07/02/2019

Advanced Regression SPH024

Lect. 4b: Non-linear modelling

www.menti.com 16 38 43

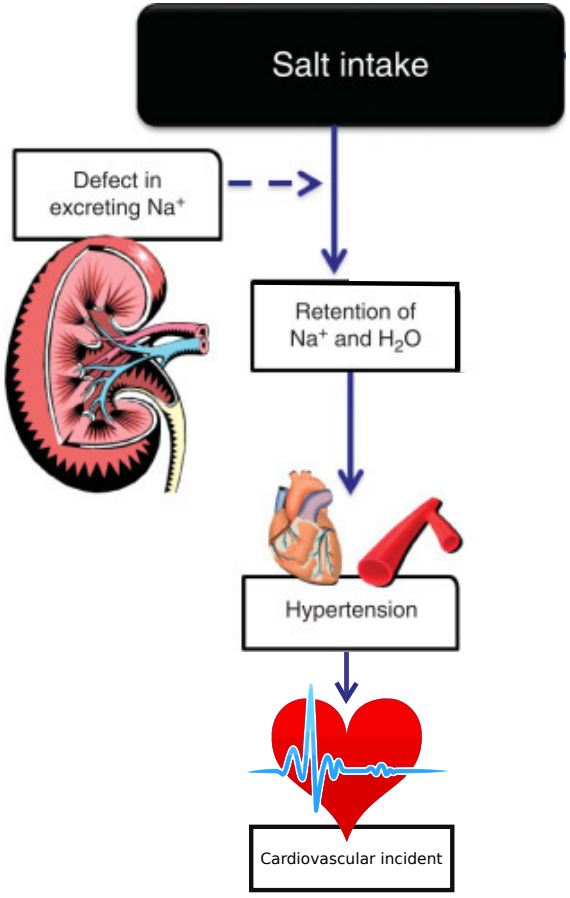
So far

1. Linear models, OLS, MLE...
2. Variable importance selection
3. High dim. analysis and regularisation
4. Mixed effects / hierarchical models

$\longrightarrow$ All GLMs: generalised linear models:

