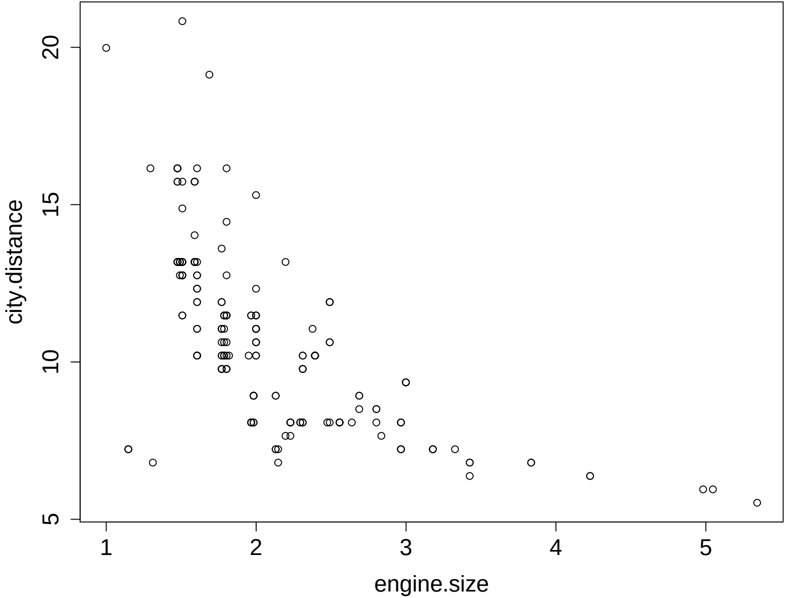
Moving Beyond Linearity

Aldo Solari

`auto`

- Reading: AS 4.1 - 4.2.4, 4.4 - 4.4.2
- $n = 203$ models of cars in circulation in 1985 in the United States but produced elsewhere
- We want identify a relationship allowing the prediction of `city.distance`, the distance covered per unit of fuel (km/L), as a function of `engine.size`, the car's engine size



city.distance
5
10
15
20
1
2
3
4
5
engine.size

Nonparametric estimation

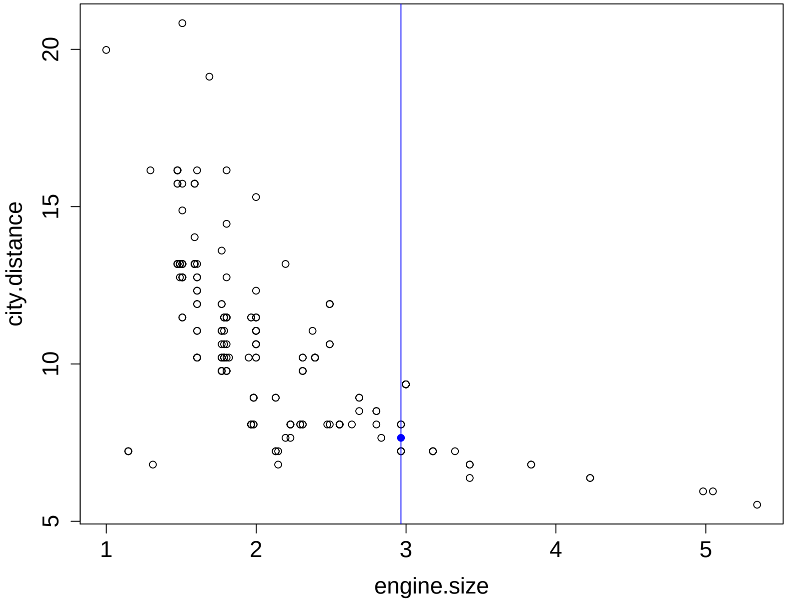
- Let's try to leave data 'speak for themselves' in a free way
- No reference to any parametric formulation for $f(x)$ (e.g. polynomials)
- The nonparametric approach to regression turns out to be particularly effective, especially when there is a considerable amount of data

