République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ KASDI MERBAH OUARGLA

FACULTÉ DES MATHÉMATIQUES ET SCIENCES DE LA MATIÉRE

DÉPARTEMENT DE MATHÉMATIQUES

MÉMOIRE PRÉSENTÉ EN VUE DE L'OTENTION DU DIPLÔME :

MASTER en Mathématiques

Option : **Probabilités Et Statistique**

Par

Benhaoued Maissa

Titre :

Etude thèorique de l'estimation de la régression dans le modèle de données censuées

Etude thèorique de l'estimation de la régression dans le modèle de données censurées

Membres du Comité d'Examen :

Dr.	.....	UMKB	Président
Dr.		UMKB	Encadreur
Dr.	.....	UMKB	Examinateur

Juin 2021

Dédicace

$\mathcal{J}$e dédie ce modeste travail :

À ceux qui m'ont tout donné sans rien en retour A ceux qui m'ont encouragée et soutenue dans les moments les plus difficiles A vous **mes chers parents,** le plus beau cadeau que Dieu puissent faire à un enfant, pour leur amour et leur support continu.Que ce travail soit le témoignage sincère et affectueux de ma profonde reconnaissance pour tout ce que vous avez fait pour moi .

À mes chers grands-parents et grand-mères .

À mes chers frères **mohamade Ziad** et **Younse**

À mes soeurs **Romaissa,Zahra,mbarka,wafa**

À mon mari **Bachir** et sa famille.

À tous les membres de ma famille

À mes amis **karima,hanan,khadidja,** que a toujours été avec moi.

Àtous mes amis et collègues.

REMERCIEMENTS

$\mathcal{J}$e tiens premièrement à me prosterner, remerciant «**Allah**» le tout puissant de m'avoir donné le force et la volonté pour terminer ce travail.

Nous tenons à remercier M **alagon rachid** pour la proposition du thème , l'encadrement de ce travail, pour ses précieux conseils et orientations.

Nous remercions également les membres du jury M.....pour avoir accepté d'examiner et d'évaluer notre travail.

Nous sincères remerciements s'adressent enfin à tous ceux qui nous ont soutenu de près ou de loin.

Table des matières

Introduction

Le la régression est de déterminer la dont l'espérance d'une variable expliquée réelle Y que dépend d'une variable explicative X scalaire,vectorielle ou même fonctionnelle (qu'elle prend ses valeurs dans un espace de dimension infinie). Nous cherchons le lien entre X et Y modélisé par la fonction r vérifiant :

$$Y = r(X) + \varepsilon$$

où ε est l'erreur supposée centrée et indépendante de X, ce qui permet de montrer que $r(X) = E(Y/X)$. Le problème consiste donc à déterminer ou à estimer pour chaque réalisation $x$de la variable X , la valeur de $r(x)$.

Pour caractériser cette fonction, une première approche consiste à utiliser un modèle de régression paramétrique. On suppose que cette fonction peut s'écrire comme une fonction explicite des valeurs de X. Cette dernière peut être linéaire, par exemple

$$r(x) = \alpha + \beta x$$

et on cherche alors à déterminer les meilleures valeurs des paramètres α et β un utilisées, par exemple la mèthode des moindres carrés. Nous nous ramenons alors à l'estimation d'un nombre fini de paramètres.Dans certains cas nous pouvons disposer pour cette estimation d'un échantillon$\{(X_i, Y_i), i = 1, ..., n\}$ de couples indépendants et ayant chacun la même loi que (X, Y) . Souvent, l'utilisation d'un modèle paramétrique n'est pas justifiée ; il est alors possible qu'il suiffit d'avoire la seule donnée de l'échantillon pour réaliser une estimation

.Ce sera à l'aide d'un modèle nonparamétrique.Dans ce cas on ne dispose d'aucune forme paramétrique pour r mais seulement d'hpothèses générales de régularité comme la continuer par exemple.

Le lien entre deux variables a générallement pour but de prédire la variable réponse Y étant donné une valeur de l'autre (variable explicative X). Il y a plusieurs façons d'aborder un problème de prévision et l'une des plus utilisées est la régression qui est basée sur l'espérance conditionnelle. La prédiction au moyen de la médiane conditionnelle nécessite l'estimation préalable de la fonction de répartition conditionnelle. Celle du mode conditionnel nécessite l'estimation de la densité conditionnelle . Ces deux méthodes de prévision ont été largement étudiées. Citons,parmi les innombrables travaux qui leur ont été consacrés *Gannoun et al (2003),Samanta et Thavaneswaran(1990),Khardani et al.(2011)etCollomb et al.(1987).*

Plusieurs mèthode d'estimation non-paramétrique de la régression sont disponibles comme , l'estimation des moindres carrés , et l'estimation des moindres carrés génalisés, ou spline de lissage., l'estimateur à noyau de Nadaraya-Watson. Ce dernier, qui a éte introduit indépendamment par Nadaraya(1964) et Watson(1964), est l'un des plus poplaires des modèles de régression non-paramétriques. Son expression est :

$$r(x) = \sum_{i=1}^{n} Y_i \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)},$$

où K est une fonction de $\mathbb{R}$ dans $\mathbb{R}$ et h_n est un paramètre réel strictement positif, appelé paramètre de lissage et dont le choix est essentiel.

Malheureusement dans la pratique, il n'est pas toujours possible d'avoir à disposition des données complètes. La fixation du temps de l'étude par exemple peut empêcher l'observation de la variable d'intérêt pour laquelle on saura seulement qu'elle dépasse la valeur observée, c'est le cas de la censure à droite. C'est pourquoi cet estimateur a été généralisé au le cas où la variable réponse est censurée à droite par Kohler et al.(2002).Guessoum et Ould Saïd(2008,2010,2012) ont étudié cet estimateur aussi dans le cas où les X_i .

Un phénomène de censurae à gauche (symétrique du précédent) peut aussi empêcher l'observation .Généralement , la censure à gauche s'accompagne de la censure à droite c' est le cas de censure mixte à laquelle nous nous intéressons dans cette mémoire.

Chapitre 1

Estimation non paramétrique sur des données censurées

1.1 Introduction

Dans cette partie nous introduisons la notion de censure dans les données, nous faisons quelques rappels basiques pour ce type de modèle, et nous donnons quelques résultats principaux concernant le comportement asymptotique de l'estimateur de la fonction de régression. Nous nous intéressons en particulier, à un estimateur à noyau non symétrique de la fonction de régression pour lequel nous établissons la vitesse de convergence.

Dans cette section, on va établir quelques résultats sur l'estimateur de la fonction de régression pour un modèle censuré, mais avant nous rappelons ce qu'est un modèle de survie, un modèle de censure.

1.2 Modèle de survie

Un modèle de survie s'appuie sur des durées de vie . Le terme de durée de vie est utilisé pour indiquer le temps qui passe jusqu'à la survenue d'un évènements, par exemple , l'apparition d'une maladie, la guérison d'un maladie, la panne d'une machine, etc...

Les modèles de survie sont souvent caractérisés par la préssence de censure : une censure se produit lorsque l'évènement étudié n'intervient pas période d'observation pour une raison ou une autre. Cette censure est dite "censure à droite" et est la plus courante mais n'est pas la seule censure que l'on peut rencontrer sur des données de survie

1.3 Données de survie et censure

Un donnée de survie représente le temps écoulé entre le début d'une observation et l'arrivée d'un événement. Historiquement, cette théorie a démarré dans Le cadre biomédical, d'où l'utilisation du terme décès. Cependant, le terme "données de survie" couvre d'autres événements , comme l'apparition d'une maladie ou une épidémie. Dans l'industrie , il peut s'agir de la panne d'une machine. En économie, du temps écoulé pour qu'une personne trouve un travail. Dans plusieurs cas, l'événement est la transition d'un état à un autre . Par exemple, le décès est la transition de l'état "vivant" vers l'état "mort" . L'apparition d'une maladie est la transition de l'état " en santé" vers l'état "malade".

