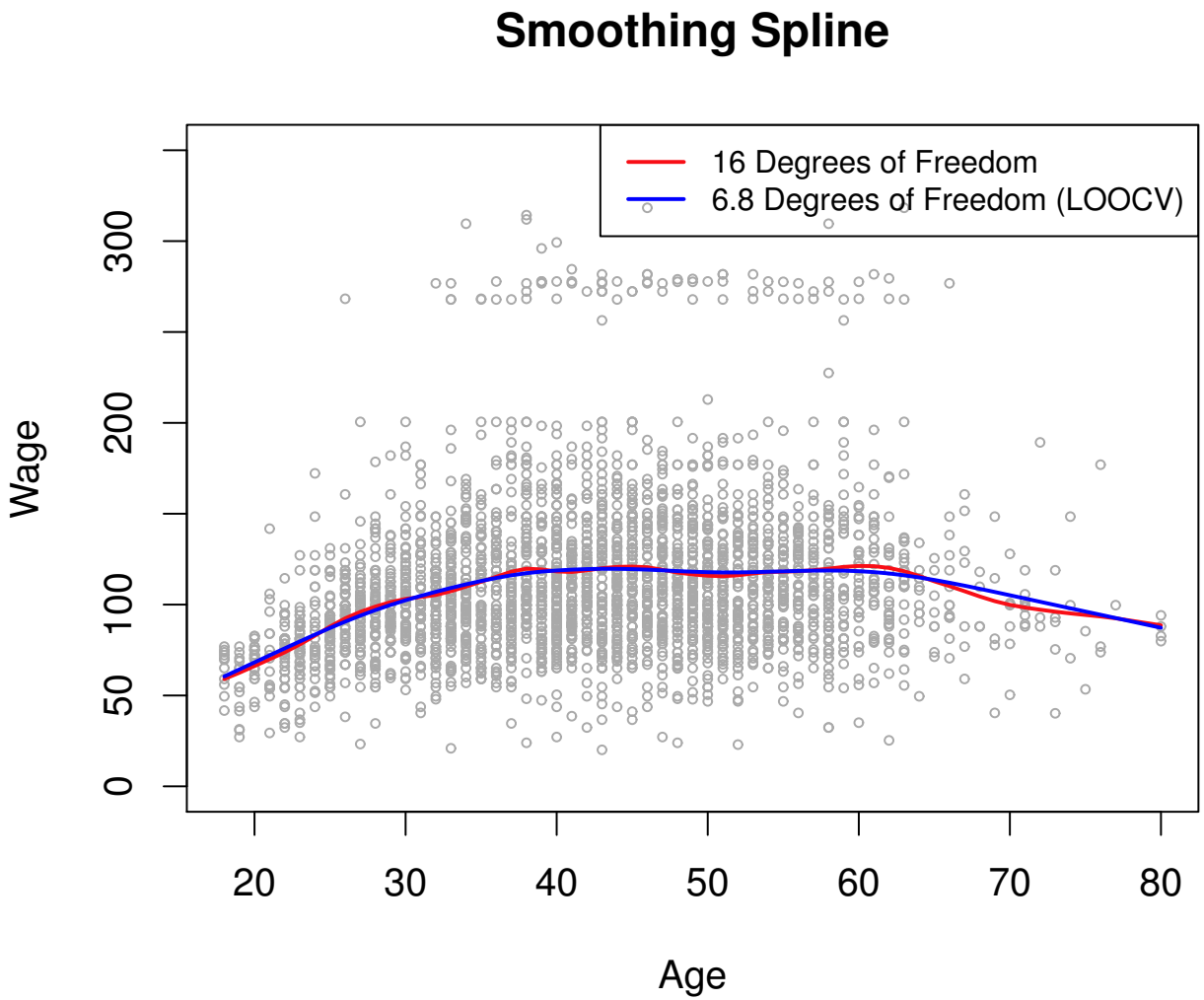
- For smoothing splines no need to choose knot number and locations, each predictor observation x_i is a knot.
- Remaining problem is choice of smoothing parameter λ.
- Obvious option: choose λ to minimize CV estimates of RSS.
- For smoothing splines LOOCV error can be computed at nearly the cost of single fit:

$$\mathrm{RSS}_{CV}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - [\boldsymbol{S}_\lambda]_{i,i}}\right]^2,$$

$\hat{g}_\lambda^{(-i)}(x_i)$: value of smoothing spline fitted with all but i-th observation,

$\hat{g}_\lambda(x_i)$: value of smoothing spline using all observations.

- Similar "magic formula" in (5.1) for LS regression.