Averaging

- To predict Y at $x = x_0$, gather all the training points (x_i, y_i) having $x_i = x_0$, then
- to estimate $f(x_0) = \mathbb{E}(Y|X = x_0)$, use the mean of their y_i:

$$\hat{f}(x_0) = \mathrm{Average}\{y_i : x_i = x_0\} = \frac{1}{\sum_{i=1}^{n} I\{x_i = x_0\}} \sum_{i:x_i=x_0} y_i$$

- Problem: in the training data, there may be no observations having $x_i = x_0$



city.distance
20
15
10
5
1
2
3
4
5
engine.size

Outline

1. k-Nearest-Neighbour
2. Local Regression
3. Regression Splines

Nearest Neighbour Averaging

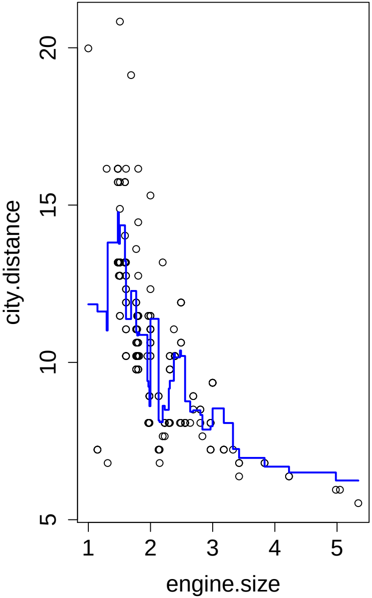
- Estimate $f(x_0) = \mathbb{E}(Y|X = x_0)$ by averaging those y_i whose x_i are in a neighbourhood of x_0
- e.g. define the neighbourhood $\mathcal{N}_k(x_0)$ to be the set of k observations having values x_i closest to x_0 in euclidean distance $\|x_i - x_0\|$

$$\hat{f}(x_0) = \mathrm{Average}\{y_i : x_i \in \mathcal{N}_k(x_0)\} = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x_0)} y_i$$

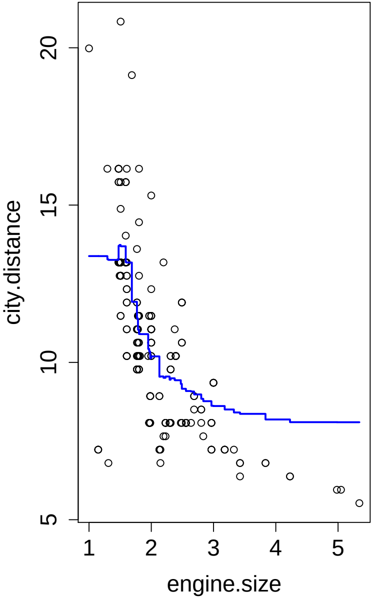
- This method is called k-nearest-neighbour regression

Choice of k

- By varying the number of neighbours k, we can achieve a wide range of flexibility in the estimated function $\hat{f}(x)$
- Small k corresponds to a more flexible fit: the k points are closer to target x (low bias), but averages based on a small sample have high variance
- Large k corresponds to a less flexible fit: it includes points far from x (high bias), but have smaller variance
- Best value for k depends on how smooth the true function $f(x)$ is, and how noisy y is
- Can try different values of k and use cross-validation


city.distance
engine.size
5
10
15
20
1
2
3
4
5

$k = 10$


city.distance
engine.size
5
10
15
20
1
2
3
4
5

$k = 60$

Extensions of the linear model

- Polynomial regression
- Local regression
- Step functions
- Regression splines
- ...

