

Régression robuste

Gilbert Saporta

Conservatoire National des Arts et Métiers

<http://cedric.cnam.fr/~saporta>

Janvier 2012

1. Les moindres carrés

- Fonction de perte: $L(y; f(x)) = (y - f(x))^2$
- Sur la population: f optimal: $f(x) = E(Y/x)$
- Hypothèse de régression linéaire:
 $E(Y/x) = \beta_0 + \beta_1 x$
en régression multiple: $E(Y/\mathbf{x}) = \mathbf{x}' \beta$

- Estimateur des moindres carrés

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{A}\mathbf{y}$$

$$\mathbf{y} - \mathbf{X}\mathbf{b} \perp W \quad (\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{X}\mathbf{u} = 0 \quad \forall \mathbf{u}$$

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b} \quad \text{Equations normales}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

$$\text{projecteur } \mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$$

- \mathbf{b} estimateur de variance minimale de β parmi les estimateurs linéaires sans biais
- estimateur du maximum de vraisemblance si résidus gaussiens iid

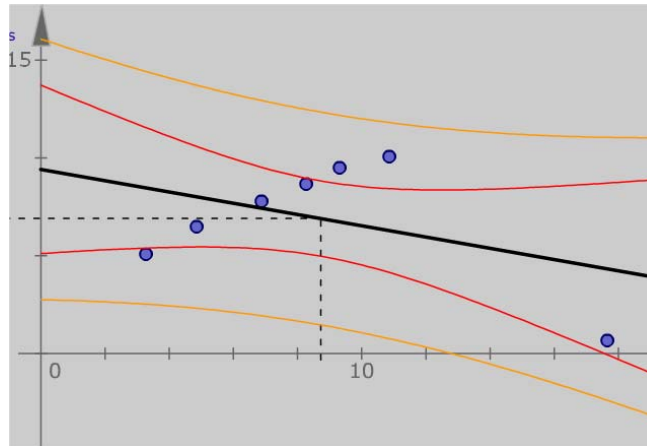
$$V(\mathbf{b}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- Démo

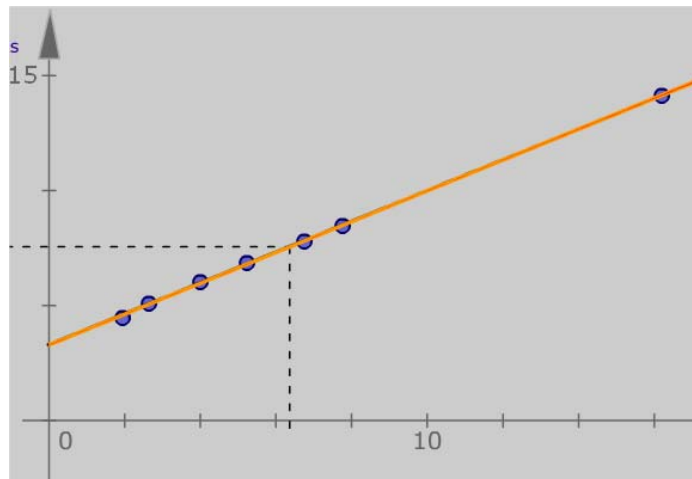
[cours de Marc Bourdeau \(Montréal\) sur le modèle linéaire](#)

http://www.agro-montpellier.fr/cnam-lr/statnet/cours_autre.htm

- Sensibilité aux valeurs extrêmes



- Ne pas confondre observations aberrantes et observations influentes (au sens de l'écart au modèle)



Résidus et influence des observations

- Résidu : vecteur $\mathbf{y} - \hat{\mathbf{y}}$
- La « hat matrix » ou projecteur

$$\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \quad \hat{\mathbf{y}} = \mathbf{A}\mathbf{y}$$

- Les termes diagonaux h_i

$$\frac{1}{n} \leq h_i \leq 1 \quad \sum_{i=1}^n h_i = p + 1$$

- Résidu:
 - espérance nulle, $V(y_i - \hat{y}_i) = \sigma^2(1 - h_i)$
- Résidu studentisé $\frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_i}}$
- Résidu prédit (en enlevant i) $y_i - \hat{y}_{(-i)} = \frac{y_i - \hat{y}_i}{1 - h_i}$
- **Press**: somme des carrés des résidus prédits
- Influence d'une observation sur les estimations des coefficients: la **distance de Cook**

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{(-i)})' \mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(-i)})}{(p + 1)\hat{\sigma}^2} = \frac{1}{p + 1} (\hat{y}_i - \hat{y}_{-i})^2 \frac{h_i}{1 - h_i}$$

- Devrait rester <1

2. Régression en norme L_1 (LAD)

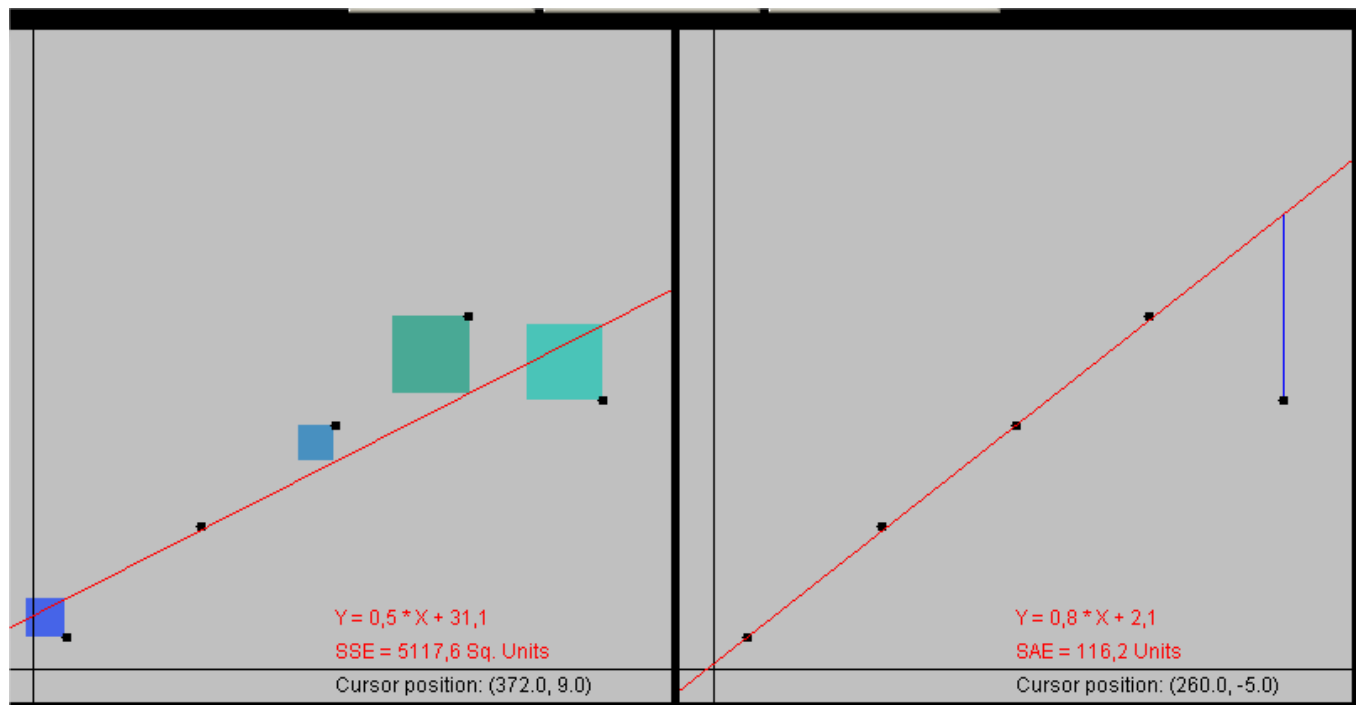
- Utiliser les valeurs absolues

$$L(y; f(x)) = |y - f(x)|$$

- Sur la population f : optimal $f(x) = \text{Med}(Y/x)$
- En régression linéaire simple sur échantillon

$$\min \sum_{i=1}^n |y_i - a - bx_i|$$

- http://www.math.wpi.edu/Course_Materials/SAS/tablets/7.3/7.3c/index.html



- 50 ans plus ancien que les moindres carrés: **Boscovitch 1757**
- Résolution plus difficile (surtout en régression multiple)
 - Pas de solution analytique
 - Nécessité d'un algorithme
- Propriété:
 - La droite de régression L_1 passe par deux des points

