Modèles additifs

C. Ambroise

Laboratoire Statistique et Génome
UMR CNRS 8071

Plan

Plan

Introduction I

- Supponsons que l'on cherche à expliquer une variable réponse y par une fonction des prédicteurs $X_1, X_2, ..., X_p$
- Le modèle linéaire pose

$$y = \beta_0 + \sum_j \beta_j X_j + \epsilon$$

- Pour coller plus aux données, on peut chercher une fonction f

$$y = f(X_1, X_2, ..., X_p) + \epsilon.$$

Introduction II

- Le modèle additif est un compromis

$$y = \beta_0 + \sum_j \beta_j f_j(X_j) + \epsilon$$

avec

- $f_j(X_j) = X_j$ donne le modèle linéaire
- $f_j(X_j) = \log X_j$ permet des transformations non linéaire des variables
- $f_j(X_j)$ peut aussi être une fonction non paramétrique.

Plan

Régression non paramétrique univariée I

- Le modèle considère

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

avec

  - (X, Y) un couple de variable aléatoires,
  - $E[\epsilon] = 0$ et $var[\epsilon] = \sigma^2$
  - $X \perp\!\!\!\perp \epsilon$

- La fonction f est estimée à partir d'un n-échantillon $\{(x_i, y_i)\}_{i=1}^{n}$

Plan

Estimateurs à noyaux

- Estimation de f :au point x

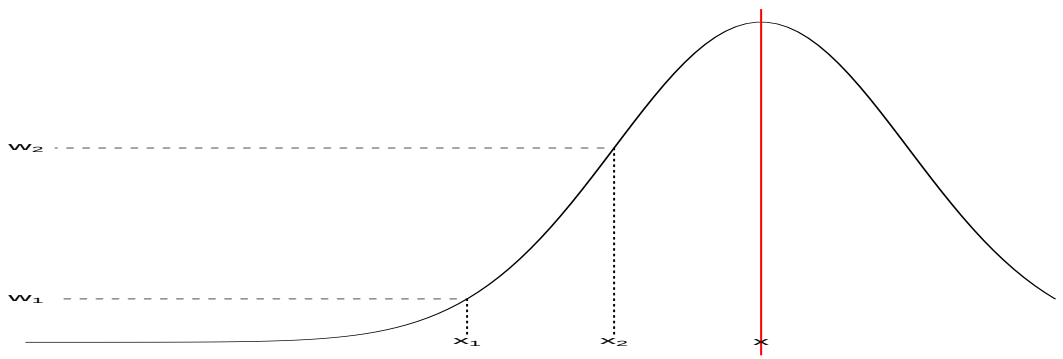
$$\hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{j=1}^{n} K(\frac{x - x_j}{\lambda}) y_j = \frac{1}{n} \sum_{j=1}^{n} w_j(x) y_j,$$

where

- $w_j(x) = K(\frac{x-x_j}{\lambda})/\lambda$.
- $\int K = 1$

Exemples de noyaux

- La fonction K permet de donner de l'importance aux voisins x_j "proches" du point x d'intérêt
- Si les x_j sont placés de manière très inégale $\hat{f}_\lambda$ est susceptible d'être peut performante
- Noyau Gaussien : $K(t) = \frac{1}{2\pi} \exp(-t^2/2)$



- Noyau d'Epanechnikov : $K(t) = 3/4(1 - t^2)\mathbb{I}_{\|t\| \leq 1})$
- ...

