Nonlinear Methods

Yudong Chen
School of ORIE, Cornell University
ORIE 4740 Lec 18–19

Recap: What we covered so far

- Concepts: model flexibility; bias-variance tradeoffs
- Linear regression: fitting and evaluation models
- Classification: Logistic regression; KNN
- Model selection and regularization: subset selection; Ridge; Lasso; principal component regression
- Unsupervised techniques: PCA; *k*-means and hierarchical clustering
- Cross-validation

Recap: Supervised vs. Unsupervised

Supervised learning:

- Regression
- Classification
- Regularization & variable selection: apply to both
- CV: estimate test errors to choose models (tunning parameters)

Unsupervised learning:

- PCA
- Clustering

Recap: Linear vs. Nonlinear

Linear techniques:

- Linear regression
- Logistic regression
- *k*-means
- PCA

Simple extensions of linear techniques:

- Adding high-order and interaction terms
- Converting to dummy variables

Nonlinear techniques:

- KNN

Next:

- More extensions to linear & logistic regression
- Decision Trees & Random Forest

Beyond Linear Regression and Logistic Regression

- Nonlinear models with 1 predictor: $Y = f(X)$
  - The basis function approach
  - Regression Splines
  - Smoothing Splines
  - Local Regression (not covered)

- Nonlinear models with p predictors: $Y = f(X_1, X_2, \ldots, X_p)$
  - Generalized Additive Models (GAMs)

The Basis Function Approach (ISLR 7.1–7.3)

Linear regression: $Y \approx \beta_0 + \beta_1 X$

Logistic regression: $\log\left(\frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)}\right) \approx \beta_0 + \beta_1 X$

Adding high order terms:

$$Y \quad \text{or} \quad \text{log-odds}(Y) \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots$$

Logarithmic terms:

$$\cdots\cdots \approx \beta_0 + \beta_1 \log(X)$$

More generally:

$$Y \quad \text{or} \quad \text{log-odds}(Y) \approx \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \cdots + \beta_K b_K(X)$$

- $b_1(\cdot), \ldots, b_K(\cdot)$: basis functions (pre-specified)

Polynomial Basis Functions

$$Y \quad \text{or} \quad \text{log-odds}(Y) \approx \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \cdots + \beta_K b_K(X)$$

Polynomial functions:

$$b_j(x) = x^j, \quad j = 1, \ldots, K$$

This leads to a polynomial model

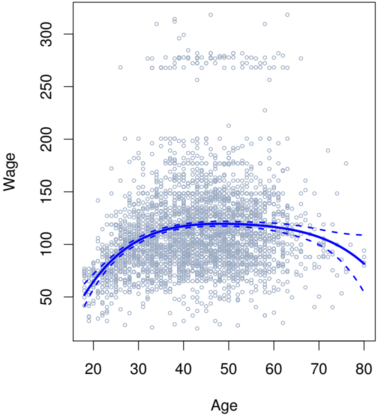
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_K X^K$$

Example: Wage Dataset

y_i = wage of individual i $\qquad$ x_i = age of individual i

Regression with polynomial basis functions up to degree 4:

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$$



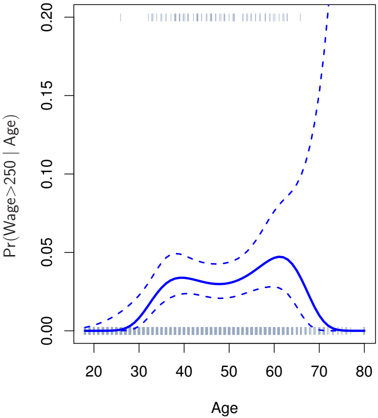
- Dotted lines: 95% confidence intervals of $\hat{y}_i$

Example: Wage Dataset

y_i = wage of individual i $\qquad$ x_i = age of individual i

Classification with polynomial basis functions up to degree 4:

$$\log\left[\frac{\hat{\Pr}(y_i > 250)}{1 - \hat{\Pr}(y_i > 250)}\right] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$$



Step Basis Functions

$$Y \quad \text{or} \quad \text{log-odds}(Y) \approx \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \cdots + \beta_K b_K(X)$$

