Introduction to Statistical Learning
with applications in Python

Based on "Introduction to Statistical Learning, with applications in R" by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Beyond Linearity

Polynomial Regression, Basis Functions, Splines, Generalised Additive Models

Kurt Rinnert

Physics Without Frontiers

The Abdus Salam International Centre for Theoretical Physics

UNIVERSITY OF LIVERPOOL

Abstract

Linear models are often applicable and easy to interpret. However, as we have stressed several times, the *true* functional relationships are rarely linear.
In this lecture we will explore several ways to drop the assumption of linearity.

Overview

- Polynomial Regression
- Basis Functions
- Splines
- Smoothing Splines
- Generalised Linear Models (GAMs)

We'll see again that model flexibility is related to the bias-variance trade-off.

Polynomial Regression

- We have already seen some ways to extend linear regression to include non-linear terms.
- In particular we have fitted and interpreted polynomial models:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \epsilon_i$$

- Or models with interaction terms:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon_i$$

- We will not go through all this again, but some exercises will still refer to these ideas.

In this lecture we will explore more powerful concepts to move beyond linearity.

Basis Functions

- The idea is quite simple: we don't limit ourselves to polynomials or interactions.
- Instead we allow arbitrary functions of the predictors:

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

- We can think of this as a linear model with predictors

$$b_1(x_i), b_2(x_i), \ldots, b_K(x_i)$$

- This means we can use the regression methods we learned about to estimate the parameters $\boldsymbol{\beta}$.

All techniques for evaluating performance and model selection are still applicable.

Piecewise Polynomials

- Fitting high-degree polynomials is not ideal as it can lead to severe over-fitting.
- We can instead fit lower-degree *piecewise polynomials*, for example:

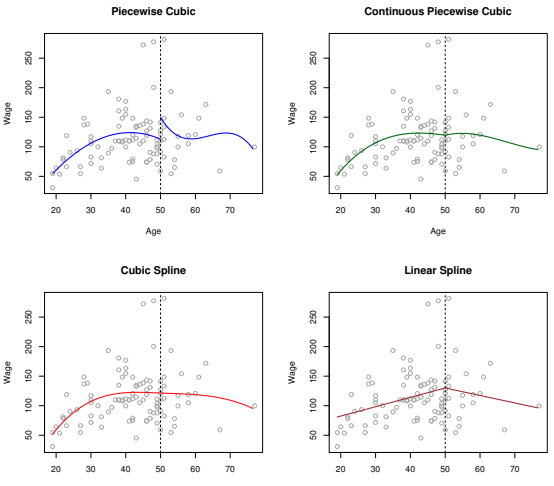
$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c \end{cases}$$

- Keeping the degree low mitigates the chance of over-fitting.
- In this example the piecewise polynomial has one *knot* at $x_i = c$.
- We can increase the flexibility by introducing more knots.

Here, and in the following we assume on single predictor X.

Piecewise Polynomials

- The piecewise polynomial has a discontinuity at c (age = 50).
- This can be removed by requiring *continuity*.
- That still leaves us with a kink.
- The situation is further improved by requiring the derivatives to be continuous.



Putting constraints on the knots of piecewise polynomials leads to splines.

Constraints and Splines

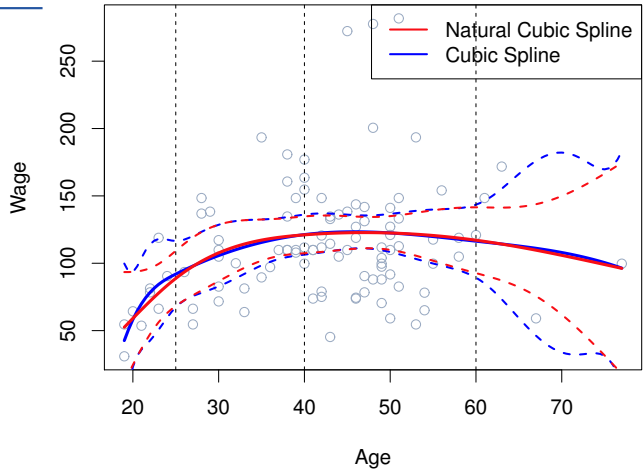
- To construct a *cubic spline* we start from a piecewise polynomial of degree 3 and impose the following constraints.

constraint	continuity
$b(x_i)$	value
$b'(x_i)$	slope
$b''(x_i)$	curvature

- Each constraint frees up one degree of freedom.
- In general, cubic splines with K knots use $K + 4$ degrees of freedom.

