

Estimation non paramétrique : Quelques (bonnes ?) pratiques dans l'

Christophe Bontemps
Toulouse School of Economics (INRA)



Séminaire joint :
*Séminaire Statistique TSE et Réseau des Ingénieurs Statisticiens
Toulousains
13 mai 2014*

PLAN

Pourquoi ce “non”

Définition par le “ non”

Estimer une densité

Des boîtes et des bosses

En pratique avec R

La fenêtre !

Critères

La régression

La fenêtre !

Cas pratiques avec R

Les cas moins simples

Cas pratiques avec R

A quoi ça sert tout ça ?

Estimation d'une
probabilité conditionnelle
Ajustement, prévisions et
simulations

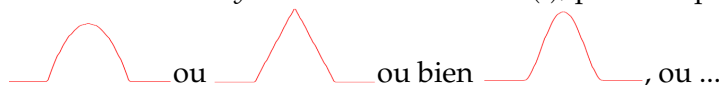
Vous avez demandé un
test ?

Une définition par le “ *non*”

- ▶ *Non*-paramétrique ne s’oppose pas vraiment à paramétrique
 - ▶ C’est l’objet d’intérêt qui *n’est pas* un paramètre
 - ▶ On parle aussi d’estimation fonctionnelle, de paramètre fonctionnel
 - ▶ Une estimation non-paramétrique comporte des choix de paramètres
 - ▶ \exists de multiples façon d’estimer non-paramétriquement
- Focus sur les méthodes “à noyau”
- ▶ Beaucoup de méthodes sont programmées dans *R*

→ Demo 1

- Pour l'estimation non-paramétrique de la densité :
- On choisit un "noyau" i.e. une fonction $K(\cdot)$, par exemple :



$K(\cdot)$ est une sorte de "bosse" et vérifie :

$$\int K(u)du = 1, \int u K(u)du = 0, \text{ et } \int u^2 K(u)du = \kappa_2 < \infty$$

- L'estimateur à noyau de Parzen-Rosenblatt est :

$$\widehat{f_h(x)} = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$$

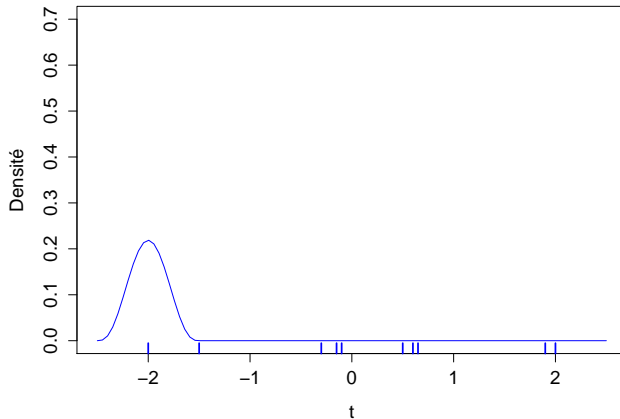
- Ca ressemble à l'histogramme non ?

$$\widehat{f_h(x)} = \frac{1}{nh} \sum_{i=1}^n \mathbb{1}_{[X_i \text{ dans le meme segment que } x]}$$

- Peut être vu comme une "somme de bosses"

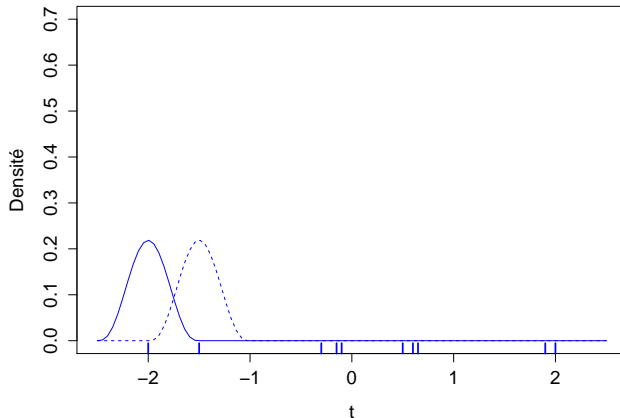
Comment ça marche ? Exemple sur 10 points