Selon le contexte, les termes décès , événement, échec ou transition peuvent être utilisés pour désigner l'événement d'intérêt.

Les données censurées sont des observations ne correspondant pas à de vraies valeurs de la variable d'intérêt . Cependant, nous disposons tout de même d'une information partielle permettant par exemple de fixer une borne inférieure (censure à droit) ou une ou une borne supérieure (censure à gauche).

Les raisons de cette censure peuvent être le fait que le patient soit toujours vivant ou non malade à la fin de l'étude , ou qu'il se soit retiré de l'étude pour des raisons personnelles (immigration , mutation, ...).

il existe différents types de censures :

censure de type 1 : **fixée**

Soit C un nombre positif fixé .Au lieu d'observer les variables $X_1, X_2, ..., X_n$ qui nous in-

téressent, on observe X_i que lorsque $X_i \leq C$, si non on sait seulement que $Y_i > C$. L'observation est alors $Y_i = \min(X_i, C) = X_i \wedge C$. C'est le cas lorsqu'on décide à l'avance que le nombre C est la durée de l'étude.

censure de type 2 :

On décide d'observer les durées de survie de n patients jusqu'à ce que r d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Si l'on ordonne les durées de survie $X_1, X_2, ..., X_n$,soit $X_{(1)}$ la plus petite, $X_{(i)}$ la ième on a :

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(i)} \leq ... \leq X_{(n)}$$

On dit que les $X_{(i)}$ sont les statistiques d'ordre des X_i . La date de censure est alors $X_{(r)}$ et on observe :$Y_i = X_i \Lambda X_{(r)}$

$$\begin{cases} Y_1 = X_{(1)} \\ Y_2 = X_{(2)} \\ ..... \\ Y_r = X_{(r)} \\ Y_{r+1} = X_{(r+1)} \\ Y_n = X_{(n)} \end{cases}$$

ce cas est fréquemment utilisé en fiabilité lorsqu'on observe jusqu'à la première panne.

censure de type 3 :(aléatoire)

A chaque individu i, est associé un couple de v.a (X_i, C_i) positives où X_i est son temps de survie et C_i son temps de censure, tel que seule la plus petite est observée, c'est-à-dire $Y_i = X_i \Lambda C_i$

$$\delta_i =_{I(Y_i=X_i)}= I_{(X_i\leq C_i)} = \begin{cases} 1 & \text{si} \quad \text{non censeure} \\ 0 & \text{si} \quad \text{censeure} \end{cases}$$

En pratique la censure aléatoire peut avoir plusieurs causes : par exemple perte de vue , arrêt du traitement ou bien fin de l'étude ,Alors ce qu'on observe c'est le couple (Y_i, δ_i) et $\delta_i = ¶_{(Y_i\leq X_i)}$ (l'indicatrice de non censure).

1.4 Censure à droite

La durée de vie est dit censurée à droite si l'individu n'a pas subi l'événement à sa dernère observation .En présence de censure à droite , les durées de vie (Y) ne sont pas toutes observées ; pour certaines d'entre elles , on sait seulement qu'elles sont supérieures à une certaine valeur connue .Soit R une variable aléatoire de censure , au lieu d'observer la variable Y qui nous intéresse , on observe le coupe de variables (Z,δ) avec Z=min(Y, R) et $\delta = 1_{\{Y\leq R\}}.\delta$ est appelé indicateur de censure puisque ses valeurs nous informent sur le fait que l'observation est complète (si $\delta = 1$) ou censurée à droite (si $\delta = 0$).

Un exemple illustratif est lorsqu'on s'intéresse à la durée de vie d'un genre de machines précis mais que ces dernières tombent en panne s'il se produit une surtension d'électricité .Ici la durée de vie de la machine est censurée à droite par l'instant auquel se produit la surtension.

1.5 Censure à gauche

Le censure à gauche correspond au cas où l'individu a déjà subi l'événement avant qu'il ne soit observé. On sait uniquement que la valeur d'intérêt est inférieure à une certaine valeur connue représentée par une variable aléatoire L. Pour chaque individu , on peut associer un couple de variables aléatoires (Z, ∂) telles que Z=max (Y, L) et $\partial = 1_{\{Y\geq L\}}$. Un des premiers exemples de censure à gauche rencontré dans la littérature concerne le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger

(les babouins passent la nuit dans les arbres). Le temps d'événement (descente de l'arbre) est observé si le babouin descend de l'arbre après l'arrivée des observateurs . Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs : dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs .On observe donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs .

Dans un même échantillon peuvent être présentés des données censurées à droite et d'autres censurées à gauche , comme c'est le cas dans ce qui suit .

1.6 Censure double

par exemple , une étude s'est intéressée à l'âge auquel les enfants d'une communauté africaine apprennent à accomplir certaines tâches .Au début de l'étude , certains enfants savaient déjà effectuer les tâches étudiées , on sait seulement alors que l'âge où ils ont appris est inférieur à leur âge à la date du début de l'étude . A la fin de l'étude, certains enfants ne savaient pas encore accomplir ces tâches et on sait alors seulement que l'âge auqule ils apprendront éventuellement ont appris est supérieur à leur âge à la fin de l'étude. L'âge au début de l'étude (varaible de censure à gauche L) est évidemment inférieure à la fin de l'étude (varaible de censure à droite$R$). L'âge d'intérêt est observé ssi il se trouve dans la période d'étude . Nous observons $Z = \max(\min(Y, R), L)$ avec un indicateur de censure . Ce modèle à été étudié dans Turnbul qui a introduit un estimateur implicite de la fonction de survie de Y donné comme solution d'une équation de self-consistance.

1.7 Censure mixte

Nous disons qu'il y a censure mixte lorsque deux phénomènes de censure (l'un à gauche et l'autre à droite) peuvent empêcher l'observation du phénomène d'intérêt sans qu'on puisse nécessairement déterminer un intervalle auquel il appartient. Comme dans le modèle décrit

dans l'article de [Patilea et Rolin (2006)] , au lieu d'observer un échantillon de la variable d'intérêt Y, on observe un échantillon du couple $(Z; A)$ avec $Z = \max(\min(Y, R), L)$ et

$$A = \begin{cases} 0 & \text{si} \quad \mathrm{L} \prec Y \leq R \\ 1 & \text{si} \quad \mathrm{L} \prec R \prec Y \\ 2 & \text{si} \quad \min(Y, R) \leq L \end{cases}$$

où L et R sont des variables de cenceure et A est l'indiateur de censure. Un exemple de ce modèle est donné par un système formé par trois composants , dont deux sont placés en série (le composant dont le temps de fonctionnement nous intéresse et un autre).Un troisième est placé en parallèle avec ce système en série .Ici, il est clair qu'il n'est pas raisonnable de supposer que le temps de fonctionnement d'un composant soit inférieur à un autre .

L'analyse de surive a connu un développement important dans la seconde moitié du vingtième siècle après que Kaplan et Meier aient introduit leur célèbre estimateur de la fonction de survie pour des données censurées à droite .Estimateur qui généralise le complément à un de la fonction de répartition empirique et que nous rappelons ci dessous.

