Everything you ever wanted to know about splines but were too afraid to ask

Martyn Plummer

University of Warwick

03 June 2023

Overview

Categorization and its discontents

Join the dots

Smoothing splines

Splines in R

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

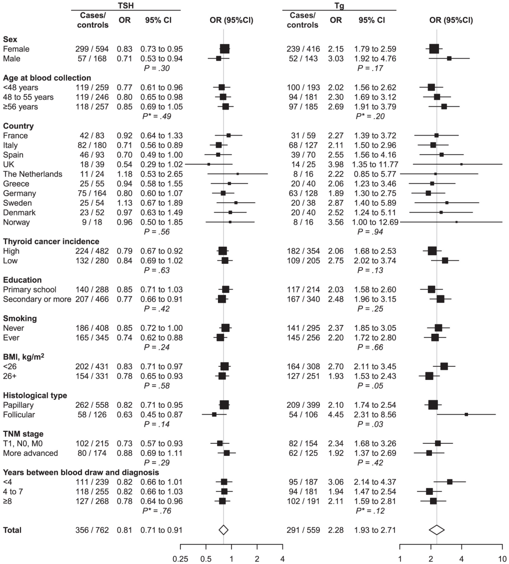
Categorization and its discontents

Join the dots

Smoothing splines

Splines in R

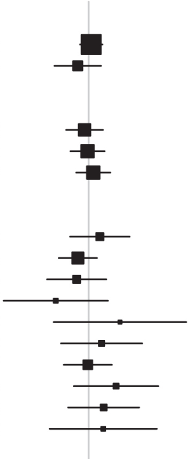
Rinaldi et al, JNCI. 2014 Jun;106(6):dju097



	TSH Cases/controls	TSH OR	TSH 95% CI	Tg Cases/controls	Tg OR	Tg 95% CI
Sex						
Female	299 / 594	0.83	0.73 to 0.95	239 / 416	2.15	1.79 to 2.59
Male	57 / 168	0.71	0.53 to 0.94	52 / 143	3.03	1.92 to 4.76
			P = .30			*P* = .17
Age at blood collection						
<48 years	119 / 259	0.77	0.61 to 0.96	100 / 193	2.02	1.56 to 2.62
48 to 55 years	119 / 246	0.80	0.65 to 0.98	94 / 181	2.30	1.69 to 3.12
≥56 years	118 / 257	0.85	0.69 to 1.05	97 / 185	2.69	1.91 to 3.79
			*P** = .49			*P** = .20
Country						
France	42 / 83	0.92	0.64 to 1.33	31 / 59	2.27	1.39 to 3.72
Italy	82 / 180	0.71	0.56 to 0.89	68 / 127	2.11	1.50 to 2.96
Spain	46 / 93	0.70	0.49 to 1.00	39 / 70	2.55	1.56 to 4.16
UK	18 / 39	0.54	0.29 to 1.02	14 / 25	3.98	1.35 to 11.77
The Netherlands	11 / 24	1.18	0.53 to 2.65	8 / 16	2.22	0.85 to 5.77
Greece	25 / 55	0.94	0.58 to 1.55	20 / 40	2.06	1.23 to 3.46
Germany	75 / 164	0.80	0.60 to 1.07	63 / 128	1.89	1.30 to 2.75
Sweden	25 / 54	1.13	0.67 to 1.89	20 / 38	2.87	1.40 to 5.89
Denmark	23 / 52	0.97	0.63 to 1.49	20 / 40	2.52	1.24 to 5.11
Norway	9 / 18	0.96	0.50 to 1.85	8 / 16	3.56	1.00 to 12.69
			P = .56			*P* = .94
Thyroid cancer incidence						
High	224 / 482	0.79	0.67 to 0.92	182 / 354	2.06	1.68 to 2.53
Low	132 / 280	0.84	0.69 to 1.02	109 / 205	2.75	2.02 to 3.74
			P = .63			*P* = .13
Education						
Primary school	140 / 288	0.85	0.71 to 1.03	117 / 214	2.03	1.58 to 2.60
Secondary or more	207 / 466	0.77	0.66 to 0.91	167 / 340	2.48	1.96 to 3.15
			P = .42			*P* = .25
Smoking						
Never	186 / 408	0.85	0.72 to 1.00	141 / 295	2.37	1.85 to 3.05
Ever	165 / 345	0.74	0.62 to 0.88	145 / 256	2.20	1.72 to 2.80
			P = .24			*P* = .66
BMI, kg/m²						
<26	202 / 431	0.83	0.71 to 0.97	164 / 308	2.70	2.11 to 3.45
26+	154 / 331	0.78	0.65 to 0.93	127 / 251	1.93	1.53 to 2.43
			P = .58			*P* = .05
Histological type						
Papillary	262 / 558	0.82	0.71 to 0.95	209 / 399	2.10	1.74 to 2.54
Follicular	58 / 126	0.63	0.45 to 0.87	54 / 106	4.45	2.31 to 8.56
			P = .14			*P* = .03
TNM stage						
T1, N0, M0	102 / 215	0.73	0.57 to 0.93	82 / 154	2.34	1.68 to 3.26
More advanced	80 / 174	0.88	0.69 to 1.11	62 / 125	1.92	1.37 to 2.69
			P = .29			*P* = .42
Years between blood draw and diagnosis						
<4	111 / 239	0.82	0.66 to 1.01	95 / 187	3.06	2.14 to 4.37
4 to 7	118 / 255	0.82	0.66 to 1.03	94 / 181	1.94	1.47 to 2.54
≥8	127 / 268	0.78	0.64 to 0.96	102 / 191	2.11	1.59 to 2.81
			*P** = .76			*P** = .12
Total	356 / 762	0.81	0.71 to 0.91	291 / 559	2.28	1.93 to 2.71

