Lecture 10
Polynomial regression

BIOST 515

February 5, 2004

Polynomial regression models

$$y = X\beta + \epsilon$$

is a general linear regression model for fitting any relationship that is linear in the unknown parameters, β. For example, the following polynomial

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2 + \beta_5 x_2^2 + \epsilon$$

is a linear regression model because y is a linear function of β.

Polynomial models in one variable

A kth order polynomial in one variable is defined as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon.$$

Polynomial models are useful

- in situations where the analyst knows that curvilinear effects are present in the true response function
- as approximating functions to unknown and possibly very complex nonlinear relationships.

We can think of the polynomial model as the Taylor series expansion of the unknown function.

Important considerations

- Order of the model
- Model building strategy
- Extrapolation
- Ill-conditioning
- Hierarchy

Piecewise polynomials

A low-order polynomial may provide a poor fit to the data, and increasing the order of the polynomial may not help. Transformations of x or y may solve this problem, but sometimes we may prefer to use more flexible approaches. One such approach is to use **splines**.

- piecewise polynomials used in curve fitting
- polynomials within intervals of x that are connected acoress different intervals of x

The piecewise linear spline function is given by

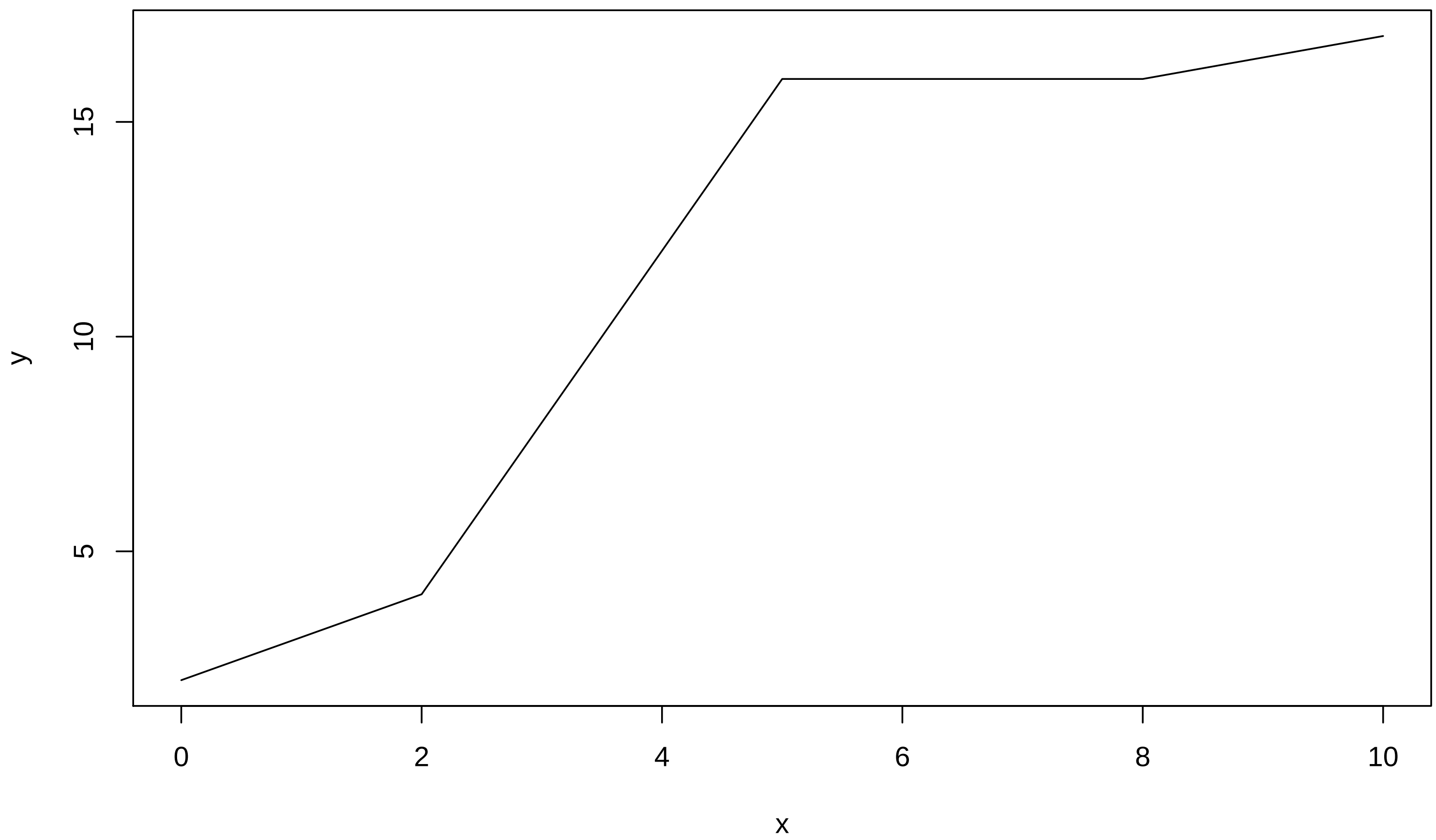
$$f(x) = \beta_0 + \beta_1 x + \beta_2(x-a)_+ + \beta_3(x-b)_+ + \beta_4(x-c)_+,$$

