Non-linear Regression: Smoothing Splines and Generalized Additive Models

(ISLR 7.5 - 7.7)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

1. Smoothing Splines
2. Local Regression
3. General Additive Models

Smoothing Splines

Fit a smooths function $g(x)$ that minimizes

$$\underbrace{\sum_{i=1}^{n}[y_i - g(x_i)]^2}_{\text{“RSS”}} + \underbrace{\lambda \int g''(t)^2 dt}_{\text{roughness penalty}}$$

- The first derivative $g'(t)$: the slope of $g(\cdot)$ at t
- The second derivative $g''(t)$: the amount by which the slope is changing; it's a measure of roughness.
- $\int g''(t)^2 dt$: a measure of the total change in the function $g'(t)$, over its entire range.
- An extreme case: if $g(t)$ is a straight line, then $\int g''(t)^2 dt = 0$.

Smoothing Splines

- $\lambda \geq 0$: tuning parameter.
  - The smaller λ, the more wiggly the function, eventually interpolating y_i when $\lambda = 0$.
  - As $\lambda \to \infty$, the function $g(x)$ becomes linear.

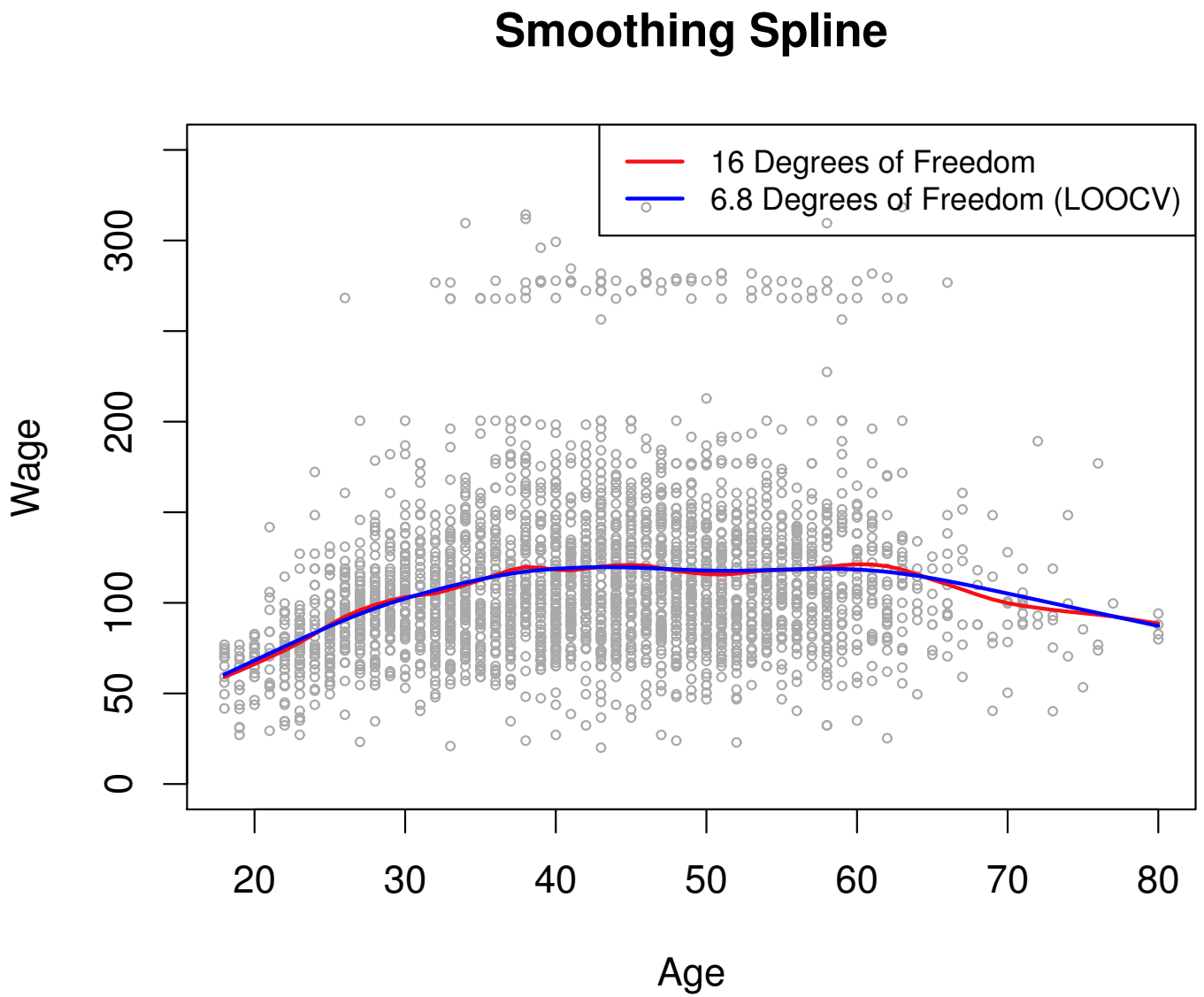
- Smoothing splines solution: a natural cubic spline, with a knot at every unique value of x_i.
  - It is not the same natural cubic spline that one would get if one applied the basis function approach without the penalty.
  - It is a shrunken version of such a natural cubic spline, where the value of the tuning parameter λ controls the level of shrinkage.

