Section 11: More non-linear models

STA 35C – Statistical Data Science III

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 12:10 PM – 1:00 PM, Olson 158
University of California, Davis

Overview

Based on Chapter 7 of ISL book James et al. (2021).

Motivation

Recall regression problem:

$$Y = f(X_1, \dots, X_p) + \varepsilon \tag{1}$$

- So far, we mostly focused on models that assumed that f is a linear function of the predictors $X_1, X_2, \dots, X_p$:

$$f(X_1, X_2, \dots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{2}$$

- Linearity assumption is sometimes a poor approximation.
- Ridge regression and LASSO improve upon ordinary least squares, but they still assume linearity.
- This section introduces models that relax the assumption of linearity while maintaining as much interpretability as possible.

Polynomial regression

Idea and definition

We saw polynomial regression previously in "Overview of statistical learning" section:

$$f(X_1) = \beta_0 + \beta_1 X + \beta_2 X_1^2 + \cdots + \beta_d X_1^d \tag{3}$$

- Uses $X_1, X_1^2, \ldots, X_1^d$ as predictors; each adds to the *global* structure of $f(X)$.
- Each coefficient β_j affects function at any value of X_1. (draw graph)
- It is unusual to use d greater than 3 or 4; a very high order polynomial can become overly flexible and can take on some very strange shapes. This is especially true near the boundary of the X variable.

Step functions

Idea

Can instead *localize* effect of β_j to a small range of X_1 by using *step functions*.

- Recall the definition of an *indicator function:* e.g., for an interval B, we have

$$1_B(a) = \begin{cases} 1 & \text{if } a \in B \\ 0 & \text{if } a \notin B \end{cases}.$$

Definition

Model: create cutpoints $c_1 < c_2 < \cdots < c_K$ in X_1's range, then model $f(X_1)$ in (1) by

$$\beta_0 + \beta_1 \mathbf{1}_{(-\infty, c_1)}(X_1) + \beta_2 \mathbf{1}_{[c_1, c_2)}(X_1) + \cdots + \beta_{K-1} \mathbf{1}_{[c_{K-1}, c_K)}(X_1) + \beta_K \mathbf{1}_{[c_K, \infty)}(X_1). \quad (4)$$

- The $K+1$ intervals partition the real line $(-\infty, \infty)$, so the sum

$$\mathbf{1}_{(-\infty, c_1)}(X_1) + \mathbf{1}_{[c_1, c_2)}(X_1) + \mathbf{1}_{[c_2, c_3)}(X_1) + \cdots + \mathbf{1}_{[c_{K-1}, c_K)}(X_1) + \mathbf{1}_{[c_K, \infty)}(X_1)$$

  equals 1, since X_1 must be in exactly one of the $K+1$ intervals.
- Thus (4) is a *piecewise-constant function* of X_1.

Example ($c_1 = 2, c_2 = 4, c_3 = 7$)

Estimation

- Use least squares to fit a linear model using indicators as predictors.
- Cutpoints $c_1, \ldots, c_K$ must be stated/estimated; might be different from any actual breakpoints in the data.

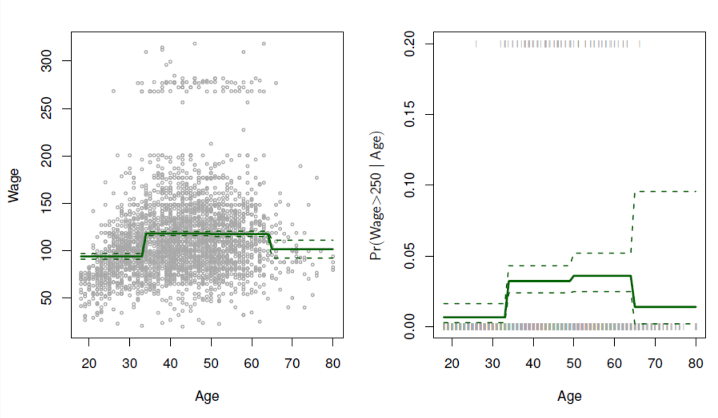

Figure 1: From James et al. (2021). The `Wage` data set. Left: The solid curve displays the fitted values from a least squares regression of wage (in thousands of dollars) using step functions of age, and the dashed curves indicate an estimated 95% confidence interval. Right: We model "wage $>$ 250" using logistic regression with step functions of age. The fitted posterior probability of wage exceeding $250,000 is shown, along with an estimated 95% confidence interval.