Une bosse



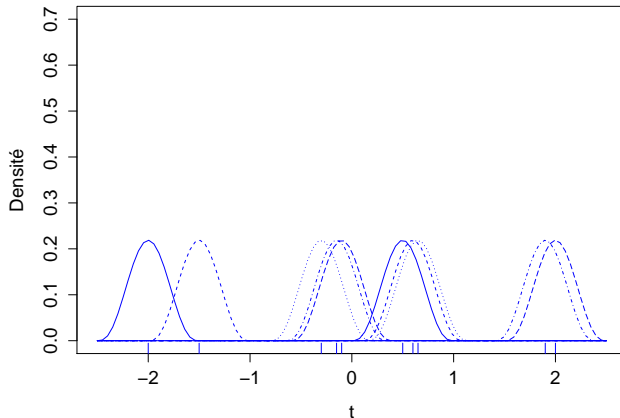
Comment ça marche ? Exemple sur 10 points

Une bosse autour de chaque point



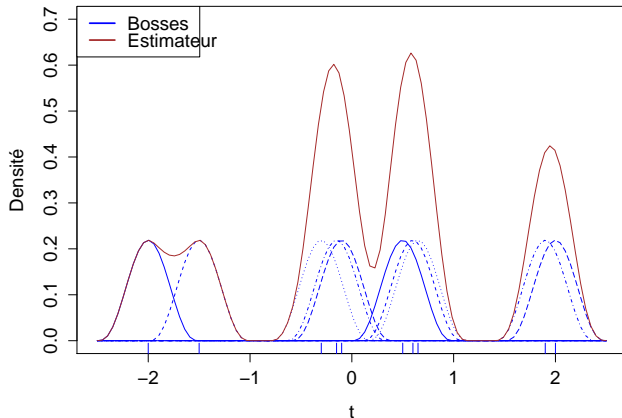
Comment ça marche ? Exemple sur 10 points

Une somme de bosses



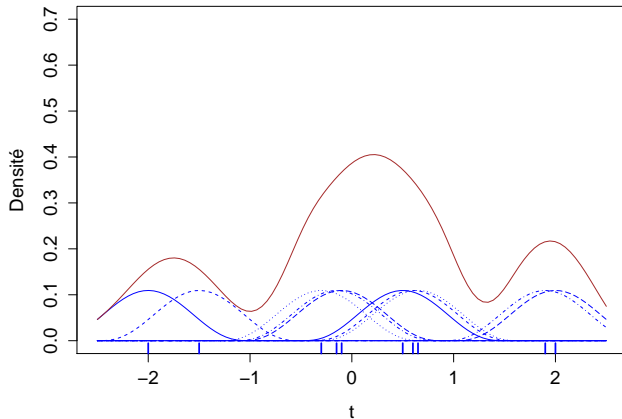
Comment ça marche ? Exemple sur 10 points

L'estimateur = somme de bosses



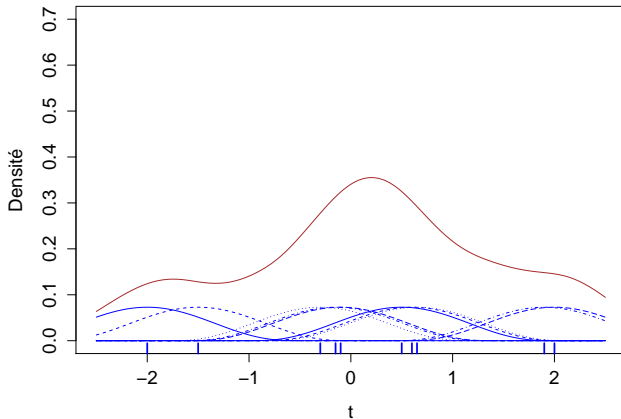
Comment ça marche ? Si j'agrandis " h "

Une somme de bosses ($h=1$)



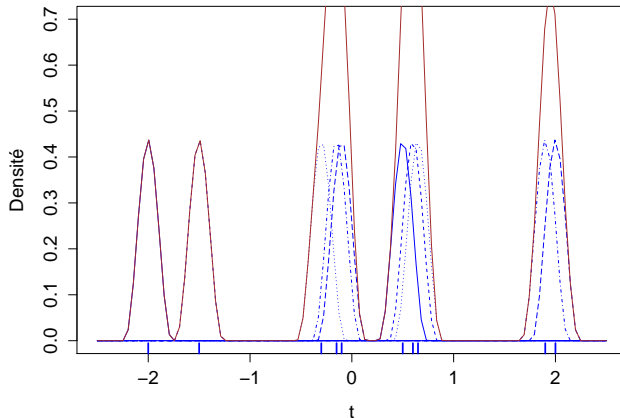
Comment ça marche ? Si j'agrandis " h " encore

Une somme de bosses ($h=1.5$)



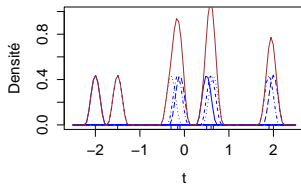
Comment ça marche ? Si je réduits " h "

