UNIVERSITÉ KASDI MERBAH
OUARGLA

Faculté des mathématiques et sciences de la matière

DÉPARTEMENT DE MATHÉMATIQUES

MASTER

Spécialité : Mathématiques

Option : Probabilités et statistiques

Par : NOURREDDINE GEURRICHA

Thème

L'estimateur de la régression en utilisant la méthode de noyau dans les modèles de censurées

Version de : 20/10/2020

Devant le jury composé de :

Prof. ABBASSI.H.	Université KASDI Merbah- Ouargla	Président
Dr.HALILI.R.	Universié KASDI Merbah- Ouargla	Examinateur
Dr. AGOUNE.R	Université KASDI Merbah- Ouargla	Rapporteur

Année Universitaire : 2019/2020

Dédicace

Je dédie ce modeste travail
A ma très cher Mère et mon très cher Père
A mes chers sours et frères
A toutes la familles GUERRICHA
A ceux qui ont veillé pour mon bien être
A ceux qui m'ont toujours encouragé pour que je réussisse dans mes études
A tout ce qui m'ont encouragé lors de la réalisation de mon travail
Sans oublier de dédier ce mémoire à mes très chères amies intimes
A tous mes collègues de ma promotion de Proba-Stat 2020
Finalement à tous ceux qui m'on aider de proche ou de loin

Remerciements

Je remercie Dieu le tout-puissant de m'avoir donné la volonté, la force et le courage pour bien mener et finir mon travail de mémoire.
Je voudrais d'abord et avant tout remercier ma encadreur vertueuse Dr.RACHID AGOUNE pour tous leurs efforts en vue d'établir cette mémoire, elle a eu le rôle fondamental et essentiel et la grand mérite dans tout ce qui a été réalisé, comme cela avait toujours été, près de moi et me guider et corriger mes erreurs et me donner de précieux conseils et les conseils appropriés et les alertes considérés, je la répète mes remerciements pour tout ce qu'elle a fait pour moi à travers la mise en plan à toutes les étapes de la préparation de ce mémoire depuis le début jusqu à la fin, étape par étape, était que le premier et dernier facteur pour le succès de ce travail, et je lui dis encore une fois merci beaucoup à pour votre appréciation profonde.

Avec un grand honneur, j'aimerais présenter mes remerciements et ma gratitude aux membres du jury, Monsieur Dr: ABASSI.H, et Monsieur Dr: HALILI.R tout d'abord d'avoir accepté d'exa- miner mon mémoire, qui sans eux ce mémoire ne pourra jamais voir le jour pour ntérêt et apport qu'ils ont apporté à mon travail.

J'exprime ma gratitude à ma famille qui m'a toujours soutenue et encouragée dans la voie que je m'étais fixèe. Je remercie particulièrement mes parents qui m'ont stimulée et encouragé pendant mes études. qui étaient tou- jours prêts à fournir tous les moyens physique et morale pour la réussite de ce projet.

Contents

Introduction

Les méthodes de régression sont utiles pour modéliser la corrélation entre 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.
la première approche consiste à utiliser un modèle de régression paramétrique. peut s'écrire comme une fonction explicite des valeurs de X. Cette dernière peut être linéaire, sous forme:

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

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 ramenons alors à l'estimation d'un nombre fini de paramètres. Dans certains cas nous pouvons disposer pour cette estimation d'un échantillon $\{(X_i, Y_i), i = 1 \cdots n\}$ de couples indépendants et ayant chacun la même
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 nonparamétrique. Dans ce cas on ne dispose d'aucune forme paramétrique pour r.
Etudier le lien entre deux variables a généralement pour but de prédire 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 . Plusieurs paradigmes d'estimation non-paramétrique de la régression sont disponibles dont, l'estimation des moindres carrés, et l'estimation des moindres carrés pénalisés ou spline de lissage.

L'estimation à poids ,parmi les estimateurs utilisés de cette derniére les k plus proches voisin et l'estimateur à noyau de Nadaraya-Watson. Ce dernier, qui a été introduit indépendamment par Nadaraya (1964) et Watson (1964), est l'un des plus populaires des estimateurs de régression non-paramétriques. Son expression est

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

où K est une fonction de R dans 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. C'est pourquoi cet estimateur a été généralisé dans le cas où la variable réponse est censurée à droite ,ont étudié cet estimateur aussi bien dans le cas où les X_i sont iid que dans le cas où les X_i sont αmélangeantes.
Un phénomène de censure à gauche peut aussi empêcher l'observation du phénomène d'intérêt pour lequel on saura seulement qu'il est inférieur à la valeur observée. Généralement, la censure à gauche s'accompagne de la censure à droite comme cela est le cas pour la censure mixte à laquelle nous nous intéressons dans ce mémoire.
Dans ce contexte, Patilea et Rolin (2006) donnent un estimateur produitlimite de la fonction de survie de la durée d'intérêt X qui généralise le célèbre estimateur de Kaplan et Meier (1958) et démontrent sa convergence uniforme presque sure ainsi que sa normalité asymptotique sous certaines . En ce qui concerne la régression, Messaci (2010) a introduit des estimateurs à poids de la régression, dont l'estimateur à noyau. Kebabi et al. (2011) ont donné des estimateurs des moindres carrés et montré leur convergence vers la valeur optimale presque sûrement.

