Nonlinear Methods

Dr. Du, Lan

Faculty of Information Technology, Monash University, Australia

FIT5149 week 7

Outline

Polynomial Regression: the basics

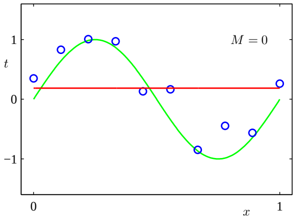
- A polynomial function

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i \\ &= \sum_{j=0}^{d} \beta_j x_i^j \end{aligned}$$

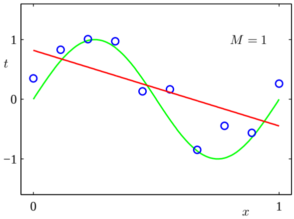
  where d is the order (or degree) of the polynomial.
  - Constant polynomial: β_0
  - linear polynomial: $\beta_0 + \beta_1 x_i$
  - quadratics: $\beta_0 + \beta_1 x_i + \beta_2 x_i^2$
  - cubics: $\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$
- Curve fitting: minimise the following error function

$$E(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{j=0}^{d} \beta_j x_i^j - y_i\right)^2$$

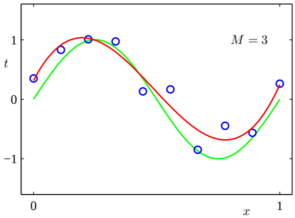
Polynomial Regression: fit data generated with sin()



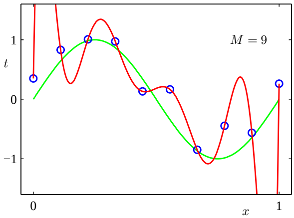
(a)



(b)

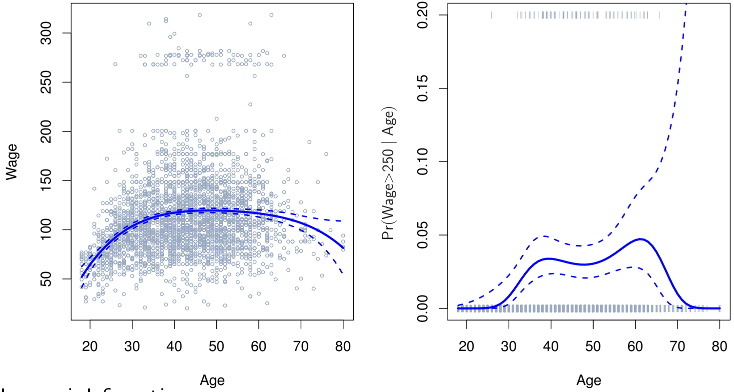


(c)



(d)

Polynomial Regression: predict Wage with Age

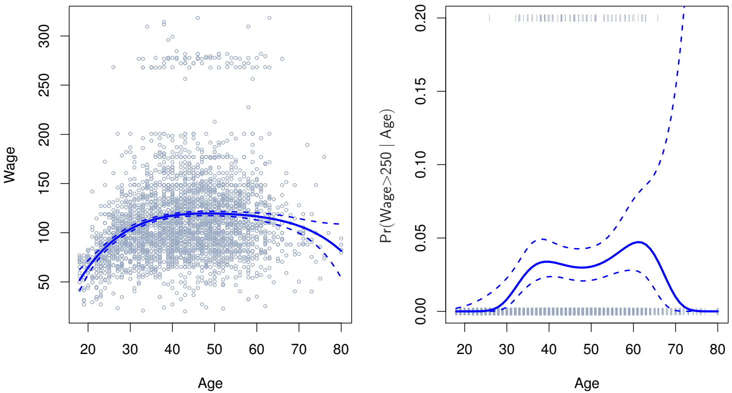


- Polynomial function :

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

  - $\hat{f}(x_0)$ is a linear function of the $\hat{\beta}_l$
  - Pointwise-variance $Var[\hat{f}(x_0)]$ at any value x_0. The two plots show $\hat{f}(x_0) \pm 2 \cdot se[\hat{f}(x_0)]$ (approximately 95% confidence interval).