Une somme de bosses ($h=0.25$)

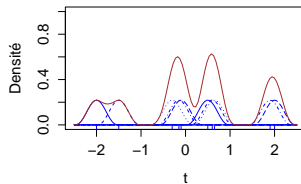


Comment ça marche ? En résumé

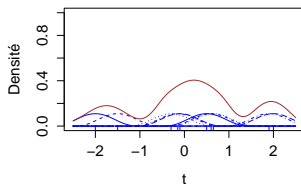
$h = 0.25$



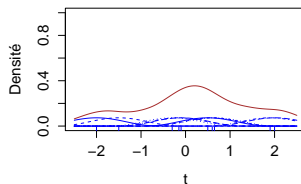
$h = 0.5$



$h = 1$



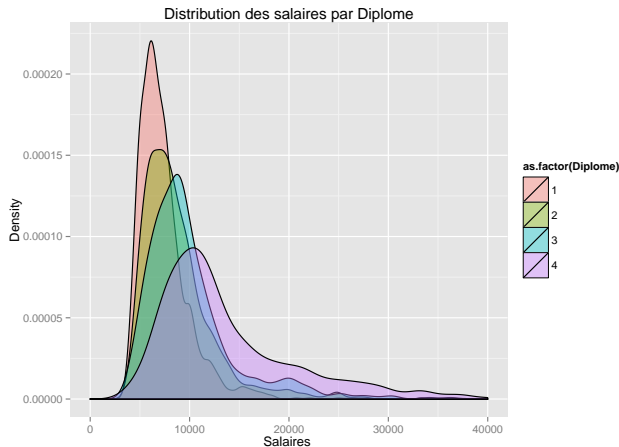
$h = 1.5$



En pratique avec R

- ▶ La commande `plot(density(x))` permet de représenter graphiquement la densité
 - ▶ Plusieurs packages permettent rapidement d'estimer une densité
 - ▶ *KernSmooth*, *np*
 - ▶ *ggplot2* permet également de faire des représentations (très jolies)
- Focus sur *np* ici pour des raisons explicitées plus tard.

► Exemple de graphique avec *ggplot2* (fonction *qplot*)



Comment choisir sa fenêtre ?

- Visuellement ...
- Calculer un critère pour différentes valeurs de h et prendre le minimum...
- Directement avec l'erreur quadratique en un point $MSE(\hat{f}_h(x))$:

$$MSE(\hat{f}_h(x)) = E \left[(\hat{f}_h(x) - f(x))^2 \right] = Var(\hat{f}_h(x)) + \left\{ \text{Biais}(\hat{f}_h(x)) \right\}^2$$

- Mieux encore, l'IMSE $\hat{f}_h = \int MSE(\hat{f}_h(x)) dx$

$$\begin{aligned} &\simeq \frac{1}{nh} \int K^2(z) dz + \frac{h^4}{2} \cdot \kappa_2^2 \cdot \int (f''(z))^2 dz \\ &= \frac{1}{nh} \cdot \Phi_0 + \frac{h^4}{2} \cdot \kappa_2^2 \cdot \Phi_1 \end{aligned}$$

- Et ça c'est vachement utile !

Comment choisir sa fenêtre (suite)

► $l'IMSE(\hat{f}_h) = \frac{1}{nh} \cdot \Phi_0 + \frac{h^4}{2} \cdot \kappa_2^2 \cdot \Phi_1$

et donc :

↗ si $nh \nearrow \infty$ le premier terme disparaît

↘ et si $h \searrow 0$; c'est le second !

► La fenêtre qui minimise $l'IMSE(\hat{f}_h)$ est : $h_{opt} = c \cdot n^{-1/5}$

avec $c = \left[\frac{\int K^2(z) dz}{(\int z^2 K(z) dz)^2 \cdot (\int (f''(z))^2 dz)} \right]^{1/5}$

► On a ensuite le choix :

► "Faire comme si" on connaissait κ_2 , Φ_0 , et Φ_1

→ Règle du pouce : $h_{RoT} = 1.059 \cdot \sigma(x) \cdot n^{-1/5}$

► Estimer toutes ces choses là : $\int (f''(x))^2 dx, \dots \rightsquigarrow \hat{c}$

→ Méthode de Plug-in : $h_{plug} = \hat{c} \cdot n^{-1/5}$

Comment choisir sa fenêtre (*validation croisée*)