$$Y = \beta X + \gamma Z + \epsilon$$

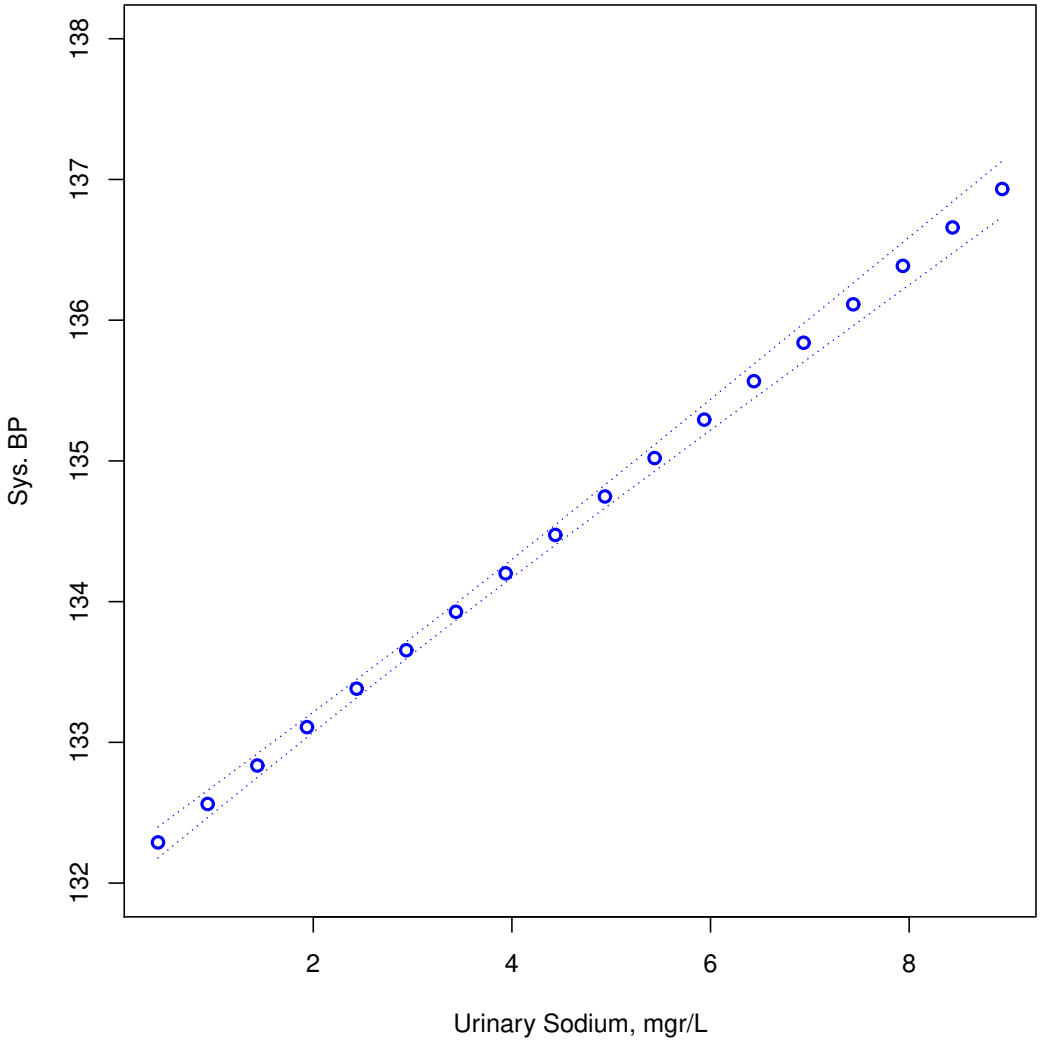
Beyond linear models



Case study: urinary sodium vs CVD in UKBiobank

- Outcomes: Systolic blood pressure
- Exposure: Na^+
- Confounders: Age, Sex, K, BMI

Beyond linear models



Beyond linear models

Linear framework

$$Y = \beta X + \epsilon$$

New framework

$$Y = f(X) + \epsilon$$

where $f : \mathbb{R}^p \to \mathbb{R}$**:**

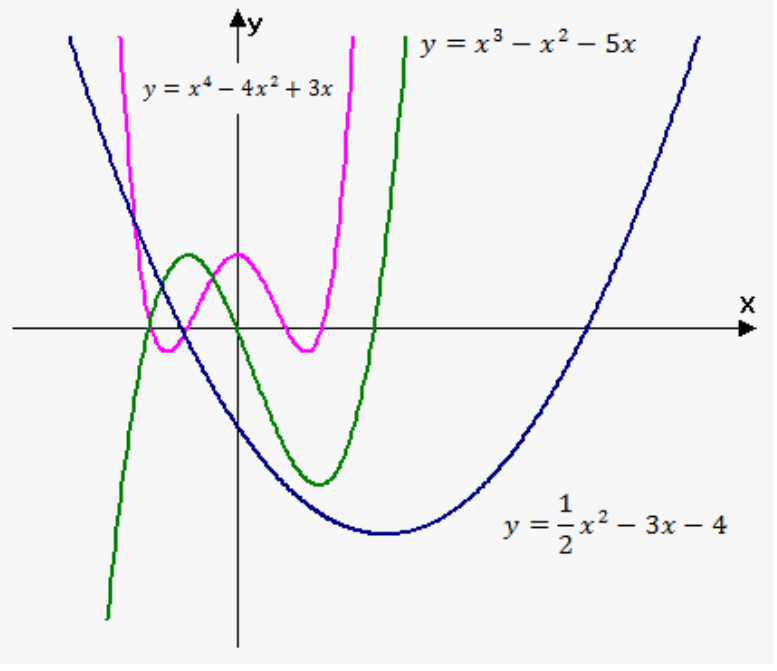
$$f(X) = f(X_1, X_2, ..., X_p) + \epsilon,$$
$$\hat{Y} = \hat{f}(X_1, X_2, ..., X_p)$$

Aim: Model & estimate f **+ perform inference on** f

Polynomial regression

For one predictor x_k, $k \leq p$:

$$
\begin{aligned}
Y_i &= \beta_0 + \sum_d^D \beta_d x_{k,i}^d + \epsilon \\
&= \beta_0 + \beta_1 x_{k,i} + \beta_2 x_{k,i}^2 \\
&+ \beta_3 x_{k,i}^3 + \ldots + \beta_D x_{k,i}^D + \epsilon
\end{aligned}
$$



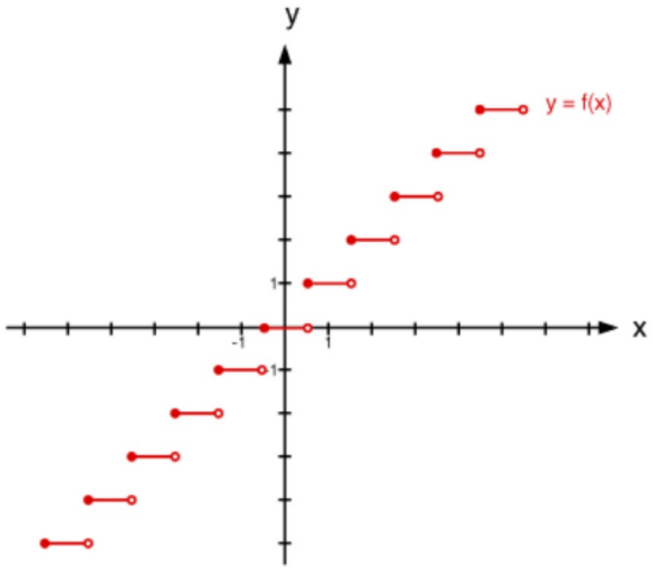
Piecewise regression

Cut support of X_k into "nice / simple" M pieces at $c_1, c_2, ..., c_{M-1}$

$$Y = f(X_k) + \epsilon,$$

$$f(X_k) = \begin{cases} f_1(X_k), & X_k \leq c_1 \\ f_2(X_k), & c_1 < X_k \leq c_2 \\ \vdots & \vdots \\ f_M(X_k), & c_{M-1} < X_k \end{cases}$$

