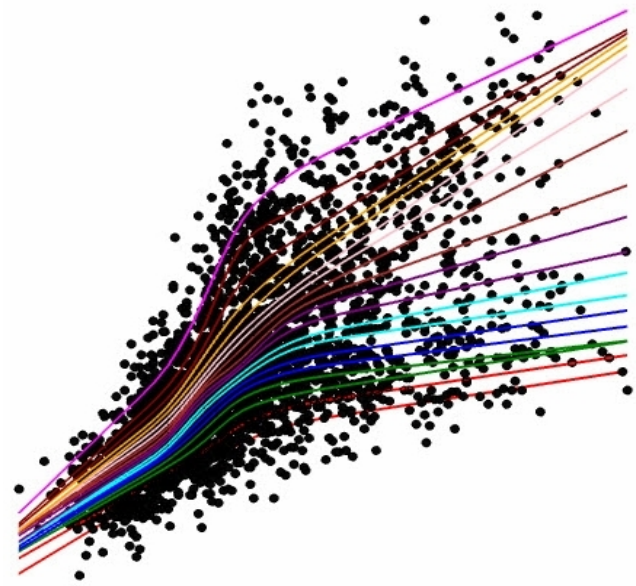
Méthodes statistiques pour l'ingénierie financière

Régression quantile non paramétrique localement linéaire

Ruijie YUAN et Fabrice DURAND – Master 2 de statistique mathématique



Année 2012-2013

Table des matières

1. Introduction à la régression quantile

[1] Lorsque l'on dispose d'un jeu d'observations $\{(X_i, Y_i)\}_{1\leq i\leq n} \in (\mathbb{R}^p \times \mathbb{R})^n$, on peut selon la forme du nuage de points ou d'après une expertise chercher une explication linéaire, quadratique ou autre de $\{Y_i\}_{1\leq i\leq n}$ en fonction de $\{X_i\}_{1\leq i\leq n}$ par exemple. La méthode la plus connue est celle des **moindres carrés** consistant à minimiser la somme des carrés des erreurs d'estimation. Si l'on considère des variables aléatoires au lieu de jeux de données, cette méthode consiste à estimer $E[Y|X=x]$ par la valeur de $a\in\mathbb{R}$ qui minimise $E[(Y-a)^2|X=x]$. Cette méthode est facile à mettre en œuvre et conduit à une erreur d'estimation de moyenne ou d'espérance nulle. Par contre, elle a l'inconvénient de ne pas être robuste aux valeurs extrêmes ou aberrantes, tout comme un calcul de moyenne est sensible aux valeurs extrêmes. A contrario, dans un jeu de données, un calcul de médiane ou d'un certain quantile n'est pas sensible aux valeurs extrêmes puisque ces dernières ne changent pas les pourcentages de données supérieures et inférieures aux dits quantiles. Il en sera de même pour la régression quantile, dont la première idée est d'avoir une erreur d'estimation non plus de moyenne nulle mais de médiane nulle. On peut alors généraliser à des régressions quantiles dont les courbes ne passeraient nécessairement pas par la médiane par un certain quantile. Ainsi, la régression quantile autour de la médiane demande de minimiser $E[|Y-a||X=x]$, alors que celle autour d'un quantile d'ordre p implique la fonction « check » définie par :
$\forall u\in\mathbb{R}, \rho_p(u)=p.u.I_{[0,+\infty[}(u)-(1-p).u.I_{]-\infty,0[}(u)$ et implique la minimisation de $E[\rho_p(Y-a)|X=x]$.

2. Deux approches non paramétriques de la régression quantile localement linéaire [2]

2.1. Lissage par simple noyau

a) La méthode

Le quantile d'ordre p de Y conditionnellement à $X=x$ est donné par :

$q_p(x)=\underset{a\in\mathbb{R}}{argmin}\, E[\rho_p(Y-a)|X=x]$. L'idée d'une régression quantile localement linéaire est d'approcher le quantile le quantile inconnu $q_p(x)$ par par une fonction affine $q_p(t)\approx q_p(x)+q_p'(x).(t-x)\approx a+b.(t-x)$ pour t dans un voisinage de x . Ceci conduit à définir un estimateur de $q_p(x)$ en posant $\hat{q}_p(x)=\hat{a}$ avec :

$(\hat{a},\hat{b}):=\underset{a,b}{argmin}\sum_{i=1}^{n}\rho_p\left(Y_i-a-b(X_i-x)\right)K\left(\frac{x-X_i}{h}\right)$ où $K(.)$ est un noyau et h une largeur de fenêtre associée à ce noyau et au jeu de données.

