Introduction à la régression
cours n°4

ENSM.SE – axe MSA

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

Retour sur TP tailles

- Compréhension du problème
  - Rôle des variables sexe et poids
- Analyses graphiques unidimensionnelles
  - Valeurs à corriger
  - Enfants trop jeunes à supprimer
  - Adultes plus âgés
  - Prédicteurs ou prédicateurs ??

Autres analyses graphiques

- Corrélation des prédicteurs
  - poids/taille → utilisation de l'IMC
  - taille père/taille mère → (tp+tm)/2, tp-tm, tp/tm…
- Analyses bidimensionnelles :
  - Effet apparent de sexe, taille père, taille mère.
  - Interaction sexe*taille père
- Corrélations et termes d'interaction

Modèle linéaire

- Un modèle par sexe ou un modèle global ?
- Utiliser la fonction lm
- Attention aux tables d'ANOVA de R
- Penser à utiliser R^2 et $\hat{\sigma}$

Retour sur l'étude critique

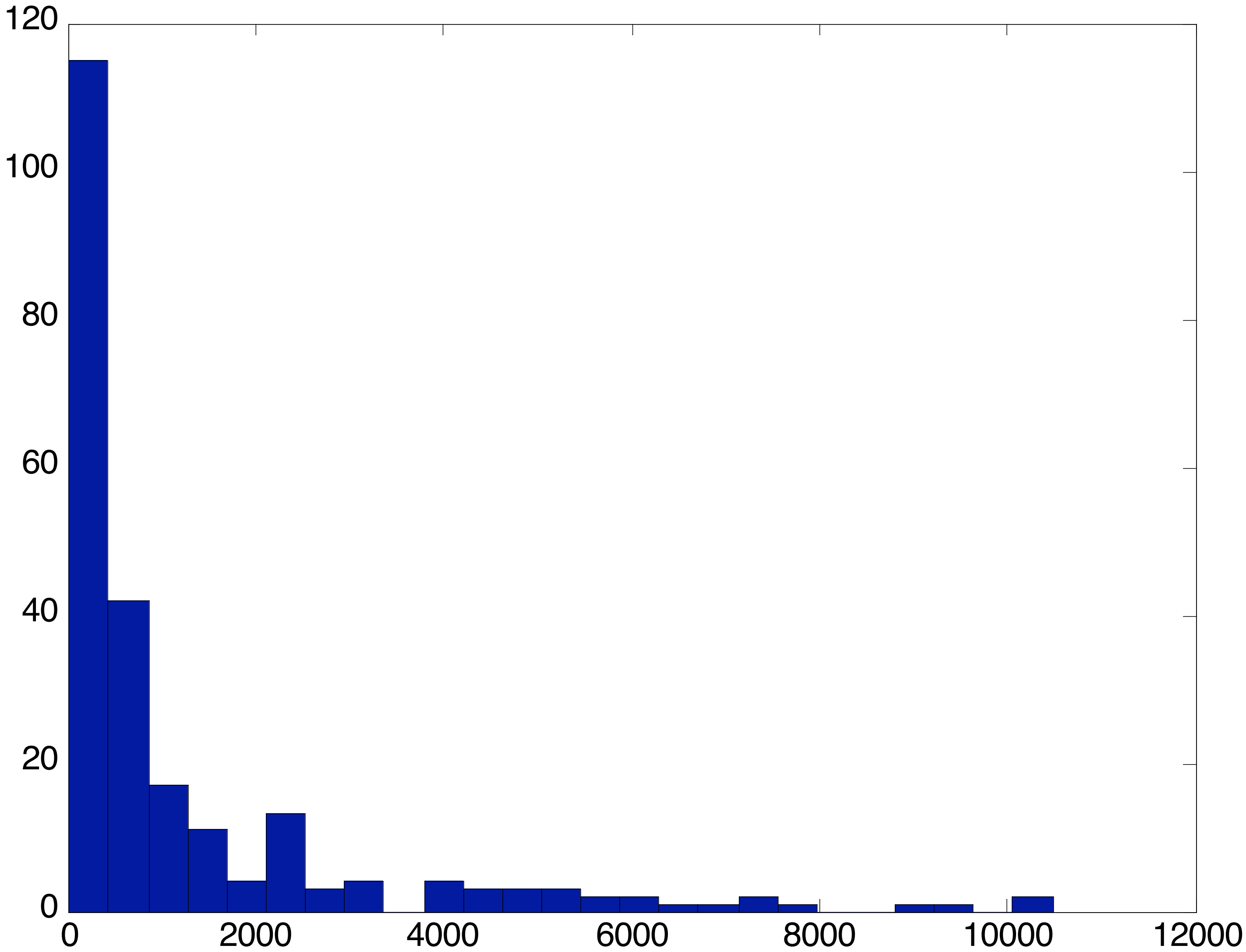
- Difficultés :
  - Gestion du temps
  - Travail en groupe
  - Critiquer sans refaire
  - Ouvrir des voies sans les explorer
  - Critique pas à pas vs critique globale