- On peut aussi décomposer l'ISE($\hat{f}_h(x)$) :

$$\begin{aligned}
 ISE(\hat{f}_h(x)) &= \int \left(\hat{f}_h(x) - f(x) \right)^2 dx \\
 &= \underbrace{\int \hat{f}_h(x)^2 dx}_{\text{calculable}} - 2 \underbrace{\int \hat{f}_h(x) f(x) dx}_{E(\hat{f}_h(x))} + \underbrace{\int f(x)^2 dx}_{\text{pas de } h!}
 \end{aligned}$$

- Quelques calculs plus tard... on minimise un critère empirique basé sur l'estimation de ces valeurs

$$CV(h) = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n K^{(2)} \left(\frac{X_j - X_i}{h} \right) - \frac{2}{n} \sum_{i=1}^n \hat{f}_h^{-i}(X_i)$$

où $\hat{f}_h^{-i}(X_i) = \text{leave-one-out}$ et $K^{(2)}(u) = \int K(u-t)K(t)dt$.

- Et la fenêtre choisie $\hat{h}_{CV} = \arg \min_h CV(h)$

Comment choisir sa fenêtre en pratique

- ▶ Plusieurs critères :
 - ▶ Dans *KernSmooth*, on peut utiliser la commande $dpik(x)$ pour calculer une fenêtre qui sera directement “pluggée” dans l’estimateur .
 - ▶ Dans *np*, on privilégie une approche *data-driven* : la validation croisée.
- ▶ On procédera donc toujours en deux étapes dans **R** :
 1. On estime la (ou les) fenêtre(s)
 2. On estime la fonction (densité, regression ou autre) avec cette (ces) fenêtre(s)
- 2-bis On peut ensuite visualiser le résultat en estimant les valeurs de $\hat{f}_h(x)$ sur un ensemble de points régulièrement espacés (séquence ou grille)

↪ **Demo 2**

Pour la régression, on a fait le plus dur !

- L'objet statistique à étudier est :

$$m(x) \equiv E(Y|X = x) = \int y f(y|x) dy = \int y \frac{f(x, y)}{f(x)} dy$$

- On met des chapeaux partout !

$$\widehat{m(x)} = \int y \frac{\widehat{f(x, y)}}{\widehat{f(x)}} dy$$

- On montre que : (Estimateur de Nadaraya-Watson)

$$\widehat{m}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

- C'est une somme pondérée des Y_i

$$\widehat{m}(x) = \sum_{i=1}^n Y_i W(X_i, x, h)$$

Comment choisir sa fenêtre pour la régression ?

- ▶ Même logique, calculs différents
- ▶ Fenêtre optimale :

$$h_{opt} = \left[\frac{\int \sigma^2(x) f^{-1}(x) dx \int K^2(u) du}{\int \{2m'(x) f'(x) f^{-1}(x) + m''(x)\}^2 dx \kappa_2^2} \right]^{1/5} n^{-1/5}$$

- ▶ Plug-in : Ben "YAKA" estimer tout ces trucs et remplacer...
- ▶ Règle du pouce : $h_{RoT} \propto \sigma(x) \cdot n^{\frac{-1}{5}}$
- ▶ Validation croisée :

$$h_{CV} = \arg \min_h \frac{1}{n} \sum_{i=1}^N \left(Y_i - \widehat{m}^{-i}(X_i) \right)^2$$

Démo dans un cas simple (avec *Shiny*)

Oui, mais dans la vraie vie :

- Et si on a plusieurs variables ?

↪ Une fenêtre par variable, noyaux multiplicatifs :

$$\hat{m}(x, z) = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h_x}\right) \cdot K\left(\frac{Z_i - z}{h_z}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h_x}\right) \cdot K\left(\frac{Z_i - z}{h_z}\right)}$$

- Et si on a une variable discrète, x^d , avec c catégories ?

↪ Il existe des noyaux généralisés (Aitchison and Aitken)

$$l(X_i^d, x^d) = \begin{cases} 1 - \lambda & \text{if } X_i^d = x^d \\ \frac{\lambda}{c-1} & \text{otherwise.} \end{cases}$$

ou $\lambda \in [0, (c-1)/c]$.

- Oui : on peut mixer les deux !

↪ cf exemple dans une minute !!

Oui, mais dans la vraie vie :

- ▶ Et si on a beaucoup d’observations, la CV ça prend du temps ?
 - ↪ Oui !
 - ↪ Package *npRmpi* permet de paralléliser les calculs
 - ↪ On peut aussi “*ruser*” ...
- ▶ Et comment on compare avec un modèle linéaire ?
 - ↪ \exists des tests :
 - ▶ *npcmstest* pour tester la correcte spécification d’un modèle linéaire (Hsiao, Li, and Racine (2007))
 - ▶ *npsigtest* pour tester la significativité des régresseurs (Racine, Hart, and Li (2006))
- ▶ On peut aussi invoquer une représentation graphique pour comparer...
 - ↪ cf exemple dans 30 secondes !