- **Algorithme LAD simple** (Birkes & Dodge, 1993)
 - Prendre un point (x_1, y_1) , trouver la meilleure droite passant par ce point. Elle passe par au moins un autre point noté (x_2, y_2)
 - Trouver la meilleure droite passant par (x_2, y_2) . Elle passe par noté (x_3, y_3)
 - Continuer jusqu' à ce que $(x_k, y_k) = (x_{k-1}, y_{k-1})$
- **Possibilité de non-unicité ou de dégénérescence**
- **Régression LAD multiple**
 - Programmation linéaire
 - $p+1$ résidus sont nuls

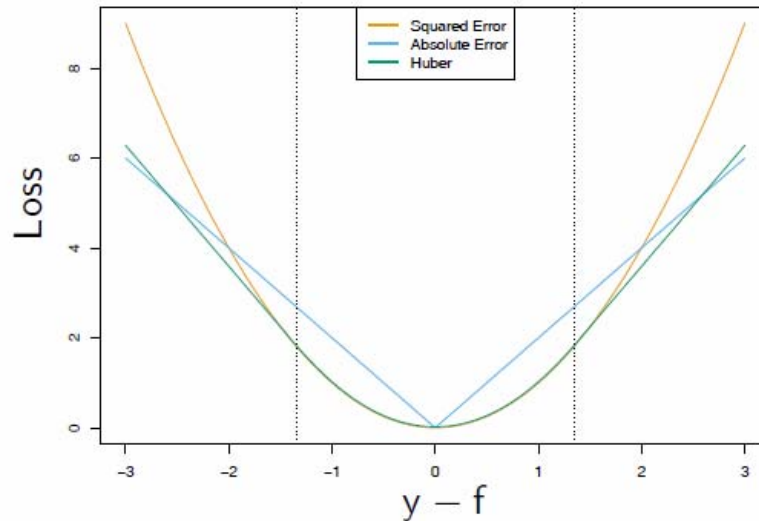
- Erreurs standard asymptotiques

$$V(\hat{\boldsymbol{\beta}}) \simeq \frac{1}{4(f'(0))^2} (\mathbf{X}'\mathbf{X})^{-1}$$

avec f densité de ε

3.M-régression

- Issue des M-estimateurs de **Huber**
- Exemple: fonction de perte quadratique jusqu'à c , linéaire au delà



©Hastie, Tibshirani & Friedman 2009 Chap 10

- Pour c grand, on retrouve les mco, pour $c=0$ la régression L_1
- Si $\sigma = 1$, $c=1.5$ donne une estimation d'efficacité asymptotique 95% dans le cas gaussien

- $\min \sum_{i=1}^n \rho(y_i - \mathbf{x}_i \boldsymbol{\beta})$ ρ fonction convexe paire

– M-estimateur: maximum de vraisemblance avec erreurs de densité proportionnelle à $\exp(-\rho(u))$

- En dérivant: $\sum_{i=1}^n \rho'(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) \mathbf{x}'_i = 0$
- Notation usuelle $\psi = \rho'$
- Moindres carrés pondérés

$$\sum_{i=1}^n w_i (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}) \mathbf{x}'_i = 0 \quad \text{avec } w_i = \frac{\rho'(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}})}{y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}}$$

Propriétés

- Si σ est inconnu on minimise $\sum_{i=1}^n \left[\rho \left(\frac{y_i - \mathbf{x}_i \boldsymbol{\beta}}{\sigma} \right) + a \right] \sigma$
avec $a > 0$
- La matrice de covariance des estimateurs est asymptotiquement proportionnelle à celle des moindres carrés

Asymptotic Covariance and Confidence Intervals

The following three estimators of the asymptotic covariance of the robust estimator are available in PROC ROBUSTREG:

$$\text{H1: } K^2 \frac{[1/(n-p)] \sum (\psi(r_i))^2}{[(1/n) \sum (\psi'(r_i))]^2} (X^T X)^{-1}$$

$$\text{H2: } K \frac{[1/(n-p)] \sum (\psi(r_i))^2}{[(1/n) \sum (\psi'(r_i))]} W^{-1}$$

$$\text{H3: } K^{-1} \frac{1}{(n-p)} \sum (\psi(r_i))^2 W^{-1} (X^T X) W^{-1}$$

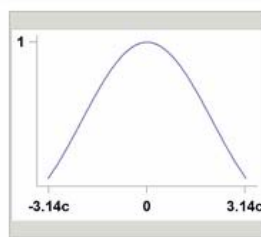
where $K = 1 + \frac{p}{n} \frac{\text{Var}(\psi')}{(E\psi')^2}$ is a correction factor and $W_{jk} = \sum \psi'(r_i) x_{ij} x_{ik}$. Refer to Huber (1981, p. 173) for more details.

4. Fonction de poids

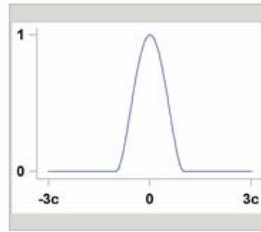
Proc ROBUSTREG de SAS

Weight Function	Option	Default a, b, c
andrews	WF=ANDREWS<(C=c)>	1.339
bisquare	WF=BISQUARE<(C=c)>	4.685
cauchy	WF=CAUCHY<(C=c)>	2.385
fair	WF=FAIR<(C=c)>	1.4
hampel	WF=HAMPEL<(<A=a> <B=b> <C=c)>	2, 4, 8
huber	WF=HUBER<(C=c)>	1.345
logistic	WF=LOGISTIC<(C=c)>	1.205
median	WF=MEDIAN<(C=c)>	0.01
talworth	WF=TALWORTH<(C=c)>	2.795
welsch	WF=WELSCH<(C=c)>	2.985

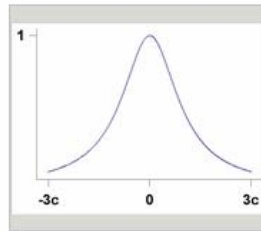
$$\text{andrews } W(x, c) = \begin{cases} \frac{\sin(\frac{x}{c})}{\frac{x}{c}} & \text{if } |x| \leq \pi c \\ 0 & \text{otherwise} \end{cases}$$



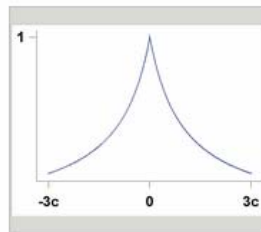
$$\text{bisquare } W(x, c) = \begin{cases} (1 - (\frac{x}{c})^2)^2 & \text{if } |x| < c \\ 0 & \text{otherwise} \end{cases}$$



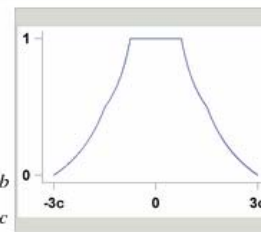
$$\text{cauchy } W(x, c) = \frac{1}{1 + (\frac{|x|}{c})^2}$$



