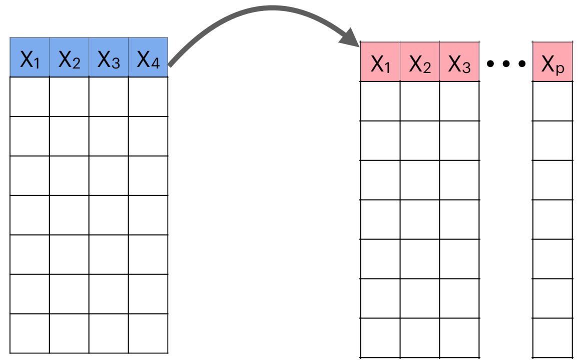
“Non-linear” Linear Regression

Lecture 8

Termeh Shafie

Recall: Feature Engineering



when do we do this and why?

Basis Function

A family of functions/transformations that can be applied to a variable X: $f(X_1)$, $f(X_2)$, $f(X_3)$, …

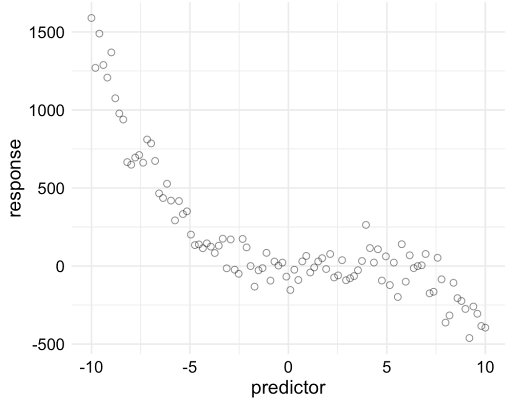
$$Y = \beta_0 + \beta_1 f(X_1) + +\beta_2 f(X_2) + \beta_3 f(X_3) + \cdots + \beta_k f(X_k) + \epsilon$$

Polynomial Regression Models

```
Polynomial regression fits a nonlinear relationship by modeling
the response as a polynomial function of the predictor
```

The Assumption of Linearity

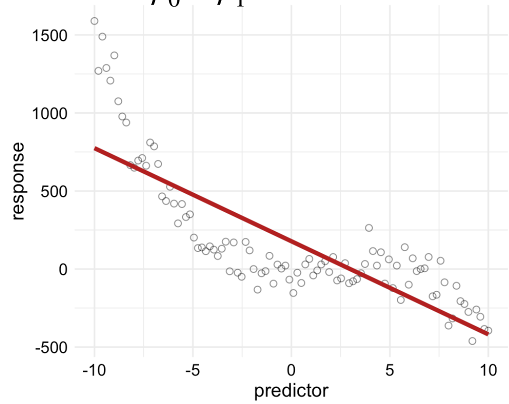
in reality the relationships between predictors and the response
are almost never exactly (first order) linear...



5

Polynomial Regression Models

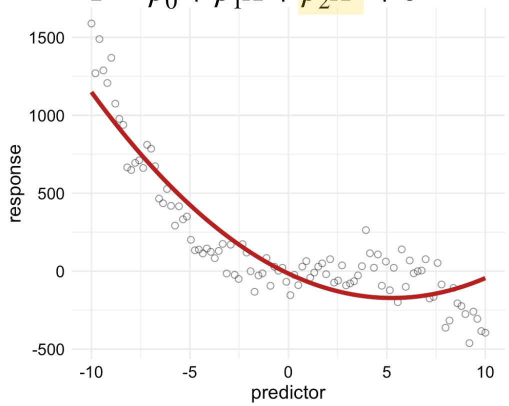
$$Y = \beta_0 + \beta_1 X + \epsilon$$



6

Polynomial Regression Models

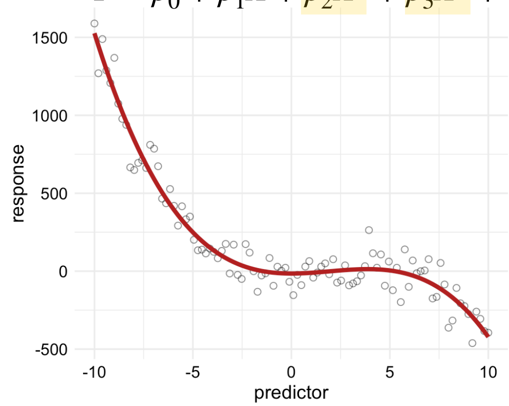
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$