Chapter 1

Rappel sur l'analyse de survie

1.1 Données censurées

Cette mémoire s'intéressant à l'estimation dans un modèle de censure, commençons par présenter cette notion de censure. En analyse de survie et en fiabilité, on s'intéresse au temps qui s'écoule jusqu'à la réalisation d'un certain événement. On appelle ce temps un temps de défaillance, la durée de vie, la durée de survie, ou simplement une durée. C'est une variable aléatoire positive et souvent supposée bornée. Cela peut être la durée de vie d'un patient après un traitement, la durée de chômage, le temps de panne d'un appareil, l'âge auquel un enfant apprend à accomplir une tâche donnée, etc. Il arrive souvent, pour diverses raisons, que la durée d'intérêt ne puisse pas être observée. Cela peut être dû à la perte de vue d'un patient, au début ou à la fin de la période d'étude, etc. Ces valeurs sont censurées. Les valeurs censurées, bien qu'inconnues, doivent être prises en compte pour obtenir des estimations correctes et des conclusions précises. Selon la situation spécifique, la littérature statistique contient un grand nombre de procédures qui permettent de tenir compte des observations censurées. Il existe plusieurs types de censure.

Définition (Variable de censure).

La variable de censure Y est défnie par la non-observation de l'événement étudié. Si au lieu d'observer X, on observe Y , et que l'on sait que $(X > Y)$ (respectivement $X < Y$, $Y_1 < X < Y_2$) , on dit qu'il y a censure à droite (respectivement censure à gauche, censure par intervalle) . Pour un individu donné j, on va considérer.
- son temps de survie X_j ,

- son temps de censure Y_j ,
- la durée réellement observée Z_j .

Types de Censures

La censure des données se fait selon plusieurs mécanismes tel que la censure à droite, la censure à gauche, la censure double (ou mixte).

1. Censure à droite

Il y a censure à droite lorsque la durée de survie d'intérêt est supérieure à la durée observée. Un exemple typique est celui où l'événement d'intérêt est le décès d'un patient malade et la durée d'observation est une durée totale d'hospitalisation (on n'a plus de nouvelles après l'hospitalisation) . On peut aussi observer ce genre de phénomène dans des études de fiabilité quand la panne d'un appareil ne permet pas de poursuivre l'observation pour l'appareil objet de notre étude. Pour ce type de censure, tout ce que l'on sait est que la vraie durée est supérieure à la durée observée.

2. Censure à gauche

Il y a censure à gauche lorsque nous observons la censure C (et non pas la durée de vie T) et que nous savons que $T < C$. Un phénomène symétrique au précédent se produit, le patient à déjà subi l'événement avant l'instant où on commence l'étude. Ce modèle est par exemple adapté au cas où l'on s'intéresse à l'âge auquel un individu commence à accomplir une tâche. Tout ce qu'on sait chez l'individu censuré est que le véritable âge est inférieur à la valeur observée (l' âge au moment de l'étude par exemple).

3. Censure par intervalle

Il y a censure par intervalles 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 périodiques de patients et généralise aussi bien le modèle de censure à droite que celui de censure à gauche.

Censure de type I : fixe

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 : attente

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 de type III : aléatoire

C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type d'expériences, la date d'inclusion du patient dans l'étude est fixée, mais la date de n d'observation est inconnue (celle-ci correspond, par exemple, à la durée d'hospitalisation du patient). Soit $\{X_1, ..., X_n\}$ un échantillon d'une va positive X, on dit qu'il y a cen- sure aléatoire de cet échantillon s'il existe une autre va positive elle aussi Y d'échantillon $Y_1, \cdots, Y_n$ dans ce cas au lieu d'observer les $X_j^{'}(s)$, on observe un couple de va's (Z_j, δ_j) avec

$Z_j := \min(X_j, Y_j)$ et$\delta_j := 1I\{X_j \leq Y_j\}$ pour $j = 1, \cdots, n$

où δ_j l'indicateur de censure, qui détermine si X a été censuré ou non

- si $\delta_j = 1$, la durée d'intérêt est observée $(Z_j = X_j)$
- si $\delta_j = 0$, elle est censurée $(Z_j = Y_j)$.

On observe des durées incomplètes. Dans cette thèse, on s'intéresse uniquement au cas des censures à droite du type aléatoire. Celui-ci correspond à un modèle fréquemment utilisé en pratique, ce qui justife amplement qu'on y attache à l'intérêt.

1.2 Estimateur de Kaplan-Meier

Soit $X_1, ..., X_n$ un échantillon représentant les durées d'intérêt (ces variables sont donc supposées positives), de fonction de répartition F, et $C_1, ..., C_n$ un échantillon représentant les temps de censure, que l'on suppose indépendants des durées d'intérêt, de fonction de répartition G. Dans le modèle de censure aléatoire à droite, on observe non pas la durée d'intérêt X_i mais plutôt la plus petite des deux valeurs $Z_i = \min(X_i, C_i)$, ainsi

que l'indicateur de censure δ_i qui vaut 1 si la durée d'intérêt est observée, et 0 si elle est censurée, $\delta_i = 1_{\{X_i \leq C_i\}}$.

Dans ce genre de données, qui sont souvent des durées de survie ou des données de fiabilité, la fonction de répartition F est estimée par l'estimateur introduit par Kaplan et Meier (1958), donné pour $z < Z_n$ où

$Z(n) = \max(Z_1, ..., Z_n)$ par

$F_n(Z) = 1 - \prod_{i:Z_i \leq z} \left(\frac{N_i(Z_i)-1}{N_i(Z_i)}\right)^{\delta_i}$

avec $N_n(x) = \sum_{i=1}^{n} 1_{Z_i \geq x}$. Pour $z \geq Z(n)$, il y a plusieurs conventions pour définir $F_n(z)$: Soit on le définit par $F_n(Z_{(n)})$, ce qui fait que F_n peut ne pas être une fonction de répartition si Z_n est une donnée censurée, soit on le définit par 0, soit on le laisse non défini.

Cet estimateur a des propriétés assez similaires à celles la fonction de répartition empirique : la convergence uniforme presque sûre (Stute et Wang, 1993; Winter et al., 1978), la normalité asymptotique (Breslow et Crowley, 1974; Gill, 1983), et la loi du logarithme itéré (Földes et Rejt"o, 1981).
Le cas de la censure mixte fera l'objet des sections suivantes.

