Non-linear Regression: Polynomial Regressions, Regression Splines
(ISLR 7.1 - 7.4)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

Moving Beyond Linearity

The truth is never linear! Or almost never!
But often the linearity assumption is good enough.

When its not ...

- polynomials
- step functions
- splines
- generalized additive models

offer a lot of flexibility,
without losing the ease and interpretability of linear models.

In this chapter, all (but the last) methods are illustrated using one response Y and a single X.

Polynomial Regression

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

- In general, it is unusual to use d greater than 3 or 4.
- This is just a standard linear model with predictors $x_i, x_i^2, \ldots, x_i^d$.
- Coefficients 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:

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

- Since $\hat{f}(x_0)$ is a linear function of the $\hat{\boldsymbol{\beta}}$, we can get a simple expression for *point-wise variances* $\text{Var}[\hat{f}(x_0)]$ at any value x_0.
- Confidence interval: $\hat{f}(x_0) \pm 2 \cdot \text{se}[\hat{f}(x_0)]$.

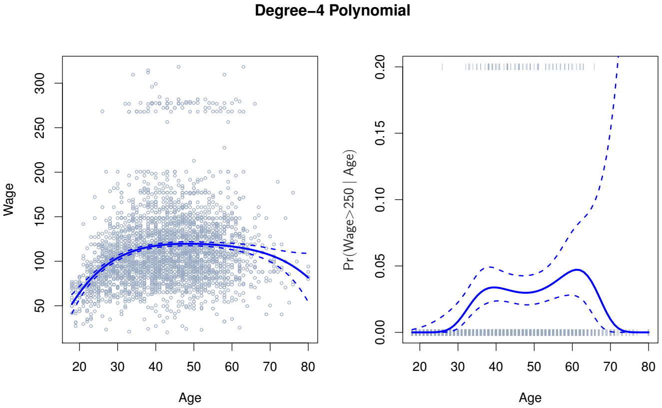
Polynomial Logistic Regression

Logistic regression follows naturally:

$$P(y_i = 1 \mid x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}}$$

- To get confidence intervals, compute upper and lower bounds on on the logit scale, and then invert to get on probability scale.

The Wage Data



- Left: regression; right: classification.
- Confidence bands (dashed): point-wise $\hat{f}(x_0) \pm 2 \cdot \text{se}[\hat{f}(x_0)]$ on a grid of values for x_0.
- Caveat: polynomials have notorious tail behavior.

Step Functions

- Polynomial regression imposes a global structure on X.
- We can instead cut the variable into distinct regions: converts a continuous variable into an ordered categorical variable.
- With K *knots* (cutpoints) $c_1, c_2, \ldots, c_K$ in the range of X, we construct $K+1$ indicator variables (dummy variables)

$$\begin{aligned}
C_0(X) &= \mathbf{1}(X < c_1) \\
C_1(X) &= \mathbf{1}(c_1 \leq X < c_2) \\
C_2(X) &= \mathbf{1}(c_2 \leq X < c_3) \\
&\vdots \\
C_K(X) &= \mathbf{1}(c_K \leq X)
\end{aligned}$$

Here $\mathbf{1}(A)$ is an 0-1 indicator function for the event A.

Step Functions in Linear Regression

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

Note that $C_0(X)$ is not included. Why?

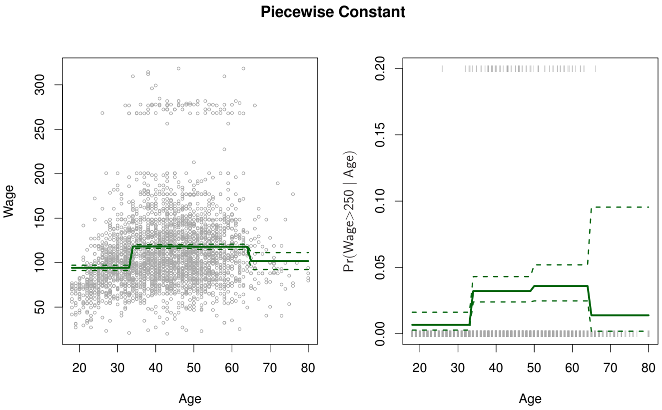
Intepretation

- β_0: the mean value of Y for $X < c_1$.
- $\beta_0 + \beta_j$: the mean value of Y for $c_j \le X < c_{j+1}$, for $j = 1, \ldots, K-1$
- $\beta_0 + \beta_K$: the mean value of Y for $X \ge c_K$.

Useful way of creating interactions that are easy to interpret. For example, interaction effect of `Year` and `Age`:

$$\mathbf{1}(\texttt{Year} < 2005) \cdot \texttt{Age}, \quad \mathbf{1}(\texttt{Year} \ge 2005) \cdot \texttt{Age}$$

The Wage Data



- Left: regression; right: classification.
- Knots: 35, 50, 65

Basis Functions

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

- The basis functions $b_1(\cdot), b_2(\cdot), \ldots, b_K(\cdot)$ are chosen ahead of time.
- Special cases:
  - polynomial regression: $b_j(x) = x^j$
  - step functions: $b_j(x) = \mathbf{1}(c_j \leq x < c_{j+1})$

