

Everything You Ever Wanted to Know about Splines...

Martyn Plummer

Infection and Cancer Epidemiology Group, IARC

15 June 2018

Overview

Categorization and its discontents

Join the dots

Brownian motion

Smoothing splines

Conclusions

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

Categorization and its discontents

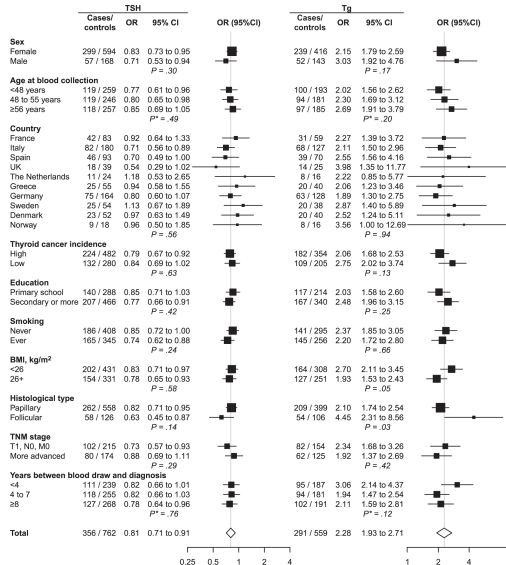
Join the dots

Brownian motion

Smoothing splines

Conclusions

Rinaldi et al, JNCI. 2014 Jun;106(6):dju097



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

Categorization and its discontents

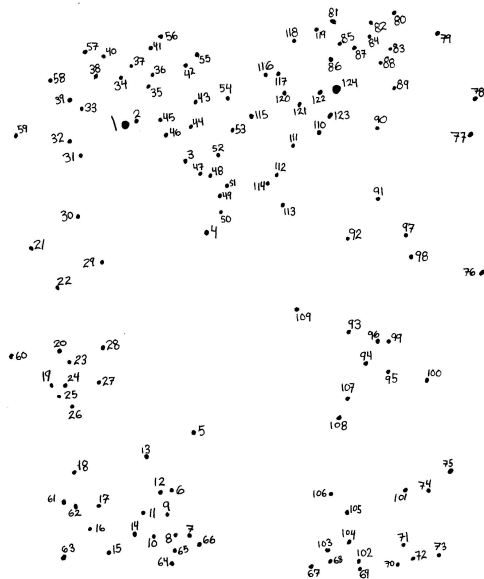
Join the dots

Brownian motion

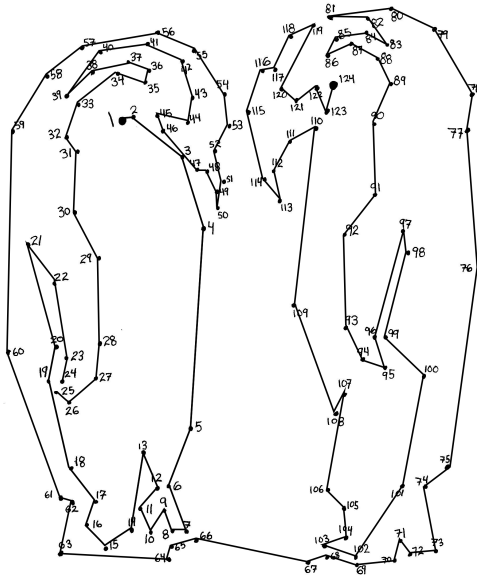
Smoothing splines

Conclusions

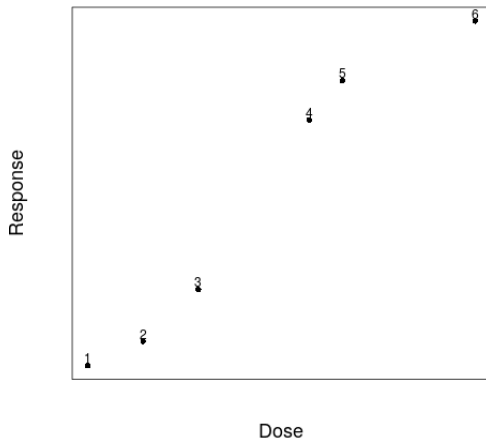
Join the dots



Join the dots

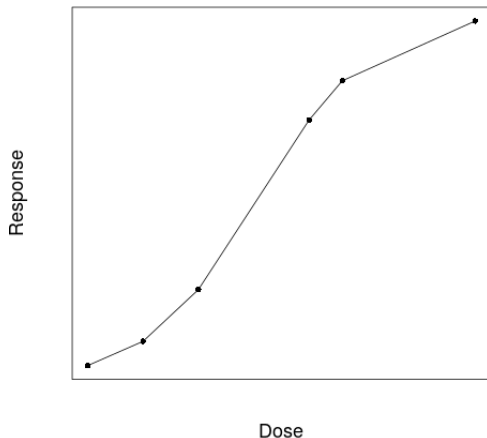


Linear interpolation



- Suppose a dose response curve is known exactly at certain points
- We can fill in the gaps (interpolate) by drawing a straight (linear) line between adjacent points