Pour calculer $\hat{q}_p$, les auteurs ont utilisé un algorithme itératif basé sur pondération ajustée des moindres carrés, les détails étant fournis dans le manuscrit de thèse de Yu (1997).

b) L'erreur quadratique moyenne

Sous certaines conditions, entre autres que $q_p(x)$ soit suffisamment lisse et que x ne soit pas trop près des bords, l'erreur quadratique moyenne de l'estimateur $\hat{q}_p(x)$ est donnée par : $MSE\left(\hat{q}_p(x)\right)\approx\frac{1}{4}h^4\mu_2(K)^2q_p''(x)^2+\frac{R(K)p(1-p)}{n.h.g(x)f(q_p(x)|x)^2}$ où :

$\mu_2(K)=\int u^2K(u)du$, $R(K)=\int K(u)du$ et g est la densité marginale de X .

Ainsi, pour un point proche de la frontière $x=c.h$ avec $0<c<1$, et si K et g ont pour support respectif $[-1,1]$ et $[0,1]$, alors l'erreur pour ce point est :

$MSE\left(\hat{q}_p(c.h)\right)\approx\frac{1}{4}h^4\alpha_c^2(K)q_p''(0^+)^2+\frac{\beta_c(K)p(1-p)}{n.h.g(0^+)f(q_p(0^+)|0^+)^2}$ ou :

$\alpha_c(K)=\frac{a_2^2(c,K)-a_1(c,K)a_3(c,K)}{a_0(c,K)a_2(c,K)-a_1^2(c,K)}$, $\beta_c(K)=\frac{\int_{-1}^{c}\left[a_2(c,K)-a_1(c,K)u\right]^2K(u)du}{\left[a_0(c,K)a_2(c,K)-a_1^2(c,K)\right]^2}$

et $\forall l=1,2,a_l(c,K)=\int_{-1}^{c}u^lK(u)du$, en notant $g(0^+)=\lim_{z\to 0}g(z)$

Avantages : Ces expressions de l'erreur quadratique moyenne reflètent deux avantages :

- l'erreur ne dépend pas de g, la densité marginale X.
- elle présente un bon comportements aux frontières, sans besoin de corrections.

c) Choix de la largeur de fenêtre

Le point de départ est la relation fournissant une largeur de fenêtre asymptotiquement optimale, pour certaine valeur de p : $h_p^5 = \frac{R(K)\,p(1-p)}{n\mu_2(K)^2 (q_p''(x))^2 g(x) f(q_p(x)|x)^2}$.

Ceci fournit un rapport de largeurs de fenêtres pour deux valeurs différentes de p :

$$\left(\frac{h_{p_1}}{h_{p_2}}\right)^5 = \frac{p_1(1-p_1)}{p_2(1-p_2)} \frac{(q''_{p_2}(x))^2 g(x) f(q_{p_2}(x)|x)^2}{(q''_{p_1}(x))^2 g(x) f(q_{p_1}(x)|x)^2} \quad .$$

Maintenant, en considérant φ et Φ la densité et la fonction de répartition de loi normale standard, on peut montrer que : $\left(\frac{h_{p_1}}{h_{p_2}}\right)^5 = \frac{p_1(1-p_1)}{p_2(1-p_2)} \frac{\varphi(\Phi^{-1}(p_2))}{\varphi(\Phi^{-1}(p_1))}$.

En particulier si $p_2 = 1/2$, alors : $h_p^5 = \pi^{-1} 2p(1-p)\varphi(\Phi^{-1}(p))^{-2} h_{1/2}^5$ (1).

Par ailleurs, en utilisant la méthode Fan (1993) [4] pour choisir h_{mean} , nous avons :

$h_{mean}^5 = \frac{R(K)\sigma^2(x)}{n\mu_2(K)^2 (m''(x))^2 g(x)}$ (2) où $\sigma^2(x)$ et $m(x)$ sont l'espérance et la variance conditionnelle. On en déduit : $\left(\frac{h_{mean}}{h_{1/2}}\right)^5 = \frac{2}{\pi}$ (3).

