Lecture 17: Moving Beyond Linearity/Nonparametric Regression

Nayel Bettache

Department of Statistics and Data Science, Cornell University

Moving Beyond Linearity

The linearity assumption is almost always an approximation, and sometimes a poor one.

We can improve upon least squares using regularization $\rightarrow$ reducing the complexity of the linear model. But we are still using a linear model.

We consider the following extensions to relax the linearity assumption.

- Polynomial regression
- Step functions
- Regression splines
- Smoothing splines
- Local regression
- Generalized additive models

Polynomial Regression

- The **polynomial regression**

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d + \epsilon_i,$$

  where ϵ_i is the error term.

- The coeffcients can be estimated using least squares linear regression.

- Not really interested in the coefficients; more interested in the fitted function values at any value x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + ... + \hat{\beta}_d x_0^d.$$

- There is a simple formula to calculate the pointwise standard error of $\hat{f}(x_0)$. The pointwise confidence interval is $\hat{f}(x_0) \pm 2 \cdot se[\hat{f}(x_0)]$.

- We either fix the degree d at some reasonably low value (≤ 3 or 4), else use cross-validation to choose d.

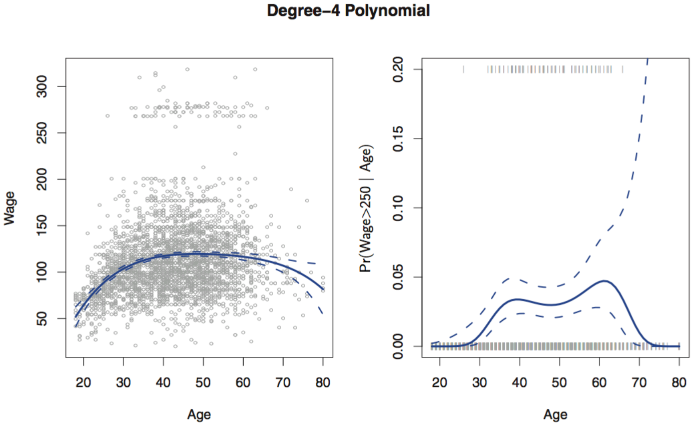
Polynomial Regression

- The polynomial regression can be used for logistic regression

$$logit\ P(y_i = 1|x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d.$$

- Can do separately on several variables (see GAMs later).

The Wage Data



Left: The solid blue curve is a degree-4 polynomial of wage 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, with a degree-4 polynomial.

Step Functions

- The polynomial regression imposes a global structure on the non-linear function of X.
- The **step function** approach avoids such a global structure. Here we break the range of X into bins, and fit a different constant in each bin. Define

$$C_0(X) = I(X < c_1),\ C_1(X) = I(c_1 \leq X < c_2), \ldots, C_K(X) = I(c_K \leq X),$$

where $c_1, c_2, ..., c_K$ are K cutpoints in the range of X. Basically, $C_0(X), \ldots, C_K(X)$ are $K + 1$ dummy variables, and the summation is 1.

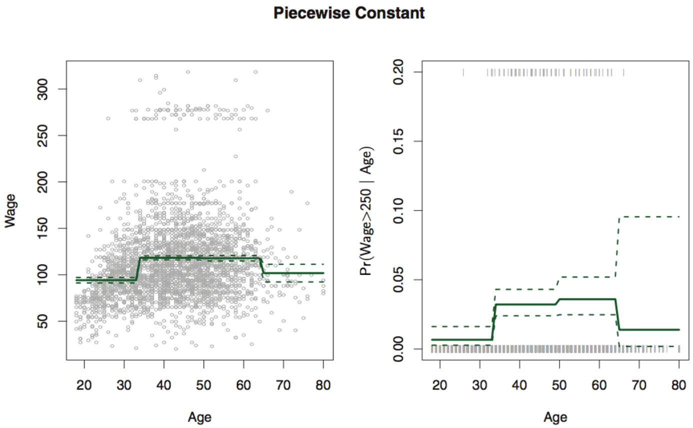
- We then use least squares to fit a linear model using $C_1(X)$, $C_2(X), \ldots, C_K(X)$ as predictors

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + ... + \beta_K C_K(x_i) + \epsilon_i,$$

where ϵ_i is the error term. (Why there is no $C_0(X)$ in the model?)

- β_j represents the average increase in the response for X in $c_j \leq X < c_{j+1}$ relative to $X < c_1$.