Pourquoi ce “non”



Estimer une densité



La fenêtre !



La régression



A quoi ça sert tout ça ?



Démo complète sur un jeu de données (pas à pas !)

Autre exemple d'application : To drink or not to drink (tap water) ?

- ▶ 4.623 ménages en France déclarant tous leurs achats alimentaires (y compris boissons)
- ▶ Classifiés en "*buveur d'eau du robinet*" (*irob* = 1) ou non (*irob* = 0)
- ▶ Information individuelle composée de variables continues :
 - ▶ Revenu déclaré *Income* et indice mesurant la qualité de l'environnement local : **Poor Raw Water Quality** (PRWQ)
 $\hookrightarrow PRWQ \nearrow$ si l'environnement est dégradé
- ▶ et de variables discrètes (ordonnées ou pas)
 - ▶ *diplome, region, habitant en milieu rural, retired.*
- ▶ Modèle estimé (probit) (cf Bontemps & Nauges (2009))

TABLE: Modèle probit estimé (cf Bontemps & Nauges (2009))

	Estimate	$Pr(Z > z)$
(Intercept)	2.3296	0.0000
PRWQ	-1.8113***	0.0021
Income	-0.5492**	0.0155
diploma	-	-
diplo.L	-0.1328	0.0464
diplo.Q	0.0435	0.4433
diplo.C	-0.0229	0.5703
Region	-	-
Region2	-0.0284	0.7376
Region3	-0.5879***	0.0000
Region4	-0.0590	0.3836
Region5	-0.0468	0.5887
Region6	0.3706***	0.0000
Region7	0.1486***	0.0576
Region8	0.2974***	0.0005
deleg	-0.0178	0.6966
rural	0.2397	0.0095
iret	-1.3491***	0.0089
PRWQ × Income	0.5789**	0.0166
PRWQ × iret	0.9461*	0.0871

Estimation non-paramétrique d'une probabilité conditionnelle

- L'objet statistique est la probabilité conditionnelle de Y (0/1) conditionnelle à $X = (X^c, X^d, \tilde{X}^d)$

$$g(Y = y|X = x) = \frac{f(x, y)}{f(x)} \quad (1)$$

- Pour un $x = (x^c, x^d, \tilde{x}^d)$ donné, l'estimateur de $f(x)$ est :

$$\begin{aligned} \hat{f}(x) &= \hat{f}(x^c, x^d, \tilde{x}^d) \\ &= n^{-1} \sum_{i=1}^n \prod_{j=1}^p W(X_{ij}^c, x_j^c) \prod_{j=1}^q l(X_{ij}^d, x_j^d) \prod_{j=1}^r \tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d) \end{aligned}$$

On a là 3 types de noyaux différents suivant la nature des variables (Li & Racine, 2003).

- Pour une **variable continue** x_j^c , on retrouve $W(\cdot)$:

$$W(X_{ij}^c, x_j^c) = \frac{1}{h_j} K \left(\frac{X_{ij}^c - x_j^c}{h_j} \right)$$

avec $K(\cdot)$ notre noyau "classique" et h_j la fenêtre associée.

- Pour une **variable discrète** x_j^d avec c_j categories, on a :

$$l(X_{ij}^d, x_j^d) = \begin{cases} 1 - \lambda_j & \text{if } X_{ij}^d = x_j^d \\ \frac{\lambda_j}{c_j - 1} & \text{sinon.} \end{cases}$$

avec la "fenêtre" $\lambda_j \in [0, (c_j - 1)/c_j]$.

- Pour une **variable discrète ordonnée** \tilde{x}_j^d , on a :

$$\tilde{l}(\tilde{X}_{ij}^d, \tilde{x}_j^d) = \begin{cases} 1 & \text{if } \tilde{X}_{ij}^d = \tilde{x}_j^d \\ \gamma_j^{|\tilde{X}_{ij}^d - \tilde{x}_j^d|} & \text{sinon.} \end{cases}$$

avec la "fenêtre" $\gamma_j \in [0, 1]$.

