الجمهوريـة الجزائـريـة الديمقراطيـة الشعبيـة
République Algérienne Démocratique et Populaire
وزارة التعليـم العالـي والبحـث العلمـي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Nº Réf :……

Centre Universitaire
Abd Elhafid Boussouf Mila

Institut des Sciences et Technologie **Département de Mathématiques et Informatiques**

Mémoire préparé en vue de l'obtention du diplôme de Master

En: Mathématiques

Spécialité: Mathématiques fondamentales et appliquées

Estimation de la fonction de régression en analyse de survie

Préparé par :

- Brik Kenza
- Khelifa Beghedouche Safia

Soutenue devant le jury

Encadré par : Mehazem Allal……...M. A. A
Président : Boueden Rabah ……..M. A. A
Examinateur : Elhadj Ali Thouria …….M. A. A

Année Universitaire : 2015/2016

TABLE DES MATIÈRES

Introduction **iii**

1 Estimation de la fonction de régression pour des données complètes **1**

1.1 Rappels sur les modes de convergence 1

1.2 L'estimation non paramétrique pour des données complètes 4

1.2.1 Estimation de la fonction de répartition 5

1.2.2 Estimation de la fonction de densité 7

1.2.3 Estimation de la fonction de régression 13

2 Introduction à l'analyse des durées de survie **21**

2.1 Définitions 22

2.2 Distributions de la durée de survie 22

2.2.1 Cas absolument continue 23

2.2.2 Cas discret 28

2.2.3 Cas général 28

2.3 Données incomplètes (censurées) 29

2.4 L'estimateur de Kaplan-Meier de la fonction de survie 32

3 Estimation de la fonction de régression pour des données censurées **36**

3.1 Censure à droite 37

3.1.1 Principe de l'estimation 38
3.1.2 Propriétés de l'estimateur 40
3.2 Censure à gauche 44
3.2.1 Propriétés de l'estimateur 46

Bibliographie **49**

INTRODUCTION

Le but de la régression est de déterminer la manière dont l'espérance d'une variable expliquée réelle Y dépend d'une variable explicative X scalaire, vectorielle ou même fonctionnelle (ce qui veut dire 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

$$Y = r(X) + \epsilon,$$

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 plutôt à estimer) pour chaque réalisation x de la variable X, la valeur de la fonction $r(x)$.
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 compte tenu d'un critère, par exemple celui 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 de se suffire de la seule donnée de l'échantillon pour réaliser une estimation. Ce sera à l'aide d'un modèle non paramétrique. Dans ce cas on ne dispose d'aucune forme paramétrique pour r mais seulement d'hypothèses générales de régularité comme la dérivabilité. Etudier le lien entre deux variables a généralement pour but de prédire l'une d'entre elles (variable réponse) étant donné une valeur de l'autre (variable explicative).

Une des façons les plus utilisées pour résoudre un problème de prévision est la régression qui est basée sur l'espérance conditionnelle.

Parmis les paradigmes disponibles d'estimation non-paramétrique de la régression c'est l'estimateur à noyau de Nadaraya-Watson (1964) que nous l'étudions au premier chapitre. Ce problème a suscité un grand intérêt et a conduit à la proposition de plusieurs estimateurs sur la base de l'observation d'un échantillon du couple (X, Y). En analyse de survie, il est connu que l'observation de Y n'est pas toujours possible. Y peut être le temps de survie à une maladie . Certains sujets, sous étude, peuvent disparaître de l'étude fortuitement (accident) ou d'une manière planiée (fin de l'étude). Y est alors censuré à droite : on ne connaît pas sa valeur exacte, on sait seulement quelle dépasse l'observation recueillie.

Carbonez (1992) et Carbonez, Györ et van der Meulen (1995) ont introduit un estimateur à partitions consistant pour Y bornée et censurée à droite que nous l'étudions au troixème chapitre. Améliorant ce travail, Kohler, Màthé et Pintér (2002) ont simplié la preuve du résultat précédent et ont même étendu le travail en proposant dautres estimateurs non paramétriques . Cependant dautres types de censure existent. Pour certaines valeurs de Y , on peut seulement savoir quelles sont inférieures aux observations. Cest la censure à gauche. Cela peut être le cas lorsqu'on sintéresse à l'âge auquel des personnes ont commencé à accomplir une tâche. Certains individus peuvent seulement se rappeller quils ont commencé cette tâche avant un certain âge sans savoir exactement lequel.

Afin d'estimer la fonction de régression le présent mémoire est subdivisé en trois chapitres. Dans le premier chapitre nous introduisons quelques rappels sur les différents modes de convergences et nous évoquons, dans ce même chapitre, l'estimation non paramétrique pour des des données complètes des trois fonctions : la fonction de répartition, la fonction

densité et la fonction de régression.

Le deuxième chapitre décrit les notions de base de l'analyse de survie. Et nous définissons les différents types de censure, et nous terminons ce chapitre par introduire l'estimateur de Kaplein -Meier de la fonction de survie que nous l'utilisons au dernier chapitre.

Dans le troisième chapitre nous présenterons l'estimation de la fonction de régression pour ds données censurées (à droite et à gauche), puis nous étudions quelques propriétés asymptotiques de ces estimateurs .

❖

Notations et abréviations

$a \wedge b$	:	$\inf(a, b)$.
$a \vee b$	:	$\max(a, b)$.
$E(X)$	:	L'espérance de X
$Var(X)$	:	La variance
L^2	:	L'espace des fonctions de carrées intégrable
$c - \text{à} - d$	:	Continue à droite .
$m.q$	:	Moyenne quadratique .
$p.s$	:	Presque sure.
$i.i.d$	:	Indépendantes et identiquement distribuées .
$v.a.r$	:	Variable aléatoire réel .
$E(Y \backslash X = x)$	:	L'espérance conditionnelle de Y sachant X.
F'	:	La drivé de F.

CHAPITRE 1

ESTIMATION DE LA FONCTION DE RÉGRESSION POUR DES DONNÉES COMPLÈTES

Le but de ce chapitre est d'introduire des estimateurs non paramétriques des fonctions de répartition, de densité et de régression, basés sur des données complètes, et de donner quelques unes de leurs propriétés. Ces dernières s'énoncent par un ou plusieurs des modes de convergence, que nous rappelons ci dessous.

1.1 Rappels sur les modes de convergence

Il existe de nombreuses notions de convergence de variables aléatoires. Elle sont essentielles pour les applications. Elles servent surtout à montrer que les phénomènes aléatoires présentent certaines régularités, à partir desquelles on peut identifier certaines de leurs propriétés. Dans toute cette section, on suppose que les suites de variables aléatoires réelles (X_n) sont supposées construites sur un espace de probabilité (Ω, A, P).

La convergence en loi (ou en distribution)

On dit que (X_n) converge en loi vers X si la suite des fonctions de répartition des (X_n) converge vers la fonction de répartition de X en tout point de continuité de F.

$$\lim_{n\to\infty} F_n(x) = F(x)$$

On note $X_n \xrightarrow{L} X$.
Bien que cette convergence est la plus faible, elle est très utilisée en pratique car elle permet d'approximer la fonction de répartition de X_n par celle de X.

La convergence en probabilité

On dit que (X_n) converge en probabilité (en mesure) vers X si :

$$\forall \delta > 0, \lim_{n\to\infty} (P|X_n - X| > \delta) = 0$$

on note $X_n \xrightarrow{P} X$.

La convergence en moyenne quadratique

Si $(X_n) \in L^2$ et $X \in L^2$, on dit que (X_n) convergence en moyenne quadratique vers X si :

$$E|X_n - X|^2 = \int |X_n - X|^2 dP \to 0, \;\; quand\, n \to \infty$$

on note $X_n \xrightarrow{m.q} X$, elle correspond à la convergence dans L^2

La convergence presque sûre (ou la convergence forte)

On dit que la suite (X_n) converge vers X presque sûrement s'il existe un ensemble P négligeable $N \subset \Omega$ tel que :

$$\forall \omega \in \Omega / N \Rightarrow \lim_{n\to\infty} X_n(\omega) = X(\omega)$$

On note $X_n \xrightarrow{P.S} X$, elle correspond à la convergence P presque partout.

Et on a aussi

$$X_n \xrightarrow{P.S} X \Leftrightarrow P(\lim_{n\to\infty} X_n = X) = 1$$

La convergence presque sûre est utilisée dans la loi forte des grands nombres, donné par the théorème ci-dessous.

Thoérème 1.1. *(La loi forte des grands nombres)*
Soit $(X_n, n \in N)$ *une suite de variables aléatoires i.i.d (indépendantes et identiquement distribuées) définie sur un espace de probabilité* (Ω, A, P) *et tel que :*
$E|X_1| < \infty$, *pour* $n \in N$ *posons* $S_n = X_1 + \ldots + X_n$ *alors*

$$\frac{S_n(\omega)}{n} \xrightarrow{P.S} E[X_1].$$

La convergence presque complète

On dit que $(X_n)_{n\in N}$ converge presque complètement vers la v.a.r X si

$$\forall \delta > 0, \sum_{n\in\mathbb{N}} P(|X_n - X| > \delta) < \infty.$$

on note $X_n \xrightarrow{P.Co} X$

La relation entre les différents modes de convergence

Les différents modes de convergence rappelés ci-dessus sont liés dans le sens où certains entraîns d'autres, et en résumé on a :

$$X_n \xrightarrow{P.Co} X \Rightarrow X_n \xrightarrow{P.S} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X \text{ et } X_n \xrightarrow{m.q} X \Rightarrow X_n \xrightarrow{P} X$$

Notons que la convergence P.Co entraine à la fois la convergence P.S et la convergence en probabilité.

1.2 L'estimation non paramétrique pour des données complètes

La théorie de l'estimation joue un rôle principale en statistique. Cette théorie est divisée en deux volets principaux, à savoir l'estimation paramétrique et l'estimation non-paramétrique. Cette dernière consiste, généralement, à estimer à partir des observations, une fonction inconnue, élément d'une certaine classe fonctionnelle. Une procédure non-paramétrique est défnie indépendamment de la loi de l'échantillon d'observation. En fait, on parle de méthode d'estimation non-paramétrique lorsque celle-ci ne se ramène pas à l'estimation d'un paramétre appartenant à un espace de dimension finie. Plus généralement un modèle de statistique semi paramétrique comporte à la fois une composante paramétrique et une autre non paramétrique (exemple : le modéle de Cox).

Un des problèmes centraux en statistique est celui de l'estimation de caractéristiques fonctionnelles associées à la loi des observations, comme par exemple, la fonction de répartition, la densité ou la fonction de régression. Dans le modéle de régression non-paramétrique, on suppose l'existence d'une fonction r(x) qui exprime la valeur moyenne de la variable réponse Y en fonction de la variable d'entrée X.

Par ailleurs des donées sont dites complètes si elles correspondent à de véritables réalisations de la variable d'intérêt .

Exemples d'estimation paramétrique et non paramétrique :

1. On observe une suite $Y = (y_1, y_2, \ldots, y_n)$ de variables i.i.d gaussiennes. Estimer leur moyenne.

2. On observe une suite $Y = (y_1, y_2, \ldots, y_n)$ de variables i.i.d uniformes sur $[0, a]$, a est inconnu. Estimer leur moyenne.

3. On observe une suite $Y = (y_1, y_2, \ldots, y_n)$ de variables i.i.d. Estimer leur densité.

-Les problèmes d'estimation 1 et 2 sont paramétriques.

-Le problème d'estimation 3 est non paramétrique car il s'agit d'estimer un vecteur de dimension infinie et dont la loi est inconnue.

les propriétés d'un estimateur :

On suppose que $\Theta_n(x)$ est un estimateur du paramètre $\Theta(x)$, On dit que $\Theta_n(x)$:

1. est sans biais si :
$$E(\Theta_n(x)) = \Theta(x)$$

2. est asymptotiquement sans biais si :
$$\forall x \in \mathbb{R}, \qquad \lim_{n\to\infty} E(\Theta_n(x)) = \Theta(x)$$

3. est consistant si :
$$\forall x \in \mathbb{R}, \qquad \Theta_n(x) \xrightarrow{P} \Theta(x)\ , \qquad quand\ n \longrightarrow \infty$$

   Ceci implique que : $\lim_{n\to\infty} Var(\Theta_n(x)) = 0$ et $\lim_{n\to\infty} E(\Theta_n(x)) = \Theta(x)$.