Finalement, les relations (1), (2) et (3) fournissent : $h_p = h_{mean}\left[\frac{p(1-p)}{\varphi^2\left(\Phi^{-1}(p)\right)}\right]^{1/5}$

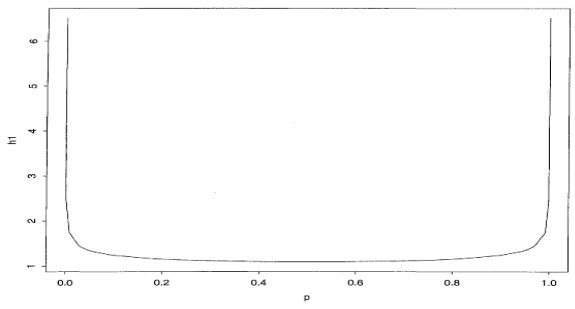


Fig. 1 – Evolution de h_p/h_{mean} en fonction de p.

La figure ci-contre montre que rapport h_p/h_{mean} , et donc également h_p , est minimal pour $p=1/2$. La courbe est également symétrique autour de $p=1/2$, conformément à l'intuition. h_p varie très peu (en augmentant) lorsque p s'éloigne modérément de ½, mais tend vers l'infini lorsque p se rapproche de 0 ou 1.

2.2. Lissage par double noyau

a) La méthode

Introduisons un deuxième noyau W à densité symétrique et Ω la fonction de répartition associée et notons que : $\int_{-\infty}^{y} W_{h_2}(Y_j-u)\,du=\Omega\left(\frac{y-Y_j}{h_2}\right)$. Ainsi, quand la largeur de fenêtre s'approche de zéro, nous avons l'approximation : $E\left[\Omega\left(\frac{y-Y}{h_2}\right)\middle| X=x\right]\underset{h_2\to 0}{\approx} F(y|x)$, où F est la fonction de répartition de Y .

De plus, une approche localement linéaire est suggérée par l'approximation :

$$E\left[\Omega\left(\frac{y-Y}{h_2}\right)\middle| X=t\right]\underset{h_2\to 0}{\approx} F(y|t)\underset{t\to x}{\approx} F(y|x)+\frac{\partial F(y|x)}{\partial x}(t-x)=a+b.(t-x)$$

Alors, on définit $\tilde{F}_{h_1,h_2}(y|x)=\tilde{a}$ ou

$(\tilde{a},\tilde{b})=\underset{a,b}{argmin}\sum_{i=1}^{n}\left(\Omega\left(\frac{y-Y_i}{h_2}\right)-a-b(X_i-x)\right)^2\times K\left(\frac{x-X_i}{h_1}\right)$. On obtient alors l'écriture explicite :

$$\tilde{F}_{h_1,h_2}(y|x)=\frac{\sum_j w_j(x,h_1)\,\Omega\left(\frac{\tilde{y}-Y_j}{h_2}\right)}{\sum_j w_j(x,h_1)}$$ avec :

$w_j(x,h_1)=K\left(\frac{x-X_j}{h_1}\right)\left[S_{n,2}-(x-X_j)S_{n,1}\right]$ où $\forall k=1,2,\ S_{n,k}=\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_1}\right)(x-X_i)^k$

Ainsi, pour retourner a l'estimation conditionnelle, on définit $\tilde{q}_p(x)$ qui satisfait $\tilde{F}_{h_1,h_2}(\tilde{q}_p(x)|x)=p$ de sorte que $\tilde{q}_p(x)=\tilde{F}^{-1}_{h_1,h_2}(p|x)$ (*).

Cependant, très occasionnellement, $\tilde{F}_{h_1,h_2}(y|x)$ peut ne pas être monotone partout en y .

Pour pallier à cette difficulté, les auteurs sélectionnent $\tilde{q}_{1/2}(x)=\tilde{F}^{-1}_{h_1,h_2}(1/2|x)$. Ensuite, pour $p>1/2$, ils choisissent pour $\tilde{q}_p(x)$ la plus grande valeur vérifiant (*), de même pour $p<1/2$, ils choisissent pour $\tilde{q}_p(x)$ la plus petite valeur vérifiant (*). La monotonie est alors assurée.