TABLE: Modèle Probit et modèle non-paramétrique estimés sur ces données - Coefficients (probit) et fenêtres optimales (CV)

	Estimate	$Pr(Z > z)$	Bandwidth	upper bound
(Intercept)	2.3296	0.0000	-	-
PRWQ	-1.8113***	0.0021	0.1801905	∞
Income	-0.5492**	0.0155	1.294752	∞
diploma	-	-	0.8634835	1
diplo.L	-0.1328	0.0464	-	-
diplo.Q	0.0435	0.4433	-	-
diplo.C	-0.0229	0.5703	-	-
Region	-	-	0.1208747	0.875
Region2	-0.0284	0.7376	-	-
Region3	-0.5879***	0.0000	-	-
Region4	-0.0590	0.3836	-	-
Region5	-0.0468	0.5887	-	-
Region6	0.3706***	0.0000	-	-
Region7	0.1486***	0.0576	-	-
Region8	0.2974***	0.0005	-	-
deleg	-0.0178	0.6966	0.5	0.5
rural	0.2397	0.0095	0.0721212	0.5
iret	-1.3491***	0.0089	3.253532e-13	0.5
PRWQ×Income	0.5789**	0.0166	-	-
PRWQ×iret	0.9461*	0.0871	-	-
irob	-	-	9.802058e-15	0.5

The bandwidths are chosen by minimizing a least-square cross-validation criterion.

The upper bound for a bandwidth, is equal to $(c_j - 1)/c_j$ in the case of an unordered discrete variable with c_j categories, and 1 in the case of an ordered one.

TABLE: Modèle Probit et modèle non-paramétrique estimés sur ces données - Coefficients (probit) et fenêtres optimales (CV)

	Estimate	$Pr(Z > z)$	Bandwidth	upper bound
(Intercept)	2.3296	0.0000	-	-
PRWQ	-1.8113***	0.0021	0.1801905	∞
Income	-0.5492**	0.0155	1.294752	∞
diploma	-	-	0.8634835	1
diplo.L	-0.1328	0.0464	-	-
diplo.Q	0.0435	0.4433	-	-
diplo.C	-0.0229	0.5703	-	-
Region	-	-	0.1208747	0.875
Region2	-0.0284	0.7376	-	-
Region3	-0.5879***	0.0000	-	-
Region4	-0.0590	0.3836	-	-
Region5	-0.0468	0.5887	-	-
Region6	0.3706***	0.0000	-	-
Region7	0.1486***	0.0576	-	-
Region8	0.2974***	0.0005	-	-
deleg	-0.0178	0.6966	0.5	0.5
rural	0.2397	0.0095	0.0721212	0.5
iret	-1.3491***	0.0089	3.253532e-13	0.5
PRWQ×Income	0.5789**	0.0166	-	-
PRWQ×iret	0.9461*	0.0871	-	-
irob	-	-	9.802058e-15	0.5

The bandwidths are chosen by minimizing a least-square cross-validation criterion.

The upper bound for a bandwidth, is equal to $(c_j - 1)/c_j$ in the case of an unordered discrete variable with c_j categories, and 1 in the case of an ordered one.

TABLE: Modèle Probit et modèle non-paramétrique estimés sur ces données - Coefficients (probit) et fenêtres optimales (CV)

	Estimate	$Pr(Z > z)$	Bandwidth	upper bound
(Intercept)	2.3296	0.0000	-	-
PRWQ	-1.8113***	0.0021	0.1801905	∞
Income	-0.5492**	0.0155	1.294752	∞
diploma	-	-	0.8634835	1
diplo.L	-0.1328	0.0464	-	-
diplo.Q	0.0435	0.4433	-	-
diplo.C	-0.0229	0.5703	-	-
Region	-	-	0.1208747	0.875
Region2	-0.0284	0.7376	-	-
Region3	-0.5879***	0.0000	-	-
Region4	-0.0590	0.3836	-	-
Region5	-0.0468	0.5887	-	-
Region6	0.3706***	0.0000	-	-
Region7	0.1486***	0.0576	-	-
Region8	0.2974***	0.0005	-	-
deleg	-0.0178	0.6966	0.5	0.5
rural	0.2397	0.0095	0.0721212	0.5
iret	-1.3491***	0.0089	3.253532e-13	0.5
PRWQ×Income	0.5789**	0.0166	-	-
PRWQ×iret	0.9461*	0.0871	-	-
irob	-	-	9.802058e-15	0.5

The bandwidths are chosen by minimizing a least-square cross-validation criterion.

The upper bound for a bandwidth, is equal to $(c_j - 1)/c_j$ in the case of an unordered discrete variable with c_j categories, and 1 in the case of an ordered one.