$$= \sum_m f_m(X_k)\ \mathrm{I}\left(X_k \in (c_{m-1}, c_m]\right)$$



Where $f_m(.)$ can be constant or linear

Polynom. and Piecewise regression

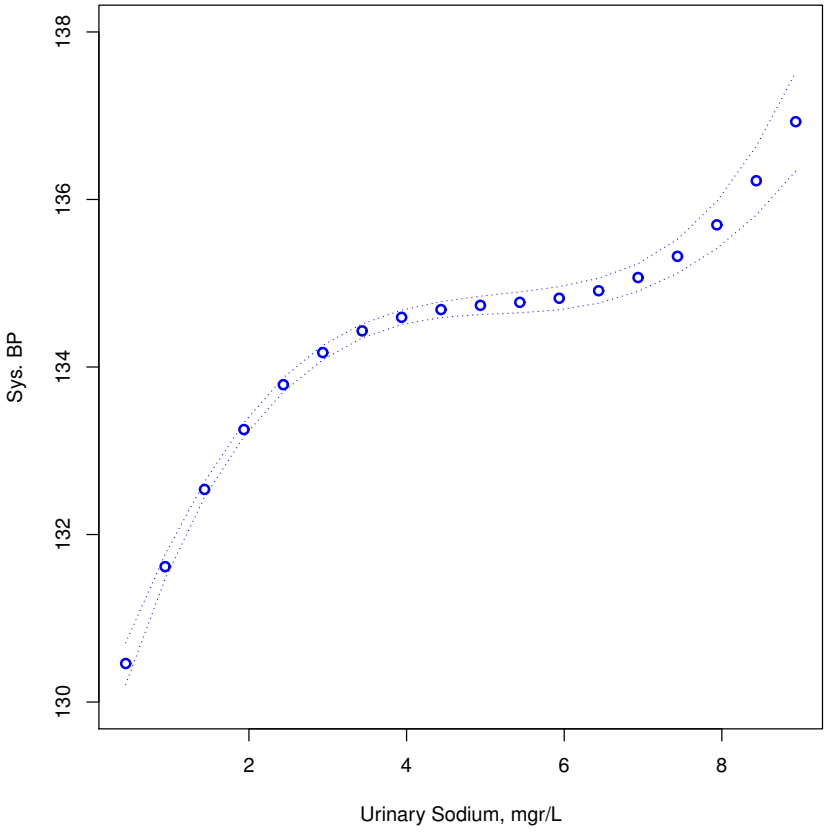


Figure 1

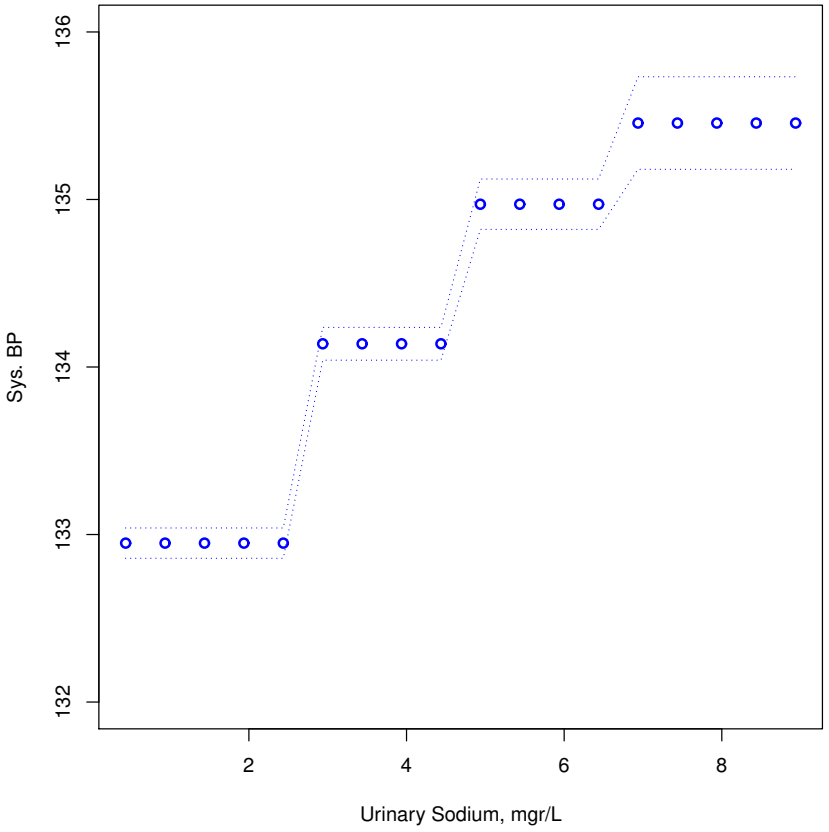


Figure 2

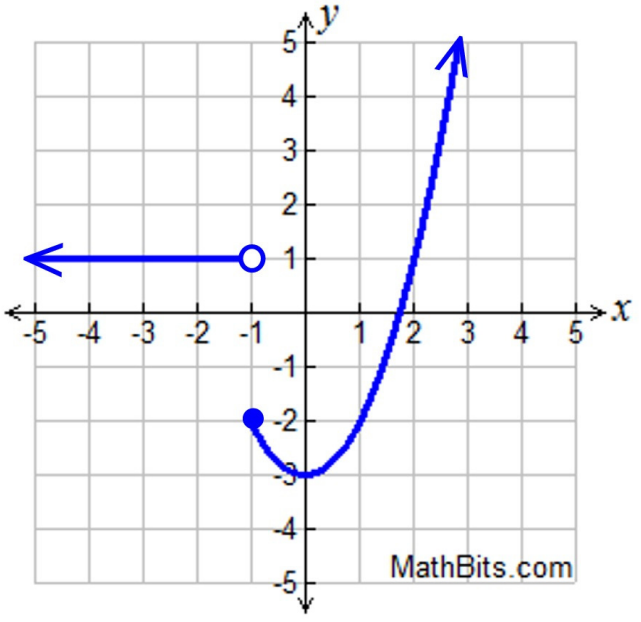
Piecewise polynom. regression

