Splines: flexible models for nonlinear effects

Martyn Plummer

University of Warwick

03 June 2023

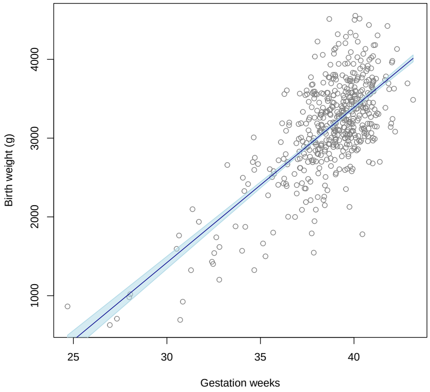
Overview

Introduction

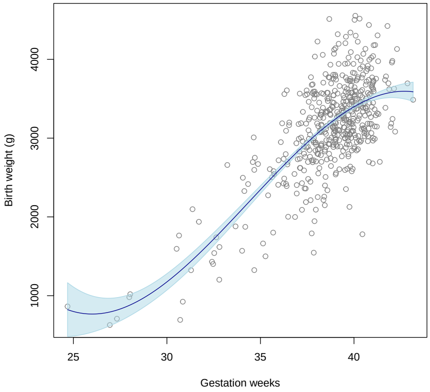
- Splines are a flexible class of models that can be helpful for representing dose-response relationships in epidemiology.
- In this course we will be using spline models extensively.
- However, spline models are widely misunderstood.
- The purpose of this lecture is to give a conceptual background on where spline models come from.

Outline

A linear model for the births data



A cubic model for the births data

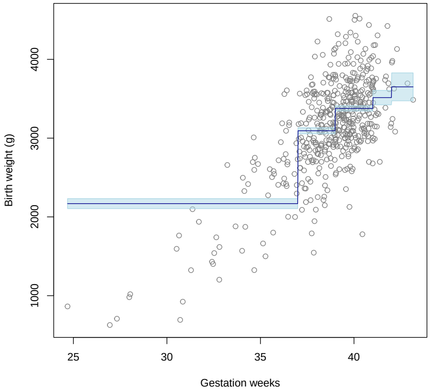


Categories

Medical doctors like to think in terms of categories

- **preterm** < 37 weeks
- **early term** 37-39 weeks
- **full term** 39-41 weeks
- **late term** 41-42 weeks
- **post term** ≥ 42 weeks

Fitting a categorical model

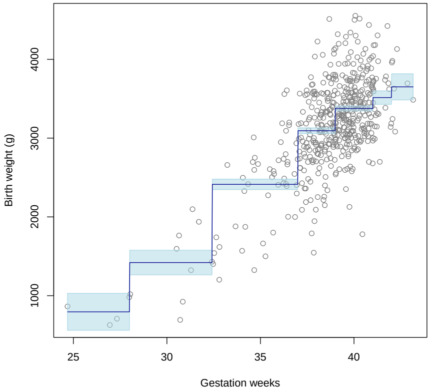


More categories

The poor fit for the category “preterm” can be improved by adding more categories:

- **extremely preterm** < 28 weeks
- **very preterm** 28-32 weeks
- **moderate to late preterm** 32-37

Plotting a more detailed categorical model



Statisticians against categorization

- Greenland S (1995) Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis, Epidemiology, **6**, 450–454.
- Senn S (2005) Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials.
- Bennette C, and Vickers A, (2012), Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology 12:21

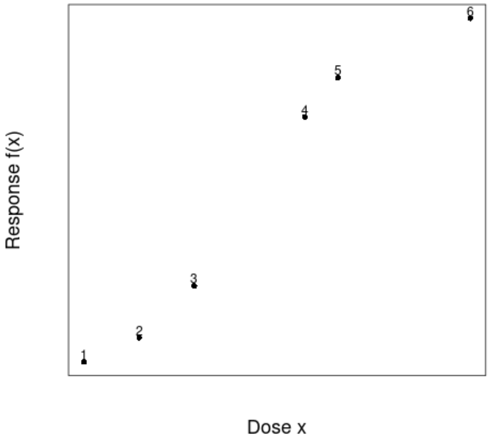
Epidemiologists against categorization

Rose, G. (1992) The Strategy of Preventive Medicine

- Many diseases are not discrete. Instead there is an underlying continuum of increasing severity (e.g. hypertension).
- In medicine, we tend to conflate a clinical action (treat vs. do not treat) with the presence/absence of disease.
- Disease prevention efforts are best targeted at shifting the distribution of risk for the whole population instead of trying to identify and target a "high risk" group.