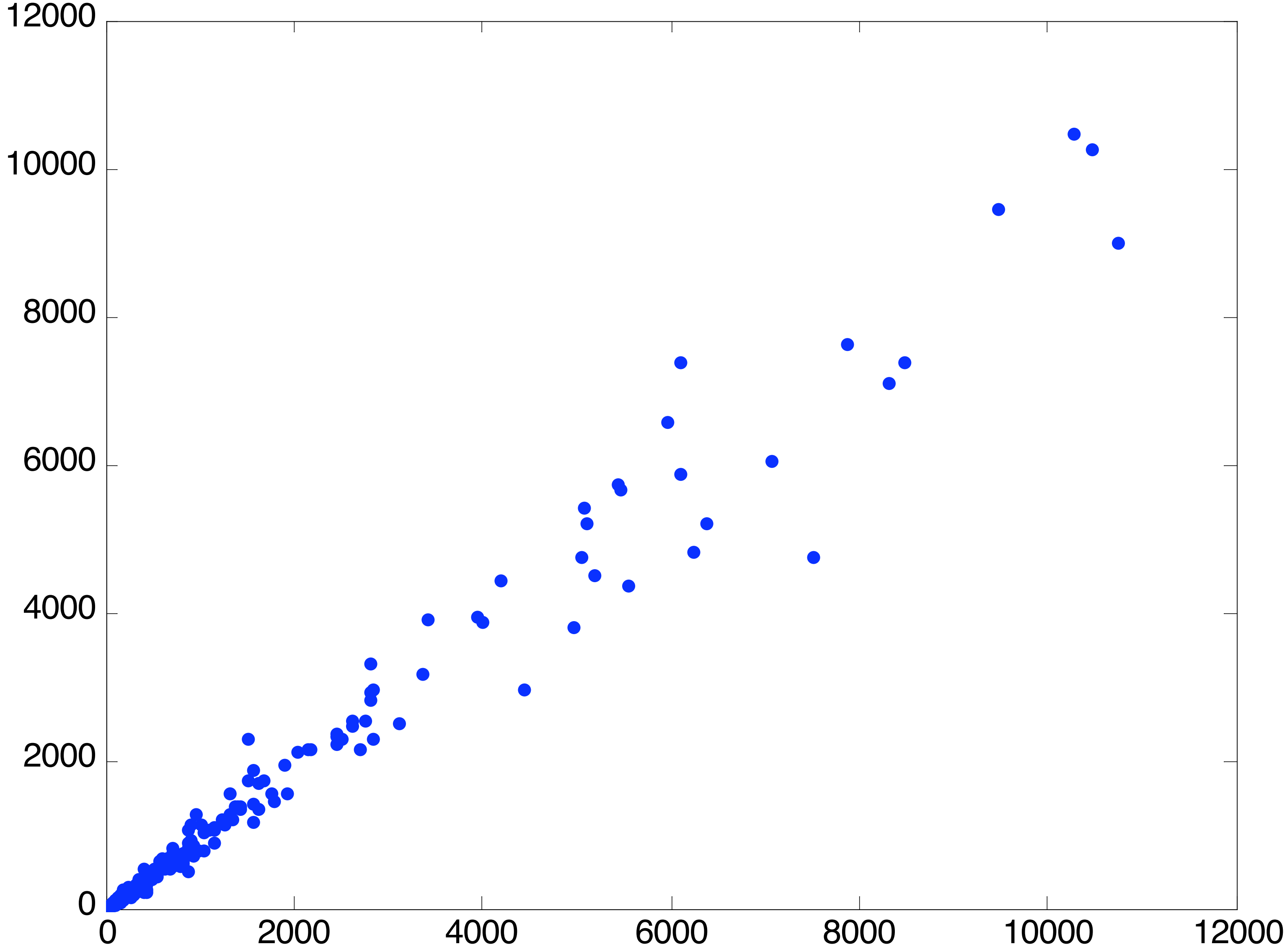
Etude critique

- Remarques globales :
  - traduction : attention aux faux amis
  - absence totale d'analyse graphique
  - démarche « en aveugle »
  - validation ?
  - analyse des prévisions ?

Etape par étape

- Step 1 (prédicteurs) :
  - Peu de renseignements sur les données (origine ? aspect temporel ?…)
  - Corrélation des prédicteurs → nouveaux prédicteurs
  - Pas d'analyse graphique
- Step 2 (stepwise) :
  - Trouver une source (internet, R…)
  - Corrélation des prédicteurs (e.g. coutest et jourest)


