Problem Statement

Not all relationships between the features and the outcome can be captured by functions that in essence can be captured in a solely linear model.

This may either lead to non-linear but parametric regression models, for which the structure of the functional relationship is known.

Alternative, semi-parametric models using spline functions in the features, wavelets or Fourier series can be used.

Progression towards Spline Functions

1. Polynomial functions with a given degree. Example 4-degree polynomial:

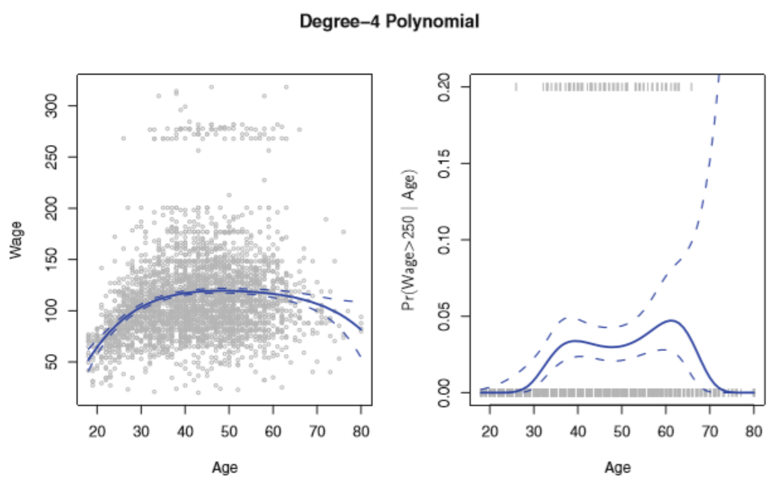

FIGURE 7.1. *The* **Wage** *data. Left: The solid blue curve is a degree-4 polynomial of* **wage** *(in thousands of dollars) as a function of* **age**, *fit by least squares. The dotted curves indicate an estimated 95 % confidence interval. Right: We model the binary event* **wage>250** *using logistic regression, again with a degree-4 polynomial. The fitted posterior probability of* **wage** *exceeding \$250,000 is shown in blue, along with an estimated 95 % confidence interval.*

The functional form is $\mathbf{y} = \beta_0 + \beta_1 \cdot \mathbf{x}^0 + \beta_1 \cdot \mathbf{x}^1 + \beta_2 \cdot \mathbf{x}^2 + \cdots + \boldsymbol{\varepsilon}$. The degrees of freedom consumed by the polynomial are equal to the polynomial degree plus one for the intercept

2. Step Linear Functions. Example: varying intercepts for given of supports of a feature:

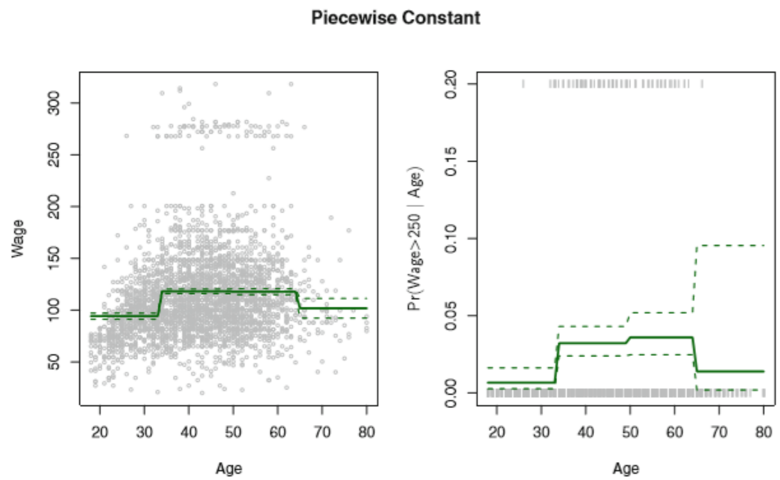

FIGURE 7.2. *The* Wage *data. Left: The solid curve displays the fitted value from a least squares regression of* wage *(in thousands of dollars) using step functions of* age*. The dotted curves indicate an estimated 95 % confidence interval. Right: We model the binary event* wage>250 *using logistic regression, again using step functions of* age*. The fitted posterior probability of* wage *exceeding \$250,000 is shown, along with an estimated 95 % confidence interval.*

The degrees of freedom are equal to the number of knots plus one for the general intercept.

