Spline Regression

MAST90083 Computational Statistics and Data Mining

Karim Seghouane

School of Mathematics & Statistics

The University of Melbourne

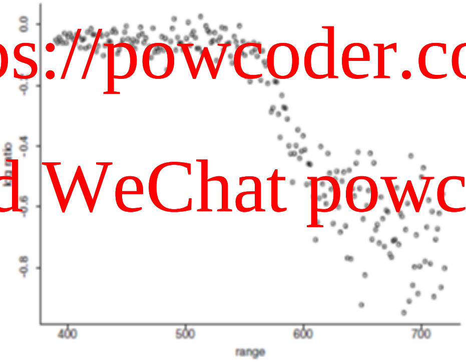
Outline

Introduction

- Some data sets are hard or impossible to model using traditional parametric techniques
- Many data sets also involve nonlinear effects that are difficult to model parametrically
- There is a need for flexible techniques to handle complicated nonlinear relationships
- Here we look at some ways of freeing oneself of the restrictions of parametric regression models

Introduction

The interest is the discovery of the underlying trend in the observed data which are treated as a collection of points on the plane



Introduction

- Alternatively, we could think of the vertical axis as a realization of a random variable y conditional on the variable x
- The underlying trend would then be a function

$$f(x) = E(y|x)$$

- This can also be written as

$$y_i = f(x_i) + \epsilon_i, \quad E(\epsilon_i) = 0$$

- and the problem is referred as nonparametric regression

Introduction

- Aim Estimate the unspecified smooth function from the pairs (x_i, y_i), $i = 1, \ldots, n$
- x here will be considered univariate
- There are several available methods, here we focus first on penalized splines

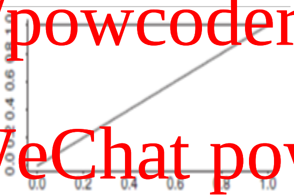
Motivation

- Let's start with the straight line regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Motivation

- The corresponding basis for this model are the functions: 1 and x



- The model is a linear combination of these functions which is the reason for use of the world basis

Motivation

- The basis functions correspond to the columns of X for fitting the regression

$$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

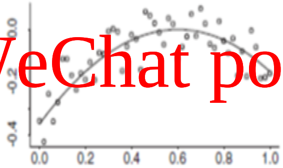
- The vector of fitted values

$$\hat{\mathbf{y}} = X\left(X^\top X\right)^{-1} X^\top \mathbf{y} = H\mathbf{y}$$

Motivation

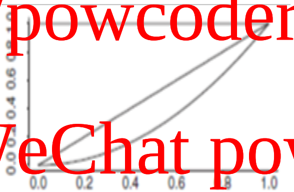
- The quadratic model is a simple extension of the linear model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$



Motivation

- There is an extra basis function x^2 corresponding to the addition of the $\beta_2 x_i^2$ term to the model



- The quadratic model is an example of how the simple linear model might be extended to handle nonlinear structure

Motivation

- The basis functions correspond to the columns of X for fitting the regression in the case of a quadratic model is given by

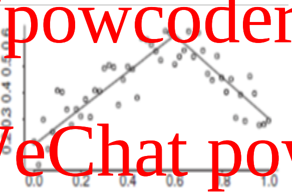
$$\begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

- The vector of fitted values

$$\hat{\mathbf{y}} = X\left(X^\top X\right)^{-1} X^\top \mathbf{y} = H\mathbf{y}$$

Spline basis function

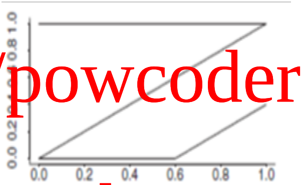
- We know look at how the model can be extended to accommodate a different type of nonlinear structure



- Broken line model: it consists of two differently sloped lines that join together

Spline basis function

- Broken line: A linear combination of three basis functions



- where we have $(x - 0.6)_+$ with

$$u_+ = \begin{cases} u & u > 0 \\ 0 & u \leq 0 \end{cases}$$

Spline basis function

- Broken line model is

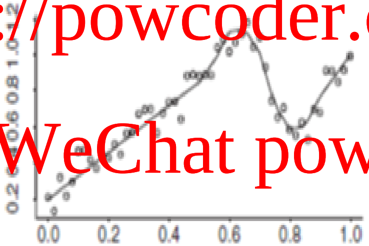
$$y_i = \beta_0 + \beta_1 x_i + \beta_{11}(x_i - 0.6)_+ + \epsilon_i$$