Outline

1. *k*-Nearest-Neighbour
2. **Local Regression**
3. Regression Splines

Local linear regression

- If $f(x)$ is a derivable function in x_0 then it is locally approximated by a line passing through $(x_0, f(x_0))$, i.e.,

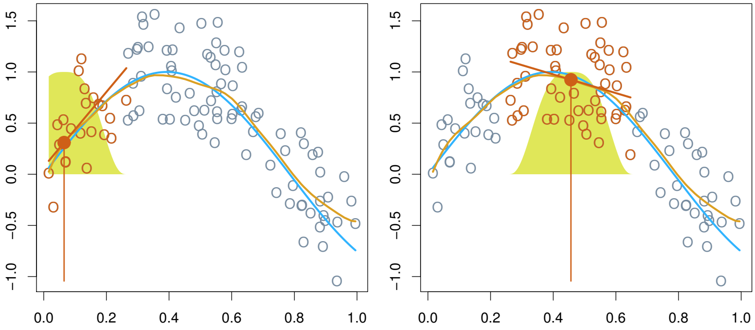
$$f(x) = \underbrace{f(x_0)}_{\alpha} + \underbrace{f'(x_0)}_{\beta}(x - x_0) + \text{error}$$

- We introduce the weighted least squares by weighting observations x_i with their distance from x_0:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left\{ y_i - \alpha - \beta(x_i - x_0) \right\}^2 w_h(x_i - x_0)$$

- h ($h > 0$) is a scale factor, called bandwidth or smoothing parameter, and
- $w_h(\cdot)$ is a symmetric density function around 0, said kernel

Local Regression



The fit $\hat{f}(x_0)$ at x_0 is obtained by fitting a weighted linear regression (orange line segment), and using the fitted value at x_0 (orange solid dot) as the estimate $\hat{f}(x_0)$

Source: ISL p. 281

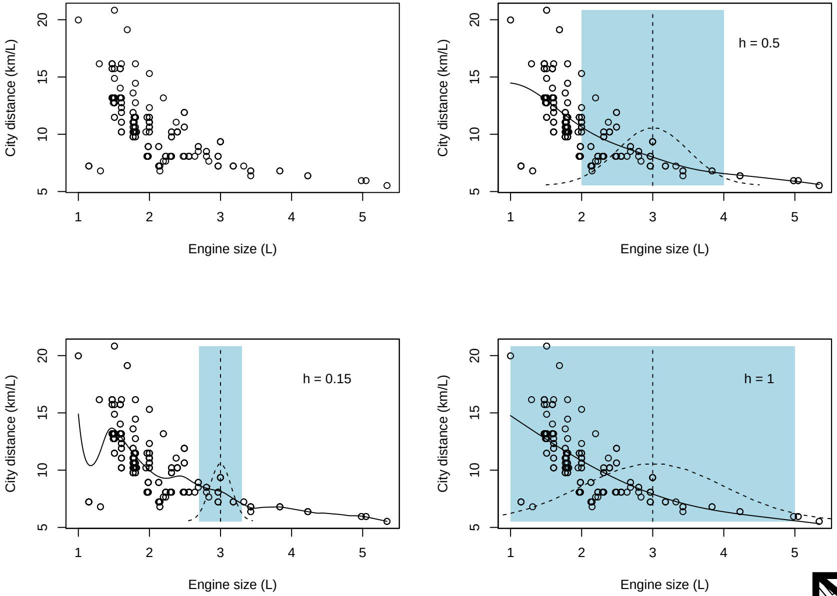
Local linear regression

- By varying x_0, we obtain a whole estimated curve $\hat{f}(x)$
- We can show that the estimate relative to a general point x can be obtained from the explicit formula

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\{a_2(x;h) - a_1(x;h)(x_i - x)\}\, w(x_i - x; h)}{a_2(x;h)\, a_0(x;h) - a_1(x;h)^2}\, y_i,$$

