## Flexible Regression

Session 1 - Introduction and Nonparametric Regression (Chs $1 \& 2$ )

Electronic notes available at: https://warwick.ac.uk/fac/sci/statistics/apts/ students/resources/

## What this course is about?

Fitting smooth relationships to describe drivers of water quality.


## What this course is about?

Fitting smooth surfaces to explain spatiotemporal variation in river nutrient levels.


$$
\mathbb{k}
$$

## What this course is about?

Fitting relationships to appropriate quantiles to explain BMI using age.


## What this course is about?

Extending the previous approaches to summarise temperature across multiple seasonal patterns.


## What this course is about?

Extending the previous approaches to summarise temperature across multiple seasonal functions.


## What this course is about? - Flexible Regression

- flexibility in the mean:

$$
Y_{i}=f\left(x_{i}, \boldsymbol{\beta}\right)+\varepsilon_{i} .
$$

minimise

$$
\sum_{i=1}^{n}\left(y_{i}-f\left(\mathbf{x}_{i}\right)\right)^{2} .
$$

- flexibility in the response quantile:
e.g. median regression minimise

$$
\sum_{i=1}^{n}\left|y_{i}-f\left(\mathbf{x}_{i}\right)\right|
$$

## What this course is about? - Flexible Regression

- flexibility in the mean:
- Nonparametric regression - Chapter 2 (Session 1);
- (Generalised) Additive Models (GAMs) - Chapter 4 (Session 3)
- flexibility in the response quantile:
- Quantile regression - Chapter 3 (Session 2);
- (Generalised) additive quantile regression - Chapter 5 (Session $2 \& 4)$


## What this course is about? - Flexible Regression

To help illustrate the ideas we'll also have 2 labs:

- Lab 1 (Tues) - nonparametric regression/quantile regression with splines
- Lab 2 (Thurs) - (Generalised) Additive (quantile) Models (GAMs)


## Chapter 2 - Session motivation

## Example 2.1 Great Barrier Reef data

Zone an indicator for the closed (1) and open (0) zones
Year an indicator of 1992 (0) or 1993 (1)
Latitude latitude of the sampling position
Longitude longitude of the sampling position
Depth bottom depth
Score1 catch score 1
Score2 catch score 2

## Chapter 2 - Session motivation

Figure 2.1: Great Barrier Reef data


## What we will be discussing in this session?

Nonparametric regression

- Approaches for nonparametric regression
- Properties of smooth functions
- Why use splines?
- How to construct splines in 1D?
- Penalty-based approaches


### 2.1 Nonparametric regression

A simple nonparametric regression model has the form

$$
Y_{i}=f\left(x_{i}\right)+\varepsilon_{i}, \quad i=1, \ldots, n
$$

where the data $\left(x_{i}, y_{i}\right)$ are described by a smooth curve $f$ plus independent errors $\varepsilon_{i}$.

Smoothing is used to estimate $f()$.

### 2.1 Nonparametric regression

Smoothers have two main uses:

Description - to aid 'visually' in the exploration of a relationship or pattern.

Estimation - to estimate the dependence of the mean of $Y$ on the predictor $x$.

### 2.1 Nonparametric regression

The two key questions that arise regarding the definition of a smoother are:

- Which smoothing method should be used?
- What level of smoothing is appropriate?


### 2.1 Nonparametric regression

Which smoothing method should be used?

- local fitting approaches;
- spline based methods.


### 2.2 A local fitting approach

For example, local linear regression. Solve the least squares problem:

$$
\min _{\alpha, \beta} \sum_{i=1}^{n}\left\{y_{i}-\alpha-\beta\left(x_{i}-x\right)\right\}^{2} w\left(x_{i}-x ; h\right)
$$

and take as the estimate at $x$ the value of $\hat{\alpha}$, as this defines the position of the local regression line at the point $x$. The weight function, $w\left(x_{i}-x ; h\right)$, is a kernel function (see preliminary material).



### 2.2 A local fitting approach

$$
\min _{\alpha, \beta} \sum_{i=1}^{n}\left\{y_{i}-\alpha-\beta\left(x_{i}-x\right)\right\}^{2} w\left(x_{i}-x ; h\right)
$$




### 2.2.2 A local fitting approach - properties

- Sometimes computational or practical reasons can constrain our choice of smoother.
- Expressions for bias and variance (derived in section 2.2.1) can help us to choose between smoothing approaches.
- There is a trade-off between following the data closely (low bias, possibly large variance) and obtaining a smooth function (low variance, possibly large bias).