120
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
12000

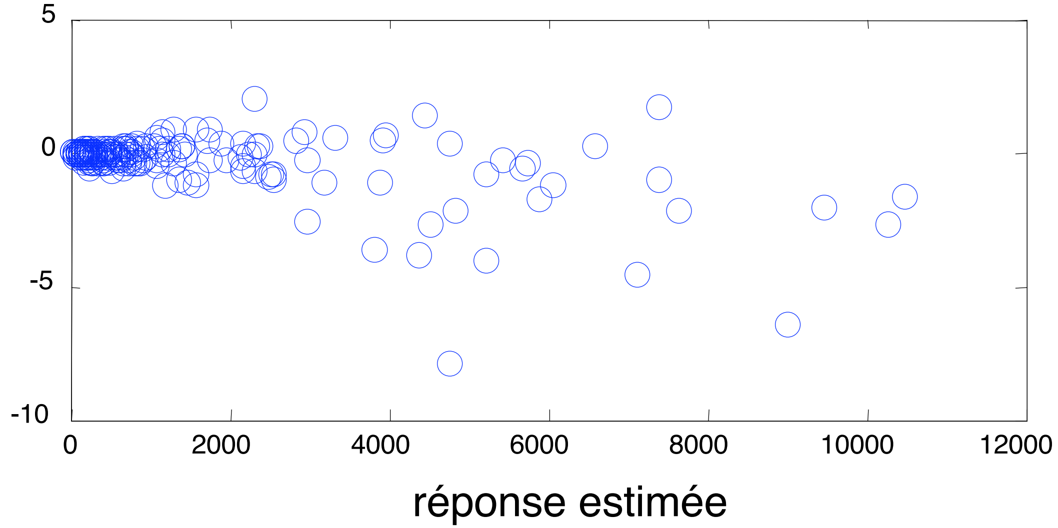

12000
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
10000
12000

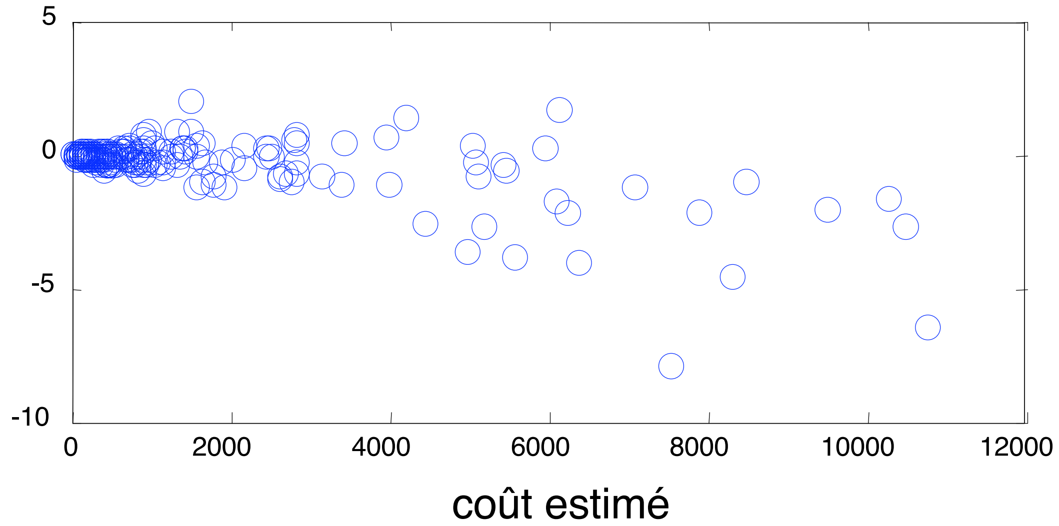
Etape par étape

- Step 3 (modèle quadratique et deux prédicteurs) :
  - Pourquoi $coutest^2$?
  - Attention : $statut^2$=statut
  - Corrélation des prédicteurs (statut*coutest et statut*$coutest^2$ par ex.). Voir centrage

- Step 4 (validation) :
  - Seulement ANOVA + R^2
  - σ_{est} = 296 k$ → précision pour de petits chantiers ?
  - Pas d'analyse de résidus
  - Pourquoi valider avant de finaliser le modèle ?

Résidus standardisés





Etape par étape

- Step 5 (modèles emboîtés) :
  - RAS
  - La p-valeur est préférable (on rejette H_0 à 1%)
- Step 6 (prévisions) :
  - Souvent non examiné par les groupes
  - Intervalles de prévision médiocres en général…
  - et ridicules pour des petits chantiers (limite inférieure négative !)

Rappel démarche

- Observations graphiques
  - prédicteurs, réponse, prédicteurs entre eux et contre réponse (dont interactions)
- Modélisation et inférence
  - estimation paramètres + corrélation, ANOVA, résidus
- Observations à problème ou influentes
- Prévisions

Observations aberrantes Observations influentes

- Retour sur l'exemple 3
  - Observation n°9 aberrante (résidus studentisés)
  - Estimation du modèle très perturbée par la suppression de l'observation n°9
- Simulation Excel

ANOVA avec 25 données

Table d'analyse de variance

Source	df	SS	MS	F	p
Regression	2	5544	2772	261.7	4.441e-016
Error	22	233	10.59		
Total	24	5777			

Coefficients

Root MSE	3.255	R-square	0.9597
		R-sq(adj)	0.956

Paramètres estimés

Predictor	Coeff	Stdev	t-ratio	p
intercept	2.353	1.095	2.149	0.04292
nb	1.615	0.1705	9.474	3.199e-009
dist	0.01437	0.003608	3.984	0.0006273

ANOVA avec 24 données

Table d'analyse de variance

Source	df	SS	MS	F	p
Regression	2	2291	1146	194.6	2.798e-014
Error	21	123.6	5.887		
Total	23	2415			

Coefficients

Root MSE	2.426	R-square	0.9488
		R-sq(adj)	0.9439

Paramètres estimés

Predictor	Coeff	Stdev	t-ratio	p
intercept	4.456	0.951	4.686	0.0001263
nb	1.497	0.13	11.52	1.552e-010
dist	0.01032	0.002849	3.621	0.001601