Avantage :

Alors que deux fonctions quantiles obtenues par la première méthode $\hat{q}_{p_1}$ et $\hat{q}_{p_2}$ peuvent se croiser, l'algorithme utilise pour la deuxième méthode assure que ce comportement ne se produit pas pour $\tilde{q}_{p_1}$ et $\tilde{q}_{p_2}$. A contrario cette méthode a l'inconvénient de devoir choisir une seconde largeur de fenêtre h_2 .

b) Erreur quadratique moyenne

Sous certaines conditions de régularité et en s'assurant que le point x ne se trouve pas sur un bord, nous avons le résultat suivant :

$$MSE(\tilde{q}_p(x)) \underset{\substack{h_1,h_2\to 0 \\ nh_1\to\infty}}{\approx} \frac{1}{4}\left[\frac{\mu_2(K)h_1^2F^{20}(q_p(x)|x)+\mu_2(W)h_2^2F^{02}(q_p(x)|x)}{f(q_p(x)|x)}\right]^2 + \frac{R(K)}{nh_1g(x)f^2(q_p(x)|x)}(p(1-p)-h_2f(q_p(x)|x)\alpha(W))+o(h_1^4+h_2^4+h_2/nh_1)$$

,

où :

$$F^{ab}(q_p(x)|x)=\frac{\partial^{ab}}{\partial z^a\partial y^b}F(y|z)\,avec\,(z=x\,et\,y=q_p(x))$$

$$\alpha(W)=\int\Omega(t)(1-\Omega(t))dt \quad .$$

Quant au comportement aux bords, pour des points situés près de la frontière gauche $x=ch_1, 0<c<1$, nous avons avec les mêmes conditions et en supposant de plus q_p bornée sur $[0,1]$ et continue à droite en 0, le résultat suivant :

$$MSE(\tilde{q}_p(ch_1)) \underset{\substack{h_1,h_2\to 0 \\ nh_1\to\infty}}{\approx} \frac{1}{4}\left[\frac{\mu_2(K)h_1^2F^{20}(q_p(0^+)|0^+)+\mu_2(W)h_2^2F^{02}(q_p(0^+)|0^+)}{f(q_p(0^+)|0^+)}\right]^2 + \frac{R(K)}{nh_1g(0^+)f^2(q_p(0^+)|0^+)}(p(1-p)-h_2f(q_p(0^+)|0^+)\alpha(W))+o(h_1^4+h_2^4+h_2/nh_1)$$

Cette erreur est utile au choix de la deuxième largeur de fenêtre, et pourra être comparée à l'erreur équivalente dans la méthode à simple noyau.

c) Choix de la largeur de fenêtre

Après justification et expérimentations, les auteurs utilisent la règle suivante pour le choix de la deuxième largeur de fenêtre pour une régression quantile à double noyau d'ordre p .

$$h_{2,p} = \begin{cases} max\left(\frac{h_{1,1/2}^5}{h_{1,p}^3}, \frac{h_{1,p}}{10}\right) & si\ h_{1,1/2} < 1 \\ \frac{h_{1/1/2}^4}{h_{1,p}^3} & sinon \end{cases} .$$

Ce choix n'est sans doute pas optimal, mais fournit cependant une formule raisonnable pour obtenir $h_{2,p}$ à partir de $h_{1,p}$.