### 2.2.3 Local linear regression in R

A local linear regression fit for the Reef data can be obtained using the R library sm.


### 2.3 Regression splines

Example 2.3 - Which function fits the data better? - bias versus variance (Figure 2.4)



### 2.3 Regression splines

Example 2.3 - Which function fits the data better? - bias versus variance (Figure 2.4)


Both have the same fitted values $\hat{y}_{i}=\hat{f}\left(x_{i}\right)$ !

### 2.3 Regression splines

Example 2.3 - Which can we learn from this example?

- Difficult to do smoothing without knowing / understanding the context.
- Family of smooth functions too rich to be able to only rely on the data.
- Alternatives to local fitting approaches
- Splines based on truncated power series and B-splines
- Penalties or Bayesian approaches to penalise wiggliness


### 2.3.1 Regression splines - polynomial regression

## Linear regression

$$
\mathbb{E}\left(Y_{i}\right)=\beta_{0}+\beta_{1} x_{i} \quad \text { for } i=1, \ldots, n
$$

In matrix-vector notation:

$$
\mathbb{E}(\mathbf{y})=\mathbf{B} \boldsymbol{\beta} \quad \text { with } \mathbf{y}=\left(Y_{1}, \ldots, Y_{n}\right)^{\top} \text { and } \mathbf{B}=\left(\begin{array}{cc}
1 & x_{1} \\
\vdots & \vdots \\
1 & x_{n}
\end{array}\right) .
$$

B - matrix of basis functions
$\boldsymbol{\beta}$ - vector of basis coefficients
Basis functions: $B_{0}(x)=1, B_{1}(x)=x$.

### 2.3.1 Regression splines - polynomial regression



Figure: 2.5 A simple linear regression line with underlying simulated data


Figure: 2.6 The basis functions for simple linear regression $1, x$

### 2.3.1 Regression splines - polynomial regression

## Polynomial regression

$$
\mathbb{E}\left(Y_{i}\right)=\beta_{0}+\beta_{1} x_{i}+\ldots+\beta_{r} x_{i}^{r} \quad \text { for } i=1, \ldots, n
$$

just corresponds to

$$
\mathbf{B}=\left(\begin{array}{cccc}
1 & x_{1} & \ldots & x_{1}^{r} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n} & \ldots & x_{n}^{r}
\end{array}\right)
$$

Polynomial regression is a basis expansion technique.

$$
\hat{\boldsymbol{\beta}}=\left(\mathbf{B}^{\top} \mathbf{B}\right)^{-1} \mathbf{B}^{\top} \mathbf{y}
$$

### 2.3.1 Regression splines - polynomial regression

Example 2.4: Glucose levels in potatoes


Polynomial regression can be a useful tool for small datasets. (if a small degree of polynomial is used)

### 2.3.1 Regression splines - polynomial regression

Simulated example 2.5: $y_{i}=1-x_{i}^{3}-2 \exp \left(-100 x_{i}^{2}\right)+\varepsilon_{i}$ with $\mathbf{x}=(-1,-0.98, \ldots, 0.98,1)$ and $\varepsilon_{i} \sim \mathrm{~N}\left(0,0.1^{2}\right)$.


Polynomial regression (degree $r=10$ )


Polynomial regression (degree $r=17$ )
2.3.1 Regression splines - polynomial regression - what is going wrong?

Let's look at the hat matrix (Figure 2.9): $\mathbf{S}=\mathbf{B}\left(\mathbf{B}^{\top} \mathbf{B}\right)^{-1} \mathbf{B}^{\top}$ ( $\hat{\mathbf{y}}=\mathbf{S y}$ )


Polynomial regression (degree $r=10$ ) Polynomial regression (degree $r=17$ )

### 2.3.1 Regression splines: spline-based model



2.3.1 Regression splines - polynomial regression - problems

- Polynomials are not a local model.
$\rightsquigarrow$ Oscillations (Runge's phenomenon), "derivative propagation"
- Likely to produce spurious edge effects on both ends of range. (polynomial has to diverge to $\pm \infty$ as $x \rightarrow \pm \infty$ )
- Also very likely to be numerically unstable. In our example:
- Condition number of $\mathbf{B}^{\top} \mathbf{B}: 1.56 \times 10^{12}$
- For a B-spline: Condition number of $\mathbf{B}^{\boldsymbol{\top}} \mathbf{B}: 32.49$
- Possible solution: Use piecewise polynomial functions.