Outline

Join the dots



- Suppose that we have a set of (x, y) points that we think come from an underlying smooth relationship between x and y.
- We want to join the dots in a way that is as smooth as possible.
- This turns out to be a mathematically well defined problem with a unique solution.

A roughness penalty

- Suppose $y = f(x)$ for some function $f(.)$.
- The *roughness* of the curve in the interval $[a, b]$ is meausured by the integral

$$\int_a^b \left(\frac{\partial^2 f}{\partial x^2} \right)^2 dx$$

- We want the roughness of f to be as small as possible.

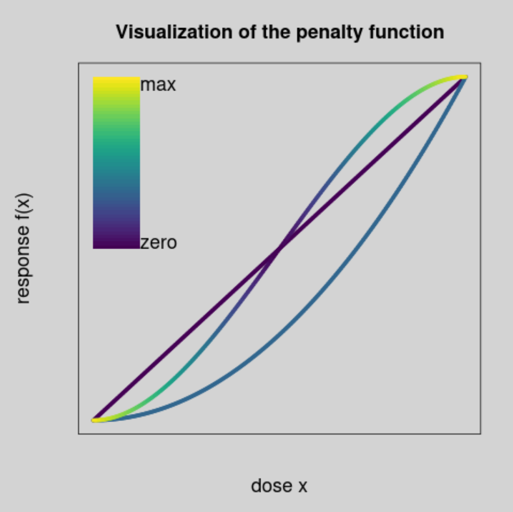
A roughness penalty

- Suppose $y = f(x)$ for some function $f(.)$.
- The *roughness* of the curve in the interval $[a, b]$ is meausured by the integral

$$\int_a^b \left(\frac{\partial^2 f}{\partial x^2}\right)^2 dx$$

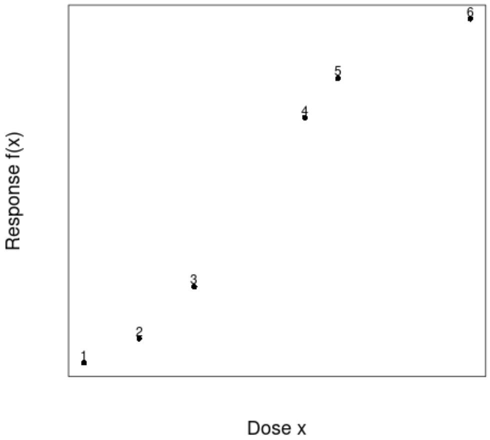
- We want the roughness of f to be as small as possible.

What does the roughness penalty mean?



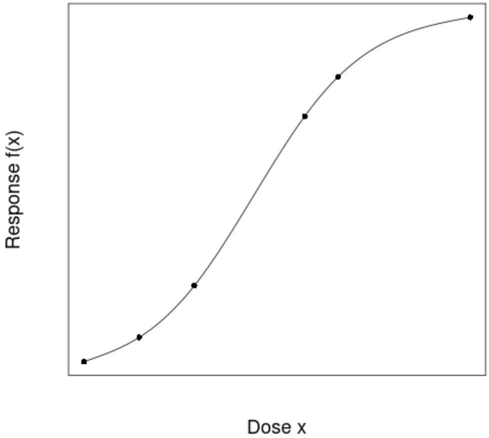
- The contribution to the penalty at each point depends on the curvature (represented by a colour gradient)
- A straight line has no curvature, hence zero penalty.
- Sharp changes in the slope are heavily penalized.

An interpolating cubic spline



- The smoothest curve that goes through the observed points is a cubic spline.

An interpolating cubic spline



- The smoothest curve that goes through the observed points is a cubic spline.

What is a cubic spline?

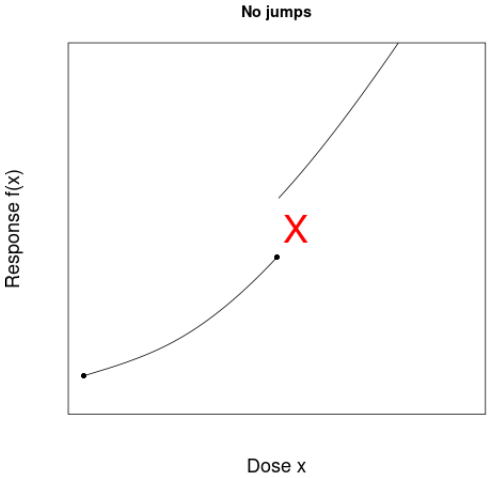
Splines are piecewise cubic curves

- Every observed point is a knot.
- The knots divide the curve into sections
- Each section is a cubic function