Linear interpolation



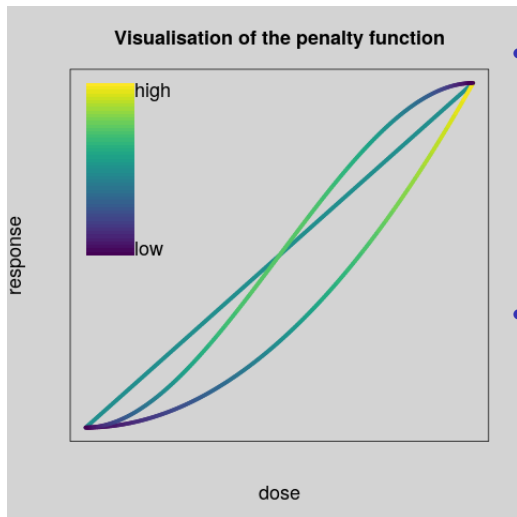
- Suppose a dose response curve is known exactly at certain points
- We can fill in the gaps (interpolate) by drawing a straight (linear) line between adjacent points

Why linear interpolation?

Out of all possible curves that go through the observed points, linear interpolation is the one that minimizes the penalty function

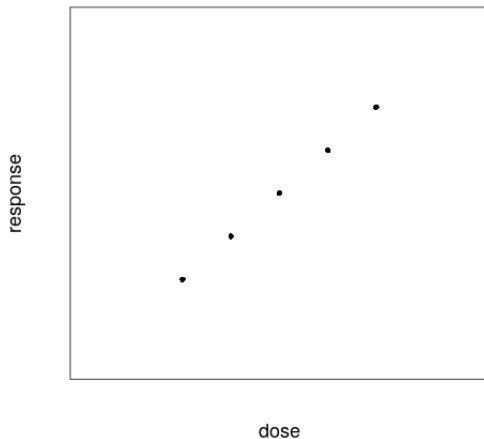
$$\int \left(\frac{\partial f}{\partial x} \right)^2 dx$$

What does the penalty mean?



- The contribution to the penalty at each point depends on the steepness of the curve (represented by a colour gradient)
- Any deviation from a straight line between the two fixed points will incur a higher penalty overall.

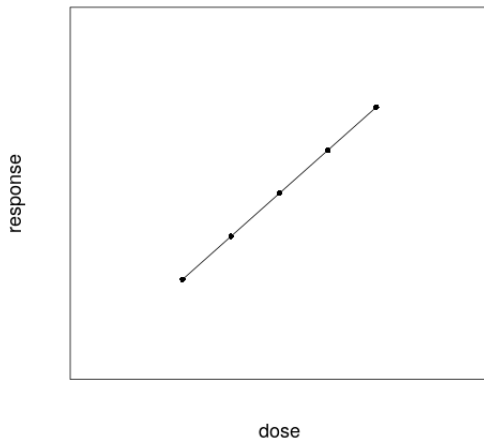
Extrapolation



- Linear interpolation fits a linear dose-response curve exactly
- But it breaks down when we try to extrapolate

Extrapolation

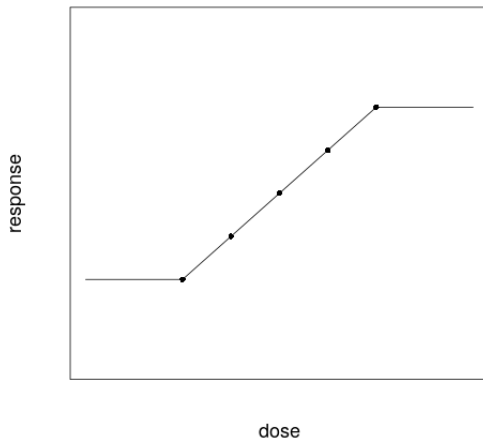
Linear interpolation



- Linear interpolation fits a linear dose-response curve exactly
- But it breaks down when we try to extrapolate