The Wage Data



Left: The solid blue curve is a step function of wage 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, with the step function.

Pros and Cons of Step Function

- The step function approach is widely used in biostatistics and epidemiology among other areas, because the model is easy to fit and the regression coefficient has a natural interpretation.
- However, unless there are natural breakpoints in the predictors, piecewise-constant functions can miss the trend of the curve. The choice of breakpoints can be problematic.
- Polynomial and piecewise-constant regression models are in fact special cases of a **basis function** approach,

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i,$$

  where $b_1(X)$, $b_2(X)$, . . . , $b_K(X)$ are known basis functions.
- In the following, we investigate a very common choice for a basis function: **regression splines**.

Piecewise Polynomials

- Instead of a single polynomial in X over its whole domain, we can rather use different polynomials in regions defined by knots,

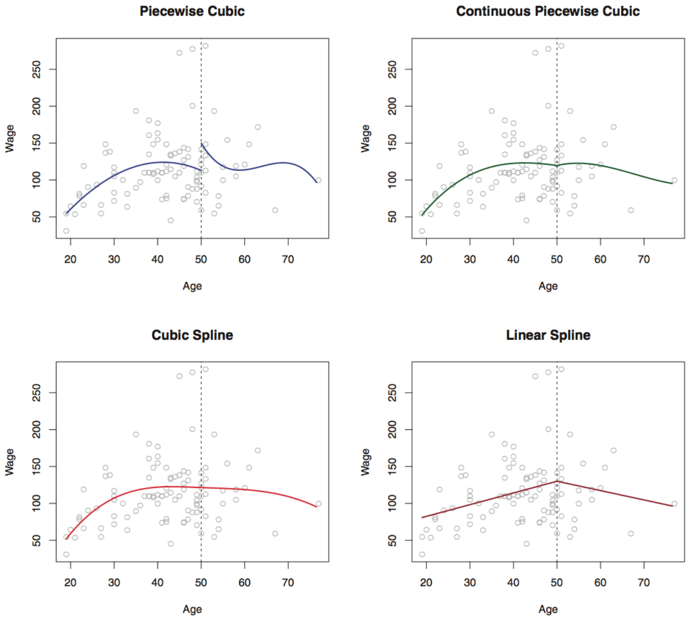
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

- Using more knots leads to a more flexible piecewise polynomial. In general, if we place K different knots throughout the range of X, then we will end up fitting $K + 1$ different cubic polynomials.

- Better to add constraints to the polynomials, e.g. continuity. This leads to **cubic splines**.

- The general definition of a degree-d spline is that it is a piecewise degree-d polynomial, with continuity in derivatives up to degree $d - 1$ at each knot.

The Wage Data



The Spline Basis Representation

- How can we construct the degree-d spline?
- A **linear spline** with knots at $\xi_k, k = 1, ..., K$ is a piecewise linear polynomial continuous at each knot. It is

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + ... + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

  where b_k are basis functions

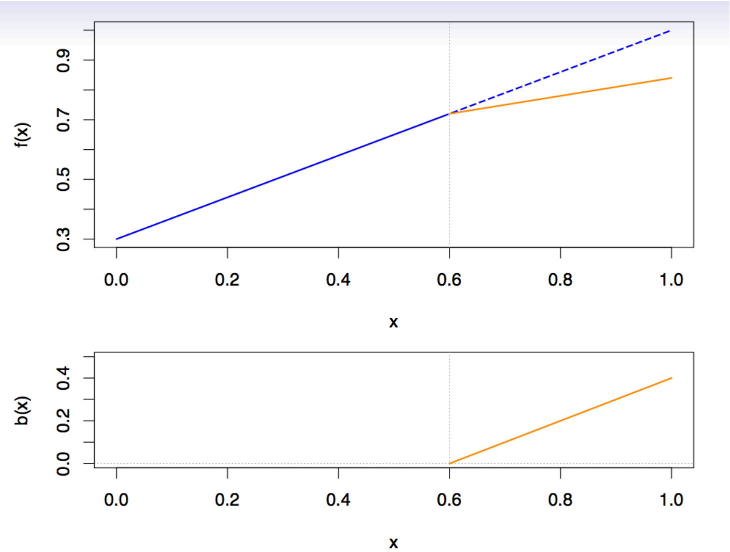
$$b_1(x_i) = x_i, b_{k+1}(x_i) = (x_i - \xi_k)_+, \quad k = 1, ..., K,$$

  here $(\cdot)_+$ means positive part,

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

- What is the interpretation of β_1? (The averaged increase of Y if we increase one unit of X when $X < \xi_1$.)