We do not concern ourselves with the details involved in imposing the constraints.

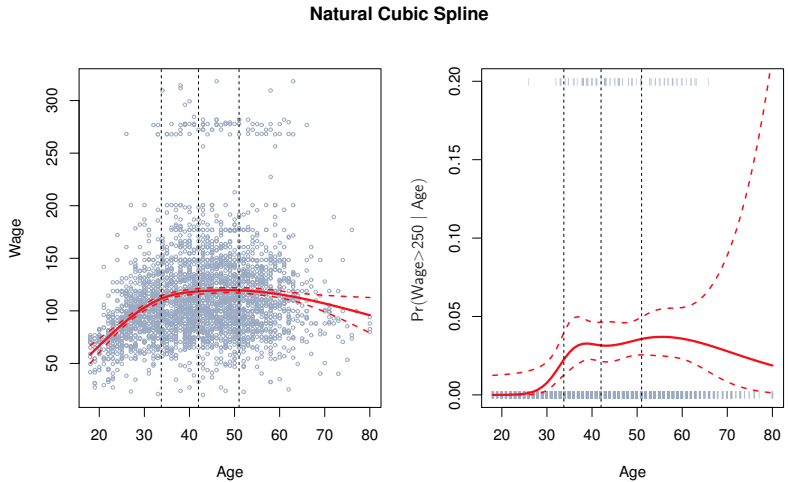
Natural Splines



- Splines can have large variance at the outer range of the predictors.
- *Natural splines* are constrained to be linear at the boundaries.

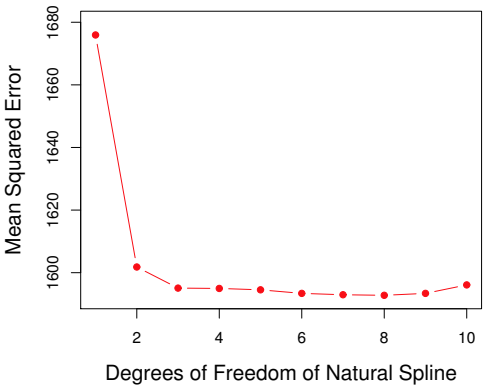
Note that the confidence interval for the natural spline is narrower.

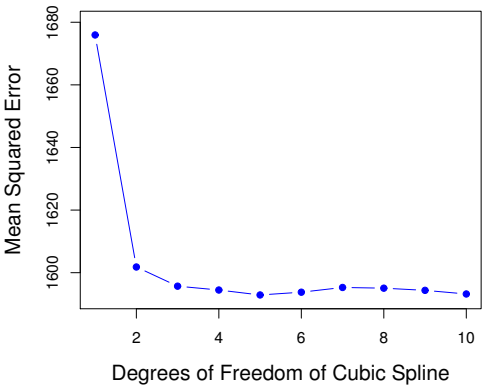
Choosing *K* & Knot Positions



We usually specify the degrees of freedom and place the knots uniformly over quantiles .

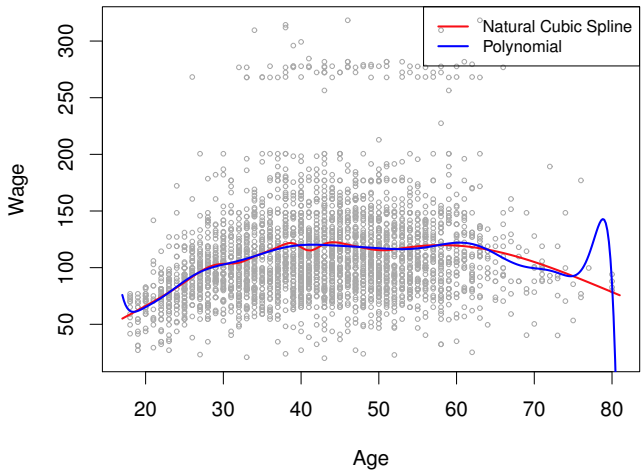
Choosing K & Knot Positions





In practice – you guessed it – we use cross-validation.