4. est fortement consistant si :
$$\forall x \in \mathbb{R}, \qquad \Theta_n(x) \xrightarrow{P.S} \Theta(x)$$

5. est fortement et uniformément consistant si :
$$\sup_{x\in\mathbb{R}} |\Theta_n(x) - \Theta(x)| \xrightarrow{P.S} 0\ , \qquad quand\ n \longrightarrow \infty$$

1.2.1 Estimation de la fonction de répartition

La fonction de répartition caractérise complètement la loi d'une variable aléatoire X et permet, par exemple, de calculer la probabilité que X appartienne à un intervalle.

Son estimation est fondamentale en statistique. Soient donc $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ la statistique d'ordre n associée à $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées, de même loi que X, de fonction de répartition inconnue F. Une estimation naturelle de F se fait par la fonction de répartition empirique donnée par

$$\begin{aligned} F_n(x) &= \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} \\ &= \frac{1}{n}\sum_{i=1}^{n} I\{X_{i,n} \leq x\} \\ &= \begin{cases} 0 & si \quad x < x_{1,n} \\ \frac{k}{n} & si \quad x_{k,n} \leq x < x_{k+1,n} \qquad k = 1, \ldots, n-1 \\ 1 & si \quad x \geq x_{n,n} \end{cases} \end{aligned}$$

où $I(.)$ désigne la fonction indicatrice.
La loi forte des grands nombres nous montre que $F_n(x)$ est un estimateur fortement consistant c'est à dire :

$$\forall x \in \mathbb{R}, \qquad F_n(x) \xrightarrow{P.S} F(x), \ \ quand\ n \longrightarrow \infty$$

De plus, la convergence presque sûre est uniforme en $x \in \mathbb{R}$, d'après le résultat suivant.

Thoérème 1.2. *(Glivenko-Cantelli) :*

$$P\Big[\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\Big] = 1$$

Preuve

Voir Renyi, chapitre 07, p. 378.

De plus le théorème central limite permet d'obtenir la normalité asymptotique de cet estimateur.

1.2.2 Estimation de la fonction de densité

Soit X une variable aléatoire continue de loi absolument continue, par apport à la mesure de Lebesgue, et soit f la densité de probabilité de X.

$$f(x) \simeq \frac{F(x+h) - F(x-h)}{2h} \quad \text{pour } h > 0, \text{ assez petit}$$

Rosemblatt (1956) 19 a donné un estimateur de f, en remplaçant F par son estimateur F_n. D'où

$$f_n(x) \simeq \frac{F_n(x+h) - F_n(x-h)}{2h}$$

où F_n est la fonction de répartition empirique. Cet estimateur peut encore s'écrire :

$$\begin{aligned} f_n(x) &= \sum_{i=1}^{n} \frac{I\{x - h_i \leq x + h\}}{2nh} \\ &= \frac{1}{2nh} \sum_{i=1}^{n} I\{-1 \leq \frac{x - X_i}{h} \leq 1\} \\ &= \frac{1}{nh} \sum_{i=1}^{n} K_0(\frac{x - X_i}{h}) \end{aligned}$$

avec

$$K_0(u) = \frac{1}{2} I\{-1 \leq u \leq 1\}$$

Dans le même article, Rosemblatt a mesuré la qualité de cet estimateur, en calculant son biai et sa variance, donnés respectivement par :

$$\begin{aligned} Ef_n(x) - f(x) &= \frac{1}{2h} E(F_n(x+h) - F_n(x-h)) - f(x) \\ &= \frac{1}{2h}(F(x+h) - F(x-h)) - f(x) \end{aligned}$$

et par :

$$Var[f_n(x)] = \frac{1}{4nh_n^2}[F(x+h_n)(1 - F(x+h_n)) + F(x-h_n)(1 - F(x-h_n))]$$

$$-\frac{1}{4nh_n{}^2}[2F(\inf(x-h_n),(x+h_n)))+2F(x+h_n)F(x-h_n)]$$

Nous remarquons que si $h_n \to 0$ et $nh_n \to 0$ quand $n \to \infty$, on a :

$$\lim_{n\to\infty} E[f_n(x)] = f(x)$$

et

$$\lim_{n\to\infty} Var[f_n(x)] = 0$$

alors $f_n(x)$ est un estimateur consistant.

Ainsi une généralisation de cet estimateur a été introduite par Parzen (1962) [16] en posant

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\Big(\frac{X_i - x}{h_n}\Big)$$

où (h_n) est une suite de réels strictement positifs, tendant vers zéro quand $n \to \infty$ (appelée fenêtre) et $K : \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction intégrable, telle que $\int K(u)du = 1$ appelée noyau.

Exemple de noyaux K les plus utilisés dans l'estimation de la densité :

♠$K(u) = \frac{1}{2}I(|u| \leq 1)$: noyau rectangulaire,

♠$K(u) = (1-|u|)I(|u| \leq 1)$: noyau triangulaire,

♠$K(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$: noyau d'Epachnikov ou parabolique,

♠$K(u) = \frac{15}{16}(1-u^2)^2 I(|u| \leq 1)$: noyau quadratique,

♠$K(u) = \frac{1}{\sqrt{2\pi}}exp(-u^2/2)$: noyau gaussien,

♠$K(u) = \frac{1}{\sqrt{2\pi}}exp(-|u|/\sqrt{2})\sin(|u|/\sqrt{2}+\pi/2)$: noyau de Silverman.

L'estimateur de Parzen-Rosenblatt a connu un très grand succès parmi les estimateurs non paramétriques, ceci est dû à sa simplicité et sa convergence vers la densité f pour tous les modes (presque sûre, en probabilité, en moyenne quadratique et presque complète) et il nous laisse aussi le choix sur le noyau K.

1.2.2.1 Propriétés de l'estimateur de Parzen-Rosenblatt

Pour étudier la convergence de cet estimateur on utilise le théorème de Bochner (1955) [3], rappelé ci dessous :

Thoérème 1.3. *(Bochner) :3*
Soit $K : (\mathbb{R}, B(\mathbb{R})) \to (\mathbb{R}, B(\mathbb{R}))$ *(*$(B(\mathbb{R})$ *étant la tribu borélienne de* $\mathbb{R}$*) une application bornée et intégrable telle que :*

$$|z|K(z) \longrightarrow 0, \ \ quand\, |z| \to \infty$$

Par ailleurs,soit $g : (\mathbb{R}, B(\mathbb{R})) \to (\mathbb{R}, B(\mathbb{R}))$ *une application intégrable.*
posons $g_n(x) = \frac{1}{h_n} \int_{\mathbb{R}} K(\frac{z}{h_n}) g(x-z)dz$*, où* $0 < h_n \longrightarrow 0$ *quand* $n \to \infty$
Si g *est continue alors* $\lim_{n\to\infty} g_n(x) = g(x) \int_{-\infty}^{+\infty} K(z)dz$.
Si g *est uniformément continue alors la convergence de* g_n *est uniforme.*

Preuve :

Soit $\epsilon > 0$, on a :

$$\begin{aligned} \Big|g_n(x) - g(x)\int_{-\infty}^{+\infty} K(z)dz\Big| &= \Big|\frac{1}{h_n}\int_{\mathbb{R}} K(\frac{z}{h_n})g(x-z)dz - \int_{-\infty}^{+\infty} g(x)K(z)dz\Big| \\ &= \Big|\int_{-\infty}^{+\infty} K(y)g(x-h_n y)dy - \int_{-\infty}^{+\infty} g(x)K(y)dy\Big| \end{aligned}$$

g étant continue,il vient

$$\forall \epsilon > 0, \exists \delta > 0 / |x-z| < \delta \Rightarrow |g(x) - g(z)| \leq \epsilon.$$

Donc :

$$\begin{aligned} &\int |(g(x-h_n y) - g(x))K(y)|dy \\ &\leq \int_{y:|y|<\frac{\delta}{h_n}} |g(x-h_n y) - g(x)||K(y)|dy + \int_{y:|y|\geq\frac{\delta}{h_n}} |g(x-h_n y) - g(x)||K(y)|dy \\ &\leq \epsilon \int_{y:|y|<\frac{\delta}{h_n}} |K(y)|dy + \int_{y:|y|\geq\frac{\delta}{h_n}} |g(x-h_n y)||K(y)|dy + |g(x)| \int_{y:|y|\geq\frac{\delta}{h_n}} |K(y)|dy \end{aligned}$$

Par l'absolue continuité de l'intégrale et le fait que $h_n \to 0$, quand $n \to \infty$:

$\exists n_0 \in \mathbb{N}/\forall n \geq n_0 : \int_{y:|y|\geq\frac{\delta}{h_n}} |K(y)|dy < \epsilon.$
De plus, sur $\{\int_{y:|y|\geq\frac{\delta}{h_n}}\}$, on a $|y|\frac{h_n}{\delta} \geq 1$. d'où
$\int_{y:|y|\geq\frac{\delta}{h_n}} |g(x-h_ny)||K(y)|dy \leq \int_{y:|y|\geq\frac{\delta}{h_n}} |y|\frac{h_n}{\delta}|g(x-h_ny)||K(y)|dy$
$\leq \frac{1}{\delta} \sup_{|y|\geq\frac{\delta}{h_n}} |y||K(y)| \int |g(z)|dz \to 0$ par l'hypothèse puisque $h_n \longrightarrow 0$.

1.2.2.2 La consistance de l'estimateur

L'estimateur à noyau de la densité dépend de deux paramètres : la fenêtre h et le noyau K. Ce dernier établit l'aspect du voisinage de x et h contrôle la taille de ce voisinage, donc h est le paramètre prédominant pour avoir de bonnes propriétés asymptotiques, néanmoins le noyau K ne doit pas être négligé, comme le montre le travail de Parzen (1962) [16] cité ci-dessous sur la consistance de cet estimateur.
Cette dernière est obtenue, en se basant sur l'étude asymptotique du biais, de la variance et de la décomposition suivante :

$$\begin{aligned} E[f_n(x) - f(x)]^2 &= E[f_n(x) - E(f_n(x)) + E(f_n(x)) - f(x)]^2 \\ &= Var(f_n(x)) + [Ef_n(x) - f(x)]^2 \\ &= Var[f_n(x)] + [Biais\{f_n(x)\}]^2 \end{aligned}$$

Dans la suite, nous supposons que le noyau K est une densité de probabilité vérifiant les conditions suivantes.

(**K.1**) **:** K est bornée

(**K.2**) **:** $\lim|x|K(x) = 0$, *quand* $|x| \to \infty$,

(**K.3**) **:** $\forall x \in \mathbb{R} :\ K(-x) = K(x)$,

(**K.4**) **:** $\int x^2K(x)dx < \infty$.