3. Apports et résultats comparés des deux méthodes [2]

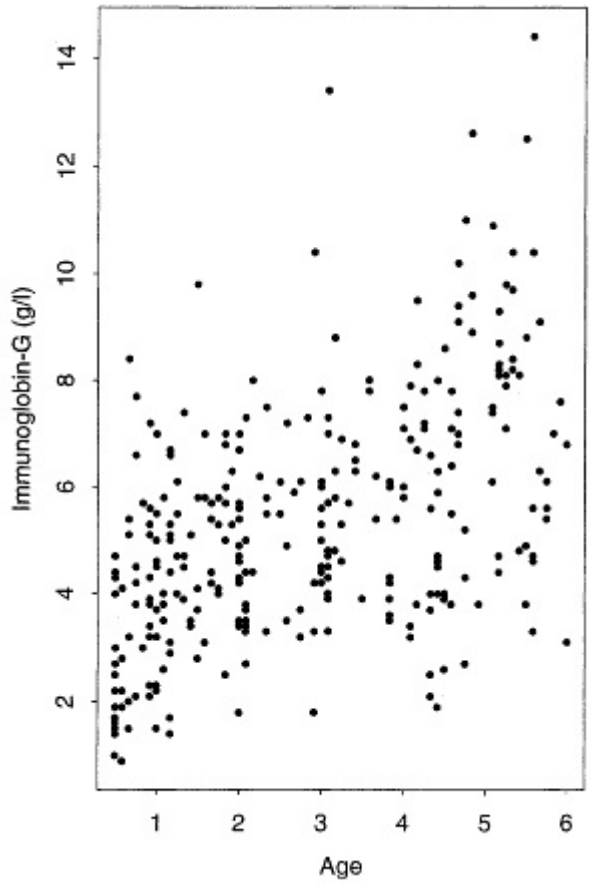


Fig. 2 – Concentration d'immunoglobulines G en fonction de l'age des enfants

Les auteurs ont testé les deux méthodes sur un jeu de données de concentration de sérum en immunoglobulines G (en g/L) chez des enfants âgés de 6 mois à 6 ans.

Nous observons ci-dessous les résultats des régressions quantile de différents ordres par la méthode à simple noyau puis à double noyau.

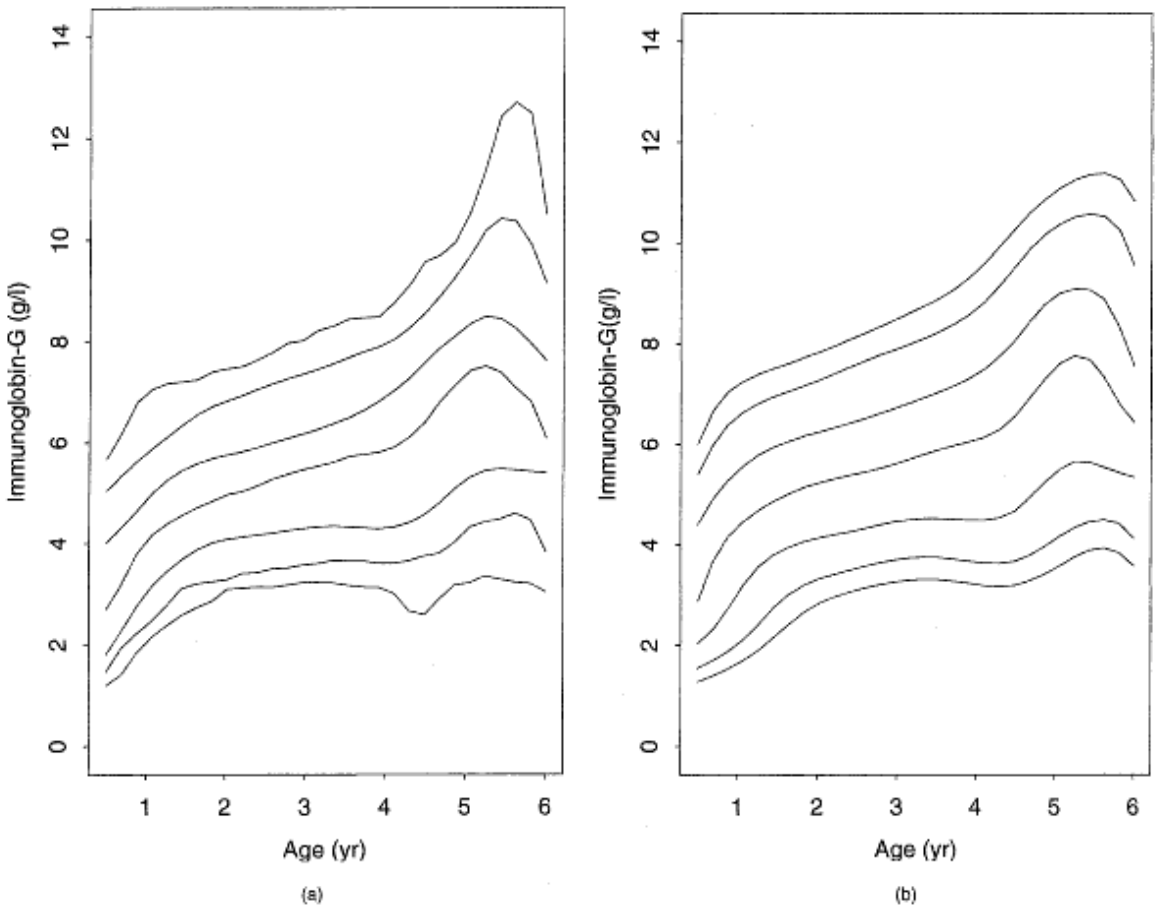


Fig. 3 – Courbes quantiles lissées d'ordres 5 %, 10 %, 25 %, 50 %, 75 %, 90 % et 95 %, pour les données de concentrations d'immunoglobulines G, utilisant (a) le lissage à simple noyau et (b) le lissage à double noyau.