Smoothing spline fit to `Wage` data. Red: specified 16 effective DOF.
Blue: λ determined by LOOCV, resulting in $df_\lambda = 6.8$.

Contents

Nonlinear Regression Models

Generalized additive models

- Up to now: single predictor X, extensions of simple linear regression.
- Here: consider extensions of multiple linear regression of response Y on predictors $X_1, \ldots, X_p$.
- Framework: **generalized additive models** (GAMs).
- Allow nonlinear functions of X_j while maintaining additivity.
- Can be applied with quantitative and qualitative responses.

- Extend standard multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

to

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{i,j}) + \varepsilon_i.$$

- Additive: separate f_j for each X_j, then add.
- Example: Consider natural splines and task of fitting model

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \varepsilon \qquad (7.4)$$

from `Wage` data set, with quantitative variables `year`, `age` and qualitative variable `education` $\in$ {`<HS`, `HS`, `<Coll`, `Coll`, `>Coll`}. Fit f_1, f_2 using natural splines, f_3 using separate constant for each value (dummy variable approach).

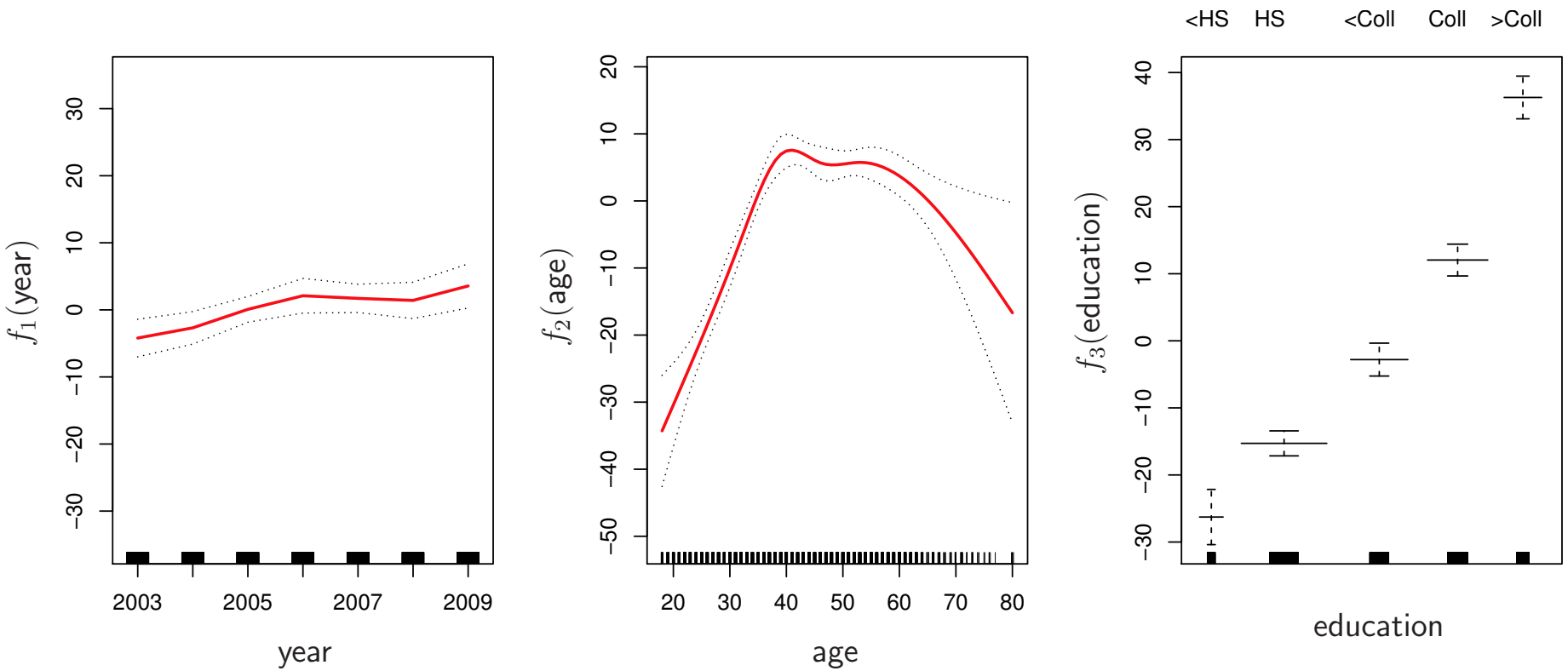
Nonlinear Regression Models

GAMs for regression

- Fit entire model (7.4) using LS, expand each function in natural spline basis or dummy variables, resulting in single large regression matrix.