TABLE: Modèle Probit et modèle non-paramétrique estimés sur ces données - Coefficients (probit) et fenêtres optimales (CV)

	Estimate	$Pr(Z > z)$	Bandwidth	upper bound
(Intercept)	2.3296	0.0000	-	-
PRWQ	-1.8113***	0.0021	0.1801905	∞
Income	-0.5492**	0.0155	1.294752	∞
diploma	-	-	0.8634835	1
diplo.L	-0.1328	0.0464	-	-
diplo.Q	0.0435	0.4433	-	-
diplo.C	-0.0229	0.5703	-	-
Region	-	-	0.1208747	0.875
Region2	-0.0284	0.7376	-	-
Region3	-0.5879***	0.0000	-	-
Region4	-0.0590	0.3836	-	-
Region5	-0.0468	0.5887	-	-
Region6	0.3706***	0.0000	-	-
Region7	0.1486***	0.0576	-	-
Region8	0.2974***	0.0005	-	-
deleg	-0.0178	0.6966	0.5	0.5
rural	0.2397	0.0095	0.0721212	0.5
iret	-1.3491***	0.0089	3.253532e-13	0.5
PRWQ×Income	0.5789**	0.0166	-	-
PRWQ×iret	0.9461*	0.0871	-	-
irob	-	-	9.802058e-15	0.5

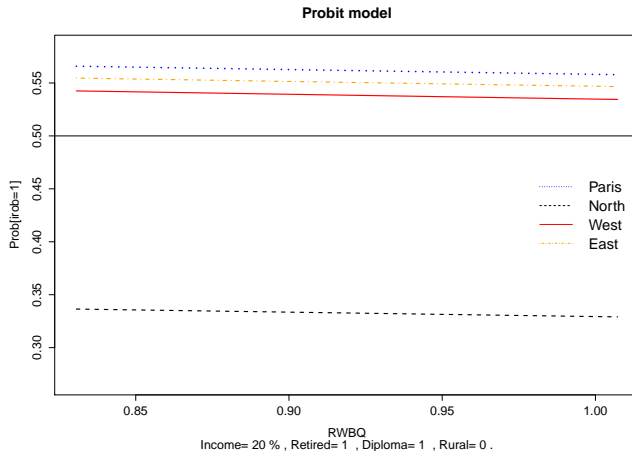
The bandwidths are chosen by minimizing a least-square cross-validation criterion.

The upper bound for a bandwidth, is equal to $(c_j - 1)/c_j$ in the case of an unordered discrete variable with c_j categories, and 1 in the case of an ordered one.

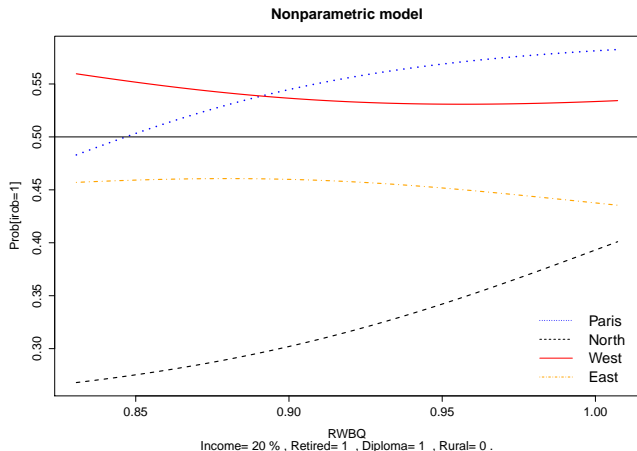
UNE AUTRE VISION DES RÉSULTATS

- ▶ On peut représenter en 2-D la probabilité de boire de l'eau du robinet comme une fonction **certaines variables** du modèle.
 - ▶ Il faut fixer les autres variables à un niveau déterminé (médiane, moyenne, autre)
 - ▶ Les interactions entre variables sont spécifiées dans le modèle paramétrique, mais sont automatiques dans le modèle non-paramétrique (fonction)
- ↪ Les graphiques 3-D mettent en lumière cela ..

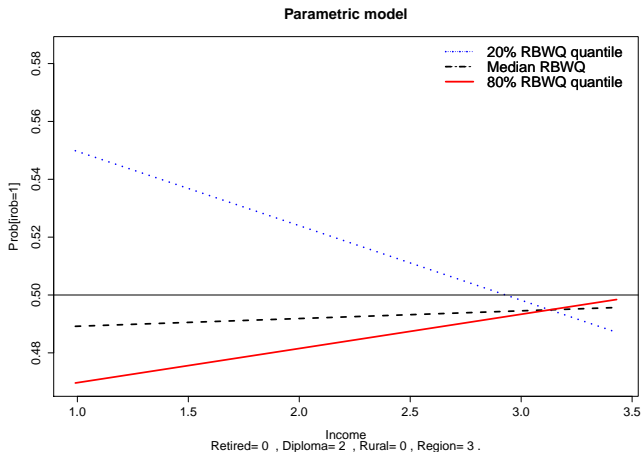
- Probabilité de boire de l'eau du robinet comme une fonction de l'indice *PRWQ* pour **différentes regions**.