Cut support of X_k into M pieces at $c_1, c_2, ..., c_{M-1}$

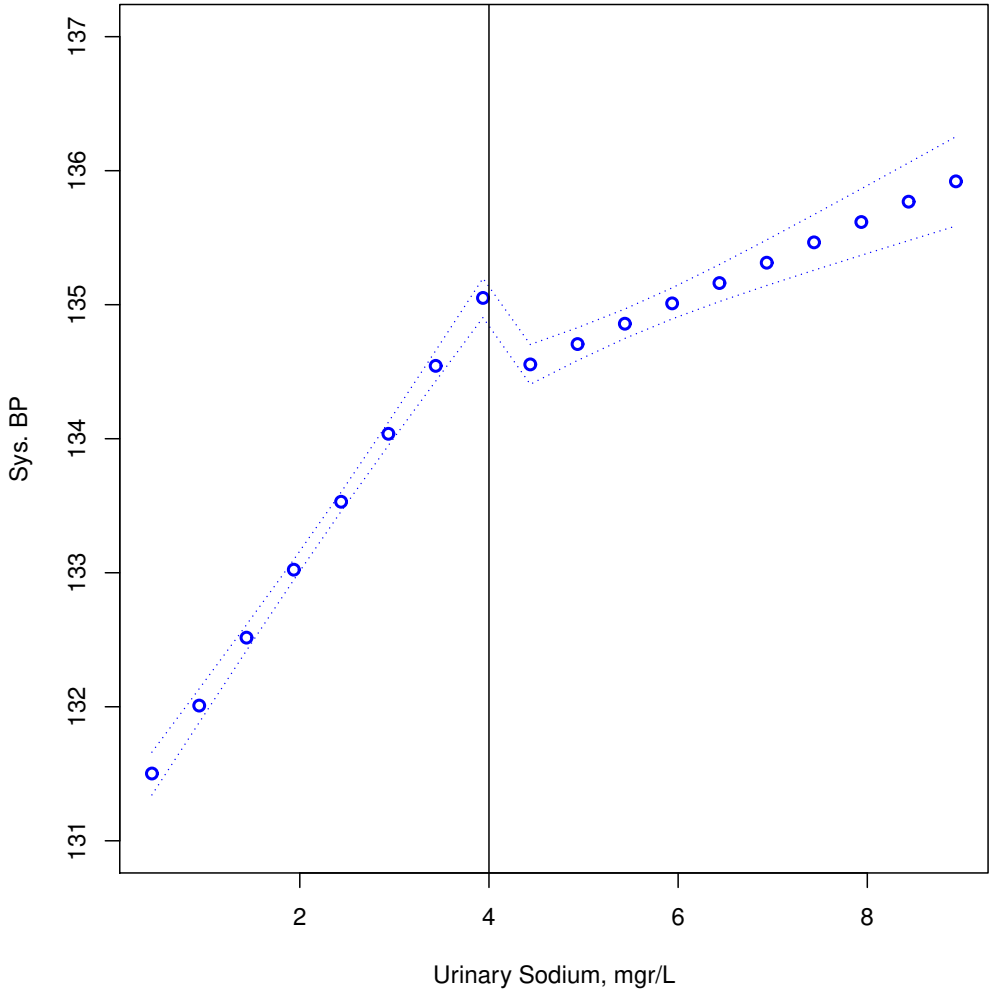
$$Y = f(X_k) + \epsilon,$$

$$f(X_k) = \sum_m f_m(X_k)\, \mathrm{I}(X_k \in (c_{m-1}, c_m])$$

Where $f_m(.)$ is a polynomial of degree $d \geq 2$



Piecewise polynom. regression



Splines regression

Two important conditions:

1. Continuous: no discontinuity in $f(X_k)$
2. Smoohtness: 1st and 2nd derivative defined and finite

Splines:

$$Y_i = \beta_0 \sum_{m}^{M} \beta_m b_m(X_{k,i}) + \epsilon$$

Where $b_m(.)$: basis functions

Splines regression

Cubic splines with K knots: $\{c_1, c_2, \ldots c_K\}$

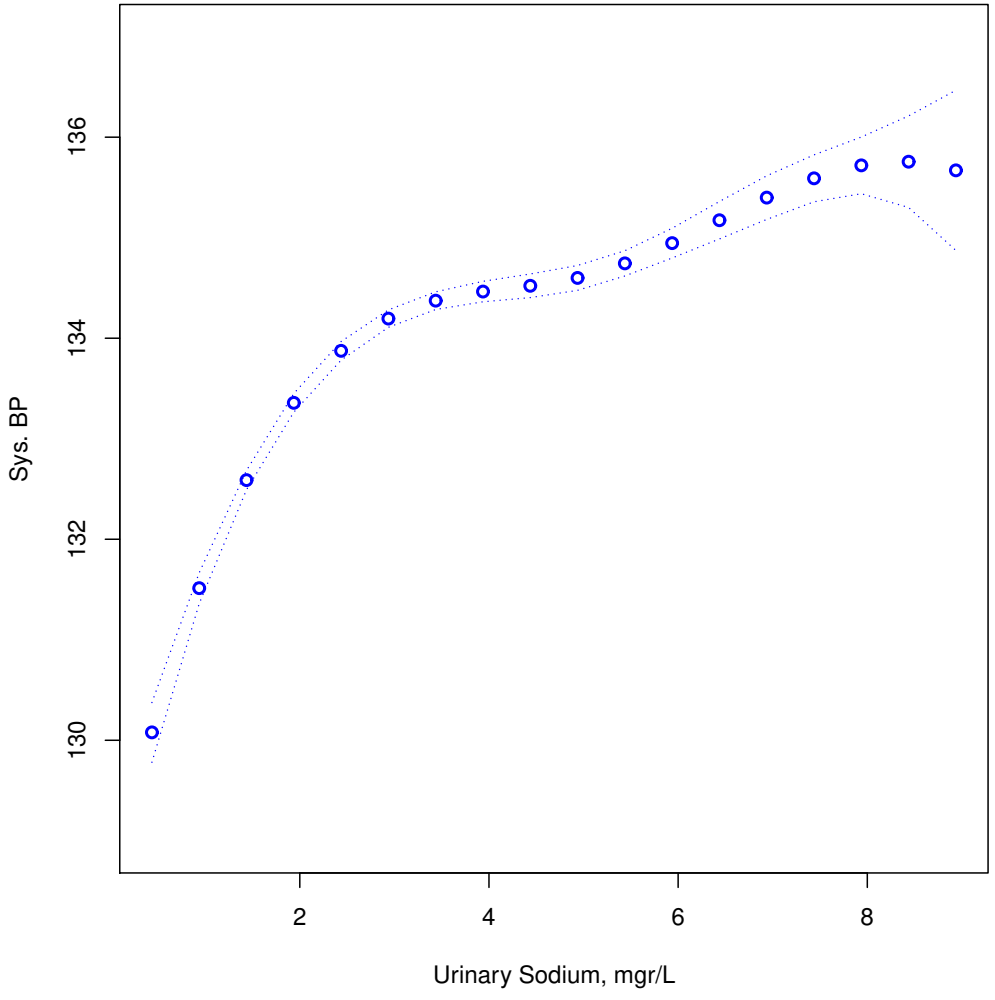
$$Y_i = \beta_0 \sum_{m}^{K+3} \beta_m b_m(X_{k,i}) + \epsilon,$$

$$b_m(X) = \begin{cases} X^m, & m \in [1,3] \\ (X - c_{m-3})_+^3, & m \in [4,K] \end{cases},$$

$$(X - c_{m-3})_+^3 = \begin{cases} (X - c_{m-3})^3 & X \geq c_{m-3} \\ 0 & \text{otherwise} \end{cases}$$

Qus: how many degrees of freedom are there in a cubic spline with K knots?