Step functions: Given knots $c_1, c_2, \ldots, c_K$

$$\begin{aligned}
b_1(x) &= C_1(x) \triangleq I(c_1 \le x < c_2) \\
b_2(x) &= C_2(x) \triangleq I(c_2 \le x < c_3) \\
&\vdots \\
b_{K-1}(x) &= C_{K-1}(x) \triangleq I(c_{K-1} \le x < c_K) \\
b_K(x) &= C_K(x) \triangleq I(c_K \le x)
\end{aligned}$$

This leads to a piecewise-constant model

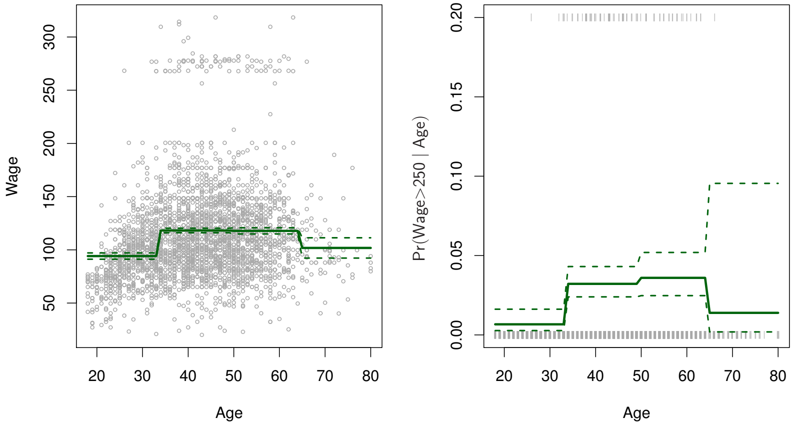
- knots need to be pre-specified (often not clear how to do so)

Example: Wage Dataset

$y_i =$ wage of individual i $\qquad x_i =$ age of individual i

Use step basis functions with 2 knots:

$$y_i \text{ or } \log\left[\hat{\Pr}(y_i > 250)/(1 - \hat{\Pr}(y_i > 250))\right] \approx \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i)$$



The Basis Function Approach

$$Y \quad \text{or} \quad \text{log-odds}(Y) \approx \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \cdots + \beta_K b_K(X)$$

Fitted by least squares

Can use all the tools from linear regression:

- Standard errors & confidence intervals for $\hat{\beta}_j$
- p-values for each $\hat{\beta}_j$
- p-values for the entire model

Other choices of basis functions:

- $b_1(x) = \sqrt{x}$
- $b_1(x) = \log(x)$
- Based on wavelets or Fourier series (not covered)
- Regression Splines (next)

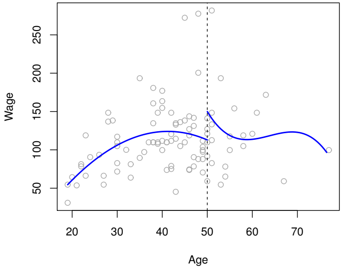
Regression Splines (ISLR 7.4)

Using step functions, we fit a piecewise constant model

$$Y \approx \beta_0 + \beta_1 C_1(X) = \begin{cases} \beta_0 & \text{if } X < c \\ \beta_0 + \beta_1 & \text{if } X \geq c \end{cases}$$

More generally, we can fit a piecewise polynomial model

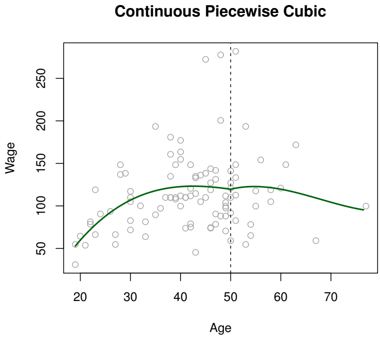
$$Y \approx \begin{cases} \beta_{01} + \beta_{11}X + \beta_{21}X^2 + \beta_{31}X^3 & \text{if } X < c \\ \beta_{02} + \beta_{12}X + \beta_{22}X^2 + \beta_{32}X^3 & \text{if } X \geq c \end{cases}$$

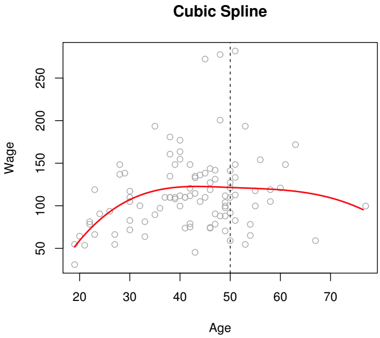


- 8 degrees of freedom (too flexible)