1.8 Détermination de la loi d'une durée de survie

Supposons que la durée de survie X soit une variable positive ou nulle , et absolument continue , alors sa loi de probabilité peut être définie par l'une des fonctions suivantes :

Définition 1.8.1 [*Fonction de survie*]

La fonction de survie noteé $S_X(t)$, est la probabilité pour un individu de vivre au moins jusqu'au temps t .

$$S_X(t) = P(X \succ t)$$

$$= 1 - P(X \leq t)$$

$$= 1 - F_X(t)$$

si la fonction de répartition a une dérivée au poit t alors la densité de X est donnée par :

$$f(t) = \lim_{h \to 0} \frac{F_X(t+h) - F_X(t)}{h} = F'_X(t) = -S'_X(t)$$

Définition 1.8.2 [***Taux de hasard***]

Le taux de hasard ou la fonction de risque λ est définie comme la probabilité qu'un indvidu fasse l'évènement considéré durant un intervalle de temps très court sachant qu'il a survécu jusqu'au début de l'intervalle

$$\lambda(t) = \lim_{h \to 0} \frac{P(X \prec t + h \,/\, X \geq t)}{h}$$

$$= \lim_{h \to 0} \frac{P(X \prec t + h\,,\ X \geq t)/P(X \geq t)}{h}$$

$$= \frac{1}{P(X \geq t)} \lim_{h \to 0} \frac{P(t \leq X \prec t + h)}{h}$$

$$= \frac{f_X(t)}{S_x(t)}$$

$$= \frac{f_x(t)}{1 - F_X(t)}$$

La fonction de risque mesure le risque instantané de survenue de l'évènement.

Définition 1.8.3 [***Taux de hasard cumulé***]

Le taux de hasard cumulé ou la fonction de risque cumulative évaluée au temps t est l'intégrale de la fonction de risque entre 0 et t

$$\Lambda(t) = \int_0^t \lambda(u) \quad du = -\log S(t)$$

On peut déduire la fonction de survie à partir du taux de hasard cumulé par la relation

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)\, du\right)$$

1.9 Estimateur de la fonction de survie

Soit (Ω, A, P) un espace probabililsé . Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires (v.a) positives indépendantes et identiquement distribuées (i.i.d) désignant des durés de survie d'un évenement donné de fonction de répartition (f.d.r) F .Sot $C_1, C_2, ..., C_n$ une suite de (v.a) de censures , posilives (i.i.d) de (f.d.r) G .Généralement , les (v.a) C_i sont supposées être indépendantes des X_i . Soit $(T_i, \delta_i)_i = 1, ..., n$ l'échantillon réellement observé .

où

$$T_i = \min(X_i, C_i) \quad et \quad \delta_i = ¶_{\{X_i \leq C_i\}}$$

1.5.1 Estimateur de Kaplan-Meier

L'estimateur de la fonction de survie le plus utilisé sans aucune hypothèse faite sur la distribution des temps de survie est l'estimateur de $Kaplan - Meier$ .Cet estimateur (que l'on notera EKM) est aussi appelé estimateur *product Limit* (PL) car il s'obtient comme limite d'un produit . L'ideé de la construction de l'EKM est la suivante, pour $t' \prec t$

$$S(t) = P(X \succ t, X \succ t')$$

$$= P(X \succ t \ / X \succ t')\, S(t')$$

On renouvelle l'opération en choisissant $t" \prec t'$ on obtint :

$$S(t') = P(X \succ t'/X \succ t")S(t")$$

D'où

$$S(t) = P\ (X \succ t\ /X \succ t')\quad P(X \succ t'/X \succ t")\, S(t")$$

Si on choisit pour dates où l'on conditionne celles où il s'est produit un évènement (décès ou censure) i.e $T_{(i)}$ on estime seulement des quantités de la forme

$$P_i = P\left(X \succ T_{(i)}/X \succ T_{(i-1)}\right)$$

P_i est la probabilité de survivre pendant l'intervalle $I_i = \left]T_{(i-1)}, T_{(i)}\right]$ quand on est vivant au début de cet intervalle . Soit R_i le nombre de sujets à risque à l'instant $T_{(i)}$.M_i le nombre de décès observées à l'instant $T_{(i)}$ et $q_i = 1 - p_i =$ probabilité de mourir pendant l'intervalle I_i sachant qu' on était vivant au début de l'intervalle .Alors un estimateur nateurel de q_i est

$$\hat{q}_i = \frac{M_i}{R_i} = \frac{\text{nombre de mort à l'instant } T_{(i)}}{\text{nombre de sujets à risque}}$$

Supposons qu'il n'y ait pas d'exequo (càd tous les $T_{(i)}$ sont différents). si $\delta_{(i)} = 0$ il ya censure à l'instant $T_{(i)}$, implique $M_i = 0$. On a alors

$$\hat{q}_i = \begin{cases} \frac{1}{R_i} & \text{si } \delta_{(i)} = 1 \\ 0 & \text{si } \delta_{(i)} = 0 \end{cases}$$

$\Longrightarrow$

$$\hat{p}_i = 1 - \hat{q}_i = \begin{cases} 1 - \frac{1}{R_i} & \text{si } \delta_{(i)} = 1 \\ 1 & \text{si } \delta_{(i)} = 0 \end{cases}$$

$\Longrightarrow$

$$\hat{p}_i = \left(1 - \frac{1}{R_i}\right)^{\delta_{(i)}}$$

il est claire que $R_i = n - i + 1$. On obtient finalement l'EKM pour la fonction de survie de la variable durée de vie X :

$$\hat{S}_{KM}(t) = 1 - \hat{F}_{KM}(t) = \begin{cases} \Pi_{T_{(i)} \leq t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} \text{ si } \ t \prec T_{(n)} \\ 0 \qquad\qquad\qquad\qquad\qquad\quad \text{si } t \geq T_{(n)} \end{cases} \tag{1.1}$$

et donc on a aussi l'EKM pour la fonction de survie de la variable de censure C

$$\bar{G}_n(t) := 1 - \hat{G}_{KM}(t) = \begin{cases} \Pi_{T_{(i)} \leq t} \left(1 - \frac{1}{n-i+1}\right)^{1-\delta_{(i)}} \text{ si } \ t \prec T_{(n)} \\ 0 \qquad\qquad\qquad\qquad\qquad\quad \text{si } t \geq T_{(n)} \end{cases} \tag{1.2}$$

où$\left(T_{(i)}, \delta_{(i)}\right)_{i=1,\ldots,n}$ sont telle que $T_{(1)} \leq T_{(2)} \leq .... \leq T_{(n)}$ et les $\delta_{(i)}$ sont les indicatrice correspondantes .

l'estimateur de $Kaplan\ Mieir$ peut aussi se mettre sous la forme suivante

$$\hat{S}_{KM}(t) = \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{1}{n-i+1}\right)^{¶\left\{T_{(i)} \leq t\right\}} \text{ si } \ t \prec T_{(n)} \\ 0 \qquad\qquad\qquad\qquad\qquad\quad \text{si } t \geq T_{(n)} \end{cases}$$

$$\bar{G}_n(t) = \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{1}{n-i+1}\right)^{¶\left\{T_{(i)} \leq t\right\}} \text{ si } \ t \prec T_{(n)} \\ 0 \qquad\qquad\qquad\qquad\qquad\quad \text{si } t \geq T_{(n)} \end{cases} \tag{1.3}$$

1.10 Estimateur à noyau de la fonction régression pour des données censurées

Dans cette partie on défini l'estimateur à noyau de la régression dans le cas d'un modèle censurée . Ensuite on donne les hypothèses utilisée pour montré la convergence uniforme presque sûr de cet estimateur .Considérant une variable aléatoire réelle $(v.a)$ Y et une suite de varaibles aléatoires réelles $(Y_i)_{i\geq 1}$ de même fonction de répartition absolument continue et inconnue $(f.r)$ F et soit $(C_i)_{i\geq 1}$,une suite de varaibles alétoire censurées de même $(f.r)$ inconnue G. Soit X un vecteur alétoire dans $\mathbb{R}^d$. Et soit $(X_i)_{i\geq 1}$ une suite de copies de vecteur alétoire X et indiqué par $X_{i;1},...,X_{i;d}$, cordonnées de X_i.Contrairement au modéle avec données completes , le modèle censuré utilise la suite des observations $(T_i,\delta_i,X_i)_{i\geq 1}$, où $T_i\wedge C_i$ et $\delta_i=¶_{\{Y_i\leq C_i\}}$ sont observées .