where

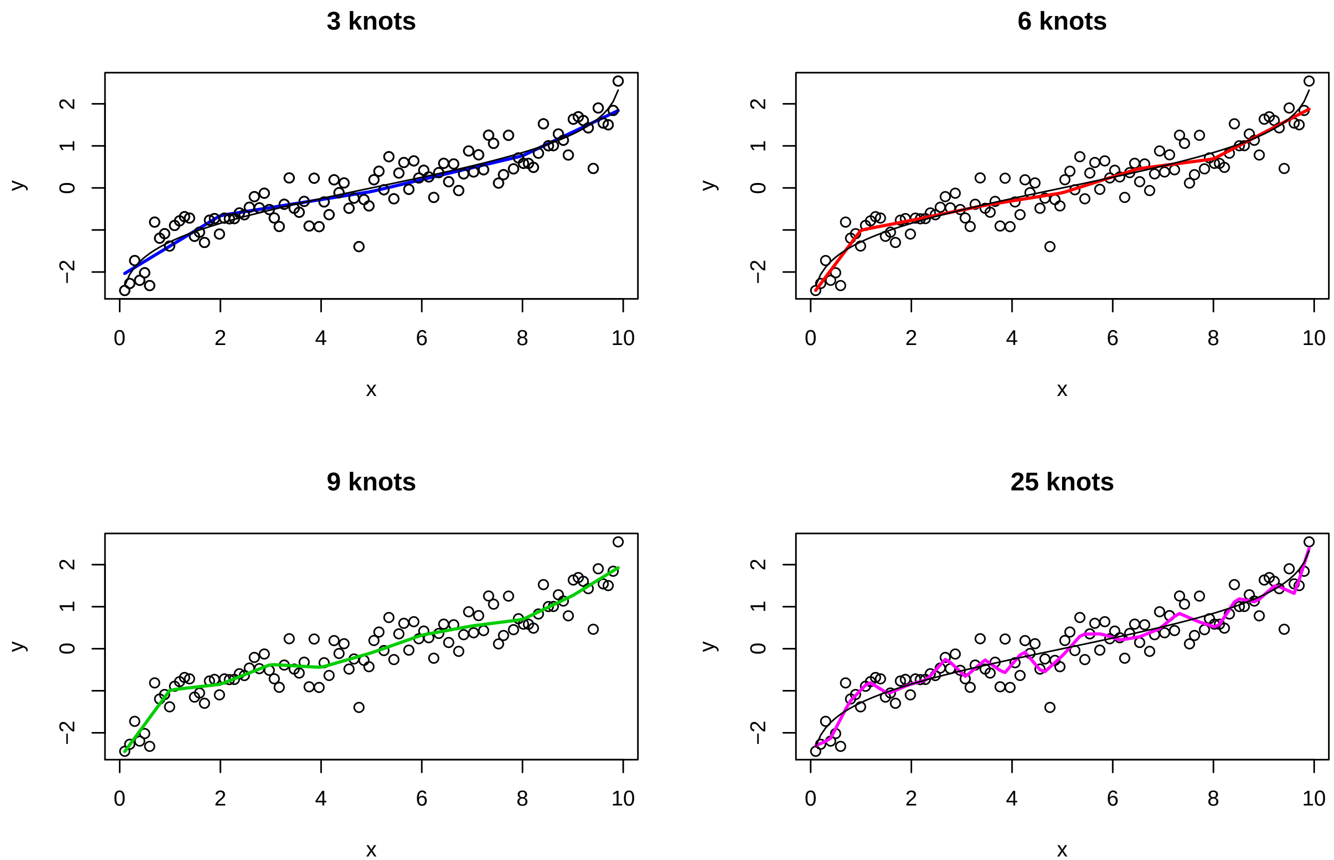
$$(u)_+ = \begin{cases} u, & u > 0, \\ 0, & u \leq 0 \end{cases}$$

and a, b and c are referred to as knots.

Example of piecewise linear spline with knots at 2, 5 and 8.



As we increase the number of knots, the piecewise linear polynomial more closely resembles a continuous line.