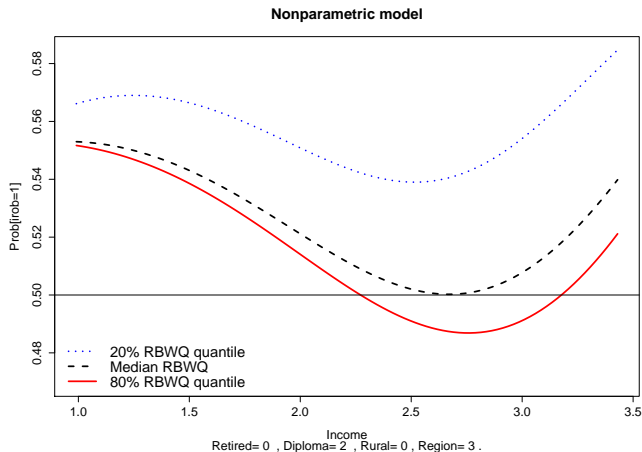
- Probabilité de boire de l'eau du robinet comme une fonction de l'indice *PRWQ* pour **différentes regions**.



- Probabilité de boire de l'eau du robinet comme une fonction du **revenu** pour différents **niveaux de l'indice PRWQ**.

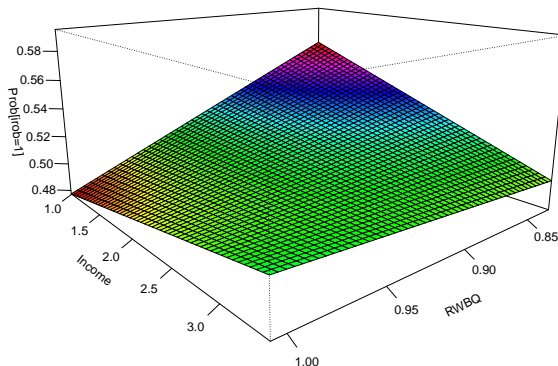


- Probabilité de boire de l'eau du robinet comme une fonction du **revenu** pour différents **niveaux de l'indice PRWQ**.



- Probabilité de boire de l'eau du robinet comme une fonction du **revenu** et de **l'indice PRWQ**.

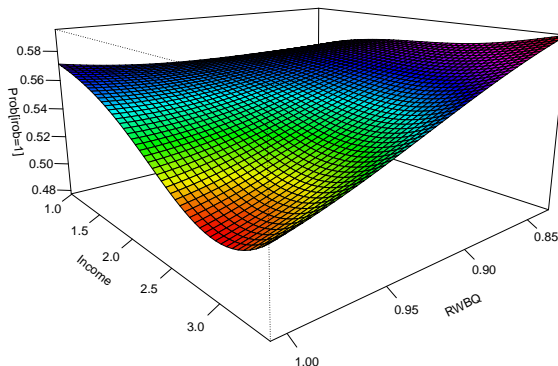
Estimated conditional prob of drinking tap water using Probit estimator



Region= 3 , Retired= 0 , Diploma= 1 , Rural= 0 .


- Probabilité de boire de l'eau du robinet comme une fonction du **revenu** et de **l'indice PRWQ**.

Estimated conditional prob of drinking tap water using NP estimator



Region= 3 , Retired= 0 , Diploma= 1 , Rural= 0 .

POUR CONCLURE EN 10 SECONDES

- ▶ L'estimation non paramétrique, c'est pas si horrible que ça !
- ▶ Les outils existent dans  et sont bien documentés
- ▶ L'estimation non paramétrique est **utile** (selon moi) si :
 - ▶ On cherche des non-linéarités (sur une variable)
 - ▶ On cherche principalement à prédire
 - ▶ On cherche à estimer une fonction dans un calcul intermédiaire (une densité par exemple)
 - ▶ On cherche à tester la pertinence de spécifications
- ▶ Les **difficultés** sont dans :
 - ▶ La compréhension des modèles estimés (pb de représentation de fonctions)
 - ▶ La diffusion des résultats en grande dimension
 - ▶ et (quand même aussi !) si on a beaucoup d'observations et/ou de grandes dimensions

VOUS AVEZ DEMANDÉ UN TEST ?

- ▶ Ne quittez pas...
- ▶ Sébastien arrive