apts.ac.uk
Academy for PhD Training in Statistics

Flexible Regression

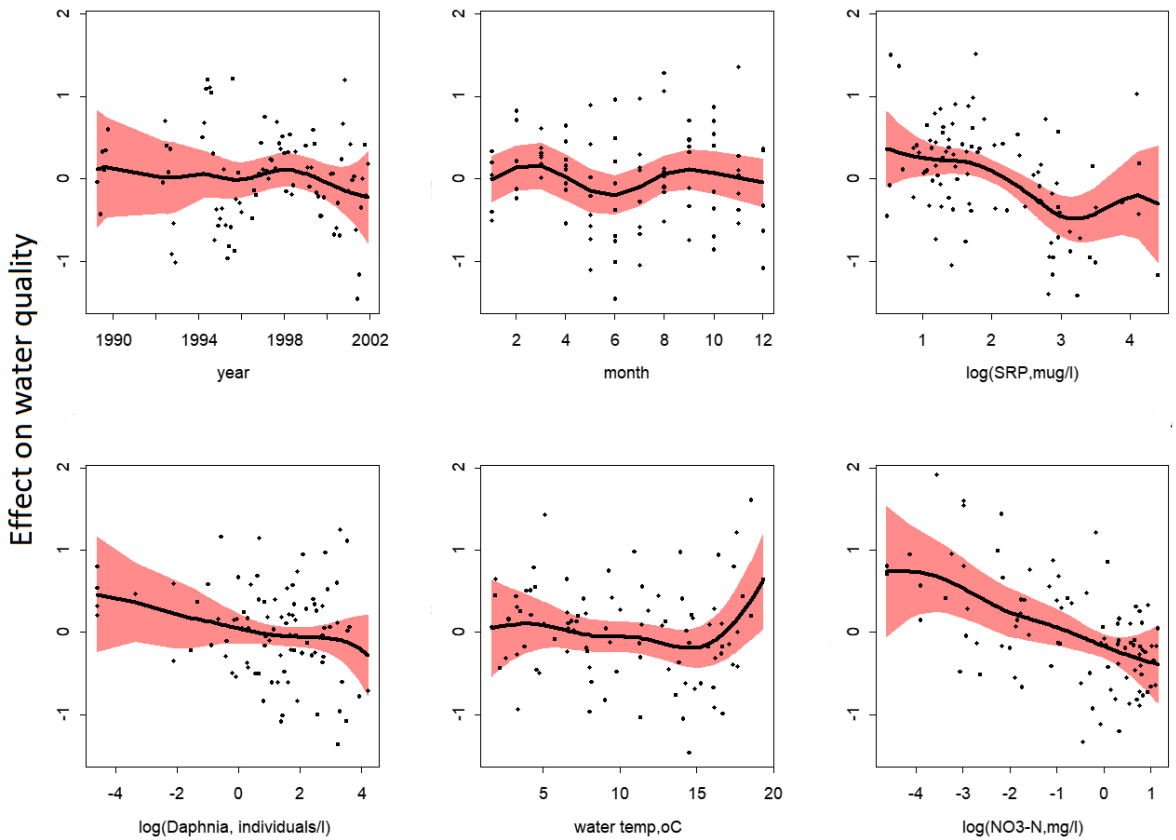
Session 1 - Introduction and Nonparametric Regression (Chs 1 & 2)

Please see the full notes for full explanation and details. The slides will help to signpost and guide you through the main points of the notes.

Claire Miller & Tereza Neocleous

What this course is about?

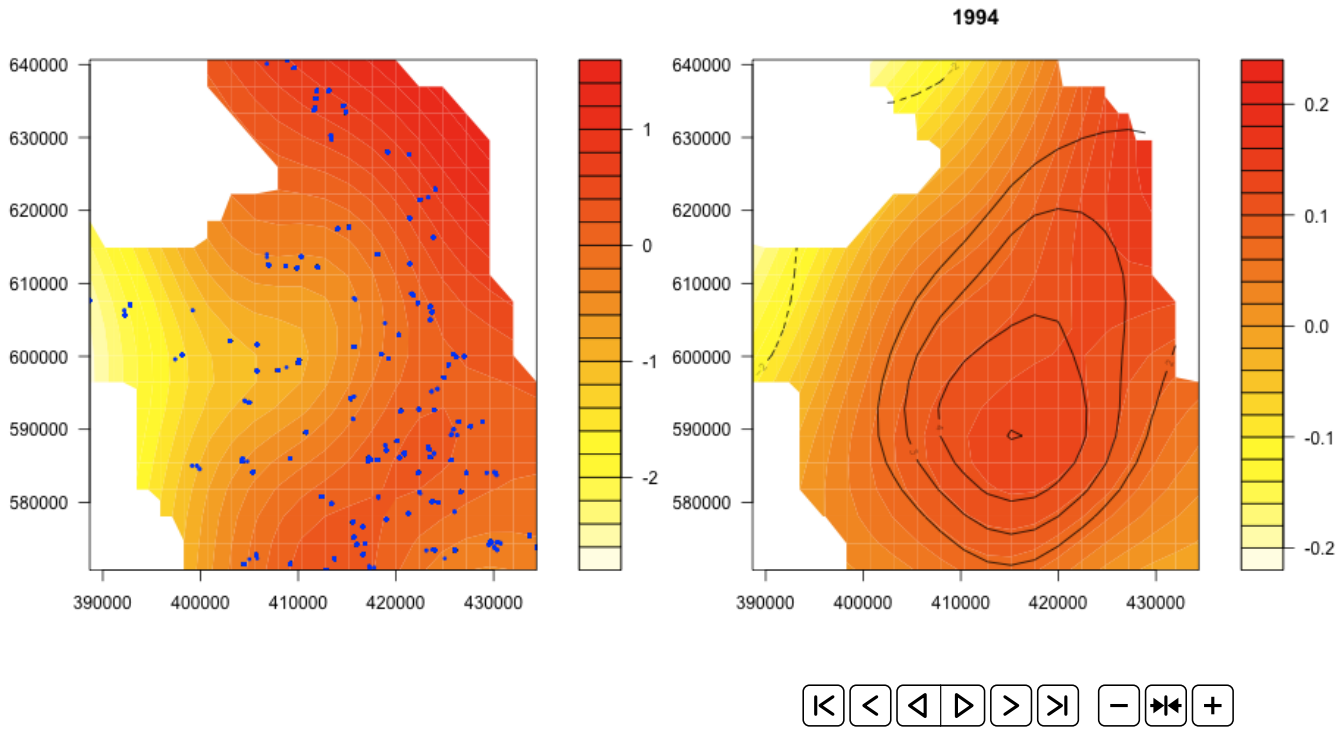
Fitting **smooth relationships** to describe temporal patterns and potential drivers (nutrients, water fleas and temperature) of water quality in a lake.



Output from an **additive model** with fitted values (black line), pink variability band and partial residuals.

What this course is about?

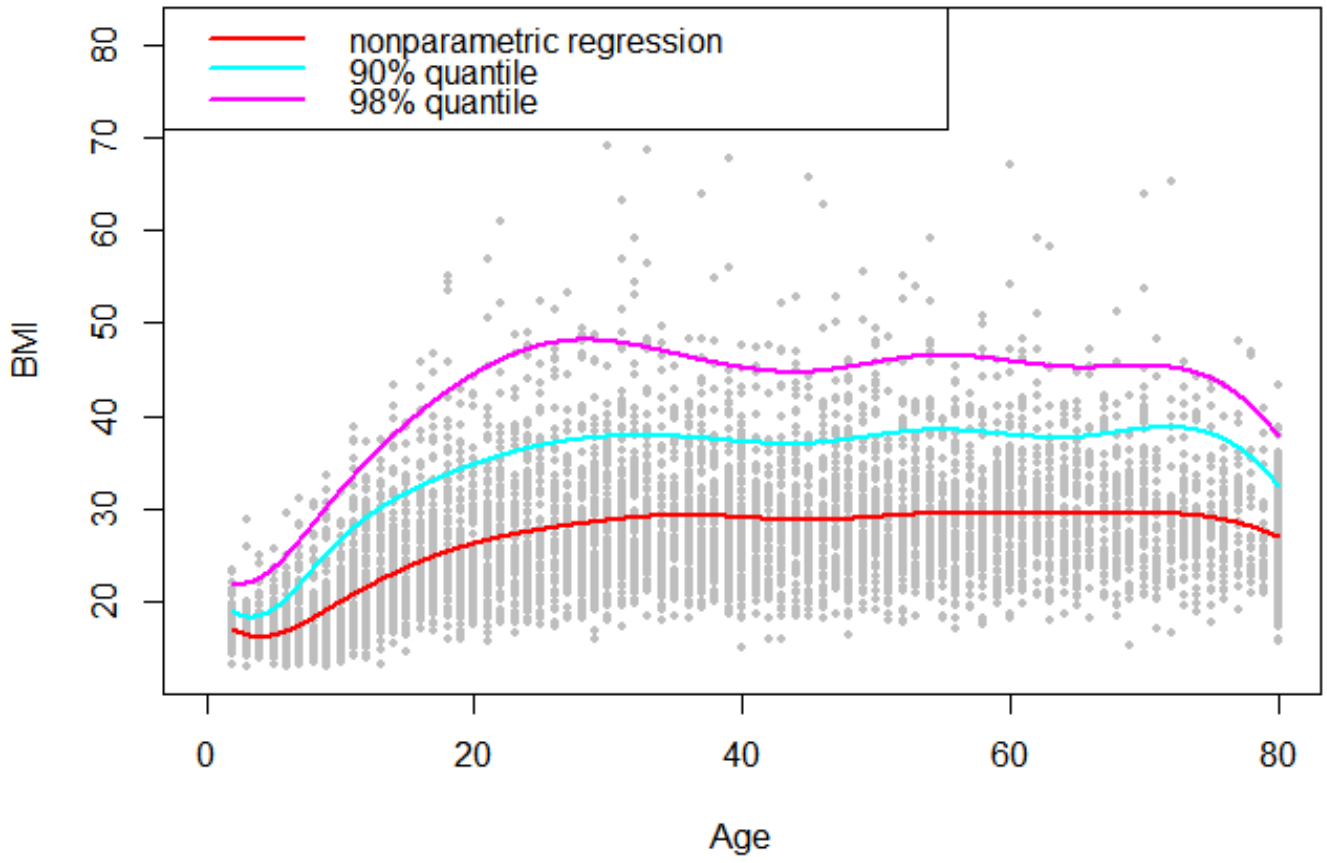
Fitting **smooth surfaces** to explain spatiotemporal variation in river nutrient levels.



The colour on the left surface provides the average levels over time, smoothed over space, with monitoring locations in blue. The video on the right shows how this average pattern changes as we move through time.

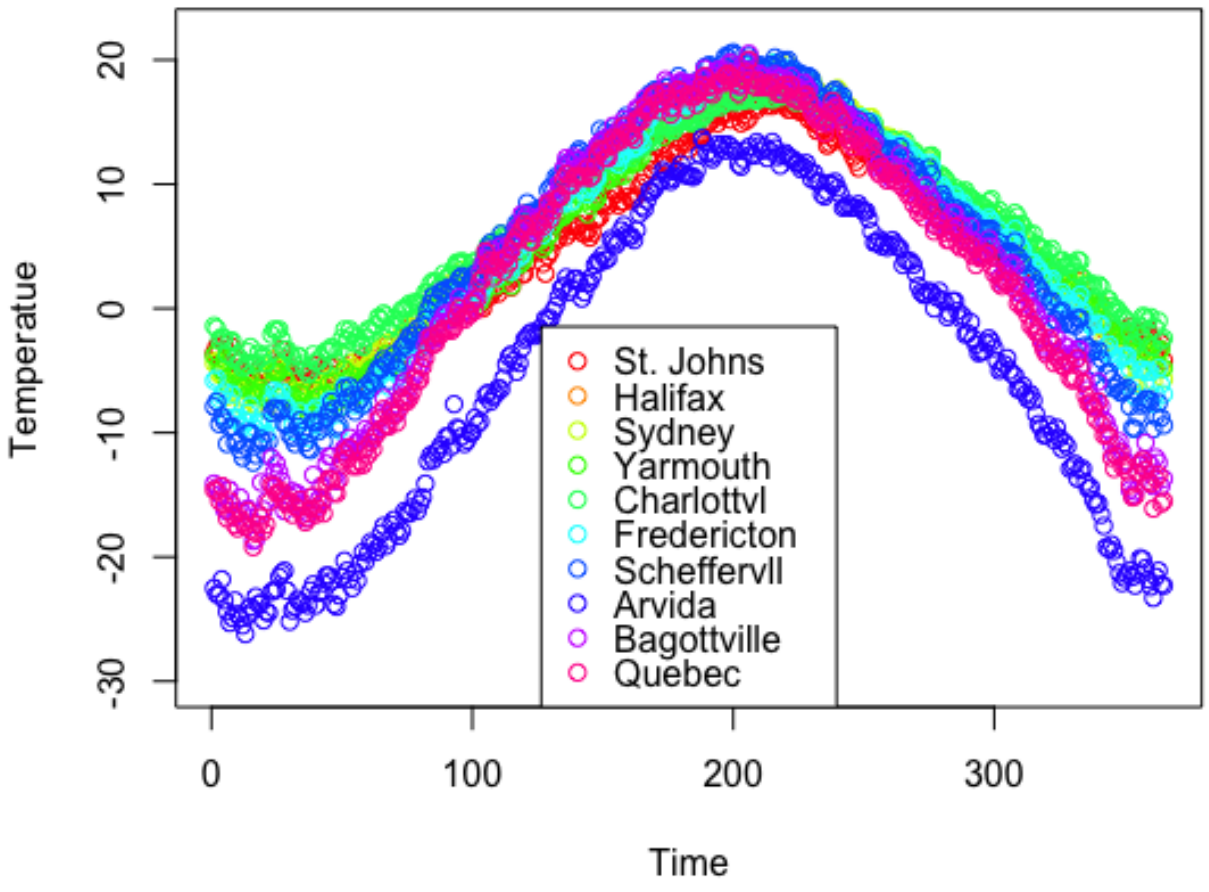
What this course is about?

