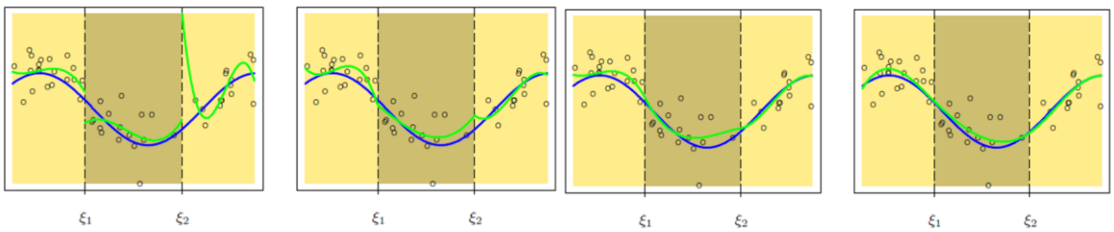
Regression Splines

Review of Step Functions & Polynomial Regression:

- Step Function: break the range of X into bins and fit a different constant in each bin. This amounts to converting a continuous variable into an ordered categorical variable
- Polynomial Regression: for large enough degree d, a polynomial regression allows us to produce an extremely non-linear curve. We generally avoid using d greater than 3 or 4 because large values of d can lead to overfitting
- Both of them are special cases of a basis function approach

Regression Spline with Constraints:



12 df | 10 df | 8 df | 6 df

- The graphs above show a series of piecewise-cubic polynomials fit to the same data, with increasing order of continuity at the knots
- Adding continuity constraints reduces the number of degrees of freedom used by 1 for each constraint at each knot

Degrees of Freedom:

- How to think of degrees of freedom under the context of regression spline: each constraint that we impose on the piecewise cubic polynomials effectively frees up one degree of freedom, by reducing the complexity of the resulting piecewise polynomial fit
- How to calculate degrees of freedom
  - A degree-d spline with K knots has K + 1 polynomials of degree d, giving (K + 1)(d + 1) coefficients. However, there are d continuity restrictions at each of the K knots, so it uses $(K + 1)(d + 1) - Kd = K + d + 1$ degrees of freedom. So a cubic spline (d = 3) with K knots uses a total of K + 4 degrees of freedom

The Spline Basis Representation:

- Recall the Basis Function's expression:

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

- Using this idea, a cubic spline with K knots can be modelled as

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

Truncated Power Basis Function:

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

, where ξ is the knot

- ξ is pronounced as 'ksi' and is the 14th letter of the modern Greek alphabet
- Adding a term of the form $\beta_4 h(x, \xi)$ to the model for a cubic polynomial will impose continuity in the first and second derivative at ξ; a discontinuity will exist in only the third derivative at ξ

Natural Spline:

- Definition: A natural spline is a spline with additional boundary constraints: the function go off linearly beyond the range of the data
- A natural spline (figure 1) with K knots has K degrees of freedom because you get back two degrees of freedom for the two constraints on each of the boundaries
  - Pros: this additional constraint gives more stable estimates at the boundaries and is therefore superior to regression spline
  - Cons: Since the natural spline takes two extra df, its flexibility decreases

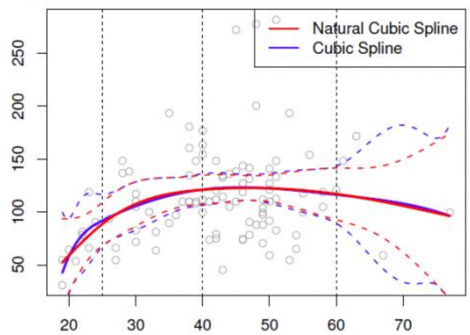


Figure 1

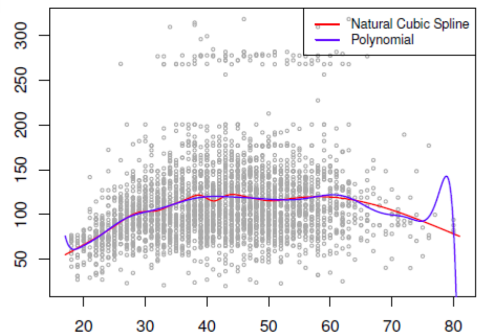


Figure 2

- Degrees of Freedom in Natural Splines = k+4 (cubic spline) + 2 (endpoint knots) - 4 (2 added constraints for linearity at each end), so df=k+2

Choosing Number & Locations of the Knots:

- One common practice is to place knots at uniform quantiles of the data
- Or, strategically place more where function might vary most rapidly and vice versa
- Use cross-validation to determine the best number of knots as well as their locations

Code in R:

- `bs(x, degree, knots or df)` creates the spline basis matrix, specifying df creates df-1 knots
- `ns(x, knots or df)` creates the natural degree 3 spline basis matrix, specifying df creates df-3 knots
- `fit.spline=lm(y ~ bs(x, degree, knots or df), data)` (can also use glm, especially for cv.glm)
- `fit.natural_spline=lm(y ~ ns(x, knots or df), data)` (can also use glm, especially for cv.glm)

Conclusion:

- Figure 2 above is a comparison of a d-14 polynomial and a natural cubic spline, each with 15df
- Advantages and disadvantages of regressions splines:
  - Regression splines often give superior results to polynomial regression because unlike polynomials, which must use a high degree to produce flexible fits, splines introduce flexibility by increasing the number of knots but keeping the degree fixed
  - Splines also allow us to place more knots, and hence flexibility, over regions where the function f seems to be changing rapidly, and fewer knots where f appears more stable