### 2.3.1 Regression splines - piecewise polynomials



Maybe we need to "glue" the polynomials together a bit $\rightsquigarrow$ splines
(Figure 2.10).

## Break

The following $R$ packages will be used in the lab tomorrow if you want to download them early:

- splines
- ggplot2
- quantreg
- rpanel
- mgcv


### 2.3.2 Regression splines

## Definition 2.1: (Polynomial) spline

Given a set of knots $a=\kappa_{1}<\kappa_{2}<\ldots<\kappa_{l}=b$, a function $f:[a, b] \rightarrow \mathbb{R}$ is called a (polynomial) spline of degree $r$ if

- $f(\cdot)$ is a polynomial of degree $r$ on each interval $\left(\kappa_{j}, \kappa_{j+1}\right)$

$$
(j=1, \ldots, l-1)
$$

- $f(\cdot)$ is $r-1$ times continuously differentiable.


### 2.3.2 Regression splines: degrees $r=0$ and $r=1$

Radiocarbon dating (Figure 2.12)



### 2.3.2 Regression splines: degrees $r=2$ and $r=3$

Radiocarbon dating



### 2.3.2 Regression splines: different numbers of knots (1)

Radiocarbon dating (Figure 2.13)

$I=3$ knots

$I=9$ knots
2.3.2 Regression splines: different numbers of knots (2)

Radiocarbon dating

$I=15$ knots

$I=31$ knots

### 2.3.2 Regression splines

## Choice of degree and number of knots

## Choice of degree $r$

- Degree $r$ controls smoothness / differentiability
- The larger the degree $r$ the more the spline behaves like a polynomial.
- Rarely necessary to go beyond $r=3$.


## Choice of number of knots /

- Number of knots / controls smoothness / flexibility
- Alternative: Use "too many" knots and control flexibility using roughness penalty (see later)


### 2.3.2 Regression splines: natural cubic splines

## Definition 2.2: Natural cubic spline

A polynomial spline $f:[a, b] \rightarrow \mathbb{R}$ of degree 3 is called a natural cubic spline if $f^{\prime \prime}(a)=f^{\prime \prime}(b)=0$.

- Key idea: Natural cubic splines extrapolate linearly.


### 2.3.2 Regression splines: natural cubic splines

## Proposition 2.3

A set of I points $\left(x_{i}, y_{i}\right)$ can be exactly interpolated using a natural cubic spline with the $x_{1}<\ldots<x_{l}$ as knots. The interpolating natural cubic spline is unique.

Natural cubic splines can be generated using the function ns in the package splines in $R$.

### 2.3.3 Regression splines - why are splines optimal?

- What does optimal mean?
- Optimal $=$ as close to the data as possible ? $\rightsquigarrow$ Objective function: $\sum_{i=1}^{n}\left(y_{i}-f\left(x_{i}\right)\right)^{2}$
- Not a good idea: Any function interpolating the data would be optimal!
- We need to balance out two aspects (remember bias and variance).
- We want $f(\cdot)$ to follow the data closely.
- We want the function $f(\cdot)$ not to be too complicated so that it generalises well to future unseen data.


### 2.3.3 Regression splines - why are splines optimal?

## Penalised / regularised fitting criterion

$$
\underbrace{\sum_{i=1}^{n}\left(y_{i}-f\left(x_{i}\right)\right)^{2}}_{\text {Fit to the data }}+\lambda \underbrace{\int_{a}^{b} f^{\prime \prime}(x)^{2} d x}_{\text {Roughness penalty }}
$$

- $\lambda \geq 0$ controls roughness.
- Natural cubic splines are minimisers of the above criterion.

Finding optimal natural cubic splines is a simple penalised-least-squares problem (more on that later)
$\rightsquigarrow$ Leads to a technique called smoothing splines (smooth.spline in R)

### 2.3.4 Regression splines - how to fit?

Minimise

$$
\sum_{i=1}^{n}\left(y_{i}-f\left(\mathbf{x}_{i}\right)\right)^{2}
$$

Represent $f\left(\mathbf{x}_{i}\right)$ as $\mathbf{B} \boldsymbol{\beta}$.
How to construct a basis?