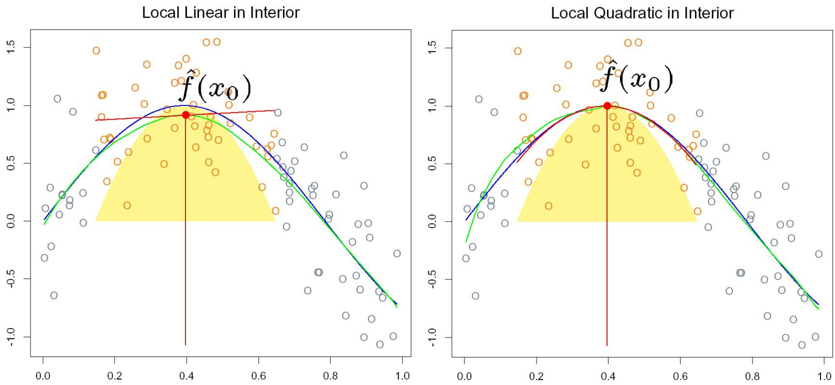
where $a_r(x;h) = \{\sum (x_i - x)^r w(x_i - x; h)\}/n$, for $r = 0, 1, 2$

- The most important component is h, which regulates the smoothness of the curve, while the choice of w is less relevant.
- We could think to w as the density of the normal distribution $N(0, h^2)$



Source: AS p. 71

Local quadratic regression



Source: ELS p. 197

Choice of kernel

- is not critical, as many studies on the subject have shown
- Let $w(t; h) = \frac{1}{h} w_0 \left(\frac{t}{h} \right)$
- The density $N(0, 1)$ is a common choice for w_0, i.e., we choose $N(0, h^2)$ for $w(t; h)$
- Many other choices are possible, in particular those with limited support as e.g. the tricubic or biquadratic ones that is

$$w_0(t) = \begin{cases} (1 - t^2)^2 & \text{if } |t| < 1, \\ 0 & \text{otherwise,} \end{cases} \qquad w_0(t) = \begin{cases} (1 - |t|^3)^3 & \text{if } |t| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

- the limited support reduces the computational burden, thanks to the many null terms

Some common choices for kernels

kernel	$w(z)$	support
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$	$\mathbb{R}$
Rectangular	$\frac{1}{2}$	$(-1, 1)$
Epanechnikov	$\frac{3}{4}(1 - z^2)$	$(-1, 1)$
biquadratic	$\frac{15}{16}(1 - z^2)^2$	$(-1, 1)$
tricubic	$\frac{70}{81}(1 - \lvert z\rvert^3)^3$	$(-1, 1)$

Choice of h

- A critical aspect is the choice of the smoothing parameter h, since we can prove that, for n sufficiently large,

$$\mathbb{E}(\hat{f}(x)) \approx f(x) + \frac{h^2}{2}\sigma_w^2 f''(x), \quad \mathbb{V}\mathrm{ar}(\hat{f}(x)) \approx \frac{\sigma^2}{nh}\frac{\alpha(w)}{g(x)},$$

  where $\sigma_w^2 = \int z^2 w(z)dz$, $\alpha(w) = \int w(z)^2 dz$ and $g(x)$ indicates the density from which the x_i where sampled;
- The bias is $\propto h^2$ and the variance is $\propto 1/(nh)$
- Therefore, although we would like to choose $h \to 0$ to bring down the bias, this makes the variance of the estimate diverge. For $h \to \infty$, the opposite occurs
- Once again, in choosing h, we have a trade-off between bias and variance

loess

- in many cases, there is an advantage in using a nonconstant bandwidth along the x-axis, according it to the level of sparseness of observed points
- variable bandwidth: it is reasonable to use larger values of h when x_i are more scattered
- Good idea! ...but how do we modify h?
- `loess`: express the smoothing parameter (`span`) defining the fraction of effective observations for estimating $f(x)$ at a certain point x_0 on the x-axis;
- this fraction is kept constant
- this imply automatically a setting of the bandwidth related to the sparsity of data