Cubic Splines



Splines regression

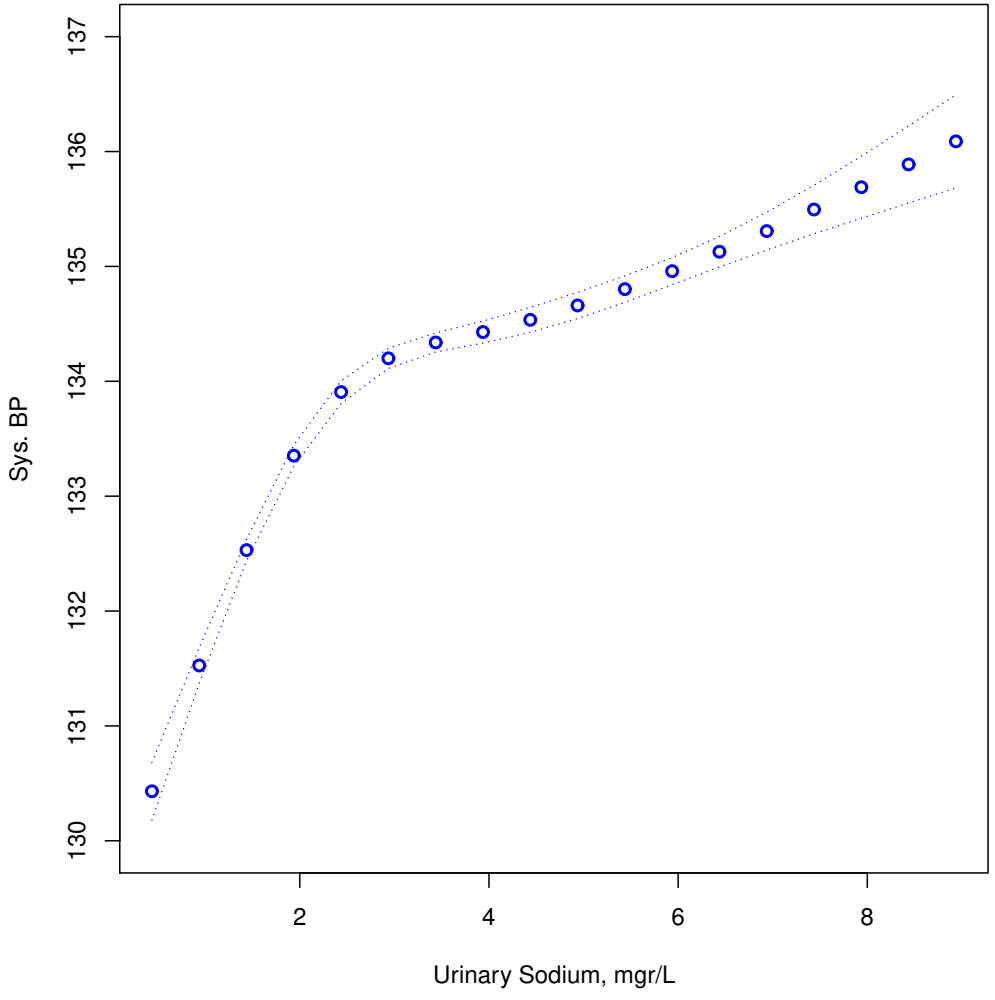
Problem: high variance in outer range of data
Solution: natural splines

1. Continuous: no discontinuity in $f(X_k)$
2. Smoohtness: 1st and 2nd derivative defined and finite
3. **Boundaries: linear functions below smallest / above largest knots**

$$Y_i = \beta_0 + g(X_{k,i}) + \epsilon,$$

$$g(X_{k,i}) = \begin{cases} \sum_m^{K+3} \beta_m b_m(X_{k,i}), & X_k \in [c_1, c_K] \\ \alpha + \gamma X_k, & X_k < c_1,\ X_k > c_K \end{cases}$$

Natural Splines



Splines regression

Problem: Poor data fit and/or overfitting

Solution: Smoothing splines

$$Y_i = \beta_0 + g(X_{k,i}) + \epsilon,\ g(.):\ \text{natural cubic splines}$$

$$g^*(x) = \arg\min_g \sum_i^n (Y_i - g(X_{k,i}))^2 \text{ s.t. } ||g''(X_k)||_2 << +\infty$$

Smoothness: "well-behaved" 1st and second derivatives

$$g^*(x) = \arg\min_g \underbrace{\sum_i^n (Y_i - g(X_{k,i}))^2}_{Loss} + \underbrace{\lambda \int g''(t)^2 dt}_{Constraint}$$

Splines regression

Problem: Poor data fit and/or overfitting

Solution: Smoothing splines

$$Y_i = \beta_0 + g(X_{k,i}) + \epsilon,\ g(.):\ \text{natural cubic splines}$$

$$g^*(x) = \arg\min_g \sum_i^n (Y_i - g(X_{k,i}))^2 \text{ s.t. } ||g''(X_k)||_2 << +\infty$$

Smoothness: "well-behaved" 1st and second derivatives

$$g^*(x) = \arg\min_g \underbrace{\sum_i^n (Y_i - g(X_{k,i}))^2}_{Loss} + \underbrace{\lambda \int g''(t)^2 dt}_{Constraint}$$

Qus: What are we actually computing with $g^*(.)$?

Smoothing splines

$$Y_i = \beta_0 + g(X_{k,i}) + \epsilon,\ g(.):\ \text{natural cubic splines}$$

$$g^*(x) = \arg\min_g \sum_i^n (Y_i - g(X_{k,i}))^2 + \lambda \int g''(t)^2 dt$$

- Knots: all points $\{X_{k,1}, X_{k,2}, ..., X_{k,n}\}$!
  - Fit the entire data set subject to smoothness constraint
- Dependent on parameter λ

Splines regression - knots

Key parameters:

- Cubic / natural splines: K (# of knots)
- Smoothing splines: λ (smoothness tuning)

How do we select K and λ?