Cubic splines

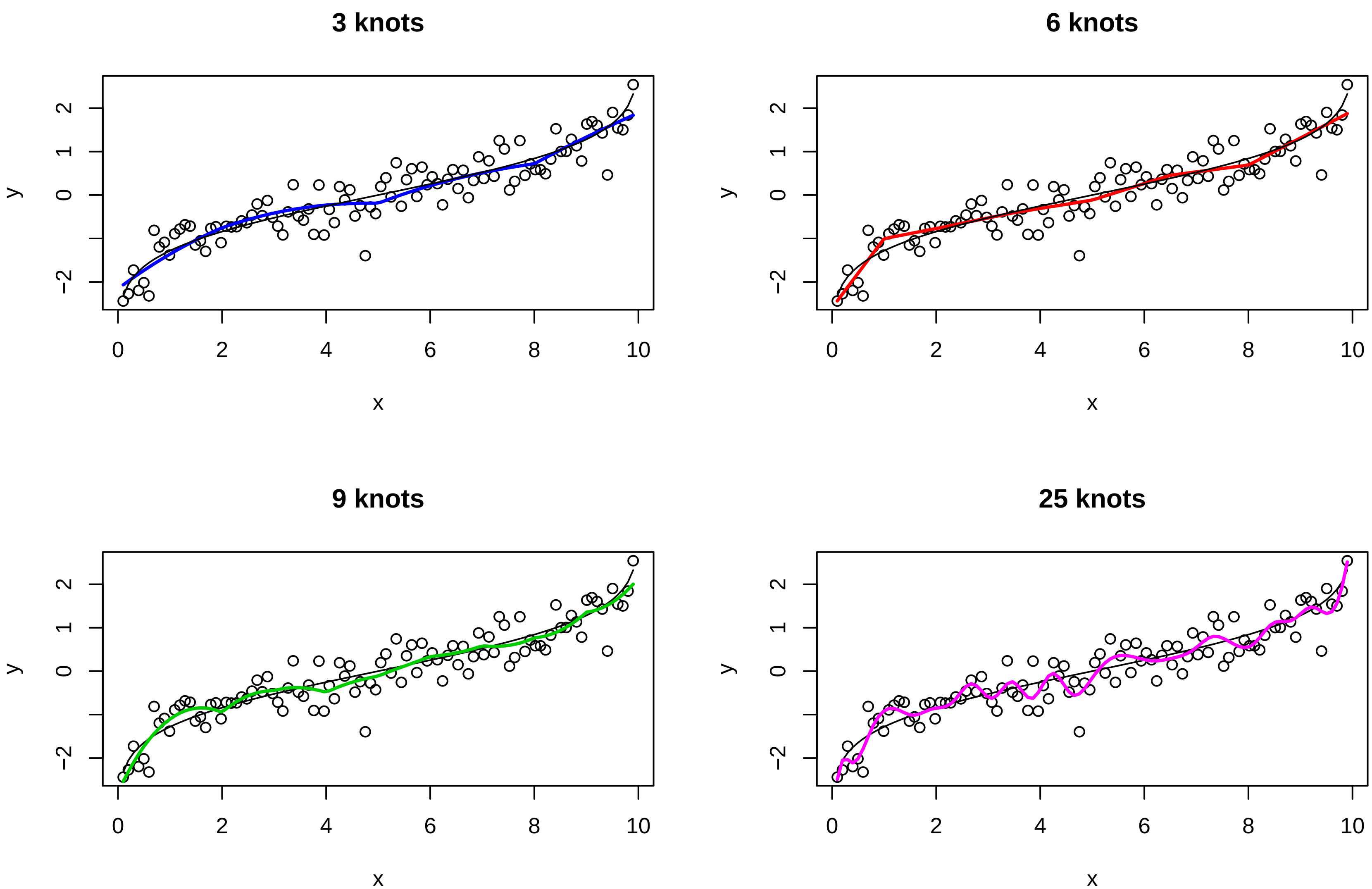
Although, linear splines may work well, they are not smooth and will not fit highly curved functions well (unless many knots are used - which requires a lot of data).

It is more common for cubic splines to be used in practice. A cubic spline function with k knots is given by

$$f(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{l=1}^{k} \beta_i (x - t_l)_+^3,$$

where $t_l, \ l = 1, \ldots, k$ are the k knots. We relate x to the outcome as

$$y_i = f(x_i) + \epsilon_i.$$


3 knots
6 knots
9 knots
25 knots
y
x
−2
0
1
2
0
2
4
6
8
10

For estimation purposes, we assume that both the locations and the number of knots are fixed. Although there are methods that allow the number and/or position of the knots to be random; these models are too complex to be fit using least squares.

The piecewise cubic splines may give us a more flexible model, but they still may be discontinuous at the knots.

Continuous cubic splines

Cubic B splines

- Given k knots at $t_1, \ldots, t_k$, a cubic B spline function is a cubic polynomial on the interval $[t_j, t_{j+1}]$,