Détection de valeurs influentes

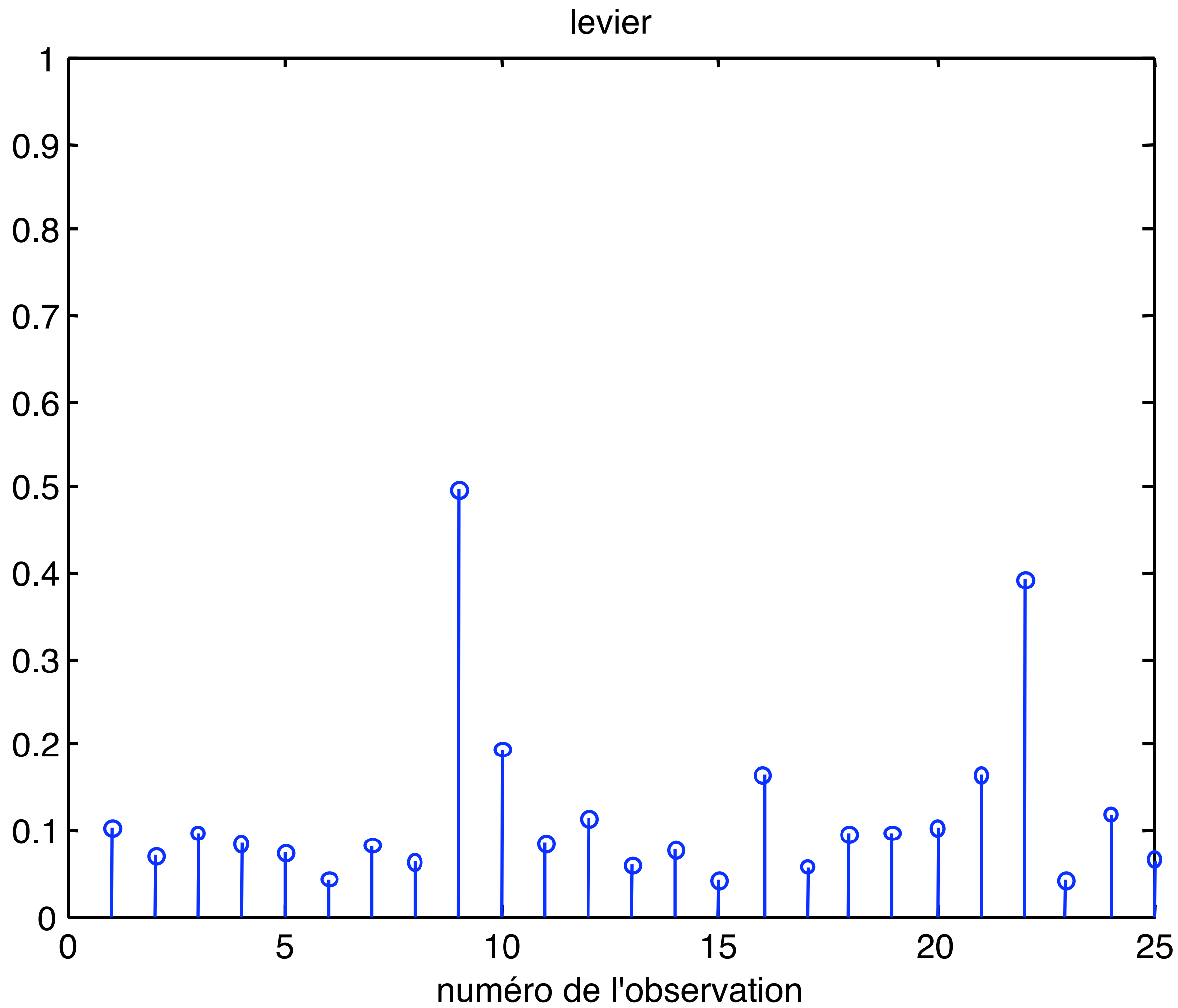
- Levier h_{ii} :

$$\hat{y}_i = \sum_{i=1}^{n} h_{ij}\, y_j \text{ et } \mathrm{Var}(\hat{\varepsilon}_i) = \sigma?(1-h_{ii})$$

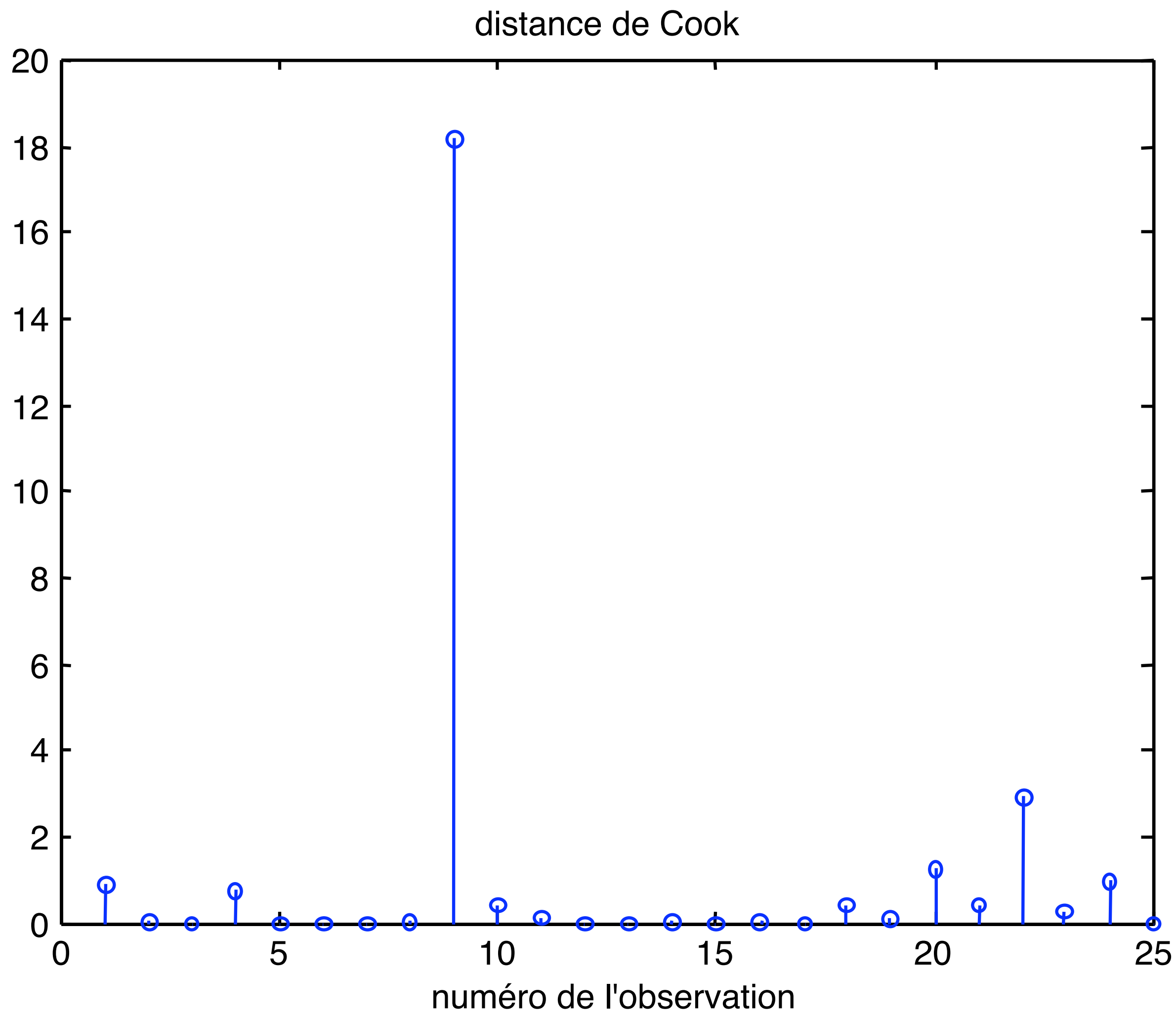
- Distance Cook D_i :