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

	TSH			OR (95%CI)
	Cases/ controls	OR	95% CI	
Sex				
Female	299 / 594	0.83	0.73 to 0.95	
Male	57 / 168	0.71	0.53 to 0.94	
			P = .30	
Age at blood collection				
<48 years	119 / 259	0.77	0.61 to 0.96	
48 to 55 years	119 / 246	0.80	0.65 to 0.98	
≥56 years	118 / 257	0.85	0.69 to 1.05	
			P = .49*	
Country				
France	42 / 83	0.92	0.64 to 1.33	
Italy	82 / 180	0.71	0.56 to 0.89	
Spain	46 / 93	0.70	0.49 to 1.00	
UK	18 / 39	0.54	0.29 to 1.02	
The Netherlands	11 / 24	1.18	0.53 to 2.65	
Greece	25 / 55	0.94	0.58 to 1.55	
Germany	75 / 164	0.80	0.60 to 1.07	
Sweden	25 / 54	1.13	0.67 to 1.89	
Denmark	23 / 52	0.97	0.63 to 1.49	
Norway	9 / 18	0.96	0.50 to 1.85	
			P = .56	



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

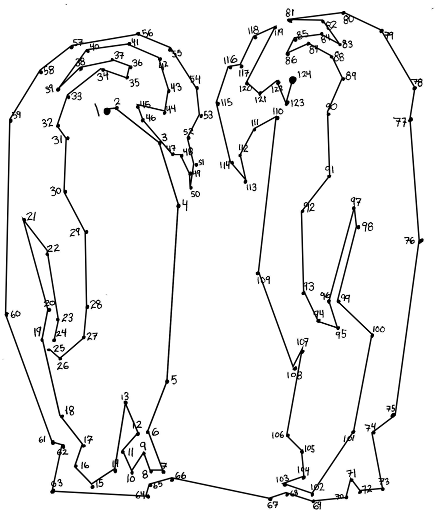
Join the dots

Smoothing splines

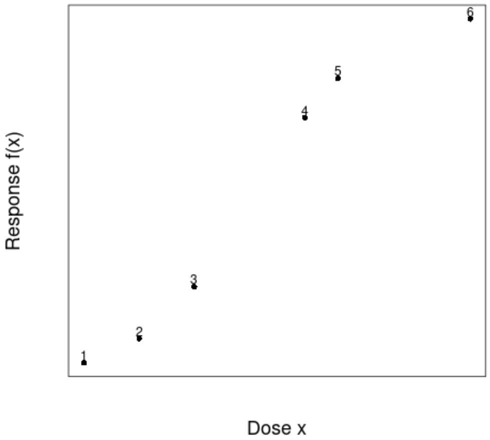
Splines in R

Join the dots

Join the dots

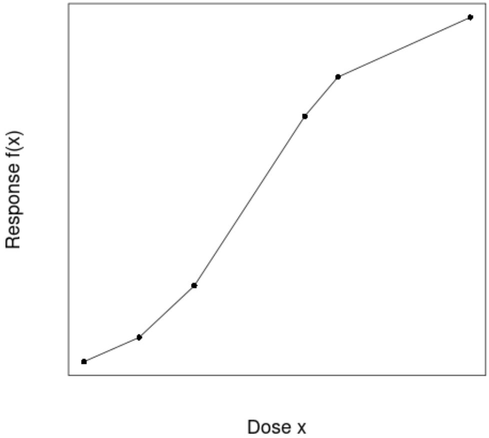


Linear interpolation



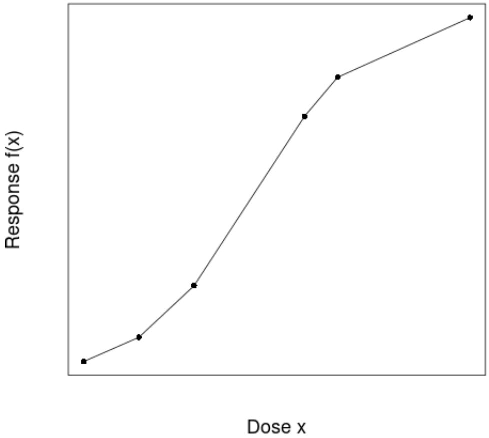
- Suppose a dose response curve is known exactly at certain points