Basis functions

Basis functions

Polynomial and piecewise-constant regression models are special cases of a *basis function* approach.

- Idea: express the response Y by K *basis functions* $b_1(\cdot), b_2(\cdot), \ldots, b_K(\cdot)$:

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \cdots \beta_K b_K(X) + \varepsilon. \tag{5}$$

- Polynomial regression: $b_j(x) := x^j$ for all j.
- Piecewise-constant regression: $b_j(x) := 1_{[c_j, c_{j+1})}(x)$ for all j and x, with certain breakpoints $c_1 < c_2 < \cdots < c_K$ for some K.
- Many possible choices for a basis function, e.g., *regression splines*.

Regression splines

Comments

Now we introduce a flexible class of basis functions that extends polynomial regression and piecewise constant regression.

- The main idea is to split the whole region into pieces, and fit a function in each region to improve the overall prediction errors.

Piecewise polynomial regression involves fitting separate low-degree polynomials in each region.

- Example: Instead of assuming that the response Y can be described by a cubic function depending on $X = X_1$ on the whole domain, i.e.,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon\,, \tag{6}$$

  we can model and fit the response below or above a certain threshold c by two different functions, so

$$\begin{aligned} Y &= \beta_{01} + \beta_{11} X + \beta_{21} X^2 + \beta_{31} X^3 + \varepsilon, \quad \text{if } X < c, \\ Y &= \beta_{02} + \beta_{12} X + \beta_{22} X^2 + \beta_{32} X^3 + \varepsilon, \quad \text{if } X \geq c. \end{aligned} \tag{7}$$

- We call c a *knot*: the threshold where the functions are separately defined.
- Each additional knot allows another cubic function to be fitted, so more knots $\rightarrow$ higher flexibility.

Piecewise polynomials with and without constraints

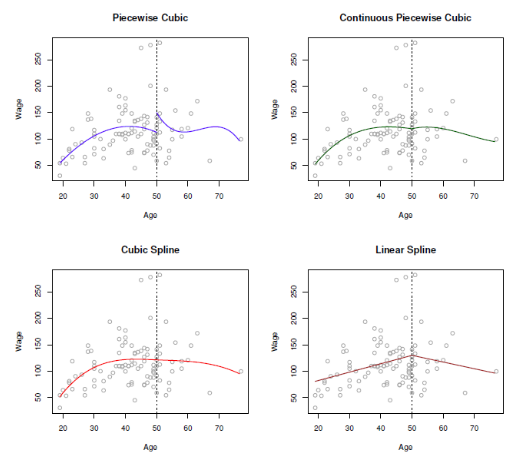

Figure 2: From James et al. (2021). Various piecewise polynomials are fitted to a subset of the `Wage` data, with a knot at `age=50`. Top Left: Cubic polynomials without constraints. Top Right: Cubic polynomials constrained to be continuous at `age=50`. Bottom Left: Cubic polynomials constrained to be continuous, and to have continuous first and second derivatives. Bottom Right: A linear spline, constrained to be continuous.

Constraints and splines

The plots on the last slide exhibit some problematic behavior.

- The top-left plot has a jump which we can avoid by adding the constraint that the function has to be continuous.
- However, continuity doesn't suffice as a smoothness condition: The top-right plot has a continuous but still unnatural "V"-shape.
- In the bottom-left plot, we added two constraints to continuity, namely that the 1st and 2nd order derivatives are also continuous (at age$=50$).
- In general, a *degree-d spline* is a piecewise degree-d-polynomial, with continuity in derivatives up to degree $d-1$ at each knot.
  - Cubic functions require continuity of up to the 2nd derivative at each knot.
  - Linear functions require only continuity at each knot.