Choosing λ

- The tuning parameter λ controls the roughness of the smoothing spline, and hence the effective degrees of freedom.
- Usually degrees of freedom refer to the number of free parameters.
- The effective df may not be an integer (we skip the definition here).
- We can specify df rather than λ!

For smoothing splines, the leave one out cross validation error can be computed with essentially the same cost as computing a single fit.

The Wage data



Local Regression

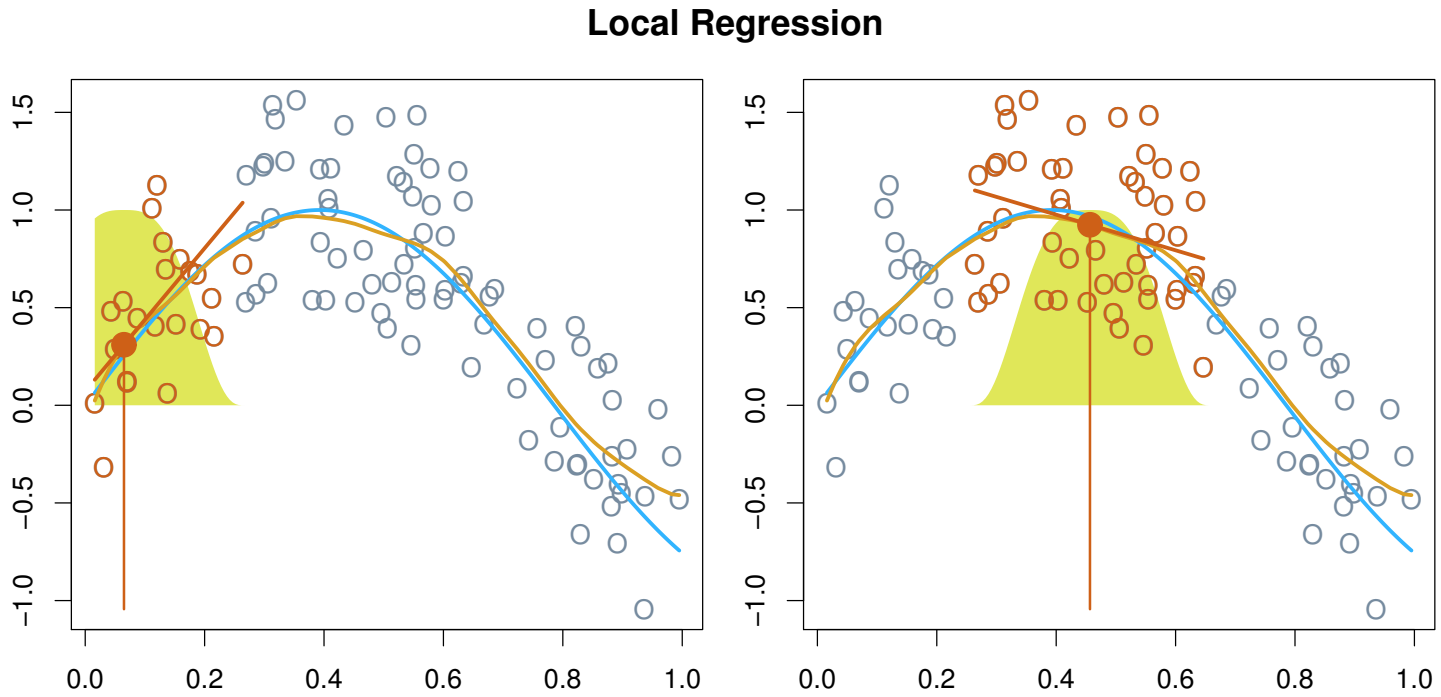
Computes the fit at a target point x_0 using only the nearby training points.