B is formulated through:

- Truncated power basis;
- B-splines.


### 2.3.4 Regression splines - truncated power series

## Definition 2.6: Truncated power basis

Given a set of knots $a=\kappa_{1}<\ldots<\kappa_{l-1}=b$ the truncated power basis of degree $r-1$ is given by

$$
\left(1, x, \ldots, x^{r-1},\left(x-\kappa_{1}\right)_{+}^{r-1},\left(x-\kappa_{2}\right)_{+}^{r-1}, \ldots,\left(x-\kappa_{l-1}\right)_{+}^{r-1}\right),
$$

where $(z)_{+}^{r}= \begin{cases}z^{r} & \text { for } z>0 \\ 0 & \text { otherwise } .\end{cases}$

### 2.3.4 Regression splines - truncated power series

Truncated power series of degree 3 (Figure 2.15)


### 2.3.4 Regression splines - truncated power series

How to fit a model using the truncated power basis?

- Use basis expansion

$$
\begin{aligned}
f(x)=\beta_{0} & +\beta_{1} x+\ldots+\beta_{r-1} x^{r-1} \\
& +\beta_{r}\left(x-\kappa_{1}\right)_{+}^{r-1}+\ldots+\beta_{r+l-2}\left(x-\kappa_{l-1}\right)_{+}^{r-1}
\end{aligned}
$$

- This is just a linear model with design matrix

$$
\mathbf{B}=\left(\begin{array}{lllllll}
1 & x_{1} & \ldots & x_{1}^{r-1} & \left(x_{1}-\kappa_{1}\right)_{+}^{r-1} & \ldots & \left(x_{1}-\kappa_{l-1}\right)_{+}^{r-1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n} & \ldots & x_{n}^{r-1} & \left(x_{n}-\kappa_{1}\right)_{+}^{r-1} & \ldots & \left(x_{n}-\kappa_{l-1}\right)_{+}^{r-1}
\end{array}\right)
$$

### 2.3.4 Regression splines - truncated power series

Illustration: Radiocarbon dating (Figure 2.19)


### 2.3.4 Regression splines - truncated power series

## Numerical problems (Figures 2.17, 2.18)




Problem: Basis functions highly correlated (0.99921) $\rightsquigarrow \mathbf{B}^{\top} \mathbf{B}$ ill conditioned (condition number: $5.85 \times 10^{9}$ )

### 2.3.4 Regression splines - B-splines

## Definition 2.7: B-spline basis

(a) Given a set of $I$ knots the $B$-spline basis of degree 0 is given by the functions $\left(B_{1}^{0}(x), \ldots, B_{I-1}^{0}(x)\right)$ with

$$
B_{j}^{0}(x)= \begin{cases}1 & \text { for } \kappa_{j} \leq x<\kappa_{j+1} \\ 0 & \text { otherwise }\end{cases}
$$

(b) Given a set of $I$ knots the B-spline basis of degree $r>0$ is given by the functions $\left(B_{1}^{r}(x), \ldots, B_{I+r-1}^{r}(x)\right)$ with

$$
B_{j}^{r}(x)=\frac{x-\kappa_{j-r}}{\kappa_{j}-\kappa_{j-r}} B_{j-1}^{r-1}(x)+\frac{\kappa_{j+1}-x}{\kappa_{j+1}-\kappa_{j+1-r}} B_{j}^{r-1}(x)
$$

2.3.4 Regression splines: B-spline basis of degree $r=1$

Figure 2.20:

2.3.4 Regression splines: B-spline basis of degree $r=2$

2.3.4 Regression splines: B-spline basis of degree $r=3$


### 2.3.4 Regression splines: B-splines

Model fitting using B-splines

- Use basis expansion

$$
f(x)=\sum_{j=1}^{I+r-1} \beta_{j} B_{j}(x)
$$

- This is just a linear model with design matrix

$$
\mathbf{B}=\left(\begin{array}{lll}
B_{1}^{r}\left(x_{1}\right) & \ldots & B_{I+r-1}^{r}\left(x_{1}\right) \\
\vdots & \ddots & \vdots \\
B_{1}^{r}\left(x_{n}\right) & \ldots & B_{I+r-1}^{r}\left(x_{n}\right)
\end{array}\right)
$$

### 2.3.4 Regression splines: B-splines

Illustration: Radiocarbon dating (Figure 2.24)