Splines regression - knots

Key parameters:

- Cubic / natural splines: K (# of knots)
- Smoothing splines: λ (smoothness tuning)

How do we select K and λ?

Answer: MSE from LOOCV - Leave-One-Out Cross-Validation

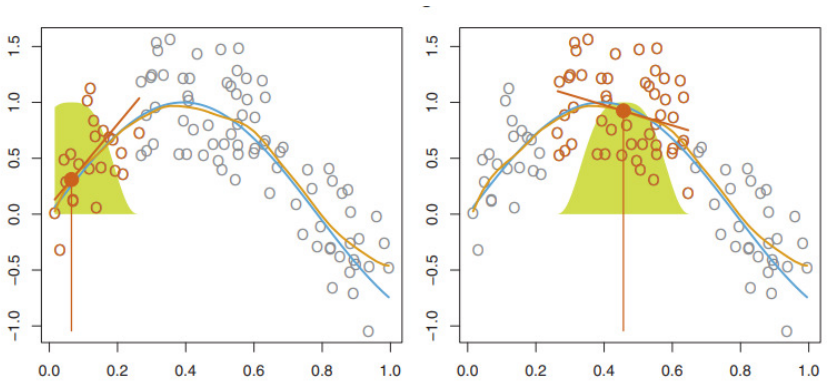
Lowess regression

Alternative to splines: Local linear regression

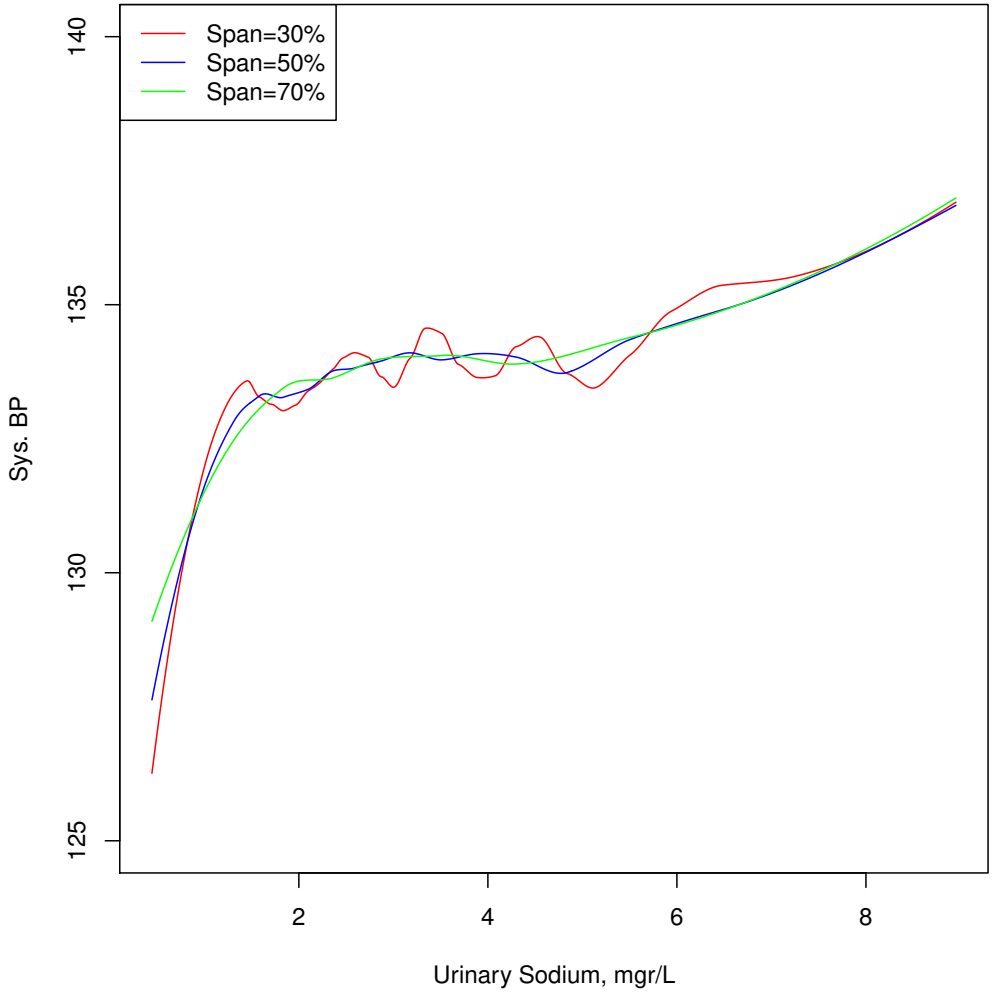
- Fit "mini" linear regressions around $X_{i,k}$, $i \leq n$
  - using only m nearest neighbours $\{X_{j,k}\}$ around $X_{i,k}$
  - weighted according to proximity to $X_{i,k}$: $\gamma_{ij} = \gamma(X_{j,k}, X_{i,k})$
- $Y_i = \beta_{0,i} + \beta_{1,i} X_{i,k} + \epsilon$
- $\{\beta_{0,i}, \beta_{1,i}\}^* = \arg\min_{\beta_0,\beta_1} \sum_j \gamma_{ij}(y_i - \beta_0 - \beta_1 x_j)^2$