Polynomial Regression: predict Wage with Age



- Logistic regression

$$Pr(y_i > 250 \mid x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4}}$$

- The confidence interval is fairly wide due to that there are only 79 high earners.

Step functions

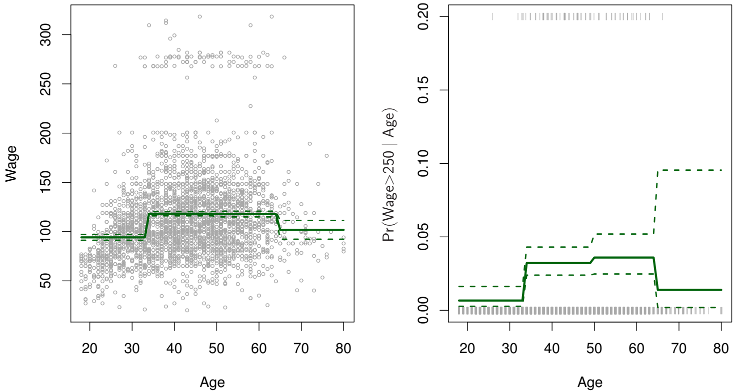
- Another way of creating transformations of a variable — cut the variable into distinct regions and avoid impose a global structure.
  - Create cut-points $c_1, c_2, c_3, \ldots, c_K$, and $K+1$ variables

$$
\begin{aligned}
C_0(X) &= I(X < c_1),\\
C_1(X) &= I(c_1 \leq X < c_2),\\
C_2(X) &= I(c_2 \leq X < c_3),\\
&\vdots\\
C_{K-1}(X) &= I(c_2K - 1 \leq X < c_K),\\
C_K(X) &= I(c_K \leq X),
\end{aligned}
$$

  - Fit a linear model using those variables

$$
y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \beta_3 C_3(x_i) + \cdots + \beta_K C_K(x_i) + \epsilon_i
$$

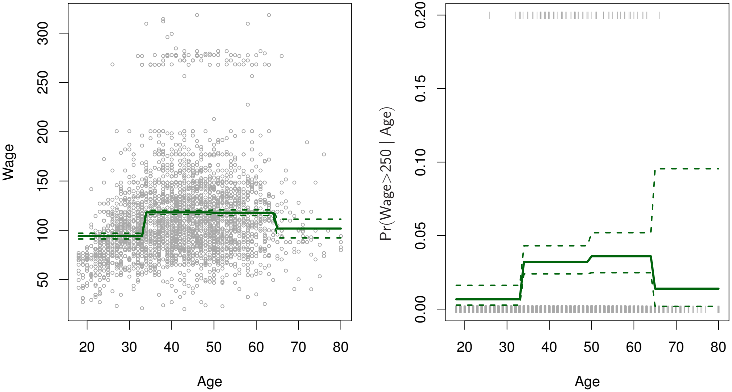
Step functions



- Easy to create a series of dummy variables representing each group

$$C_1(X) = I(X < 35) \quad C_2(X) = I(35 \leq X < 50) \quad C_3(X) = I(50 \leq X < 65), \ldots$$

Step functions

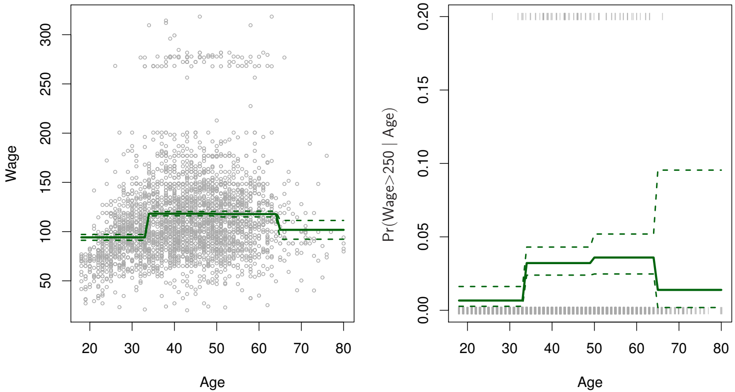


- Useful way of creating interactions that are easy to interpret. For example, interaction effect of Year and Age:

$$I(\text{Year} < 2005) \cdot \text{Age} \quad I(\text{Year} \geq 2005) \cdot \text{Age}$$

would allow for different linear functions in each age category.