Supposons que $(Y_i)_{i\geq 1}$ et $(C_i)_{i\geq 1}$ soient deux suite de variable aléatoires stationnaires indépendantes . Posons

$$m(x)=E(Y/X)=\frac{\int_{\mathbb{R}} y\ f_{X,Y}\ (x,y)\,dy}{\ell(x)}=:\frac{r_1(x)}{\ell(x)} \tag{1.4}$$

où $f_{X,Y}(.,.)$ est la densité conjointe de (X,y)et $\ell(.)$, la fonction de densité de X.

Il est bien connu que l'estimateur à noyau de la fonction régression $m(.)$ dans le cas censurées (voir , par exemple $Carbonez$ et est donné par :

$$\tilde{m}_n(x)=\sum_{i=1}^{n}W_{in}(x)\frac{\delta_i T_i}{G(T_i)} \tag{1.5}$$

où $\bar{G}$ est la fonction de survie de la v.a C,

$$W_{in}(x)=\frac{K_d\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n}K_d\left(\frac{x-X_i}{h_n}\right)}$$

, sont les poids de Watson-Nadaraya

et K_d est la fonction de densité de probabilité défini sur $\mathbb{R}^d$ et h_n une suite de nombre positif converge vers 0 quand n tend vers l'infini ∞. Ainis (1.6) peut s'écrire comme :

$$\tilde{m}_n(x) =: \frac{\tilde{r}_{1,n}(x)}{\ell_n(x)} \tag{1.6}$$

avec

$$\tilde{r}_{1,n}(x) = \frac{1}{nh^d}\sum_{i=1}^{n}\frac{\delta_i T_i}{G(T_i)}K_d\left(\frac{x-X_i}{h_n}\right) \qquad et \qquad \ell_n(x) = \frac{1}{nh^d}\sum_{i=1}^{n}K_d\left(\frac{x-X_i}{h_n}\right) \tag{1.7}$$

En pratique , puisque $\bar{G}$ est généralement inconnue , on le remplace par l'estimateur de Kaplan- Mieir EKM $\bar{G}_n$ correspondant , alors un estimateur possible de $m(x)$ est donné par :

$$m_n(x) = \frac{\frac{1}{nh^d}\sum_{i=1}^{n}\frac{\delta_i T_i}{G(T_i)}K_d\left(\frac{x-X_i}{h_n}\right)}{\frac{1}{nh^d}\sum_{i=1}^{n}K_d\left(\frac{x-X_i}{h_n}\right)} =: \frac{r_{1,n}(x)}{\ell_n(x)} \tag{1.8}$$

Chapitre 2

Estimation par la méthode du noyau

Le concept de noyau a d'abord été introduit par ROSENBLATT(1956), mais c'est CACOULOS(1966) qui a été le premier à utiliser le terme "noyau" pour désigner la fonction que l'on utilise dans les méthode non paramétriques. En hydrologie , c'est YAKOTZ(1983) et FELUCH(1983) qui ont introduit indépendamment la méthode des noyaux lors d'une conférence de l'AGU à l'automne 1983.

Dans la méthode des noyaux , une fonction K est associée à chaque observation de l'échantillon. La seule véritable restriction concernant le noyau K est que son intégration de l'échantillon .La seule véritable restriction concernant le noyau K est que son intégration sur tout le domaine de définition de x doit être égale à un .On rencontre parfois d'autres restrictions théoriques qui sont appliquées à K, comme la symétrie ou la positivté sur tout le domaine de définition du noyau (ADAMOWSKI,1989).Toutefois, ces restrictions sont surtout introduites afin desimplifier les développements téoriques. L'estimation non paramétrique de la fonction de densité peut se voir comme le cumul des fonction K de chaque observation sur tout le domaine :

Supposons que nous observons n variables aléatoires i.i.d$X_1, ..., X_n$ de densité f . L'objectif de notre étude est la construction d'un estimateur de f en un poit fixe x.

Notons $F(x) = P(X_1 \leq x)$ la fonction de répartition , ce qui permet d'écrire pour tout x :

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

Considérons la fonction répartition empirique :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}, \forall x \in \mathbb{R}$$

La loi des grands nombres permet d'affirmer que $F_n$est un estimateur de$F$,c'est-à-dire

$$F_n(x) \xrightarrow{p} F(x) \tag{2.1}$$

De plus , le théorème de Glivenko-Cantelli nous donne :

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to_{a.s} 0 \tag{2.2a}$$

Ilest même possible d'obtenir des intervalles de confiance et de tester l'adéquation des données à différentes lois. Néanmoins , il n'est pas évident d'utiliser F_n pour estimer f.

Une des premiéres idées intuitives est de considérer pour$h \succ 0$ fixé "petit" :

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{2nh} \sum_{i=1}^{n} 1_{\{-h \leq X_i - x \leq h\}}.$$

on a alors :

$$E[f_n(x)] = \frac{1}{2h}(E[F_n(x+h)] - E[F_n(x-h)]) = \frac{1}{2}(F(x+h) - F(x-h))$$

$E[f_n(x)]$ tend vers $f(x)$ quand h$\to$ 0.Ilfaut donc faire dépendre h de la taille de l'échantillon, et le faire tendre vers0 quant $n \to \infty$, de sorte que$f_n(x)$ soit un estimateur asympotiquement sans biais de $f(x)$.

L'estimateur f_n reste une fonction en escalier . Pour obtenir quelque chose de plus lisse , on peut remarquer que :

$$f_n(x) = \frac{1}{2nh_n}\sum_{i=1}^{n} 1_{]x-h_n,x+h_n[}(X_i) = \frac{1}{nh_n}\sum^{n}\frac{1}{2}1_{\{x-h_n \prec X_i \prec x+h_n\}} \tag{2.3}$$

$$= \frac{1}{nh_n}\sum_{i=1}^{n}\frac{1}{2}1_{[-1,1[}\left(\frac{x-X_i}{h_n}\right) = \frac{1}{nh_n}\sum_{i=1}^{n}k\left(\frac{x-X_i}{h_n}\right) \tag{2.4}$$

cet estimation appelé estimateur de Rosenblatt, est le premier exemple d'estimateur à noyau construit à l'aide du noyau $K(u) = \frac{1}{2}1_{\{-1\prec u\leq 1\}}$.

2.1 Noyaux

Définissons maintenant plus généralement la notion d'estimateur à noyau :

Définition 2.1.1 *soit $k : \mathbb{R} \to \mathbb{R}$. on dit que$K$est un noyau si et seulement si :*

$$\int K(u)\,du = 1$$

– K est dit positif si$K(u) \geq 0 \forall u$.

– K est dit symétrique si $K(u) = K(-u)\,\forall u$

2.1.1 Exemples de noyaux

voici quelques exemples de noyaux les plus communément utilisés :

• le **noyau rectangulaire :** $K(u) = \frac{1}{2}1_{[-1,+1[}(u)$ .C'est celui qui donne l'estimateur de type histogramme appelé **noyau de Rosenblatt.**

• le **noyau triangulaire :** $K(u) = (1-|u|)\,1_{[-1,+1[}(u)$

• le **noyau d'Epanechnikov :** $\frac{3}{4}\left(1-u^2\right)1_{[-1,+1[}\left(u\right)$

• le **noyau de Tukey ou biweight :** $\mathbf{K}\left(u\right)=\frac{15}{16}\left(1-u^2\right)^2 1_{[-1,+1[}\left(u\right)$

• le **noyau gaussien :** $\mathbf{K}\left(u\right)=\frac{1}{\sqrt{2\pi}}\exp^{-u^2/2}, u\in\mathbb{R}$

Les deux premiers ont l'avantage d'être simples , le noyau triangulaire étant continu partout et conduisant à une estimation f_n continue . Le troisiéme doit sa notoriété à une propriété d'optimalité théorique mais sans grand intérêt pratique .Le quatriéme est , à notre sens ,le plus intéressant car donnant une estimation dérivable partout , tout en étant simple à mettre en oeuvre. En fait il s'agit du noyau le plus simple parmi les noyaux de forme polynomiale dérivables partout .Ainsi il assure le lissage local de la fonction f_n .Ce noyau est d'une forme très proche du noyau Gaussien et il est donc préférable. Notons que plus la valeur de h est élevée plus on élargit la fenêtre, ce qui donne un effet de lissage globale de f_n plus important .