The spline basis representation

How can we ensure that a fitted piecewise degree-d polynomial is continuous in derivatives up to degree $d-1$?

- Consider a cubic regression, which models the regression function as

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \tag{8}$$

- We can show that adding a function of the form $\beta_4 h(x, \xi)$ to (8), where

$$h(x, \xi) = (x-\xi)^3_+ = \begin{cases} (x-\xi)^3 & \text{if } x > \xi, \\ 0 & \text{otherwise} \end{cases}$$

  will retain continuity at derivatives up to order 2. (h's derivatives, limits?)
  - Call $h(\cdot, \xi)$ a *truncated power basis function* at knot ξ.
  - Recall: a function is continuous at x if the function's left and right limits at x both equal the function's value at x.
- $f(\cdot) + \beta_4 h(\cdot, \xi)$ is the function for a cubic spline with a knot at ξ.
- For a cubic spline with $K > 1$ knots, can do least squares regression with an intercept and the $3+K$ predictors $X, X^2, X^3, h(X, \xi_1), h(X, \xi_2), \ldots, h(X, \xi_K)$.
- Estimating $K+4$ regression coefficients $\longrightarrow$ $K+4$ degrees of freedom.

Regulating variance at the outer range of predictor values

Splines can have large variance at small/large values of the predictors.

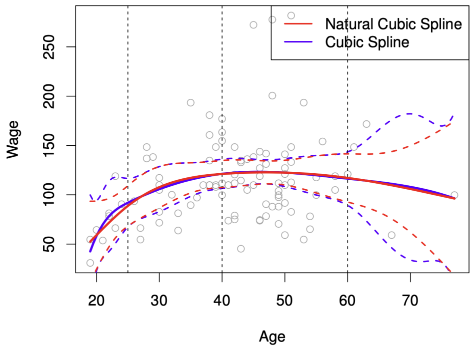

Figure 3: From James et al. (2021). Two splines fitted to a subset of the `Wage` data. Vertical dashed lines: knot locations. Colored, dashed curves: confidence bands.

- Can regulate this variance by introducing another boundary constraint.
- A *natural spline* requires the fitted piecewise function be linear at (i) its left-most piece and (ii) its right-most piece.
- This constraint generally produces more stable estimates at outer range.

Intuitively, knots should be placed where the function varies most rapidly.

- This approach can work well, but in practice it is common to place the knots in a uniform fashion.
- One way is to choose the desired degrees of freedom, then place the knots at uniform quantiles of the data.

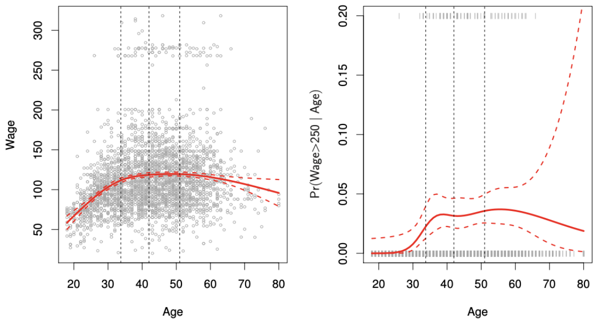

Figure 4: From James et al. (2021). A natural cubic spline function with four degrees of freedom is fit to the Wage data. Left: A spline is fit to wage (in thousands of dollars) as a function of age. Right: Logistic regression is used to model the binary event *wage* $>$ 250 as a function of age. The dashed lines denote the knot locations.

Choosing the number of knots

How many knots to use? Some options:

1. Try different numbers of knots and see which produces best looking curve.
2. Use cross-validation to estimate test error for various numbers of knots, then choose the number of knots that produces the smallest CV error.