Apport des méthodes :

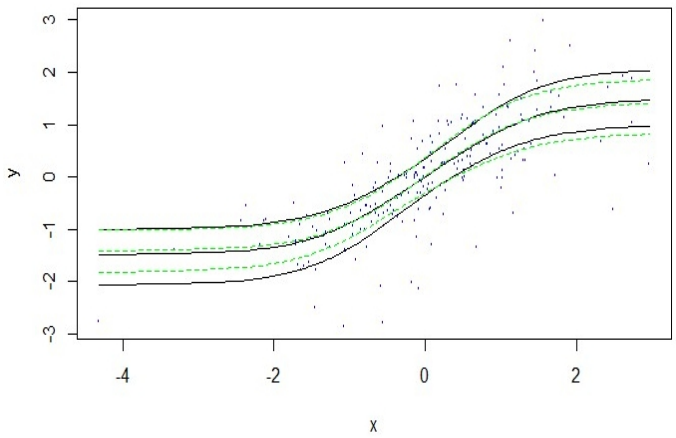
Par rapport à d'autres méthodes de régression quantile, celles-ci ont l'avantage de mettre en évidence un pic de concentration pour les enfants âgés d'environ 5 ans ½, chose qui n'était pas visible auparavant.

Simple noyau vs double noyau :

- Les courbes quantiles obtenue par la méthode à double noyau sont indéniablement **plus lisses** que celles obtenue par la méthode à simple noyau (minimisation de la fonction « check »). Ceci vient d'un **lissage** supplémentaire **vertical** en plus du lissage **horizontal**.
- Un autre avantage de la deuxième méthode est qu'elle empêche les courbes quantiles de différents ordre de **se croiser**, alors que ce problème peut arriver avec la première méthode.
- Enfin, dans la plupart des cas, l'approche par double noyau fournit une **erreur quadratique moyenne plus faible** qu'avec l'approche par simple noyau.

4. Simulations avec R

Nous avons testé une méthode de régression non paramétrique localement linéaire sur des données simulées par le modèle de Frank Copula (voir code en annexe). La figure ci-dessous présente le nuage de points des données simulées, les courbes de régression quantile théorique d'ordre 25 %, 50 % et 75 % (en bleu et traits pleins), et les courbes quantiles estimées (en vert et points tillés).



On observe que l'estimation est meilleure dans la partie centrale du nuage de points où les points sont plus concentrés. On observe, selon les quantiles, l'apparition progressive d'un léger biais quand on s'approche des bords où les points sont également moins concentrés.

Fig. 4 – comparaison des courbes quantiles théoriques et estimées

Puis, on compare les courbes des régressions quantiles à noyau d'ordre 50 % (p=0.5) pour différentes largeur de fenêtre h. On observe ici que h=1 ne convient pas : la courbe passe trop près des points et ne rend pas l'allure du nuage de points. h=3 semble proche du h optimal.

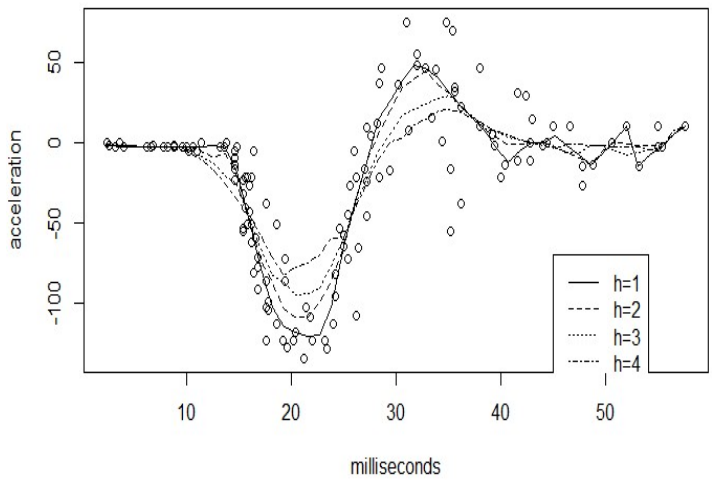


Fig. 5 – courbes de régressions quantile à noyau en fonction de la largeur de fenêtre h