1.3 Quelques propriétés sur l'estimateur de Kaplan - Meier

En analyse de survie, $\widehat{F}_n$ joue pour les données incomplétes le même rôle que la fonction de répartition empirique pour les données classique.

a)Biais et convergence
L'estimateur de Kaplan - Meier est légèrement biaisé : en général,

$$E(\widehat{F}_n(t`)) < F(t), \tag{1.1}$$

où E désigne l'espérance. Il est en revanche convergent (consistent estimator) :

$$\forall \epsilon > 0, \lim_{n\to\infty} P(|\widehat{F}_n(t) - F(t)| \geq \epsilon) = 0 \tag{1.2}$$

Il est donc asymptotiquement non biaisé :

$$\lim_{n\to\infty} E(\widehat{F}_n(t)) = F(t) \tag{1.3}$$

b)Auto-cohérence
En l'absence de censure, un estimateur de $S(t)$ est

$$\widetilde{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \Pi(t_i > t) \tag{1.4}$$

En présence de censure, on peut encore écrire

$$\widetilde{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} (\delta_i \Pi(t_i > t) + (1 - \delta_i)\Pi(t_i > t)), \tag{1.5}$$

mais la valeur de $\Pi(t_i > t)$ n'est pas connue pour les données censurées. Si un estimateur . $S_n(t)$ de $S(t)$ est connu, on peut estimer l'espérance de $\Pi(t_i > t)$
sachant que $\delta_i = 0ett_i > s_i$: on a

$$E(\Pi(t_i > t)\delta_i = 0et \quad t_i > s_i) = \frac{P(t_i > t)}{P(t_i > s_i)}, \tag{1.6}$$

donc

$$\widetilde{E}(\Pi(t_i > t)\delta_i = 0et \quad t_i > s_i) = \frac{\widetilde{S}_n(t)}{\widetilde{S}_n(s_i)}, \tag{1.7}$$

L'estimateur de Kaplan-Meier présente la propriété d'être auto-cohérent, c.- à-d. que

$$\widehat{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\delta_i \Pi(t_(i) > t) + (1 - \delta_i)\frac{\widehat{S}_n(t)}{\widehat{S}_n(s_i)}\right) \tag{1.8}$$

Si l'on part d'une fonction de survie arbitraire (mais compatible avec les contraintes sur une fonction de survie)$\widetilde{S}_n^0$, on peut calculer itérativement une estimation $\widetilde{S}_n^{(k)}$ par

$$\widetilde{S}_n^{(k)}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \Pi(t_(i) > t) + (1-\delta_i)\frac{\widetilde{S}_n^{(k-1)}(t)}{\widetilde{S}_n^{(k-1)}(s_i)}\right) \tag{1.9}$$

et l'on a $\lim_{k\to n} \widetilde{S}_n^{(k)} = \widehat{S}_n$

c)Limitations de l'estimateur de Kaplan - Meier

L'estimateur de Kaplan - Meier est discontinu. Pour certaines applications, il est nécessaire de le lisser en le convoluant avec un noyau. Il ne prend pas en compte les incertitudes sur les valeurst_i et les censuress_i. Elles sont négligeables en statistiques médicales (la date de décès ou de sortie de l'échantillon d'un patient est connue précisément). Ces incertitudes s'ajoutant à la dispersion intrinsèque de la loi de distribution (qu'on peut caractériser par l'écart-type autour de la moyenne), la dispersion apparente estimée à partir de l'estimateur de Kaplan - Meier.et les simulations, de type bootstrap par exemple, peuvent permettre de modéliser ces phénomènes et de corriger leurs effets.

1.4 Modèle de censure mixte

Considérons trois variables aléatoires positives indépendantes X, L et R de fonctions de répartition respectives F_X, F_L et F_R, et de fonctions de survie respectives S_X, S_L et S_R, où X représente la durée d'intérêt et L et R sont les durées de censure à gauche et à droite respectivement. Dans le
modèle I de Patilea et Rolin (2006), au lieu d'observer un échantillon de X on observe un échantillon du couple (Z, A) où $Z = \max(\min(X, R), L)$ et

$$A = \begin{cases} 0 & , siL < T < R \\ 1 & , siR < \max(T, L) \\ 2 & , siT \leq L \leq R \end{cases} \tag{1.10}$$

Ce modèle considère la censure à droite et la censure à gauche comme deux phénomènes qui agissent indépendamment l'un de l'autre mais l'un

peut censurer l'autre. Patilea et Rolin (2006) ont proposé d'estimer S_X, fonction de survie de la variable d'intérêt X comme suit.

1.5 Estimation

Considérons$H$ la fonction de répartition de Z, elle peut s'écrire

$\sum_{k=0}^{2} H^{(k)}(t)$

où

$H^{(k)}(t) = P(Z \leq t, A = k)$, pour $k = 0, 1, 2$.

En notant pour toute application R de R dans R, $R(t_-)$ la limite de R à gauche de t, lorsque cette limite existe, ces fonctions s'écrivent

$H^{(0)}(t) = \int_0^t F_L(u) S_R(u) dF_X(u)$,

$H^{(1)}(t) = \int_0^t F_L(u) S_X(u) dF_R(u)$,

$H^{(2)}(t) = \int_0^t \{1 - S_X(u) S_R(u)\} dF_L(u)$,

et c'est à partir de ces équations que l'estimateur est obtenu.
L'idée est de considérer dans un premier temps $Y = \min(X, R)$ et L dans un modèle de censure à gauche (c'est-à-dire que l'on considère une donnée complète si $A = 0$ ou $A = 1$ et censurée à gauche si $A = 2$), et d'estimer la fonction de répartition de Y , puis l'utiliser pour estimer la fonction de répartition de la variable d'intérêt X en considérant un modèle de censure à droite.
L'estimateur de la fonction de survie S_X ainsi obtenu, en remplaçant à la fin les fonctions $H^{(0)}$, $H^{(1)}$ et $H^{(2)}$ par leurs estimateurs empiriques, obtenus à partir d'un échantillon $(Z_i, A_i)_{1 \leq i \leq n}$, est donné par :