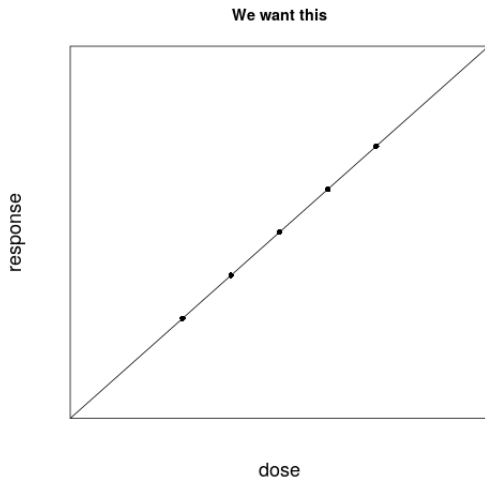
Extrapolation

Extrapolation - not what we want



- Linear interpolation fits a linear dose-response curve exactly
- But it breaks down when we try to extrapolate

Extrapolation



- Linear interpolation fits a linear dose-response curve exactly
- But it breaks down when we try to extrapolate

Why does linear interpolation break down?

- The penalty function

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

penalizes the steepness of the curve

- Minimizing the penalty function gives us the “flattest” curve that goes through the points.
 - In between two observations the flattest curve is a straight line.
 - Outside the range of the observations the flattest curve is completely flat.

A roughness penalty

- If we want a fitted curve that extrapolates a linear trend then we want to minimize the curvature.

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

- Like the first penalty function but uses the second derivative of f (i.e. the curvature).
- This is a roughness penalty.

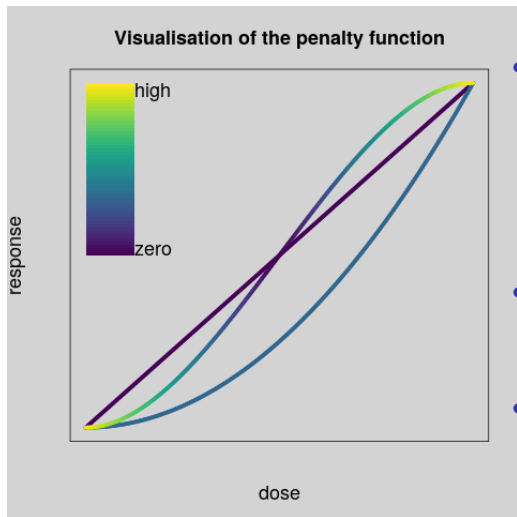
A roughness penalty

- If we want a fitted curve that extrapolates a linear trend then we want to minimize the **curvature**.

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

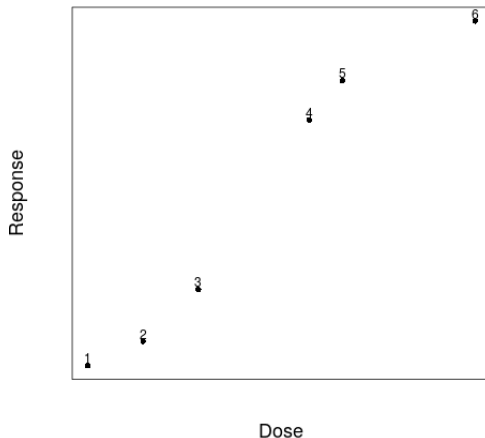
- Like the first penalty function but uses the **second derivative** of f (i.e. the curvature).
- This is a roughness penalty.

What does the roughness penalty mean?



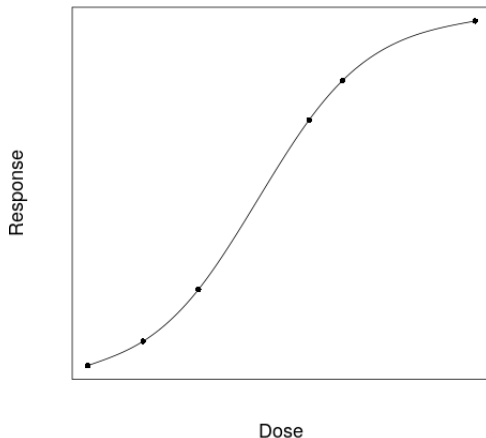
- The contribution to the penalty at each point depends on the curvature (represented by a colour gradient)
- A straight line has no curvature, hence zero penalty.
- Sharp changes in the slope are heavily penalized.

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.