- We can fill in the gaps (interpolate) by drawing a straight (linear) line between adjacent points
- This creates a mathematical function $f()$ which gives a response value $f(x)$ for every dose value x.

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
- This creates a mathematical function $f()$ which gives a response value $f(x)$ for every dose value x.

Linear interpolation



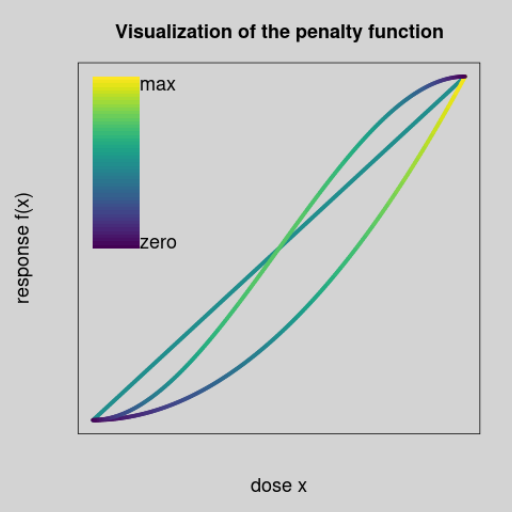
- Suppose a dose response curve is known exactly at certain points
- We can fill in the gaps (interpolate) by drawing a straight (linear) line between adjacent points
- This creates a mathematical function $f()$ which gives a response value $f(x)$ for every dose value x.

Why linear interpolation?

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

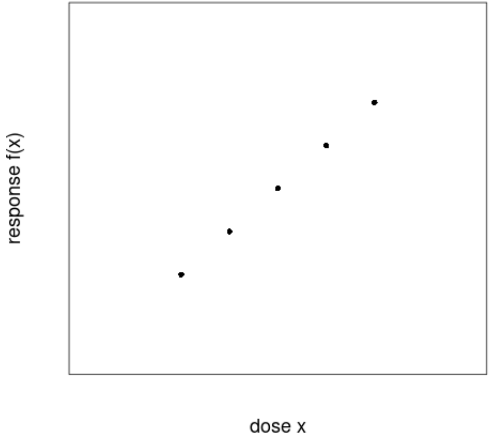
$$\int \left(\frac{\partial f}{\partial x} \right)^2 dx$$

What does the penalty mean?



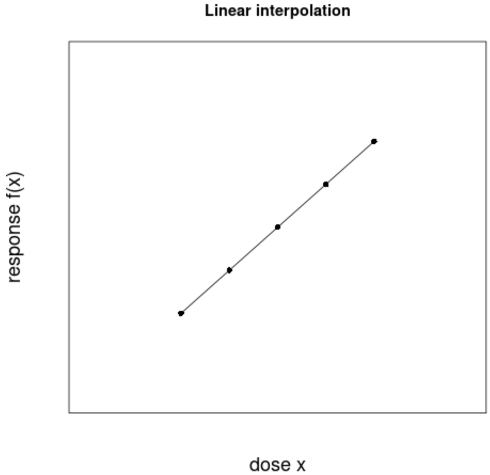
- The contribution to the penalty at each point depends on the steepness of the curve (represented by a colour gradient)
- Any deviation from a straight line between the two fixed points will incur a higher penalty overall.