Variability bands

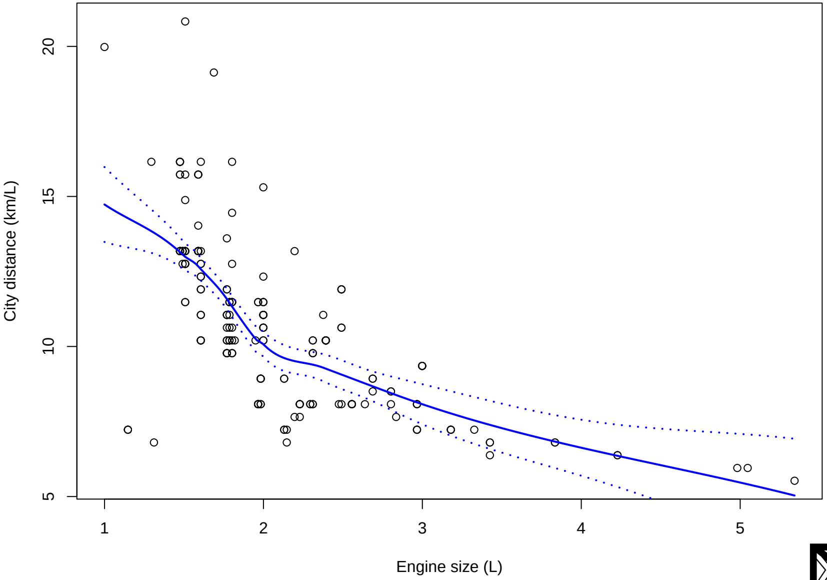
- A pivotal quantity, approximately, is

$$\frac{\hat{f}(x) - f(x) - b(x)}{\sqrt{\mathbb{V}\mathrm{ar}(\hat{f}(x))}} \sim N(0,1)$$

  where $b(x)$ indicates the bias, which cannot be neglected
- Instead of looking for complicated corrections for the unknown $b(x)$, a current solution is to construct variability bands of the type

$$\left(\hat{f}(x) - z_{\alpha/2}\,\mathrm{std.err}(\hat{f}(x)),\ \hat{f}(x) + z_{\alpha/2}\,\mathrm{std.err}(\hat{f}(x))\right)$$

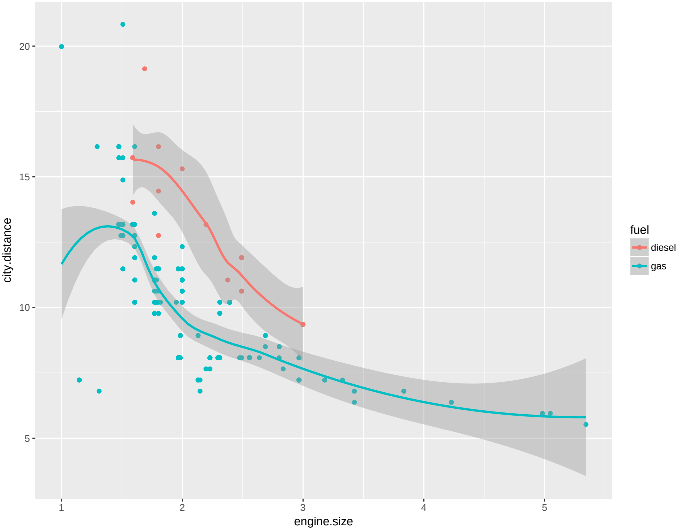
  providing an indication of the local variability of the estimate
- Variability bands do not have the coverage guarantee of confidence intervals



City distance (km/L)
5
10
15
20
1
2
3
4
5
Engine size (L)