Properties of cubic splines

- A cubic spline consists of a sequence of curves of the form

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

for some coefficients a, b, c, d , in between each observed point.

- The cubic curves are joined at the observed points (knots)
- The cubic curves match where they meet at the knots
 - Same value $f(x)$
 - Same slope $\partial f / \partial x$
 - Same curvature $\partial^2 f / \partial x^2$

Outline

Categorization and its discontents

Join the dots

Brownian motion

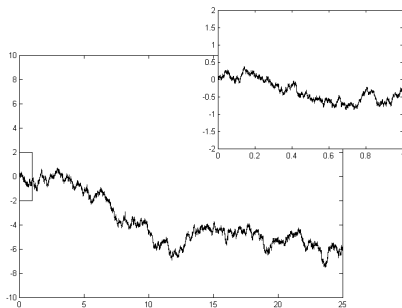
Smoothing splines

Conclusions

Brownian motion

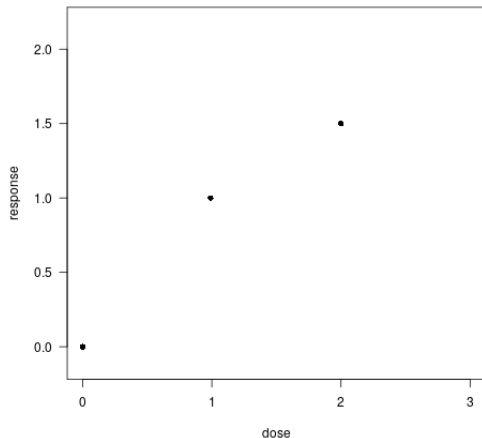
- In 1827, botanist Robert Brown observed particles under the microscope moving randomly
- Theoretical explanation by Einstein (1905) in terms of water molecules
- Verified by Perrin (1908). Nobel prize in physics 1927.

Evolution of 1-dimensional Brownian motion with time



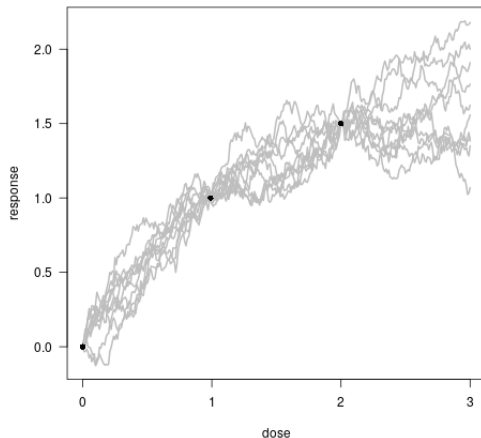
- In mathematics a Brownian motion is a stochastic process that randomly goes up or down at any time point
- Also called a Wiener process after American mathematician Norbert Wiener.
- A Brownian motion is fractal – it looks the same if you zoom in and rescale

A partially observed Brownian motion



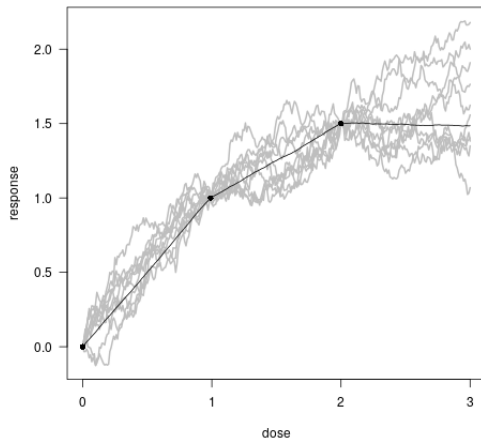
- Suppose we observe a Brownian motion at three points
- Grey lines show a sample of possible paths through the points
- The black line shows the average over all paths

A partially observed Brownian motion



- Suppose we observe a Brownian motion at three points
- Grey lines show a sample of possible paths through the points
- The black line shows the average over all paths

A partially observed Brownian motion



- Suppose we observe a Brownian motion at three points
- Grey lines show a sample of possible paths through the points
- The black line shows the average over all paths