$\widehat{S}_n(Z'_j) = 1 - F_n(Z'_j) = \prod_{1 \leq l \leq j} \left\{1 - \frac{L_{0l}}{U_{l-1} - N_{l-1}}\right\}$

où $(Z'_j)_{1 \leq j \leq M}$ sont les valeurs distinctes des Z_i prises dans l'ordre croissant,
et

$$\begin{aligned} D_{kj} &= \sum_{1\le i\le n} 1_{Z_i=Z'_j, A_i=k}, \\ N_j &= \sum_{1\le i\le n} 1_{Z_i\le Z'_j} \\ U_{j-1} &= n\prod_{j\le l\le M}\left\{1-\tfrac{D_{2l}}{N_l}\right\} \end{aligned} \tag{1.11}$$

pour $0 \le l \le 2$ et $1 \le j \le M$.

Soulignons le fait que si $L \equiv 0$ (pas de censure à gauche), S_n se réduit à l'estimateur de Kaplan-Meier qui lui-même se réduit au complément à 1 de la fonction de répartition empirique si $R \equiv \infty$.
Patilea et Rolin (2006) ont introduit cet estimateur et montré sa convergence uniforme presque sûre et sa convergence en tant que processus vers un gaussien sous des conditions d'identifiabilité du modèle.
S_n intervenant plus loin au dénominateur dans les expressions d'estimateurs de la fonction de régression, le résultat suivant donne une condition nécessaire et suffisante pour qu'il s'annule.
Dans un souci de clarté, nous notons dans la suite de ce chapitre D_{kj} par $D_{k,j}$.

Lemme 1.1.
i) Une condition nécessaire et suffisante pour que $\widehat{S}_n(Z^{'}_{k_0}) = 0$ pour la première fois et reste nul est :$D_{0,k_0} \neq 0$,$D_{1,k_0} = 0$ et $j > k_0$,$D_{0,j} = D_{1,j} = 0 si k_0 \neq M$.
ii) $\widehat{S}_n$ s'annule pour la première fois en $Z^{'}_M$ si et seulement si $D_0, M \neq 0$ et $D_{1,M} = 0$

Démonstration. 1. Commençons par montrer que pour tout $k : 0 \le k \le M-1$, nous avons

$$U_k \ge N_k \tag{1.12}$$

En effet

$$\begin{aligned} U_k &= n\big(\tfrac{N_{k+1}-D_{2,(k+1)}}{N_{k+1}}\big)\big(\tfrac{N_{k+2}-D_{2,(k+2)}}{N_{k+2}}\big)\cdots\big(\tfrac{N_M-D_{2,M}}{N_M}\big) \\ &= \Big(\tfrac{N_k+D_{0,(k+1)}+D_{1,(k+1)}}{N_{k+1}}\Big)\cdots\Big(\tfrac{N_{M-2}+D_{0,(M-1)}+D_{1,(M-1)}}{N_{M-1}}\Big) \\ &\quad\times\big(N_{M-1}+D_{0,M}+D_{1,M}\big) \\ &\ge \tfrac{N_k}{N_{k+1}}\times\cdots\times\tfrac{N_{M-2}}{N_{M-1}}\times N_{M-1} = N_k \end{aligned} \tag{1.13}$$

Remarquons que s'il existe j tel que $j > k$ avec $D_{1,j} \neq 0$ ou $D_{0,j} \neq 0$ alors

$$U_k > N_k. \tag{1.14}$$

2. Soit k_0 le premier indice k tel que $D_{0,k} = U_{k-1} - N_{k-1}$ (le premier k pour

lequel $\widehat{S}_n(Z_k^{'})$. Nous avons alors :

$D_{0,k_0} \neq 0$ et $D_{0,k_0} = U_{k_0-1} - N_{k_0-1}$

Par ailleurs

$$U_{k_0-1} = n\left(1 - \frac{D_{2,k_0}}{N_{k_0}}\right) \cdots \left(1 - \frac{D_{2,M}}{N_M}\right) = \left(1 - \frac{D_{2,k_0}}{N_{k_0}}\right) U_{k_0}. \tag{1.15}$$

D'après (1.2) et (1.3), il vient

$$\begin{aligned} D_{0,k_0} + N_{k_0-1} &= (\frac{N_{k_0} - D_{2,k_0}}{N_{k_0}}) U_{k_0} \\ &= \left(\frac{N_{k_0} - D_{0,k_0} + D_{1,k_0}}{N_{k_0}}\right) U_{k_0}, \end{aligned} \tag{1.16}$$

et en vertu de (1.1), nous devons avoir$D_{1,k_0} = 0$, donc $U_{k_0} = N_{k_0}$ alors$D_{0,k_0} \neq 0$,
$D_{1,k_0} = 0 et \forall j > k_0, D_{1,j} = D_{0,j} = 0$, (au-delà de k_0, ce qui montre que
la condition énoncée est nécessaire
Montrons que la condition nécessaire est aussi suffisante.
Supposons que

$D_{1,k_0} = 0, D_{0,k_0} \neq 0$ et $\forall j > k_0, D_{1,j} = D_{0,j} = 0$, et montrons que

$\widehat{S}_n(Z_{k_0}^{'}) = 0.$

Nous avons

$$\begin{aligned} U_{k_0-1} &= n(\frac{N_{k_0} - D_{2,k_0}}{N_{k_0}})(\frac{N_{k_0+1} - D_{2,k_0+1}}{N_{k_0+1}}) \cdots (\frac{N_M - D_{2,k_M}}{N_M}) \\ &= n\left(\frac{N_{k_0-1} + D_{0,k_0}}{N_{k_0}}\right)\left(\frac{N_{k_0}}{N_{k_0+1}}\right) \cdots \left(\frac{N_{M-1}}{N_M}\right) \\ &= N_{k_0-1} + D_{0,k_0}, \end{aligned} \tag{1.17}$$

avec $D_{0,k_0} \neq 0$, autrement dit $\widehat{S}_n(Z_{k_0}^{'})$

Chapter 2

Loi du logarithme itéré pour l'estimateur de la fonction de survie

La loi du logarithme itéré (notée LIL) pour une somme $S_n = \sum_{i=1}^{n} x_i$ de variables aléatoire indépendantes et de même loi, remonte à Khinchine et Kolmogorov dans les années 1920. Depuis, un grand nombre de travaux ont porté sur des lois du logarithme itéré pour différents estimateurs.