5. Conclusion et ouvertures

Les méthodes présentées de régression quantile localement linéaire à simple noyau et à double noyau fournissent des courbes quantiles apportant des informations supplémentaires par rapport aux méthodes déjà connues. Enfin la méthode à double noyau semble plus avantageuse que celle à simple noyau : erreur quadratique plus faible, courbes quantiles plus lisses et ne pouvant pas se croiser. Elle est présentée comme très prometteuse dans la régression quantile non paramétrique.

Les auteurs ont d'ailleurs depuis publié un nouvelle article présentant une amélioration de la régression quantile localement linéaire à double-noyau [3].

6. Bibliographie

[1] Keming Yu, Zudi Lu & Julian Stander (2003). *Quantile regression: applications and current research areas*. The Statistician, Vol. 52, No. Part 3. (2003), pp. 331-350

[2] Keming Yu & M. C. Jones (1998). *Local Linear Quantile Regression*. Journal of the American Statistical Association, 93(441), pp. 228–237.

[3] Jones, M C & Yu, Keming (2007). *Improved double kernel local linear quantile regression.* Statistical Modeling, 7(4), pp. 377–389.

[4] J. Fan. *Local Linear Regression Smoothing and Their Minimax Efficiencies.* The Annals of Statistics, 21, 196-216

[5] D. Ruppet, S. J. Sheather, & M. P. Wand. *An Effective Bandwidth Selector for Local Least Squares Regression.* Journal of the American Statistical Association, 90, 1257-1270

```r
### Regression quantile par methode non parametrique localement lineaire ###
# chargement du package quantreg
library(quantreg)

### estimating the quantile functions of the Frank copula model introduced in Koenker (2005)
n <- 200
df <- 8
delta <- 8
set.seed(4003)
x <- sort(rt(n,df))
u <- runif(n)
v <- -log(1-(1-exp(-delta))/(1+exp(-delta*pt(x,df))*((1/u)-1)))/delta
y <- qt(v,df)
plot(x,y,col="blue",cex = .25)

us <- c(.25,.5,.75)
for(i in 1:length(us)){
  u <- us[i]
  v <- -log(1-(1-exp(-delta))/
              (1+exp(-delta*pt(x,df))*((1/u)-1)))/delta
  lines(x,qt(v,df))
}

Dat <- NULL
Dat$x <- x
Dat$y <- y
deltas <- matrix(0,3,length(us))
for(i in 1:length(us)){
  tau = us[i]
  fit <- nlrq(y~FrankModel(x,delta,mu,sigma,df=8,tau=tau),
              data=Dat,tau= tau, start=list(delta=5,
                                            mu = 0, sigma = 1),trace=TRUE)
  lines(x, predict(fit, newdata=x), lty=2, col="green")
  deltas[i,] <- coef(fit)
}

### Estimation of a locally linear median regression model for the motorcycle data
# four distinct bandwidths
library(MASS)
data(mcycle)
attach(mcycle)
plot(times, accel, xlab = "milliseconds", ylab = "acceleration")
hs <- c(1, 2, 3, 4)
for (i in hs)
{
  h = hs[i]
  fit <- lprq(times, accel, h = h, tau = 0.5)
  lines(fit$xx, fit$fv, lty = i)
}
legend(45, -70, c("h=1", "h=2", "h=3", "h=4"),lty = 1:length(hs))

# functions
lprq <- function (x, y, h, tau = 0.5, m = 50)
{
  xx <- seq(min(x), max(x), length = m)
  fv <- xx
  dv <- xx
  for (i in 1:length(xx)) {
    z <- x - xx[i]
    wx <- dnorm(z/h)
    r <- rq(y ~ z, weights = wx, tau = tau, ci = FALSE)
    fv[i] <- r$coef[1]
    dv[i] <- r$coef[2]
  }
  list(xx = xx, fv = fv, dv = dv)
}

FrankModel <- function(x,delta,mu,sigma,df,tau){
  z <- qt(-log(1-(1-exp(-delta))/
                 (1+exp(-delta*pt(x,df))*((1/tau)-1)))/delta,df)
  mu + sigma*z
}
```