7

Polynomial Regression Models

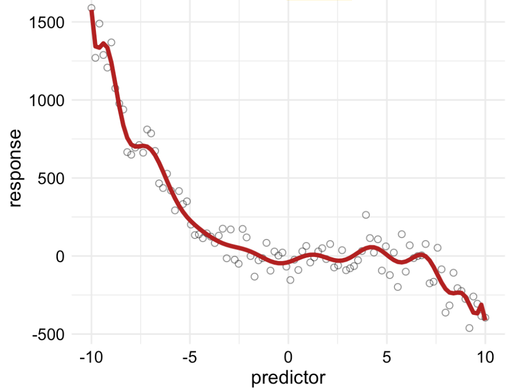
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$



8

Polynomial Regression Models

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots\cdots + \beta_{20} X^{20} + \epsilon$$



Polynomial Regression Models

in general, polynomial models are of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots\cdots + \beta_n X^n + \epsilon$$

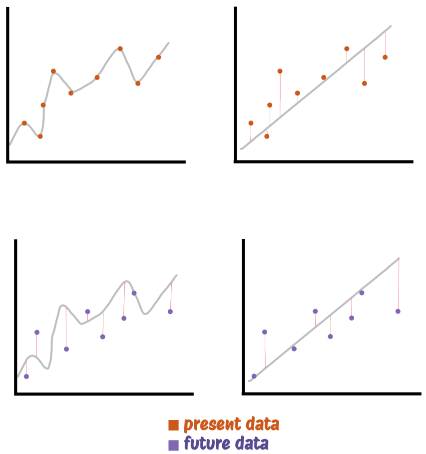
where d is called the **degree** of the polynomial

- non-linear relationship between predictors and response captured by polynomial terms but model remains linear in the parameters
- example: model can be written as

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

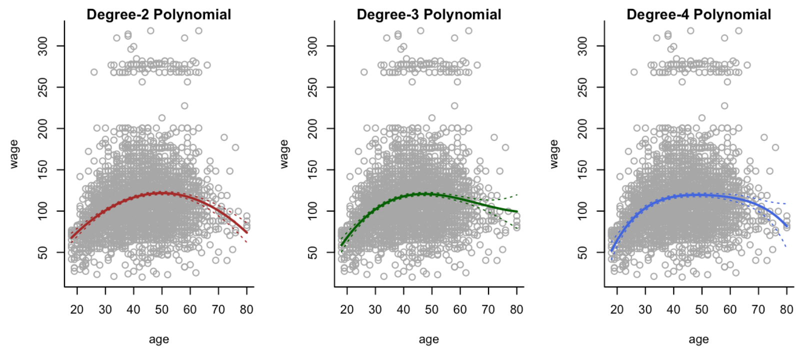
  where $X_1 = X,\ X_2 = X^2,\ X_3 = X^3$
- we can use LS for estimation

Polynomial Regression Models: Choosing d



Polynomial Regression Models

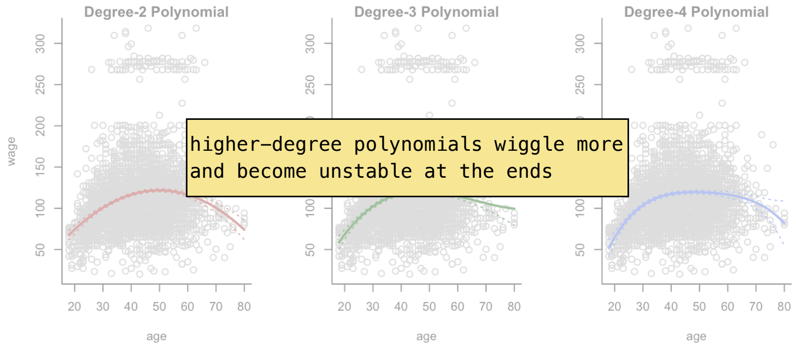
Example: Wage (ISLR2)



95% confidence interval for the mean prediction at x:

$\hat{f}(x) \pm 2 \times \mathrm{SE}[\hat{f}(x)]$ where $\mathrm{SE}[\hat{f}(x)]$ is the standard error of the mean prediction at x

Polynomial Regression Models



higher-degree polynomials wiggle more and become unstable at the ends

Polynomial Regression Models

Example: Wage (ISLR2)

```
Analysis of Variance Table

Model 1: wage ~ poly(age, 1)
Model 2: wage ~ poly(age, 2)
Model 3: wage ~ poly(age, 3)
Model 4: wage ~ poly(age, 4)
Model 5: wage ~ poly(age, 5)
  Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1   2998 5022216
2   2997 4793430  1    228786 143.5931 < 2.2e-16 ***
3   2996 4777674  1     15756   9.8888  0.001679 **
4   2995 4771604  1      6070   3.8098  0.051046 .
5   2994 4770322  1      1283   0.8050  0.369682
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA

sequential comparisons based on the F-test

For each step:

H_0 : The decrease in RSS from adding the new polynomial term is not significant.

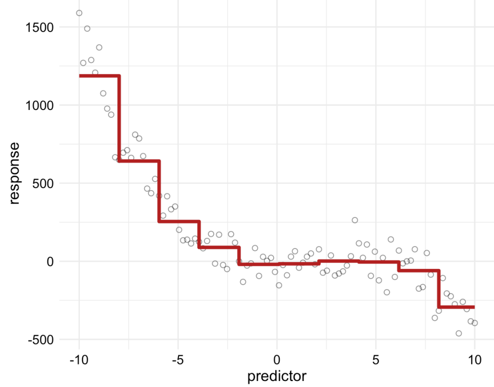
If hypothesis is rejected we move on to next comparison

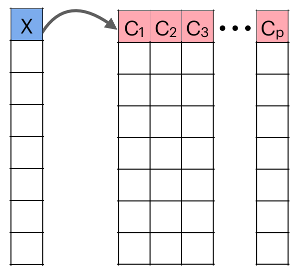
Step Functions

A step function models the predictor by dividing its range into intervals and assigning a constant fitted value within each interval

Step Functions

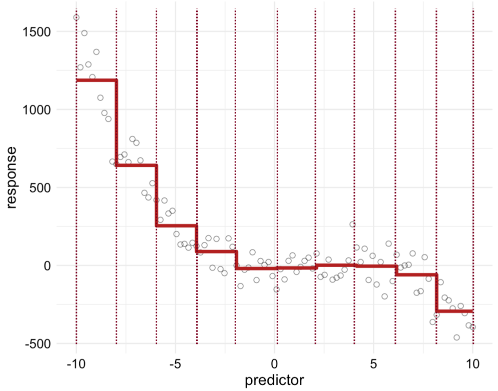
$$Y = \beta_0 + \beta_1 C_1(X) + \beta_2 C_2(X) + \cdots + \beta_K C_K(X) + \epsilon$$





Step Functions

$$Y = \beta_0 + \beta_1 C_1(X) + \beta_2 C_2(X) + \cdots + \beta_K C_K(X) + \epsilon$$

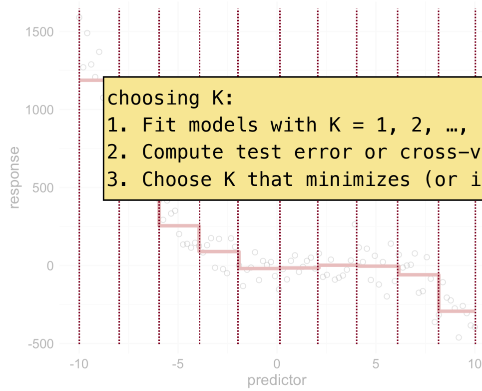


$$\begin{aligned} C_0(X) &= I(X \le c_1) \\ C_1(X) &= I(c_1 < X < c_2) \\ &\vdots \\ C_{K-1}(X) &= I(c_{K-1} < X < c_K) \\ C_K(X) &= I(c_K < X) \end{aligned}$$

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

Step Functions

$$Y = \beta_0 + \beta_1 C_1(X) + \beta_2 C_2(X) + \cdots + \beta_K C_K(X) + \epsilon$$



$C_0(X) = I(X \le c_1)$

$X < c_2)$

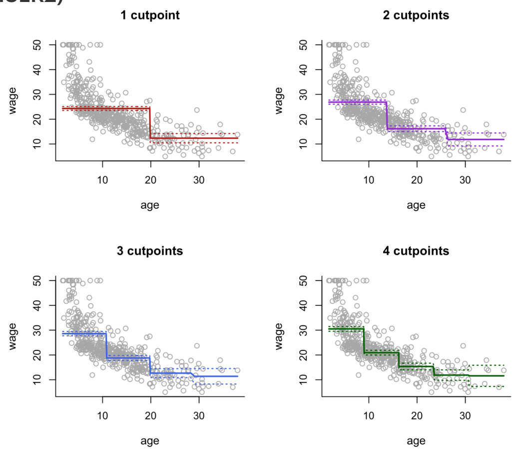
$< X < c_K)$

$C_K(X) = I(c_K < X)$

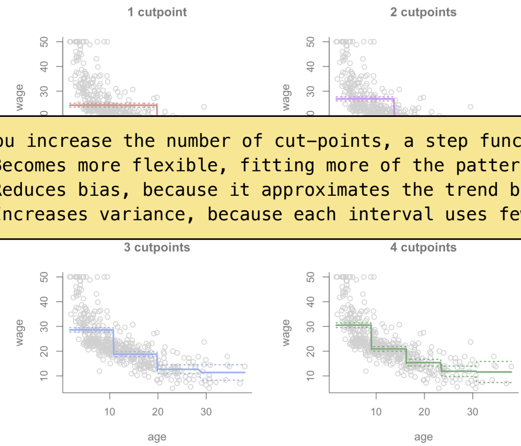
```
choosing K:
1. Fit models with K = 1, 2, …, K_max intervals
2. Compute test error or cross-validation error
3. Choose K that minimizes (or is near-minimum) error
```


Step Functions

Example: Wage (ISLR2)



Step Functions