- which can be fit using the least square estimator with

$$\begin{bmatrix} 1 & x_1 & (x_1 - 0.6)_+ \\ \vdots & \vdots & \vdots \\ 1 & x_n & (x_n - 0.6)_+ \end{bmatrix}$$

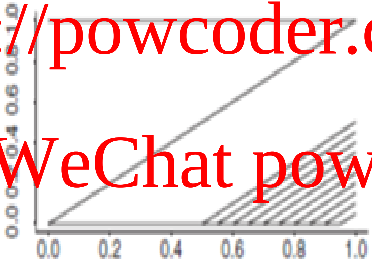
Spline basis function

- Assume a more complicated structure



Spline basis function

- If we have good reason to believe that our underlying structure is of this basic, we could change the basis ?



- where the functions: $(x - 0.5)_+$, $(x - 0.55)_+$,...,$(x - 0.95)_+$

Spline basis function

- The basis can do a reasonable job with a linear portion between $x = 0$ and $x = 0.5$

- We can use least square to fit such model with

$$\begin{bmatrix} 1 & x_1 & (x_1 - 0.5)_+ & (x_1 - 0.55)_+ & \dots & (x_1 - 0.95)_+ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n - 0.5)_+ & (x_n - 0.55)_+ & \dots & (x_n - 0.95)_+ \end{bmatrix}$$

Spline basis function

- It is possible to handle any complex type of structure by simply adding functions of the form $(x - k)_+$ to the basis
- This is equivalent to adding a column of values to the X matrix
- The value k is usually referred to as knots
- The function is made up of two lines that are tied together at $x = k$

Spline basis function

- The function $(x - 0.6)_+$ is called a linear spline basis function
- A set of such functions is called a linear spline basis
- Any linear combination of linear spline basis functions 1, x, $(x_i - k_1)_+$,...,$(x_i - k_K)_+$ is a piecewise linear function with knots $k_1, k_2, \ldots, k_K$ and called spline

Spline basis function

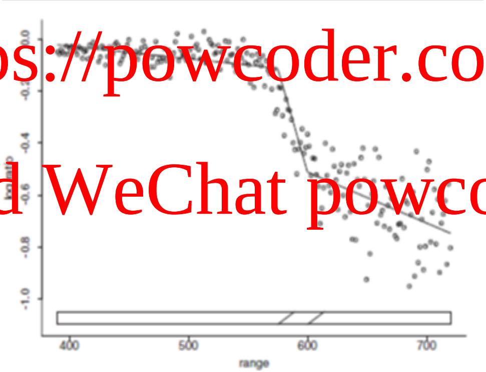
- Rather than referring to the spline basis function $(x-k)_+$ it is common to simply refer to it knots k

- We say the model has a knot at 0.35 if the function $(x-0.35)_+$ is the basis
- The spline model for a function f is

$$f(x) = \beta_0 + \beta_1 x + \sum_{i=1}^{K} \beta_{ki} (x - k_i)_+$$

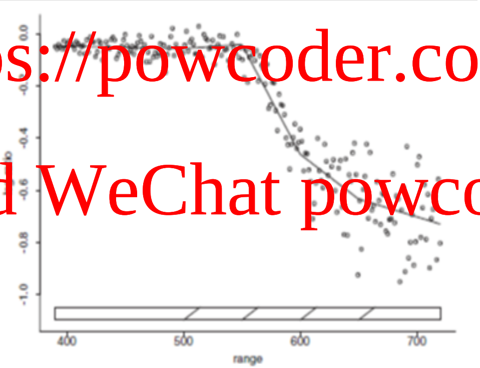
Illustration

- The selection of a good basis is usually challenging
- Start by trying to choose knots by trial