Fitting **relationships to appropriate quantiles** to explain body mass index (BMI) using age.



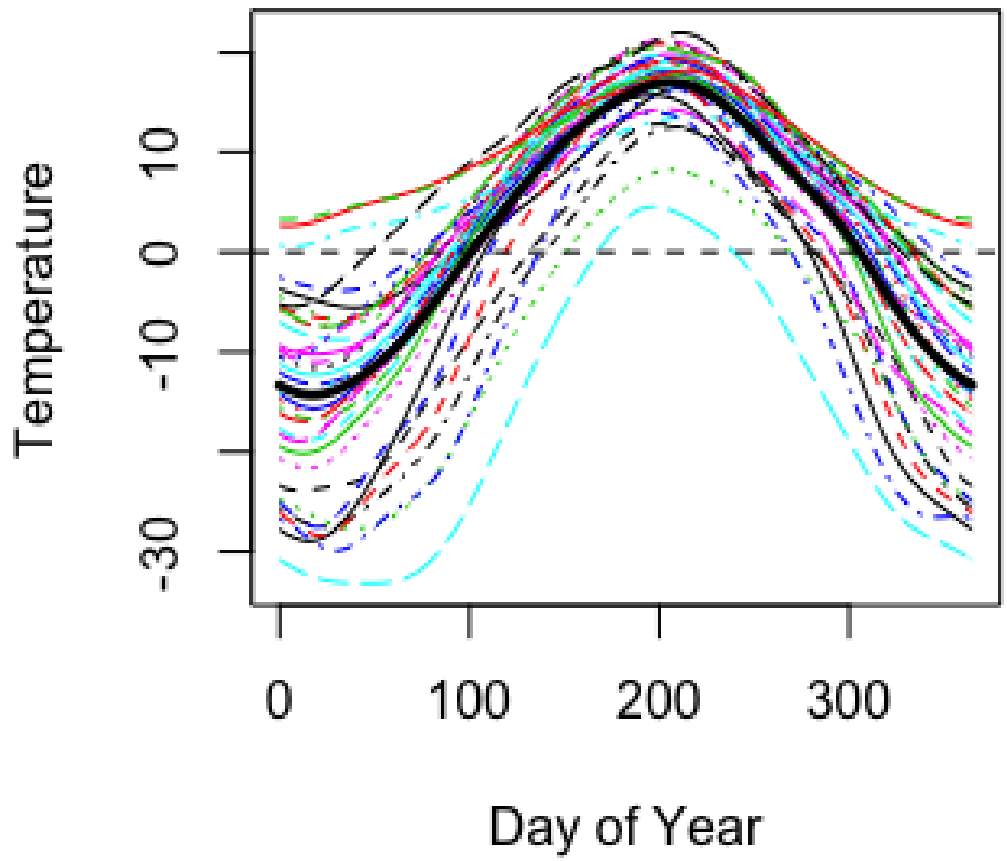
What this course is about?

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



What this course is about?

Extending the previous approaches to **summarise temperature seasonal functions for multiple Canadian weather stations.**



What this course is about? - Flexible Regression

- **flexibility in the mean:**

$$Y_i = f(x_i, \boldsymbol{\beta}) + \varepsilon_i.$$

minimise

$$\sum_{i=1}^{n} (y_i - f(\mathrm{x}_i))^2.$$

- **flexibility in the response quantile:**

e.g. median regression minimise

$$\sum_{i=1}^{n} |y_i - f(\mathrm{x}_i)|$$

.

What this course is about? - Flexible Regression

- **flexibility in the mean:**
  - Nonparametric regression - Chapter 2;
  - (Generalised) Additive Models (GAMs) - Chapter 4

- **flexibility in the response quantile:**
  - Quantile regression - Chapter 3;
  - (Generalised) additive quantile regression - Chapter 5

What this course is about? - Flexible Regression

To help illustrate the ideas there are also 2 practical lab sheets:

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

Note: The following `R` packages are required to work through the practical lab material:

- gamlss
- ggplot2
- mgcv
- quantreg
- rpanel
- splines

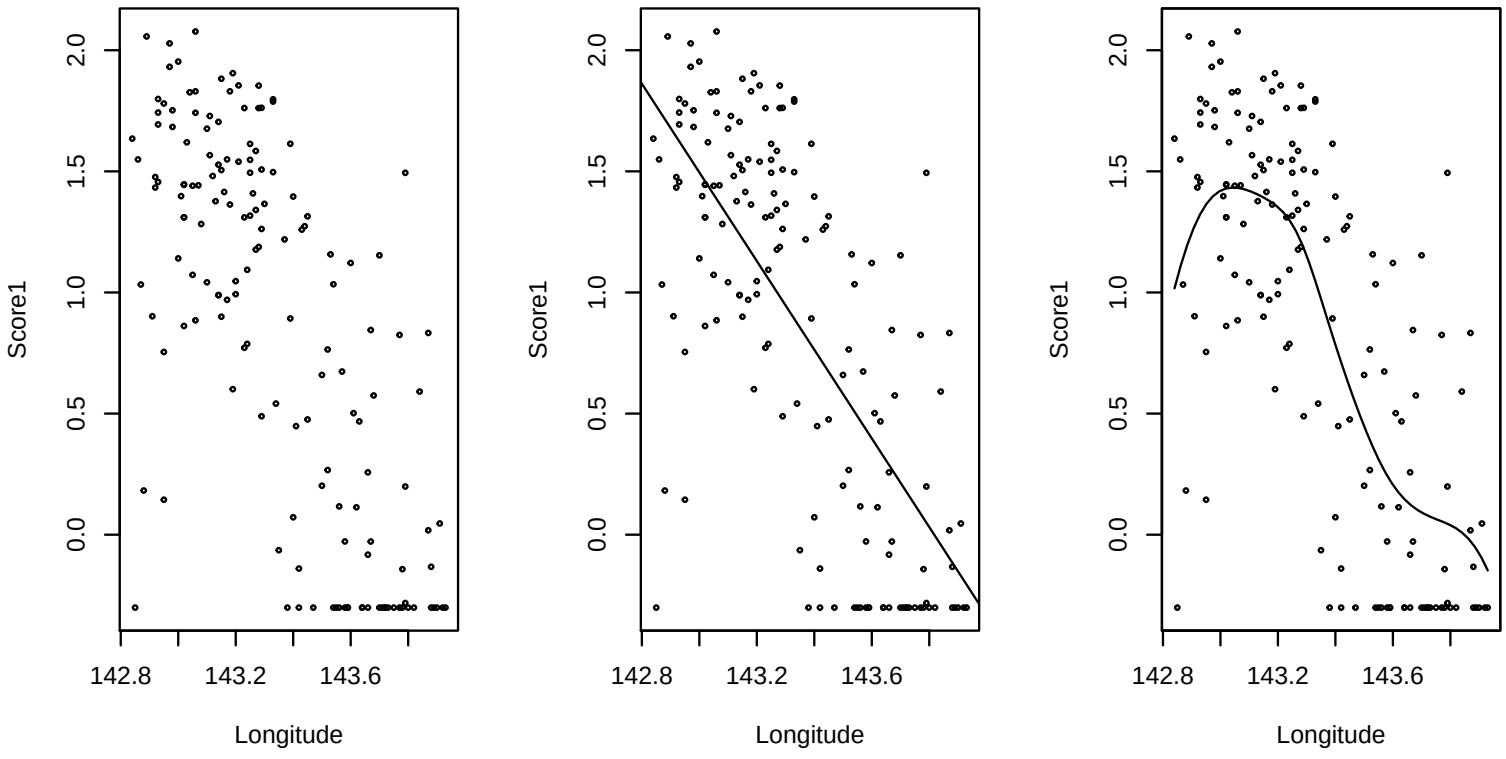
Chapter 2 - motivation

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 - motivation

Figure 2.1: Great Barrier Reef data



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

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

where the data (x_i, y_i) are described by a smooth curve f plus independent errors ε_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?

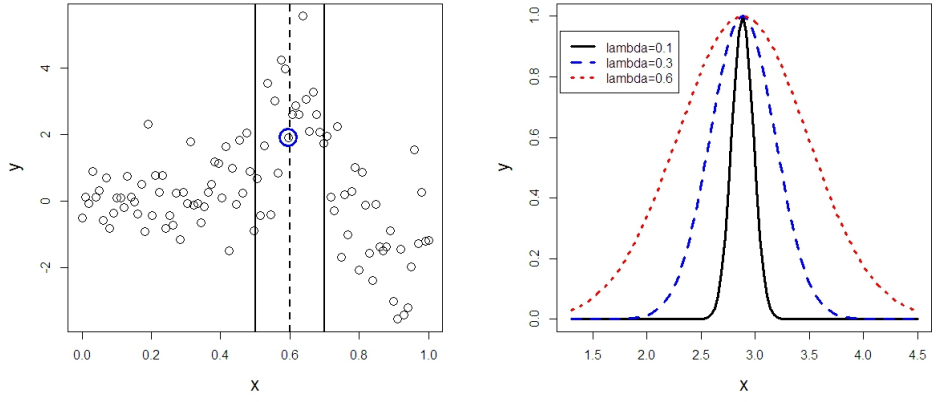
- local fitting approaches;
- spline based methods.

2.2 A local fitting approach

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

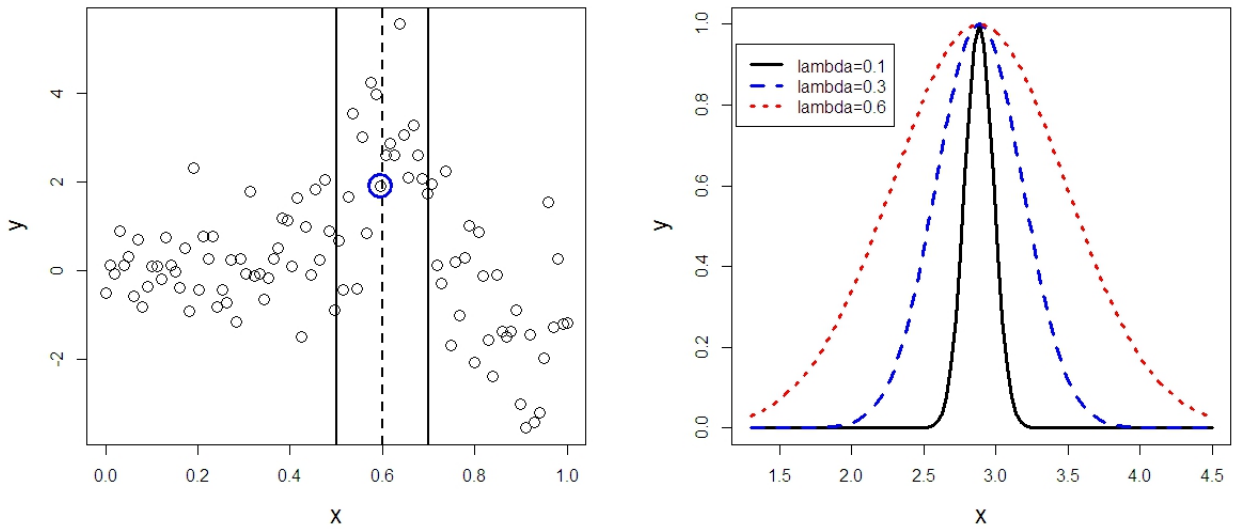
$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 \, w(x_i - x \, ; h)$$

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, $w(x_i - x \, ; h)$, is a kernel function (see preliminary material).



2.2 A local fitting approach

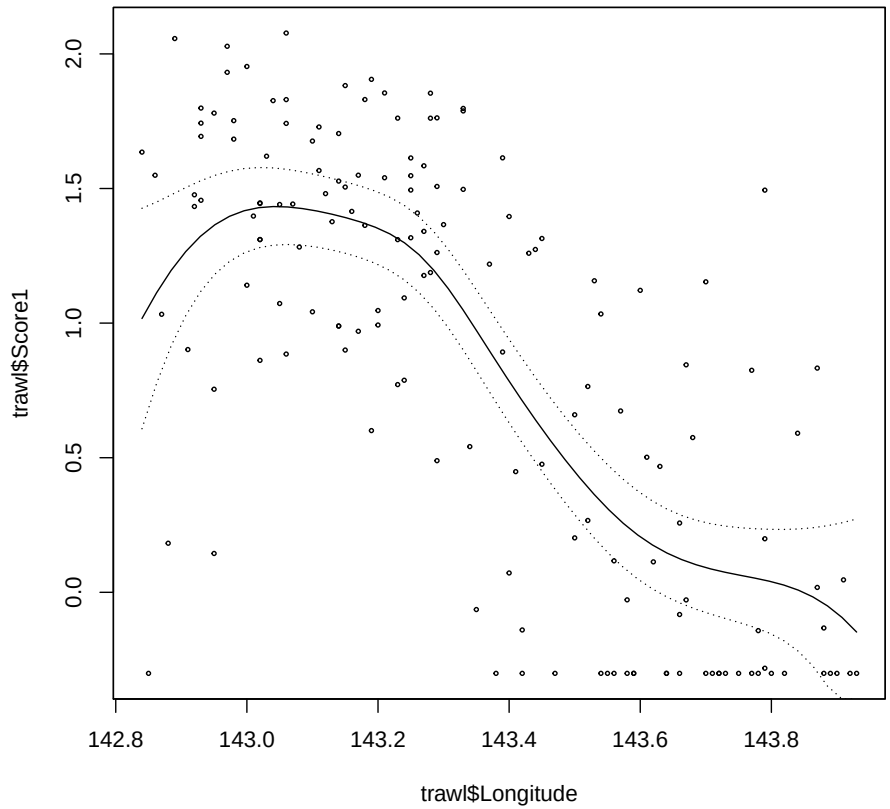
$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 \, w(x_i - x \, ; h)$$



The left plot shows an illustration of a window around a target point, the right plot gives an example of a normal kernel density with different values for standard deviation which would determine the width of the window before weights tail-off to zero.