Step functions



- Choice of cut-points or knots can be problematic. For creating nonlinearities, smoother alternatives such as splines are available.

Basic Functions

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \cdots + +\beta_K b_K(x_i) + \epsilon_i$$

- $b_k(X)$: a function or transformation that can be applied to variable X.
- What is $b_k(x_i)$ for the polynomial regression?
- What is $b_k(X_i)$ for the piecewise constant functions?

Basic Functions

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \cdots + + \beta_K b_K(x_i) + \epsilon_i$$

- $b_k(X)$: a function or transformation that can be applied to variable X.
- What is $b_k(x_i)$ for the polynomial regression?

$$b_k(x_i) = x_i^k$$

- What is $b_k(X_i)$ for the piecewise constant functions?

$$b_k(x_i) = I(c_k \leq x_i^k \leq c_{k+1})$$

Outline

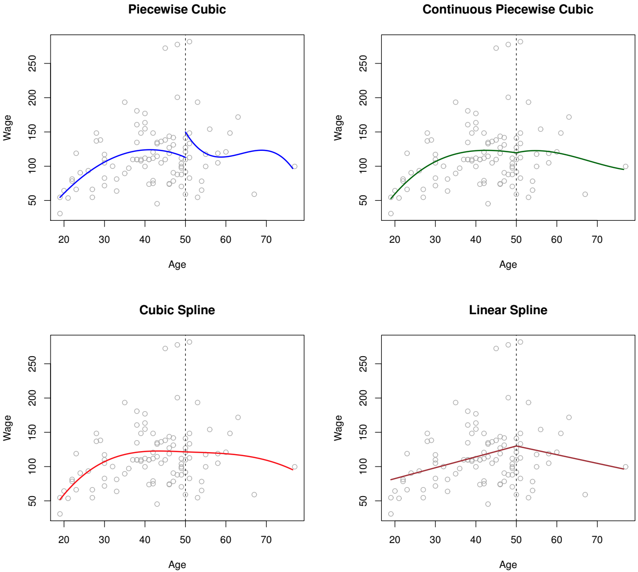
Piecewise Polynomials

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

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

- Using more knots leads to a more flexible piecewise polynomial.
  - K different knots $\rightarrow$ $K + 1$ different cubic polynomials

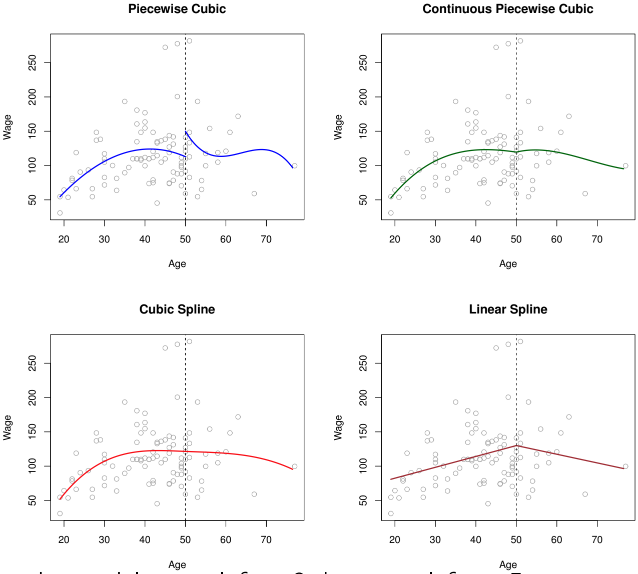
Piecewise Polynomials: continuity and smooth