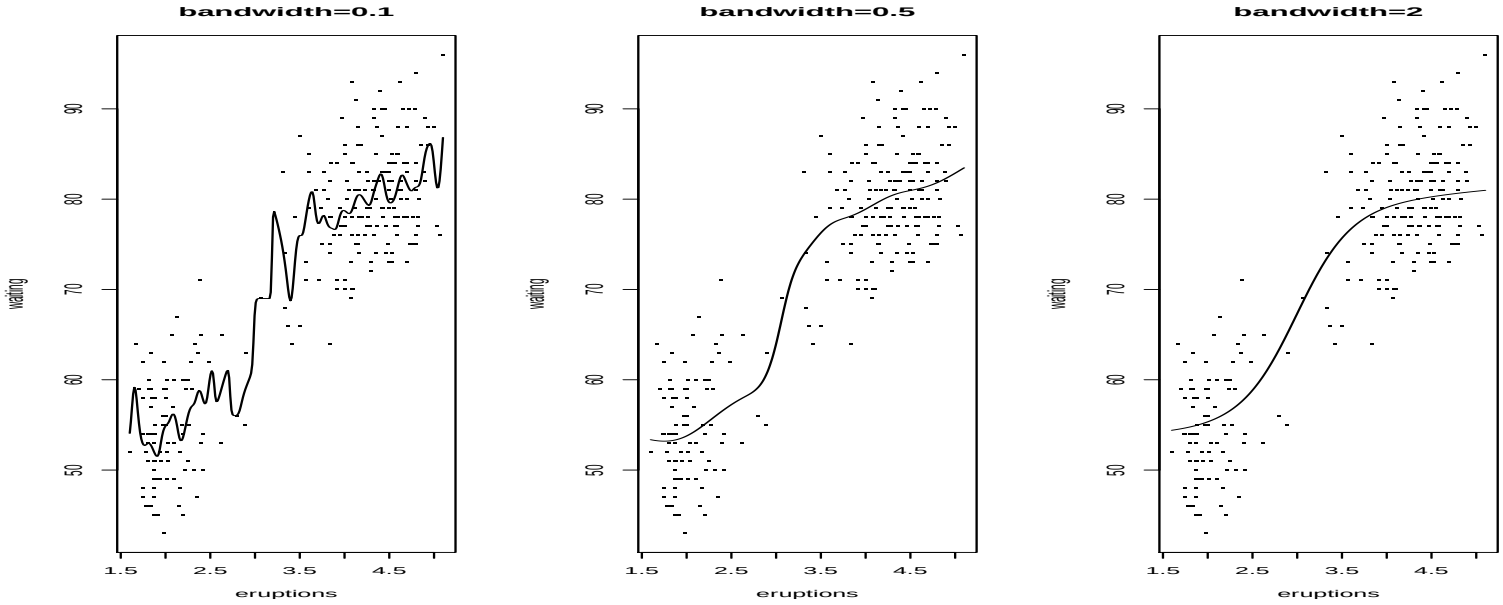
Estimateur de Nadaraya-Watson

- Pour améliorer la robustesse vis à vis de la répartition des x_j, il est préférable d'utiliser l'estimateur de Nadaraya-Watson

$$\hat{f}_\lambda(x) = \frac{\sum_{j=1}^n w_j(x) y_j}{\sum_{j=1}^n w_j(x)}.$$

Implémentation pratique

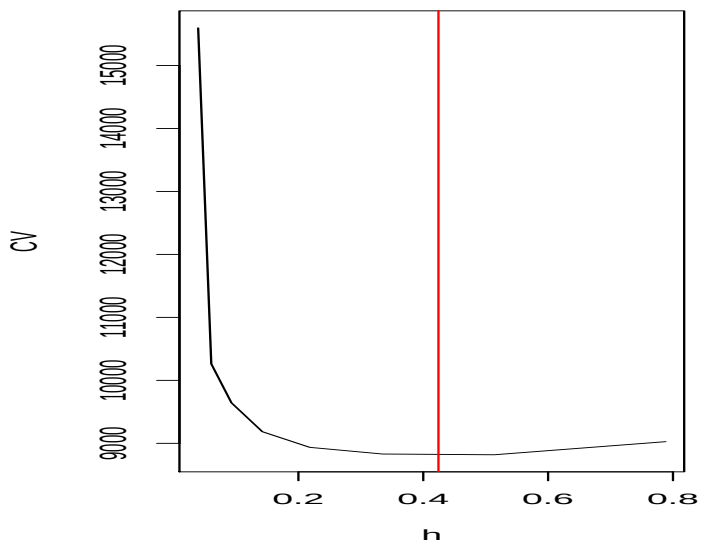
- L'implémentation de la régression par noyaux requiert deux choix :
  - le noyau
  - la largeur de bande λ