Regression Splines

Regression splines:
Piecewise polynomial models that are continuous and smooth at the knots
(smoothness = continuity of derivatives)





Most popular: Cubic splines

- Continuous piecewise cubic models with continuous first two derivatives
- K knots: $K + 4$ degrees of freedom (instead of $4K + 4$)
- Reduce flexibility/variance; increase bias

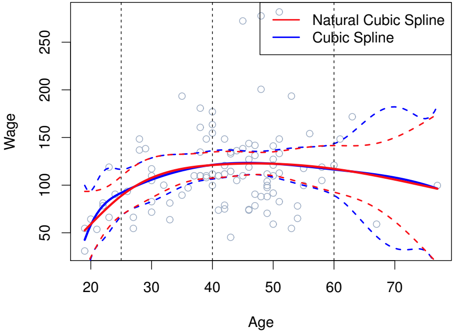
Cubic Splines

- A cubic splines with K knots ($K + 4$ DF) can be written as

$$Y \approx \beta_0 + \beta_1 b_1(X) + \cdots + \beta_{K+3} b_{K+3}(X)$$

with appropriate basis functions $b_j(\cdot)$ (cf. ISLR 7.4.3)

- So can be fitted using least squares

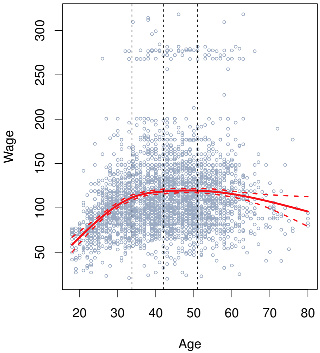


- Natural cubic splines: linear at the boundary (further reduce df/flexibility/variance)

Cubic Splines: Choosing the Knots

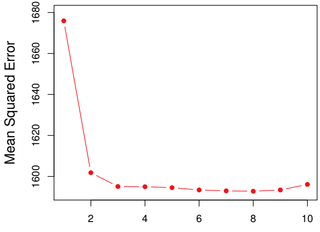
Locations of knots:

- Placed at uniform quantiles of data



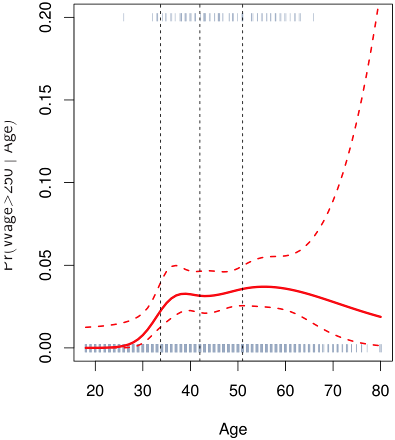
Number of knots:

- Equivalent to choosing degrees of freedom
- Choose the best-looking curve, or...
- By cross-validation



Cubic Splines

Apply to classification (logistic regression) as well



Polynomial Regression vs. Cubic Splines

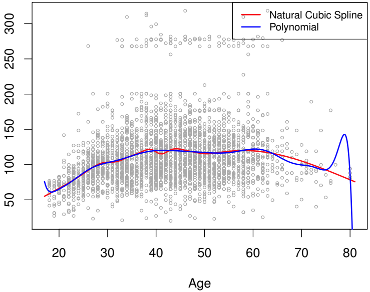
Polynomial regression (the basis function approach with polynomial basis)

$$Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_K X^K$$

- Flexibility/DF determined by degree of polynomials K

Cubic splines

- Flexibility/DF determined by number of knots K



Same degrees of freedom (=15)

Cubic splines often more stable (esp. at the boundaries)

Smoothing Splines (ISLR 7.5)

Recall:

(Cubic) Regression splines:

- Specify knots (or DF)
- Cubic polynomials between knots
- Require smoothness at knots
- Fitting: convert to a basis function model and solved by LS

Smoothing splines: Another way of fitting a smooth curve $g(\cdot)$

- Specify tuning parameter λ
- Find curve as the solution to the optimization problem

$$\min_{g} \quad \underbrace{\sum_{i=1}^{n}(y_i - g(x_i))^2}_{\text{Loss (RSS)}} + \underbrace{\lambda \int g''(t)^2 dt}_{\text{Regularization}}$$