2.1 Loi du logarithme itéré classique

La loi du logarithme itéré (ou LIL pour "Law of the Iterated Logarithm") est un des théorème limites importants de la statistique. Sa version initiale, donnée pour une somme de variables aléatoires indépendantes et de même loi, remonte à Khinchine et Kolmogorov dans les années 1920. Depuis, un grand nombre de travaux ont porté sur des lois du logarithme itéré pour la fonction de répartition empirique et d'autres estimateurs de la fonction de répartition. D'autres lois du logarithme itéré relatives à différentes statistiques ont fait l'objet de plusieurs travaux. Citons, sans prétendre à l'exhaustivité, Hall (1981) qui a montré une LIL pour des estimateurs non-paramétriques de la densité, Hardle (1984) qui a montré une LIL pour des estimateurs non-paramétriques de la régression, et Földes et Rejtő (1981a) qui ont montré une LIL pour l'estimateur de Kaplan-Meier de la fonction de survie pour des données censurées à droite, résultat que nous rappelons ci-dessous avec la LIL de Kiefer (1961) puisque nous allons nous en servir. Sous sa forme la plus simple, le théorème peut être exprimé ainsi.

Théorème 2.1 (Loi du logarithme itéré). Soit (X_n) une suite de v.a. i.i.d. centrées de variance 1, et $S_n = \sum_{i=1}^{n} X_i$. Alors :

$$\lim_{n\to\infty} \sup \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} p, s. \tag{2.1}$$

2.2 Loi du logarithme cas des données censurées à droite — Estimateur de Kaplan-Meier

Le résultat suivant est une loi du logarithme itéré pour l'estimateur de Kaplan-Meier de la fonction de répartition, qui est d'une certaine manière similaire au théorème de Chung-Smirnov. Dans le cas de la censure à droite, nous définissons l'estimateur de Kaplan- Meier, que nous notons par $\widehat{F}_n$, en posant
$\widehat{F}_n(z) = 0$ pour $z > Z_{(R)}$. Pour toute variable aléatoire V, on note par I_V et T_V le point
initial et le point terminal du support de la loi de V. En notant par F (resp. G)
la fonction de répartition de la variable d'intérêt X (resp. de la variable de censure C),
nous pouvons énoncer le résultat suivant.

Théorème 2.2 (Földes et Rejtő (1981a)). On suppose que F et G sont continues, et que $T_X < T_C$, alors

$$P\left(\sup_{I<u<\infty} |\widehat{F}_n(u) - F(u)| = O\left(\sqrt{\frac{\log\log n}{n}}\right)\right) = 1 \tag{2.2}$$

La condition $T_X < T_C$ pouvant paraître restrictive, citons le théorème autrement

Théorème 2.3 (Földes et Rejtő (1981a)). Si F et G sont continues, et si le réel T est tel que $G(T) < 1$, alors.

$$P\left(\sup_{I<u<T^*}|\widehat{F}_n(u)-F(u)|=O\left(\sqrt{\frac{\log\log n}{n}}\right)\right)=1 \tag{2.3}$$

où $T^*=\min\{T,T_X\}$.

Remarquons que dans le cas de la censure à gauche, on observe $X\vee C$ et $\delta=1_{\{X\geq C\}}$. Dans ce cas, l'estimateur de la fonction de répartition F, noté $\widetilde{F}_n$, se déduit de celui de Kaplan-Meier en inversant le temps.

$$\widetilde{F}_n(Z)=\prod_{i:Z_i>z}\left(\frac{\widetilde{N}_n(Z_i)-1}{\widetilde{N}_n(Z_i)}\right)^{\delta_i} \tag{2.4}$$

avec $\widetilde{N}_n(x)=\sum_{i=1}^n 1_{\{Z_i\leq x\}}$ Nous avons alors le résultat suivant.

Corollaire 2.2. Si F et G sont continues et si le réel I vérifie $G(I)>0$, alors

$$P\left(\sup_{I^*<u<\infty}|\widehat{F}_n(u)-F(u)|=O\left(\sqrt{\frac{\log\log n}{n}}\right)\right)=1 \tag{2.5}$$

où $I^*=\max\{I,I_X\}$,

Passons maintenant au cas de la censure mixte.

2.3 Loi du logarithme itéré de la fonction de survie en présence d'une censure mixte

Le modèle étudié ici est celui de la censure mixte introduit au chapitre 1, nous utilisons donc les mêmes notations, et nous noterons pour toute variable aléatoire U, F_U (resp. S_U) sa fonction de répartition (resp. de survie).
Nous noterons aussi $T_U = \sup\{t : F_U(t) < 1\}\, et I_U = \inf\{t : F_U(t) \neq 0\}$ les points terminaux du support de F_U, et nous supposons que les fonctions de répartition de X, R et L sont continues. Posons

$$\bar{S}_n(t) = \prod_{J/Z_J \leq t} \{1 - D_{0J}/(U_{J-1} - N_{J-1} + 1)\} \tag{2.6}$$