Extrapolation



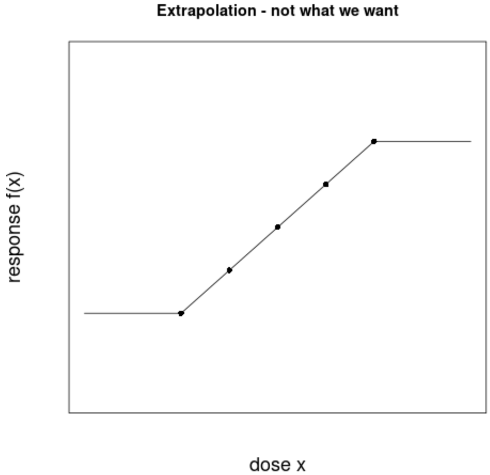
- Linear interpolation fits a linear dose-response curve exactly.
- But it breaks down when we extrapolate.

Extrapolation



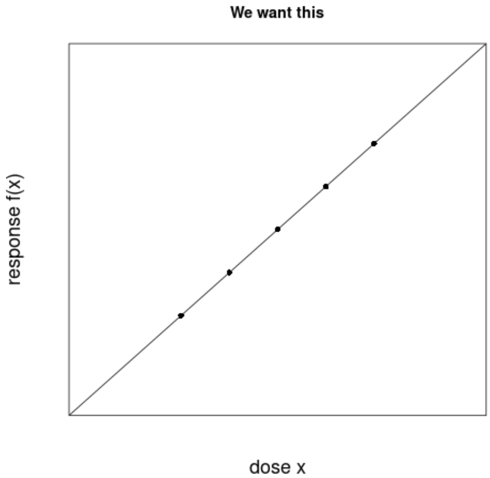
- Linear interpolation fits a linear dose-response curve exactly.
- But it breaks down when we extrapolate.

Extrapolation



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

Extrapolation



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

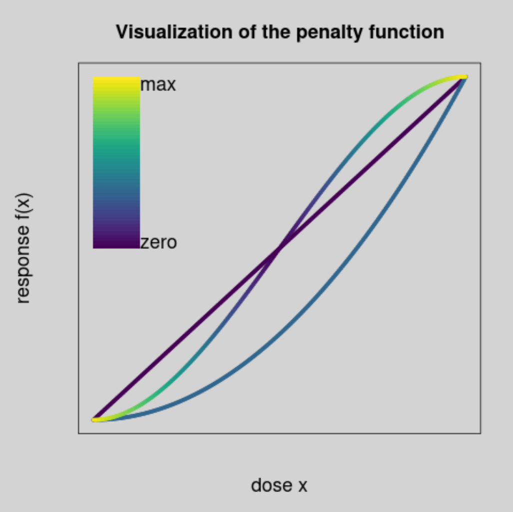
A roughness penalty

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

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

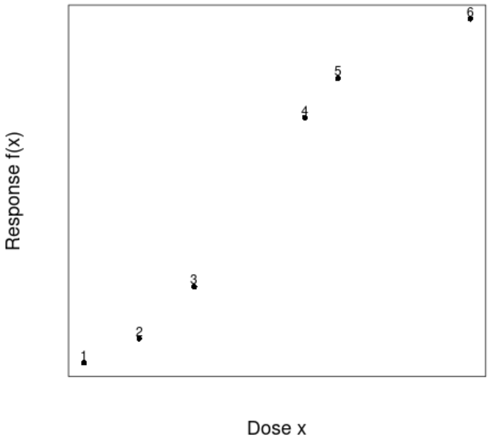
- Like the first penalty function but uses the second derivative of f (i.e. the curvature).
- This is a roughness penalty.

What does the roughness penalty mean?