Distance a-dimensionnelle entre les réponses estimées avec ou sans l'observation n° i.

Exemple 3 - leviers



Exemple 3 – distances de Cook



Prévisions

- Nouvelle valeur des prédicteurs x_{new}
- Prévision pour y :

$$\hat{y}_{new} = x_{new}\,\hat{\beta}$$

  - Intervalle de confiance pour $x_{new}\,\beta$ (réponse espérée) ?
  - Intervalle de prévision pour la réponse y_{new} ?

Intervalles de confiance/prévision

- De la forme :

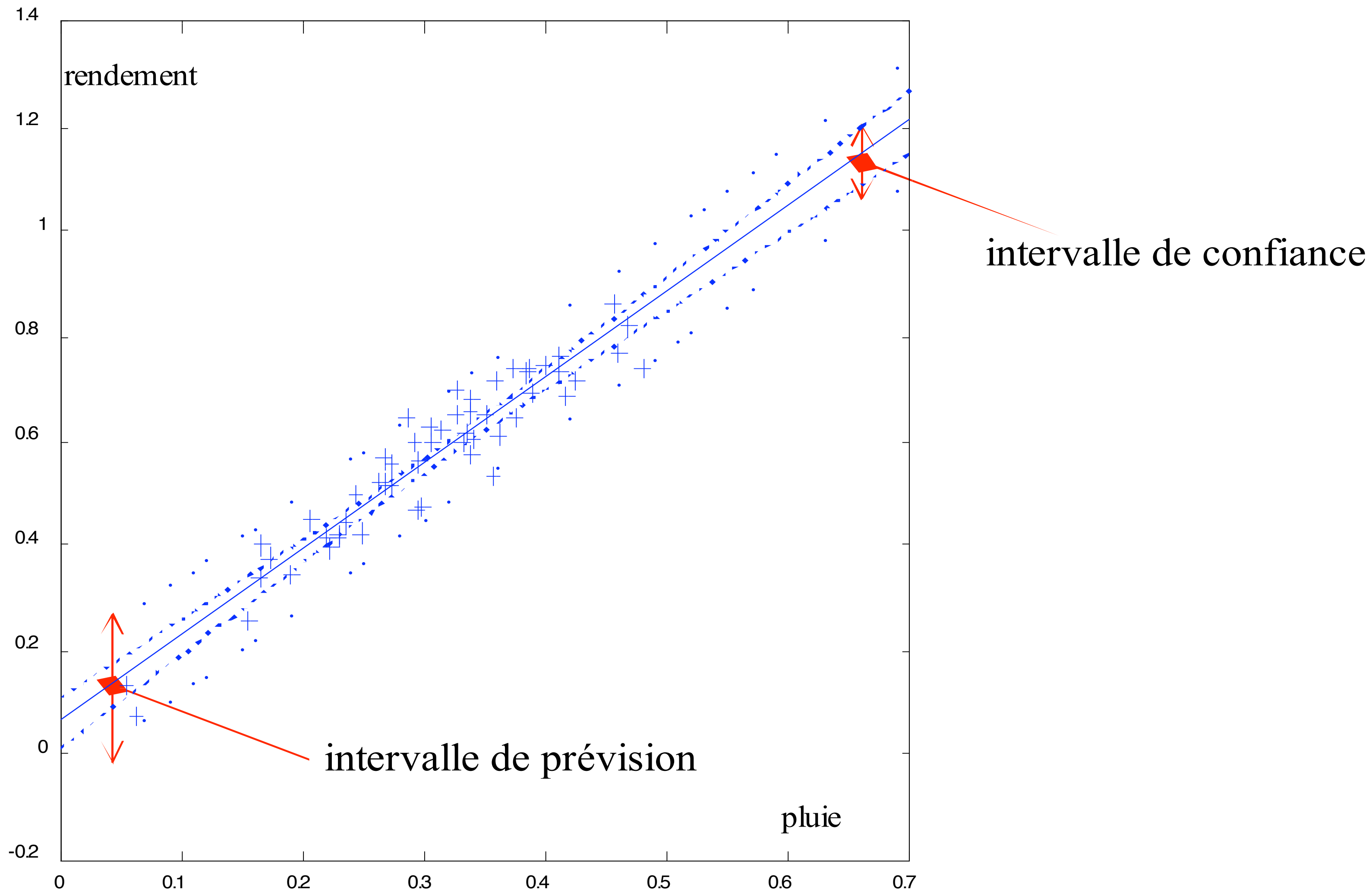
$$[x_{new}\beta - s(x_{new})t^{-1}_{n-p-1}(1-\alpha/2),\ x_{new}\beta + s(x_{new})t^{-1}_{n-p-1}(1-\alpha/2)]$$

- Confiance :

  La pente est-elle >1 ? la droite passe-t-elle par 0 ?