### 2.3.4 Regression splines: B-splines

No more numerical problems (Figures 2.22, 2.23)



Problem solved: Basis functions not highly correlated (0.8309 at most)
$\rightsquigarrow \mathbf{B}^{\top} \mathbf{B}$ not ill-conditioned (condition number: 358.263)

### 2.3.4 Regression splines: B-splines in $R($ Figure 2.21)

```
library(splines)
model <- lm(Rc.age~bs(Cal.age, df=10), data=radiocarbon)
with(radiocarbon, {
    plot(Cal.age, Rc.age)
    lines(Cal.age, predict(model))
})
```



### 2.3.5 Penalised regression splines (P-splines) - Idea

- Positioning of knots can have large influence of fitted function (especially if number of knots is small);
- Solution: Use "too many" knots and control flexibility using roughness penalty;
- We will only consider quadratic roughness penalties of the form $\|\mathbf{D} \boldsymbol{\beta}\|^{2}$.


### 2.3.5 Penalised regression splines (P-splines) - Idea

- Objective function ( ${ }^{* *}$ see ridge regression) :

$$
\sum_{i=1}^{n}\left(y_{i}-f\left(x_{i}\right)\right)^{2}+\lambda\|\mathbf{D} \boldsymbol{\beta}\|^{2}
$$

- $\lambda$ controls the trade-off between following the data $(\lambda \downarrow)$ and a strongly regularised curve $(\lambda \uparrow)$.


### 2.3.5 Penalised regression splines (P-splines) - Solution

- Solution for P-splines is

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

- Numerically more stable to use a QR decomposition to minimise augmented system

$$
\left\|\binom{\mathbf{y}}{\mathbf{0}}-\binom{\mathbf{B}}{\sqrt{\lambda} \mathbf{D}} \boldsymbol{\beta}\right\|^{2}
$$

### 2.3.6 Penalised regression splines - how to choose D?

## Smoothing splines

One can show that

$$
\int_{a}^{b} f^{\prime \prime}(x)^{2} d x=\boldsymbol{\beta}^{\top}\left(\begin{array}{ccc}
\int_{a}^{b} B_{1}^{\prime \prime}(x) B_{1}^{\prime \prime}(x) d x & \cdots & \int_{a}^{b} B_{1}^{\prime \prime}(x) B_{1+r-1}^{\prime \prime}(x) d x \\
\vdots & \ddots & \vdots \\
\int_{a}^{b} B_{1}^{\prime \prime}(x) B_{l+r-1}^{\prime \prime}(x) d x & \cdots & \int_{a}^{b} B_{l+r-1}^{\prime \prime}(x) B_{l+r-1}^{\prime \prime}(x) d x
\end{array}\right) \boldsymbol{\beta}
$$

$\rightsquigarrow$ Set $\mathbf{D}^{\top} \mathbf{D}$ equal to this matrix of cross-products.

### 2.3.6 Penalised regression splines - difference penalties

For equally-spaced knots we can also use difference penalties (much simpler).
2.3.6 Penalised regression splines - second-order difference penalty

$$
\mathbf{D}_{2}=\left(\begin{array}{ccccc}
1 & -2 & 1 & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & 0 \\
0 & \ldots & 1 & -2 & 1
\end{array}\right)
$$

$\left\|\mathbf{D}_{2} \boldsymbol{\beta}\right\|^{2}=\sum\left(\beta_{j+2}-2 \beta_{j+1}+\beta_{j}\right)^{2}$
(second-order differences)

### 2.3.6 Penalised regression splines - second-order difference penalty

- Shrinks the coefficients towards a linear sequence.
$\rightsquigarrow$ Shrinks the regression function $f(\cdot)$ towards linear function.
- Adding a linear function to $f(\cdot)$ does not change the penalty.
- Natural choise for spline basis of degree $r=3$.



### 2.3.7 Penalised regression splines in $R$

```
library(mgcv)
model <- gam(Rc.age~s(Cal.age), data=radiocarbon)
model
plot(model, residuals=TRUE)
```



## Summary

## Nonparametric regression

- Approaches for nonparametric regression
- Properties of smooth functions
- Why use splines?
- How to construct splines in 1D? (truncated power and B-splines)
- Penalty-based approaches (P-splines)

Coffee \& Tea now back in the Maths \& Stats building..... (start Design at 4pm).