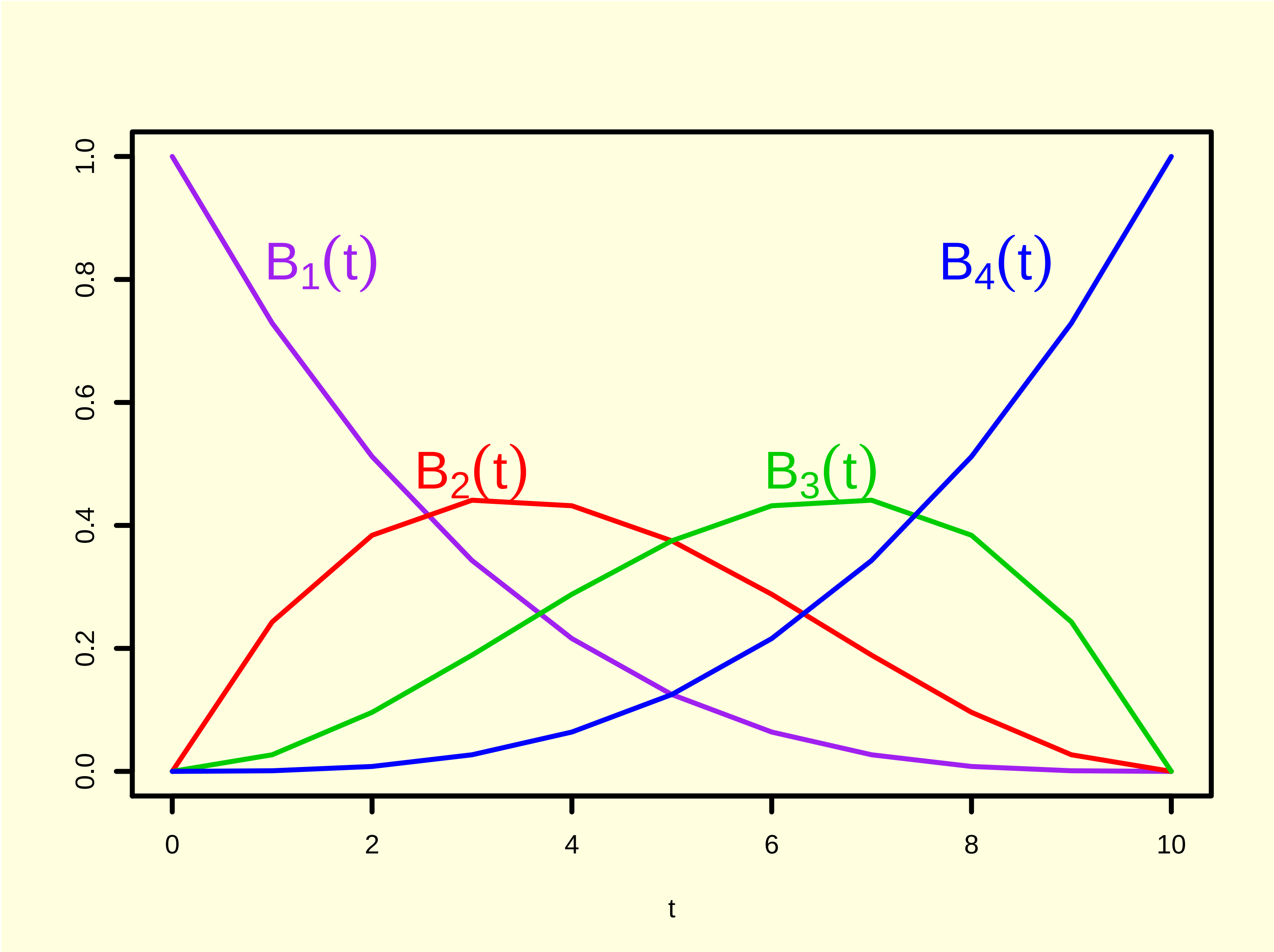
- It has continuous first and second derivatives, imposing 3 conditions at each knot.

- With k knots, $k + 1$ parameters are needed to represent the cubic spline.

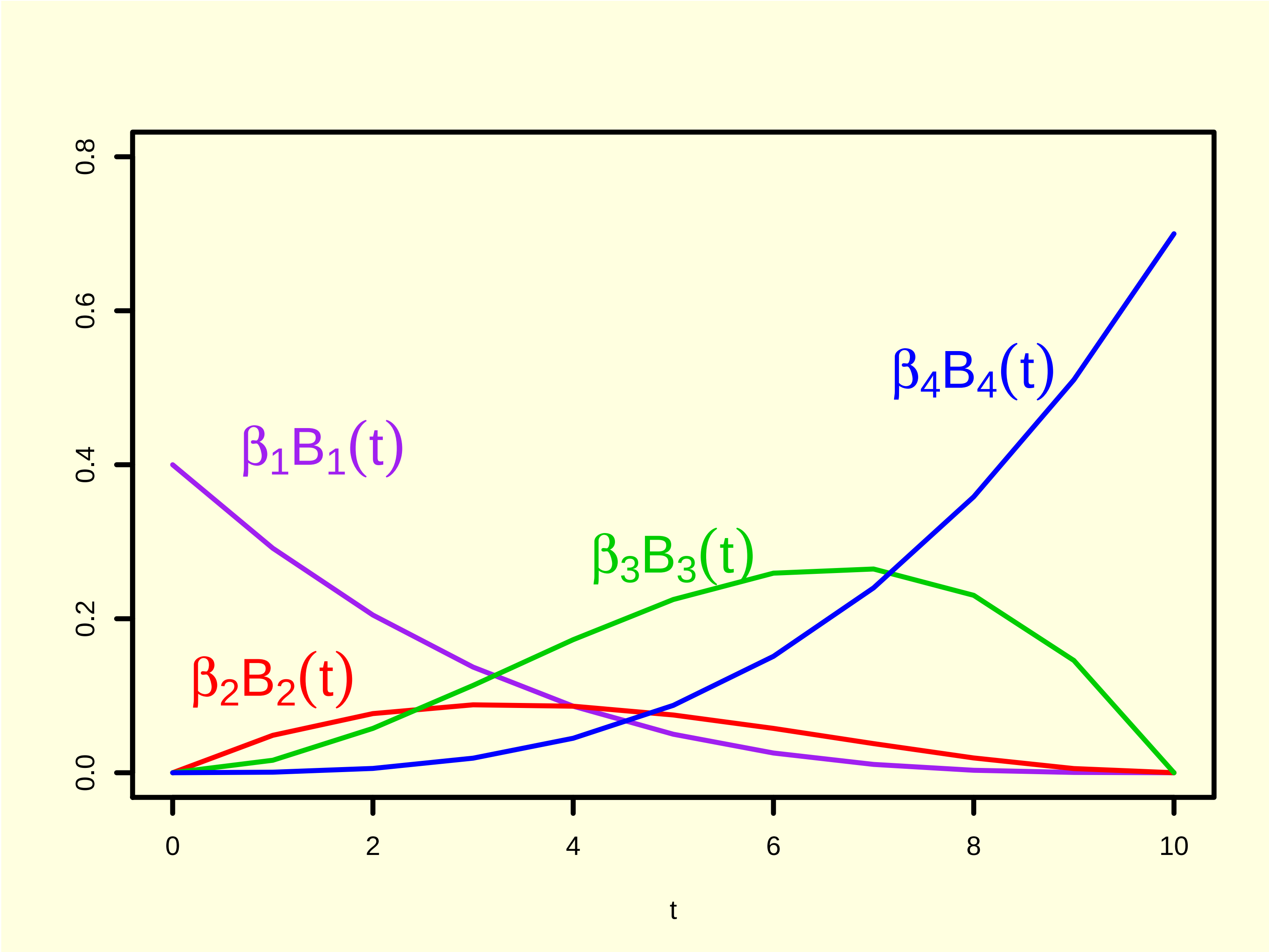
A cubic B-spline function with k knots is given by