Data visualization

Reading: GW 3.1 - 3.6

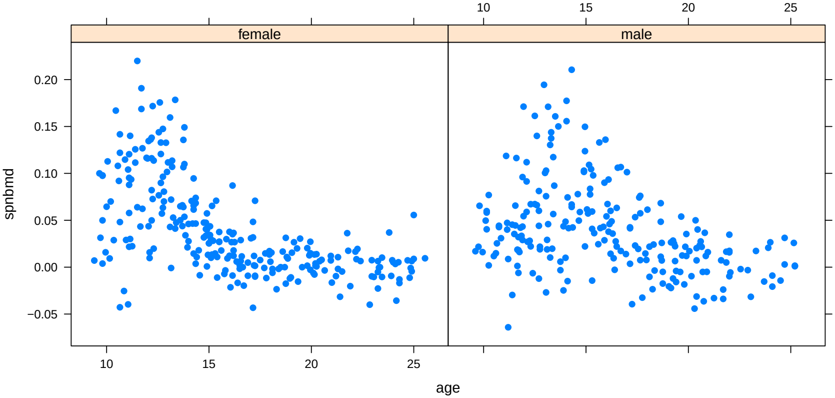


Outline

1. *k*-Nearest-Neighbour
2. Local Regression
3. **Regression Splines**

bmd data

Reading: ISL 7.1-7.4, 7.6, 7.8.1, 7.8.2



Basis functions

- Consider
$$Y = f(X) + \varepsilon$$
where
$$f(X) = \sum_{j=0}^{q} \beta_j b_j(X)$$
- $b_j(\cdot)$ are known functions called *basis functions*
- E.g. 3rd degree polynomial regression:

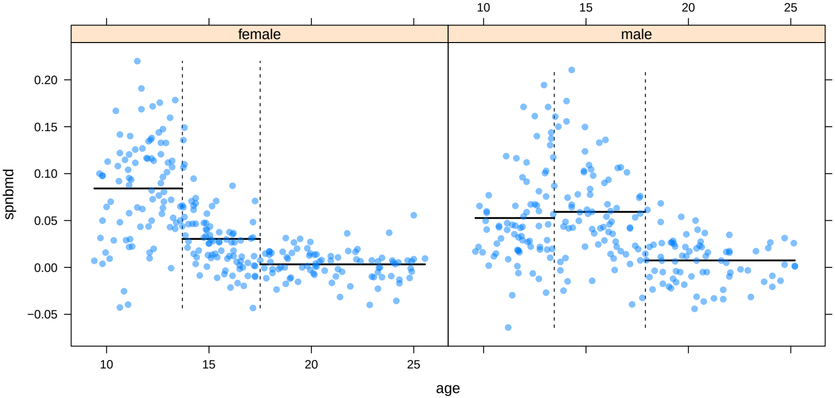
$$b_0(x) = 1, b_1(x) = x, b_2(x) = x^2, b_3(x) = x^3$$

Step functions

- Define cutpoints $\xi_1, \ldots, \xi_K$ in the range of X, called *knots*
- Basis functions: $K + 1$ *step functions*

$$
\begin{aligned}
b_0(X) &= I\{X < \xi_1\} \\
b_1(X) &= I\{\xi_1 \leq X < \xi_2\} \\
&\vdots \\
b_{K-1}(X) &= I\{\xi_{K-1} \leq X < \xi_K\} \\
b_K(X) &= I\{X \geq \xi_K\}
\end{aligned}
$$

Step functions

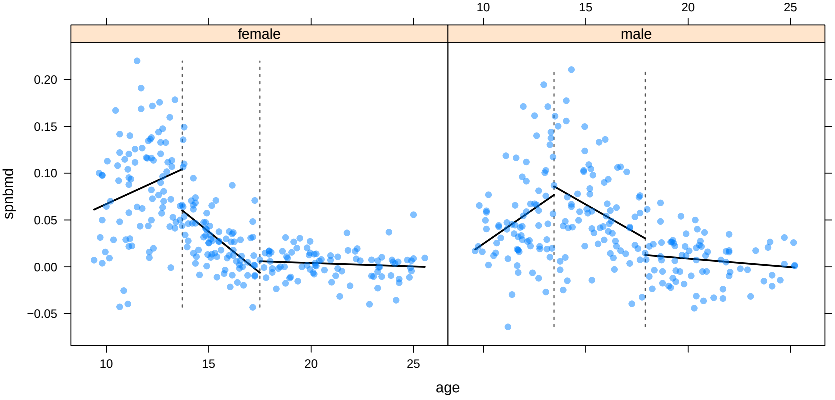


Piecewise linear regression

- Define K knots $\xi_1, \ldots, \xi_K$
- Basis functions: $2(K+1)$, fitting $K+1$ simple regressions for each partition of the data