Piecewise Polynomials: continuity and smooth

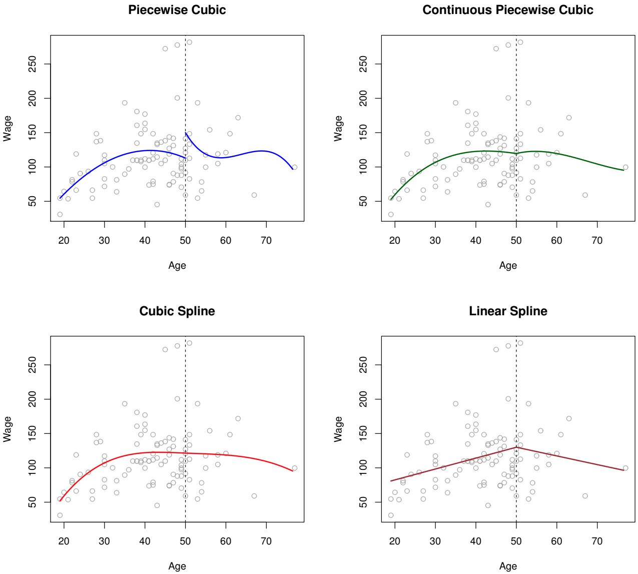
- Continuity
- Continuity of the first derivative.
  what is the geometry of the first derivative?
  - if $\frac{df}{dx}(p) > 0$, then f(x) is an increasing function at $x = p$.
  - if $\frac{df}{dx}(p) < 0$, then f(x) is an decreasing function at $x = p$.
  - if $\frac{df}{dx}(p) = 0$, then $x = p$ is called a critical point of $f(x)$, and we do not know anything new about the behaviour of $f(x)$ at $x = p$.
- Continuity of the second derivative.
  - if $\frac{d^2f}{dx^2}(p) > 0$ at $x = p$, then f(x) is concave up at $x = p$.
  - if $\frac{d^2f}{dx^2}(p) < 0$ at $x = p$, then f(x) is concave down at $x = p$.
  - if $\frac{d^2f}{dx^2}(p) = 0$ at $x = p$, we do not know anything new about the behaviour of $f(x)$ at $x = p$.