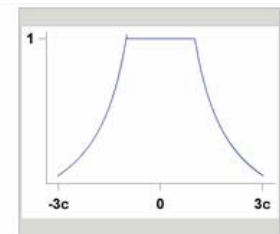
$$\text{fair } W(x, c) = \frac{1}{(1 + (\frac{|x|}{c})^3)}$$



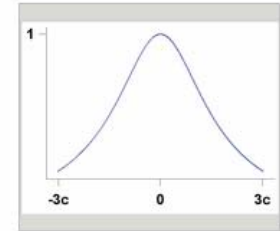
$$\text{hampel } W(x, a, b, c) = \begin{cases} 1 & |x| < a \\ \frac{a}{|x|} & a < |x| \leq b \\ \frac{a}{|x|} \frac{c-|x|}{c-b} & b < |x| \leq c \\ 0 & \text{otherwise} \end{cases}$$



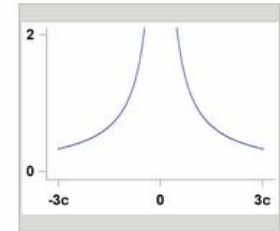
$$\text{huber } W(x, c) = \begin{cases} \frac{1}{c} & \text{if } |x| < c \\ \frac{1}{|x|} & \text{otherwise} \end{cases}$$



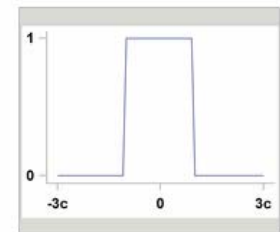
$$\text{logistic } W(x, c) = \frac{\tanh(\frac{x}{c})}{c}$$



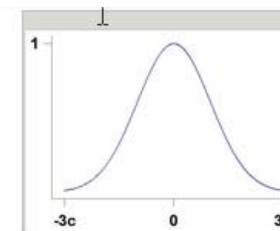
$$\text{median } W(x, c) = \begin{cases} \frac{1}{c} & \text{if } x = 0 \\ \frac{1}{|x|} & \text{otherwise} \end{cases}$$



$$\text{talworth } W(x, c) = \begin{cases} 1 & \text{if } |x| < c \\ 0 & \text{otherwise} \end{cases}$$



$$\text{welsch } W(x, c) = \exp(-\frac{1}{2}(\frac{x}{c})^2)$$



5. Régression LTS (Least Trimmed Squares) de Rousseuw

- Basée sur le sous ensemble de h individus (parmi n) où les mco donnent la plus petite somme des carrés des résidus. h est choisi entre $n/2$ et n . La valeur $h=(3n+p+1)/4$ est recommandée.

Complément: notion de « point de rupture » d'un estimateur

- Fraction des données qui peuvent être arbitrairement changées sans changer arbitrairement la valeur de l'estimateur.
- Deux cas : n fini, n infini (pt de rupture asymptotique)
 - Ne peut être > 0.5
 - Asymptotiquement:
 - Nul pour la moyenne (si une valeur devient infinie, la moyenne aussi)
 - 0.5 pour la médiane
 - n fini
 - $1/n$ pour la moyenne, $(n-1)/2n$ pour la médiane, $(n-h)/h$ pour la régression LTS

Bibliographie

- Birkes, D., Dodge, Y. (1993) *Alternative methods of regression*, Wiley
- Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley & Sons, Inc.

Régression non-paramétrique

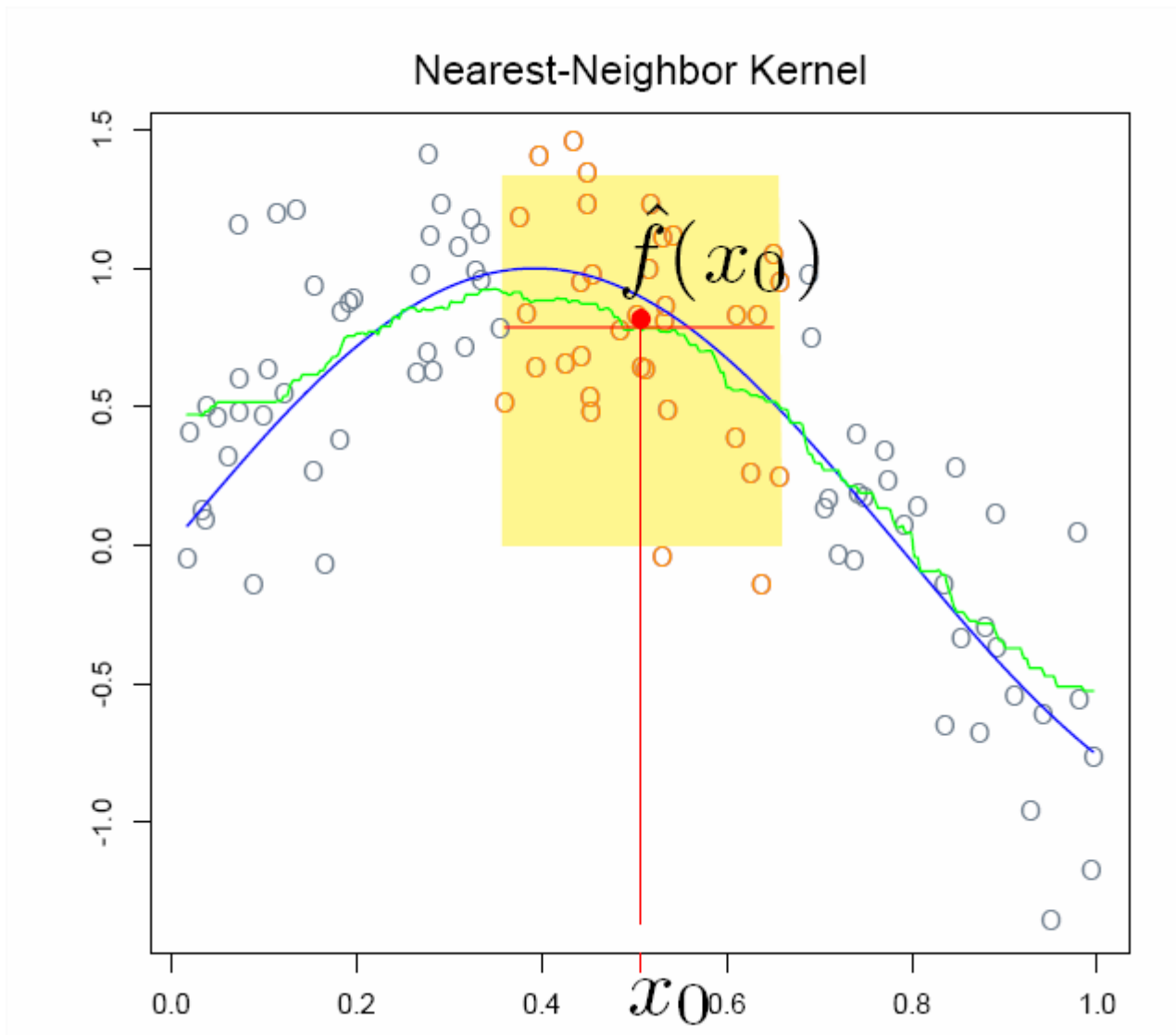
Gilbert Saporta

Conservatoire National des Arts et Métiers

<http://cedric.cnam.fr/~saporta>

Janvier 2011

- Estimation de l'espérance conditionnelle $E(Y/x_0) = f(x_0)$, ou fonction de régression.
 - Approche similaire à l'estimation de densité par la méthode du noyau
 - Premières tentatives inspirées des moyennes mobiles:
 - les k plus proches voisins: moyenne de y pour les k-ppv de x_0
 - moyenne des y sur une fenêtre de largeur fixe centrée sur x_0
- Inconvénient majeur: discontinuité**