C'est une modification nécessaire de$\widehat{S}_n$. Soit $H(t) = P(Z < t)$ la fonction
de répartition continue de l'observation Z, sa fonction de répartition empirique
est $H_n(t) = \sum_{i=1}^{n} 1_{\{Z_i<t\}}/n$. La sous-distribution de Z,

$$\widetilde{F}(t) = P(Z \leq t, A = 0) = \int_0^t F_L(u) S_R(u) dF_X(u) \tag{2.7}$$

est la fonction de répartition du vecteur aléatoire à trois dimensions $(X, X - R, L - X)$ au point $(t, 0, 0)$. Sa fonction de répartition empirique est $\widetilde{F}_n(t) = \sum_{i=1}^{n} 1_{\{Z_i \leq t, A_i = 0\}}/n$.
La LIL de Kiefer (théorème 2.2) s'applique à $\widetilde{F}_n$ et donne

$$P\left(\lim_{n\to\infty}\sup\frac{\sup|\widetilde{F}_n(u)-\widetilde{F}(u)}{\sqrt{\frac{\log\log n}{2\pi}}}\leq 1\right)=1 \tag{2.8}$$

En inversant le temps, l'estimateur produit-limite de F_L, qui est continue, est donné par $G_n(u)=\prod_{J/Z_J\geq u}\{1-D_{2J}/N_J\}$. Remarquons que la LIL de, Kiefer (1961) s'applique à, H_n et que de plus le corollaire 2.2 s'applique, à G_n (sous l'hypothèse $\sup(I_R,I_L)<I_X$) pour obtenir que pour presque tout, ω il existe n_1 et un nombre fixé A tel que pour tout $n>n_1$

$$\sup_{I_X\leq u}|G_n(u)-F_L(u)\left(H_n(u)-H(u)\right)|\leq A\sqrt{\frac{\log\log n}{2n}} \tag{2.9}$$

Maintenant en tenant compte du fait que $(F_L(u)-H(u))\geq F_L(I_X)S_R(T_X)S_X(u)$ dès que $I_X\leq u\leq T_X$, nous déduisons sous les hypothèses $I_L<I_X$ et $T_R<T_X$ que pour tout $n>n_1$,

$$G_n(u)-H_n(u)\geq(F_L(u)-H(u))=2p.s.; \tag{2.10}$$

pour tout $I_X\leq u\leq u_n$ avec

$$U_n=S_X^{-1}\left(\frac{2A}{F_L(I_X)S_R(T_X)}\sqrt{\frac{\log\log n}{2n}}\right) \tag{2.11}$$

où $S_X^{-1}(s)=\sup\{x/S_X(x)>s\}$.
La mesure de hasard associée à X est $d\Lambda(t)=dF_X(t)/S_X(t)$ qui peut être écrite $d\widetilde{F}(t)/\left(F_L(t)-H(t)\right)$ pour tout t tel que $I_X\leq t<T_X$. Pour $I_X\leq u<T_X$, on pose

$$T(u)=\int_{I_X}^{u}d\Lambda(t)=-\log(S_X(u)),T_n(u)=\int_{I_X}^{u}d\widetilde{F}_n(t)/\left(G_n(t)-H_n(t)\right) \tag{2.12}$$

où $T_n(u)$ est obtenue en remplaçant $\widetilde{F}$, F_L et H par leurs estimateurs dans
l'expression de $T(u)$. Le théorème suivant donne la loi du logarithme itéré pour $\widehat{S}_n$, estimateur de Patilea et Rolin (2006).

Théorème 2.4 (Messaci et Nemouchi (2011, 2013)). Si S_X, S_R et S_L sont des
fonctions de survies continues, et si $\sup(I_L, I_R) < I_X et T_X < T_R$. Alors

$$P\left(\sup_{-\infty<u<\infty} |\widehat{S}_n(u) - S(u)| = 0\left(\sqrt{\frac{\log\log n}{n}}\right)\right) = 1 \tag{2.13}$$

Remarquons que l'hypothèse $\sup(I_L, I_R) < I_X$ et $T_X < T_R$ assure l'iden tifiabilité du modèle étudié (cf. Patilea et Rolin, 2006).
La preuve du théorème est basée sur la décomposition suivante

$$\begin{array}{c} |\widehat{S}_n(u) - S(u)| \leq |\widehat{S}_n(u) - \bar{S}_n(u)| \\ +|\bar{S}_n(u) - S_X(u)| \end{array} \tag{2.14}$$

et nous avons

$$\begin{array}{c} \bar{S}_n(u) - S_X(u) = \left(\exp\log\bar{S}_n(u) - \exp\left(-T_n(u)\right)\right) \\ +\left(\exp\left(-T_n(u)\right) - \exp\left(-T(u)\right)\right) \end{array} \tag{2.15}$$

En appliquant le développement de Taylor aux deux derniers termes de l'expression précédente, nous obtenons

$$\begin{array}{c} \bar{S}_n(u) - S_X(u) = \exp\left(-\theta_n(u)\right)\left(\log\bar{S}_n(u) + T_n(u)\right) \\ +S_X(u)\left(T(u) - T_n(u)\right) \\ +\frac{1}{2}\exp\left(-\theta_n(u)\right)\left(T(u) - T_n(u)\right)^2 \end{array} \tag{2.16}$$

où

$$\min\left\{-\log \bar{S}_n(u), T_n(u)\right\} \leq \theta_n(u) \leq \max\left\{-\log \bar{S}_n(u), T_n(u)\right\} \tag{2.17}$$

et

$$\min\left\{T(u), T_n(u)\right\} \leq \theta_n^{'}(u) \leq \max\left\{T(u), T_n(u)\right\} \tag{2.18}$$

Nous allons maintenant énoncer et démontrer les quatre lemmes suivants. Le lemme 2.1 nous fournit un outil pour démontrer les lemmes 2.2, 2.3 et 2.4, nous permettant de traiter le premier terme du membre de droite de (2.6), et les membres de droite (2.7) et (2.8) respectivement.