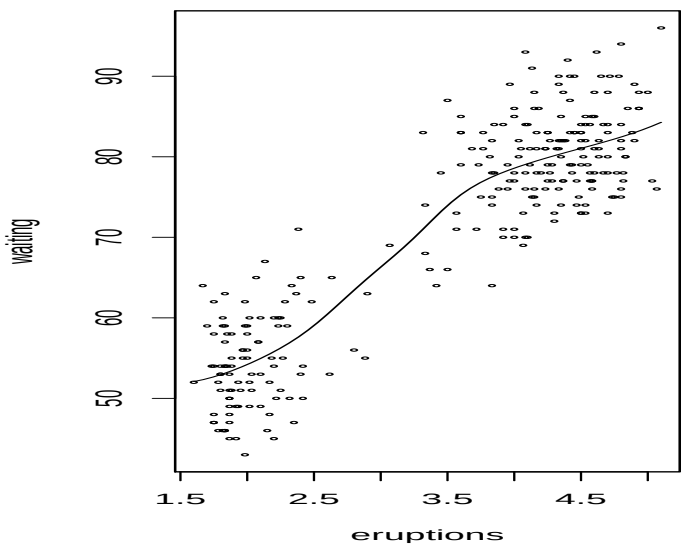


Choix de λ

- La validation croisée représente une possibilité

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{\lambda(-j)}(x_j))^2$$





La regression univariée par noyaux en R

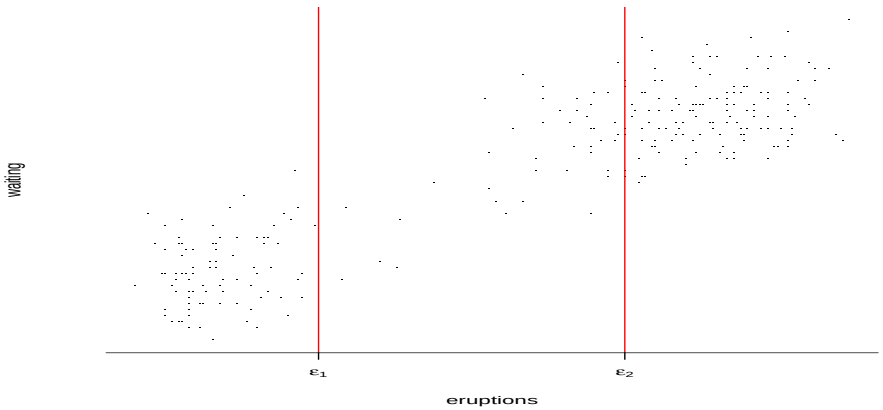
- fonction `ksmooth()`,
- fonction `sm.regression()` de la `library(sm)`

```
library(sm)
par(mfcol=c(1,2))
hm<-hcv(faithful$eruptions,
        faithful$waiting,display='lines')
abline(v=hm,col=2)
sm.regression(faithful$eruptions,
              faithful$waiting,h=hm,
              xlab="eruptions",ylab="waiting")
```

Plan

Splines de regression

- L'idée consiste à construire des polynômes de regression par morceaux qui se raccordent de façon lisse
- Les points de raccord sont les **nœuds**
- Aux nœuds $\{\epsilon_k\}_{k=1..K}$ on observe la continuité des dérivées à un certain ordre (Splines cubiques, ordre 2)