©Hastie, Tibshirani & Friedman 2009 Chap 6

Méthode de la fenêtre mobile

- Moyenne des y_i dans un voisinage autour de x_0 : $[x_0-h/2; x_0+h/2]$

$$\hat{f}(x_0) = \frac{\sum_{i=1}^n y_i \mathbf{1}_{[x_0-h/2; x_0+h/2]}(x_i)}{\sum_{i=1}^n \mathbf{1}_{[x_0-h/2; x_0+h/2]}(x_i)}$$

$$\hat{E}(Y / X = x_0) = \frac{\sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right)}$$

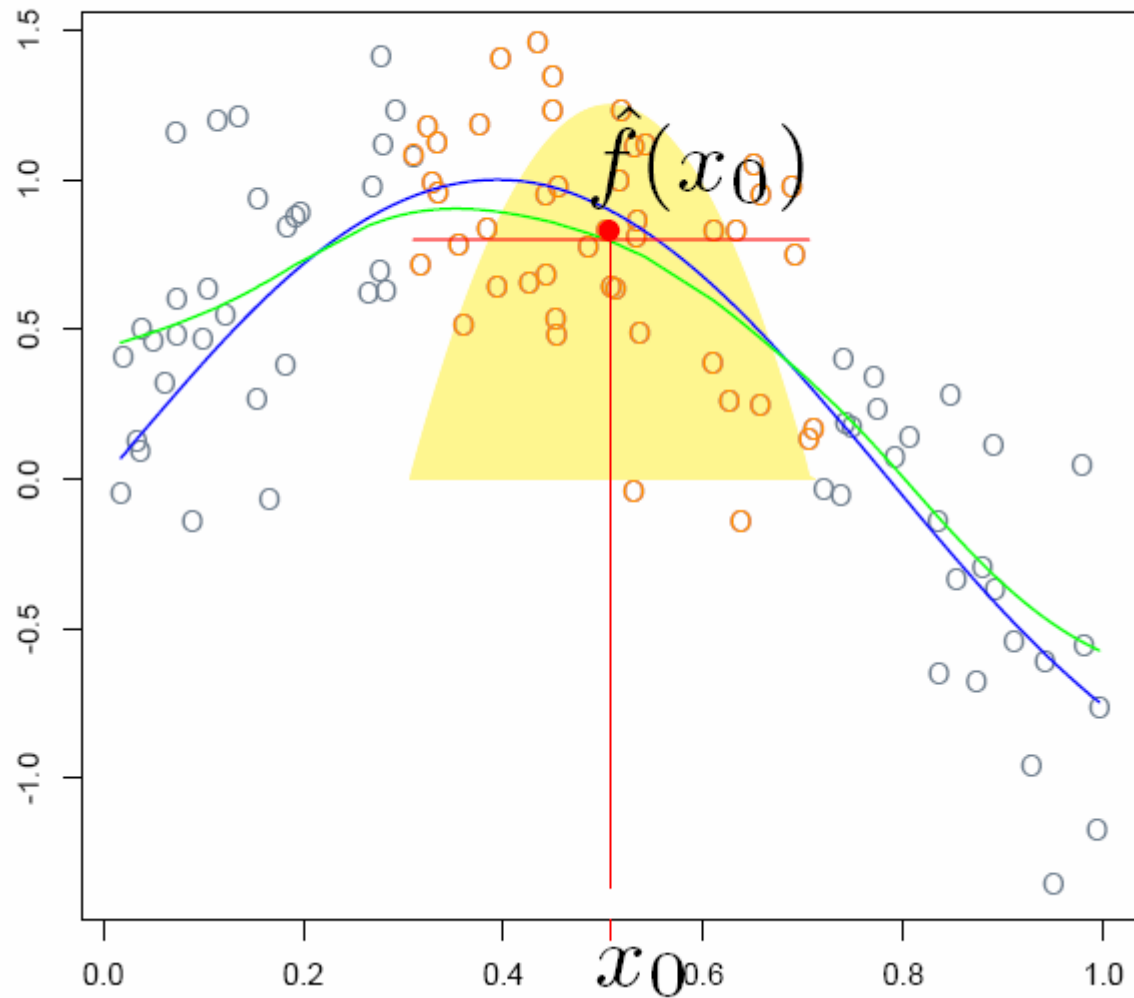
Utilisation d'un noyau continu: l'estimateur de **Nadaraya-Watson**

Noyaux classiques:

– Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)$ si $|u| \leq 1$, 0 sinon

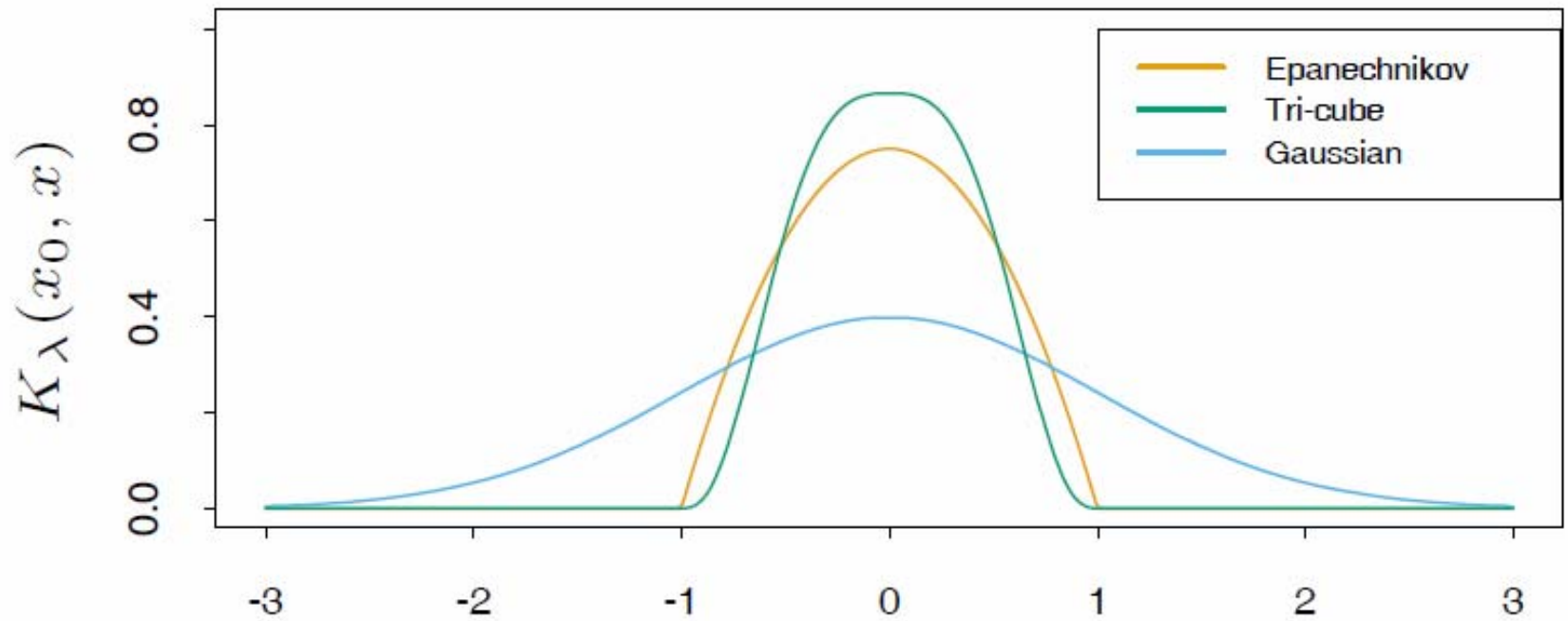
– Tricube $K(u) = \left(1 - |u|^3\right)^3$ si $|u| \leq 1$, 0 sinon