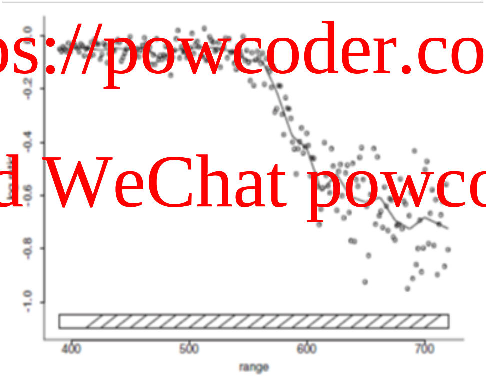
Illustration

- The fit lacking in quality for low values of range
- An obvious remedy is to use more knots



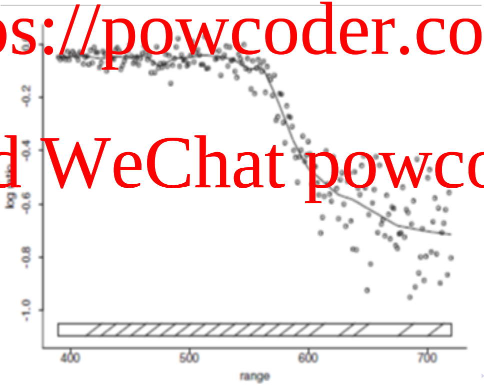
Illustration

- Larger set of knots, the fitting procedure has much more flexibility
- The plots is heavily overfitted



Illustration

- Pruning the knots to overcome the overfitting issue
- This fits the data well without overfitting
- This was arrived at, after a lot of time consuming trial and error



Knot selection

- A natural attempt at automatic selection of the knots is to use a model selection criterion

- If there are K candidate knots then there are 2^K possible models assuming the overall intercept and linear term are always present

- Highly computational intensive

Penalized spline regression

- Too many knots in the model induces roughness of the fit
- An alternative approach: retain all the knots but constrain their influence
- Hope: this will result in a less variable fit
- Consider a general spline model with K knots, K large

Penalized spline regression

- The ordinary least square fit is written as

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} \text{ where } \hat{\boldsymbol{\beta}} \text{ minimizes } \|\mathbf{y} - X\boldsymbol{\beta}\|^2$$

- and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_{11}, ..., \beta_{1K}]$ with β_{1k} the coefficient of the k^{th} knot.

- Unconstrained estimation of the $\boldsymbol{\beta}$ leads to a wiggly fit

Penalized spline regression

Constraints on the β_{1k} that might help avoid this situation are

- $\max|\beta_{1k}| < C$
- $\sum|\beta_{1k}| < C$
- $\sum\beta_{1k}^2 < C$

Penalized spline regression

Define the matrix $\mathbf{D}$ if size $(K+2) \times (K+2)$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & \mathbf{I}_{K\times K} \end{bmatrix}$$

Penalized spline regression

- The third constraint is easier to implement than the first two
- The minimization problem

$$\text{Minimize } \|\mathbf{y} - \mathbf{X}\beta\|^2 \text{ subject to } \beta^\top \mathbf{D}\beta \le C$$

- This is equivalent to choosing β to minimize

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\beta^\top \mathbf{D}\beta$$

- with solution

$$\hat{\boldsymbol{\beta}}_\lambda = \left(X^\top X + \lambda \mathbf{D}\right)^{-1} X^\top \mathbf{y}$$

Penalized spline regression

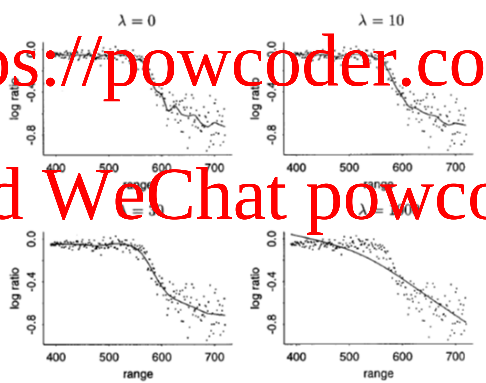
- The term $\lambda\boldsymbol{\beta}^\top \mathbf{D}\boldsymbol{\beta}$ is called a roughness penalty since it penalizes fits that are too rough, thus yielding smoother result

- The amount of smoothness is controlled by λ, which is therefore referred to as a smoothing parameter
- The fitted values for penalized spline regression are

$$\hat{\mathbf{y}} = X\left(X^\top X + \lambda^2\mathbf{D}\right)^{-1} X^\top \mathbf{y}$$

Illustration

- Linear penalized spline regression fits for different values of the smoothing parameter



Quadratic spline bases

- We have discussed linear splines, that is continuous, piecewise linear functions
- The reason for the piecewise linear nature of the functions ?