Smoothing Splines

$$\min_{g} \quad \underbrace{\sum_{i=1}^{n}(y_i - g(x_i))^2}_{\text{Loss (RSS)}} + \lambda \underbrace{\int g''(t)^2 dt}_{\text{Regularization}}$$

- Loss term: encourage $g(\cdot)$ to fit data well
- Regularization: encourage smoothness
- $g''(t)$: second derivative
- Small $g''(t)$: less wiggly near t
- Larger $\lambda \Rightarrow$ Smaller $g''(t) \Rightarrow g(\cdot)$ more smooth

The optimal solution

- Can show: the optimal $g(\cdot)$ is a natural cubic spline
- with knots at $x_1, x_2, \cdots, x_n$
- n knots , but less than $n + 4$ DF (b/c of λ)

Smoothing Splines: Choosing λ

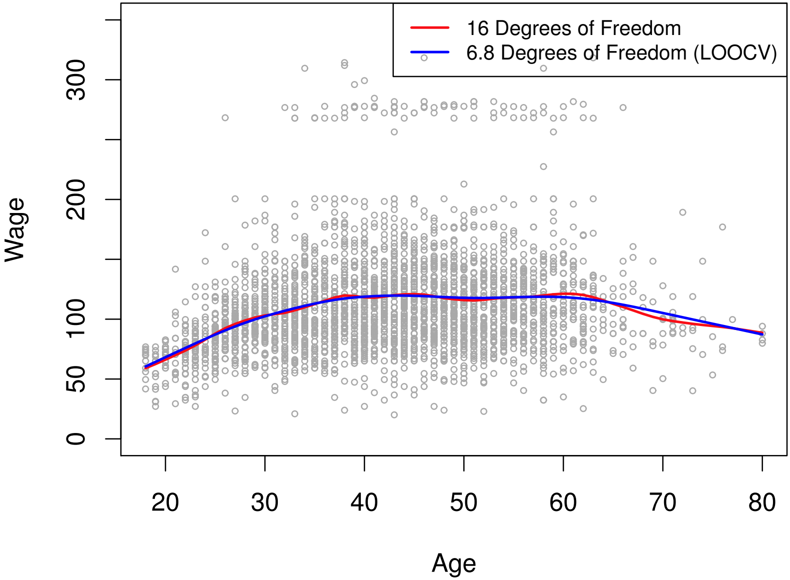
$$\min_g \quad \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

(Recall: In regression splines, flexibility determined by $\#$ knots K, or DF $K+4$)

For smoothing splines:

- Flexibility determined by λ
- Corresponding to an effective degree of freedom, df_λ
- Closed form expression for df_λ (cf. ISLR 279)
- Choose λ (or df_λ) by CV
- LOOCV can be done very efficiently

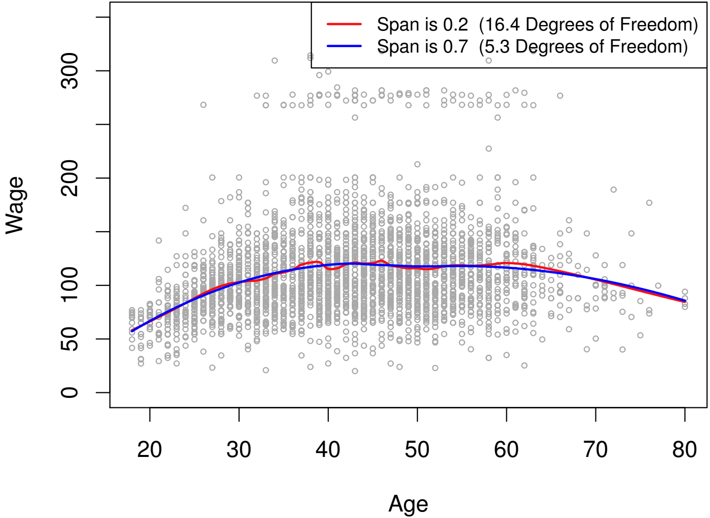
Smoothing Splines: Choosing λ



Local Regression (Not covered; ISLR 7.6)

A third way of fitting smooth curves

- Flexibility determined by a tuning parameter s (span)
- Corresponding to some effective DF