Epanechnikov Kernel



©Hastie, Tibshirani & Friedman 2009 Chap 6

- <http://www.cs.uic.edu/~wilkinson/Applets/smoothers.html>



©Hastie, Tibshirani & Friedman 2009 Chap 6

- Biais et variance pour des x_i fixés

$$w_i = K\left(\frac{x_0 - x_i}{h}\right)$$

$$E\left(\hat{f}(x_0) - f(x_0)\right) = \frac{\sum_{i=1}^n w_i (f(x_i) - f(x_0))}{\sum_{i=1}^n w_i}$$

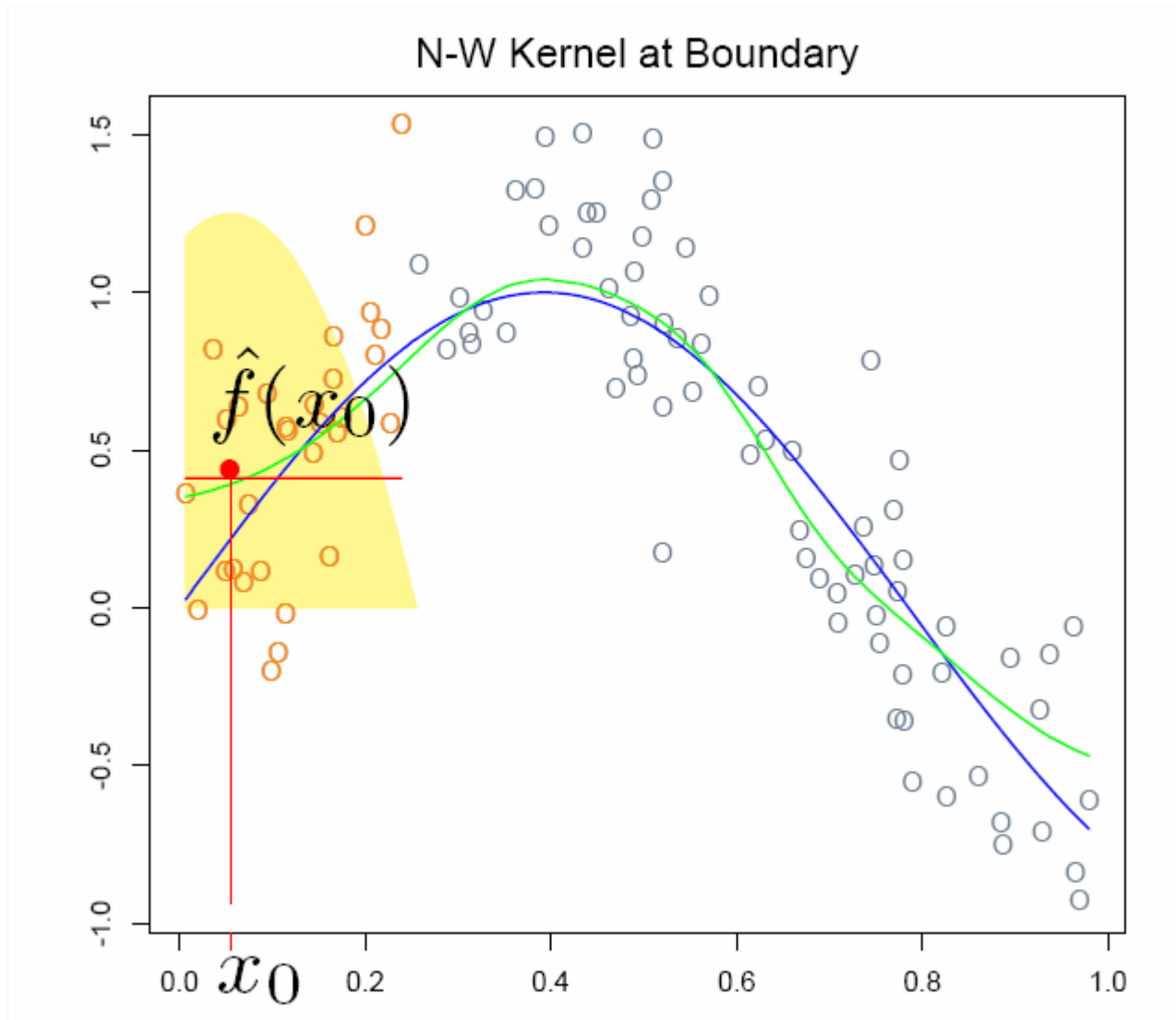
$$V\left(\hat{f}(x_0)\right) = \sigma^2 \frac{\sum_{i=1}^n w_i^2}{\left(\sum_{i=1}^n w_i^2\right)^2}$$

Si les x_i sont nombreux et équidistants sur $[a;b]$

$$\begin{aligned} E\left(\hat{f}(x_0) - f(x_0)\right) &\simeq \int K(u) (f(x_0 + uh) - f(x_0)) du \\ &\simeq \frac{h^2}{2} g''(x_0) \int u^2 K(u) du \end{aligned}$$