Lemme 2.1. Pour presque tout ω, il existe $n_0(\omega)$ tel que si $n > n_0$, alors pour
tout $I_X \leq u \leq u_n, k_1 > 0$ et $k_2 \geq 0$ où $k = k_1 + k_2 > 1$, nous avons.

$$\int_{I_X}^{u} \frac{d\widetilde{F}_n(t)}{(G_n(t)-H_n(t))^{k_1}(F_L(t)-H(t))^{k_2}} = 0\left(\frac{n}{\log\log n}\right)^{\frac{k-1}{2}} \tag{2.19}$$

Démonstration. Nous avons par (2.4), pour tout $n > n_1$, pour tout $I_X \leq u \leq u_n$ et tout $I_X \leq t \leq u$

$$(G_n(t) - H_n(t))^{k_1} \geq \left(\frac{F_L(t)-H(t)}{2}\right)^{k_1} \text{ p.s,}$$

c'est à dire,

$$\begin{gathered}\frac{1}{(G_n(t)-H_n(t))^{k_1}} \cdot \frac{1}{(F_L(t)-H(t))^{k_2}} \leq \\ \frac{2^{k_1}}{(F_L(t)-H(t))^{k_1}} \cdot \frac{1}{(F_L(t)-H(t))^{k_2}} = \\ \frac{2^{k_1}}{(F_L(t)-H(t))^{k_1+k_2}}\end{gathered} \tag{2.20}$$

Posons $k = k_1 + k_2$, nous obtenons alors

$$\int_{I_X}^{u} \frac{2^{k_1} d\widetilde{F}_n(t)}{(F_L(t)-H(t))^k} = \int_{I_X}^{u} \frac{2^{k_1} d\widetilde{F}_n(t)}{(F_L(t)-H(t))^k} + \int_{I_X}^{u} \frac{2^{k_1} d\left(\widetilde{F}_n(t)-\widetilde{F}(t)\right)}{(F_L(t)-H(t))^k} \tag{2.21}$$

Étudions chacun des deux termes du membre de droite de l'expression précédente.

i) Rappelons que_

$$d\widetilde{F}(t) = -F_L(t)S_R(t)dS_X(t) \tag{2.22}$$

De plus, un calcul élémentaire montre que

$$F_L(t) - H(t) = F_L(t)S_R(t)S_X(t) \tag{2.23}$$

Nous pouvons donc majorer le premier terme comme suit

$$\begin{aligned} \int_{I_X}^{u} \frac{-2^{k_1} d\widetilde{F}_n(t)}{(F_L(u)-H(u))^k} &= \int_{I_X}^{u} \frac{2^{k_1} F_n(t)S_R(t)d(S_X(t))}{(F_L(t)S_R(t)S_X(t))^k} \\ &\leq -\frac{2^{k_1}}{F_L^{k-1}(I_X)S_R^{k-1}(T_X)} \int_{I_X}^{u} \frac{d(S_X(t))}{S_X^k(t)} \\ &\leq \frac{2^{k_1}}{F_L^{k-1}(I_X)S_R^{k-1}(T_X)S_X^{k-1}(u)} \end{aligned} \tag{2.24}$$

ii) Quant au second terme, nous avons

$$\begin{aligned} \int_{I_X}^{u} \frac{2^{k_1} d\left(\widetilde{F}_n(t)-\widetilde{F}(t)\right)}{(F_L(t)-H(t))^k} &\leq \frac{2^{k_1}}{F_L^k(I_X)S_R^k(T_X)S_X^k(u)} \int_{I_X}^{u} |d\left(\widetilde{F}_n(t) - \widetilde{F}(t)\right) \\ &\leq \frac{2^{k_1+1}}{F_L^k(I_X)S_R^k(T_X)S_X^k(u)} \sup_{t\in R} |\widetilde{F}_n(t) - \widetilde{F}(t)| \end{aligned} \tag{2.25}$$

L'application de (2.5) montre que

$$S_X(u_n) \geq \frac{2A}{F_L(I_X)S_R(T_X)} \sqrt{\frac{\log\log n}{2n}} \tag{2.26}$$

Puisque $u \leq u_n$, tenant compte de (2.2) et regroupant les deux termes, il vient

$$\begin{aligned}
&\int_{I_X}^{u} \frac{d\widetilde{F}_n(t)}{(G_n(t)-H_n(t))^{k_1}(F_L(t)-H(t))^{k_2}} \leq \\
&\frac{2^{k_1}}{F_L^{k-1}(I_X)S_R^{k-1}(T_X)S_X^{k-1}(u)} \times \left(\frac{2}{A}+\frac{1}{k-1}\right) \\
&\leq \frac{2^{k_1}F_L^{k-1}(I_X)S_R^{k-1}(T_X)}{F_L^{k-1}(I_X)S_R^{k-1}(T_X)2^{k-1}A^{k-1}} \times \left(\frac{2n}{\log\log n}\right)^{\frac{k-1}{2}} \left(\frac{2}{A}+\frac{1}{k-1}\right) \\
&= 0\left(\frac{2n}{\log\log n}\right)^{\frac{k-1}{2}},
\end{aligned} \tag{2.27}$$

en tenant compte encore une fois de la relation (2.12).
Nous obtenons bien le résultat annoncé dans le lemme.