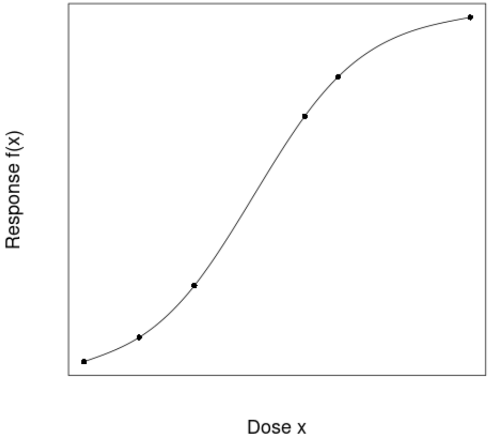
- The contribution to the penalty at each point depends on the curvature (represented by a colour gradient)
- A straight line has no curvature, hence zero penalty.
- Sharp changes in the slope are heavily penalized.

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.

What is a cubic spline?

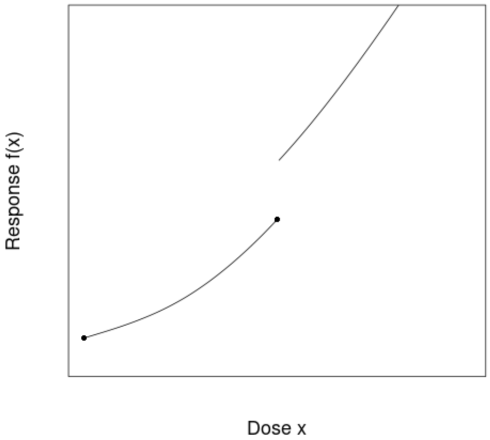
Splines are piecewise cubic curves

- 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

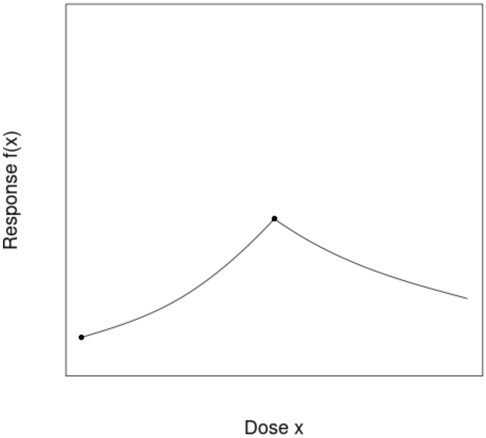
Boundary conditions



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

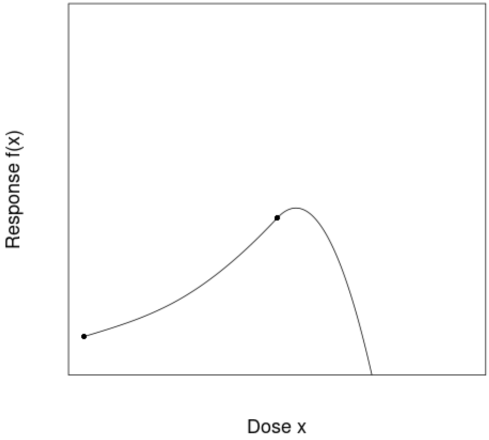
Boundary conditions



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

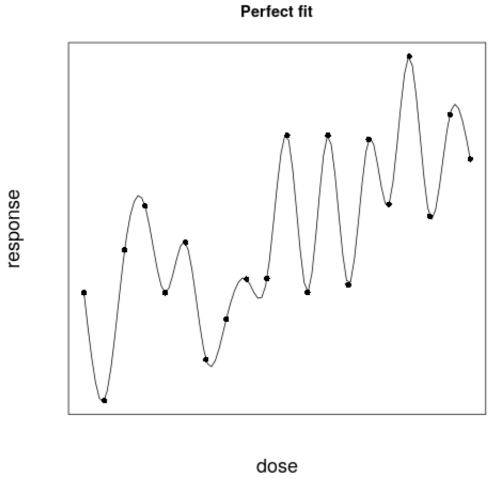
Categorization and its discontents

Join the dots

Smoothing splines

Splines in R

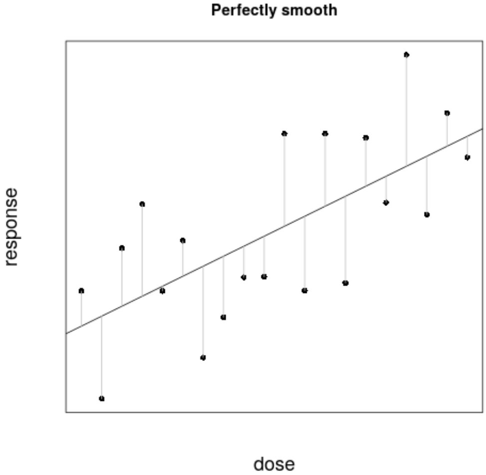
Dose response with error



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