Splines versus Polynomial Regression



High degree polynomials tend to behave badly at the boundaries.

Smoothing Splines

- *Smoothing splines* are a different approach to fitting splines.
- We want to find some *smooth* function $g(x)$ that fits the data well, that is

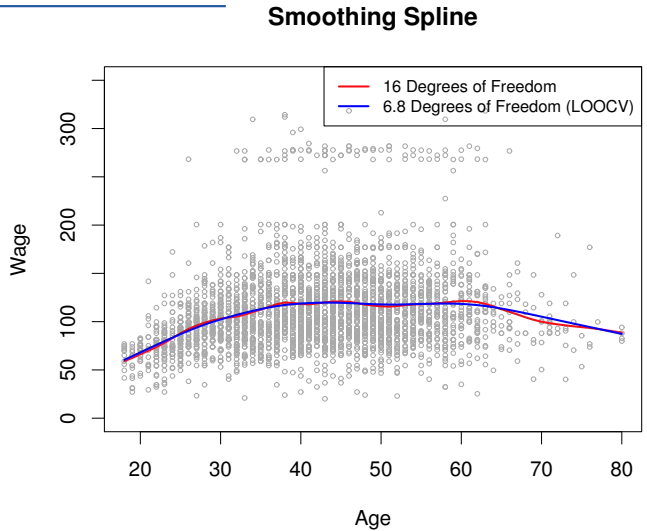
$$\text{RSS} = \sum_{i=1}^{n} (y_i - g(x_i))^2$$

  is small.
- The smoothness can be imposes by minimising

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

The *loss + penalty* approach is very common in machine learning practice.

Choosing the λ Parameter



The λ parameter is continuous, leading to the concept of *effective degrees of freedom*.

Generalised Additive Models

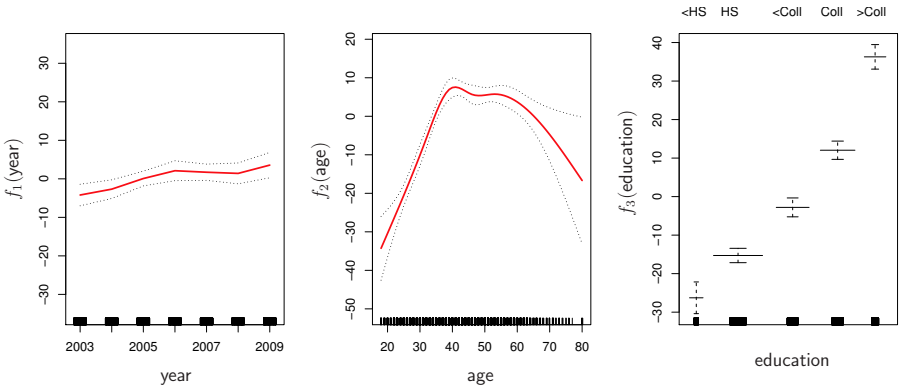
- We now release the condition of having a single predictor in the model.
- This leads to the concept of *generalised additive models* (GAMs).
- This is analogous of moving from simple linear regression to multiple linear regression.
- That is, the models can now take the form

$$y_i = \beta_0 + \sum_{j=1}^{p} f_i(x_{ij}) + \epsilon_i$$
$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

with *smooth* functions f_j.

The model still *additive* because we *add* separate f_j for each X_j.