3. Piecewise Polynomials under constraints with one knot

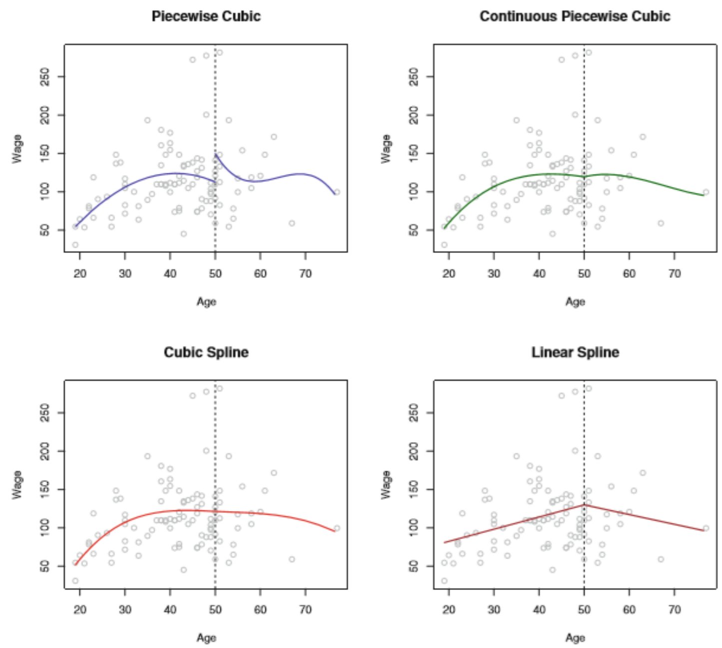

FIGURE 7.3. *Various piecewise polynomials are fit to a subset of the* Wage *data, with a knot at* age=50. Top Left: *The cubic polynomials are unconstrained.* Top Right: *The cubic polynomials are constrained to be continuous at* age=50. Bottom Left: *The cubic polynomials are constrained to be continuous, and to have continuous first and second derivatives.* Bottom Right: *A linear spline is shown, which is constrained to be continuous.*

Cubic splines are cubic polynomials within each its binding knots. These splines, in order to generate smooth graphs, need to satisfy the conditions that [a] at each knots their predicted values are identical to the value of its neighboring polynomials, [b] the slopes are identical and [c] their curvatures are identical. The second derivative only is relevant for polynomials of the third degree.

The degrees of freedom consumed by cubic splines is $df = 4 + K$ where K is the number of knots

4. Natural spline constraint

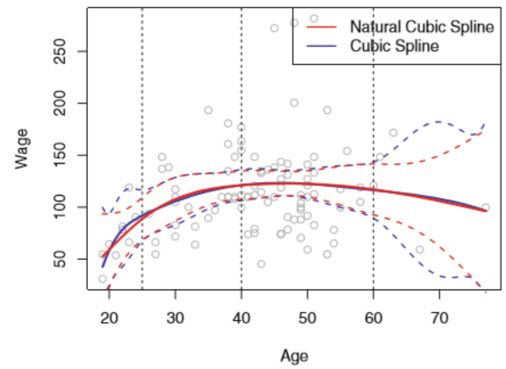

FIGURE 7.4. *A cubic spline and a natural cubic spline, with three knots, fit to a subset of the* **Wage** *data.*

Natural splines control the standard error by adding a linearity constraint at the end of the feature support. These edge constraint reduce the consumed degrees of freedom compared to a cubic spline by two: $df = 4 + K - 2$