Voici quelques courbes de noyaux usuels présentées ci-dessous :

*FIG.*1.2*−Les courbes des noyaux les plus communs*

2.2 Estimateur à noyau

L'estimateur à noyau est probablement l'estimateur le plus utilisé et certainement le plus utilisé et certainement le plus étudié mathématiquement, car il posséde des propriétés qui le rendent fort intéressant.

Définition 2.2.1 . *Un* ***estimateur à noyau*** *noté* f_n *de la fonction f est défini par :*

$$f_n\left(x\right)=\frac{1}{nh_n}\sum_{i=1}^{n}k\left(\frac{x-X_i}{h_n}\right) \tag{2.5}$$

où$\{h_n\}_{n\geq 1}$ est une suite de réels positifs appelés **Paramètres de lissage** ou **largeur de la fenêtre ,** qui tend vers 0 quand n tend vers l'infini .

Comme nous allons le voir par la suite, si le noyau K est une fonction de densité alors l'estimateur à noyau f_n est lui aussi une fonctin de densité . De plus , ce dernier posséde les propriétés de continuité et de différentiabilité.De sorte que si , par exemple , K est la densité normale alors f_n possède des dérivées de tout ordre.

Proposition 2.2.1 . *Un estimateur à noyau est une densité*

Démonstration :

$$\int_{-\infty}^{+\infty} f_n(x)\,dx = \frac{1}{nh_n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} k\left(\frac{x-X_i}{h_n}\right)dx$$

$$= \frac{1}{nh_n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} k(u)\,h_n du \left(\text{changement de variable u}=\left(\frac{x-X_i}{h_n}\right)\right)$$

$$= \frac{1}{nh_n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} k(u)\,du = \frac{1}{n}n = 1$$

Pour mieux saisir l'intuition derrière l'estimateur à noyau , nous avons construit cet estimateur à partir de l'équation(2.5) en utilisation un ensemble de données constitué seulement

de 7 observations . Le noyau K a été choisi comme étant la densité d'une loi normale de moyenne 0 et de variance 1 et le paramètre de lissage h égale à 4. On centre d'abord un noyau individuel sur chacune des 7 observations et la valeur de l'estimateur à noyau $f(z)$ au point z est simplement la somme des ordonnées de chacun des 7 noyaux individuels à ce point x comme représenté à la figure $(2,2)$ .Dans une région où l'on a plusieurs observations, la vraie densité a une valeur relativement grande et l'estimateur de la densité,par la méthode du noyau, nous donne effectivement une valeur relativement grande ce qui est observé dans la figure $(2,2)$

FIG.2.2−Estimateur à noyau basé sur 7 observation (h=4)

par exemple si x=5 on a $\hat{f}(x) = 0.03$ qui est égale a la somme des densités des 7 noyaux gaussiens au même point x=5

2.3 2.3 Propriétés de l'estimateur à noyau

Nous allons maintenant donner quelques propriétés statistiques élémentaires de l'estimateur de la densité à noyau ainsi que différentes méthodes pour choisir le paramètre de lissage .

2.3.1 Etude du biais

supposons que l'on dispose d'un échantillon d'observation $X_1, ..., X_n$, issu d'une v.a X possédant pour fonction de densité la fonction f que l'on désire estimer .On suppose que f_n est l'estimateur à noyau obtenu en utilisant le noyau K et le paramètre de lissage h et h et f_n défini par l'équation (2.5) . Supposons que :

$$K(u) \geq 0, \int k(u)\,du = 1, \int k(u)\,udu = 0, \int u^2 K(u)\,du \prec \infty$$

et en supposant que la densité de probabilité f admet les deux premières dérivées (continues) nécessaire.

$$E[f_n(x)] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}E\left[k\left(\frac{x-X_i}{h}\right)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\int_{-\infty}^{+\infty} k\left(\frac{x-t}{h}\right) f(t)\,dt$$

$$= \frac{1}{h}\int_{-\infty}^{+\infty} k\left(\frac{x-t}{h}\right) f(t)\,dt$$

$$\text{La transformation} z = \frac{x-t}{h}, i.e. t = -hz + x, \left|\frac{dz}{dt}\right| = \frac{1}{h}$$

$$E[f_n(x)] = \int_{-\infty}^{+\infty} k(z)\, f(x-hz)\,dz$$

Un développement de Taylor de $f(x-hz)$ nous donne :

$$f(x-hz) = f(x) - hzf'(x) + \frac{1}{2}(hz)^2 f''(x) + o\left(h^2\right)$$

$$E\left[\left(f_n(x)\right)\right] = \int_{-\infty}^{+\infty} k(z) f(x)\, dz - \int_{-\infty}^{+\infty} k(z) hzf'(x)\, dz + \int_{-\infty}^{+\infty} k(z) \frac{(hz)^2}{2} f''(z)\, dz + o\left(h^2\right)$$

$$= f(x) \int_{-\infty}^{+\infty} k(z)\, dz - hf'(x) \int_{-\infty}^{+\infty} zk(z)\, dz + \frac{h^2}{2} f''(x) \int_{-\infty}^{+\infty} z^2 k(z)\, dz + o\left(h^2\right)$$

$$= f(x) + \frac{h^2}{2} k_2 f''(x) + o\left(h^2\right)$$

$$Biais\left(f_n(x)\right) \approx \frac{h^2}{2} k_2 f''(x) + o\left(h^2\right) \tag{2.6}$$

Le biais dépend de :h :paramètre de lissage. k_2 : la variance du noau. $f''(x)$ la seconde dérivée de la fonction de densité au point x.

2.3.2 Etude de la variance

La variance de$f_n(x)$ est donnée par :

$$Var\left(f_n(x)\right) = var\left(\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right)\right)$$

$$= \frac{1}{n^2 h^2} \sum_{i=1}^{n} Var\left(k\left(\frac{x-X_i}{h}\right)\right)$$

car les $X_i, i = 1, ..., n$, sont i.i.d

$$Var\left(k\left(\frac{x-X_i}{h}\right)\right) = E\left[k\left(\frac{x-X_i}{h}\right)^2\right] - \left(E\left[k\left(\frac{x-X_i}{h}\right)\right]\right)^2$$

$$= \int k\left(\frac{x-X_i}{h}\right)^2 f(t)\, dt - \left(\int k\left(\frac{x-X_i}{h}\right) f(t)\, dt\right)^2$$

$$Var\left(f_n(x)\right) = \frac{1}{n} \int \frac{1}{h^2} k\left(\frac{x-X_i}{h}\right)^2 f(t)\, dt - \frac{1}{n}\left(\frac{1}{h} \int k\left(\frac{x-t}{h}\right) f(t)\, dt\right)^2$$

$$= \frac{1}{n}\int \frac{1}{h^2}k\left(\frac{x-X_i}{h}\right)^2 f(t)\,dt - \frac{1}{n}(f(x) + Biais\,(f_n(x))^2$$

En effectuant le changement de variable suivant $z = \frac{x-t}{h}$, on obtient :

$$Var\,(f_n(x)) = \frac{1}{nh}\int k(x)^2 f(x-hz)\,dz - \frac{1}{n}\left(f(x) + o\left(h^2\right)\right)^2$$

Et en effectuant un développement limité à l'ordre 2, il vient :

$$Var(f_n(x)) = \frac{1}{nh}\int K(z)^2 (f(x) - hzf'(x) + o(h))\,dz - \frac{1}{n}\left(f(x) + o\left(h^2\right)\right)^2$$

$$Var(f_n(x)) = \frac{1}{nh}f(x)\int K^2(z)\,dz + o\left(\frac{1}{nh}\right)$$

d'où :

$$Var(f_n(x)) \approx \frac{1}{nh}f(x)\int K^2(z)\,dz$$

Discussion du comportement du biais et de la variance :

1. – Le biais décroit si h diminue mais la variance augmente.
   – La variance diminue si h audmentemais le biais augmente.
   – Pour que la variance tende vers zéro ,il faut que $nh \to \infty$.
   – Plus la courbure de la densité est haute en x , plus le biais est grand.
   – La variance est plus grande pour des valeurs plus grandes de la densité.