Piecewise Polynomials

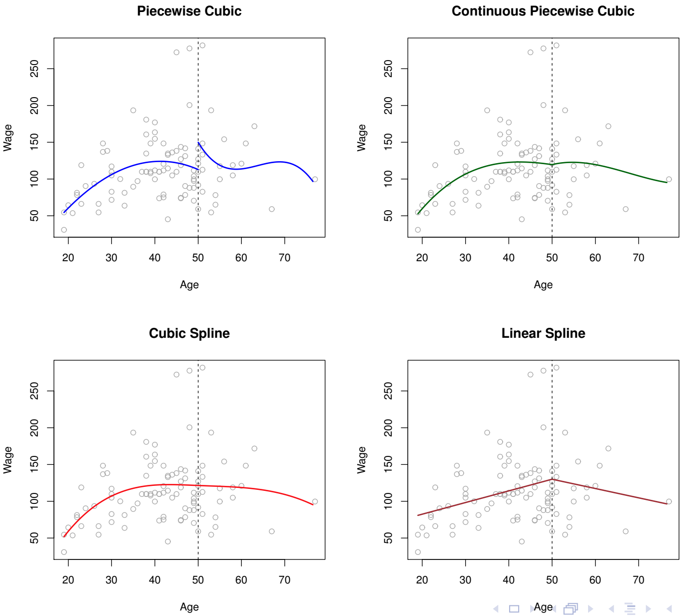
- Instead of a single polynomial in X over its whole domain, we can rather use different polynomials in regions defined by knots.

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

- Degrees of freedom of the piecewise cubic model with K knots:

$$4(K+1)$$

- Step function is a special case when $d = 0$.
- Better to add constraints to the polynomials, e.g. continuity.
- Splines have the "maximum" amount of continuity.


Piecewise Cubic
Continuous Piecewise Cubic
Cubic Spline
Linear Spline
Wage
Age
50
100
150
200
250
20
30
40
50
60
70

Regression Splines

Add continuity constrains on piecewise polynomials:

- For cubic model, the piecewise polynomial funciton and its first and second derivatives are continuous at all knots.
- Each constraint that we impose on the piecewise cubic polynomials effectively frees up one degree of freedom.
- A *cubic spline* with K knots uses a total of

$$4(K+1) - 3K = K + 4$$

  degrees of freedom.
- In general, a *degree-d spline* is a piecewise degree-d polynomial, with continuity in derivatives up to degree $d-1$ at each knot.

The Spline Basis Representation: Linear Splines

A linear spline with knots at $\xi_k, k = 1, \ldots, K$ is a piecewise linear polynomial continuous at each knot.

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

where the basis functions are

$$\begin{aligned} b_1(x_i) &= x_i \\ b_{k+1}(x_i) &= (x_i - \xi_k)_+, \quad \text{for } k = 1, 2, \ldots, K. \end{aligned}$$

Here the $(\cdot)_+$ means *positive part*:

$$(a)_+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases}$$

Note: the slope of X when $\xi_k \leq X < \xi_{k+1}$ is $\beta_0 + \beta_1 + \cdots + \beta_{k+1}$.

The Spline Basis Representation: Cubic Splines

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

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

where the basis functions are

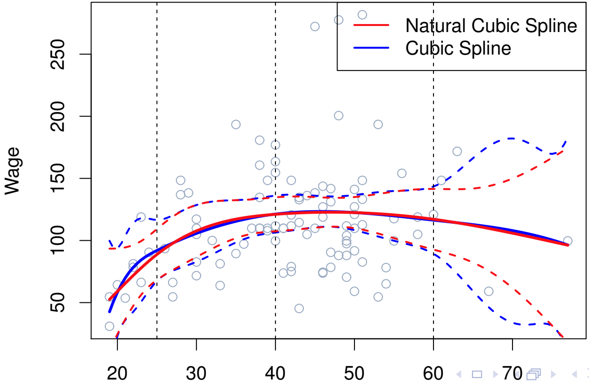
$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3, \quad \text{for } k = 1, 2, \ldots, K, \end{aligned}$$

where

$$(a)_+^3 = \begin{cases} a^3 & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases}$$

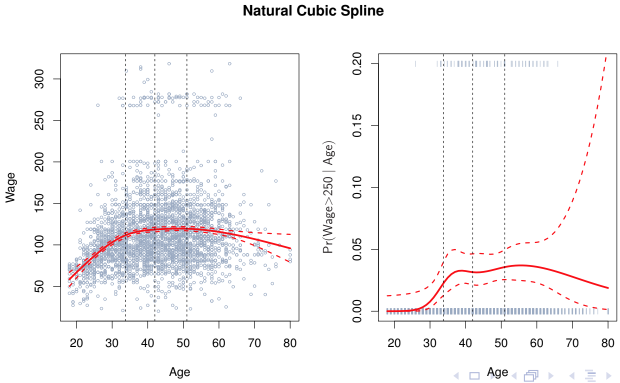
Natural Cubic Splines

- A *natural spline* is a regression spline with additional boundary constraints: the function is linear where X is smaller than the smallest knot, or larger than the largest knot.
- Degrees of freedom of a natural cubic spline: K.



How to choose the number of knots K and their locations?

- Place more knots in places where we feel the function might vary most rapidly, and to place fewer knots where it seems more stable.
- In practice, it is common to place knots in a uniform fashion. E.g., if $K = 3$, set knots at the 25th, 50th, and 75th percentiles of X.
- Use cross validation to choose K.



Polynomial Regression Vs. Natural Cubic Spline

- A degree-14 polynomial, which has 15 degrees of freedom.
- A natural cubic spline with 15 degrees of freedom.