Drawbacks:

- Requires many data points: large n
- Hard to interpret



Lowess regression



GAMs

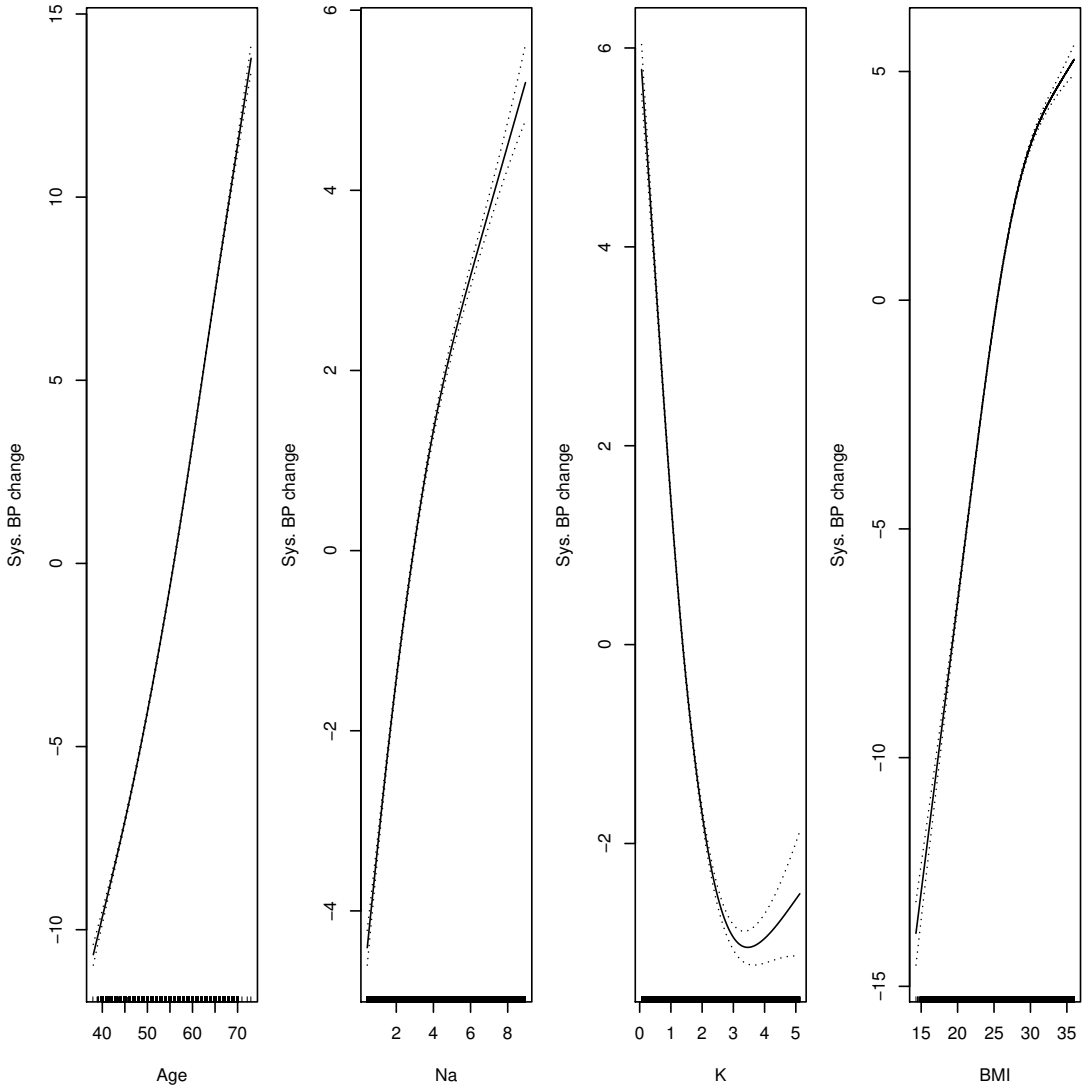
Generalised additive models: extension of non-linear framework to multiple linear models
$X = \{X_{k,i}\}$, $i \leq n$, $k \leq p$:

$$Y = \beta_0 + \sum_{k}^{p} f_k(X_k) + \epsilon$$

Where $f_k(.)$: linear / polynomial / splines .... for each $k \leq p$

- Non-linear multivariate analysis, easily interpretable
- Can miss out interactions between variables
  - Can include $f_{k,h}(X_k, X_h)$

GAMs


Sys. BP change
15
10
5
0
−5
−10
40
50
60
70
Age
Sys. BP change
6
4
2
0
−2
−4
2
4
6
8
Na
Sys. BP change
6
4
2
0
−2
0
1
2
3
4
5
K
Sys. BP change
5
0
−5
−10
−15
15
20
25
30
35
BMI

Important takeaways

- Polynomial regression
- Piecewise regression
- Splines
  - cubic / natural splines
  - smooth splines
  - selection of knots / λ
- Lowess - local linear regress.
- GAM: generalised additive models