Valable si la fenêtre est incluse dans $[a;b]$

- Problèmes d'estimation aux bornes



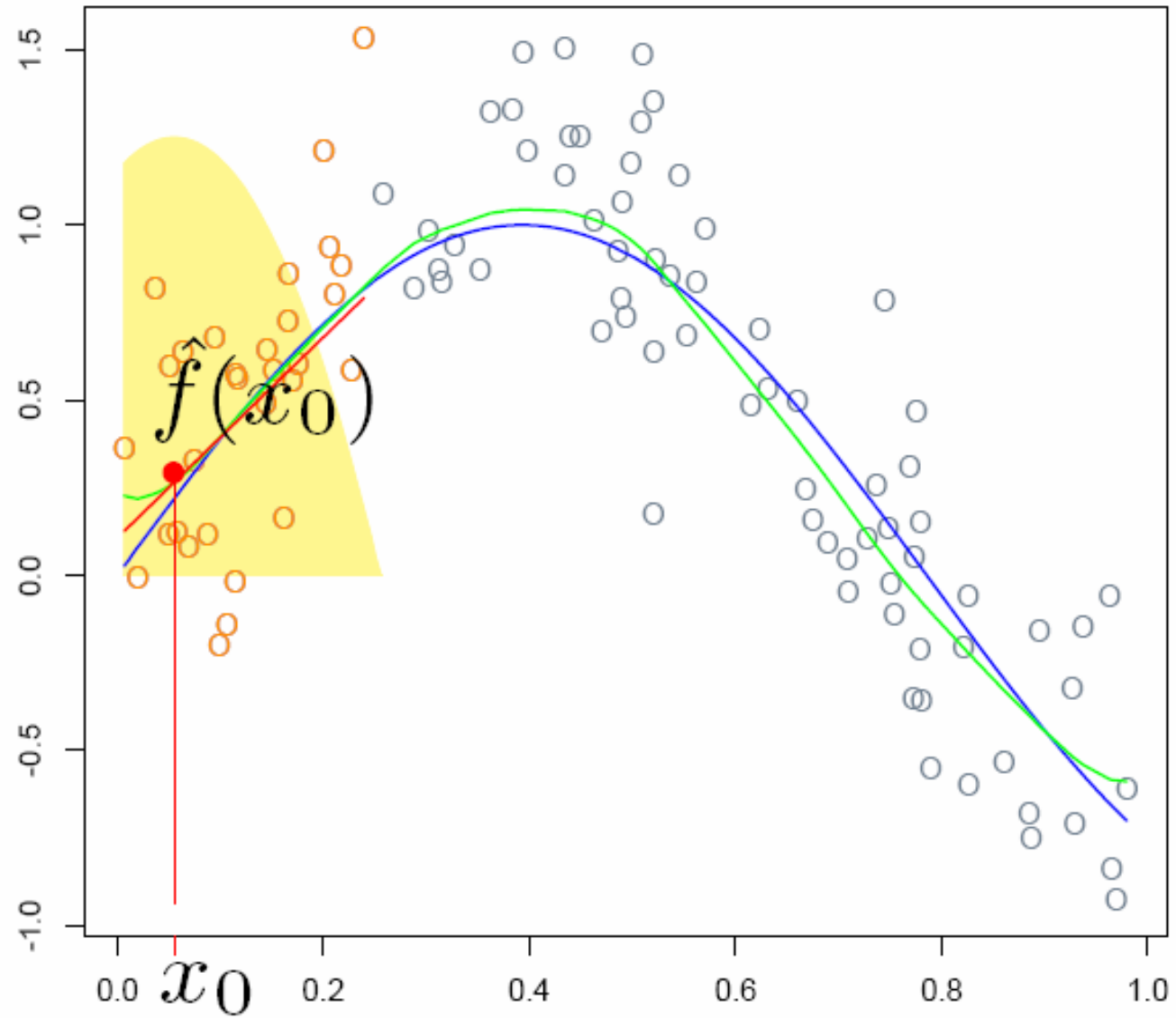
- Régression linéaire locale

- Résout le problème de l'asymétrie du noyau tronqué aux bornes. Enlève le biais à l'ordre 1
- On résout en chaque point x_0 le problème de moindres carrés pondérés:

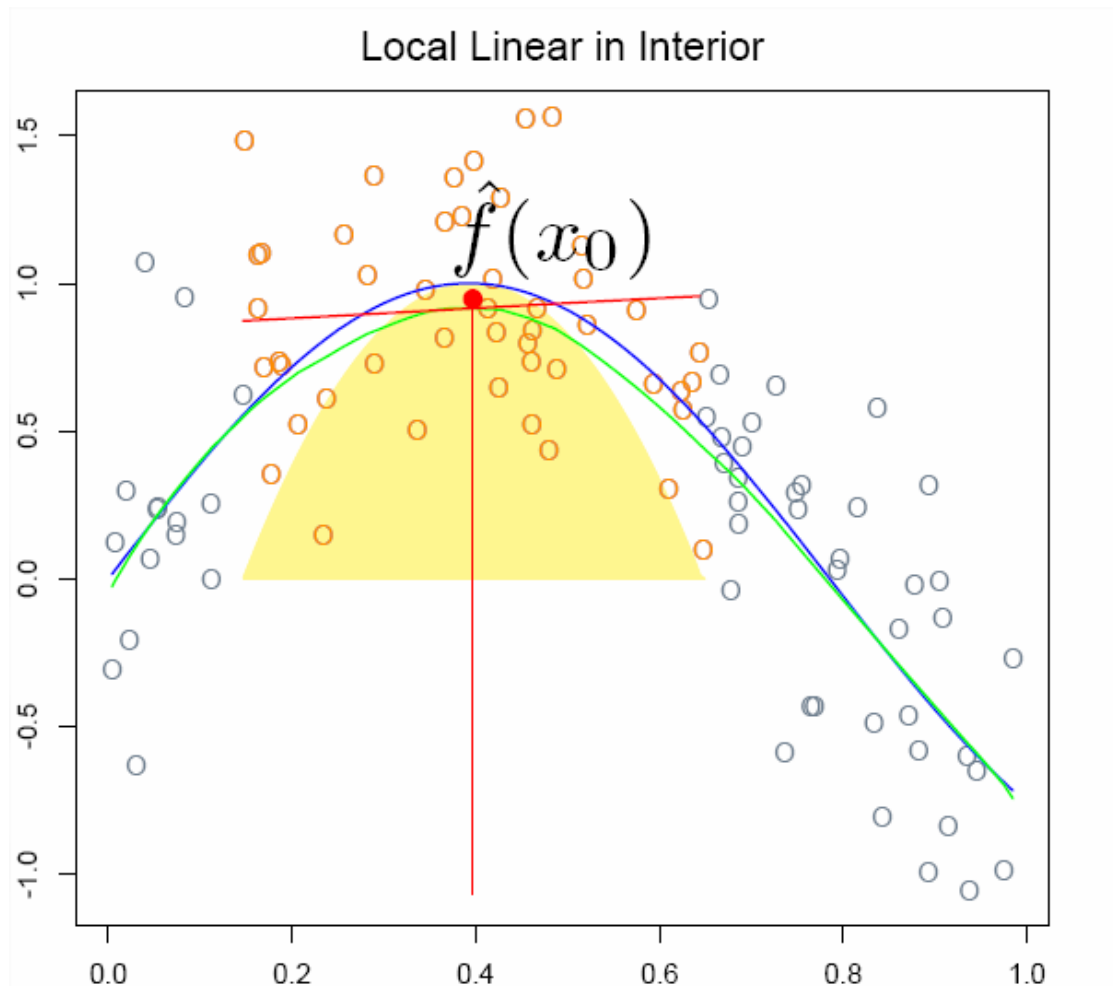
$$\min_{\alpha(x_0), \beta(x_0)} \sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right) [y_i - \alpha(x_0) - \beta(x_0)x_i]^2$$