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`.

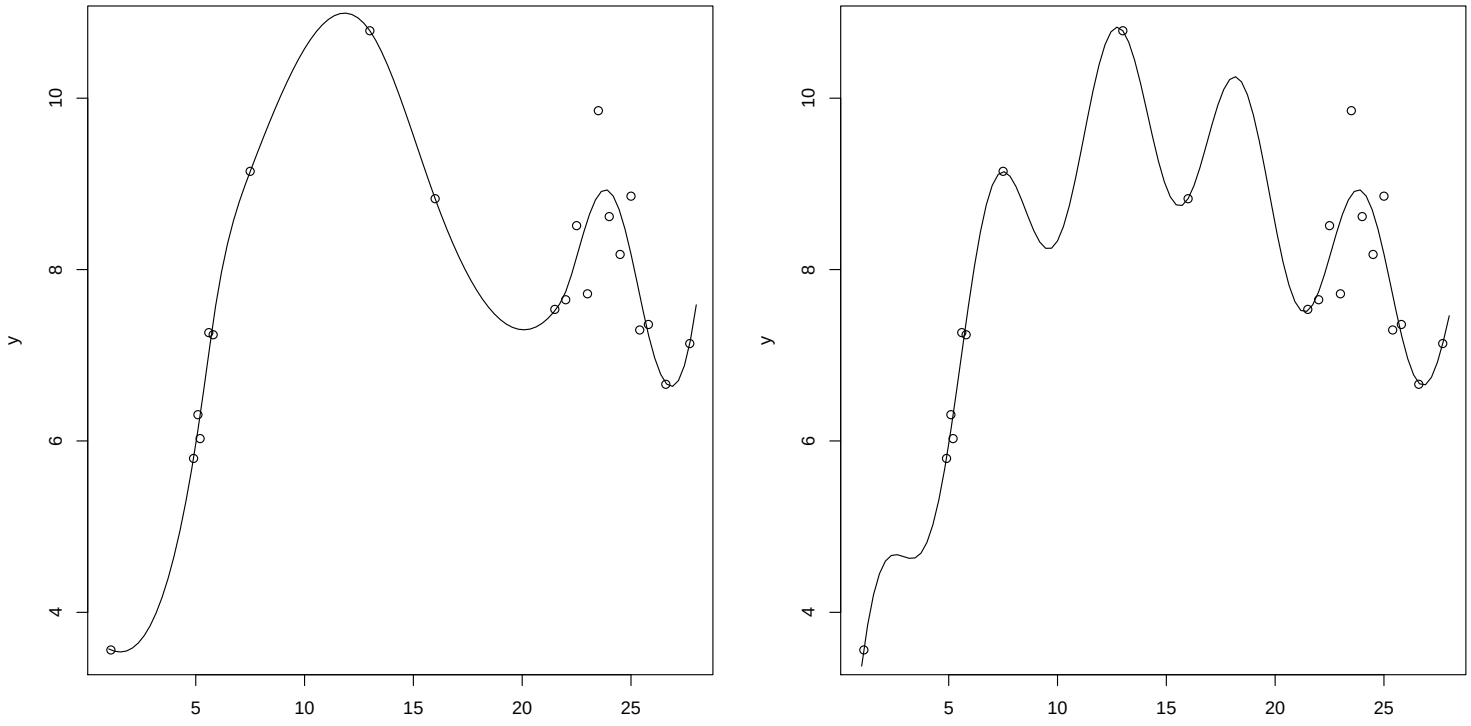


2.2.2 A local fitting approach - properties

- 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).

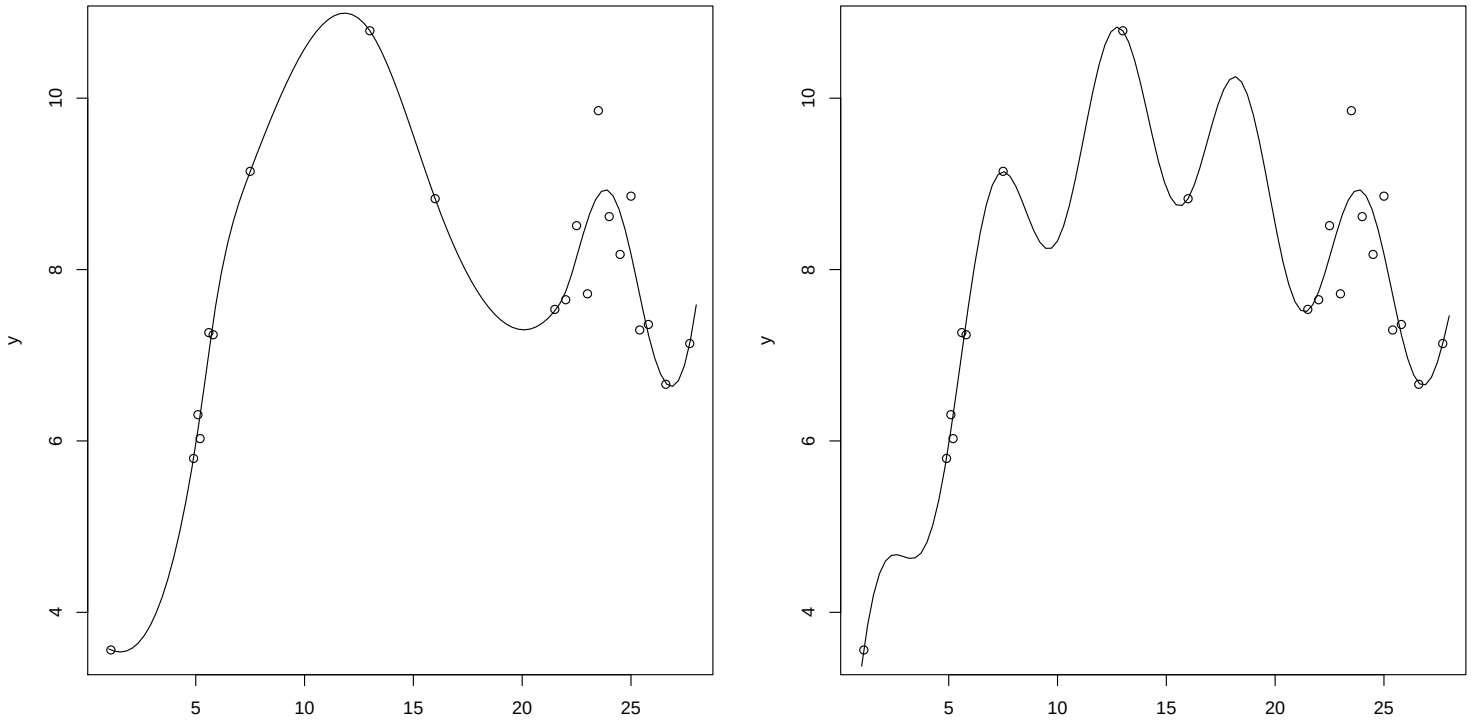
2.3 Regression splines

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 $\hat{y}_i = \hat{f}(x_i)$!

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

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i \qquad \text{for } i = 1, \ldots, n,$$

In matrix-vector notation:

$$\mathbb{E}(\mathrm{y}) = \mathrm{B}\boldsymbol{\beta} \qquad \text{with } \mathrm{y} = (Y_1, \ldots, Y_n)^\top \text{ and } \mathrm{B} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

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

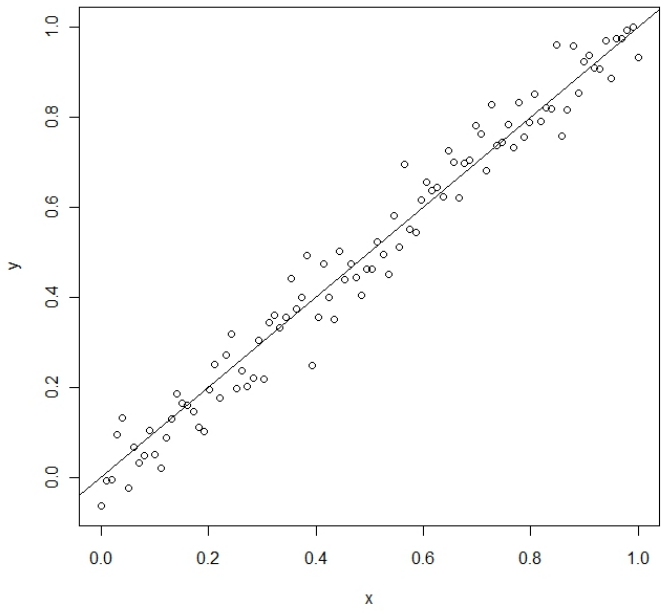


Figure: 2.5 A simple linear regression line with underlying simulated data

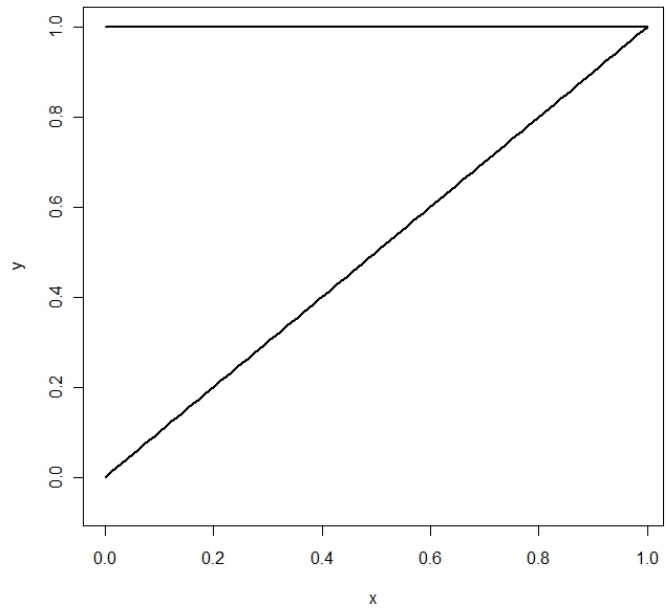


Figure: 2.6 The basis functions for simple linear regression $1, x$

2.3.1 Regression splines - polynomial regression

Polynomial regression

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \ldots + \beta_r x_i^r \qquad \text{for } i = 1, \ldots, n,$$

just corresponds to

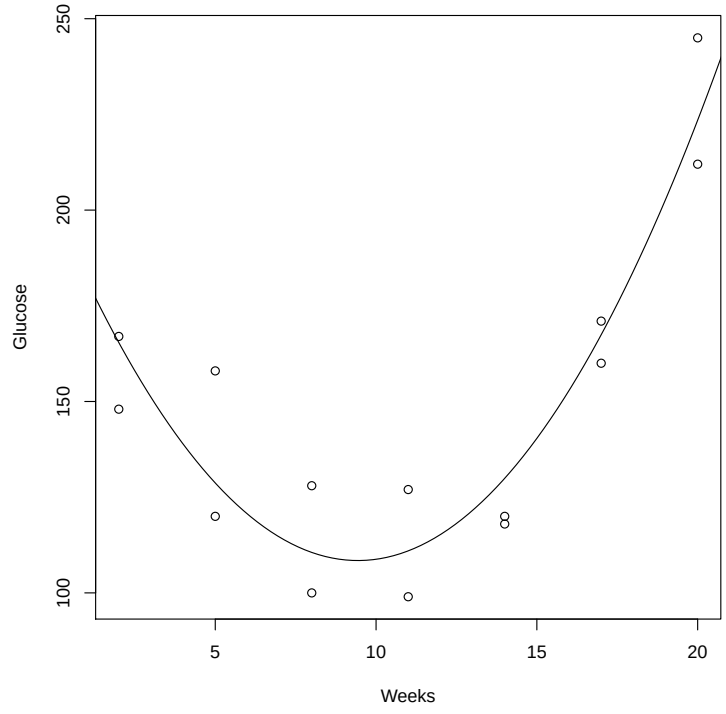
$$\mathsf{B} = \begin{pmatrix} 1 & x_1 & \ldots & x_1^r \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^r \end{pmatrix}.$$

Polynomial regression is a basis expansion technique.