$$
\begin{aligned}
b_0(X) &= I\{X < \xi_1\} \\
b_0'(X) &= X \cdot I\{X < \xi_1\} \\
b_1(X) &= I\{\xi_1 \leq X < \xi_2\} \\
b_1'(X) &= X \cdot I\{\xi_1 \leq X < \xi_2\} \\
&\vdots \\
b_K(X) &= I\{X \geq \xi_K\} \\
b_K'(X) &= X \cdot I\{X \geq \xi_K\}
\end{aligned}
$$

Piecewise linear regression



Continuos piecewise linear regression

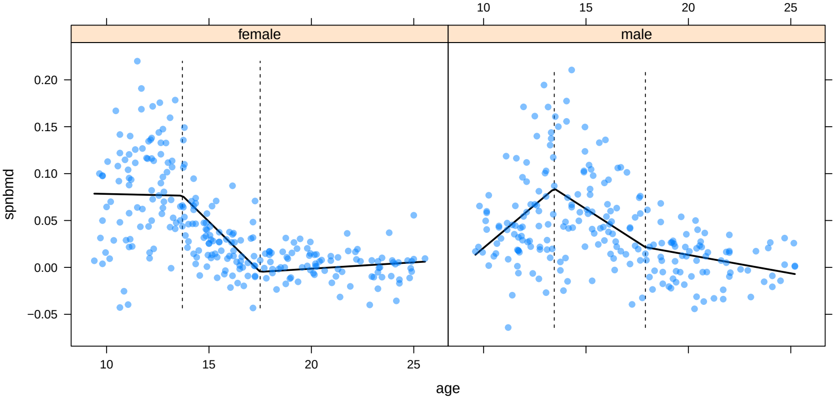
- Define K knots $\xi_1, \ldots, \xi_K$
- Basis functions: $K + 2$, giving a continuos piecewise linear model

$$
\begin{aligned}
b_0(X) &= 1 \\
b_1(X) &= X \\
b_2(X) &= (X - \xi_1)_+ \\
&\vdots \\
b_{K+1}(X) &= (X - \xi_{K-1})_+ \\
b_{K+2}(X) &= (X - \xi_K)_+
\end{aligned}
$$

where $(\cdot)_+$ defines the positive portion of its argument

$$
(a)_+ = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{if } a < 0 \end{cases}
$$

Continuos piecewise linear regression



Splines

- The preceding is an example of a *spline*: a piecewise polynomial with the constraint that the fitted curve must be continuous
- The set of basis functions introduced earlier is an example of what is called the truncated power basis