1) Étude du biais :

Proposition 1.1. *Sous les hypothèses* $(K.1) - (K.2)$ *et comme* K *est une fonction intégrable, telle que* $\int K(u)du = 1$, *et si* f *est continue* $h_n \to 0$, *quand* $n \to 0$ *et* $nh_n \to \infty$

alors $f_n(x)$ est asymptotiquement sans biais de $f(x)$.

$$\forall x \in \mathbb{R} \quad \lim_{n\to\infty} E[f_n(x)] = f(x)$$

Preuve

$$\begin{aligned} E(f_n(x)) &= E\Big[\frac{1}{nh_n}\sum_{i=1}^{n} K\big(\frac{x-X_i}{h_n}\big)\Big] \\ &= \frac{1}{h_n} E\Big[K\big(\frac{x-X}{h_n}\big)\Big] \\ &= \frac{1}{h_n}\int_{\mathbb{R}} K\big(\frac{x-t}{h_n}\big) f(t)dt \end{aligned}$$

En posant $x - t = z$, on arrive à :

$$E[f_n(x)] = \frac{1}{h_n}\int_{\mathbb{R}} K\big(\frac{z}{h_n}\big) f(x-z)dz$$

Comme K et f vérifient les conditions du théorème de Bochner, et $\lim_{n\to\infty} h_n = 0$,
On a alors

$$\lim_{n\to\infty} \frac{1}{h_n}\int_{\mathbb{R}} K\big(\frac{x-t}{h_n}\big) f(t)dt = f(x)\int_{\mathbb{R}} K(z)dz$$

d'où

$$\lim_{n\to\infty} E[f_n(x)] = f(x)$$

Remarque 1. *Nous constatons que le biais de l'estimateur converge vers zéro quand la fenêtre tend vers zéro, de plus vu son expression, on constate qu'il ne dépend pas du nombre des variables, il dépend surtout du noyau K.*

2) Étude de la variance de $f_n(x)$

Proposition 1.2. *Sous les conditions $(K.1) - (K.2)$ et comme K est une fonction intégrable, telle que $\int K(u)du = 1$, et si f est continue en tout point $x \in \mathbb{R}$ et $nh_n \to \infty$, et $h_n \to 0$ alors :*

$$\lim_{n\to\infty} Var[f_n(x)] = 0$$

En effet

$$\begin{aligned} Var[f_n(x)] &= E[f_n(x)]^2 - [Ef_n(x)]^2 \\ &\leq E[f_n(x)]^2 \\ &\leq \frac{1}{nh_n^2} E\Big[K^2\Big(\frac{x-X_i}{h_n}\Big)\Big] \\ &\leq \frac{1}{nh_n^2}\int K^2\Big(\frac{x-t}{h_n}\Big) f(t)dt \\ &\leq \frac{1}{nh_n^2}\int K^2\Big(\frac{z}{h_n}\Big) f(x-z)dz \end{aligned}$$

Remarquons que k est une fonction bornée et intégrable, telle que $\int K(u)du = 1$ *impliquent que le noyau est de carré intégrable et les hypothèses sur* h_n, K *et* f *assurent que :*

$$\frac{1}{nh_n^2}\int K^2\Big(\frac{z}{h_n}\Big) f(x-z)dz \to \frac{1}{nh_n} f(x)\int K^2(z)dz$$

Alors, par application, encore une fois, du théorème de Bochner, on obtient

$$\lim_{n\to\infty} Var[f_n(x)] = 0, \ \ quand\, nh_n \to \infty$$

Ces deux propositions impliquent la convergence en moyenne quadratique et donc à fortiori, la consistance de l'estimateur.

Thoérème 1.4. *Si* f_X *est continue,* $h_n \to 0$, *quand* $n \to \infty$. $nh_n \to \infty$ *et si la densité* K *vérifie les condition* $K_1 - K_2$, *alors* $f_n(x)$ *est un estimateur convergeant en moyenne quadratique vers* $f(x)$.

Preuve : Résulte des deux propositions précédentes.

1.2.2.3 Consistance forte

Dans le même article, Parzen a établi la normalité asymptotique, ainsi que la convergence uniforme en probabilité.
Son travail est un outil important et a été largement développé par plusieurs chercheurs (Devroye et Györfi(1985), $Silverman$(1986), $Izenman$(1991), $Scott$(1992)).

Pour une étude plus détaillée, voir par exemple (Prakasa-Rao (1983))15.

En (1976) Nadaraya a énoncé le théorème suivant sur la consistance forte de l' estimateur.

Thoérème 1.5. *Si K est à variation bornée ($K = K_1 - K_2$ ou $K_1 \nearrow$ et $K_2 \nearrow$) et si pour tout $\forall\theta > 0, \sum_{i=1}^{\infty} exp(-n\theta h_n^2) < \infty$, alors*

$$\sup|f_n(x) - f(x)| \xrightarrow{P.s} 0, \ quand\, n \longrightarrow \infty,$$

ssi f est uniformément continue.

Preuve : *se referer à Prakasa-Rao(1983) [15].*

Remarque 2. 1) *Cette méthode d'estimation (à noyau)se généralise au cas de $\mathbb{R}^p$. Ainsi, si $X_1, ..., X_n$ sont n vecteurs aléatoires i.i.d. de $\mathbb{R}^p$, de même densité f_X inconnue, alors on peut l'estimer par $f_n(x) = \frac{1}{nh_n^p}\sum_{i=1}^{n} K(\frac{X_i - x}{h_n})$, $x \in \mathbb{R}^p$ le théorème de Bochner reste valable si K et g sont définies sur $\mathbb{R}^p$ et si on pose $g_n(x) = \frac{1}{h_n^p}\int K(\frac{z}{h_n})g(x-z)dz$, et permet de déduire la consistance de $f_n(x)$, sous des conditions adéquates sur f_X, h_n, et K.*

2) *Il existe bien d'autres méthodes d'estimation de la densité, à savoir :*

la méthode de projection et celle des delta suites, une des plus récentes et des plus générales (elle englobe les précédentes) étant la méthode des ondelettes (cf Doukhan and Léon (1990), Kerkyacharian and Picard (1992)et Walter(1992)), mais l'estimation par la méthode des noyaux reste très utilisée.

4) *Il est connu que le choix du noyau n'influe pas beaucoup sur l'estimateur à noyau, mais ce dernier est très sensible au choix de la fenêtre.*

Il existe donc des méthodes qui aident à un choix plus ou moins optimum, parmi lesquelles la méthode du plug-in ou la méthode du pouce.

1.2.3 Estimation de la fonction de régression

Soient $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ des couples de variables aléatoires indépendantes et de même loi que (X, Y).

Dans le modèle de régression non paramétrique on suppose l'existence d'une fonction "r" qui exprime la valeur moyenne de la variable à expliquer Y en fonction de la variable

explicative X, c'est à dire :

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

où ϵ est une variable centrée et indépendante de X.

Le couple de variables aléatoires (X, Y) est à valeurs dans $\mathbb{R}^2$ et admet une densité jointe $f(x, y)$ sur $\mathbb{R}^2$ et une densité marginale $f_X(.) > 0$, donnée par :

$$f_X(x) = \int_{\mathbb{R}} f(x, y)dy$$

La variable Y est supposée intégrable, c'est à dire $E(|Y|) < \infty$.

On peut alors définir l'espérance conditionnelle ou la fonction de régression $r(x)$ de Y en $X = x$ par :

$$r(x) = E(Y \backslash X = x) = \int_{\mathbb{R}} y f_{Y \backslash X=x}(y)dy = \int_{\mathbb{R}} \frac{y f(x, y)}{f_X(x)}dy$$

La fonction $r(x)$ réalise la meilleure approximation de Y sachant $X = x$, au sens des moindres carrés.

En effet, on cherche une fonction f telle que $f(X)$ soit la plus proche possible de Y, dans le sens qu'on veut trouver une fonction f^* qui minimise l'erreur moyenne quadratique $E|f(x) - Y|^2$, i.e f^* doit vérifier

$$E|f^*(x) - Y|^2 = \min_f E|f(x) - Y|^2$$

Et on a

$$\begin{aligned} E|f(x) - Y|^2 &= E|f(x) - r(x) + r(x) - Y|^2 \\ &= E|f(x) - r(x)|^2 + E|r(x) - Y|^2 + 2E|(f(x) - r(x))(r(x) - Y)| \\ &= E|f(x) - r(x)|^2 + E|r(x) - Y|^2 \end{aligned}$$

Car

$$\begin{aligned} E|(f(x)-r(x))(r(x)-Y)| &= E(E(f(x)-r(x)(r(x)-Y)\backslash X) \\ &= E(f(x)-r(x)E((r(x)-Y)\backslash X)) \end{aligned}$$

Et comme

$$E((r(x)-Y)\backslash X) = r(x) - E(Y\backslash X) = 0$$

On obtient

$$E|(f(x)-r(x))(r(x)-Y)| = 0$$

Donc

$$E|f(x)-Y|^2 = E|r(x)-Y|^2 + \int_{\mathbb{R}^d} |f(x)-r(x)|^2 d(\mu(x))$$

La fonction $r(x)$ minimise l'erreur moyenne quadratique, En effet :

$$E|f(x)-Y|^2 = E|r(x)-Y|^2 + \int_{\mathbb{R}^d} |f(x)-r(x)|^2 d(\mu(x))$$

où μ est la loi de X.
Le terme $\int_{\mathbb{R}^d} |f(x)-r(x)|^2 d(\mu(x))$ est appelé l'erreur quadratique moyenne. Pour que $E|f(x)-Y|^2$ soit minimum il faut que l'erreur égale à zéro, c-à-d il faut que $f(x) = r(x)$, ce qui démontre que la fonction de régression minimise l'erreur quadratique moyenne.
Comme r est généralement inconnue, nous nous intéressons à son estimation.

1.2.3.1 L'estimateur à noyau de la fonction de régression

l'estimation de la fonction $r(x)$ que nous considérons est l'estimteur à noyau indroduit par Nadaraya [4]- Watson [7] en 1964.
Une des multiples façons de construire l'estimateur de Nadaraya-Watson que nous abordons est d'utiliser les estimateurs de $f(x,y)$ et de $f_X(x)$.
Soient $J(x,y)$ une densité de probabilité sur $\mathbb{R}^2$

Posons $K(x) = \int_{\mathbb{R}} J(x,y)dy$.

si $h_n \longrightarrow 0$, *quand* $n \to \infty$ alors un estimateur de $f(x,y)$ est donné par

$$f_n(x,y) = \frac{1}{nh_n^2}\sum_{i=1}^{n} J\Big(\frac{x-X_i}{h_n}, \frac{y-Y_i}{h_n}\Big)$$

Et de même, on trouve l'estimateur de la densité marginale donné par la formule suivante

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h_n}\Big)$$

un estimateur naturel de $r(x)$ est alors donné par :

$$\begin{aligned} r_n(x) &= \int_{\mathbb{R}} \frac{y f_n(x,y)}{f_n(x)}dy \\ &= \int_{\mathbb{R}} \frac{y\frac{1}{nh_n^2}\sum_{i=1}^{n} J\Big[\frac{(x-X_i)}{h_n}, \frac{(y-Y_i)}{h_n}\Big]}{f_n(x)}dy. \end{aligned}$$

En posant $z = \frac{y-Y_i}{h_n} \Rightarrow dy = h_n dz$, il vient

$$r_n(x) = \frac{h_n\sum_{i=1}^{n} m(\frac{x-X_i}{h_n})}{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})} + \frac{\sum_{i=1}^{n} Y_i K(\frac{x-X_i}{h_n})}{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})}$$

où

$$m(x) = \int_{\mathbb{R}} yJ(x,y)dy$$

peut être considéré comme, un estimateur de $r(x)$. Pour un choix de

$$J(x,y) = K(x)L(y)$$

avec

$$\int yL(y)dy = 0$$

alors $m(x) = 0$.

D'où l'expression de l'estimateur de la fonction de régression :

$$r_n(x) = \begin{cases} \dfrac{\sum_{i=1}^n Y_i K(\frac{x-X_i}{h_n})}{\sum_{i=1}^n K(\frac{x-X_i}{h_n})} = \frac{\Phi_n(x,h_n)}{f_n(x,h_n)}, & si \quad \sum_{i=1}^n K(\frac{x-X_i}{h_n}) \neq 0 \\ \\ 0 & sinon \end{cases}$$

$$\Phi_n(x, h_n) = \frac{1}{nh_n} \sum_{i=1}^n Y_i K\Big(\frac{x - X_i}{h_n}\Big)$$

Plus généralement, des estimateurs de $r(x)$ peuvent s'écrire

$$r_n(x) = \sum_{i=1}^n W_{n,i}(x) Y_i,$$

où $W_{n,i}(x)$ sont des poids dépendant de x et de $(X_1, \ldots, X_n)$. La valeur de $W_{n,i}$ dépend du type d'estimateur considéré, nous citons ci-dessous des exemples connus.

Estimateur des plus proches voisins

Soit k_n un paramètre de l'estimation, on pose

$$W_{n,i}(x) = \begin{cases} \frac{1}{K_n} & si\ X_i\ est\ parmi\ le\ voisin\ le\ plus\ proche\ de\ k_n\ de\ x\ dans\ \{X_1, \ldots, X_n\}, \\ 0 & sinon \end{cases}$$

Pour plus des détails (voir Devroye et all. 1994) 1.