La figure suivante nous permet de mieux voir le comportement du biais et de la variance .

FIG.1.3-Le"trad-off"biais-variance en fonction de h.

La variance est repésentée par la courbe en pointillé et le biais par la courbe fine, la courbe en gras repésente le MSE.

Chapitre 3

Convergence presque complète de l'estimateur de la fonction de régression

On va présenter en bref quelques outils probabilistes. Parmi ces outils, ceux qui sont reformuler dans des nouveaux types a. . . n de les rendre simplement appliquable pour les modèles non paramétrique fonctionnels. Ces nouvelles formulations seront

également utiles pour toute personne intéressée pour le développement de nouvelles avancées sur l'etude asymptotiques en statistiques fonctionnels non paramétriques. Cependant, l'obtention des résultats asymptotiques
nécessite l'utilisation des outils de probabilité de base pour variables aléatoires réelles et de nombreux résultats présentés ci-dessous concernent les variables aléatoires réelles. La notion de convergence presque complèteet met l'accent sur le lien entre ce mode de con vergence et d'autres modes standards (tels que la convergence presque sûre ou la convergence en probabilité).

Définition 3.0.1 *On dit que la suite* $(Xn)_{n\in\mathbb{N}}$ *converge presque complète vers la variable*

aléatoire réelle X , si et seulement si :$\exists\exists$

$$\forall n \in \mathbb{N}, \sum_{n\in\mathbb{N}} (P\left|X_n - X\right| > \varepsilon) < \infty \tag{3.1}$$

et on note la converge presque complète de $(Xn)_{n\in\mathbb{N}}$ vers X par :

$$\lim X_n = X \ p.co \tag{3.2}$$

Définition 3.0.2 *On dit que la vitesse de convergence presque complète de* $(Xn)_{n\in\mathbb{N}}$ *vers* X *est d'ordre* u_n *si et seulement si :*

$$\exists \epsilon > 0, \sum_{n\in\mathbb{N}} (P\left|X_n - X\right| > \varepsilon u_n) < \infty \tag{3.3}$$

et on écrit :

$$X_n - X = O(u_n) p.co \tag{3.4}$$

En se basant sur la preuve donnée dans Ferraty et Vieu (2003), nous traitons dans ce paragraphe la convergence presque comlète de l'estimateur à noyau de la fonction de régression , auxquelles nous rajoutons les hypothèses suivantes :

h1.**f. r** sont des fonctions continues au voisinage de x, un point fixé de **R** .

h2. le paramètre de lissage h_n est tel que :

$$\lim_{n\to\infty} h_n = 0 et \lim_{n\to\infty} \frac{\log n}{n h_n} = 0$$

h3.La densité f et la variable Ysont telles que :$f(x) \succ 0$

h4.$|Y| \prec M \prec \infty$, où M est une constante réelle positive.

h5.$\int_R k(x)\, dx = 1$

h6.k est borné , intégrable et à support compact ,

h7.$\leq Z_n p\left[\frac{1}{n}\sum |u_i - E[u_i]| \succ \varepsilon\right]$,

Sous les hypothèses$h1.h2.h3.h4.h5.h6$

$$\lim_{n\to\infty} \hat{r}_{h_n}(x) = r(x)$$

Démonstration

La démonstration de ce théorme est basée sur la décompsition suivante :

$$\hat{r}_{h_n}(x) - r(x) = \frac{1}{\hat{f}_n(x)} \left[(\hat{\phi}_{h_n}(x) - E\left(\hat{\phi}_{h_n}(x)\right)) + (E\left(\hat{\phi}_{h_n}(x)\right) - \phi(x)) \right]$$

$$+ \left[\left(f(x) - E\left(\hat{f}_{h_n}(x)\right) \right) + \left(E\left(\hat{f}_{h_n}(x)\right) - \hat{f}_{h_n}(x) \right) \right] \frac{r(x)}{\hat{f}_{h_n}(x)},$$

où $\phi(x) = f(x) r(x)$ .Le résultat énoncé découle des lemmes suivants : ■

Lemme 3.0.1

D'après les hypothèses **h1,h4,h5,h6** on a

$$\lim_{n\to\infty} E\left(\hat{\phi}_{h_n}(x)\right) = \phi(x)$$

Démonstration

on a :

$$E\left(\hat{\phi}_{h_n}(x)\right) = E\left[\frac{1}{nh_n} \sum_{i=1}^{n} Y_i k\left(\frac{x - X_i}{h_n}\right) \right]$$

$$= \frac{1}{h_n} E\left[Y k\left(\frac{x - X_i}{h_n}\right) \right].$$

Le conditionnement par rapport à $X = x$ nous donne :

$$E\left[\hat{\phi}_{h_n}(x)\right] = \frac{1}{h_n} E\left[r(x) k\left(\frac{x - X}{h_n}\right) \right],$$

où :

$$E\left(\hat{\phi}_{h_n}(x)\right) = \frac{1}{h_n}\int r(t)\, k\left(\frac{x-t}{h_n}\right) f(t)\, dt$$

$$= \frac{1}{h_n}\int \phi(t)\, k\left(\frac{x-t}{h_n}\right) dt,$$

en utilisant le changement de variable $u = \frac{x-t}{h_n}$,on obtient ;

$$E(g(x)) = \int \phi(x - uh_n)\, k(u)\, du,$$

comme k est à support compact , la continuité uniforme de ϕ et l'hpothèse **h5** ,

nous donnent :

$$\lim_{n\to\infty} E\left(\hat{\phi}_{h_n}(x)\right) = \phi(x).$$

■

Lemme 3.0.2 *l*

D'après les hypothèses **h1,h4,h5,h6** on a :

$$\lim_{n\to\infty} E\left(\hat{\phi}_{h_n}(x)\right) - \hat{\phi}_n(x) = 0$$

Démonstration

on a

$$\hat{\phi}_{h_n}(x) - E\left(\hat{\phi}_{h_n}(x)\right) = \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

où

$$Z_i = \frac{1}{h_n}\left[Y_i k\left(\frac{x-X_i}{h_n}\right) - E\left(Y_i k\left(\frac{x-X_i}{h_n}\right)\right)\right].$$

De plus , les hypothèses **h4,h6**, nous donnent :

$$\left|k\left(\frac{x-X_i}{h_n}\right)\right| \leq M \Longrightarrow Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right| \leq M^2$$

$$\Longrightarrow Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right| - E\left(Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right|\right) \leq M^2 - E\left(Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right|\right)$$

$$\Longrightarrow \frac{Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right| - E\left(Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right|\right)}{h_n} \leq \frac{M^2 - E\left(Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right|\right)}{h_n}$$

$$\Longrightarrow |Z_i| \leq \frac{M^2 - E\left(Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right|\right)}{h_n}$$

$$|Z_i| \leq \frac{C}{h_n}$$

où

$$C = M^2 - E\left(Y_i\left|k\left(\frac{x-X_i}{h_n}\right)\right|\right),$$

d'autre part

$$E\left(Z_i^2\right) = var\left(\frac{1}{h_n}Y_i k\left(\frac{x-X_i}{h_n}\right)\right) \leq E\left(T_i\right),$$

où :

$$T_i = \frac{1}{h_n}Y_i k\left(\frac{x-X_i}{h_n}\right).$$

En utilisant le conditionnement par rapport à la variable X, on obtient :

$$E\left[T_i^2\right] = \frac{1}{h_n^2} E\left[\phi\left(x\right) k^2\left(\frac{x - X_i}{h_n}\right)\right]$$

$$= \frac{1}{h_n^2} \int \phi\left(x\right) k^2\left(\frac{x - t}{h_n}\right) f\left(t\right) dt,$$

où :

$$\phi\left(x\right) = E\left(Y^2 / X = x\right),$$

en utilisant le changement de variable $z = \frac{x-t}{h_n}$ on obtient :

$$E\left(T_i^2\right) = \frac{1}{h_n} \int \phi\left(x - z h_n\right) k^2\left(z\right) f\left(x - z h_n\right) dz,$$