Piecewise Polynomials: continuity and smooth



DF for fitting the model: top_left →8; bottom_left → 5

Piecewise Polynomials: continuity and smooth



DF for fitting the model with K knots: 4 + K

Linear Splines

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

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

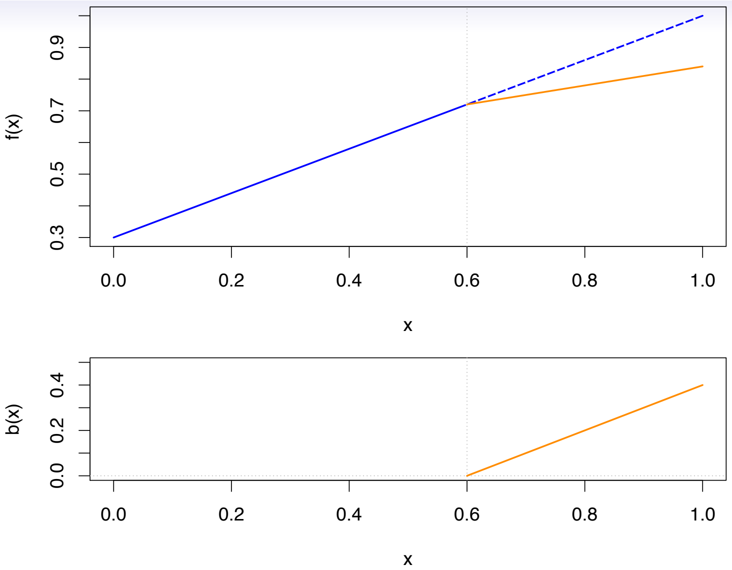
where the b_k are basis functions

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

and

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

Linear Splines



Cubic Splines

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

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

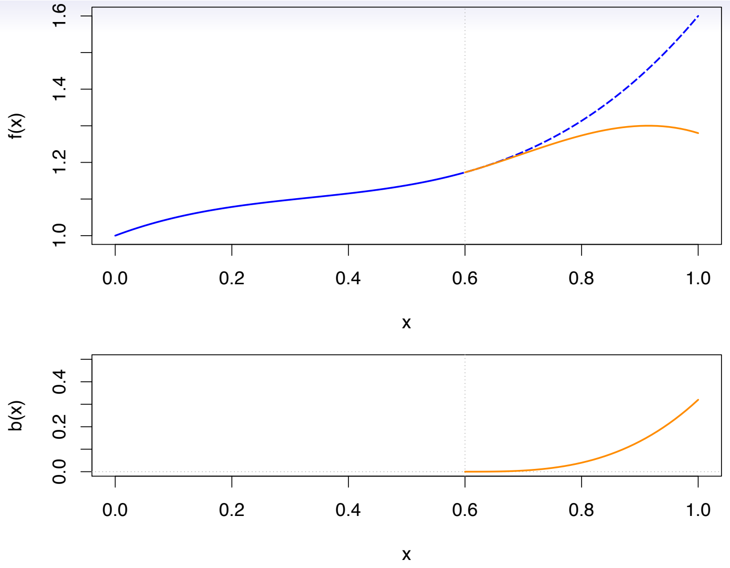
where

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

and

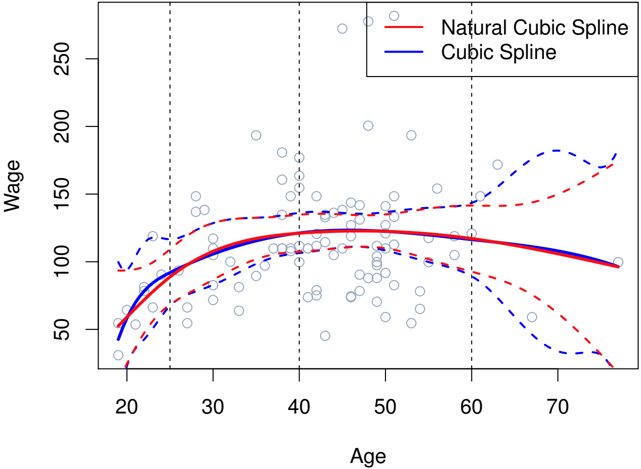
$$(x_i - \xi_k)^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

Cubic Splines



Natural Cubic Splines

- A natural cubic spline extrapolates linearly beyond the boundary knots.
  - Add extra constraints to the end, i.e., the second derivatives at the two outer knots are zero.



#Knots and Knot placement

- One strategy is to decide K, and then place them at appreciate quantiles of the observed X

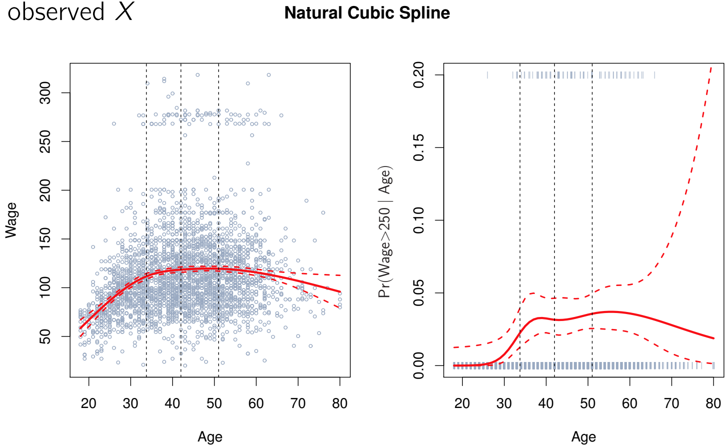


Figure: 3 knots are chosen automatically as the 25th, 50th and 75th percentiles. (3 knots correspond to 4 DF for fitting the natural cubic splines.)