Mini Summary

1 predictor: $Y = f(X)$

- Basis function approach: $f(X) = \sum_j \beta_j b_j(X)$
- Regression Splines: $f(X) =$ piecewise polynomials joint smoothly
- Smoothing Splines: $f(X) =$ solution to $f''(\cdot)$-regularized least squares
- Local Regression (not covered)

p predictors: $Y = f(X_1, X_2, \ldots, X_p)$

- Generalized Additive Models (GAMs)

Generalized Additive Models (ISLR 7.7)

Recall: Multiple linear regression

$$Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Generalized Additive Model: Maintains only additivity

$$Y \approx \beta_0 + f_1(X_1) + \cdots + f_p(X_p)$$

- $f_j(\cdot)$: Any of the univariate nonlinear functions we just learned
- E.g. polynomials, linear combination of basis functions, cubic/smoothing splines
- Build multivariate nonlinear models by adding up univariate ones

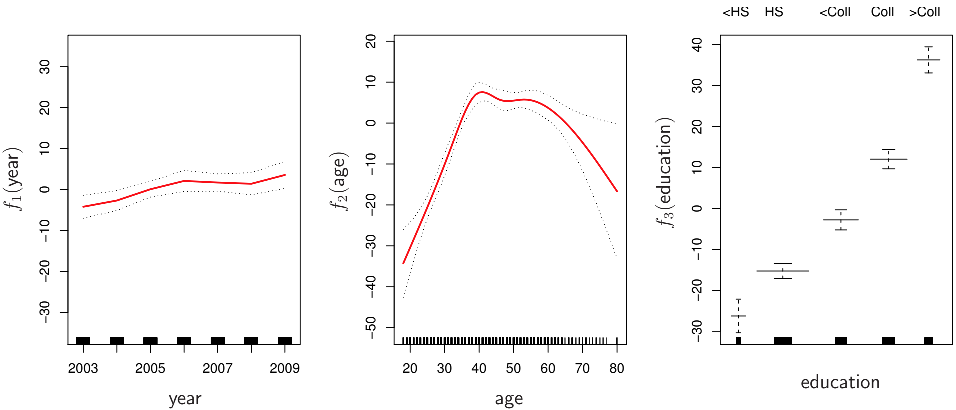
Example: Wage Dataset

Fit a GAM of the form

$$\text{wage} \approx \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$

where

- $f_1(\cdot)$, $f_2(\cdot)$: natural cubic splines
- education: categorical w/ 5 levels `<HS, HS, <Coll, Coll, >Coll`
- $f_3(\cdot) =$ a different value for each level of education
  - i.e., encode education w/ 4 four dummy variables and fit a usual linear model



30
20
10
0
−10
−20
−30
$f_1(\text{year})$
2003
2005
2007
2009
year
20
10
0
−10
−20
−30
−40
−50
$f_2(\text{age})$
20
30
40
50
60
70
80
age
<HS
HS
<Coll
Coll
>Coll
40
30
20
10
0
−10
−20
−30
$f_3(\text{education})$
education

GAMs for Classification

Recall: Logistic regression

$$\log\left(\frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)}\right) \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Logistic regression GAM:

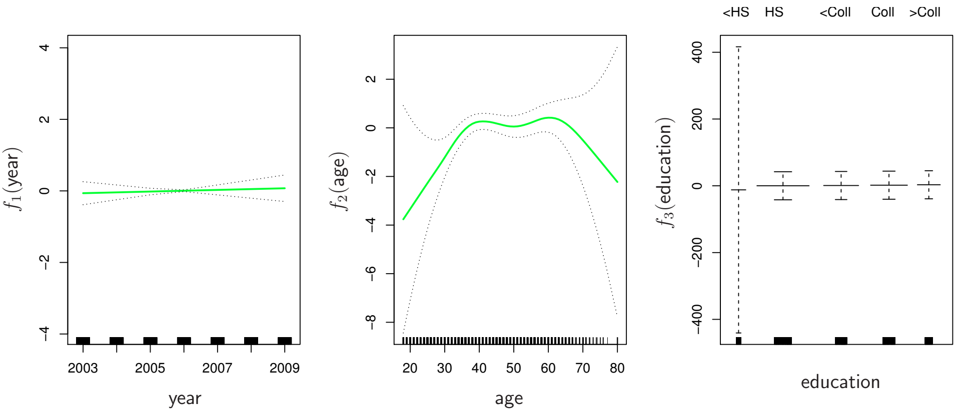
$$\log\left(\frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)}\right) \approx \beta_0 + f_1(X_1) + \cdots + f_p(X_p)$$