La continuité de f sur le support compact k , les hypothèses **h4,h5** impliquent :

$$E\left(T_i^2\right) \leq \frac{C}{h_n}.$$

comme les conditions du corollaire 7 étant satisfaites , alors nous déduisons qui :

$$\frac{1}{n} \sum_i Z_i = O_{p.co}\left(\sqrt{\frac{\log n}{n h_n}}\right).$$

La convergence presque complète de la densité établie dans le paragraphe précédent assure les convergences suivantes :

$$\lim_{n \to \infty} E\left(\hat{f}_{h_n}\left(x\right)\right) = f\left(x\right)$$

$$\lim_{n \to \infty} E\left(\hat{f}_{h_n}\left(x\right)\right) - f\left(x\right) = 0$$

■

Lemme 3.0.3 *l*

D'après les hypothèses **h1,h2,h5,h6** on a

$$\exists \delta \succ 0, \sum_{n \in N} p\left(\hat{f}_{h_n}(x) \leq \delta\right) \prec \infty.$$

entraîne la convergence presque complète de$\hat{f}_{h_n}(x)$ vers $f(x)$, c'est à dire;

$$\forall \varepsilon \succ 0, \sum_{n \in N} p\left(\left|\hat{f}_{h_n}(x) - f(x)\right| \succ \varepsilon\right) \prec \infty,$$

on a

$$\left|\hat{f}_{h_n}(x)\right| \leq \frac{f(x)}{2} \Longrightarrow \left|\hat{f}_{h_n}(x) - f(x)\right| \succ \frac{f(x)}{2}$$

d'ou

$$P\left[\left|\hat{f}_{h_n}(x) \leq \frac{f(x)}{2}\right|\right] \leq P\left[\left|\hat{f}_{h_n}(x) - f(x)\right| \succ \frac{f(x)}{2}\right]$$

comme $f(x) \succ 0$, en posant $\delta = \varepsilon = \frac{f(x)}{2}$, on arrive au résultat.

h8. r et f est k fois continûment dérivable autour du point x.

h9.Le noyau K est tel que :

k est d'ordre i au sens de Gasser c'est à dire :

$$\int t^i k(t)\, dt = 0, \forall j = 1, 2, ..., i-1.et0 \prec \left|\int t^i k(t)\, dt\right| \prec \infty.$$

Théorème 3.0.1 *t*

Considérons le modèle 8 avec $k \succ 0$,et supposone que les hypothèses 2,9,4,3,6 soient réalisées , alors on a :

$$|\hat{r}_{h_n}(x) - r(x)| = O\left(h_n^k\right) + O\left(\sqrt{\frac{\log(n)}{n}}\right) p.co$$

En utilisant la décomosition précédente , Le résultat énoncé découle des lemmes suivants :

Lemme 3.0.4 *l*

Sous les hypothèses 9,8,6 on a :

$$E\left(\hat{\phi}_{h_n}(x)\right) - \phi(x) = O\left(h_n^k\right). \tag{3.5}$$

Démonstration

L'expression de $\hat{\phi}_{h_n}(x)$ est analogue à la précédente .En effet , on a :

$$E\left(\hat{\phi}_{h_n}(x)\right) = \int_{-\infty}^{+\infty} k(z)\,\phi(x - zh_n)\,dz.$$

Le modèle 8, nous permet de développer ϕ au voisinage de x, ceci nous permet d'écrire :

$$\phi(x - zh_n) = \phi(x) + \sum_{i=1}^{k-1} \frac{(-1)^i (zh_n)^i}{i!}\phi^i(x) + \frac{(-1)^k (zh_n)^k}{k!}\phi^k(\theta_z),$$

où θ_z entre x et $x - zh_n$.

L'hypothèse 9 sur k , implique :

$$E\left(\hat{\phi}_{h_n}(x)\right) = \phi(x) + (-1)^k h^k \frac{\int z^k \phi^k k(z)(\theta_z)\,dz}{k!}.$$

La convergence uniforme de $\phi^k(\theta_z)$ vers $\phi^k(x)$ (assurée par le modèle 8) et la condition 6, nous donnent

$$E\left(\hat{\phi}_{h_n}(x)\right) - \phi(x) = O\left(h_n^k\right).$$

■

Les lemmes (3.02) (3.03) (3.04)

$$E\left(\hat{f}_{h_n}(x) - f(x)\right) = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

$$E\left(\hat{g}_{h_n}(x) - g(x)\right) = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Et :

$$\exists \delta \succ 0, \sum_{n\in\mathbb{N}} p\left(\hat{f}_{h_n}(x) \leq \delta\right) \prec \infty$$

En combinant tous les résultats cités précédemment, on arrive au résultat cherché .

Nous essayons maintenant d'établir une vitesse convergence presque comlète uniforme. Il suffit d'une part de considérer un compact S de R tel que l' hypothèse 7 soit remplacée par le modèle suivant .

B1. r et f sont k continûment dérivables sur S .

Et de supposer d'autre part l'existence de $\theta \succ 0$, tel que :

$$\inf_{x\in S} f(x) \succ \theta \tag{3.6}$$

B2.Choisir un compact S de R sur lequel f est k fois continûment dérivable.

♣ il existe $C \prec 1, \forall x \in S, \forall y \in S$

$$|k(x) - k(u)| \leq C|x - y| \tag{3.7}$$

Nous avons aussi besoin de l'hypothèse suivante sur le paramètre de lissage,

♣ il existe $\varepsilon \prec \infty$, telle que :

$$\lim_{n\to\infty} n^{2\varepsilon-1}h_n = +\infty \tag{3.8}$$

Nous gardons toutes les conditions citées précédemment , auxquelles nous rajoutons la condition Lipschitzienne(3.3)surle noyau K et l'hypothèse (3.4) sur le paramètre de lissage h .

Théorème 3.0.2 *t*

soient les modèles **B1**, avec$k \succ 0$ et les hypothèse $h6$, $h2, h3$,(3.3) et$h4$, on a

$$\sup_{x\in s}\left(\left|\hat{r}_{h_n}(x) - r(x)\right|\right) = O\left(h_n^k\right) + O\left(\sqrt{\frac{\log n}{n}}\right), p.co$$

Une décomposition similaire au cas ponctuel, nous permet d'écrire :

$$\sup_{x\in S}\left(\left|\hat{r}_{h_n}(x) - r(x)\right|\right) \leq \frac{\sup_{x\in s}\left(\left|\hat{\phi}_{h_n}(x) - \phi(x)\right|\right)}{\inf_{x\in S}\left|\hat{f}_{h_n}(x)\right|} + \sup_{x\in S}\left(\left|f(x) - \hat{f}_{h_n}(x)\right|\right)\frac{\sup_{x\in S}\left(\left|\hat{r}_{h_n}(x)\right|\right)}{\inf_{x\in S}\left|\hat{f}_{h_n}(x)\right|}$$

$$\leq \frac{\sup_{x\in s}\left|\hat{\phi}_{h_n}(x) - E(\hat{\phi}_{h_n}(x))\right|}{\inf_{x\in S}\left|\hat{f}_{h_n}(x)\right|} + \frac{\sup_{x\in s}\left|E(\hat{\phi}_{h_n}(x)) - \hat{\phi}_{h_n}(x)\right|}{\inf_{x\in S}\left|\hat{f}_{h_n}(x)\right|}$$

$$+\left\{\sup_{x\in S}\left|f(x) - E(\hat{f}_{h_n}(x))\right| + \sup_{x\in S}\left(\left|E(\hat{f}(x)) - \hat{f}_{h_n}(x)\right|\right)\right\}\frac{\sup_{x\in S}\left(\left|\hat{r}_{h_n}(x)\right|\right)}{\inf_{x\in S}\left|\hat{f}_{h_n}(x)\right|}$$