1. Gather the fraction $s = \frac{k}{n}$ of training points x_i closest to x_0.
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood.
   - the closest has the highest weight,
   - the points further from x_0 have lower weights,
   - all but these k nearest neighbors get weight zero.
   - For example, $K(x_i, x_0) \propto e^{-\frac{(x_i - x_0)^2}{2\sigma^2}}$.
3. Fit a weighted least squares regression:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

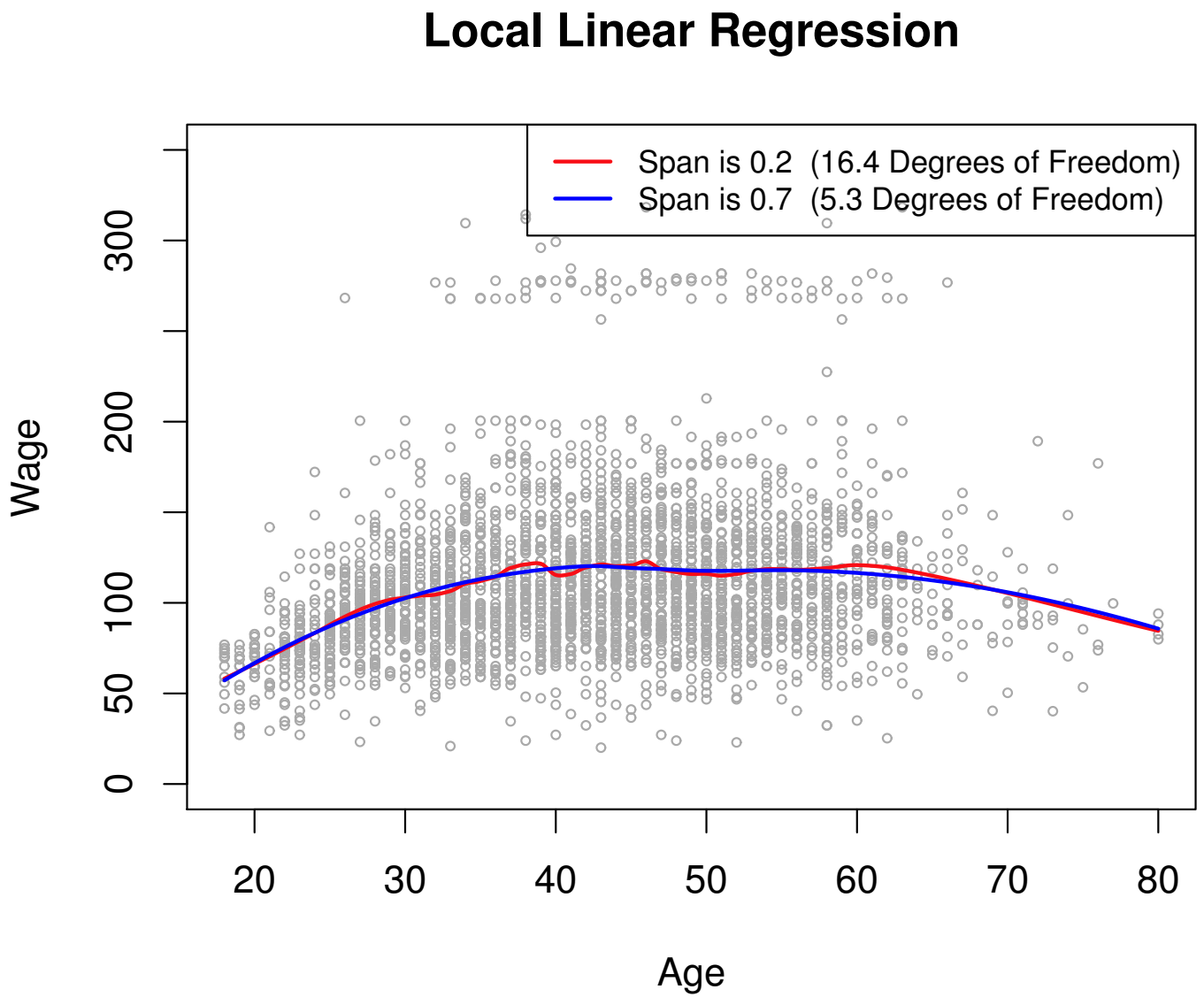
4. The fitted value at x_0 is given by

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$



- Blue curve: true $f(x)$.
- Yellow curve: local regression $\hat{f}(x)$.
- Vertical orange line: left: $x_0 = 0.05$, right: $x_0 = 0.4$
- Orange dots: local points
- Yellow bell-shaped region: weights

The span s controls flexibility



Extension to Multiple X

Now we have learned several approaches for flexibly predicting a response Y on the basis of a single predictor X:

- Polynomial regression
- Step functions
- Regression splines
- Smoothing splines
- Local regression

Next we will using these as building blocks, to flexibly predict Y on the basis of several predictors, $X_1, \ldots, X_p$.