Estimateur à partitions

On utilise une partition $\pi_n = \{A_{n,j} : j\}$ de $\mathbb{R}^2$ et on pose

$$W_{n,i}(x) = \sum_j \frac{I_{A_{n,j}}(X_i)}{\sum_{k=1}^n I_{A_{n,j}}(X_k)} I_{A_{n,j}}(x)$$

Pour plus des détails (voir Devroye et Györfi, 1983[11]).

• Signalons que d'autres types d'estimateurs existent, parmi les quels nous citons l'estimateur des moindres carrés, ou plus généralement les estimateurs spline lissage (cf, par exemple Kohler and Krzyzak (2001)[14]).

• La construction de l'estimateur à noyau de Nadaraya-Watson dépend de deux paramètres, le paramètre de lissage h dont le choix est crucial pour obtenir de bonnes propriétés asymptotiques (citées ci-dessous) et le noyau K dont on ne peut pas négliger le rôle pour la réduction du biais.

1.2.3.2 Les propriétés de l'estimateur de la fonction de régression :

D'une manière analogue aux propriétés asymptotiques de l'estimateur de Parzen Rosenblatt, nous étudions dans cette partie la convergence en moyenne quadratique.

La consistance

En vu de la décomposition suivante :

$$E(r_n(x) - r(x))^2 = Var(r_n(x)) + (Er_n(x) - r(x))^2.$$

L'étude asymptotique du biais et de la variance de l' estimateur de Nadaraya-Watson détermine les conditions suffisantes à la consistance de cet estimateur.

Thoérème 1.6. *Soit K un noyau vérifiant les hypothèses $K_1 - K_4$, si $E(y^2) < \infty, f_X(x)$ est strictement positive, r, f et $m(x) = \int_{\mathbb{R}} y^2 f(x,y)dy$ sont continues au point x et si $h_n \longrightarrow 0$ et $nh_n \longrightarrow +\infty$ ($quand\ \ n \to \infty$), alors $r_n(x)$ est un estimateur consistant de $r(x)$.*

Preuve :

puisque $f_n(x)$ est un estimateur consistant de $f_X(x)$, il suffit donc de montrer que $\Phi_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i K(\frac{x - X_i}{h_n})$ est un estimateur consistant de

$$\Phi(x) = \int_{\mathbb{R}} y f(x,y)dy$$

1) Étude asymptotique du biais

Nous avons

$$\begin{aligned} E(\Phi_n(x)) &= E\Big[\frac{1}{nh_n}\sum_{i=1}^{n} Y_i K\big(\frac{x-X_i}{h_n}\big)\Big] \\ &= \frac{1}{h_n} E\Big[Y K\big(\frac{x-X}{h_n}\big)\Big] \\ &= \frac{1}{h_n}\int_{\mathbb{R}} E(Y/X=t) K(\frac{x-t}{h_n}) f(t) dt \\ &= \frac{1}{h_n}\int_{\mathbb{R}} K(\frac{x-t}{h_n}) r(t) f(t) dt \\ &= \frac{1}{h_n}\int_{\mathbb{R}} K(\frac{x-t}{h_n}) \Phi(t) dt \to \Phi(x) \end{aligned}$$

Par le théorème de Bochner.

2) Étude de la variance

On a

$$\begin{aligned} Var(\Phi_n(x)) &= Var\big(\frac{1}{nh_n}\sum_{i=1}^{n} Y_i K(\frac{x-X_i}{h_n})\big) \\ &= \frac{1}{n^2 h_n^2}\sum_{i=1}^{n} Var\big(Y_i K(\frac{x-X_i}{h_n})\big) \\ &= \frac{1}{nh_n^2} Var\big(Y K(\frac{x-X}{h_n}\big) \\ &\leq \frac{1}{nh_n^2} E\big(Y^2 K^2(\frac{x-X}{h_n})\big) \\ &= \frac{1}{nh_n^2}\int y^2 K^2(\frac{x-u}{h_n}) f(u,y) du dy \\ &= \frac{1}{nh_n}\big(\frac{1}{h_n}\int m(u) K^2(\frac{x-u}{h_n}) du\big) \to \lim_{n\to\infty}\frac{1}{nh_n} m(u)\int_{\mathbb{R}} K^2(t) dt \to 0 \end{aligned}$$

Ceci implique que

$$\Phi_n(x) \xrightarrow{P} \Phi(x)$$

D'où

$$r_n(x) = \frac{\Phi_n(x)}{f_n(x)} \xrightarrow{P} \frac{\Phi(x)}{f(x)} = r(x).$$

Résultat de Devroye et Krzyzak(1989)

Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ *et* Y une variable aléatoire réelle de carré intégrable ($EY^2 < \infty$). L'objet de l'analyse de régression est d'estimer Y, après l'observation de X.

Sur la base d'un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de la loi de (X, Y), on veut construire un estimateur $r_n(x)$ de r telle que $\int_{\mathbb{R}} |r_n(x) - r(x)|^2 d(u(x))$ soit petite. Pour cela nous avons besoin que le noyau soit régulier, notion rappelée ci dessous.

Définition 1.1. *On dit que le noyau positif K est régulier s'il existe une boule S_r, de rayon r et de centre l'origine, et une constante positive b, telles que $K(x) \geq b1_{S_r}(x)$ et $\int \sup_{y \in x+S_r} K(y)dx < \infty$.*

Énoncé du résultat

Thoérème 1.7. *Soit K est un noyau régulier et soit $r_n(x)$ l'estimation à noyau de la fonction de régression $r(x)$. Alors les propositions suivantes sont équivalentes*

(A) Pour chaque loi de (X, Y) avec $|Y| \leq M < \infty$. et pour toute $\epsilon > 0$, il existe deux constantes c et n_0 tels que pour tout $n \geq n_0$.

$$P\Big[\int_{\mathbb{R}^d} [r_n(x) - r(x)]\mu(d(x)) > \epsilon\Big] \leq e^{-cn}$$

. (B) Pour toute loi de (X, Y) avec $|Y| \leq M < \infty$,

$$\int_{\mathbb{R}^d} [r_n(x) - r(x)]\mu(d(x)) \to 0 \quad (n \to \infty) \quad p.s.$$

(C) Pour toute loi de (X, Y) avec $|Y| \leq M < \infty$.

$$\int_{\mathbb{R}^d} [r_n(x) - r(x)]\mu(d(x)) \to 0 \quad (n \to \infty) \quad \textit{en probabilité.}$$

(D) $\lim_{n\to\infty} h_n = 0, \qquad \lim_{n\to\infty} nh_n^d = \infty.$

Pour plus des détails, voir l'article de (Luc Devroye, Laszlo Gyorfi, Adam Krzyzak et Gabor Lugosi,(1994) [9]).

CHAPITRE 2

INTRODUCTION À L'ANALYSE DES DURÉES DE SURVIE

L'inférence statistique se base sur l'observation d'un échantillon constitué par des vraies observations de la variable d'intérêt (données complètes). Mais dans la pratique, il n'est pas toujours possible de recueillir un tel échantillon. Certains observations ne sont pas nécessairement des vraies réalisations des variables d'intérêt et c'est là où on introduit l'analyse de survie.

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenue d'un évènement précis. L'événement étudié (appelé "décès") est le passage irréversible entre deux états (nommé "vivant" et "décès"). L'événement terminal n'est pas forcement la mort : il peut s'agir de l'apparition d'une maladie (par exemple, le temps avant une rechute), d'une guérison (temps entre la diagnostic et la guérison), la panne d'une machine (durée de fonctionnement d'une machine).On dit qu'une variable de censure C a empêché l'observation de la vraie valeur d'intérêt et ne nous fournit alors qu'une information partielle sur elle.

L'analyse des données (durées) de survie est l'étude du délai de la survenue de cet évènement.

2.1 Définitions

Quelques définitions sont couramment utilisées dans les études de survie.

• **Événement d'intérêt :** événement auquel on s'intéresse au cours de l'étude → Décès décès lié à un complication, rechute, disparition de symptômes. On utilisera L'« analyse de survie » dès qu'il y aura une notion de durée jusqu'à la survenue de l'évènement d'intérêt (qu'on nommera « décès »).

• **Date d'origine :** elle correspond à l'origine de la durée étudiée. Elle peut être la date de naissance, le début d'une exposition à un facteur de risque, la date d'une opération chirurgicale, la date de début d'une maladie ou la date d'entrée dans l'étude. Chaque individu peut donc avoir une date d'origine différente (pas important car c'est la durée qui nous intéresse).

• **Date des dernières nouvelles :** c'est la date la plus récente où des informations sur un sujet ont été recueillies.

• **Date de point :** c'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

• **Temps de recul :** Délai entre la date d'origine et la date de point.

• **Durée de survie :** Délai entre la date d'origine et la date de survenue ou la date des dernières nouvelles.

2.2 Distributions de la durée de survie

En analyse de survie on s'intéresse à une variable positive (car on s'intéresse au temps que met un événement pour se produire).

Et avant d'introduire la notion de mesure de hasard, nous allons d'abord étudier deux cas particulièrement intéressants dans la pratique : **le cas absolument continue**, **le cas discret** :

2.2.1 Cas absolument continue

Soit T une variable aléatoire (v.a) positive ou nulle, et absolument continue représente la durée de survie, alors sa loi de probabilité peut être définie par l'une des fonctions équivalentes (au sens que chacune peut être obtenue à partir de l'une des autres fonctions) suivantes :

1. **Fonction de répartition** F

   La fonction de répartition $F(t)$, pour t fixé, mesure la probabilité de mourir avant l'instant t, c'est-à-dire :

   $$F(t) = P(T \leq t) = \int_0^t f(v)dv$$

   où f est la densité de la v.a T.

   La fonction de répartition est croissante.

2. **Fonction de survie** S

   La fonction de survie est, pour t fixé, la probabilité de survivre au delà de t c'est-à-dire :

   $$\begin{aligned} S(t) &= P(un \text{ individu survive au-delà du temps } t) \\ &= P(T > t). \end{aligned}$$

   A partir de la définition de la fonction de répartition $F(t)$ de T

   $$\begin{aligned} S(t) &= 1 - P(\text{ un individu décédé entre 0 et } t) \\ &= 1 - F(t) \end{aligned}$$

   Ici $S(t)$ est une fonction décroissante, continue à droite avec :

   $$\begin{aligned} \lim_{t\to 0} S(t) &= 1 \\ \lim_{t\to \infty} S(t) &= 0 \end{aligned}$$

3. **Densité de probabilité f**

C'est la fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$.

$$F(t) = \int_0^t f(v)dv.$$

Si la fonction de répartition F admet une dérivée au point t alors

$$f(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h)}{h} = F^{'}(t) = -S^{'}(t).$$

Pour t fixé, la densité de probabilité représente la probabilité de mourir dans un petit intervalle de temps après l'instant t.

4. **Risque instantané λ (ou taux de hasard)**

Le risque instantané, pour t fixé caractérise la probabilité de mourir dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h \backslash T > t)}{h} = \frac{f(t)}{S(t)} = -(\ln S(t))^{'}.$$

La fonction "taux de hasard" est une des fonctions les plus intéressantes car elle donne une version probabiliste du futur d'un sujet n'ayant pas encore connu l'évènement d'intérêt et elle est appelée aussi la fonction de mortalité.

Preuve :

On a

$$\lim_{h \to 0} \frac{P(t \leq T < t + h \backslash T > t)}{h} = \lim_{h \to 0} \frac{1}{h} \frac{P(t \leq T < t + h)}{S(t)} = \lim_{h \to 0} \frac{1}{h} \frac{F(t+h) - F(t)}{S(t)} = \frac{F^{'}(t)}{S(t)} = \lambda(t) \blacksquare.$$

Si f est continue, alors λ est aussi continue, et il est possible de définir la fonction de hasard cumulé.