```
As you increase the number of cut-points, a step function:
  ‣ Becomes more flexible, fitting more of the pattern
  ‣ Reduces bias, because it approximates the trend better
  ‣ Increases variance, because each interval uses fewer data points
```

Regression Splines

```
a spline is a piecewise function
where each segment is a polynomial
```


Regression Splines

The basis of regression splines is **piecewise polynomial regression**

- Standard polynomial regression

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots\cdots + \beta_n X^n + \epsilon$$

- Piecewise polynomial regression:

$$Y = \begin{cases} \beta_{01} + \beta_{11}X + \beta_{21}X^2 + \beta_{31}X^3 + \cdots + \beta_{d1}X^d + \epsilon & \text{if } X < c \\ \beta_{02} + \beta_{12}X + \beta_{22}X^2 + \beta_{32}X^3 + \cdots + \beta_{d2}X^d + \epsilon & \text{if } X \geq c \end{cases}$$

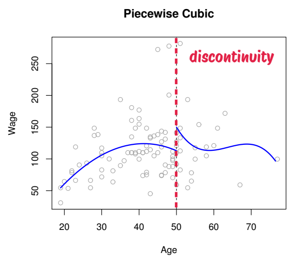
- The c is called a **knot**
- When there is no knot we have standard polynomial regression.
- When we include only the intercepts terms, we have step function regression.
- If we have K knots we are fitting $K + 1$ polynomial models

Regression Splines

Example: Wage (ISLR2)

Piecewise cubic polynomial with a single knot placed a age = 50:

$$\text{wage} = \begin{cases} f_1(\text{age}) = \beta_{01} + \beta_{11}X + \beta_{21}X^2 + \beta_{31}X^3 & \text{if age} < 50 \\ f_2(\text{age}) = \beta_{02} + \beta_{12}X + \beta_{22}X^2 + \beta_{32}X^3 & \text{if age} \geq 50 \end{cases}$$

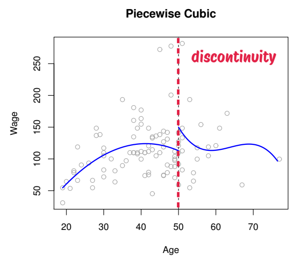


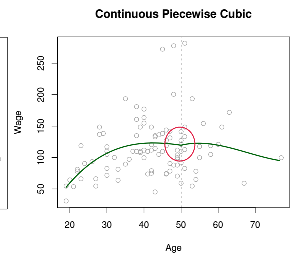
Regression Splines

Example: Wage (ISLR2)

Piecewise cubic polynomial with a single knot placed a age = 50. Constraints:

1. $f_1(\text{age} = 50) = f_2(\text{age} = 50)$





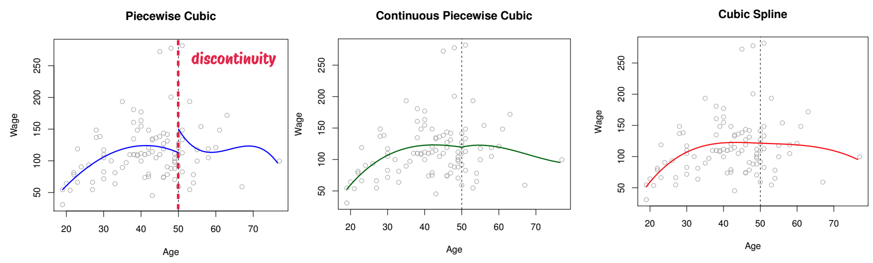
Regression Splines

Example: Wage (ISLR2)

Piecewise cubic polynomial with a single knot placed a age = 50. Constraints:

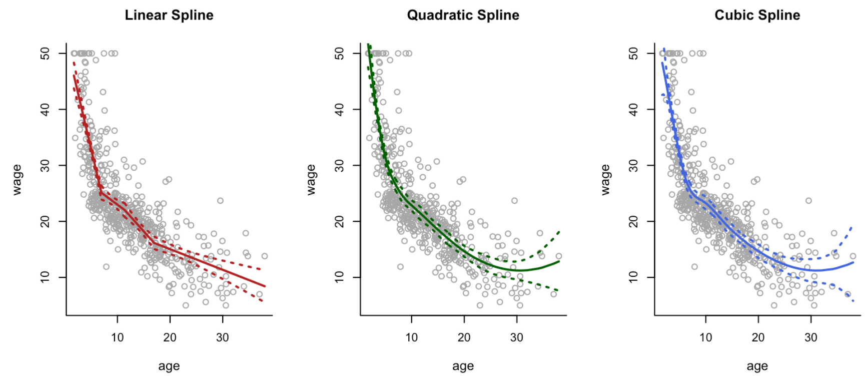
1. $f_1(\text{age} = 50) = f_2(\text{age} = 50)$
2. $f_1'(\text{age} = 50) = f_2'(\text{age} = 50)$
3. $f_1''(\text{age} = 50) = f_2''(\text{age} = 50)$

`splines are meant to be continuous and have continuous derivatives!`



Regression Splines

Example: Wage (ISLR2)



Regression Splines

Constraints and Degrees of Freedom

- In the previous example, we started with a cubic piecewise polynomial with 8 unconstrained parameters, so we started with 8 **degrees of freedom** (df)
- We initially imposed one constraint, which restricted one parameter, so we lost a degree of freedom $8 - 1 = 7$
- With the further two constraints: $8 - 3 = 5$ df
- In general, a cubic spline with K knots has $4 + K$ degrees of freedom. In R we can we can specify either the number of knots or just the degrees of freedom.

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

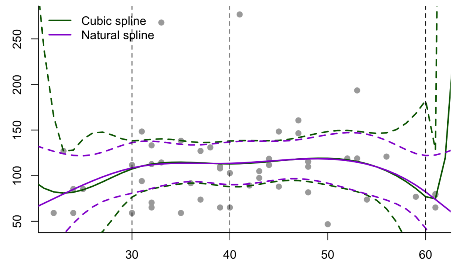
Natural Splines

- Regression splines have high variance at the outer range of the predictor (the tails)
- The confidence intervals at the tails can be wiggly (especially for small samples)

Natural splines are extensions of regression splines which remedy these problems

Two additional constraints at each boundary region:

1. The spline function is constrained to be close to linear when $X <$ smallest knot
2. The spline function is constrained to be close to linear when $X >$ largest knot



How Many Knots?

- Provided there is evidence from the data we can do it empirically:
  - Place knots where it is clearly obvious there is a distributional shift in direction
  - Place more knots on regions where we see more variability
  - Place fewer knots in places which look more stable
- Alternatively, we can place knots in a uniform fashion (25th, 50th, 75th percentiles)

29

Smoothing Splines