Statistical model for linear interpolation

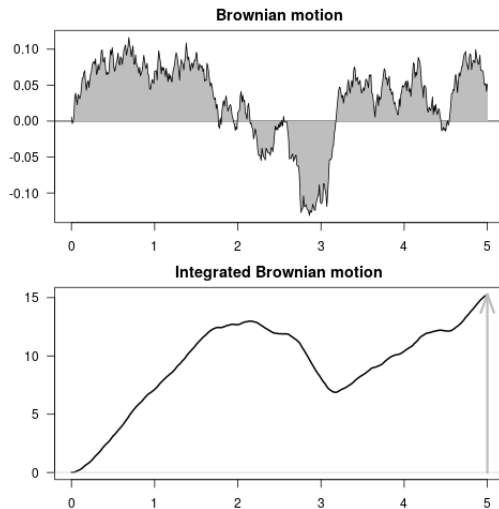
- Suppose the curve f is generated by the underlying model

$$f(x) = \alpha + \sigma W(x)$$

where W (for Wiener process) is a Brownian motion

- Then given points $(x_1, f(x_1)) \dots (x_n, f(x_n))$ the *expected value* of f is the curve we get from linear interpolation.

Integrated Brownian motion



- The value of an integrated Brownian motion is the area under the curve (AUC) of a Brownian motion up to that point.
- AUC goes down when the Brownian motion takes a negative value.

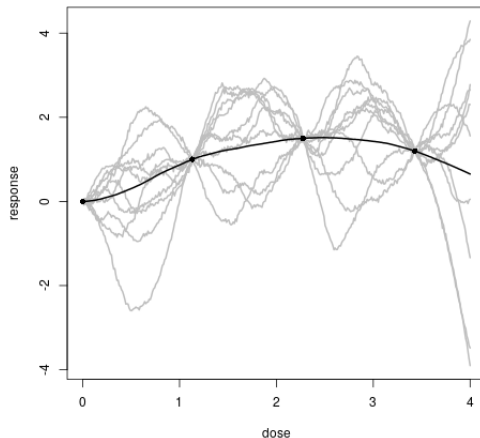
Integrated Brownian motion with drift

Add a mean parameter and a linear trend (drift) to the integrated Brownian motion:

$$f(x) = \alpha + \beta x + \sigma \int_0^x W(z) dz$$

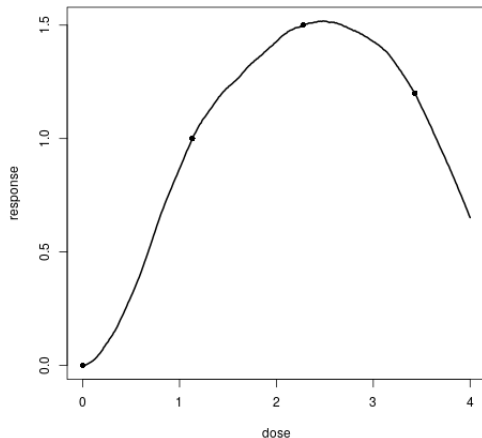
This more complex model is capable of modelling smooth curves.

A partially observed integrated Brownian motion with drift



- Grey lines show a sample of possible paths through the points
- The black line shows the average over all paths

Zoom on the expected value



- The expected value is a cubic spline.
- Extrapolation beyond the boundary of the points is linear (natural spline).

The smoothness paradox

- A cubic natural spline is the smoothest curve that goes through a set of points.
- But the underlying random process $f(x)$ is nowhere smooth.
- $f(x)$ is constantly changing its slope based on the value of the underlying Brownian motion.

The knot paradox

- There are no knots in the underlying model for a cubic natural spline.
- Knots are a result of the observation process.