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

2.3.1 Regression splines - polynomial regression

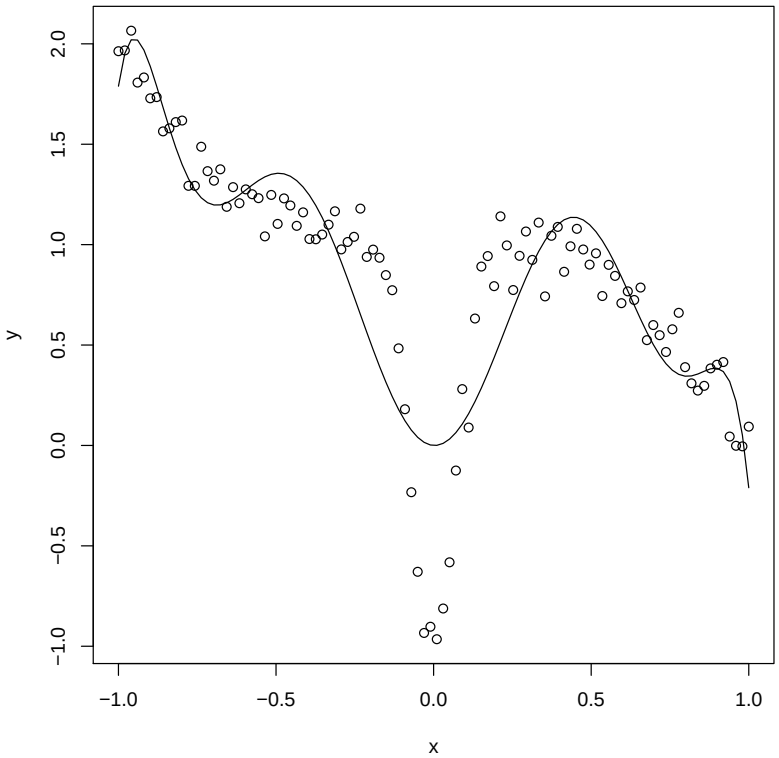
Example 2.4: Glucose levels in potatoes



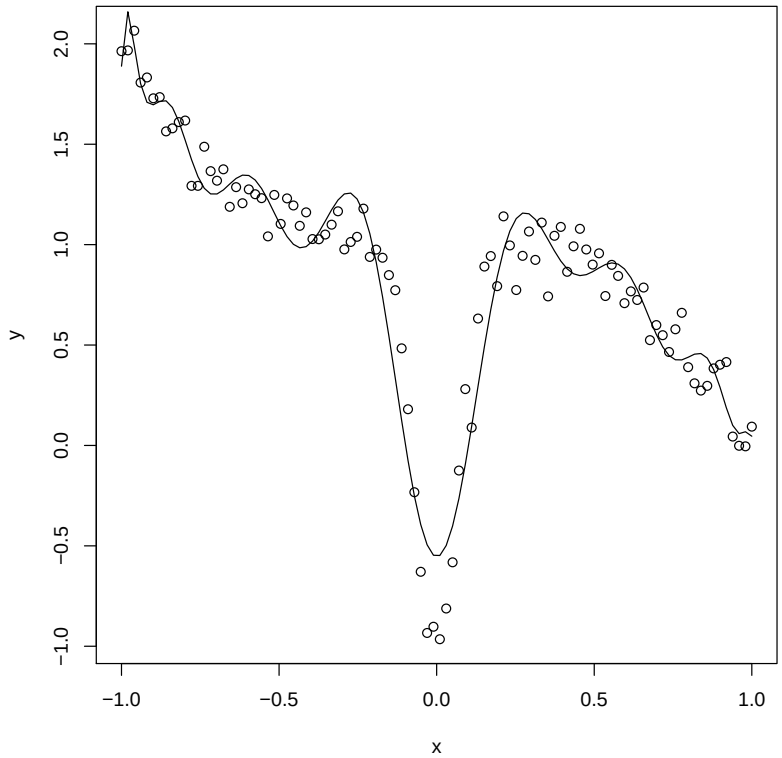
Polynomial regression can be a useful tool for small datasets.
(if a small degree of polynomial is used)

2.3.1 Regression splines - polynomial regression

Simulated example 2.5 : $y_i = 1 - x_i^3 - 2\exp(-100x_i^2) + \varepsilon_i$ with $\mathrm{x} = (-1, -0.98, \ldots, 0.98, 1)$ and $\varepsilon_i \sim \mathrm{N}(0, 0.1^2)$.



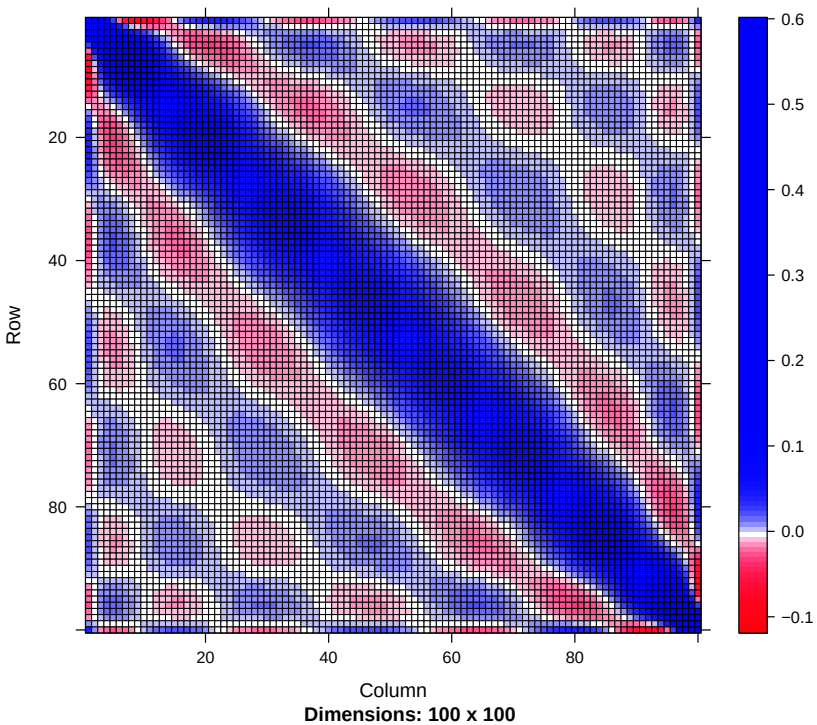
Polynomial regression (degree $r = 10$)



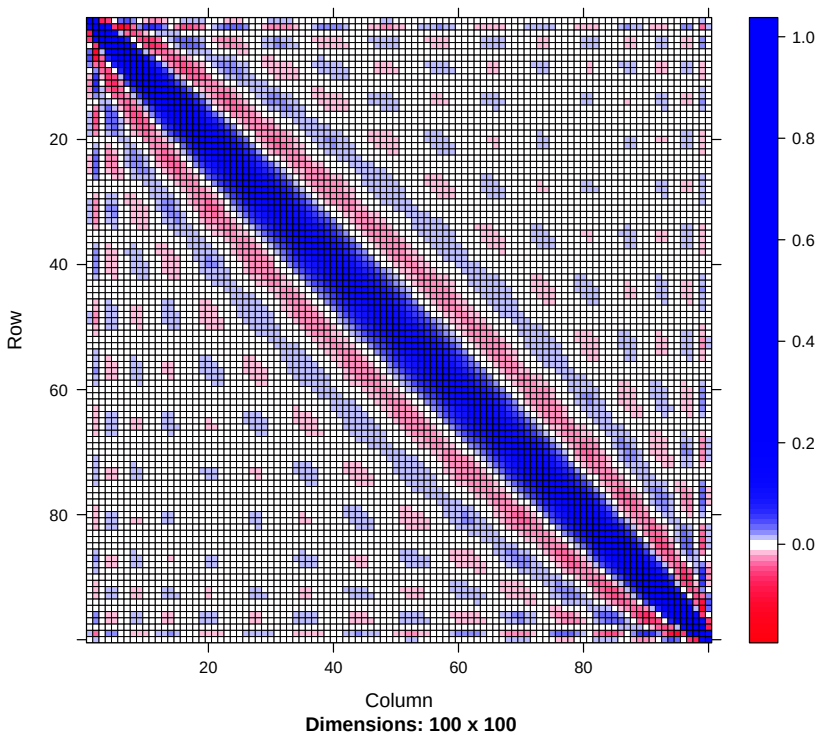
Polynomial regression (degree $r = 17$)

2.3.1 Regression splines - polynomial regression - what is going wrong?

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

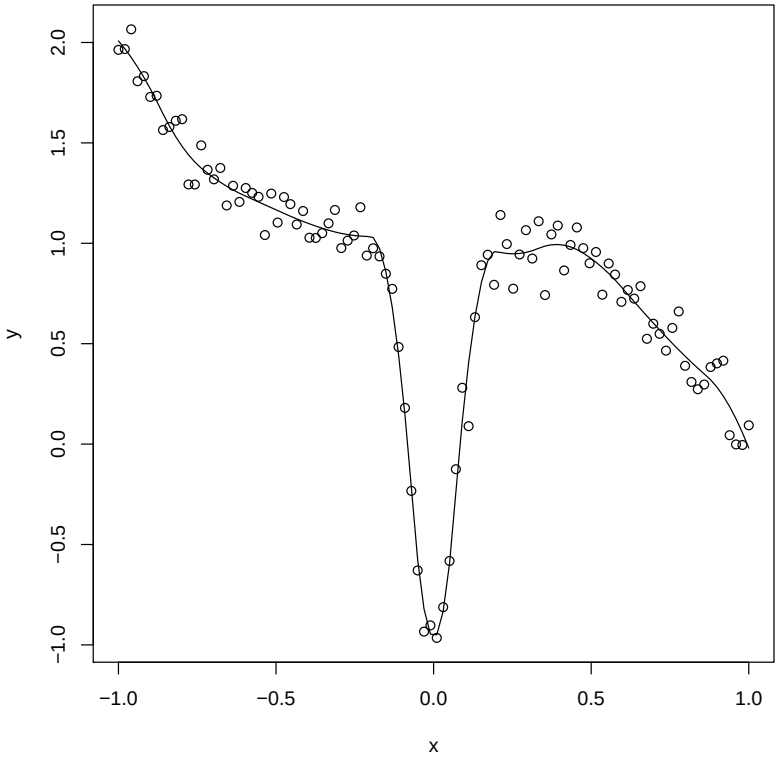


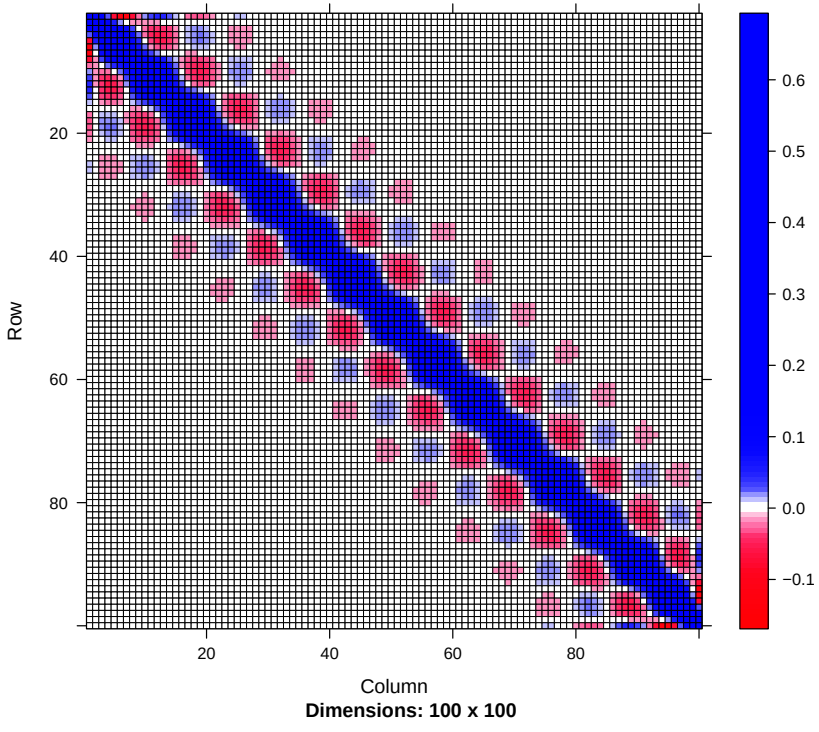
Polynomial regression (degree $r = 10$)



Polynomial regression (degree $r = 17$)

2.3.1 Regression splines: spline-based model

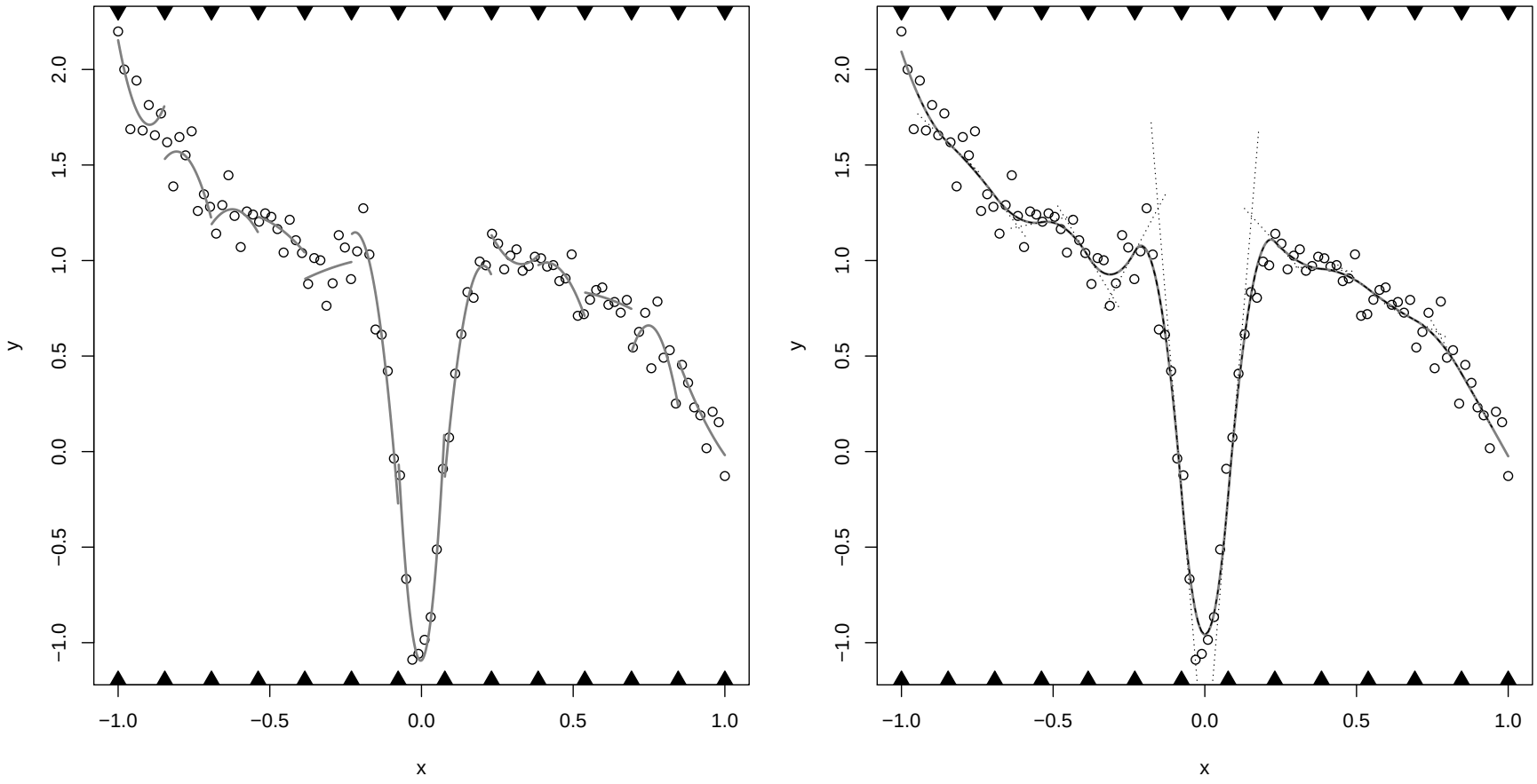




2.3.1 Regression splines - polynomial regression - problems

- 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 $x \to \pm\infty$)
- Also very likely to be numerically unstable.
  In our example:
  - Condition number of $\mathbf{B}^\top\mathbf{B}$: 1.56×10^{12}
  - For a B-spline: Condition number of $\mathbf{B}^\top\mathbf{B}$: 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).