Les approximations en $O\left(h_n^k\right)$ traitées précédemment peucent se généraliser via $(h, 6)$ et h8 comme suit :

$$\sup_{x\in S} E(\hat{f}(x)) - f_{h_n}(x)) = O\left(h_n^k\right)$$

et

$$\sup_{x\in S} E(\hat{g}(x)) - g_{h_n}(x)) = O\left(h_n^k\right)$$

comme r est borné.la preuve de ce théorème s'achèvera à partir des lemmes suivants :

Lemme 3.0.5 *l*

Sous les hypothèses **h2**,**B1**,(3.3) (3.4) on a

$$\sup_{x\in S}\left|E\left[\hat{\phi}_{h_n}(x)\right]-\hat{\phi}_{h_n}(x)\right|=O\left(\sqrt{\frac{\log(n)}{n}}\right),p.co$$

démonstration

S est un compact de $\mathbf{R}$, il existe un recouvrement fini de S tel que :

$$S\subset\cup_{k=1}^{z_n}S_k$$

où

$$S_k=]t_k-l_n,t_k+l_n[\text{ et }l_n=n^{-\beta}$$

Posons

$$t_x=\arg\min_{t\in t_1,t_2,....,t_{z_n}}|x-t|$$

avec

$$l_n=n^{-2\epsilon},l_n=Cz_n^{-1}$$

On a

$$Sup_{x\in S}\left|E\left[\hat{\phi}_{h_n}(x)\right]-\hat{\phi}_{h_n}(x)\right|\leq A_1+A_2+A_3$$

où

$$A_1=Sup_{y\in S}\left|\left[\hat{\phi}_{h_n}(x)\right]-\hat{\phi}_{h_n}(t_x)\right|,$$

$$A_2=Sup_{y\in S}\left|\left[\hat{\phi}_{h_n}(t_x)\right]-E\left[\hat{\phi}_{h_n}(t_x)\right]\right|,$$

$$A_2=Sup_{y\in S}\left|\left[E\left[\hat{\phi}_{h_n}(t_x)\right]-\hat{\phi}_{h_n}(x)\right]\right|.$$

Concernant le terme A_1, comme le noyau k est lipschitzien et la variable $\mathbf{Y}$ est bornée , on a

$$\left|\hat{\phi}_{h_n}(t_x) - \hat{\phi}_{h_n}(x)\right| = \frac{1}{nh_n}\sum_{i=1}^{n}|Y_i|\left|\left[k\left(\frac{t_x - X_i}{h_n}\right) - k\left(\frac{x - X_i}{h_n}\right)\right]\right|$$

$$\leq \frac{C}{h_n}\frac{|t_x - x_i|}{h_n^2}$$

$$= \frac{Cl_n}{h_n^2}.$$

L'hypothèse (3.4) implique

$$A_1 = o\left(\frac{\log n}{nh_n}\right).$$

Une manière de démonstration analogue à la précédente, nous permet d'écrire

$$A_3 = o\left(\frac{\log n}{nh_n}\right).$$

On ce qui concerne le terme A_2,on a :$\forall \varepsilon \succ 0$.

$$p\left[Sup_{y\in S}\left|\left[\hat{\phi}_{h_n}(t_x)\right] - E\left[\hat{\phi}_{h_n}(t_x)\right]\right|\rangle\varepsilon\right] = p\left[\max_{j=1,\ldots,z_n}\left|\hat{\phi}_{h_n}(t_j) - E\left[\hat{\phi}_{h_n}(t_j)\right]\right|\rangle\varepsilon\right]$$

$$\leq z_n p\left[\left|\hat{\phi}_{h_n}(t_j) - E\left[\hat{\phi}_{h_n}(t_j)\right]\right|\rangle\varepsilon\right]$$

$$\leq z_n p\left[\frac{1}{nh}\sum_{i=1}^{n}|u_i - E[u_i]|\rangle\varepsilon\right].$$

Où

$$u_i = Y_i k\left(\frac{X_i - t_k}{h_n}\right)$$

Il suffit de trouver des majorants pour u_i et $E[u_i^2]$,pour pouvoir appliquer le corollaire() .

D'après la démonstration du lemme () , on a

$$u_i \leq \frac{C}{h_n} \text{ et } E\left[u_i^2\right] \leq \frac{C}{h_n}.$$

Maintenant nous sommes en mesure d'appliquer le corollaire() :

$$p\left[Sup_{y\in S}\left|\hat{\phi}_{h_n}(t_x) - E\hat{\phi}_{h_n}(x)\right|\rangle\varepsilon\right] \leq n^{2\varepsilon}\exp\left(-Cn\varepsilon^2 h_n\right).$$

comme :

$$\lim_{n\to\infty}\sqrt{\frac{\log(n)}{nh_n}} = 0.$$

En posant $\varepsilon = \varepsilon_0\sqrt{\frac{\log(n)}{nh_n}}$, on obtient alors :

$$\forall\varepsilon\rangle 0, \sum p\left[Sup_{y\in S}\left|\hat{\phi}_{h_n} - E(\hat{\phi}_{h_n}(x)\right|\rangle\varepsilon\right] \leq \infty$$

pour ε_0 choisi suffisamment grand ■

Conclusion

La littérature sur l'estimation non- paramétrique est abandantes sur la densité ou la régression, en citant les travaux de Collomb (1983).

Ce mémoire est essentiellement basé sur l'estimation non paramétrique de la fonction de régression, pour laquelle nous avons donné quelques aspects particuliers qui tournent autour des méthodes à noyaux et nous avons mis en évidence le rôle du paramètre de lissage. L'estimateur de Nadarata-Watson a été particulièrement étudié aussi bien dans le cas complet qu'en présence de censures.

Nous avons ensuite rappelé les propriétés fondamentales de certains estimateurs de la fonction de régression en insistant sur l'estimateur de Nadarata-Watson étant donné que nos résultats portent sur cet estimateur. Nous avons donc donné ses propriétés asymptotiques dans le cas d'un modèle complet, et nous avons donné quelques références sur le modèle censuré ce qui nous a permis de situer l'apport de notre travail. Les techniques utilisées dans notre travail sont inspirées des travaux de guessoum (2013) pour déterminer la vitesse de convergence. Nous faisons remarquer que les résultats obtenus l'ont été pour une covariable X dans $\mathbb{R}$; nous pouvons les étendre au cas $\mathbb{R}^d$,

Bibliographie

[1] *Gannoun et al (2003)*

[2] *Samanta et Thavaneswaran(1990)*

[3] *Khardani et al.(2011)etCollomb et al.(1987*

[4] Guessoum et Ould Saïd(2008,2010,2012

[5] Kohler et al.(2002)

[6] Carbonez et al (1995)

[7] Kim et Cox (2001)

Résumé

L'objet de ce mémoire est l'étude asymptotique de l'estimateur de la fonction de régression non paramétrique sous les modèles de données censurées. En étudiant la convergence presque complète et en déterminant les vitesses de convergence. La construction de cet estimateur est basée sur la méthode du noyau crée par Rosanblat.

Mots clés : La convergence presque complète ,la vitesse de convergence , les données censurées et la méthode du noyau.

Abstract

The object of this thesis is the asymptotic study of the estimator of the nonparametric regression function under censored data models. By studying almost complete convergence and determining the convergence speeds. The construction of this estimator is based on the kernel method created by Rosanblat.

Keywords : Convergence almost complete convergence speed, censored data and kernel method.