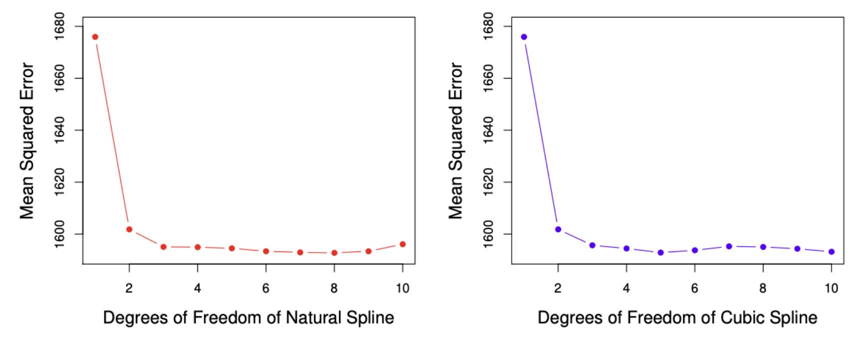

Figure 5: From James et al. (2021). Ten-fold cross-validated MSEs for selecting the degrees of freedom when fitting splines to `Wage` data.

Comparison to polynomial regression

- Regression splines often give superior results to polynomial regression.
    - Polynomial regression introduces flexibility by using high degree polynomials (which affect global behavior of function).
    - Splines introduce flexibility by increasing the number of knots, but keep the degree fixed. (Allows more “surgical” changes in function behavior.)
- This produces more stable estimates, and splines also allow placing more knots, and also precisely at specific regions.

Smoothing splines

Idea

Recall: in regression we try to find a function (let's call it g) that fits observed data $(x_1, y_1), \ldots, (x_n, y_n)$ well, i.e., that makes $RSS = \sum_{i=1}^{n}(y_i - g(x_i))^2$ small.

- Can always make RSS zero by having g interpolate all n data points, but such a function would overfit the data (poor generalization).
- In regression splines, we regulate the flexibility of g by specifying the number of knots and flexibility of basis functions before fitting to data.
- Instead, what if we regulate flexibility of g by penalizing its "wigglyness":

$$\arg\min_{g} \left\{ RSS + \lambda \int (g''(t))^2 \, \mathrm{d}t \right\}, \tag{9}$$

    - RSS is a *loss function* that encourages g to fit the data well.
    - $\lambda \int (g''(t))^2 \, \mathrm{d}t$ is a *penalty term* that penalizes g's variability/wigglyness.
- The function g minimizing the objective in (9) is called a *smoothing spline*.

Meaning

We saw this “Loss+Penalty” formulation in ridge regression and in LASSO.

- Let’s examine the penalty term in (9) more closely.

Integral term $\int (g''(t))^2 \, \mathrm{d}t$:

- g'' describes how much g' changes (i.e., how much slope of g changes), and thus can be interpreted as a measure of *roughness*.
  - If $g(t)$ is very rough (wiggly) near t, then $|g''(t)|$ is large.
  - If $g(t)$ is very smooth (stable) near t, then $|g''(t)|$ is small.
- Thus $\int (g''(t))^2 \, \mathrm{d}t$ measures the total change in g' over its entire range.

Tuning parameter λ:

- When $\lambda = 0$, smoothing spline will perfectly interpolate the training data.
- As $\lambda \to \infty$, smoothing spline turns into OLS line of best fit (infinitely smooth).
- For intermediate λ, smoothing spline will approximate training observations but will be somewhat smooth; λ controls bias-variance trade-off.

What does a smoothing spline look like?

A smoothing spline can be shown to be a piecewise cubic polynomial with knots at the unique values of $x_1, \ldots, x_n$, and continuous first and second derivatives at each knot.

- I.e., *it is a natural cubic spline with knots at $x_1, \ldots, x_n$!*
- Is a 'shrunken' version of the natural cubic spline that would be obtained using the basis function approach in slide 13 with knots at $x_1, \ldots, x_n$.
- λ controls the shrinkage, hence controls the *effective degrees of freedom*.
  - Usually *degrees of freedom* refers to the number of free parameters, e.g., the number of coefficients fit in a regression.
  - A smoothing spline has n parameters, but they are heavily constrained or shrunk down.
  - The formal definition of effective degrees of freedom is somewhat technical.
  - Intuitively it is a measure of flexibility of the smoothing spline.
  - The larger the effective df, the more flexible the smoothing spline.
  - As λ increases from 0 to ∞, the effective df decrease from n to 2.