Linear Splines



Cubic Splines

- A **cubic spline** with knots at $\xi_k, k = 1, ..., K$ is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot. It is

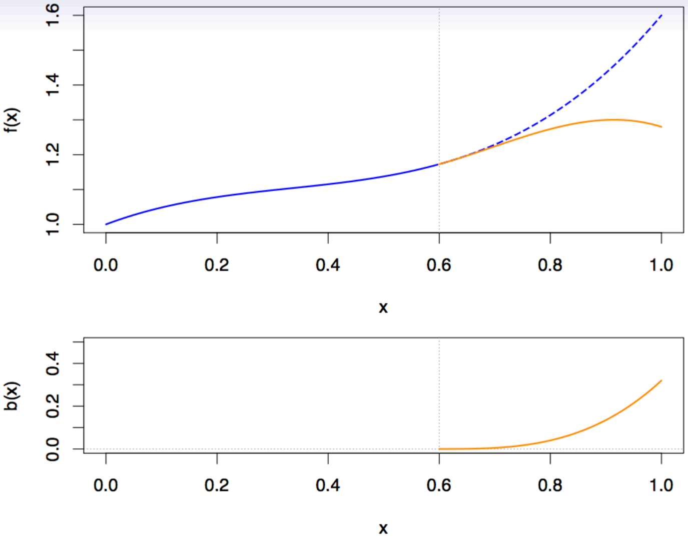
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + ... + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

where b_k are basis functions

$$b_1(x_i) = x_i, b_2(x_i) = x_i^2, b_3(x_i) = x_i^3,$$

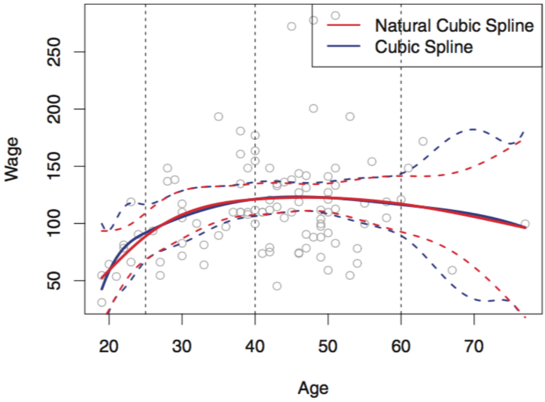
$$b_{k+3}(x_i) = (x_i - \xi_k)_+^3, \quad k = 1, ..., K,$$

Cubic Splines



Natural Cubic Splines

A natural spline is a regression spline with additional boundary constraints: the function is required to be linear at the boundary.



Choosing the Number and Locations of the Knots

- Typically, we place K knots at the corresponding quantiles of the data or place on the range of X with equal space. Usually, the placement of knots is not very crucial.
- We use cross-validation to choose K. Specifically, given a fixed K, we use cross-validation to estimate the test RSS, and then we choose K with smallest estimated test RSS.

Smoothing Splines

- The **smoothing spline** is the minimizer of the following objective function

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

  where λ is a nonnegative tuning parameter.

- The first term $\sum_{i=1}^{n}(y_i - g(x_i))^2$ is RSS which tries to make $g(x)$ match the data at each x_i.

- Broadly speaking, the second derivative of a function is a measure of its roughness. So the second term $\int g''(t)^2 dt$ is a roughness penalty on the entire range of X.

- The tuning parameter λ determines the importance between the model fit and the smoothness of the estimated function (bias-variance trade-offi).

- It can be shown that the minimizer is a shrunken version of the natural cubic spline with knots at $x_1, ..., x_n$. The math is beyond this lecture, and we will not pursue this approach.

Local Regression

Local regression is a different approach for fitting flexible non-linear functions, which involves computing the fit at a target point x_0 using only the nearby training observations.

Algorithm 7.1 *Local Regression At* $X = x_0$

1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.