#Knots and Knot placement

- A cubic spline with K knots has K + 4 parameters or degrees of freedom.
- A natural spline with K knots has K degrees of freedom.

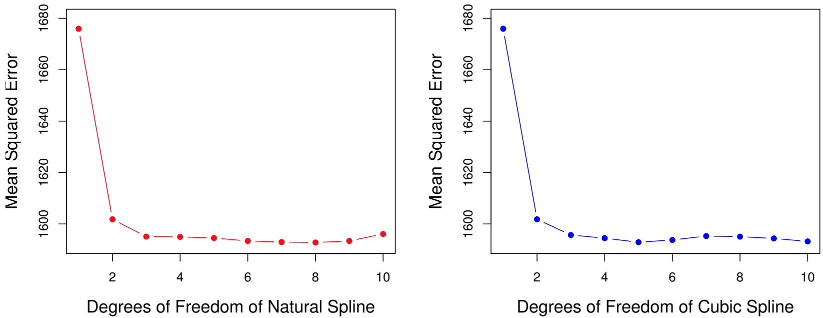


Figure: 10-fold cross-validation for selecting the degrees of freedom.

#Knots and Knot placement

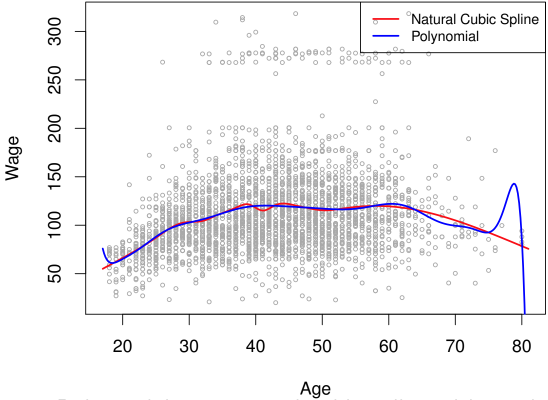


Figure: A degree-15 Polynomial vs a natural cubic spline with 15 degrees of freedom.

- Polynomial uses a high degree to produce flexible fit.
- Splines introduce flexibility by increasing K but keeping the degree fixed.
- Splines allow us to place more knots, and hence flexibility, over regions where the function f seems to be changing rapidly.

Outline

Smooth Splines: overview

- Regression splines:
  - specify a set of knots,
  - produce a sequence of basis functions, and
  - use least squares in fitting

$$RSS = \sum_{i=1}^{n} (y_i - g(x_i))^2$$

- Problem: If we don't have any constraints on $g(x_i)$, we can always make RSS zero simply by choosing g such that it interpolates all of y_i.
- What we want
  - Small RSS, and
  - Smooth fitted function

Smooth Splines: target function

- Smoothing spline

$$\underset{g \in \mathcal{S}}{minimize} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

  - The first term is RSS, and tries to make $g(x)$ match the data at each x_i
  - The second terms is a roughness penalty and controls how wiggly $g(x)$ is.
  - Tuning parameter $\lambda \geq 0$:
    - 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.
  - Note on derivatives
    - $g'(t)$: measures the slope of a function at t
    - $g''(t)$: measures the amount by which the slope is changing. If it is large in absolute value if $g(t)$ is very wiggly near t; it is close to zero otherwise.
  - $\int g''(t)^2 dt$: a measure of the total change in the function $g'(t)$.
    - If $g(t)$ is smooth, $\int g''(t)^2 dt$ will take on a small value.
    - If $g(t)$ is jumpy and variable, $\int g''(t)^2 dt$ will take on a large value.