Example: Wage Dataset
Fit a GAM of the form

$$\log\left(\frac{\Pr(\text{wage} > 250)}{\Pr(\text{wage} \leq 250)}\right) \approx \beta_0 + \beta_1 \times \text{year} + f_2(\text{age}) + f_3(\text{education})$$

where

- $f_2(\cdot)$: smoothing splines with $df = 5$
- $f_3(\cdot)$ constant for each level of education



$f_1(\text{year})$
year
2003
2005
2007
2009
$f_2(\text{age})$
age
20
30
40
50
60
70
80
$f_3(\text{education})$
education
<HS
HS
<Coll
Coll
>Coll

GAMs: Pros and Cons

$$Y \quad \text{or} \quad \text{log-odds}(Y) \;\approx\; \beta_0 + f_1(X_1) + \cdots + f_p(X_p)$$

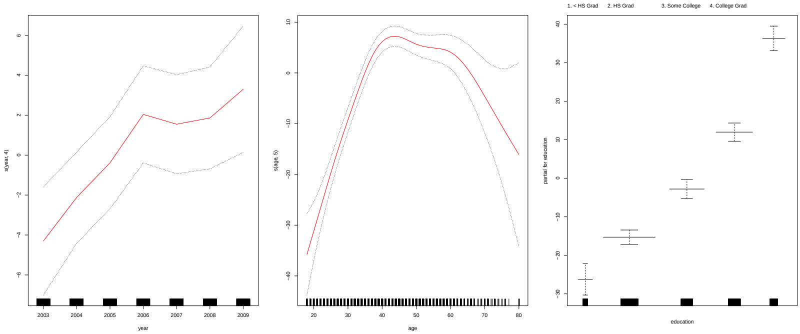
- Combine simple univariate nonlinear models $f_j(\cdot)$ to build p-variate models
- Flexible choices for each $f_j(\cdot)$
- (Natural) Cubic Spline is a popular choice
- Control flexibility by specifying degrees-of-freedom

- Interaction/synergy effects b/w predictors not captured

Nonlinear Modeling in R (ISLR 7.8)

- GAM with smoothing splines

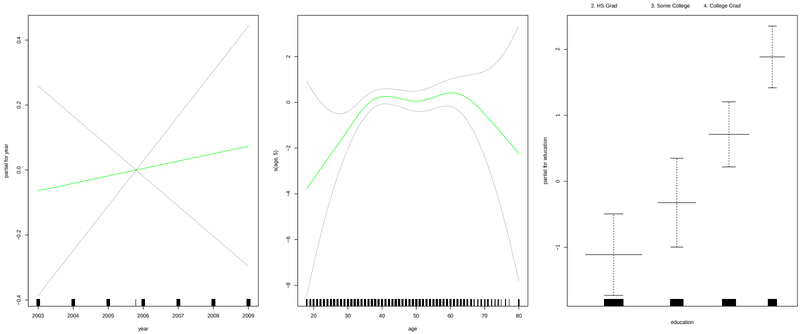
```
> library(gam)
> gam2 = gam(wage~s(year,4)+s(age,5)+education, data=Wage)

> plot(gam2, se=TRUE, col="red")
```



Nonlinear Modeling in R (ISLR 7.8)

- Logistic regression GAM with smoothing splines
Excluding observations with less than a high school education

```
> library(gam)
> gam.lr = gam(I(wage>250)~year+s(age,5)+education, family=
+ binomial, data=Wage, subset=(education!="1. < HS Grad"))
> plot(gam.lr, se=TRUE, col="green")
```



Nonlinear Modeling Summary

1 predictor: $Y = f(X)$

- Basis function approach: $f(X) = \sum_j \beta_j b_j(X)$
- Regression Splines: $f(X) =$ piecewise polynomials joint smoothly
- Smoothing Splines: $f(X) =$ solution to $f''(\cdot)$-regularized least squares
- Local Regression

p predictors: $Y = f(X_1, X_2, \ldots, X_p)$

- Generalized Additive Models (GAMs)

$$Y \quad \text{or} \quad \text{log-odds}(Y) \;\approx\; \beta_0 + f_1(X_1) + \cdots + f_p(X_p)$$

where $f_j(\cdot)$ is a polynomial, step function, cubic/smoothing spline, local regression, ...