2.3.2 Regression splines

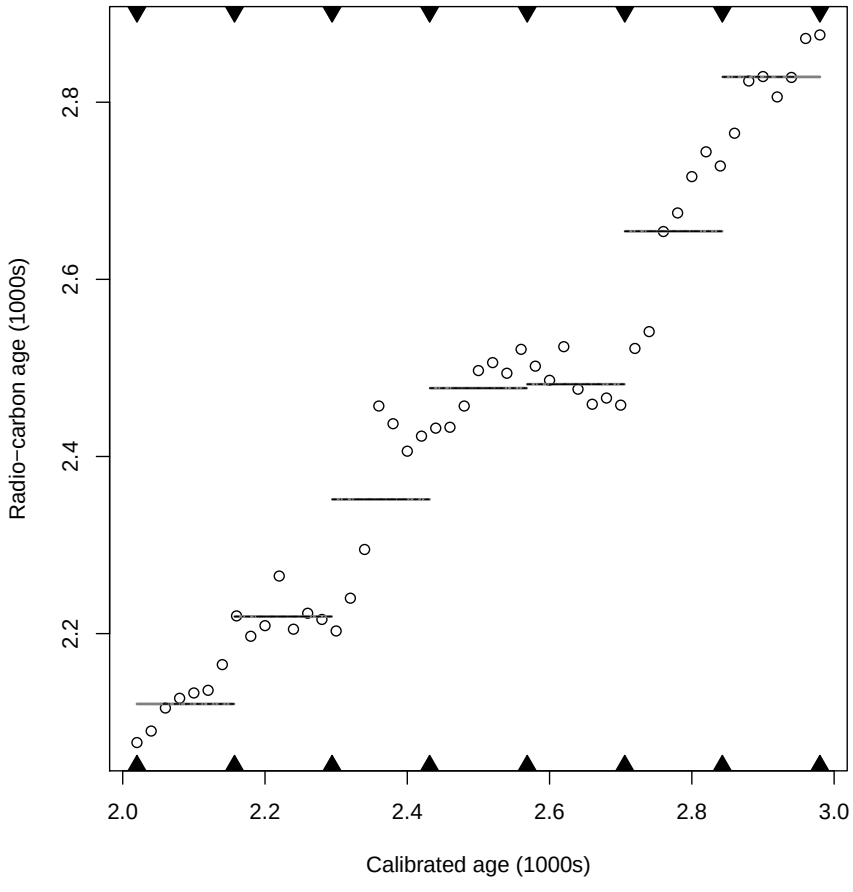
Definition 2.1: (Polynomial) spline

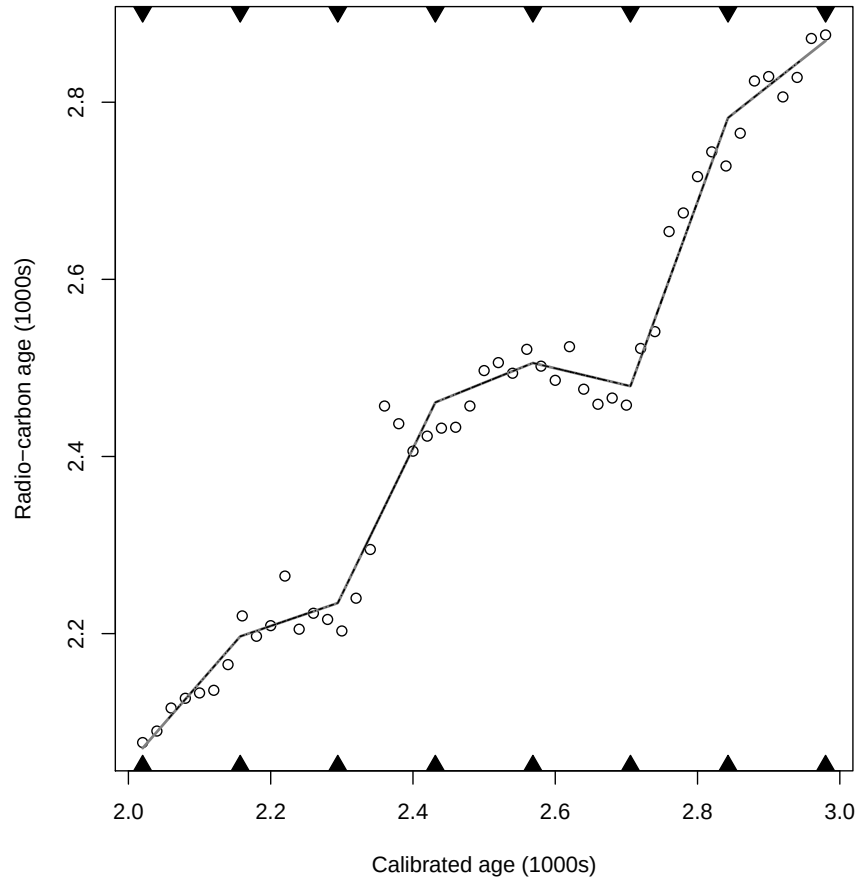
Given a set of knots $a = \kappa_1 < \kappa_2 < \ldots < \kappa_l = b$, a function $f : [a, b] \to \mathbb{R}$ is called a *(polynomial) spline* of degree r if

- $f(\cdot)$ is a polynomial of degree r on each interval (κ_j, κ_{j+1}) $(j = 1, \ldots, l - 1)$.
- $f(\cdot)$ is $r - 1$ times continuously differentiable.

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

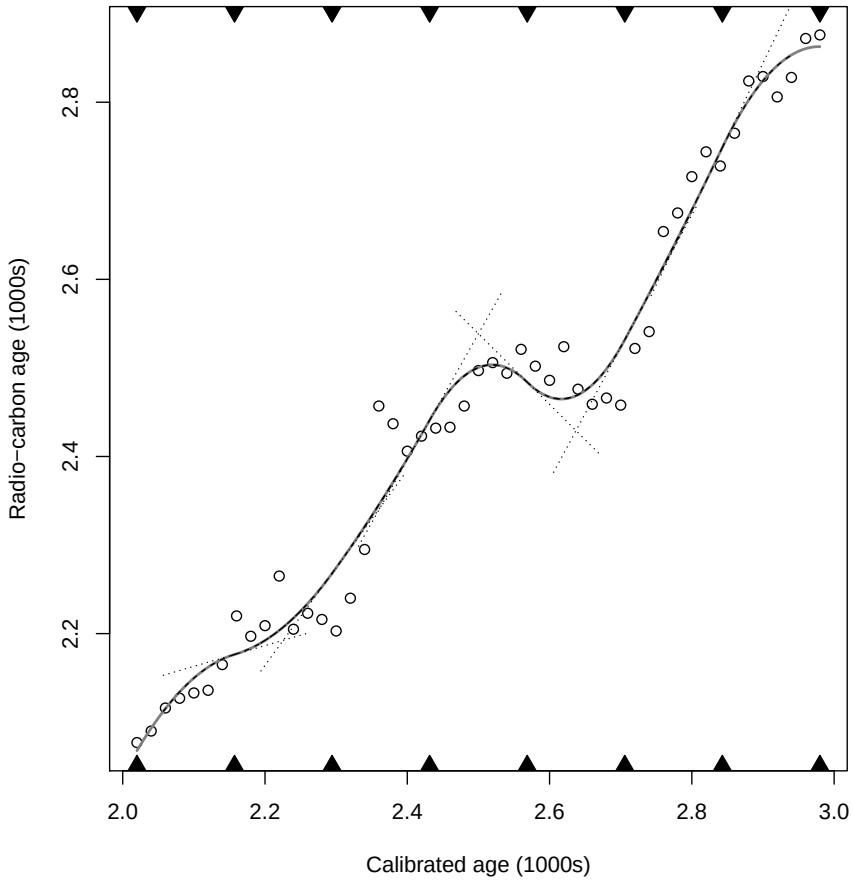
Radiocarbon dating (Figure 2.12)