$$f(x) = a + bx + cx^2 + dx^3$$

- The parameters a, b, c, d are different for different sections

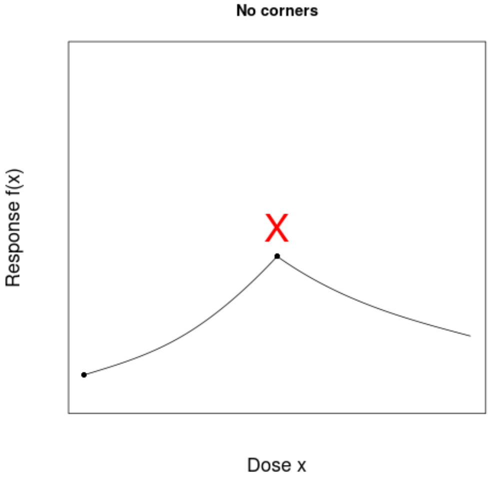
Boundary conditions



Sections need to join up smoothly.

- Both sides must go through the knot.
- The slope cannot change at a knot
- The curvature cannot change at a knot

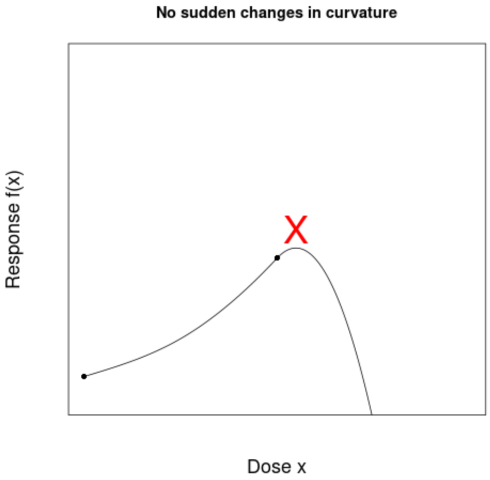
Boundary conditions



Sections need to join up smoothly.

- Both sides must go through the knot.
- The slope cannot change at a knot
- The curvature cannot change at a knot

Boundary conditions

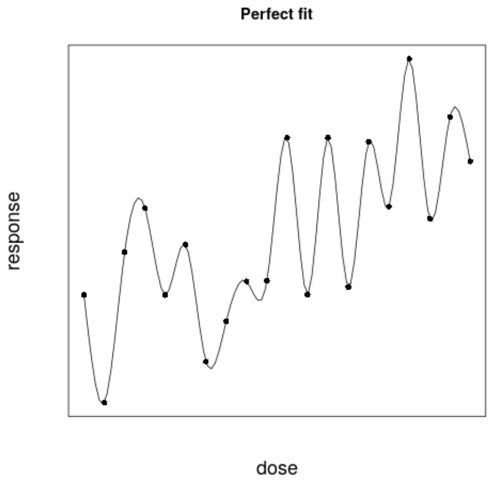


Sections need to join up smoothly.

- Both sides must go through the knot.
- The slope cannot change at a knot
- The curvature cannot change at a knot

Outline

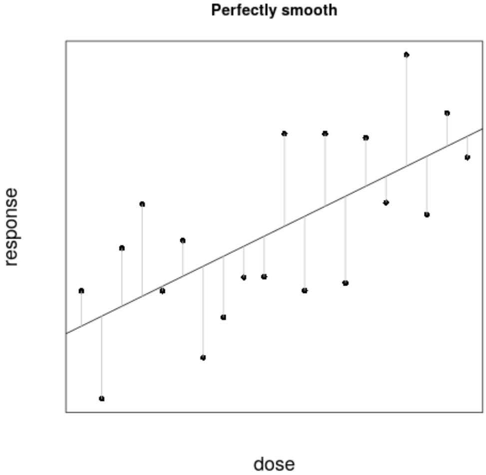
Dose response with error



In practice we never know the dose response curve exactly at any point but always measure with error. A spline model is then a compromise between

- Model fit
- Smoothness of the spline

Dose response with error



In practice we never know the dose response curve exactly at any point but always measure with error. A spline model is then a compromise between

- Model fit
- Smoothness of the spline

Fitting a smoothing spline

Minimize

$$\sum_i [y_i - f(x_i)]^2 + \lambda \int \left(\frac{\partial^2 f}{\partial x^2} \right)^2 dx$$

Or, more generally

$$\text{Deviance } + \lambda \times \text{Roughness penalty}$$

Size of tuning parameter λ determines compromise between model fit (small λ) and smoothness (large λ).

Smoothing and degrees of freedom

Software will choose the smoothing parameter λ for you automatically using cross-validation.

The smoothing parameter is adapted to the data.