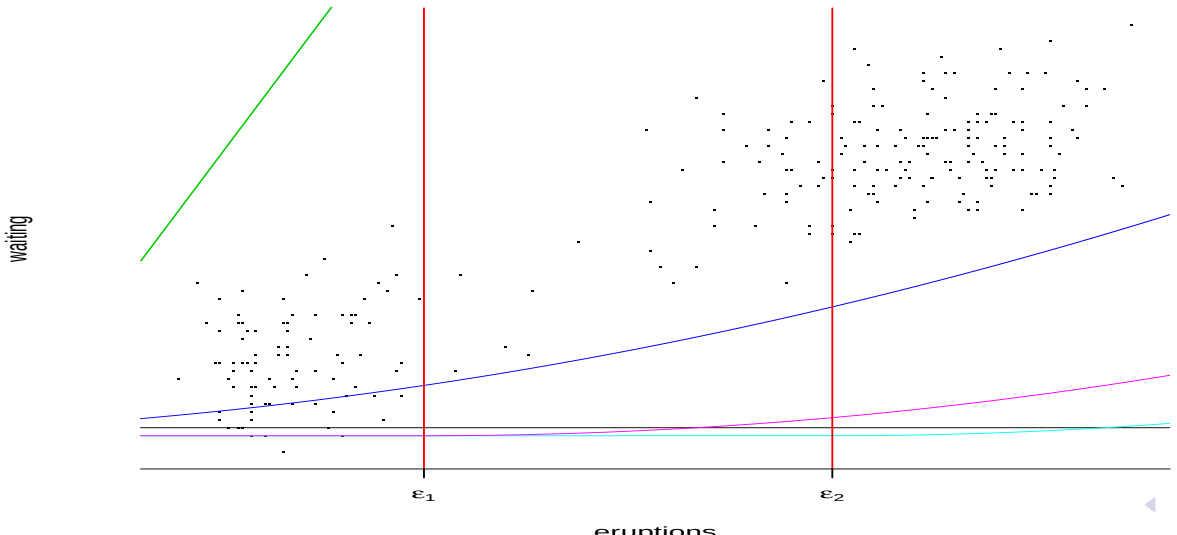
Base de pôlynomes

- Une base de pôlynomes

$$\{N_j(x)\}_{j=1}^{K+q+1} = \{1, x, x^2, x^3, ..., x^q, (x-\epsilon_1)_+^q, ..., (x-\epsilon_K)_+^q\}$$

- Un estimateur combinaison des fonctions de la base

$$f(x) = \sum_{j=1}^{K+q+1} \beta_j N_j(x)$$



Les nœuds

- Les nœuds sont placés de manière régulière (également espacés)
- ou bien placés sur des percentiles réguliers (5%, 10%, ...)
- Le nombre de nœuds détermine le degré de lissage (régularité de la fonction)
- La fonction $\hat{f}$ est choisie (par estimation des β_j) pour minimiser l'erreur quadratique en respectant les contraintes aux nœuds

$$MSE = \frac{1}{n} \sum_i \| y_i - \sum_{j=1}^{K+q+1} \beta_j N_j(x_i) \|^2$$

Splines de lissage

- Supposons que les observations x_i soient toutes considérées comme des nœuds.
- $f()$ qui minimise MSE serait trop irrégulière pour
- on considère donc $f()$ minimisant un critère pénalisé (fidélité aux données + régularité)