- Prévision :

  A quel rendement s'attendre pour 20 mm de pluie ?


rendement
pluie
intervalle de confiance
intervalle de prévision
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Régression non paramétrique

- Principe : estimer une relation sans fixer de modèle a priori (data driven relation) :
- Données = n couples (x_i, y_i) tels que :

  $y_i = f(x_i) + \varepsilon$
- Problème : obtenir une estimation de f(x), notée $\hat{f}(x)$
- Exposé pour un seul prédicteur

Kernel smoothing

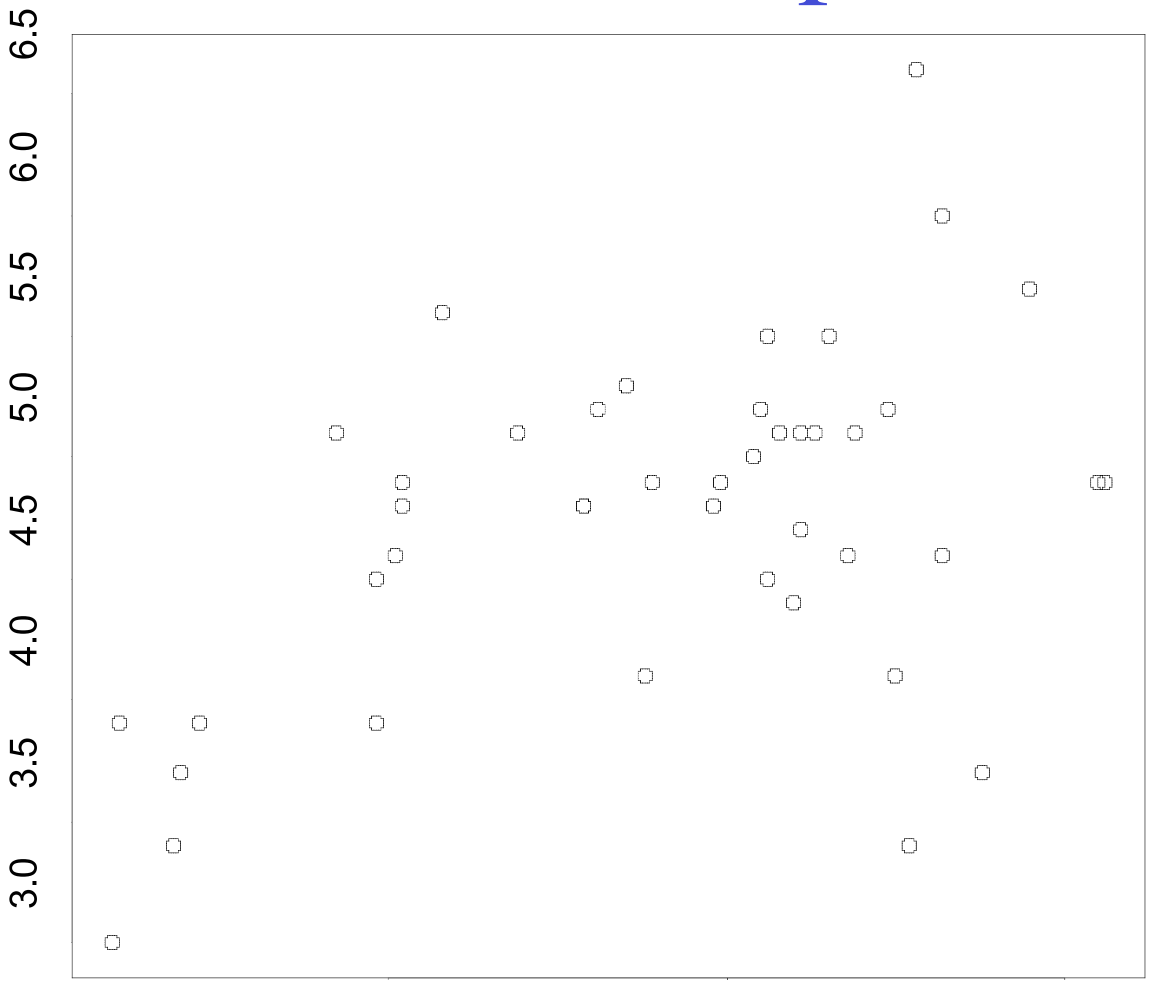
- Principe de moyenne mobile
- En plus lisse :