5. **Fonction de hasard cumulé Λ**

La fonction de hasard cumulé est l'intégrale du risque instantané λ :

$$\Lambda(t) = \int_0^t \lambda(u)du = -\ln(S(t)).$$

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé (ou du risque instantané) :

$$S(t) = exp(-\Lambda(t)) = exp\Big(-\int_0^t \lambda(u)du\Big)$$

On en déduit que

$$f(t) = \lambda(t)exp\Big(-\int_0^t \lambda(u)du\Big)$$

Proposition 2.1. *: Si f est continue sur $\mathbb{R}_+^*$ et si $A = \{t > 0/S(t) \neq 0\}$ alors les propositions suivantes sont équivalentes.*

a) $\forall t \in A, \lambda(t) = \dfrac{f(t)}{S(t)}$

b) $\forall t \in A, \lambda(t) = (-\ln S(t))'$

c) $\forall t \in A, S(t) = exp\{-\Lambda(t)\}$

d) $\forall t \in A, f(t) = \lambda(t)exp\{-\Lambda(t)\}$

Preuve :

a) $\Leftrightarrow$ *b)* car

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \\ &= \frac{-S(t)'}{S(t)} \\ &= (-\ln S(t))' \end{aligned}$$

$b) \Leftrightarrow c)$ car

$$\begin{aligned}\lambda(t) &= (-\ln S(t))' \Rightarrow \int_0^t \lambda(s)ds \\ &= -\ln S(s)]_0^t \\ &= -\ln S(t)\,.\end{aligned}$$

$c) \Rightarrow d)$car

$$\begin{aligned}S(t) &= exp\{-\Lambda(t)\} \Rightarrow S'(t) \\ &= -f(t) \\ &= -\lambda(t)exp\{-\Lambda(t)\}\end{aligned}$$

$d) \Rightarrow c)$ car

$$\begin{aligned}f(s) &= \lambda(s)exp\{-\Lambda(s)\} \Rightarrow 1 - S(t) \\ &= F(t) = \int_0^t f(s)ds \\ &= \int_0^t \lambda(s)exp\{-\Lambda(s)\} = -exp\{-\Lambda(s)]_0^t \\ &= -exp\{-\Lambda(t) + 1\}.\end{aligned}$$

Exemple :

1) loi exponentielle :

Soit T une v.a de loi exponentielle de paramètre α, alors sa densité est

$$f(t) = \alpha exp\{-\alpha t\}1_{]0,\infty[}(t)$$

- Calculons sa fonction de survie, et son taux de hasard.

On a : $S(t) = \int_t^\infty \alpha exp\{-\alpha s\}1_{]0,\infty[}(s)ds = exp\{-\alpha t\}$ et donc $\lambda(t) = \frac{f(t)}{S(t)} = \alpha$.

Nous remarquons que ce taux de hasard est constant (la loi exponentielle est alors dite sans mémoire), en fait nous pouvons montrer que la loi exponentielle est la seule loi, à densité, qui soit sans mémoire.

2)Loi de Weibull :

Soit T une v.a de loi de Weibull de paramètre $(\alpha, \beta), \alpha(>0)$ est le paramètre d'échelle et $\beta(>0)$ est le paramètre de formeSa densité est $f(t) = \frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}exp\{-(\frac{t}{\alpha})^{\beta}\}1_{]0,+\infty[}(t)$.

- Calculons sa fonction de survie, et son taux de hasard.

On a : $S(t) = \int_t^{\infty} \frac{\beta}{\alpha}(\frac{s}{\alpha})^{\beta-1}exp\{-(\frac{s}{\alpha})^{\beta}\}1_{]0,+\infty[}(s)ds = -exp\{-u^{\beta}\}]_{\frac{t}{\alpha}}^{\infty} = exp\{-(\frac{t}{\alpha})^{\beta}\}$ et $\lambda(t) = \dfrac{f(t)}{S(t)} = \frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}$.

- ✓ Nous en déduisons que λ croit, si $\beta > 1$.
- ✓ Décroit si $\beta < 1$
- ✓ Reste constante si $\beta = 1$ (on retrouve la loi exponentielle).

2.2.1.1 Quelques quantités associes à loi de la survie

Moyenne et variance de la durée de survie

Le temps moyen de survie $E(T)$ et la variance de la durée de survie $V(T)$ sont définis par les quantités suivantes :

$$E(T) = \int_0^{\infty} S(t)dt \quad , \; V(T) = 2\int_0^{\infty} tS(t)dt - (E(T))^2.$$

Ainsi on peut déduire l'espérance et la variance à partir de n'importe laquelle des fonctions $F, S, f, \lambda, \Lambda$ (mais pas l'inverse).

Quantiles de la durée de survie

- La fonction quantile de la durée de survie est définie par :

$$\begin{aligned} q(p) &= \inf(t : F(t) \geq p), \qquad 0 < p < 1. \\ &= \inf(t : S(t) \leq 1 - p). \end{aligned}$$

Lorsque la fonction de répartition F est strictement croissante et continue alors

$$\begin{aligned} q(p) &= F^{-1}(p), \qquad 0 < p < 1. \\ &= S^{-1}(1 - p). \end{aligned}$$

Le quantile $q(p)$ est le temps où une proportion p de la population a disparu.

2.2.2 Cas discret

Soit T une variable aléatoire discrète prenant les valeurs ordonnées, par ordre croissant, $x_1, \ldots, x_n \ldots$

⋆ Sa fonction de survie est donnée par :

$$S(t) = P(T > t) = \sum_{j:x_j>t} P(T = x_j).$$

⋆ Son taux de hasard est :

$$\lambda(x_j) = P(T = x_j \backslash T \geq x_j) = \frac{P(T = x_j)}{S(x_{j-1})} = 1 - \frac{S(x_j)}{S(x_{j-1})}.$$

Mais puisque $S(x) = \prod_{j:x_j \leq x} \frac{S(x_j)}{S(x_{j-1})}$

Il vient :

$$S(x) = \prod_{j:x_j \leq x} (1 - \lambda(x_j))$$

Exemple :

Soit T une v.a. prenant les valeurs $0, 5$ et 8 avec probabilité $\frac{1}{3}$ pour chacune.

Alors : $S(x) = \begin{cases} 1 & si \quad x < 0 \\ 2/3 & si \quad 0 \leq x < 5 \\ 1/3 & si \quad 5 \leq x < 8 \\ 0 & si \quad x \geq 8 \end{cases}$ et $\lambda(x) = \begin{cases} 1/3 & si \quad x = 0 \\ 1/2 & si \quad x = 5 \\ 1 & si \quad x = 8 \\ 0 & ailleurs \end{cases}$

2.2.3 Cas général

Mesure de hasard

Si T est une v.a.r. de fonction de répartition quelconque F, la notion de taux de hasard se généralise en la notion de mesure de hasard $d\Lambda$ définie par :

$$d\Lambda = \frac{dF(s)}{1 - F(s-)}$$

où $F(s-)$ est la limite de F à gauche de s et dF est la loi de probabilité de T.
Nous pouvons remarquer que :

⋆ Si $dF(s) = f(s)ds$ alors $d\Lambda(s) = \lambda(s)ds = \frac{f(s)ds}{S(s)}$.

⋆ Si dF est une loi discrète nous retrouvons $\lambda(s) = \dfrac{P(T=s)}{S(s-)}$.

2.3 Données incomplètes (censurées)

Comme nous avons mentionné plus tôt, en analyse de survie, les données recueillies sont la plupart du temps incomplètes : **la censure** est un mécanisme inhérent des données de survie. Dans la suite la variable durée de survie T est définie comme le délai écoulé entre la date d'origine T_0 et la date de survenue de l'événement.

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Pour l'individu i, considérons
- Son temps de survie X_i,
- Son temps de censure C_i,
- La durée réellement observée T_i.

Définition 2.1. *: La durée T est dite censurée si :*

• *le patient est toujours vivant à la fin de l'étude (exclus-vivants).*

• *le patient est perdu de vue : il a quitté l'étude avant qu'on ait pu observer l'événement d'intérêt.*

Censure à droite :

Il y a censure à droite lorsque nous observons la censure C (et non pas la durée de vie d'intérêt T) et que nous savons que $T > C$.

les observations sont des répliques du couples (X, δ), où

$$X = \inf(T, C)$$

et

$$\delta = \begin{cases} 1 & si \quad T \leq C \\ 0 & si \quad T > C \end{cases}$$

où δ vaut 1 quand l'observation est complète (elle correspond à une donnée de la variable d'intérêt) et vaut 0 si la donnée est censurée.

Ce modèle adapté au cas où l'événement d'intérêt est le temps de survie à une maladie et où la date de fin de l'étude est préalablement fixé ; les patients vivants à la fin de l'étude fournissent des données censurées à droite.

Censure à gauche :

Une durée de survie est dite censurée à gauche si l'individu a déjà connu l'évènement d'intérêt avant l'entrée dans l'étude. Formellement, la durée de survie pour un individu est définie par le couple (X, δ), où

$$X = \max(T, C)$$

et

$$\delta = \begin{cases} 1 & si \quad T > C \\ 0 & si \quad T \leq C \end{cases}$$

Avec la durée de vie et le temps de censure C supposés indépendants.

Si $\delta = 1$, le sujet subit l'évènement et est observé. Si $\delta = 0$, le sujet est dit censuré à gauche : au lieu d'observer T, on observe une valeur C avec pour seule information le fait que T soit inférieur à C.

Ce modèle est par exemple adapté au cas où l'on s'intéresse à l'age auquel un individu commence à accomplir une tache.

Tout ce qu'on sait chez l'individu censuré est que le véritable age est inférieure à la valeur observée (l'age au moment de l'étude par exemple).

- On ne peut pas toujours savoir le moment exact du déclenchement de la maladie pour certaines personnes, on sait seulement que leur âge est inférieur à leur âge au moment de l'étude. Ces données sont censurées à gauche.

Censure par intervalles :

Il y a censure par intervalle lorsque la censure à droite et la censure à gauche sont conjuguées.

Dans ce cas l'information apportée par l'expérience se traduit par l'appartenance de la durée de vie à un intervalle de temps ($C_1 < T < C_2$).

Ce modèle est adapté au cas de suivis périodique de patients et généralise aussi bien le modèle de censure à droite que celui de censure à gauche.

Exemple :

T durée entre l'infection par le VIH et le décès par le sida.

Censure mixte :

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'intérêts ans qu'on puisse nécessairement déterminer un intervalle auquel il appartient. Dans le modèle il décrit dans l'article de Patilea et Rolin (2006), au lieu d'observer un échantillon de la variable d'intérêt T, on observe un échantillon du couple (Z, A) avec $Z = \min(\max(T, L), R)$ et

$$A = \begin{cases} 0 & si \quad L < T < R \\ 1 & si \quad R \leq \max(T, L) \\ 2 & si \quad T \leq L \leq R \end{cases}$$

ou L et R sont des variables de censure et A est l'indicateur de censure.

Censure de type I :

L'expérimentateur fixe une date (non aléatoire) de fin d'expérience. La durée de participation maximale est alors fixée (non aléatoire) et vaut, pour chaque observation, la différence entre la date de fin d'expérience, et la date d'entrée du patient dans l'étude. Le nombre d'événements observés est, quant à lui, aléatoire. Ce modèle est souvent utilisé dans les études épidémiologiques.

Censure de type II :

L'expérimentateur fixe a priori le nombre d'événements à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'événements étant, quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité.

Censure aléatoire :

C'est typiquement ce modèle qui est utilisé pour les essais thérapeutique. Dans ce type d'expérience, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient). Ici, le nombre d'événements observés et la durée totale de l'expérience sont aléatoires.