$$MSE + \lambda \int (f^{(2)}(t))^2 dt$$

avec λ choisi par validation croisée

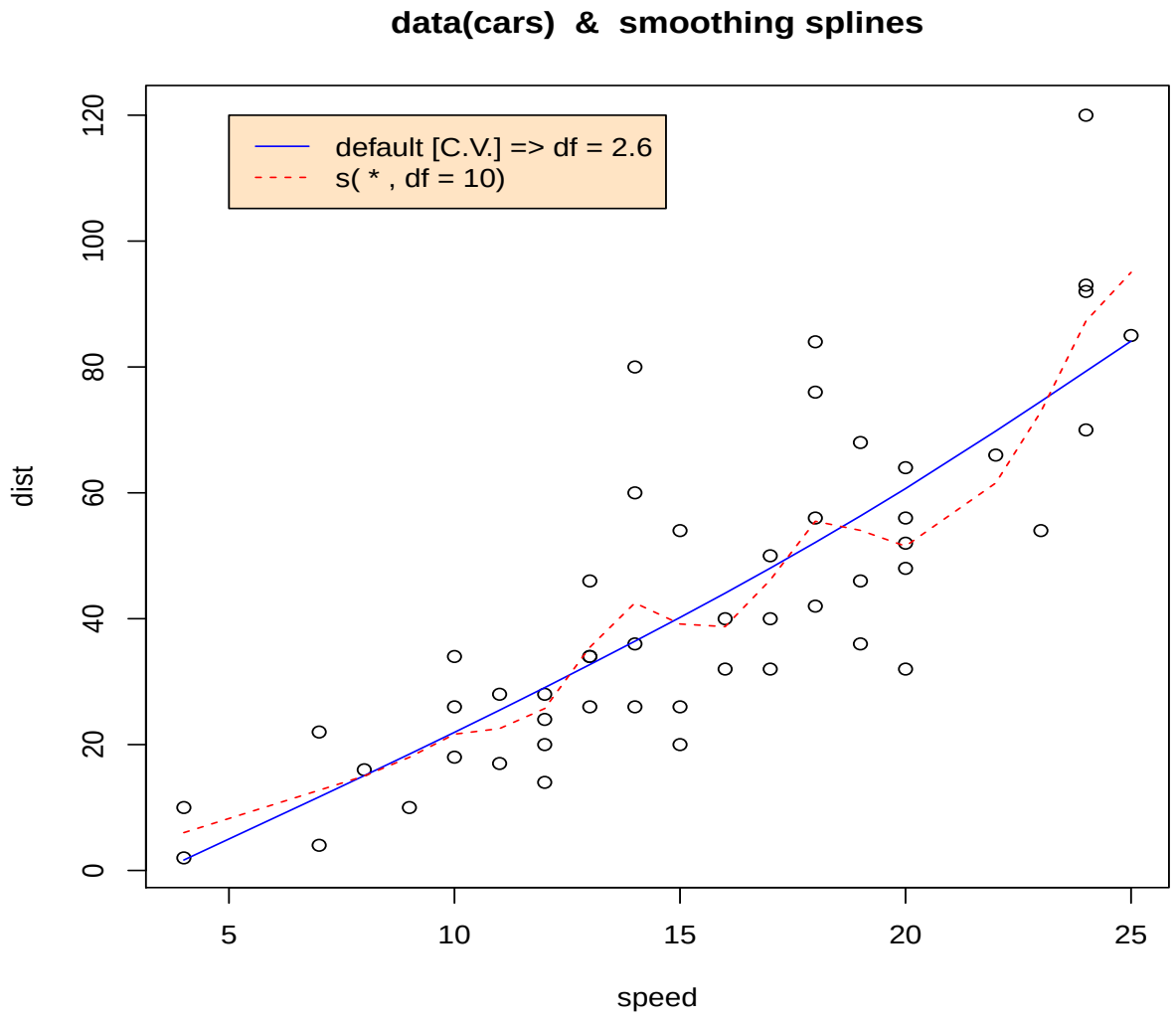
- La solution optimale pour

Splines de lissage en R

- fonction `smooth.spline`,

```
attach(cars)
plot(speed, dist, main = "data(cars) &
     smoothing splines")
cars.spl <- smooth.spline(speed, dist)
(cars.spl)
lines(cars.spl, col = "blue")
lines(smooth.spline(speed, dist, df=10),
       lty=2, col = "red")
legend(5,120,c(paste("default [C.V.] => df =",
        round(cars.spl$df,1)),
                "s( * , df = 10)"),
       col = c("blue","red"), lty = 1:2,
       bg='bisque')
detach()
```

Splines de lissage en R



Plan

Modèles additifs

$$y = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \epsilon$$

- Les fonction f_j peuvent des splines de lissage, de régression, noyau, ...

Algorithme de Backfitting

1. Initialisation des fonctions $\hat{f}_j$ (nulles, régression linéaire, ...)
2. Cycle sur les fonctions $j = 1, ..., p$ jusqu'à convergence
   - $\hat{f}_j$ est ajustée pour prédire les résidus $(y_i - \beta_0 - \sum_{k \neq j} f_k(x_{ik}))$ en utilisant l'échantillon d'apprentissage $\{(y_i, x_{i1}, ..., x_{ip}))\}_{i=1,...,n}$

- En fait l'hypothèse de résidus Gaussiens on peut utiliser les tests d'hypothèses des modèles emboîtés (`anova` en R)

Modèles additifs en R

- package `gam` de Hastie et Tibshirani,
- package `mgcv` de Wood.

```
> library(faraway)
> data(ozone)
> olm <- lm(O3~temp + ibh + ibt, ozone)
> summary(olm)

Call:
lm(formula = O3 ~ temp + ibh + ibt, data = ozone)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3224  -3.1913  -0.2591   2.9635  13.2860

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.7279822  1.6216623  -4.765 2.84e-06 ***
temp         0.3804408  0.0401582   9.474  < 2e-16 ***
ibh         -0.0011862  0.0002567  -4.621 5.52e-06 ***
ibt         -0.0058215  0.0101793  -0.572    0.568
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.748 on 326 degrees of freedom
Multiple R-squared: 0.652, Adjusted R-squared: 0.6488
F-statistic: 203.6 on 3 and 326 DF,  p-value: < 2.2e-16
```