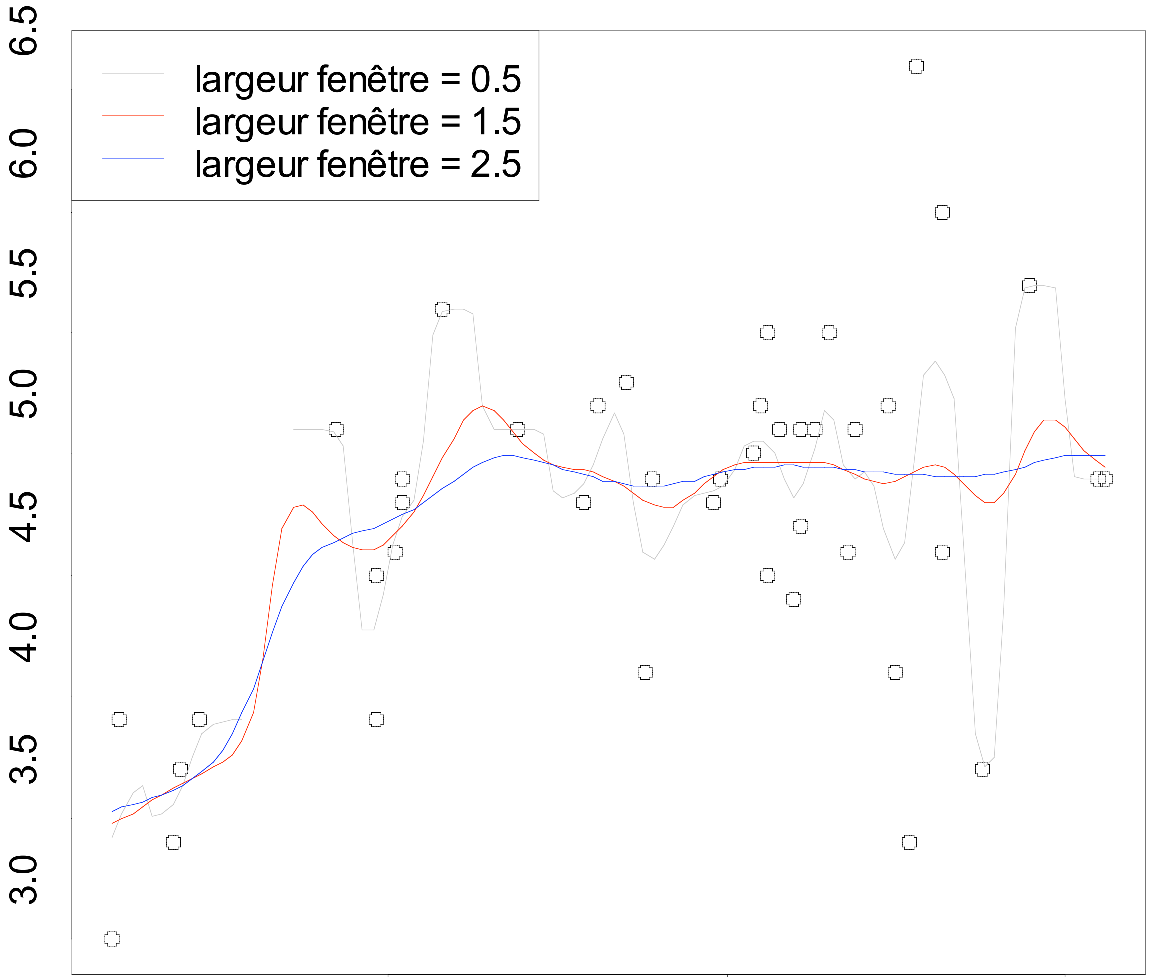
  Estimateur de Nadaraya-Watson
- Noyaux K :
  - Gaussien : $\exp(-x^2/2h^2)$
  - Epanechnikov : $(1-x^2/h^2)$ si $|x| \leq h$
  - Tricubique : $(1-|x/h)|^3)^3$ si $|x| \leq h$
- Choix de la largeur de fenêtre h
- Sous R : ksmooth

Exemple 4

- **Les données :** 43 enfants atteints de diabète insulino-dépendant. On connaît leur âge x_i exprimé en années et leur concentration sanguine en peptide-C en pmol/ml.
- **Intérêt :** la concentration en peptide-C reflète le potentiel d'un individu à sécréter l'insuline et donc à métaboliser les glucides.
- **Problème :** prévoir cette concentration