Quadratic spline bases

- We have discussed linear splines, that is continuous, piecewise linear functions

- The reason for the piecewise linear nature of the functions ?
- is that they are a linear combination of piecewise linear functions of the form $(x-k)_+$

- A simple way of escaping from piecewise linearity ?

Quadratic spline bases

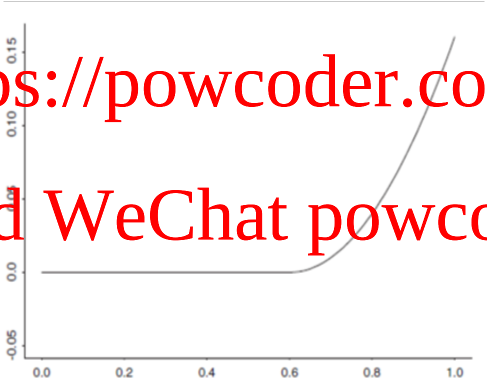
- We have discussed linear splines, that is continuous, piecewise linear functions
- The reason for the piecewise linear nature of the functions ?

- is that they are a linear combination of piecewise linear functions of the form $(x-k)_+$
- A simple way of escaping from piecewise linearity ?

- is to add x^2 to the basis and also to replace each $(x-k)_+$ by its square $(x-k)_+^2$

Illustration of a quadratic spline basis function

- Illustration of the function $(x - 0.6)^2_+$



Quadratic spline bases

- The function doesn't have a sharp corner like $(x-0.6)_+$ does
- The function $(x-0.6)_+^2$ has a continuous first derivative
- Any linear combination of the functions

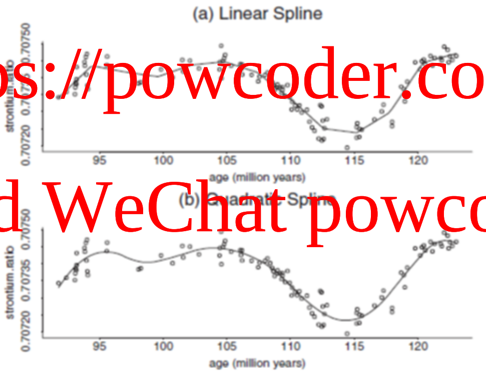
$$1, x, x^2, (x-k_1)_+^2, \dots, (x-k_K)_+^2$$

- also have a continuous first derivative and not have any sharp corner

- This result in better fit
- This is called a quadratic spline basis with knots at $k_1, \dots, k_K$

Illustration of quadratic spline basis functions

- Quadratic spline do a better job of fitting peaks and valleys



Other spline bases

- We discussed linear and quadratic spline models
- One reason for considering other models is to achieve smoother fits $\rightarrow$ important if one plans to differentiate the fit to estimate derivative of the regression function

- In principle a change of basis does not change the fit but some bases are more stable and allow computation of a fit with better accuracy

- Besides numerical stability: ease of implementation is another reason for selecting one basis over another
- An obvious generalization is given by

$$1, x, \ldots, x^p, (x - k_1)_+^p, \ldots, (x - k_K)_+^p$$

- know as the truncated power basis of degree p

Other spline bases

- since the function $(x-k)_+^p$ has $p-1$ continuous derivatives, higher values of p lead to smoother spline functions
- The p^{th} degree spline model is

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{i=1}^{K} \beta_{ki} (x - k_i)_+^p$$

- The expression for the fitted values is given by

$$\hat{\mathbf{y}} = X \left(X^\top X + \lambda \mathbf{D} \right)^{-1} X^\top \mathbf{y}$$

$$\mathbf{D} = \mathrm{diag}\left(\mathbf{0}_{p+1}, \mathbf{1}_K\right)$$