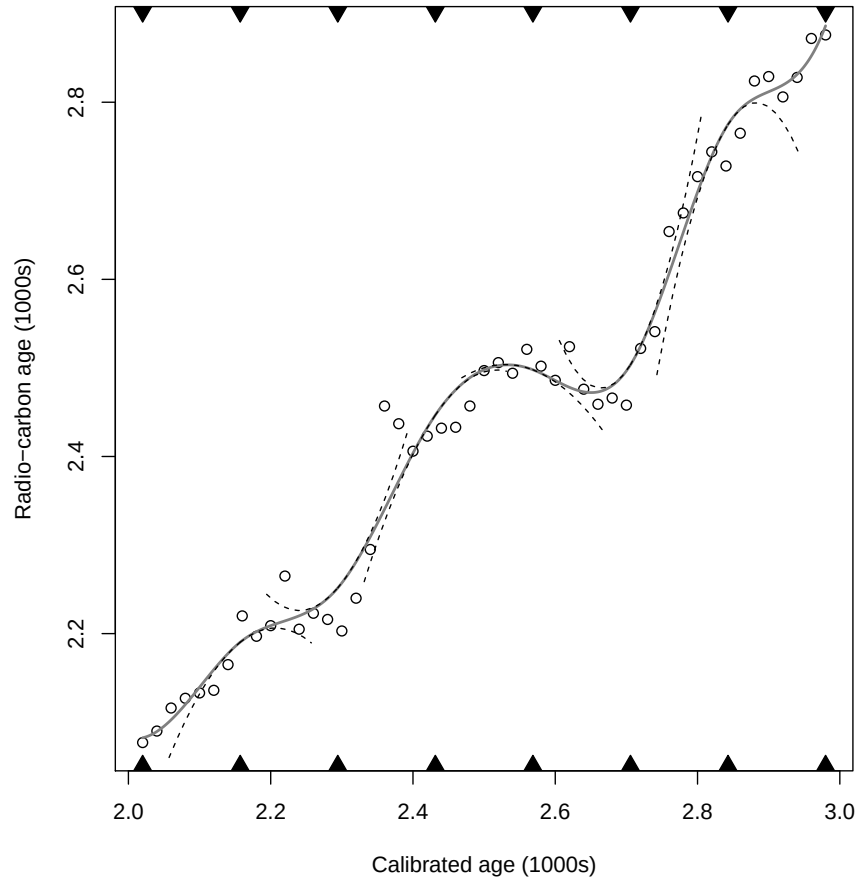




2.3.2 Regression splines: degrees $r = 2$ and $r = 3$

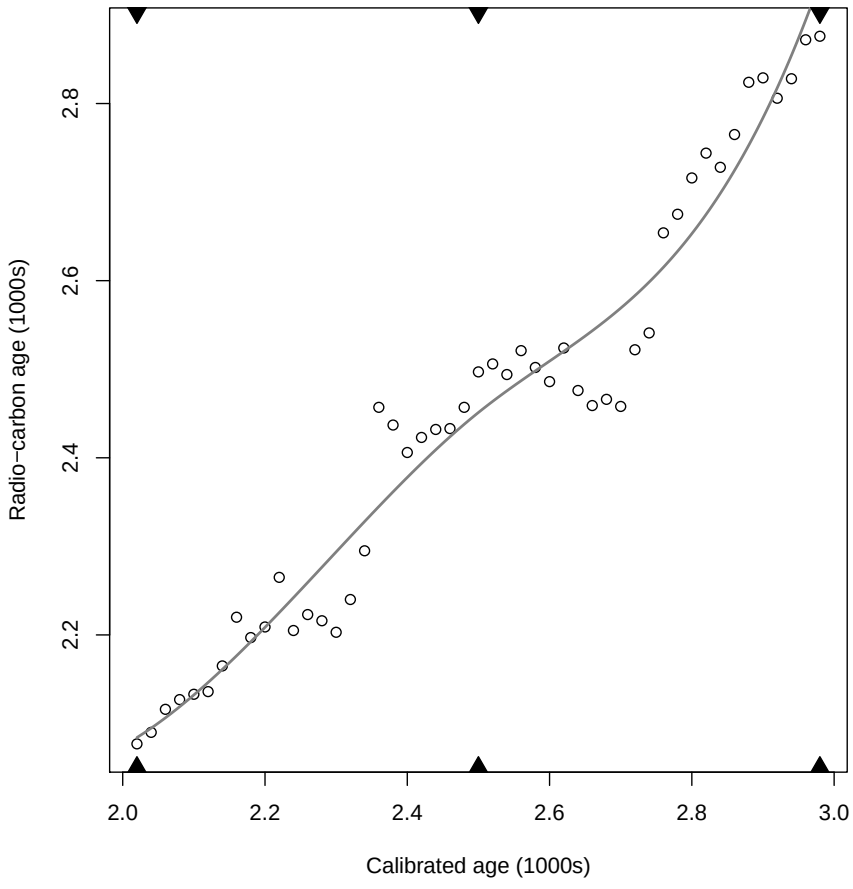
Radiocarbon dating



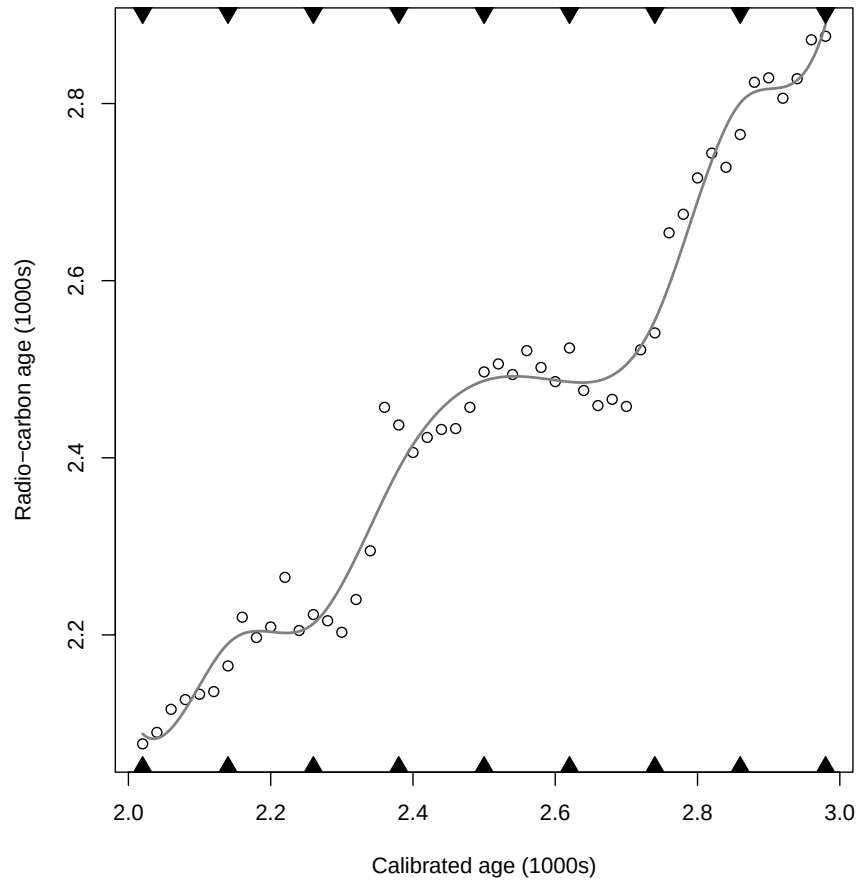


2.3.2 Regression splines: different numbers of knots (1)

Radiocarbon dating (Figure 2.13)



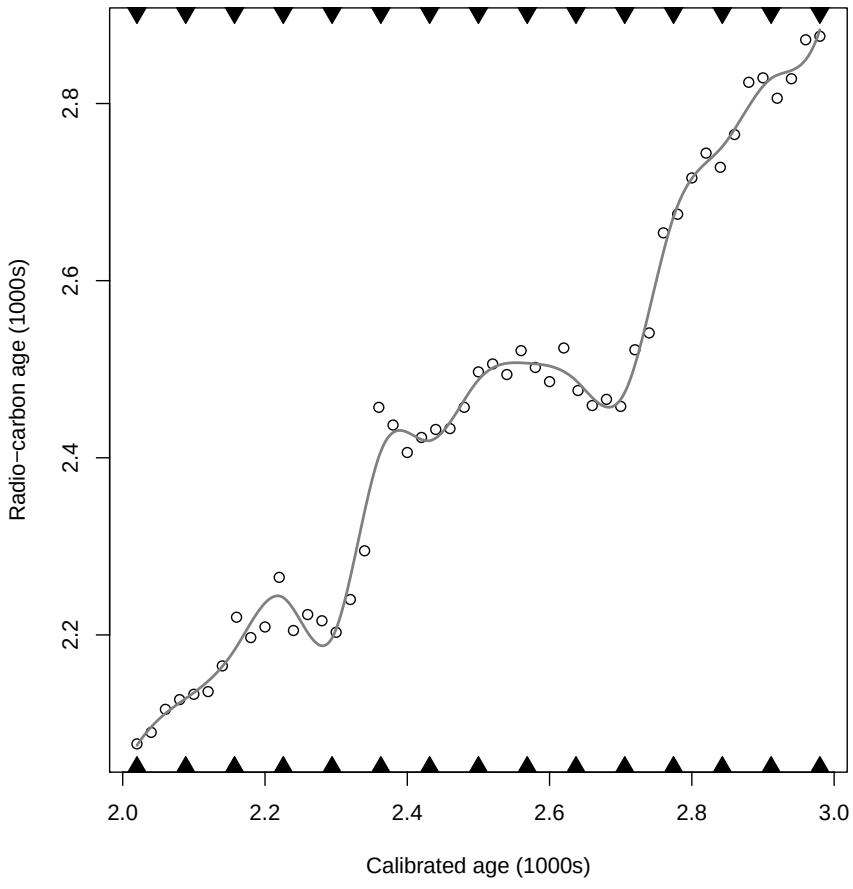
$l = 3$ knots



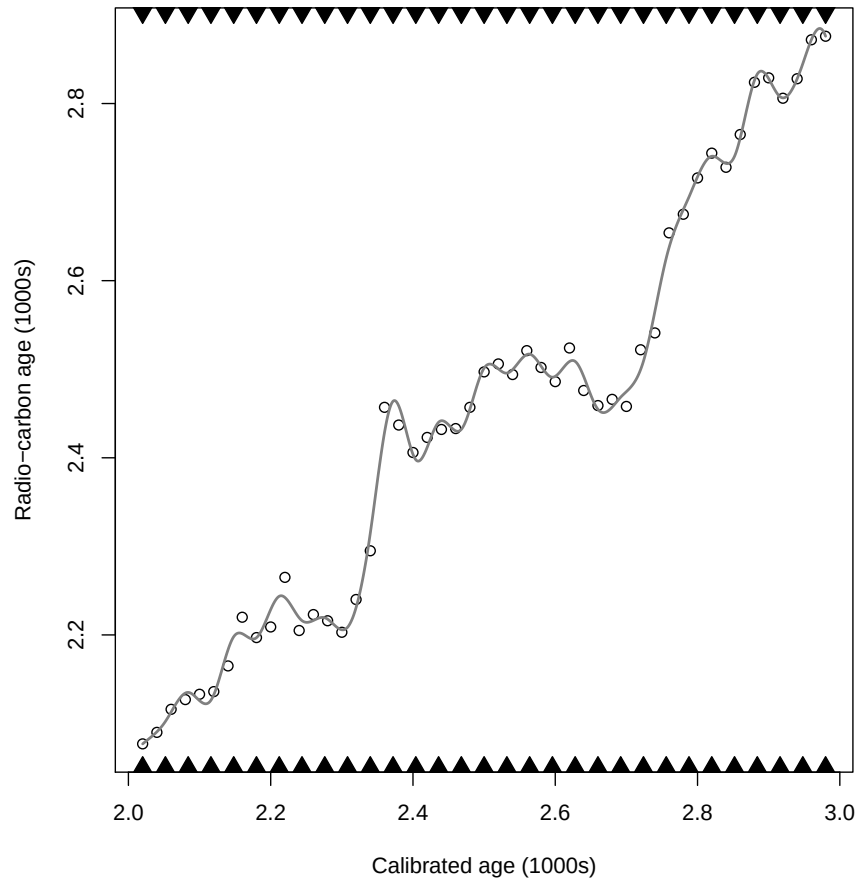
$l = 9$ knots

2.3.2 Regression splines: different numbers of knots (2)

Radiocarbon dating



$l = 15$ knots



$l = 31$ 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 l

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

2.3.4 Regression splines - how to fit?

Minimise

$$\sum_{i=1}^{n}(y_i - f(\mathrm{x}_i))^2.$$

Represent $f(\mathrm{x}_i)$ as $\mathrm{B}\boldsymbol{\beta}$.

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

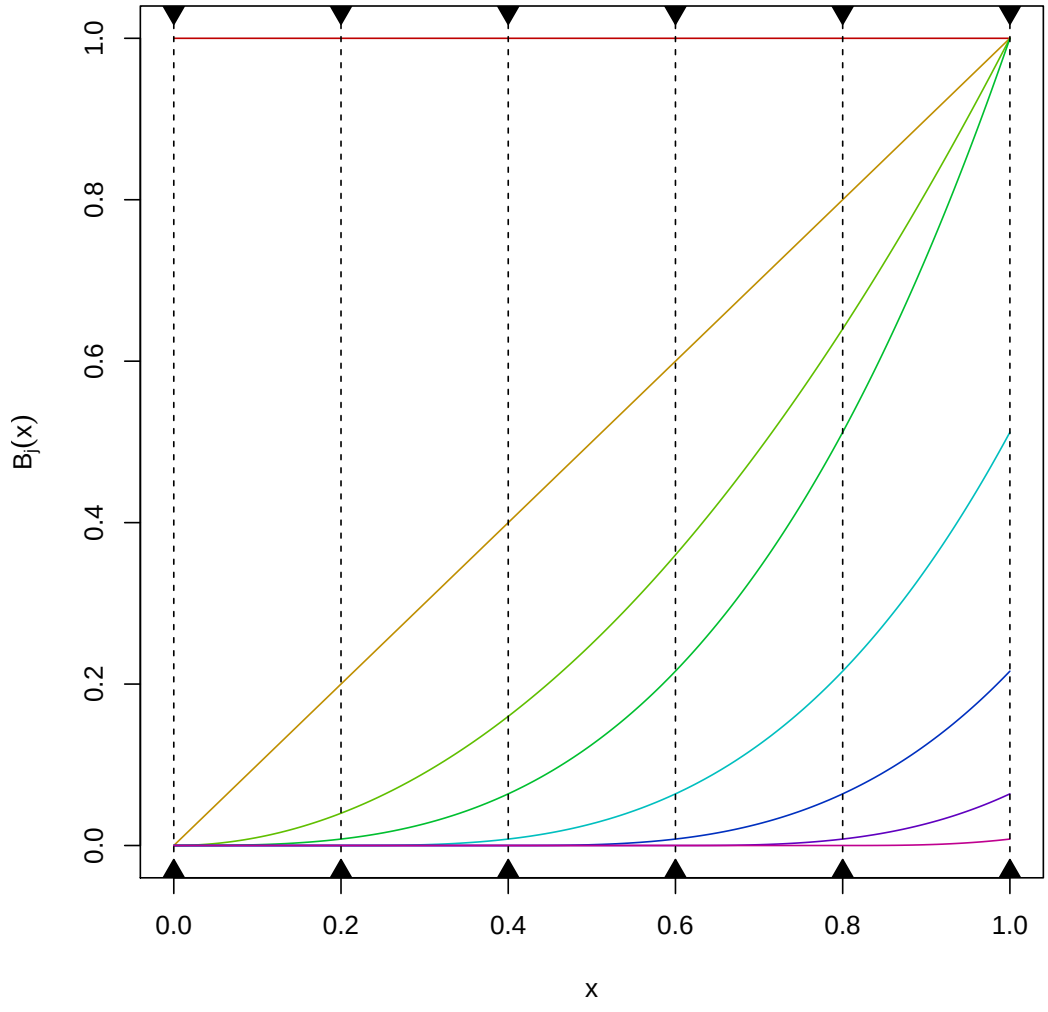
Given a set of knots $\kappa_1 < \ldots < \kappa_m$ the truncated power basis of degree r is given by

$$\left(1, x, \ldots, x^r, (x-\kappa_1)_+^r, (x-\kappa_2)_+^r, \ldots, (x-\kappa_m)_+^r\right),$$

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

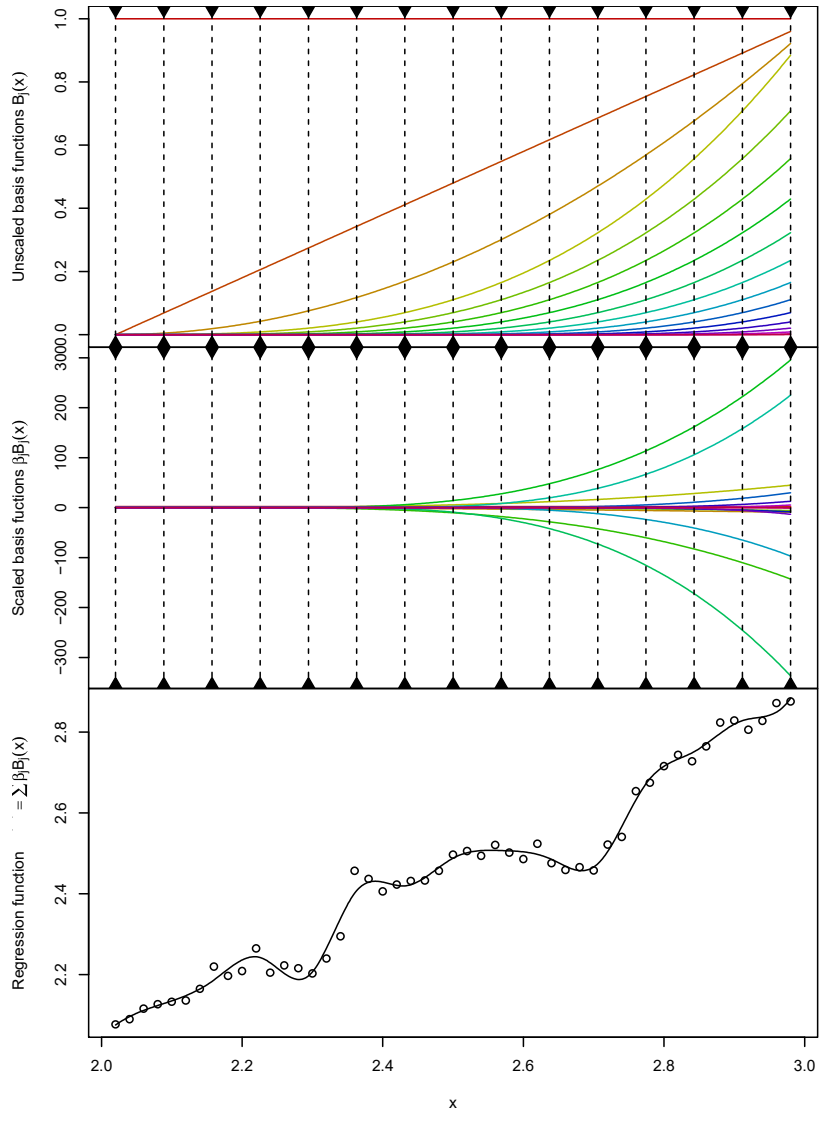
$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_r x^r + \beta_{r+1}(x-\kappa_1)_+^r + \ldots + \beta_{r+m}(x-\kappa_m)_+^r$$