Categorization and its discontents

Join the dots

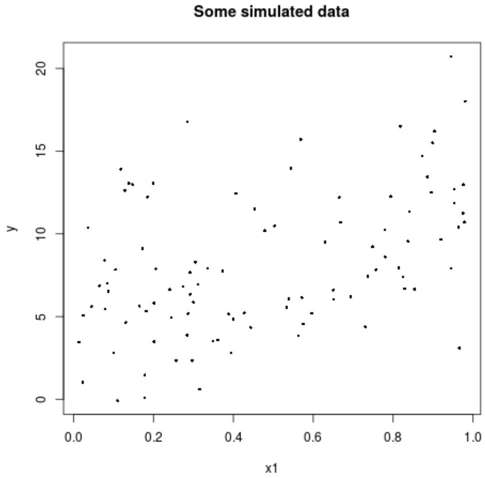
Smoothing splines

Splines in R

Spline models in R

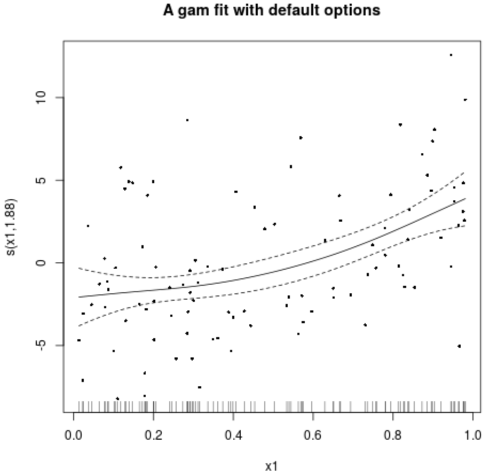
- Do not use the splines package.
- Use the `gam` function from the mgcv package to fit your spline models.
- The `gam` function chooses number and placement of knots for you and estimates the size of the tuning parameter λ automatically.
- You can use the `gam.check` function to see if you have enough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

Penalized spline



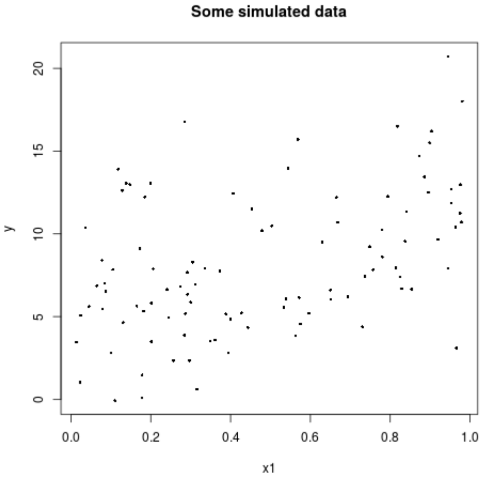
- A gam fit to some simulated data
- Model has 9 degrees of freedom
- Smoothing reduces this to 2.88 effective degrees of freedom

Penalized spline



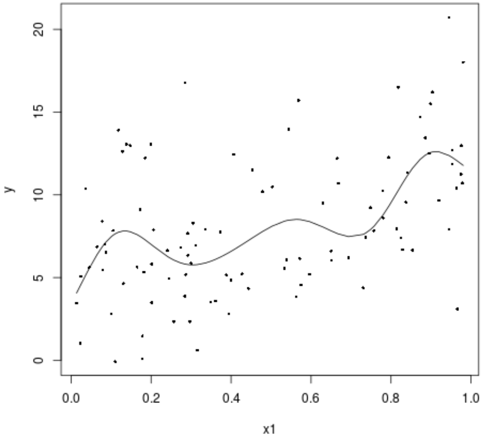
- A gam fit to some simulated data
- Model has 9 degrees of freedom
- Smoothing reduces this to 2.88 effective degrees of freedom

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