$$f(x) = \sum_{i=1}^{k+4} \beta_k B_k(x),$$

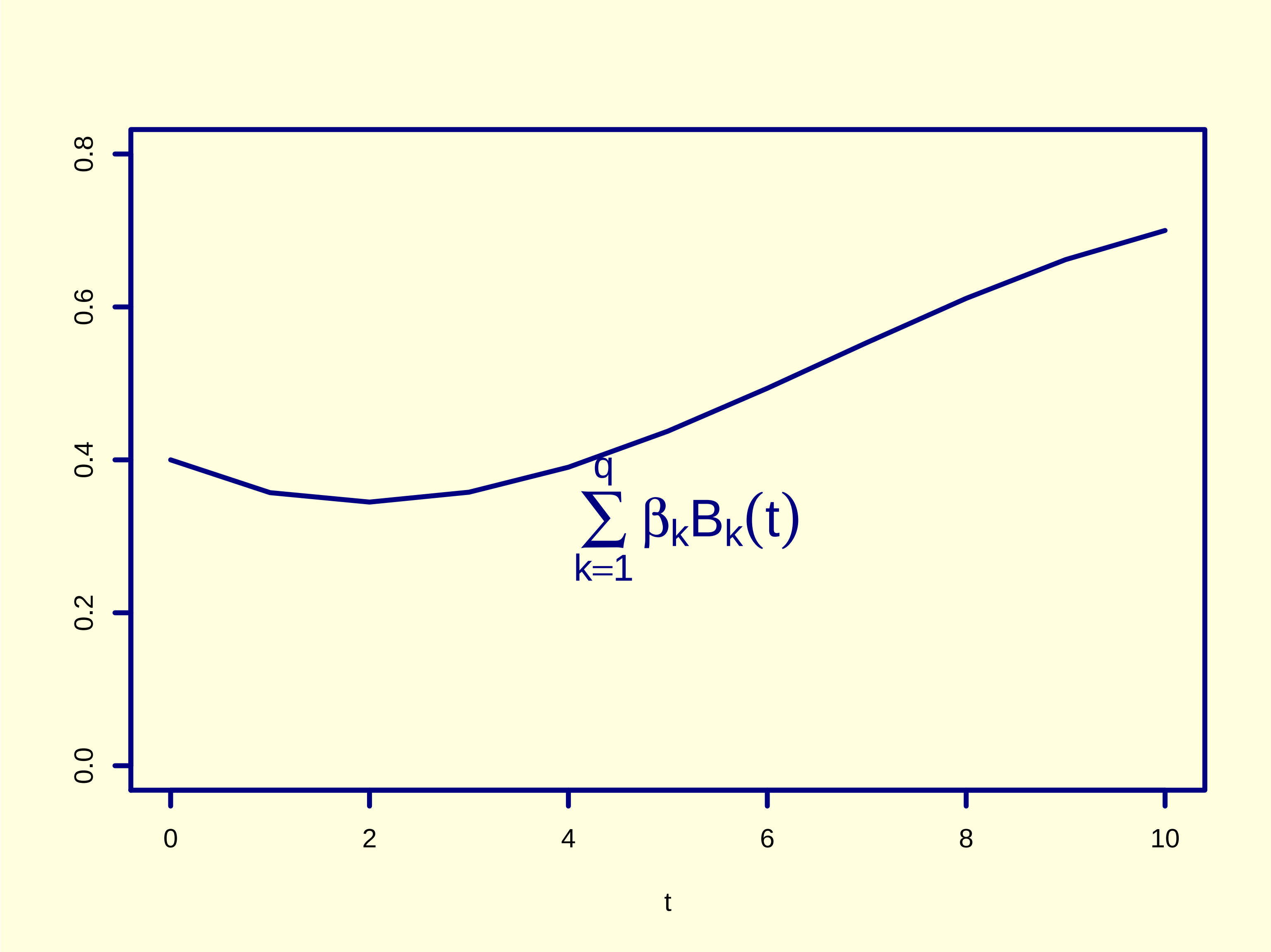
where $B_k(x)$ is the kth B-spline basis function.