5. Comparison of natural splines and polynomials of equal degrees of freedom

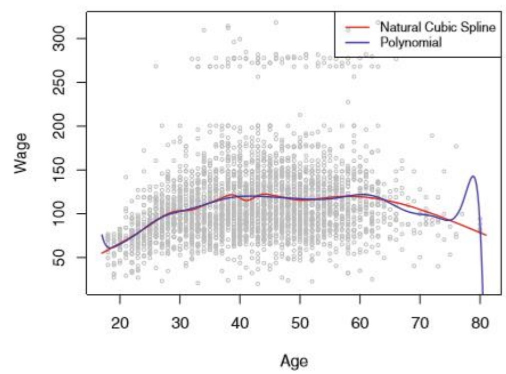

FIGURE 7.7. *On the* **Wage** *data set, a natural cubic spline with 15 degrees of freedom is compared to a degree-15 polynomial. Polynomials can show wild behavior, especially near the tails.*

Polynomial functions are very erratic at the edge of the feature support.

6. Smoothing Splines

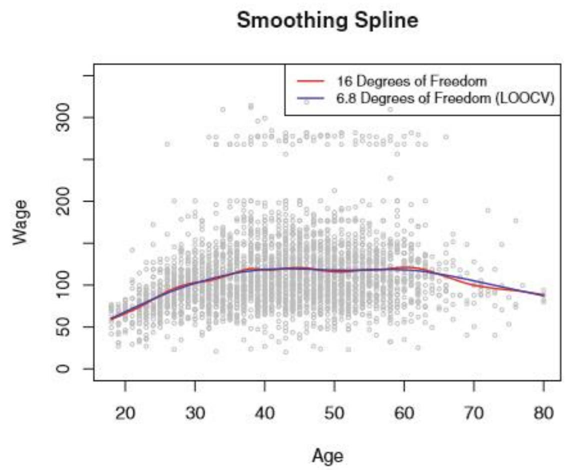

FIGURE 7.8. *Smoothing spline fits to the* `Wage` *data. The red curve results from specifying 16 effective degrees of freedom. For the blue curve, λ was found automatically by leave-one-out cross-validation, which resulted in 6.8 effective degrees of freedom.*

Smoothing splines impose a shrinkage constraint similar to ridge regression:

$$\underbrace{\sum_{i=1}^{n}\left(y_i - g(x_i)\right)^2}_{loss} + \underbrace{\lambda \cdot \int g''(t)^2 \cdot dt}_{penalty}$$

For $\lambda \to \infty$ the function $g(\,)$ is perfectly smooth with zero knots, whereas for $\lambda = 0$ the function $g(\,)$ has a knot at each x_i. The optimal number of knots, i.e., the λ-value, is determined by LOOCV.
The λ-value is inversely proportional to the degrees of freedom lost by the smoother.

7. Generalized Additive Models

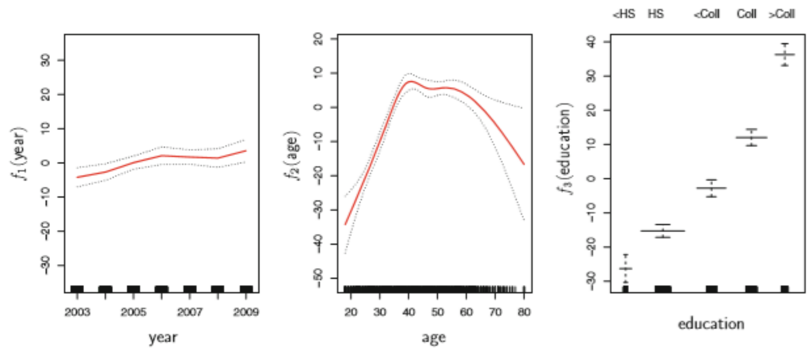

FIGURE 7.11. *For the* Wage *data, plots of the relationship between each feature and the response,* wage*, in the fitted model (7.16). Each plot displays the fitted function and pointwise standard errors. The first two functions are natural splines in* year *and* age*, with four and five degrees of freedom, respectively. The third function is a step function, fit to the qualitative variable* education.

GAM calibrates by *backfitting* one smoother $g_k(\,)$ for each feature $\mathbf{x}_k$ and combines them linearly:

$$\mathbf{y} = \beta_0 + g_1(\mathbf{x}_1) + g_2(\mathbf{x}_2) + \cdots + \boldsymbol{\varepsilon}$$