General Additive Models

- We extend the multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- by replacing $\beta_j x_{ij}$ with a smooth nonlinear function $f_j(x_{ij})$

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

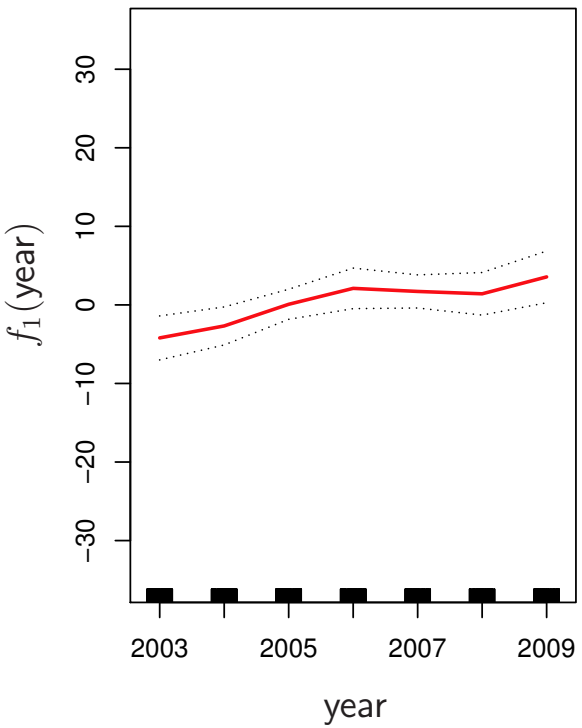
- Allows for flexible nonlinearities in multiple variables, but retains the *additive structure* of linear models.

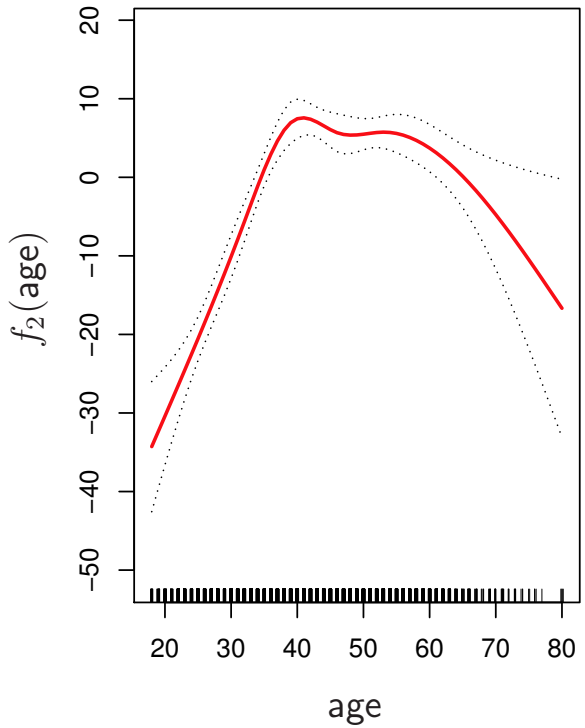
The Wage Example

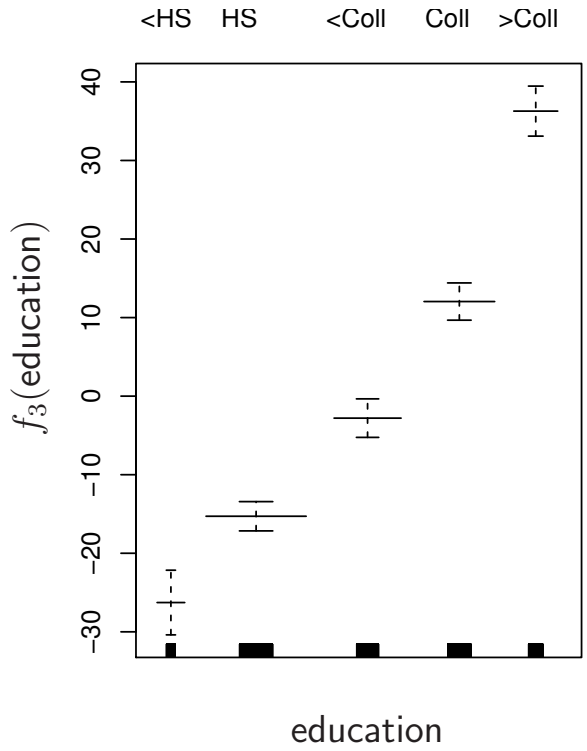
$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$