- Nota: formules globales (pour tout x), mais chacune utilisée seulement en x_0

Local Linear Regression at Boundary

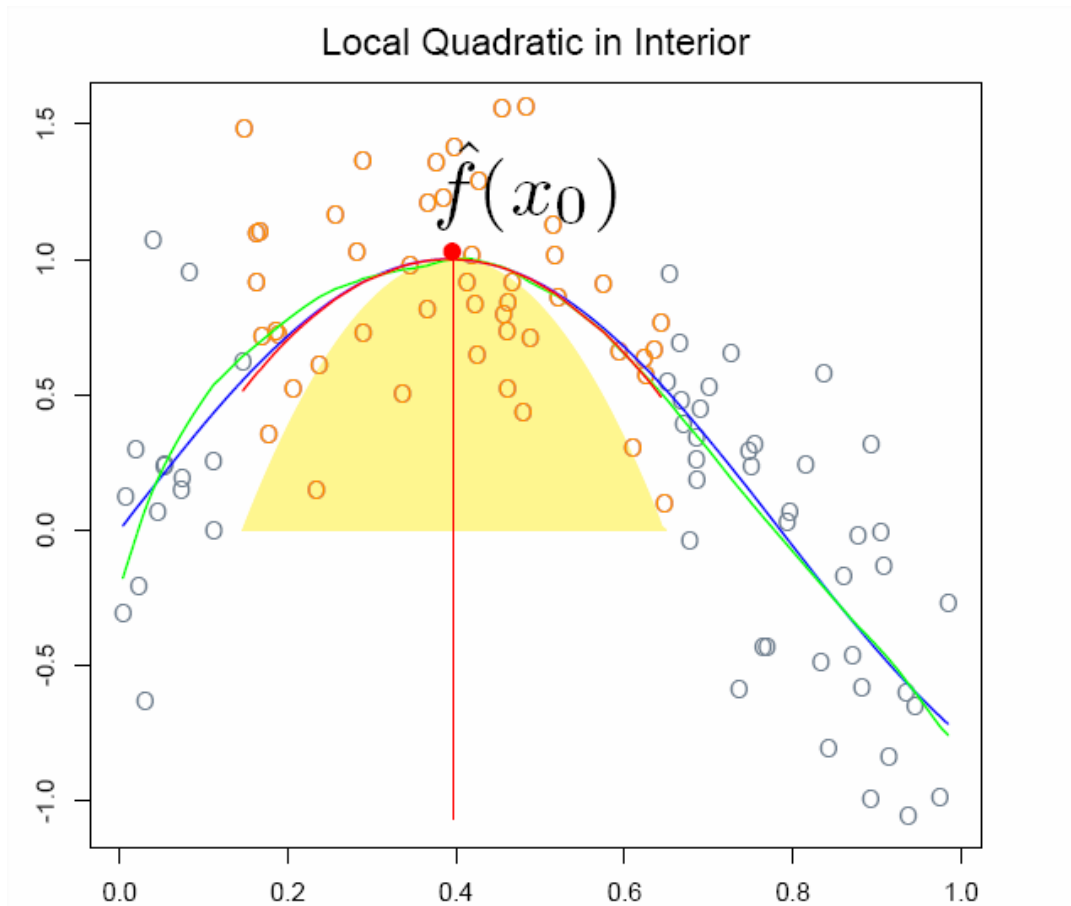


- Les creux et les bosses
 - La régression linéaire locale est biaisée dans les zones de courbure forte



- Régression polynomiale locale

$$\min_{\alpha(x_0), \beta(x_0)} \sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right) \left[y_i - \alpha(x_0) - \sum_{j=1}^d \beta_j(x_0) x_i^j \right]^2$$



- Estimation linéaire et noyau équivalent
 - moindres carrés pondérés
 - \mathbf{X} matrice à n lignes et $d+1$ colonnes des x^j

$\mathbf{W}(x_0)$ matrice diagonale n, n de terme $K\left(\frac{x_0 - x_i}{h}\right)$

$$\begin{aligned}\hat{f}(x_0) &= \mathbf{x}_0' (\mathbf{X}' \mathbf{W}(x_0) \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}(x_0) \mathbf{y} \\ &= \sum_{i=1}^n l_i(x_0) y_i\end{aligned}$$

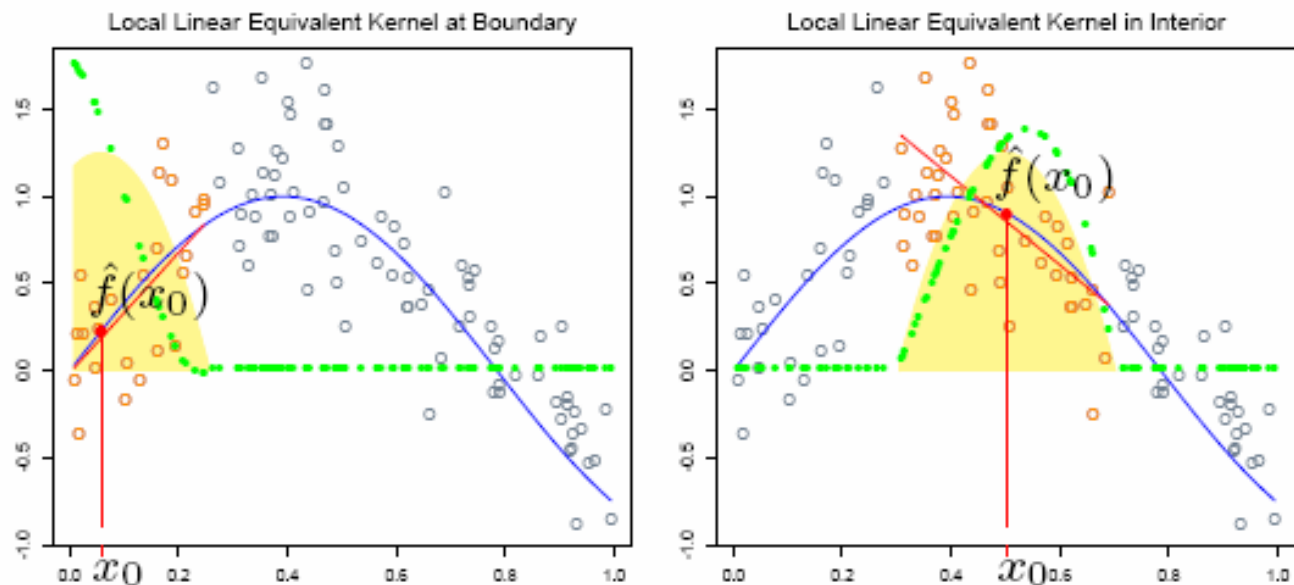
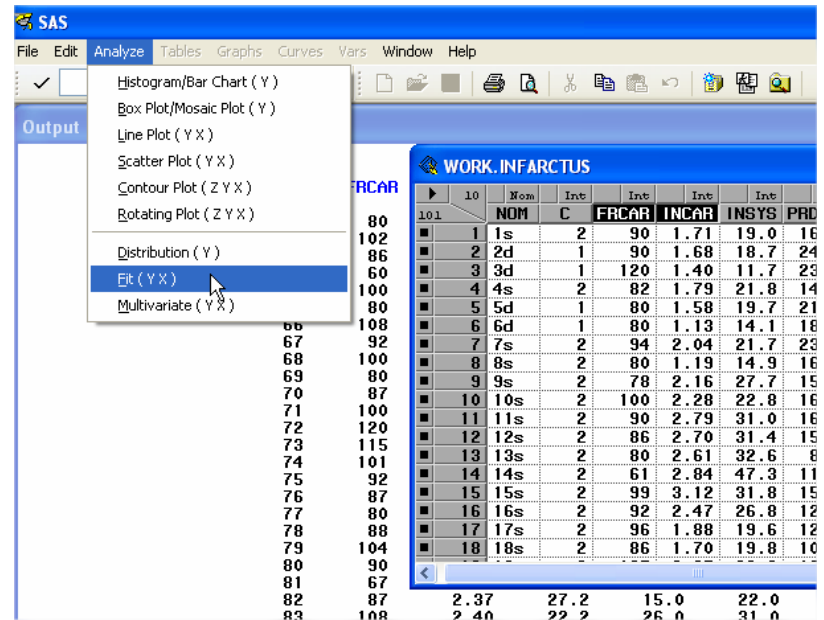
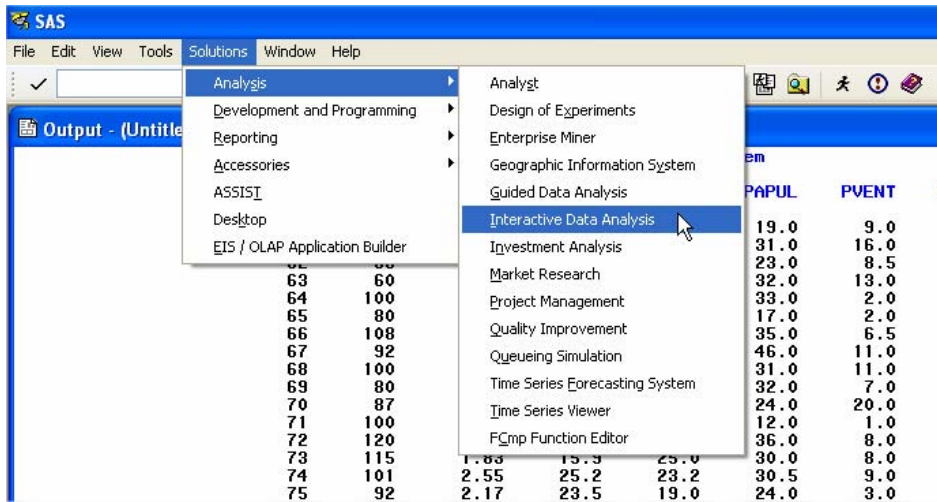


FIGURE 6.4. The green points show the equivalent kernel $l_i(x_0)$ for local regression. These are the weights in $\hat{f}(x_0) = \sum_{i=1}^N l_i(x_0)y_i$, plotted against their corresponding x_i . For display purposes, these have been rescaled, since in fact they sum to 1. Since the yellow shaded region is the (rescaled) equivalent kernel for the Nadaraya–Watson local average, we see how local regression automatically modifies the weighting kernel to correct for biases due to asymmetry in the smoothing window.