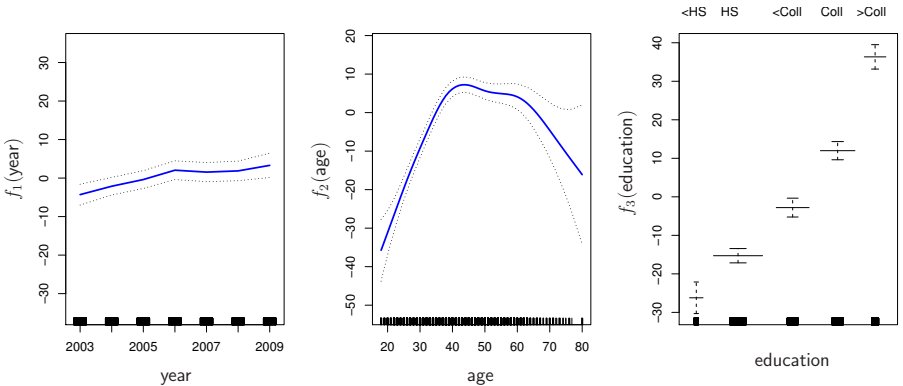
Generalised Additive Models



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

A GAM model on the Wage data set with two cubic splines and a qualitative variable.

Generalised Additive Models



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

A GAM model on the Wage data set with two smoothing splines and a qualitative variable.

GAMs for Classification

- GAMs can also be used for classification.
- Just like with ordinary linear models, we use logistic regression in this scenario:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

- For example:

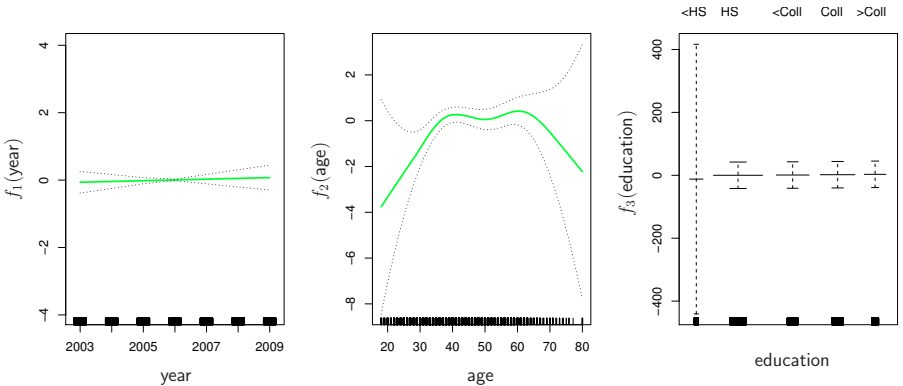
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times \texttt{year} + f_2(\texttt{age}) + f_3(\texttt{education})$$

where

$$p(X) = P(\texttt{wage} > 250 | \texttt{year}, \texttt{age}, \texttt{education})$$

That is, we want to predict whether the yearly income is greater than $250,000.

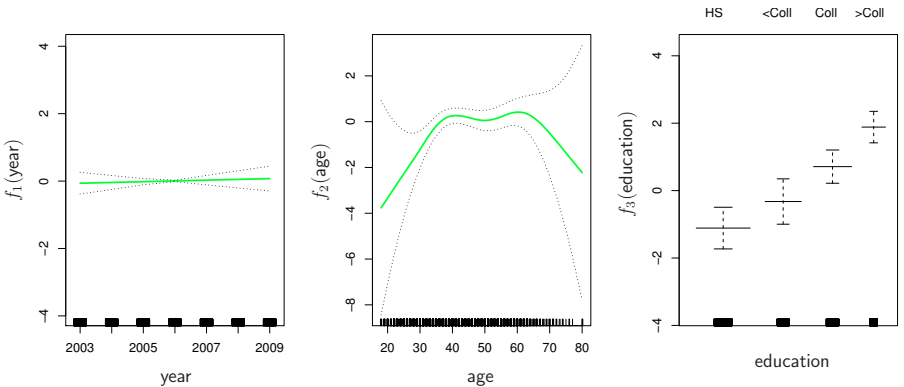
GAMs for Classification



$$p(X) = P(\texttt{wage} > 250 | \texttt{year}, \texttt{age}, \texttt{education})$$

There seems to be a problem with the qualitative variable.

GAMs for Classification



$$p(X) = P(\texttt{wage} > 250 | \texttt{year}, \texttt{age}, \texttt{education})$$

It turns out high school drop-outs never earn more than $250,000/year.