Nonlinear Regression Models

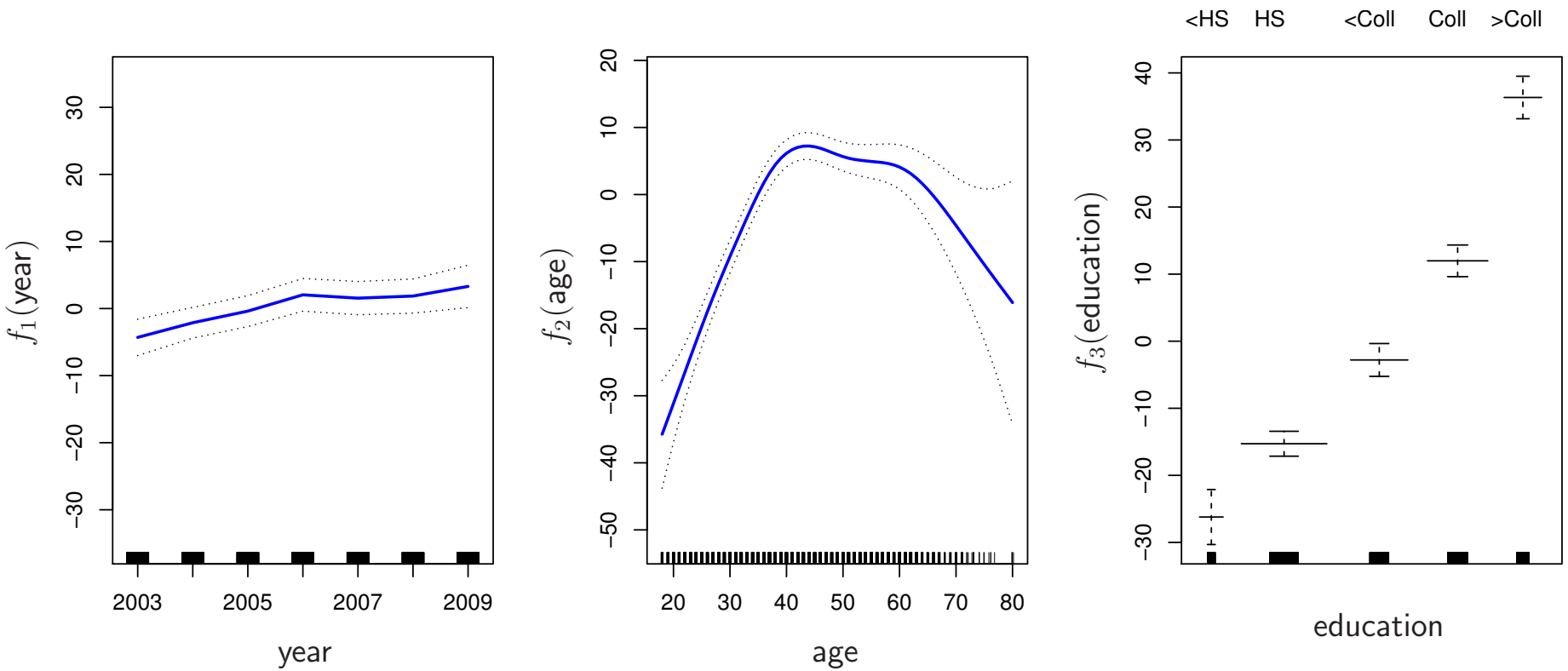
GAMs for regression



Relationship of each feature and response (`wage`). f_1 and f_2 are natural splines in `year` and `age` with 4 and 5 DOF, respectively. f_3 is a step function fit to qualitative predictor `education`.

Nonlinear Regression Models

GAMs for regression



Same as before except f_1 and f_2 smoothing splines with 4 and 5 DOF, respectively. Fit of smoothinmg splines more difficult than for natural splines, standard software solves an optimization problem via an algorithm known as **backfitting**.

Nonlinear Regression Models

GAMs: benefits and shortcomings

+ GAMs allow fitting nonlinear f_j to each X_j in order to capture nonlinear dependencies.
+ Potentially more accurate predictions of response Y.
+ Model still additive, effect of each X_j can be examoned separately, useful for inference.
+ Smoothness of each f_j can be summarized via (effective) DOF.
- Additivity is a restriction, interactions can be missed. Can add interaction terms manually by adding predictors $X_j \times X_k$ or low degree interaction functions $f_{j,k}(X_j, X_k)$.

GAMs are a useful compromise between linear and fully nonparametric methods such as random forests and boosting (later).