2.4 L'estimateur de Kaplan-Meier de la fonction de survie

Nous nous intéressons à l'estimation de la fonction de survie S de la v.a.r T censurée à droite. C'est à dire que nous observons n variables indépendantes
$(X_i = T_i \wedge C_i, \delta_i = 1_{T_i \leq C_i})$ de même loi que le couple $(X = T \wedge C, \delta = 1_{T \leq C})$, où C est une variable de censure et δ est l'indicateur de censure.
Cet estimateur est aussi appelé P-L (Produit-Limite) car il s'obtient comme la limite d'un produit.
Il est fondé sur la remarque suivante : Si $t' < t$, la probabilité de survivre au-delà de l'instant t est égale au produit suivant :

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(T > t \backslash T > t')S(t') \end{aligned}$$

Si l'on renouvelle l'opération en choisissant une date t'' antérieure à t' on aura de même

$$\begin{aligned} S(t') &= P(T > t') \\ &= P(T > t' \backslash T > t'')S(t'') \end{aligned}$$

et ainsi de suite. Donc, la méthode d'estimation se base sur la relation suivante.

$$S(t) = \prod_{j=1}^{n} P(T > t_j \backslash T > t_{j-1})S(t_n)$$

où $t_0 < t_1 < t_2 \cdots < t_n$.

Soit donc $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ la statistique d'ordre associée $X_1, X_2, \ldots, X_n$

Nous avons :

$$P(T > X_{(i+1)}) = \prod_{j=1}^{i} P(T > t_j \backslash T > t_{j-1}) = \prod_{j=1}^{i} p_j = \prod_{j=1}^{i} (1 - q_j)$$

où

⋆ p_j est la probabilité qu'un individu survive pendant l'intervalle de temps $I_j =]X_{(j)}, X_{(j+1)}]$ sachant qu'était vivant au début de cet intervalle.

⋆ q_j est la probabilité de mourir durant l'intervalle I_j sachant que l'individu était vivant au début de cet intervalle. Alors l'estimateur naturel de q_j est donnée par

$$q_j = \frac{M(X_{(j)})}{R(X_{(j)})} = \frac{nombre\ de\ morts\ \text{à}\ l'instant X_j}{nombre\ de\ sujets\ \text{à}\ risque\ juste\ avant\ l'instant\ X_j}$$

où

$$M(X_{(j)}) = \sum_{i=1}^{n} \delta_i 1_{X_i = X(j)} \quad et \quad R(X_{(j)}) = \sum_{i=1}^{n} 1_{\{X_i \geq X_j\}}$$

l'estimateur de Kaplan-Meier est donné par :

$$S_n(t) = \prod_{\{X(j) \leq t\}} \left(1 - \frac{M(X_{(j)})}{R(X_{(j)})}\right)$$

Remarque 3. *:*

Supposons qu'il n'y ait pas d'ex-æquo. Soit $\delta_{(i)}$ l'indicateur de censure associée à $X_{(j)}$ Si $\delta_i = 0$, c'est qu'il y a eu une censure en $X_{(j)}$ et donc $M(X_{(j)}) = 0$. si $\delta_(j) = 1$, c'est qu'il y a eu un mort en $X_{(j)}$ et donc $M(X_{(j)}) = 1$. dans ce dernier cas $R(X_{(j)}) = n-j+1$, on obtient alors :

$$S_n(t) = \prod_{\{x(j)\leq t\}} (1 - \frac{1}{n-j+1})^{\delta(j)}.$$

Exemple 1 :

On ne considère pour l'instant que des données complètes, c-à-d, non censurées relatives à des temps de réalisation d'un évènement mesurés en jours et observé sur 10 individus :

$$6, 19, 32, 42, 42, 43, 94, 105, 105, 120$$

t_i	$M(X_j)$	$R(X_j)$	$1 - \frac{M(X_j)}{R(X_j)}$	$S_n(t_i)$	$F_n(t_i) = 1 - S_n(t_i)$
0	0	10	1	1	0
6	1	10	1-1/10=0.90	0.90	0.10
19	1	9	1-1/9=0.889	0.80	0.20
32	1	8	1-1/8=0.875	0.70	0.30
42	2	7	1-2/7=0.7143	0.50	0.50
43	1	5	1-1/5=0.80	0.40	0.60
94	1	4	1-1/4=0.75	0.30	0.70
105	2	3	1-2/3=0.330	0.10	0.90
120	1	1	1-1=0	0	1

Exemple 2 :

On introduit des données censurées. Dans ce cas la fonction de survie n'est estimée que pour les temps observés mais il faut naturellement ajuster le nombre d'individus à risque. La règle est que pour une durée donnée t_i on ne comptabilise dans les individus risqués que ceux qui ont une date d'évènement égale ou supérieure à t_i ou une durée de censure supérieure à t_i (au passage on notera qu'une convention est que si, pour un individu quelconque, les survenues de l'évènement et de la censure sont concomitantes alors on le considère comme non censuré. En d'autres termes, on impose que la réalisation de

l'évènement précède la censure).

Dans la liste ci-dessous relatives à 19 durées mesurées en jours, les données censurées sont signalées par l'exposant * :

$6, 19, 32, 42, 42, 43^*, 94, 126^*, 169^*, 207, 211^*, 227^*, 253, 255^*, 270^*, 310^*, 316^*, 335^*,$
346^*.

On obtient alors :

t_i	$M(X_j)$	$R(X_j)$	$1 - \frac{M(X_j)}{R(X_j)}$	$S_n(t_i)$	$F_n(t_i) = 1 - S_n(t_i)$
0	0	19	1	1	0
6	1	19	0.947	0.947	0.053
19	1	18	0.944	0.895	0.105
32	1	17	0.941	0.842	0.158
42	2	16	0.875	0.737	0.263
94	1	13	0.923	0.680	0.320
207	1	10	0.90	0.612	0.388
253	1	7	0.957	0.525	0.475

Remarque 4. *: Il existe des durées supérieures à* 253 *jours mais elles sont toutes censurées.*

En conséquence dans ce deuxième exemple, et contrairement au précédent, la valeur estimée de la fonction de survie correspondant au temps d'évènement maximal observé (soit 253 *jours) ne s'annule pas. Précédemment nous avions* $S_n(120) = 0$ *du fait que la durée maximale était non censurée. En d'autres termes, le fait que l'estimateur de la fonction de survie s'annule ne signifie pas que tous les individus ont connu l'évènement étudié mais seulement que la durée maximale ne correspond pas à une censure. Et à titre d'exemple,on reprend les chiffres de ce second exemple en remplaçant* 346^* *par* 346.

CHAPITRE 3

ESTIMATION DE LA FONCTION DE RÉGRESSION POUR DES DONNÉES CENSURÉES

Introduction

Les estimateurs non paramétriques (à noyau) de la régression ont été introduits simultanément par Nadaraya (1964)[4] et Watson (1964)[7].

Ce problème a suscité un grand intérêt et a conduit à la proposition de plusieurs estimateurs sur la base de l'observation d'un échantillon du couple (X, Y). En analyse de survie, il est connu que l'observation de Y n'est pas toujours possible. Y peut être le temps de survie à une maladie. Certains sujets, sous étude, peuvent disparaître de l'étude fortuitement (accident) ou d'une manière planifiée (fin de l'étude). Y est alors censuré à droite. on ne connaît pas sa valeur exacte, on sait seulement qu'elle dépasse l'observation recueillie. Dans ce contexte de censure à droite Beran (1981)[17] a proposé une classe d'estimateurs de la fonction de survie conditionnelle permettant d'en déduire des estimateurs non paramétriques de la fonction de régression difficilement calculables. Ensuite

Carbonez (1992)[1] et Carbonez, G et Van Der Meulen (1995)[2] ont introduit un estimateur à partitions consistant pour Y bornée et censurée à droite. Améliorant ce travail, Kohler, Màthé et Pintér (2002)[12] ont simplifié la preuve du résultat précédent et ont même étendu le travail en proposant d'autres estimateurs non paramétriques (à noyau, plus proches voisins, moindres carrés et spline de lissage). Cependant d'autres types de censure existent.

Pour certaines valeurs de Y, on peut seulement savoir qu'elles sont inférieures aux observations. C'est la censure à gauche. Cela peut être le cas lorsque s'intéresse à l'âge auquel des personnes ont commencé à accomplir une tâche. Certains individus peuvent seulement se rappeler qu'ils ont commencé cette tâche avant un certain âge sans savoir exactement lequel. Nous détaillons dans ce chapitre, l'étude de l'éstimateur à noyau de Kohler et autres (2002), introduit dans le cadre de la censure à droite. IL se caractérise par sa forme très simple à calculer et il a suscité de nombreux travaux parmi lesquels Ould-said et Cai (2005)[6] et Kohler et al (2006) . Puis nous étendons le travail précédent à un modèle de censure à gauche.

3.1 Censure à droite

Dans plusieurs études, il n'est pas possible d'observer un échantillon de (X, Y). Ainsi si la variable Y est le temps de survie d'un patient, à une maladie, ce patient peut décéder d'une autre cause pendant l'étude ou être toujours vivant à la fin de celle-ci.
Dans ce cas Y n'est pas observé mais l'observation est le minimum entre Y et une variable de censure C.
Plus précisément soit Y une variable d'intérêt positive et bornée et C une variable aléatoire de censure positive.
Nous observons l'échantillon $(X_1, Z_1, \delta_1), ..., (X_n, Z_n, \delta_n)$ où $Z_i = Y_i \wedge C_i$ et
$\delta_i = I_{\{Y_i \leq C_i\}}$ (δ_i est l'indicateur de censure).
Nous nous proposons d'estimer $r(x) = E(Y \backslash X = x)$, l'estimateur donné au premier chapitre ne peut plus être utilisés puisque Y n'est pas toujours observé.

3.1.1 Principe de l'estimation

Pour toute variable aléatoire V, on note

$$T_V = \sup\{t : F(t) < 1\},$$

$$I_V = \inf\{t : F(t) \neq 0\}.$$

où F est la fonction de répartition de V.

• L'idée, introduite par (Carbonez et all (1995)[2]), et reprise par (Kohler, Màthé et Pintér (2002)[12]) comme suit :

Soient $S(t) = P(Y > t)$ *et* $H(t) = P(C > t)$ les fonctions de survie respectives de Y et C. On suppose que

$$(H_1) : \begin{cases} (H_{1.1}) & C \text{ et } (X, Y) \text{ sont indépendants et } H \text{ est continue} \\ \\ (H_{1.2}) & T_Y < \infty \text{ et } H(T_Y) > 0. \end{cases}$$

Remarquons que la condition $H(T_Y) > 0$ implique $T_Y < T_C$.

La fonction de régression $r(x)$ doit être estimée à partir des données censurées, qui est un i.i.d. séquence de variables aléatoires : $\{(X_1, Z_1, \delta_1), ..., (X_n, Z_n, \delta_n)\}$.

Et on remplace Y par une estimation de sa moyenne, car la fonction de régression est une espérance conditionnelle, son estimation devrait correspondre à une estimation de l'espérance. Considérons d'abord le cas où il n'y a pas de covariables.

Si les Y_i sont i.i.d, alors l'estimateur de l'espérance de Y est la moyenne arithmétique

$$\frac{1}{n}\sum_{i=1}^{n} Y_i$$

Soit h une fonction de $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. On se propose d'estimer la moyenne $E\{h(X, Y)\}$ sur la base de l'échantillon des données censurées à droite.