- This is just a linear model with design matrix

$$\mathrm{B} = \begin{pmatrix} 1 & x_1 & \ldots & x_1^r & (x_1-\kappa_1)_+^r & \ldots & (x_1-\kappa_m)_+^r \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^r & (x_n-\kappa_1)_+^r & \ldots & (x_n-\kappa_m)_+^r \end{pmatrix}$$

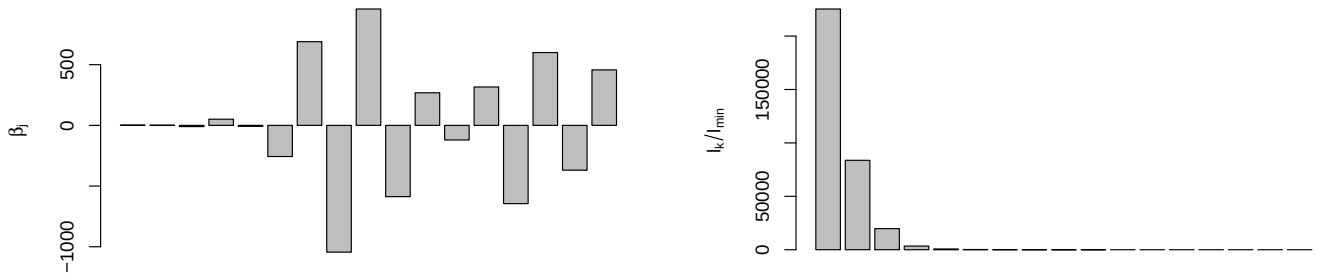
2.3.4 Regression splines - truncated power series

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)
$\rightsquigarrow$ $B^\top B$ ill conditioned (condition number: 5.85×10^9)

2.3.4 Regression splines - B-splines

Definition 2.7: B-spline basis

(a) Given a set of l knots the B-spline basis of degree 0 is given by the functions $(B_1^0(x), \ldots, B_{l-1}^0(x))$ with

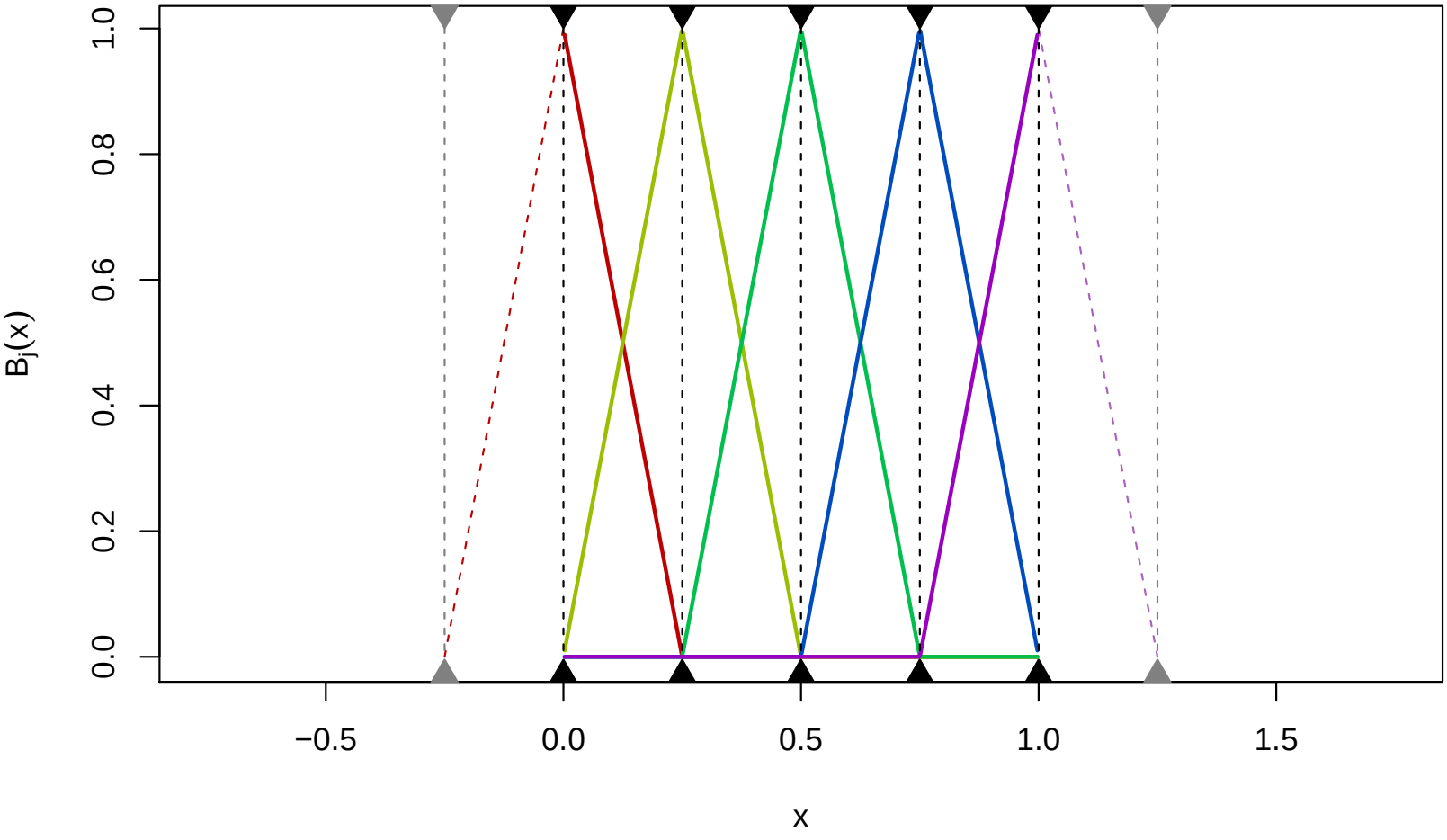
$$B_j^0(x) = \begin{cases} 1 & \text{for } \kappa_j \leq x < \kappa_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

(b) Given a set of l knots the B-spline basis of degree $r > 0$ is given by the functions $(B_1^r(x), \ldots, B_{l+r-1}^r(x))$ with

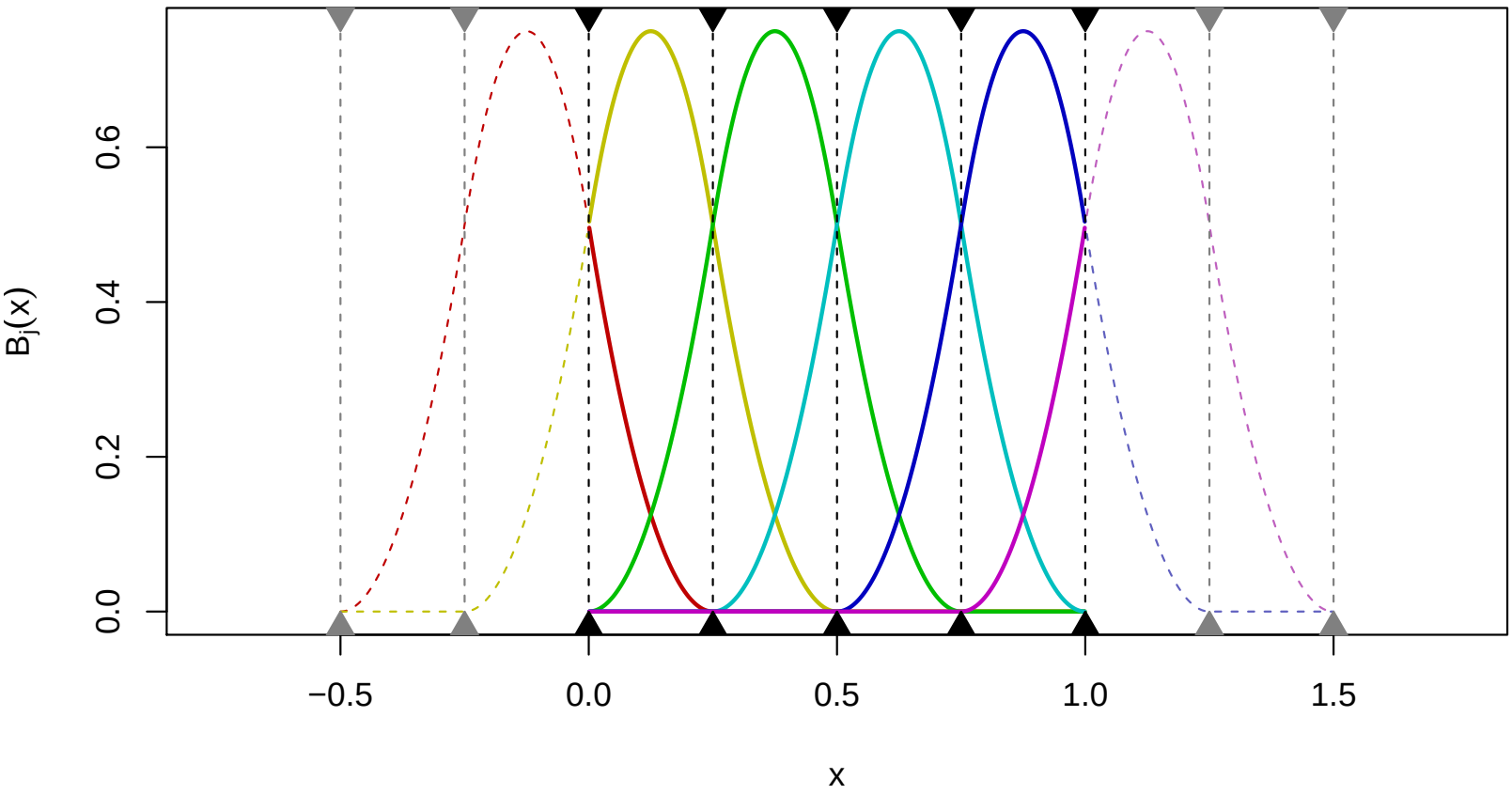
$$B_j^r(x) = \frac{x - \kappa_{j-r}}{\kappa_j - \kappa_{j-r}} B_{j-1}^{r-1}(x) + \frac{\kappa_{j+1} - x}{\kappa_{j+1} - \kappa_{j+1-r}} B_j^{r-1}(x).$$

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

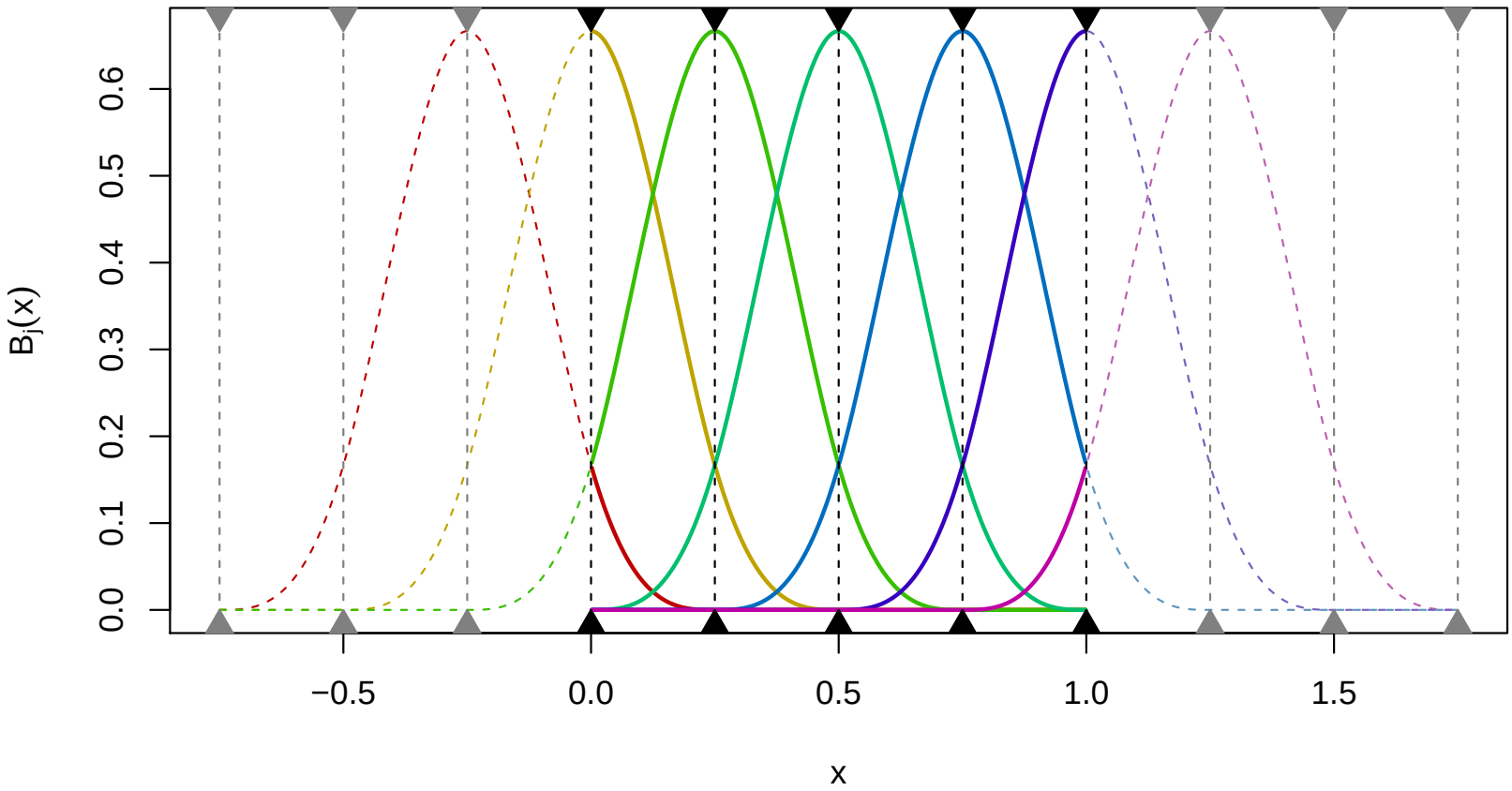
Figure 2.20:



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

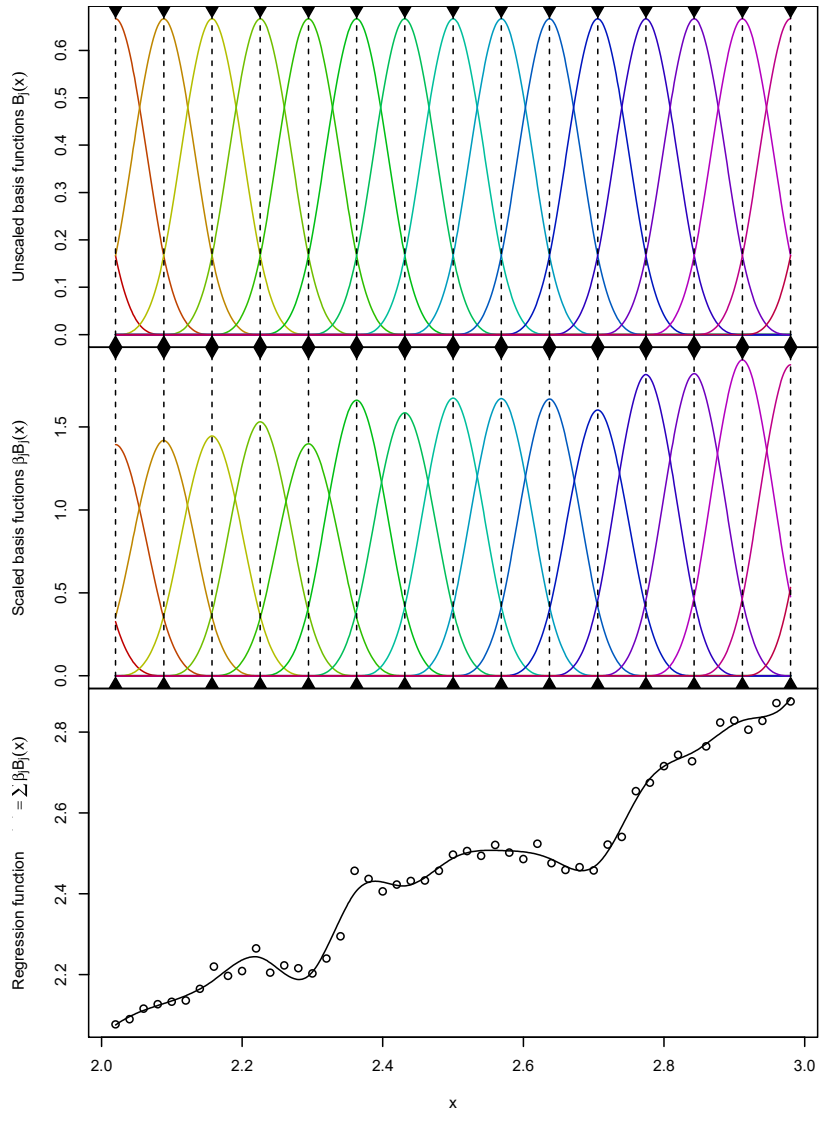
$$f(x) = \sum_{j=1}^{l+r-1} \beta_j B_j(x)$$

- This is just a linear model with design matrix

$$\mathrm{B} = \begin{pmatrix} B_1^r(x_1) & \dots & B_{l+r-1}^r(x_1) \\ \vdots & \ddots & \vdots \\ B_1^r(x_n) & \dots & B_{l+r-1}^r(x_n) \end{pmatrix}.$$

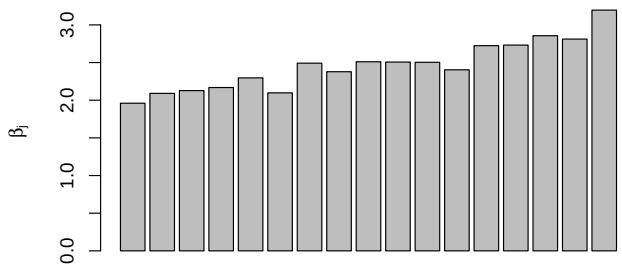
2.3.4 Regression splines: B-splines

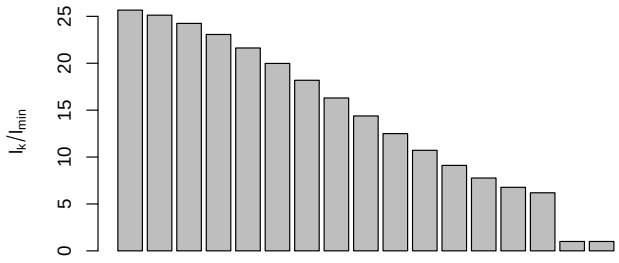
Illustration: Radiocarbon dating (Figure 2.24)



2.3.4 Regression splines: B-splines

No more numerical problems (Figures 2.22, 2.23)

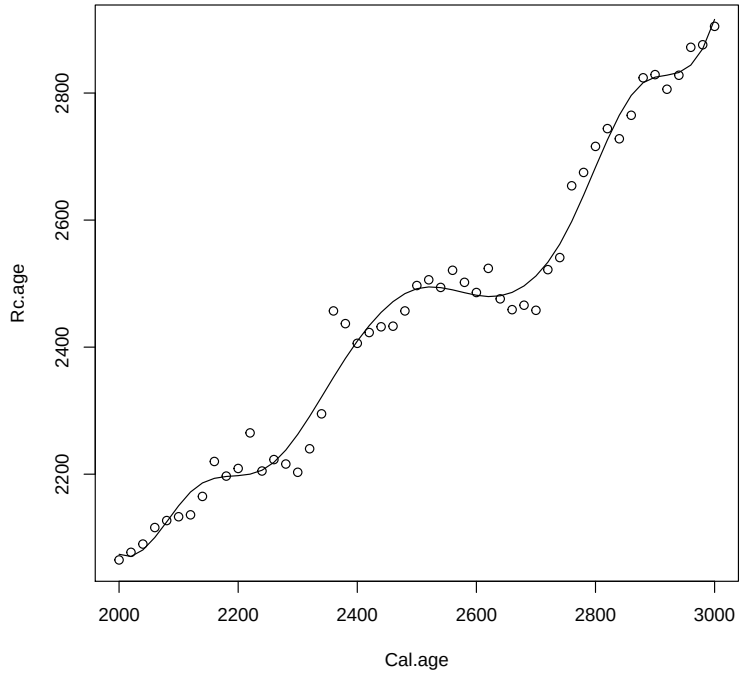




Problem solved: Basis functions not highly correlated (0.8309 at most)
$\rightsquigarrow$ $B^\top B$ not ill-conditioned (condition number: 358.263)

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 $\|\mathrm{D}\boldsymbol{\beta}\|^2$.

2.3.5 Penalised regression splines (P-splines) – Idea

- Objective function (**see ridge regression):

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \|\mathrm{D}\boldsymbol{\beta}\|^2$$

- λ 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

$$\hat{\boldsymbol{\beta}} = (\mathsf{B}^\top \mathsf{B} + \lambda \mathsf{D}^\top \mathsf{D})^{-1} \mathsf{B}^\top \mathsf{y},$$

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

$$\left\| \begin{pmatrix} \mathsf{y} \\ 0 \end{pmatrix} - \begin{pmatrix} \mathsf{B} \\ \sqrt{\lambda}\mathsf{D} \end{pmatrix} \boldsymbol{\beta} \right\|^2$$

2.3.6 Penalised regression splines - how to choose D?

Smoothing splines

One can show that

$$\int_a^b f''(x)^2 \, dx = \boldsymbol{\beta}^\top \begin{pmatrix} \int_a^b B_1''(x) B_1''(x) \, dx & \dots & \int_a^b B_1''(x) B_{l+r-1}''(x) \, dx \\ \vdots & \ddots & \vdots \\ \int_a^b B_1''(x) B_{l+r-1}''(x) \, dx & \dots & \int_a^b B_{l+r-1}''(x) B_{l+r-1}''(x) \, dx \end{pmatrix} \boldsymbol{\beta}$$

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

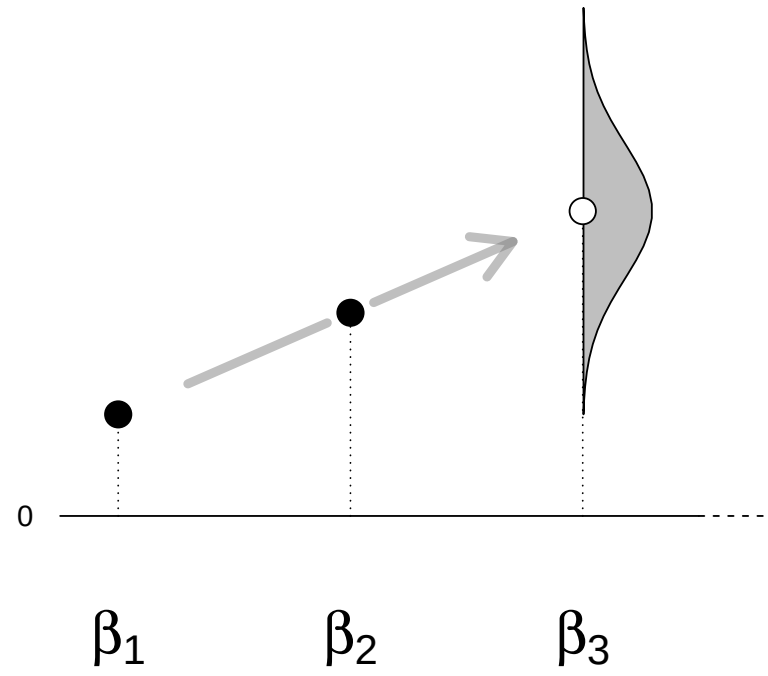
2.3.6 Penalised regression splines - difference penalties

For equally-spaced knots we can also use difference penalties (much simpler).

2.3.6 Penalised regression splines - second-order difference penalty

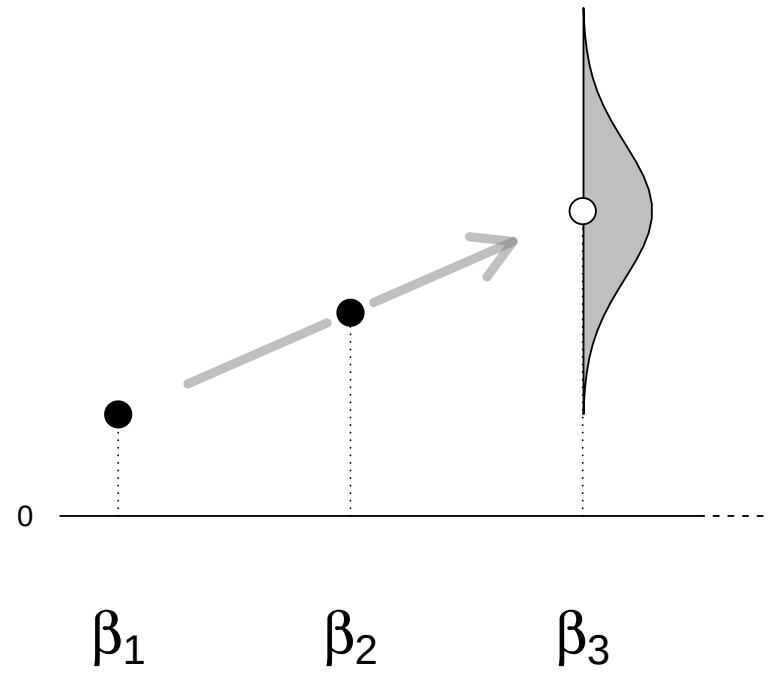
$$\mathsf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 1 & -2 & 1 \end{pmatrix}.$$

$\|\mathsf{D}_2\boldsymbol{\beta}\|^2 = \sum(\beta_{j+2} - 2\beta_{j+1} + \beta_j)^2$
(second-order differences)



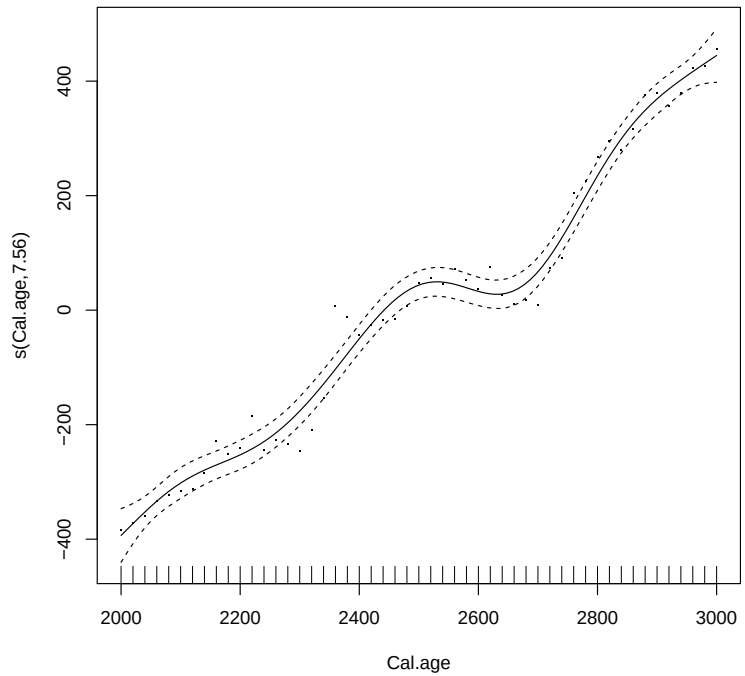
2.3.6 Penalised regression splines - second-order difference penalty

- 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)