- Choix de h
 - validation croisée
- Méthode proche: LOESS ou LOWESS
- Extension possible à la régression logistique

SAS INSIGHT

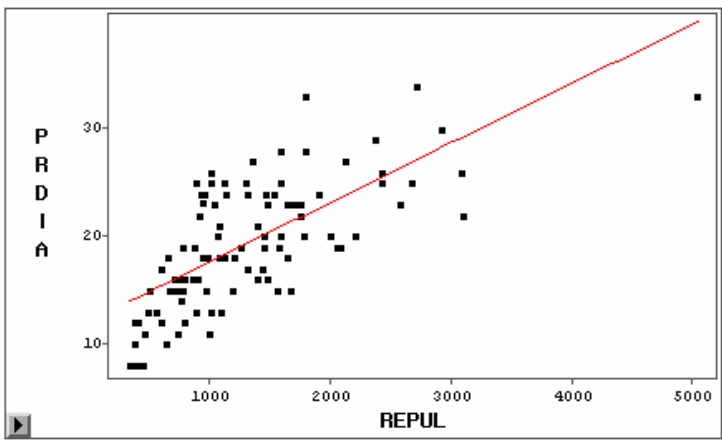


WORK.INFARCTUS

	12	Nom	Int	Int	Int	Int	Int	Int
101	NOM	C	FRCAR	INCAR	INSYS	PRDIA	PAPUL	P
30	1	1s	2	90	1.71	19.0	16.0	19.5
32	2	2d	1	90	1.68	18.7	24.0	31.0
36	3	3d	1	120	1.40	11.7	23.0	29.0
30	4	4s	2	82	1.79	21.8	14.0	17.5
30	5	5d	1	80	1.58	19.7	21.0	28.0
38	6	6d	1	80	1.13	14.1	18.0	23.5
32	7	7s	2	94	2.04	21.7	23.0	27.0
30	8	8s	2	80	1.19	14.9	16.0	21.0
30	9	9s	2	78	2.16	27.7	15.0	20.5
37	10	10s	2	100	2.28	22.8	16.0	23.0
30	11	11s	2	90	2.79	31.0	16.0	25.0
20	12	12s	2	86	2.70	31.4	15.0	23.0
15	13	13s	2	80	2.61	32.6	8.0	15.0
31	14	14s	2	61	2.84	47.3	11.0	17.0
32	15	15s	2	99	3.12	31.8	15.0	20.0
37	16	16s	2	92	2.47	26.8	12.0	19.0
30	17	17s	2	96	1.88	19.6	12.0	19.0
38	18	18s	2	86	1.70	19.8	10.0	14.0
30								
37			2.37	27.2	15.0	22.0	10.0	
38			2.40	22.2	26.0	31.0	4.0	
20			1.91	15.9	18.0	27.0	15.0	
38			1.50	13.9	28.0	43.0	16.0	
36			2.36	27.4	24.0	34.0	8.0	
12			1.56	13.9	24.0	29.0	4.0	
30			1.34	17.0	16.0	25.0	16.0	
35			1.65	17.4	20.0	33.0	7.0	
30			2.04	22.7	28.0	41.0	10.0	
30			3.03	33.6	17.0	23.5	7.0	
34			1.21	12.9	17.0	22.0	3.0	
51			1.34	26.3	11.0	17.0	6.0	
32			1.17	11.0	22.0	25.0	10.0	

PRDIA = REPUL
 Response Distribution: Normal
 Link Function: Identity

Model Equation
 PRDIA = 11.9816 + 0.0055 REPUL



Parametric Regression Fit				
Curve	Degree(Polynomial)	Model		Error
		DF	Mean Square	DF Mean
—	1	1	1660.4355	99

SAS

File Edit Analyze Tables Graphs **Curves** Vars Window Help

Output - (Untitled)

Confidence Ellipse

Confidence Curves

Polynomial...

Spline...

Kernel...

Loess...

Local Polynomial, Fixed Bandwidth...

			Int	
64	100	4	4s	2
65	80	5	5d	1
66	108	6	6d	1
67	92	7	7s	2

Loess Fit

Type:

Mean

Linear

Quadratic

Method:

GCV

DF

Alpha

Weight:

Normal

Triangular

Quadratic

Tri-Cube

DF: 3

Alpha: 0.5

Number of Intervals: 128

Convergence Criterion (DF):

Relative Difference

Absolute Difference

Relative Difference: 0.005

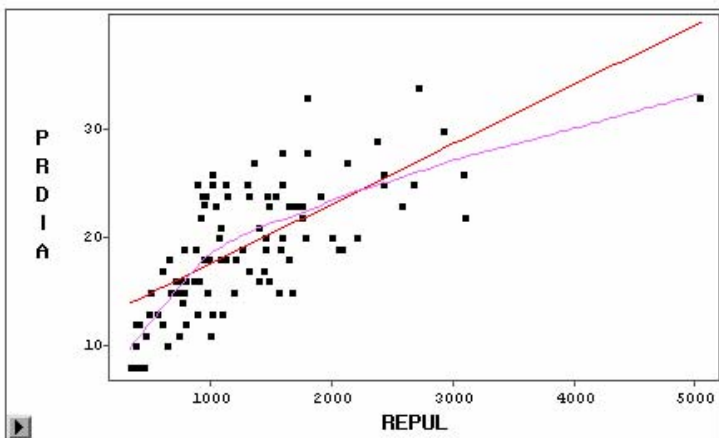
Absolute Difference: 0.05

OK Cancel



▶ PRDIA = REPUL
 Response Distribution: Normal
 Link Function: Identity

▶ Model Equation
 PRDIA = 11.9816 + 0.0055 REPUL



▶ Parametric Regression Fit									
Curve	Degree(Polynomial)	Model		Error		R-Square	F Stat	Pr > F	
		DF	Mean Square	DF	Mean Square				
	1	1	1660.4355	99	17.3173	0.4920	95.88	<.0001	

▶ Loess Fit											
Curve	Type	Weight	N Intervals	Method	Alpha		K	DF	R-Square	MSE	MSE(GCV)
	Linear	Tri-Cube	128	GCV	0.7583		76	4.339	0.5841	14.5196	15.1714

Avantages et inconvénients

- Utile si la forme de la régression est totalement inconnue
- Méthode adaptative qui s'ajuste automatiquement
- Pas de formule explicite, prévision délicate en dehors du domaine (extrapolation)

Bibliographie

- Cleveland, W.S.; Devlin, S.J. (1988). Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting . *Journal of the American Statistical Association* 83 (403): 596–610.
- Dreesbeke, J.J., Saporta G. (éditeurs) (2011) *Approches non paramétriques en régression*, Editions Technip
- Hastie, T., Tibshirani, R., Friedman, J. (2009): *The Elements of Statistical Learning* , 2nd edition, chapitre 6, Springer, <http://www-stat.stanford.edu/~hastie/Papers/ESLII.pdf>
- Lejeune, M. (1985), Estimation non paramétrique par noyaux : régression polynomiale mobile, *Revue de Statistique Appliquée*, 33, 43-68.