Outline

Categorization and its discontents

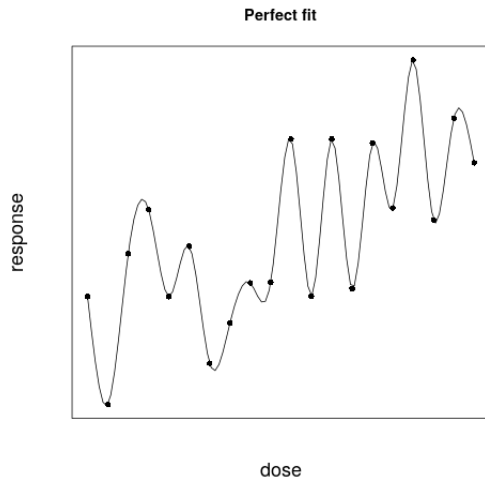
Join the dots

Brownian motion

Smoothing splines

Conclusions

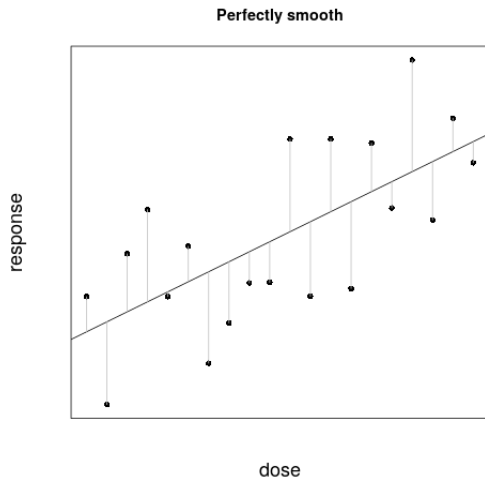
Dose response with error



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

Deviance + λ * Roughness penalty

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

How to choose the tuning parameter λ

This is a statistical problem. There are various statistical approaches:

- Restricted maximum likelihood (REML)
- Cross-validation
- Bayesian approach (with prior on smoothness)

At least the first two should be available in most software.

Outline

Categorization and its discontents

Join the dots

Brownian motion

Smoothing splines

Conclusions

Spline models done badly

- Choose number and placement of knots
- Create a spline bases
- Use spline basis as the design matrix in a generalized linear model.
- Without penalization, model will underfit (too few knots) or overfit (too many knots)
- Placement of knots may create artefacts in the dose-response relationship

Spline models done well

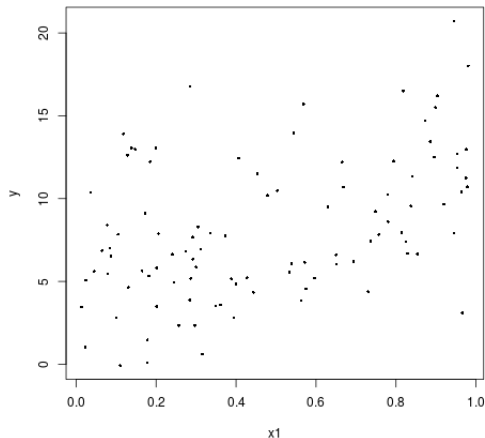
- A knot for every observed value (remember: knots are a product of the observation process).
- Use penalization: find the right compromise between model fit and model complexity.
- In practice we can get a good approximation to this “ideal” model with fewer knots.
- This assumption should be tested

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

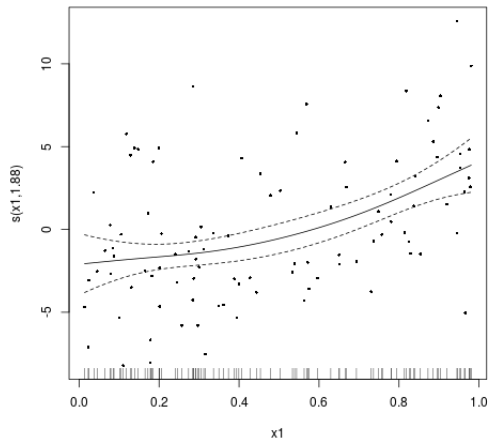
Some simulated data



- 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

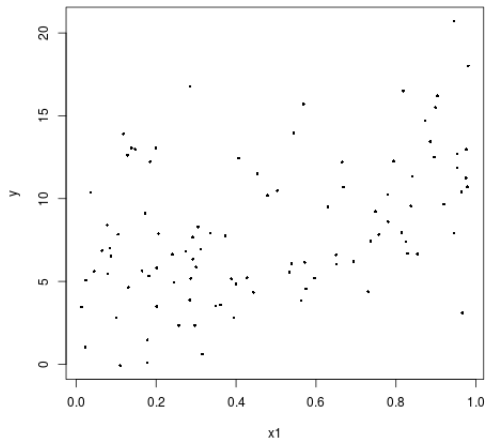
A gam fit with default options



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

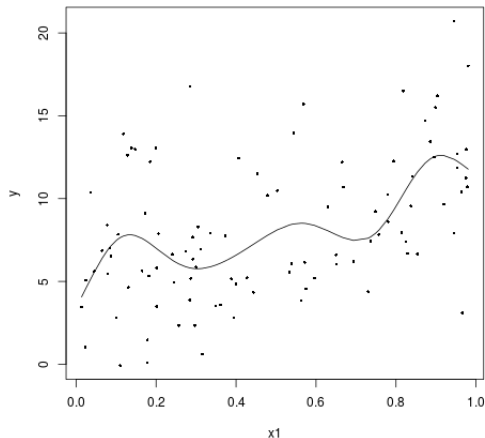
Unpenalized spline

Some simulated data



- 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