$$b_j(X) = X^j, \quad j = 0, \ldots, d$$

$$b_{d+k}(X) = (X - \xi_k)_+^d, \quad k = 1, \ldots, K$$

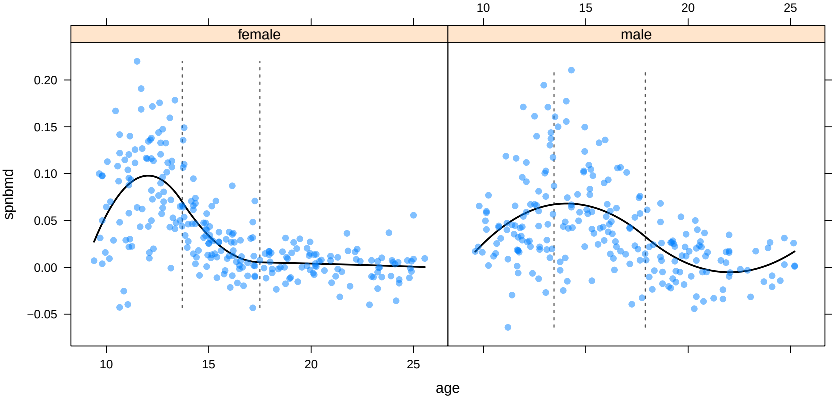
which gives $(d+1) + K$ basis functions

Quadratic Splines

- Define K knots $\xi_1, \ldots, \xi_K$
- Basis functions: $K + 3$

$$
\begin{aligned}
b_0(X) &= 1 \\
b_1(X) &= X \\
b_2(X) &= X^2 \\
b_3(X) &= (X - \xi_1)_+^2 \\
&\vdots \\
b_{K+2}(X) &= (X - \xi_{K-1})_+^2 \\
b_{K+3}(X) &= (X - \xi_K)_+^2
\end{aligned}
$$

Quadratic splines

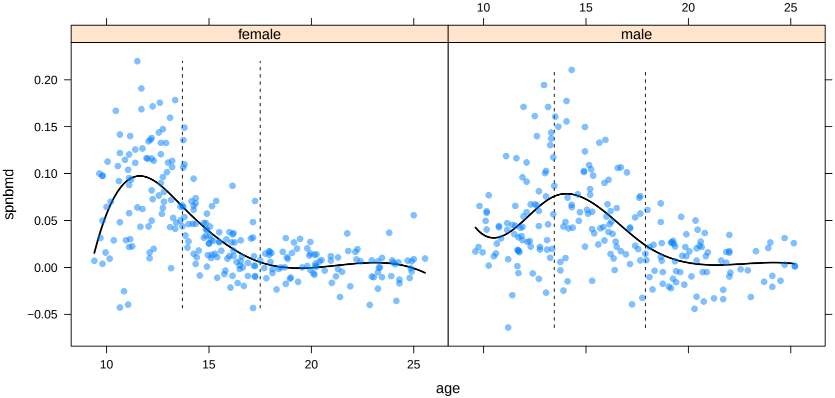


Quadratic Splines

- Define K knots $\xi_1, \ldots, \xi_K$
- Basis functions: $K + 3$

$$
\begin{aligned}
b_0(X) &= 1 \\
b_1(X) &= X \\
b_2(X) &= X^2 \\
b_3(X) &= X^3 \\
b_4(X) &= (X - \xi_1)^3_+ \\
&\vdots \\
b_{K+3}(X) &= (X - \xi_{K-1})^3_+ \\
b_{K+4}(X) &= (X - \xi_K)^3_+
\end{aligned}
$$

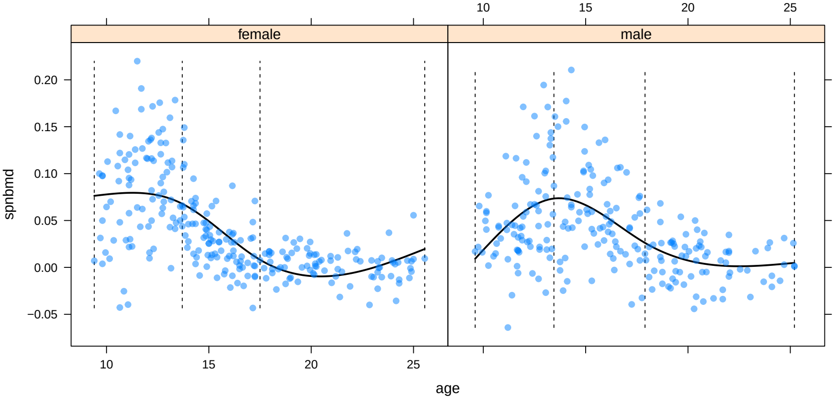
Cubic splines



Natural cubic splines

- Cubic splines to be erratic at the boundaries of the data
- Natural cubic splines ameliorate this problem by adding 4 constraints that the function is linear beyond the boundaries of the data
- A natural cubic spline with K knots has K basis functions

Natural cubic splines



Problems with knots

- Regression splines have one shortcoming: the placement of knots
- Choices regarding the number of knots and where they are located are not particularly easy to make
- Next: smoothing splines