Smoothness of the model can be measured with the *effective degrees of freedom* (EDF)

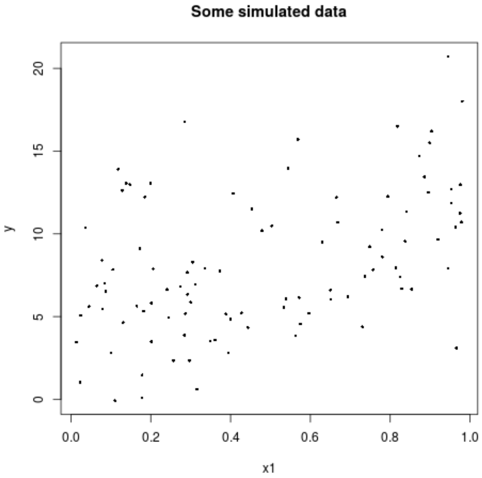
- Linear model: maximally smooth
  - EDF=2 (intercept + slope parameter)
- Intepolating mode: best fit
  - EDF=n (one parameter for every observation)

Outline

Spline models in R

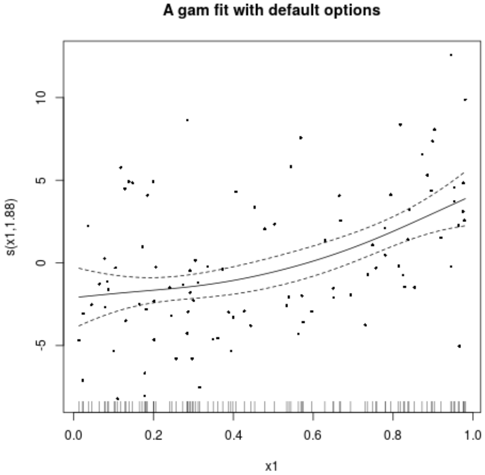
- Do not use the splines package.
- Use the `gam` function from the mgcv package to fit your spline models.
- The `gam` function chooses number and placement of knots for you and estimates the size of the tuning parameter λ automatically.
- You can use the `gam.check` function to see if you have enough knots. Also re-fit the model explicitly setting a larger number of knots (e.g. double) to see if the fit changes.

Penalized spline



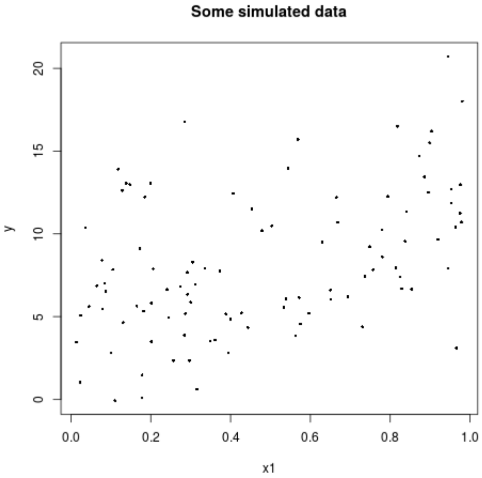
- A gam fit to some simulated data
- Model has 9 degrees of freedom
- Smoothing reduces this to 2.88 effective degrees of freedom

Penalized spline



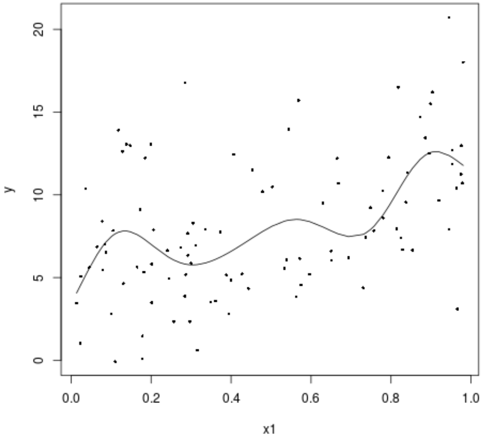
- A gam fit to some simulated data
- Model has 9 degrees of freedom
- Smoothing reduces this to 2.88 effective degrees of freedom

Unpenalized spline



- An unpenalized spline using the same spline basis as the gam fit.
- Model has 9 degrees of freedom

Unpenalized spline



- An unpenalized spline using the same spline basis as the gam fit.
- Model has 9 degrees of freedom

Conclusions

- Epidemiologists like to turn continuous variables into categories.
- Statisticians do not like categorization because it loses information.
- Splines are a flexible class of models that avoid categorization but also avoid making strong assumptions about the shape of a dose-response relationship.
- Penalized regression splines are based on compromise between goodness-of-fit and smoothness.
- Most of the decisions in fitting a penalized regression spline can be made for you
    - Degree of smoothing
    - Number of knots
    - Placement of knots