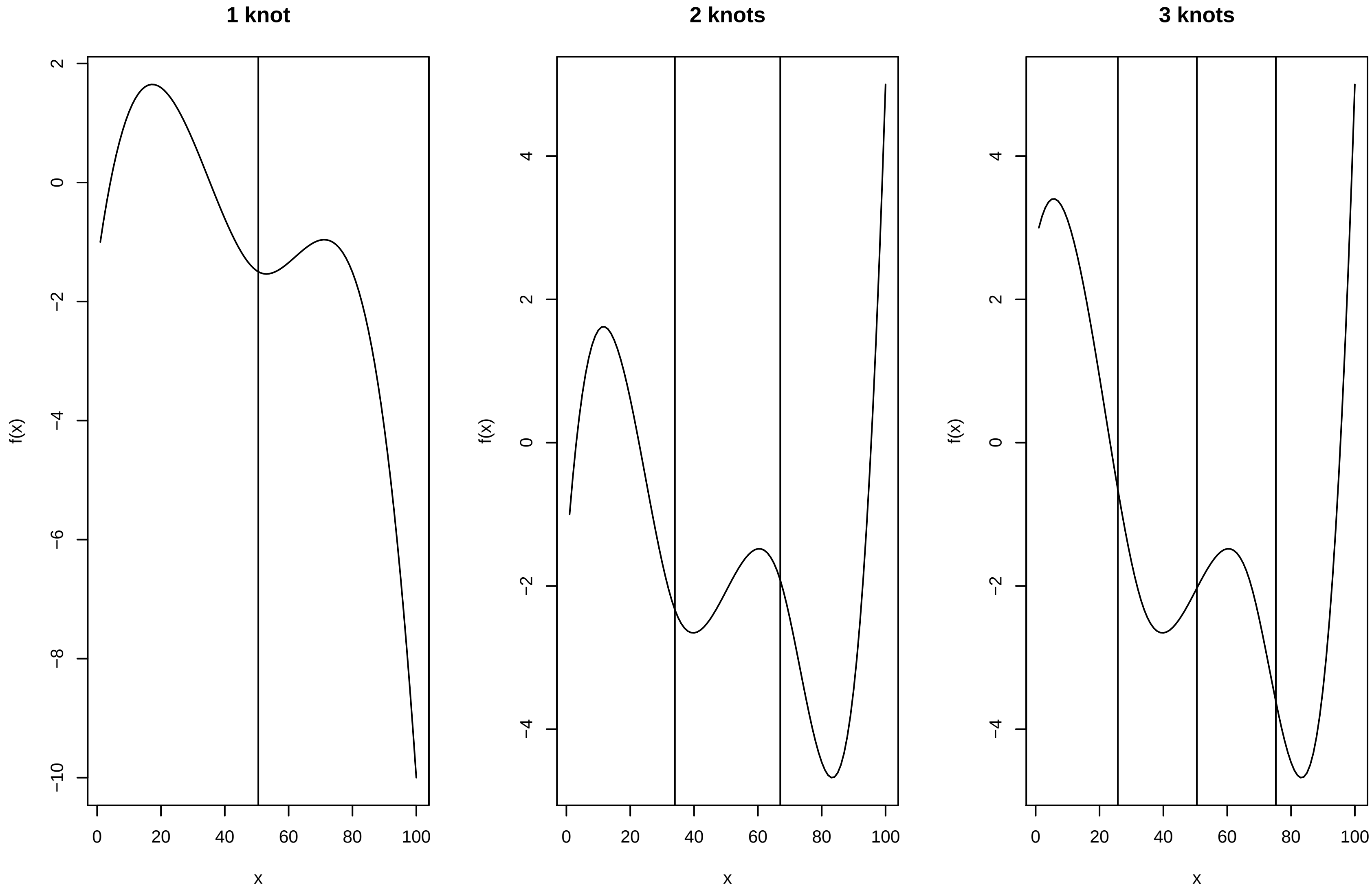


B1(t)
B2(t)
B3(t)
B4(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
t



0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
t
β4B4(t)
β1B1(t)
β3B3(t)
β2B2(t)