```
a smoothing spline is a non parametric approach
designed to balance fit with smoothness
```

30

Smoothing Splines

- Unlike regression splines and natural splines, there are no knots!
- The discrete problem of selecting a number of knots into a continuous penalization problem
- We seek a function g among all possible functions (linear + non-linear) which minimizes

$$\text{model fit} + \text{roughness penalty term} = \underbrace{\sum_{i=1}^{n}(y_i - g(x_i))^2}_{\text{model fit term}} + \lambda \int \underbrace{(g''(t))^2 dt}_{\text{catches non-linearities}}$$

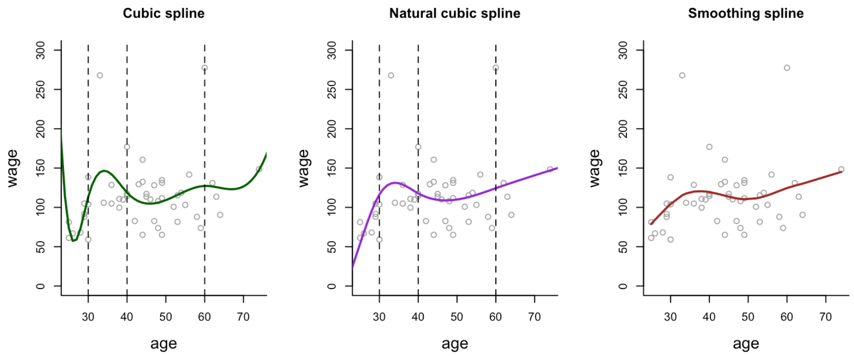
- The function g that minimizes the above quantity is called a **smoothing spline**
- $\lambda \geq 0$ is the tuning penalty parameter, also called **roughness penalty**
  - when $\lambda = 0$ we get an extremely wiggly non-linear function g (completely useless)
  - as λ increases, the function becomes smoother
  - theoretically: when $\lambda \to \infty$, g'' is zero everywhere $\Longrightarrow g(X) = \beta_0 + \beta_2 X$ i.e. linear model
- A smoothing spline is the natural cubic spline whose knots are at all the data points $x_1, x_2, x_3, \ldots, x_n$

31

Cubic vs. Natural vs. Smoothing Splines

Example: Wage (ISLR2)

Training data = 50

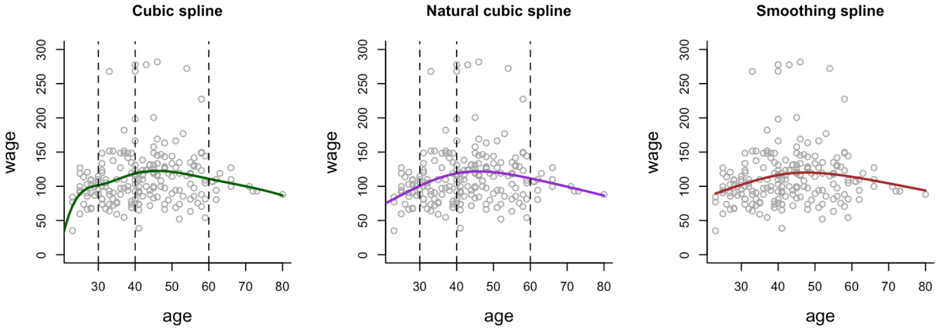


32

Cubic vs. Natural vs. Smoothing Splines

Example: Wage (ISLR2)

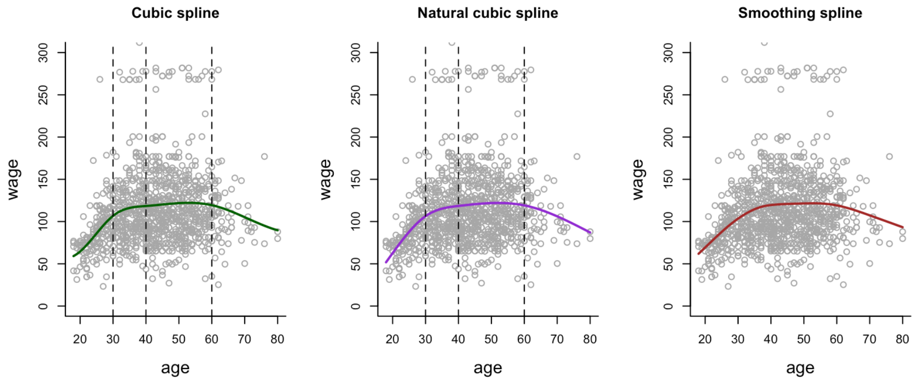
Training data = 200



Cubic vs. Natural vs. Smoothing Splines

Example: Wage (ISLR2)

Training data = 1000



Cubic vs. Natural vs. Smoothing Splines

Criterion	Polynomial Splines	Natural Splines	Smoothing Splines
Flexibility	High with more knots	Moderate	High, controlled by λ
Boundary Behavior	May behave erratically	Linear at boundaries	Smooth, but depends on λ
Noise Handling	Poor, sensitive to noise	Moderate	Excellent, balances fit and smoothness
Interpretability	Good for low degree	Good	Moderate, influenced by λ
Knot Selection	User-defined	User-defined	Not required
Computation	Fast	Fast	Slower for large data

Generalized Additive Models (GAMs)