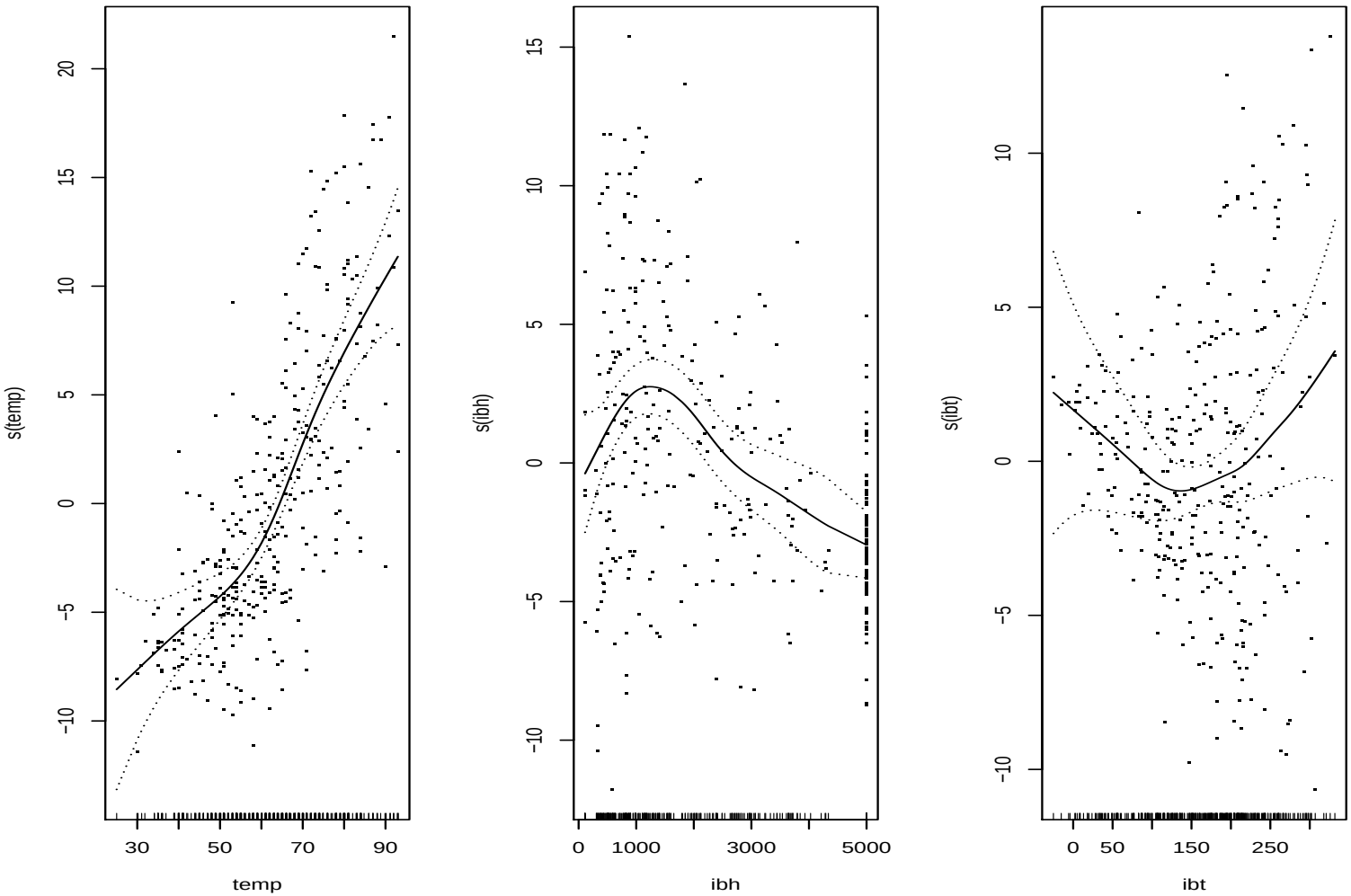
Modèles additifs en R

```
> amgamr <- gam(O3~ s(temp) + s(ibh) + s(ibt), data=ozone)
> summary(amgamr)
Call: gam(formula = O3 ~ s(temp) + s(ibh) + s(ibt), data = ozone)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-13.2725  -2.3447  -0.2385   2.1820  12.9764

(Dispersion Parameter for gaussian family taken to be 18.7364)

    Null Deviance: 21115.41 on 329 degrees of freedom
Residual Deviance: 5939.448 on 317 degrees of freedom
AIC: 1918.292
Number of Local Scoring Iterations: 2
DF for Terms and F-values for Nonparametric Effects
            Df Npar Df Npar F     Pr(F)
(Intercept)  1
s(temp)      1       3 7.2983 9.541e-05 ***
s(ibh)       1       3 6.4328 0.0003060 ***
s(ibt)       1       3 5.5790 0.0009679 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèles additifs en R



Modèles additifs généralisés

- Dans les modèles linéaire généralisés, une fonction d'un paramètre μ est expliquée de manière linéaire

$$g(\mu) = \eta = \beta_0 + \sum_j \beta_j x_j$$

- Les modèles additifs linéaire généralisés, utilise

$$g(\mu) = \eta = \beta_0 + \sum_j \beta_j f_j(x_j)$$