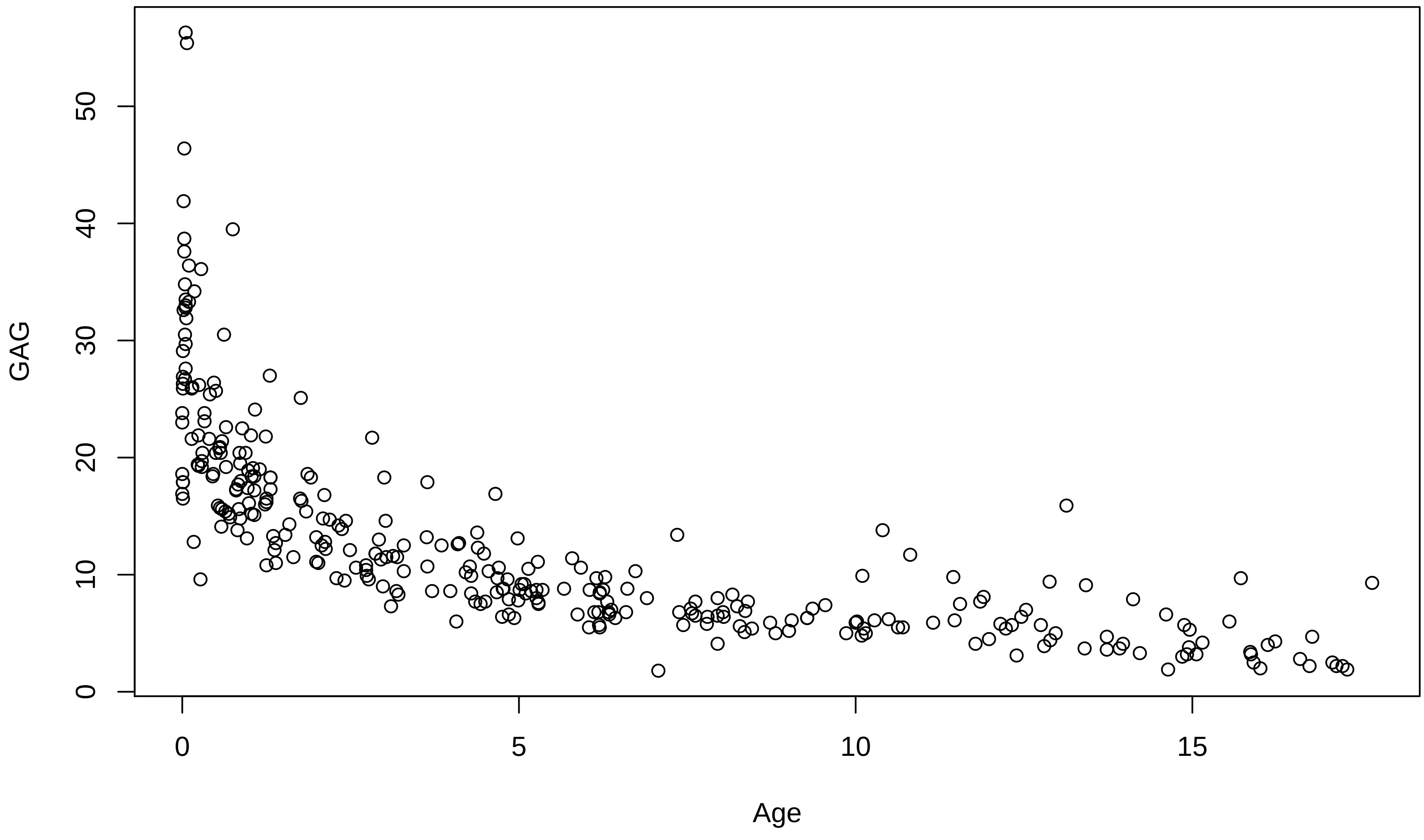

0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
t
$\sum_{k=1}^{q} \beta_k B_k(t)$