• Un "estimateur" sans biais de $E\{h(X, Y)\}$ est donné par :

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i h(X_i, Z_i)}{H(Z_i)}.$$

- En utilisant l'indépendance entre (X, Y) et C avec les propriétés de l'espérance conditionnelle, il vient

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i h(X_i,Z_i)}{H(Z_i)}\right\} &= E\left\{\frac{I_{\{Y_1\leq C_1\}}h(X_1,Y_1)}{H(Y_1)}\right\} \\ &= E\left(\frac{h(X_1,Y_1)}{H(Y_1)}E(I_{\{Y_1\leq C_1\}}\backslash(X,Y))\right) \\ &= E(h(X_1,Y_1)). \end{aligned}$$

Le problème est que H est inconnu. On l'estime par l'estimateur de Kaplan Meier (1958)[5], donné par :

$$H_n(t) = \begin{cases} \prod_{i=1}^{n}\left[1-\frac{1-\delta_i}{n-i+1}\right]^{I_{[Z_{(i)}\leq t]}}, & si \quad t < T_{K,n}, \\ \\ \lim\limits_{s\to T_{K,n}, s<T_{K,n}} H_n(s), & si \quad t \geq T_{K,n}. \end{cases}$$

où $T_{K,n} = max\{Z_1, \ldots, Z_n\}$ et les paires $(Z_{(i)}, \delta_{(i)}), i = 1, \ldots, n$ sont les n paires observées (Z_i, δ_i) ordonnées en $Z_{(i)}, i.e. Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)} = T_{K,n}$.
Remarquons que H_n a été légèrement modifié afin de ne jamais s'annuler.
Cela suggère d'estimer $r(x)$ par :

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{n,i}(x)\frac{\delta_i Z_i}{H_n(Z_i)} \tag{3.1}$$

avec la fonction poids $W_{n,i}$ définie comme suit,

$$W_{n,i}(x) = \frac{K(\frac{x-X_i}{h_n})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h_n})}.$$

Nous utilisons les notations suivantes. Pour tout $t \,/\, 0 \leq t \leq \infty$ et $x \in \mathbb{R}$, on définit :

$$T_{[0,t]}(x) = \begin{cases} t, & si \quad x > t, \\ x, & si \quad 0 \leq x \leq t, \\ 0, & si \quad x < 0. \end{cases}$$

Pour $f : \mathbb{R}^d \to \mathbb{R}$ définissons $T_{[0,t]}f : \mathbb{R}^d \to \mathbb{R}$ par $(T_{[0,t]}f)(x) = T_{[0,t]}(f(x))$.
Du fait que $0 \leq Y \leq T_Y < \infty$ p.s, on a $0 \leq r(x) \leq T_Y$, on estime donc $r(x)$ plutôt par :

$$r_n(x) = \begin{cases} T_{K,n}, & si \quad \hat{r}_n(x) > T_{K,n}, \\ \hat{r}_n(x), & si \quad 0 \leq \hat{r}_n(x) \leq T_{K,n}, \\ 0, & si \quad \hat{r}_n(x) < 0. \end{cases} \tag{3.2}$$

Par analogie avec (3.2), posons

$$r_n^*(x) = \begin{cases} T_Y, & si \quad \hat{r}_n(x) > T_Y, \\ \hat{r}_n(x), & si \quad 0 \leq \hat{r}_n(x) \leq T_Y, \\ 0, & si \quad \hat{r}_n(x) \leq 0. \end{cases}$$

3.1.2 Propriétés de l'estimateur

Le résultat suivant est montré dans Gill et Johansen [18].

Lemme 3.1. *: On a*

$$\sup_{t \leq T_Y} |H_n(t) - H(t)| \to 0 \;\; n \to \infty \;\; p.s$$

Remarquons que Kohler et Mathé ont utilisé le résultat de (Stute et Wang (1993[20])) qui exige la continuité de H pour avoir le résultat précédent.

Lemme 3.2. *: Sous l'hypothèse* $(H_{1.2})$*, On a :*

$$\int_{\mathbb{R}^d}|r_n(x)-r(x)|^2\mu(dx)\to 0\ \ (n\to\infty)\ p.s$$

si et seulement si

$$\int_{\mathbb{R}^d}|r_n^*(x)-r(x)|^2\mu(dx)\to 0\ \ (n\to\infty)\ \text{p.s}$$

Preuve :

$$\begin{aligned}|r_n^*(x)-r(x)|^2 &= |r_n^*(x)-r_n(x)+r_n(x)-r(x)|^2\\ &\leq 2|r_n^*(x)-r_n(x)|^2+2|r_n(x)-r(x)|^2\end{aligned}$$

$$\int_{\mathbb{R}^d}|r_n^*(x)-r(x)|^2\mu(dx)\leq 2\int_{\mathbb{R}^d}|r_n^*(x)-r_n(x)|^2\mu(dx)+2\int_{\mathbb{R}^d}|r_n(x)-r(x)|^2\mu(dx).$$

On a :

$$\int_{\mathbb{R}^d}|r_n(x)-r(x)|^2\mu(dx)\to 0\ \ (n\to\infty)\ \text{p.s}$$

Donc il suffit démontrer que :

$$\int_{\mathbb{R}^d}|r_n^*(x)-r_n(x)|^2\mu(dx)\to 0\ \ (n\to\infty)\ \text{p.s}$$

On a : $T_{K,n}\leq T_Y$ p.s.

$$\begin{aligned}|r_n^*(x)-r_n(x)| &= |T_{[0,T_Y]}\hat{r}_n(x)-T_{[0,T_{K,n}]}\hat{r}_n(x)|\\ &\leq T_Y-T_{K,n}.\\ \int_{\mathbb{R}^d}|r_n^*(x)-r(x)|^2\mu(dx) &\leq \int_{\mathbb{R}^d}(T_Y-T_{K,n})^2\mu(dx)\\ &= (T_Y-T_{K,n})^2,\end{aligned}$$

Ou $T_{K,n}\to T_Y(n\to\infty)$ p.s ■

Thoérème 3.1. *: (Kohler, Màthé et Pintér (2002) [12])*

Sous l'hypothèse (H_1) *et si* K *est un noyau régulier,* $\lim_{n\to\infty} h_n = 0, \lim_{n\to\infty} nh_n^d = \infty$ *alors l'estimateur* $r_n(x)$ *défini par (*3.1*) et (3.2) vérifie :*

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \ p.s$$

Preuve :

D'après le lemme précédent, il suffit de montrer que :

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \text{ p.s}$$

Posons :

$$\overline{r}_n(x) = T_{[0,T_Y]} = \left(\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_i Z_i}{H(Z_i)} \right)$$

$$\begin{aligned} |r_n^*(x) - r(x)|^2 &= |r_n^*(x) - \overline{r}_n(x) + \overline{r}_n(x) - r(x)|^2 \\ &\leq |r_n^*(x) - r_n(x)|^2 + 2|\overline{r}_n(x) - r(x)|^2. \end{aligned}$$

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \leq 2 \int_{\mathbb{R}^d} |r_n^*(x) - \overline{r}_n(x)|^2 \mu(dx) + 2 \int_{\mathbb{R}^d} |\overline{r}_n(x) - r(x)|^2 \mu(dx). \quad (3.3)$$

Commençons par majorer le second terme de l'inégalité précédente.

$$\int_{\mathbb{R}^d} |\overline{r}_n(x) - r(x)|^2 \mu(dx) \leq \int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_1 Z_1}{H(Z_1)} - r(x)|^2 \mu(dx)$$

De plus

$0 \leq \delta_1 Z_1 \setminus H(Z_1) \leq T_Y \setminus H(T_Y) \ \ p.s. \ et$

$$\begin{aligned} E\left\{\frac{\delta_1 Z_1}{H(Z_1)}\backslash X_1\right\} &= E\left\{\frac{I_{\{Y_1\leq C_1\}}Y_1}{H(Y_1)}\backslash X_1\right\} \\ &= E\left(\frac{Y_1}{H(Y_1)}E(I_{\{Y_1\leq C_1\}}(X_1,Y_1))\backslash X_1\right) \\ &= E(Y_1\backslash X_1)=r(X_1) \end{aligned}$$

Donc d'après le théorème de (Devroye et Krzyzak (1989) [10]) on obtient :

$$\int_{\mathbb{R}^d}|\overline{r}_n(x)-r(x)|^2\mu(dx) \quad \leq \quad \int_{\mathbb{R}^d} |\sum_{i=1}^{n}\frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n}K(\frac{(x-X_i)}{h_n})}\frac{\delta_1 Z_1}{H(Z_1)}-r(x)|^2\mu(dx)\to 0\ (n\to\infty)\ p.s$$

Reste à majorer le premier terme de l'inégalité donnée à la formule (3.3)

$$\begin{aligned} &\int_{\mathbb{R}^d} |r_n^*(x)-\overline{r}_n(x)|^2\mu(dx) \\ \leq\ & T_Y\int_{\mathbb{R}^d}|\sum_{i=1}^{n}\frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n}K(\frac{(x-X_i)}{h_n})}\frac{\delta_1 Z_1}{H_n(Z_1)}-\sum_{i=1}^{n}\frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n}K(\frac{(x-X_j)}{h_n})}\frac{\delta_1 Z_1}{H(Z_1)}|\mu(dx) \\ \leq\ & T_Y\int_{\mathbb{R}^d}\sum_{i=1}^{n}\frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n}K(\frac{(x-X_j)}{h_n})}T_Y|\frac{1}{H_n(Z_i)}-\frac{1}{H(Z_i)}|\mu(dx) \\ \leq\ & T_Y^2\frac{1}{H_n(T_Y)H(T_Y)}\sup_{t\leq T_Y}|H(t)-H_n(t)|\to 0\ (n\to\infty)\ p.s. \end{aligned}$$

à cause de $(H_{1,2})$ et du lemme 3.1.

le fait que Guessoum et Ould said (2009) 21 ont modifié légèrement H_n en lui imposant de s'annuler à partir de l'observation la plus grande. Ils ont alors, d'une part établi la convergence presque sûre uniforme sur des compacts de l'estimateur ainsi obtenu, donné des vitesses de convergence et prouvé d'autre part sa normalité asymptotique. Nous nous devons aussi de faire remarquer que le résultat donné au théorème précédent a été aussi

prouvé dans Kohler, Màthé et Pintér (2002) pour des estimateurs à poids (plus proches voisins et à partitions) dans un modèle de censure à droite. Dans la suite de ce chapitre, en nous inspirant de la démarche précédente, nous proposons un résultat similaire dans un modèle de censure à gauche.

3.2 Censure à gauche

Dans ce cas, l'événement d'intérêt peut se produire avant le début de l'étude, au temps C_i. Si c'est le cas, on ne connaît pas le moment exact de l'événement.
Donc les observations peuvent être représentées par le maximum entre Y et une variable de censure C.
Plus précisément soit Y une variable d'intérêt positive et bornée et C une variable aléatoire de censure positive.
Nous observons l'échantillon $D_n = \{(X_i, Z_i, \delta_i), i = 1, \ldots, n\}$, où $Z_i = Y_i \vee C_i$ et $\delta_i = I_{Y_i \geq C_i}$ (δ_i est l'indicatrice de censure).
Nous nous proposons d'estimer $r(x) = E(Y \backslash X = x)$.

Définition de l'estimateur

Soient $F(t) = P(Y \leq t)$ et $G(t) = P(C \leq t)$ les fonctions de répartition respectives de Y et C. On a besoin de l'hypothèse (H_2) englobant les hypothèses suivantes

$$(H_2) : \begin{cases} (H_{2.1}) & C \text{ et } (X, Y) \text{ sont indépendants} \\ \\ (H_{2.2}) & T_Y \vee T_c < \infty \text{ et } G(I_Y) > 0. \end{cases}$$

Ici la condition $G(I_Y) > 0$ implique $I_C \leq I_Y$. Soit l une fonction de $\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. On se propose d'estimer la moyenne $E\{l(X, Y)\}$ à partir de l'échantillon
$D_n = \{(X_i, Z_i, \delta_i), i = 1, \ldots, n\}$.
Sous l'hypothèse d'indépendance, un "estimateur" sans biais de cette quantité est donné

par :

$$E\{l(X,Y)\} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i l(X_i,Z_i)}{G(Z_i)}.$$

En effet :

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i l(X_i,Z_i)}{G(Z_i)}\right\} &= E\left\{\frac{\delta_1 l(X_1,Z_1)}{G(Z_1)}\right\} \\ &= E\left\{\frac{I_{\{Y_1\geq C_1\}} l(X_1,Y_1)}{G(Z_1)}\right\} \\ &= E\left(\frac{l(X_1,Y_1)}{G(Z_1)}E(I_{\{Y_1\geq C_1\}}\backslash(X_1,Y_1))\right) \\ &= E(l(X_1,Y_1)). \end{aligned}$$

- En pratique, G est inconnue. On ne peut donc pas utiliser " l'estimateur" tel quel, il est alors naturel de remplacer G par un estimateur, obtenu par adaptation au contexte de censure à gauche de l'estimateur de Kaplan Meier et donné par :

$$G_n(t) = \begin{cases} \prod_{i=1}^{n}\left[1-\frac{1-\delta_i}{i}\right]^{I_{[Z_{(i)}>t]}}, & si \quad t > T_{L,n}, \\ \\ \prod_{i=1}^{n}\left[1-\frac{1-\delta_i}{i}\right]^{I_{[Z_{(i)}>T_{L,n}]}}, & si \quad t \leq T_{L,n}. \end{cases}$$

où $T_{L,n} = \inf\{Z_1,\ldots,Z_n\}$. Remarquons que cet estimateur ne s'annule jamais. On peut se reporter à Kebabi et Messaci (2009)[8] pour des détails sur cet estimateur, il y est en particulier montré le résultat suivant.