```
GAM models a response as the sum of smooth,
flexible functions of each predictor
```

Generalized Additive Models (GAMs)

GAMs provide a general framework for extending a standard linear model: allowing non-linear functions of each of the variables, while maintaining additivity

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + f_3(X_3) + \cdots + f_p(X_p) + \epsilon$$

each linear component $\beta_j X_j$ can be replaced by smooth non-linear function $f_j(X_j)$

For example, a GAM may include
- non-linear polynomial method for continuous predictors
- step functions which are more appropriate for categorical predictors
- linear models if that seems more appropriate for some predictors

Generalized Additive Models (GAMs)

General model:

$$Y_i = f(x_{i,1}, \ldots, x_{i,p}) + \varepsilon_i$$

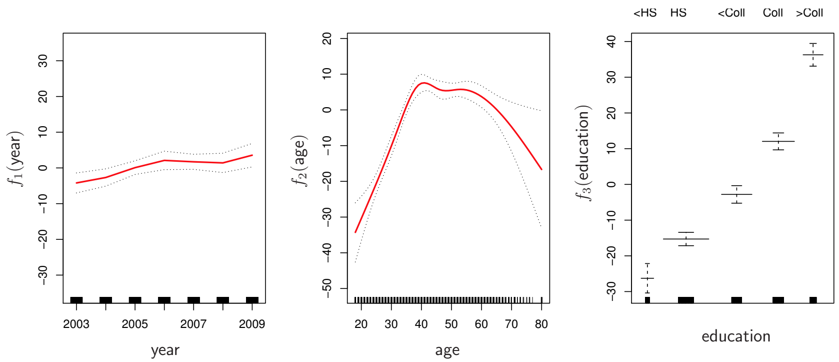
Examples:
- $f(x_1, x_2, x_3) = x_1 + x_1^2 + x_2 + x_2^2 + x_1 x_2 + \sin^2(x_3)$
- $f(x_1, x_2) = \pi + e^{5x_1} + \log(2x_2)$
- $f(x_1, x_2, x_3) = 1 + \log(0.5x_1^2) - x_2 + x_3$

Generalized Additive Models (GAMs)

Example: Wage (ISLR2)

the first two functions are natural splines in year and age

the third function is a step function, fit to the qualitative variable education



Generalized Additive Models (GAMs)

\+ Very flexible in choosing non-linear models and generalizable to different types of responses.
\+ Because of the additivity we can still interpret the contribution of each predictor while considering the other predictors fixed.
\+ GAMs can outperform linear models in terms of prediction.
\+ Built on the framework of GLMs, so can handle different response distributions

\- Additivity is convenient but it is also one of the main limitations of GAMs (independent contributions of predictors)
\- Spline fitting and penalization can be computationally intensive for large data.
\- GAMs might miss non-linear interactions among predictors.

This Week's Practical

Hands on modeling non-linearity