Smooth Splines: target function

- Smoothing spline

$$\underset{g \in \mathcal{S}}{minimize} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

  - The first term is RSS, and tries to make $g(x)$ match the data at each x_i
  - The second terms is a roughness penalty and controls how wiggly $g(x)$ is.
  - Tuning parameter $\lambda \geq 0$:
    - 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.
  - A natural cubic spline with a knot at every unique value of x_i
  - Some details
    - Smoothing splines avoid the knot-selection issue
    - Cross-validate (LOOCV) a single parameter λ

Outline

Local Regression

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.

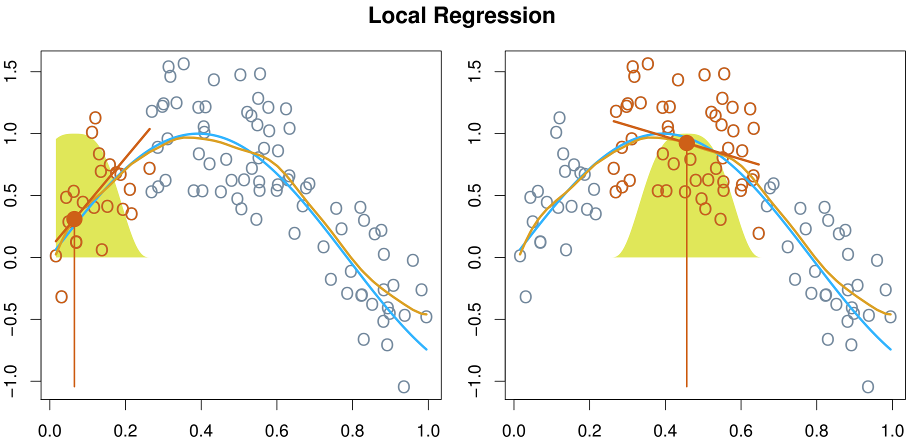
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.

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

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

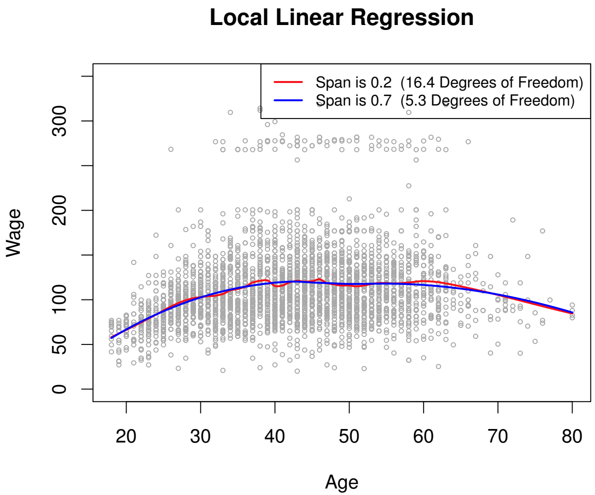
4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Local Regression



- With a sliding weight function, we fit separate linear fits over the range of X by weighted least squares.

Local Regression



- Smaller value of s, the more local and wiggly will be our fit.
- A very large value of s will lead to a global and smooth fit.

Outline

Generalised Additive Models: regression

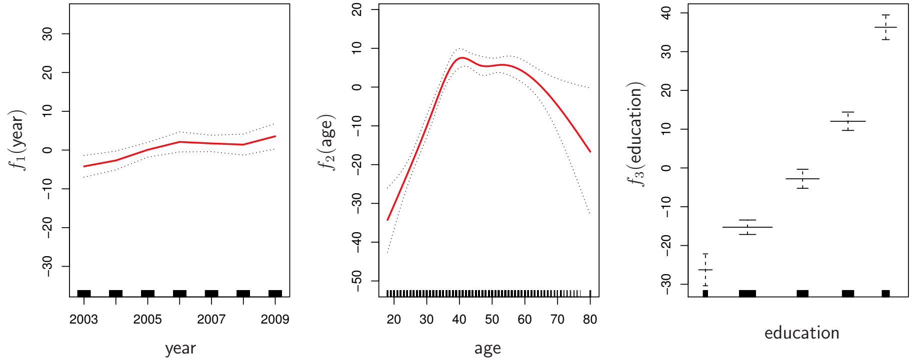
- GAM: a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity.
  - The multiple linear regression model