Lemme 3.3. *: On a*

$$\sup_{t\geq I_Y}|G_n(t)-G(t)| \to 0 \;\; n\to\infty \;\; p.s$$

Un estimateur possible de $E(l(X,Y))$ est alors donné par :

$$E\{l(X,Y)\} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Z_i}{G_n(Z_i)}.$$

Cela suggère d'estimer $r(x)$ par :

$$\hat{r}_{n,g}(x) = \sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_i Z_i}{G_n(Z_i)} \tag{3.4}$$

Du fait que $0 \leq Y \leq T_Y < \infty$ p.s, on a $0 \leq r(x) \leq T_Y$, on l'estime donc plutôt par :

$$r_{n,g}(x) = \begin{cases} Z_{(n)}, & si \quad \hat{r}_{n,g}(x) > Z_{(n)}, \\ \hat{r}_n, & si \quad 0 \leq \hat{r}_{n,g}(x) \leq Z_{(n)}, \\ 0, & si \quad \hat{r}_{n,g}(x) < 0. \end{cases} \tag{3.5}$$

où $Z(n) = \max(Z_1, \ldots, Z_n)$. Par analogie avec (3.5), posons

$$r^*_{n,g}(x) = T_{[0,T_Y \vee T_C]}(\hat{r}_{n,g}(x)) = \begin{cases} T_Y \vee T_C, & si \quad \hat{r}_{n,g}(x) > T_Y \vee T_C, \\ \hat{r}_{n,g}(x), & si \quad 0 \leq \hat{r}_{n,g}(x) \leq T_Y \vee T_C, \\ 0, & si \quad \hat{r}_{n,g}(x) < 0. \end{cases}$$

3.2.1 Propriétés de l'estimateur

Lemme 3.4. *: Sous l'hypothèse $(H_{2,2})$, on a*

$$\int_{\mathbb{R}^d} |r_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \ p.s$$

si et seulement si

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \text{ p.s}$$

Preuve :

$$\begin{aligned} |r^*_{n,g}(x) - r(x)|^2 &= |r^*_{n,g}(x) - r_{n,g}(x) + r_{n,g}(x) - r(x)|^2 \\ &\leq 2|r^*_{n,g}(x) - r_{n,g}(x)|^2 + 2|r_{n,g}(x) - r(x)|^2 \end{aligned}$$

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r(x)|^2 \mu(dx) \leq 2\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r_{n,g}(x)|^2 \mu(dx) + 2\int_{\mathbb{R}^d} |r_{n,g}(x) - r(x)|^2 \mu(dx).$$

On a :

$$\int_{\mathbb{R}^d} |r_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \text{ p.s}$$

Donc il suffit de montrer que :

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r_{n,g}(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \text{ p.s}$$

$$\begin{aligned} |r^*_{n,g}(x) - r_{n,g}(x)| &= |T_{[0,T_Y \vee T_C]} \hat{r}_{n,g}(x) - T_{[0,Z_{(n)}]} \hat{r}_{n,g(x)}| \\ &\leq T_Y \vee T_C - Z_{(n)}. \\ \int_{\mathbb{R}^d} |r^*_{n,g}(x) - r_{n,g}(x)|^2 \mu(dx) &\leq \int_{\mathbb{R}^d} (T_Y \vee T_C - Z_{(n)})^2 \mu(dx) \\ &= (T_Y \vee T_C - Z_{(n)})^2. \end{aligned}$$

et $Z_{(n)} \to T_Y \vee T_C$ $(n \to \infty)$ p.s. Alors :

$$\int_{\mathbb{R}^d} |r^*_{n,g(x)} - r_{n,g}(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \ p.s.$$

D'où le résultat ■.

Thoérème 3.2. *:*

Sous l'hypothèse (H_2) *et si* K *est un noyau régulier,* $\lim_{n\to\infty} h_n = 0, \lim_{n\to\infty} nh_n^d = \infty$ *alors l'estimateur* $r_{n,g}(x)$ *défini par (*3.4*) et (3.5) vérifie :*

$$\int_{\mathbb{R}^d} |r_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \ p.s$$

Preuve :

D'après le lemme précédent, il suffit de montrer que :

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \text{ p.s}$$

Posons

$$\overline{r}_{n,g}(x) = T_{[0,T_Y \vee T_C]} = \left(\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_i Z_i}{G(Z_i)} \right)$$

On a :

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r(x)|^2 \mu(dx) \leq 2 \int_{\mathbb{R}^d} |r^*_{n,g}(x) - \bar{r}_{n,g}(x)|^2 \mu(dx) + 2 \int_{\mathbb{R}^d} |\bar{r}_{n,g}(x) - r(x)|^2 \mu(dx). \quad (3.6)$$

Commençons par majorer le second terme de l'inégalité précédente.

Remarquons que $\int_{\mathbb{R}^d} |\bar{r}_{n,g}(x) - r(x)|^2 \mu(dx) \leq \int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_1 Z_1}{G(Z_1)} - r(x)|^2 \mu(dx)$

De plus

$0 \leq \delta_1 Z_1 \setminus G(Z_1) \leq T_Y \vee T_G \setminus G(I_Y) \ \ p.s.$

et

$$\begin{aligned} E\left\{ \frac{\delta_1 Z_1}{H(Z_1)} \backslash X_1 \right\} &= E\left\{ \frac{I_{\{Y_1 \geq C_1\}} Y_1}{H(Y_1)} \backslash X_1 \right\} \\ &= \left(\frac{Y_1}{H(Y_1)} E(I_{\{Y_1 \geq C_1\}}(X_1, Y_1)) \backslash X_1 \right) \\ &= E(Y_1 \backslash X_1) = r(X_1) \end{aligned}$$

Donc d'après le théorème de (Devroye et Krzyzak (1989)[10]) on obtient :

$$\int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_1 Z_1}{G(Z_1)} - r(x)|^2 \mu(dx) \to 0 \ (n \to \infty) \quad p.s$$

Reste à majorer le premier terme de l'inégalité donnée à la formule (3.7)

$$\begin{aligned}
&\int_{\mathbb{R}^d} |r^*_{n,g}(x) - \overline{r}_{n,g}(x)|^2 \mu(dx) \\
\leq\ & T_Y \vee T_C \int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_i)}{h_n})} \frac{\delta_1 Z_1}{G_n(Z_1)} - \sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_1 Z_1}{G(Z_1)} | \mu(dx) \\
\leq\ & T_Y \vee T_C \int_{\mathbb{R}^d} \sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} T_Y \vee T_C | \frac{1}{G_n(Z_i)} - \frac{1}{G(Z_i)} | \mu(dx) \\
\leq\ & (T_Y \vee T_C)^2 \frac{1}{G_n(I_Y) G(I_Y)} \sup_{t > I_Y} |G_n(t) - G(t)| \to 0 \ (n \to \infty) \quad p.s.
\end{aligned}$$

le lemme (3.3) et par le fait que $G(I_Y) > 0$ et $G_n(I_Y) > 0$.

Remarquons que contrairement au cas de la censure à droite, nous n'avons pas utilisé l'hypothèse de la continuité de G.

BIBLIOGRAPHIE

[1] A. Carbonez. *Non parametric Functional Estimation under Random Censoring and a New Semiparametric Model of Random Censorship.* Phd Thesis Katholieke Universiteit leuven, (1992).

[2] A. Carbonez,G, Van Der Meulen. *Partition-estimates of a regression function under random censoring,* Statist. Decisions $13, 21 - 37, (1995)$.

[3] Bochner, S. *Harmonic Analysis and the Theory of probability.* University of Chicago Press, Chicogo, Illinois. (1955).

[4] E. A. Nadaraya. *On estimating regression. Theori. Probab. Appl.* $9, 141 - 142(1964)$.

[5] E. L. Kaplan and P. Meier.*Non parametric estimation from incomplete observation.* J. Amer. Statist. Assoc. 53, 457-481, (1958).

[6] E. Ould Said and Z. Cai. *Strong uniform consistency of non parametric estimation of the consored conditional mode function.* J. Non parametric Stat. $17, No.7, 797 - 806.(2005)$.

[7] G. S. Watson. *Smooth regression analysis.* Sankhyà Ser. A $26, 359 - 372, (1964)$.

[8] K. Kebabi et F. Messaci. *Estimation non paramétrique dans un modèle de censure à gauche. Communication RAMA09.* Alger, (2009).

[9] L. Devroye, A. Krzyazak, and G. Lugosi. *On the strong universal consistency of nearest neighbor regression function estimates.* Ann. Statist. $22, 1371 - 1385, (1994)$.

[10] L. Devroye, and A. Krzyàzak. *An equivalence theorem for L1 convergence of the kernel regression estimates, J.* Statist. Plann. Inference $23, 71 - 82, (1989)$.

[11] L. Devroye, L. G. *Distribution free exponential bounds on the L*1 *error of partitioning estimates of a regression function, in Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics(F. Konecny, J. Mogyorodi, and W. Wertz, Eds.)* Akadémiai Kiado, Budapest, Hungary, pp.$67 - 76, (1983)$.

[12] M. Kohler, K. Màthé and M. Printér. *Prediction from Randomly Right Censored Data.* J. Multivariate, Anal. $80, 73 - 100, (2002)$.

[13] M. P. Wand, et M. C. Jones. Kernel Smoothing. Chapman and Hall, London, (1995).

[14] M. Kohler, and A. Krzyzak, *Non parametric regression estimation using penalized least squares, to appear in IEEE Trans.* Infrom. Theory (2001).

[15] Prakasa, Rao, B.L.S. *Non parametric Functional Estimation.* Academic Press, New York, (1983).

[16] Parzen, E. *On estimation of a probability density function and mode.* Ann. Math. Statist , $33, 1065 - 1076, (1962)$.

[17] R. Beran. *Non parametric regression with randomly censored survival data.* Technical uni- versity of Clifornia, Berkeley. (1981).

[18] R. D. Gill and S. Johansen. *A survey of product integration with a view toward application in survival analysis.* The Annals of Statistics. Vol 18, $No4, 1501-1555, (1990)$.

[19] Rosenblatt, M. *Remarks on some nonparametric estimates of a density function.* Ann. Math. Statist , $27, 832 - 837(1956)$..

[20] W. Stute and J. L. Wang. *The strong law under random censorship.* Ann. Statist, $21, 1591 - 1607, (1993)$.

[21] Z. Guessoum and E. Ould Said.*On the non parametric estimation of the regression function under censorship model.* Statist. Decisions. $26, 159 - 177, (2009)$.

Résumé

Dans ce mémoire, nous étudions les estimateurs à noyau de la fonction de régression dans différents contextes, à savoir pour des données complètes ainsi que pour des données censurées (à droite et à gauche). De plus nous étudions quelque propriétés de ces estimateurs.

Mots-clés : Estimation non paramétrique, fonction de régression, estimateur à nouyau, données censurées , convergence presque sûre.

ملخص

في هذه المذكرة تطرقنا الى دراسة تقدير دالة الانحدار بطريقة النواة في سياقات مختلفة، وهي ما تعرف بالمعطيات الكاملة وبمعطيات المراقبة على اليمين و على اليسار. اضافة الى ذلك، ندرس بعض خصائص هذه التقديرات .

الكلمات المفتاحية:التقدير غيرالوسيطي، دالة الانحدار، مقدر النواة، بيانات المراقبة، التقارب المؤكد تقريبا.

Abstract

In this memory, we study the kernel estimators of the regression function in different contexts, namely for completed data and for censored data (right and left). Furthermore we study some properties of these estimators.

Keywords : nonparametric estimation, regression function, kernel estimator, censored data, almost sure convergence.