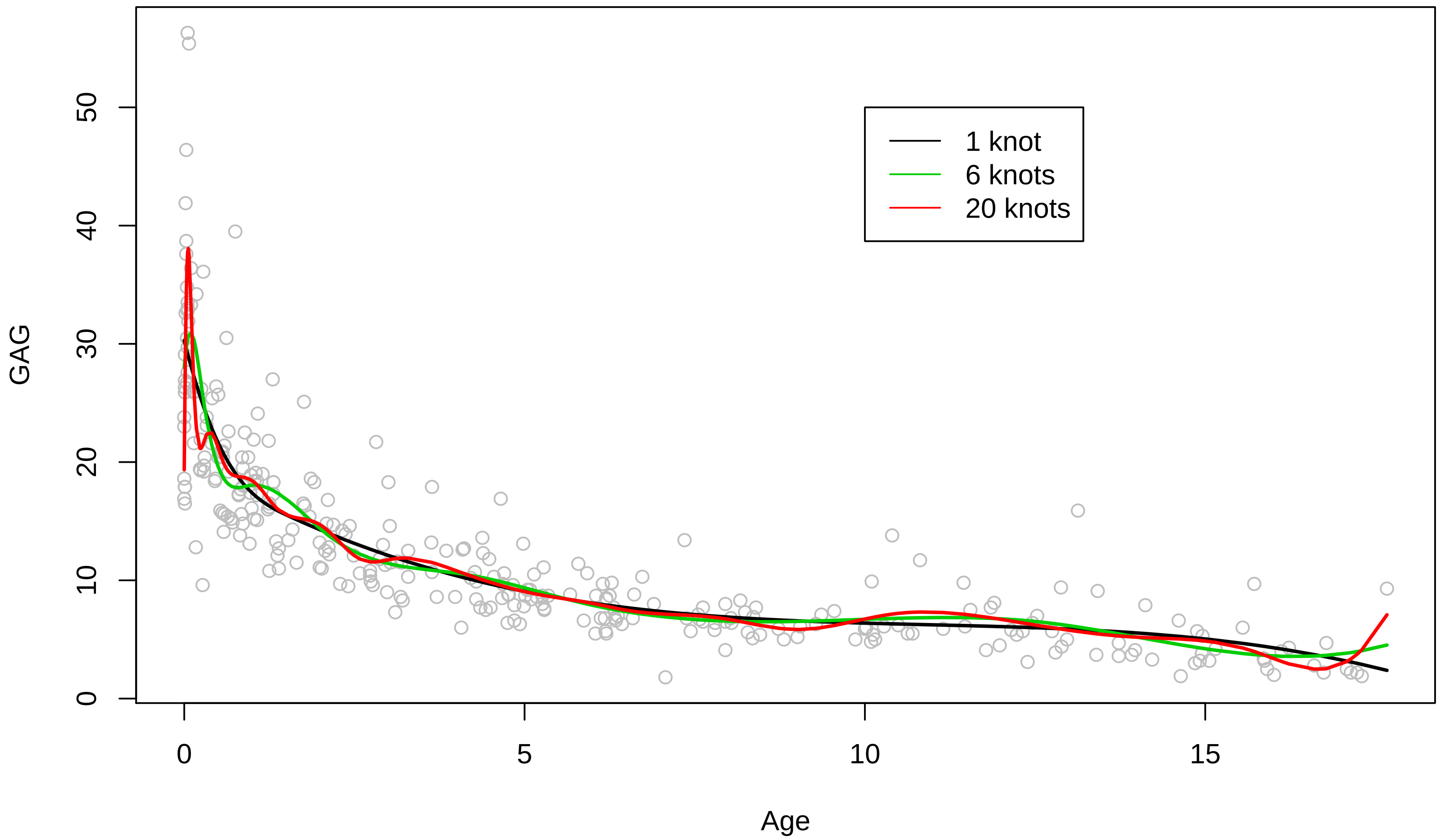

1 knot
2 knots
3 knots
f(x)
x
0
20
40
60
80
100
2
0
−2
−4
−6
−8
−10
4

Example

Venables and Ripley provide a data set, GAGurine, in the MASS library. It is described as follows:

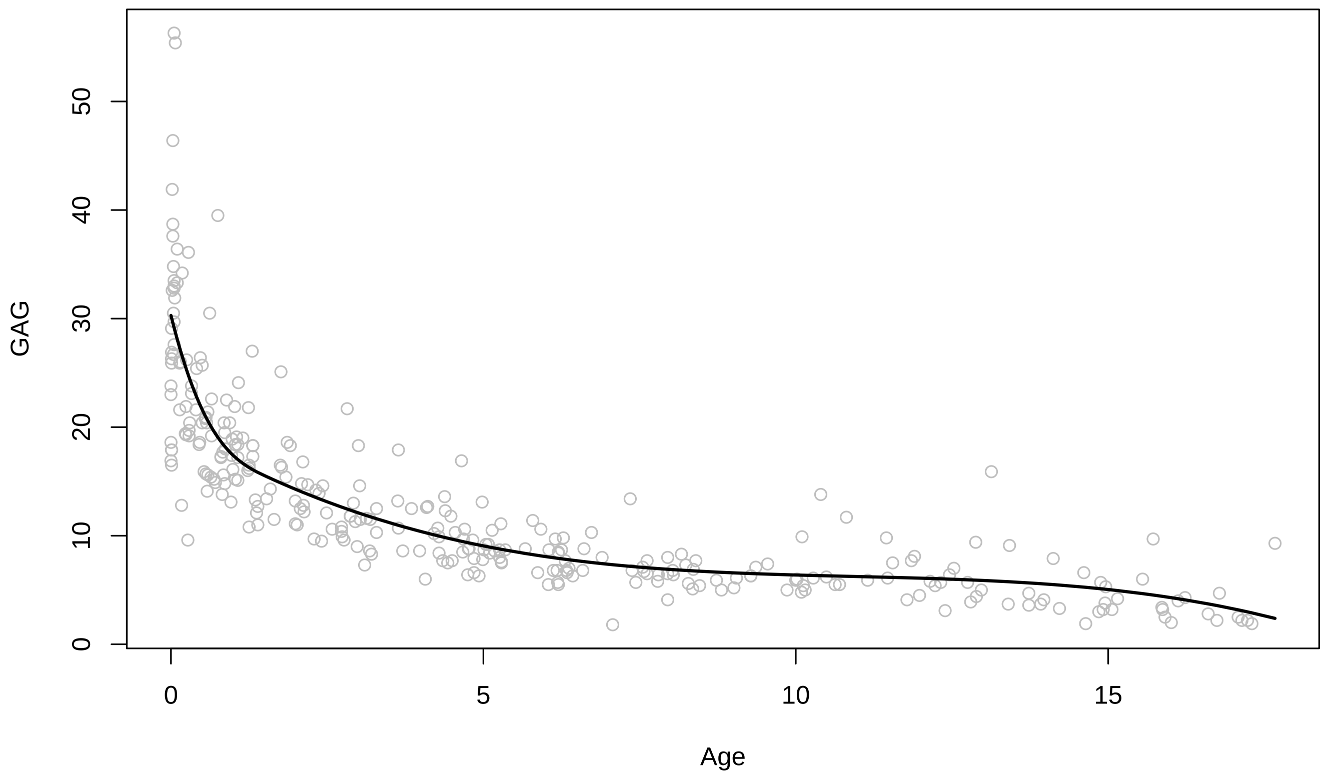
Data were collected on the concentration of a chemical GAG in the urine of 314 children aged from zero to seventeen years. The aim of the study was to produce a chart to help a paediatrican to assess if a child's GAG concentration is "normal".


GAG
Age
0
10
20
30
40
50
0
5
10
15



1 knot
6 knots
20 knots
GAG
Age
0
10
20
30
40
50
0
5
10
15

```
lmbs1=lm(GAG~bs(Age,df=5),data=GAGurine)
plot(GAGurine$Age,GAGurine$GAG,col="gray",ylab="GAG", xlab="Age")
lines(GAGurine$Age,fitted(lmbs1),lwd=2)
```



Choosing the number and position of knots

- Knots are usually placed at quantiles of the data or at regularly spaced intervals.
- Choosing the number, rather than the placement, seems to be more crucial to the fit.
- Therefore choose a number of knots that represents the curvature you believe to be present in the data. This comes with experience.
- You may also want to place knots at points in the data where you expect significant changes in the relationship between the predictor and the outcome to occur.