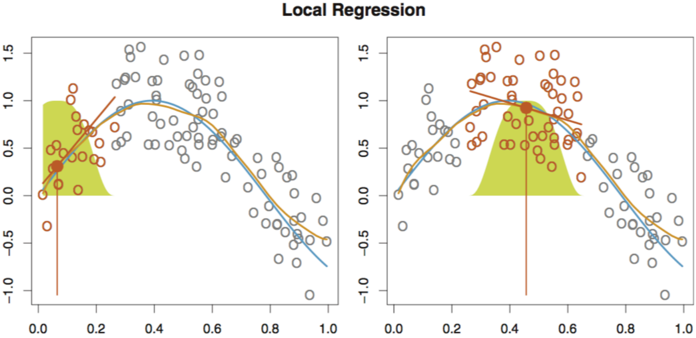
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.

3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{7.14}$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Simulated Example



The blue curve is true $f(x)$, and the light orange curve is the local regression $\hat{f}(x)$. The orange colored points are local to the target point x_0, represented by the orange vertical line. The yellow bell-shape indicates weights assigned to each point, decreasing to zero with distance from the target point. The fit $\hat{f}(x_0)$ at x_0 is obtained by fitting a weighted linear regression (orange line segment), and using the fitted value at x_0 (orange solid dot) as the estimate $\hat{f}(x_0)$.

Local Regression

- The size of the neighborhood (fraction s of training data) is a tuning parameter, which can be chosen by cross-validation.
- When we have two dimensional predictors X1 and X2, we can simply use two-dimensional neighborhoods, and fit bivariate linear regression models using the observations that are near each target point in two-dimensional space.
- However, local regression can perform poorly if p is much larger than about 3 or 4 (known as curse of dimensionality).

Generalized Additive Models

- In the previous sections, we only have a single predictor X.
- **Generalized additive models** (GAMs) provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity,

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

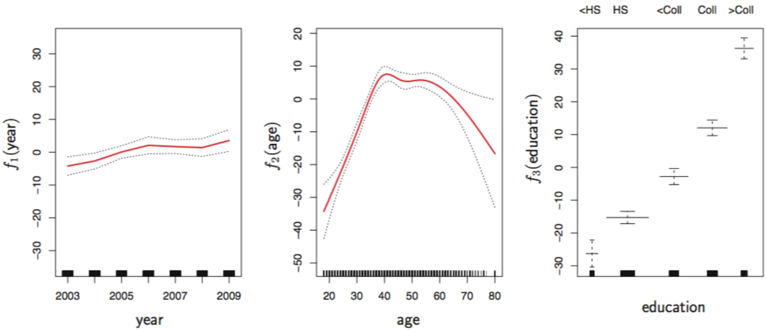
- We can fit GAMs using smoothing splines or other smoothing methods (local regression, regression splines) for a single predictor, via an approach known as backfitting.
- Coefficients not that interesting; fitted functions are.
- Can mix terms – some linear, some nonlinear.
- Can be applied to classification problems

$$logit\ P(y_i = 1|x_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}).$$

Wage Data

Consider the wage data

$$wage = \beta_0 + f_1(year) + f_2(age) + f_3(education) + \epsilon.$$

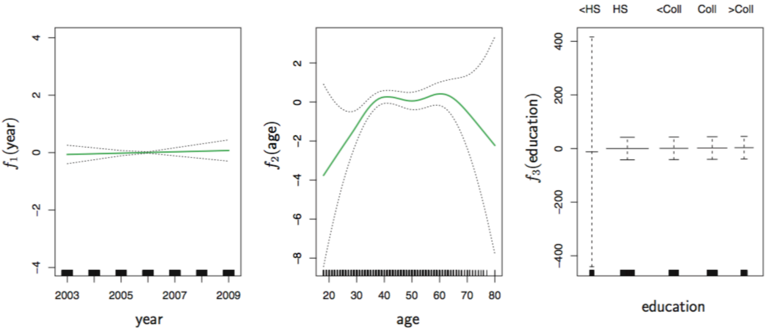


The first two functions are natural splines in year and age, with four and five degrees of freedom, respectively. The third function is a step function, fit to the qualitative variable education.

Wage Data

Consider the wage data

$$logit \; P(wage > 250) = \beta_0 + \beta_1 \times year + f_2(age) + f_3(education).$$



The first function is linear in year, the second function a smoothing spline with five degrees of freedom in age, and the third a step function for education. There are very wide standard errors for the first level <HS of education.

Pros and Cons of GAMs

- GAMs allow us to fit a non-linear f_j to each X_j, so that we can automatically model non-linear relationships that standard linear regression will miss.
- The non-linear fits can potentially make more accurate predictions for the response Y.
- 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.
- It solves the curse of dimensionality.
- However, GAMs fail to incorporate the interaction of variables.