Données exemple 4

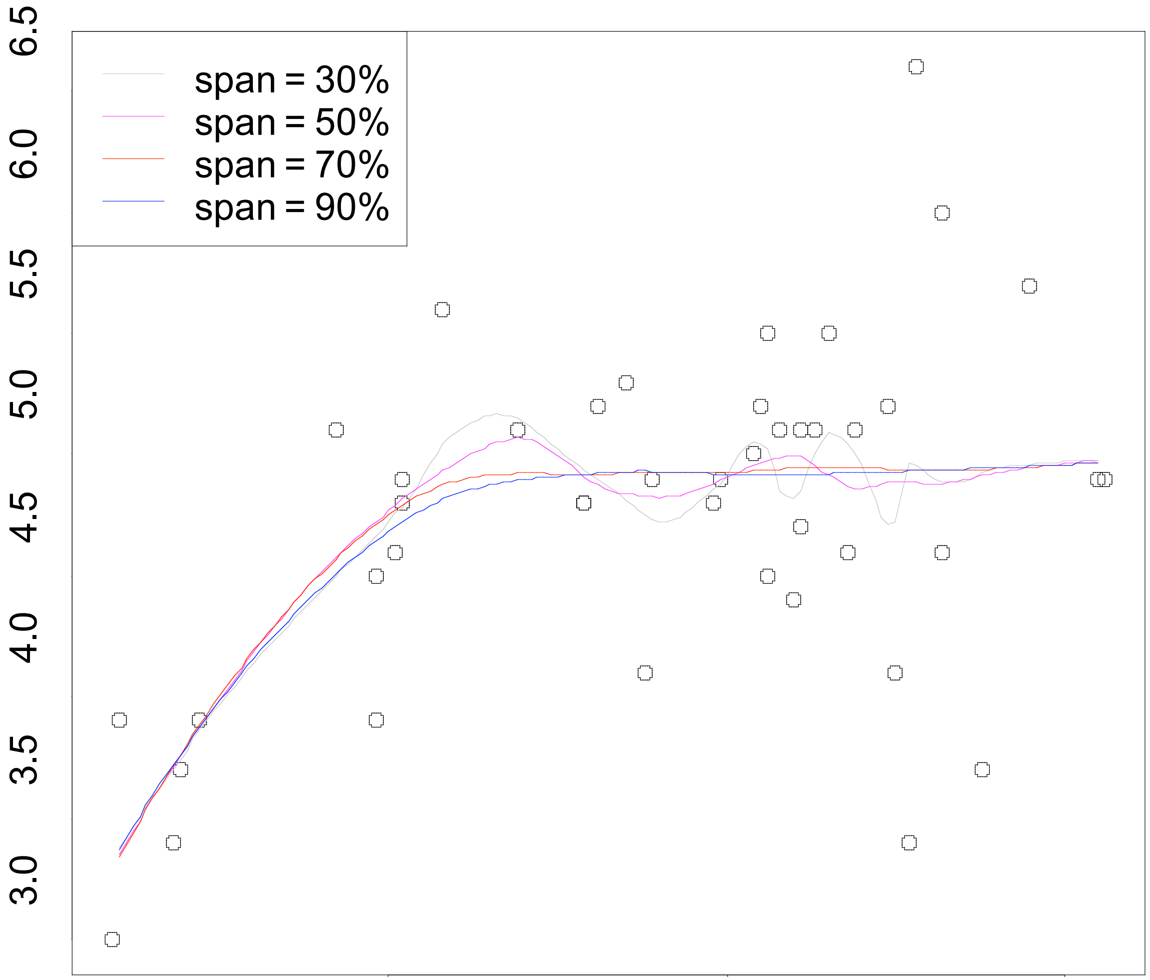


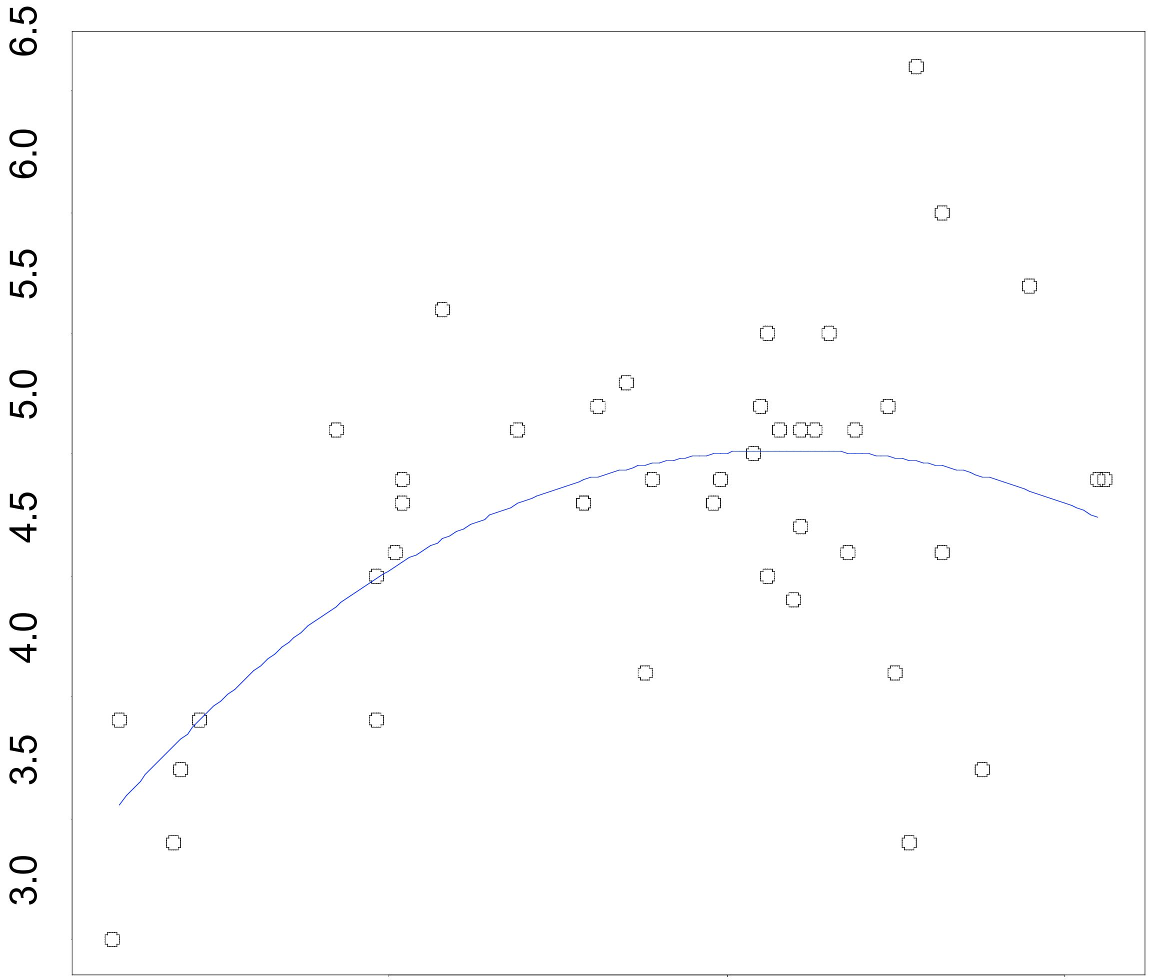

largeur fenêtre = 0.5
largeur fenêtre = 1.5
largeur fenêtre = 2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5

Modèle linéaire local

Principe : approximation locale de f par un polynôme de bas degré (≤ 2).

- On donne un poids aux données en fonction de leur éloignement du point courant
- Poids gouverné par un noyau K et une largeur de fenêtre h.
- Estimation en x du polynôme par moindres carrés pondérés.
- Choix de h
- Sous R : loess (qui contient en plus


span = 30%
span = 50%
span = 70%
span = 90%
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5


6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0