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

  - Generalised additive model

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{i,j}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i \end{aligned}$$

Generalised Additive Models: regression



- Fit the following model

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

Then, fit f_1 and f_2 using natural splines, and f_3 using a step function.

$$lm(wage \sim ns(year, 4) + ns(age, 5) + education, data = Wage)$$

Generalised Additive Models: classification

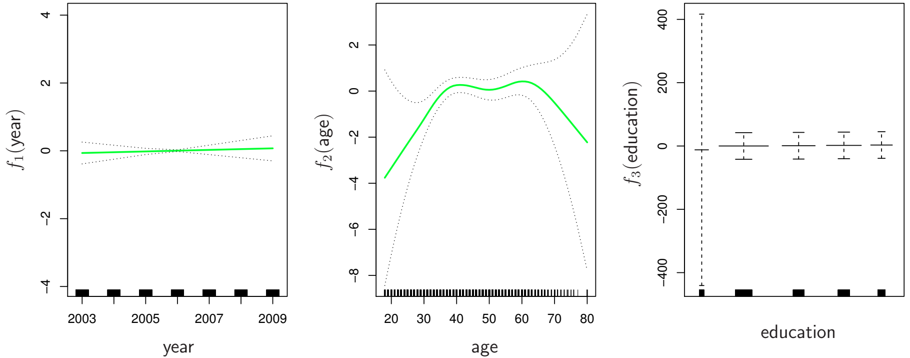
- The logistic regression model

$$log(\frac{P(X)}{1 - P(X)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- GAM extension

$$log(\frac{P(X)}{1 - P(X)}) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

Generalised Additive Models: classification



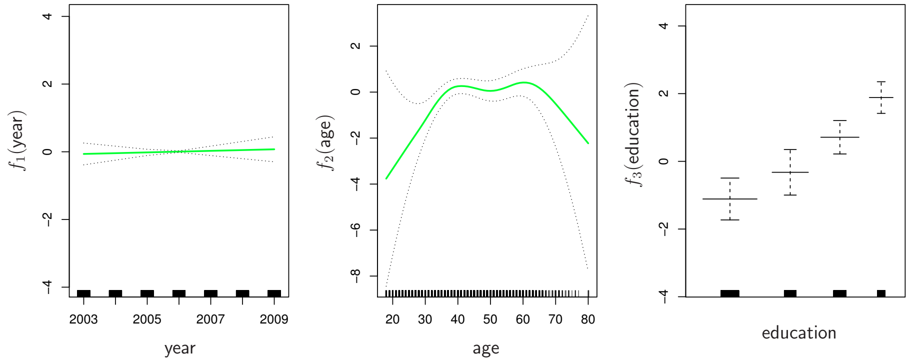
- Fit the following model

$$log(\frac{P(X)}{1-P(X)}) = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$

where

$$P(X) = P(\text{wage} > 250 \mid \text{year}, \text{age}, \text{education})$$

Generalised Additive Models: classification



- Fit the following model

$$log(\frac{P(X)}{1-P(X)}) = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$

where

$$P(X) = P(\text{wage} > 250 \mid \text{year}, \text{age}, \text{education})$$

Summary

- Simple extensions of linear models, e,g, polynomial regression, step functions, etc.
- Splines: Regression splines & Smooth splines
- Local regression
- Generalised additive models
- Reading materials:
  - "Moving Beyond Linearity", Chapter 7 of "Introduction to Statistical Learning", 6th edition
    - Skip section 7.5.2
- Acknowledgement:
  - Figures in this presentation were taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
  - Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani
  - Figures on the slide 4 are from Christopher Bishop's website.