- Variable `year`, `age`: continuous.
  $f_1(\cdot), f_2(\cdot)$: natural splines.
- Variable `education`: categorical with 5 levels.
  $f_3(\cdot)$: constants for each level (dummy variables).
- The entire model is just a big regression onto spline basis variables and dummy variables.
- Coefficients not that interesting; fitted functions are.

The Wage Example: Regression



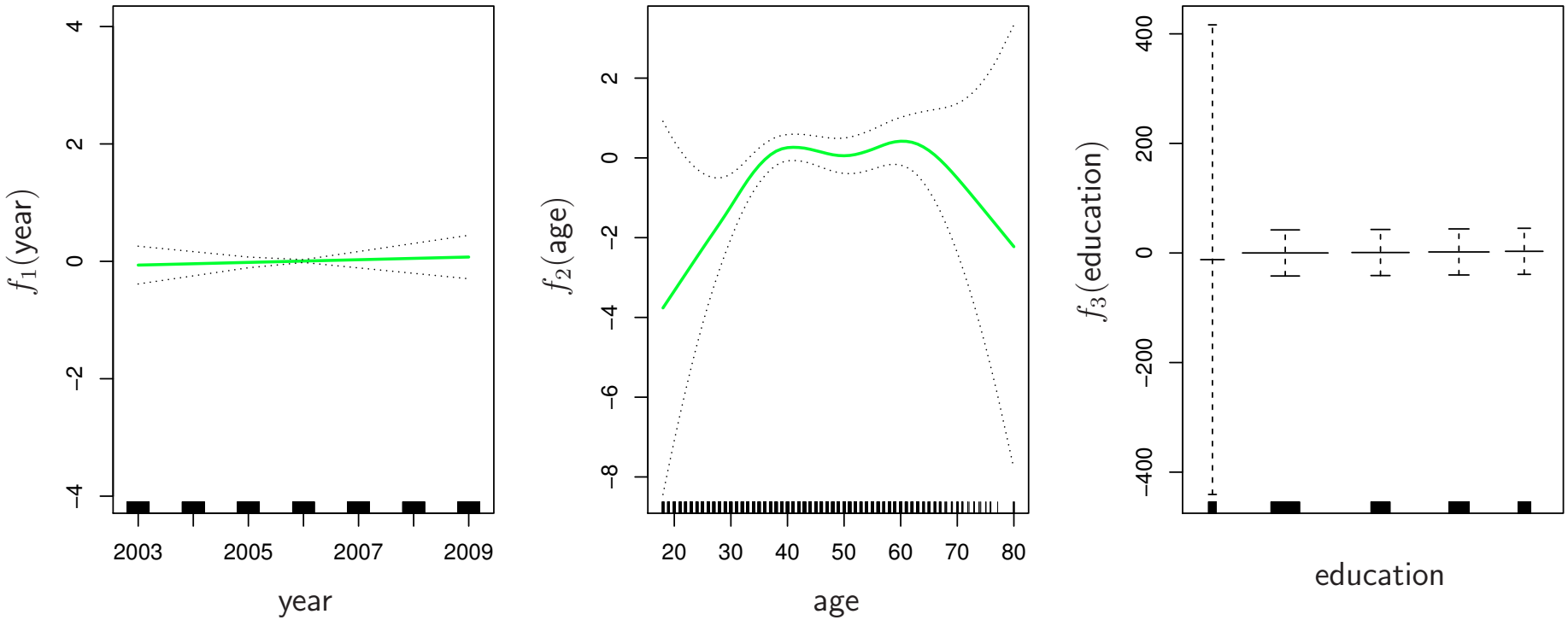




For Classification: Logistic Regression GAM

$$\log\left[\frac{p(x_i)}{1-p(x_i)}\right] = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip})$$

Response: $\mathbf{1}(\texttt{wage} > 250)$



GAM: Pros and Cons

- A non-linear $f_j(\cdot)$ to each X_j
- More accurate predictions (than linear models)
- Easy interpretation:
  Because the model is additive, we can still examine the effect of each X_j on Y individually while holding all of the other variables fixed.
- The smoothness of the function f_j can be summarized via df.

- Restricted to be additive: add interaction terms manually.