Lemme 2.2. Nous avons

$$\sup_{I_X \leq u \leq U_n} |\widetilde{S}_n(u) - \bar{S}_n(u)| = 0\left(\sqrt{\frac{1}{n\log\log n}}\right) p.s. \tag{2.28}$$

Démonstration. Rappelons l'inégalité suivante, dont nous allons nous servir pour majorer$|\widehat{S}_n(u) - \bar{S}_n(u)|$. Si pour tout $1 \leq i \leq n$, $|a_i| \leq 1$ et $|b_i| \leq 1$ alors

$$|\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i| \leq \sum_{i=1}^{n} |a_i - b_i|, \tag{2.29}$$

Nous avons donc

$$\begin{aligned}
|\widetilde{S}_n(u) - S_n(u)| &= |\prod_{j/z_j \leq u}\left\{\frac{1-D_{0j}}{U_{j-1}-N_{j-1}}\right\} - \prod_{j/z_j \leq u}\left\{\frac{1-D_{0j}}{U_{j-1}-N_{j-1}+1}\right\}| \\
&\leq \sum_{j/z_j \leq u} \frac{1-D_{0j}}{(U_{j-1}-N_{j-1})^2} = \sum_{j/z_j \leq u} \frac{nd\widetilde{F}_n(z_j)}{(nG_n(z_j)-nH_n(z_j))^2} \\
&= \int_{I_X}^{u} \frac{nd\widetilde{F}_n(t)}{(nG_n(t)-nH_n(t))^2} \\
&= 0\left(\sqrt{\frac{1}{(n\log\log n)}}\right) p.s.
\end{aligned} \tag{2.30}$$

en appliquant le lemme 2.1 pour $k_1 = 2$ et $k_2 = 0$.

Lemme 2.3. Nous avons

$$\sup_{I_X \leq u \leq U_n} S_X(u)|T_n(u) - T(u)| = 0\left(\sqrt{\frac{\log\log n}{n}}\right) p.s. \tag{2.31}$$

Démonstration. Remarquons que,

$$\begin{gathered} |T_n(u) - T(u)| \leq |\int_{I_X}^{u} \frac{(G_n(t)-H_n(t))-(F_L(t)-H(t))}{(G_n(t)-H_n(t))(F_L(t)-H(t))} d\widetilde{F}_n(t)| + \\ |\int_{I_X}^{u} \frac{d(\widetilde{F}_n(t))-\widetilde{F}(t)}{F_L(t)-H(t)}| \\ \leq \sup_{I_X} \left|(G_n(t) - H_n(t)) - (F_L(t) - H(t)) \int_{I_X}^{u} \frac{d\widetilde{F}_n(t)}{(G_n(t)-H_n(t))(F_L(t)-H(t))}\right| \\ + \frac{2\sup|\widetilde{F}_n(t)-\widetilde{F}(t)|}{F_L(I_X)S_R(T_X)S_X(U)} \end{gathered} \tag{2.32}$$

En vertu de (2.3) et (2.13), il s'ensuit que pour n assez grand

$$|T_n(u) - T(u)| \leq 2\sqrt{\log\log n/2n}\frac{2A\left(\frac{2}{A}+1\right) + (1+\epsilon)}{F_L(T_X)S_R(T_X)S_X(U)} \tag{2.33}$$

Compte tenu de (2.5), nous voyons qu'il existe une constante K, telle que

$$\sup_{I_X \leq u \leq u_n} |T_n(u) - T(u)| \leq K p.s. \tag{2.34}$$

En revenant encore à (2.14), nous déduisons que

$$\sup_{I_X \leq u \leq u_n} S_X(u)|T_n(u) - T(u)| = 0\left(\sqrt{\frac{\log\log n}{n}}\right) p.s. \tag{2.35}$$

Nous sommes maintenant en mesure de donner la démonstration du Théorème 2.3.

Démonstration du Théorème 2.3. En vertu de (2.11), nous voyons que

$$\begin{array}{c} \frac{1}{2}\exp\left(-\theta_n^{'}(u)\right)|T_n(u)-T(u)|^2 \\ \leq \frac{1}{2}S_X(u)|T_n(u)-T(u)|^2\exp\left(|T_n(u)-T(u)|\right) \\ \leq \frac{K}{2}S_X(u)|T_n(u)-T(u)|\exp(K) \end{array} \tag{2.36}$$

L'application immédiate du lemme 2.4 donne alors

$$\frac{1}{2}\exp\left(-\theta_n^{'}(u)\right)|T_n(u)-T(u)|^2 = O\left(\sqrt{(\log\log n)/n}\right) \tag{2.37}$$

Par ailleurs, tenant compte de (2.10), il vient

$$\exp\left(-\theta_n(u)\right)|\log\bar{S}_n(u)+T_n(u)| \leq |\log\bar{S}_n(u)+T_n(u)|. \tag{2.38}$$

Combinant les relations (2.9) et (2.15) avec les lemmes 2.2 et 2.3, nous pouvons conclure que, pour $I_X \leq u \leq u_n$

$$|S_n(u)-S_X(u)| = 0\left(\sqrt{(\log\log n)/n}\right) p.s. \tag{2.39}$$

Cette dernière relation, combinée avec l'inégalité

$$|\widehat{S}_n(u)-S_X(u)| \leq |\widehat{S}_n(u)-\bar{S}_n(u)| + |S_X(u)-\bar{S}_n(u)|, \tag{2.40}$$

montre que

$$\sup_{I_X\leq u\leq u_n}|\widehat{S}_n(u)-S_X(u)| = 0\left(\sqrt{(\log\log n)/n}\right) \tag{2.41}$$

pour n suffisamment grand.
La preuve du théorème est maintenant immédiate en combinant le dernier résultat avec la relation suivante

$$\sup_{u_n \leq u \leq +\infty} |\widehat{S}_n(u) - S_X(u)| \leq |S_X(u_n)| + |\widehat{S}_n(u_n) - S_X(u_n)|. \tag{2.42}$$

Nous terminons ce chapitre en donnant un taux de convergence presque complète de $\widehat{S}_n$ (se reporter à l'appendice pour un rappel sur la convergence presque complète et le taux associé).
Sous les mêmes conditions que le théorème 2.3, nous avons

$$\sup_{-\infty \leq u \leq +\infty} \left| \widehat{S}_n(u) - S_X(u) \right| = O\left(\sqrt{(\log n)/n} \right) \tag{2.43}$$

Démonstration. Rappelons que $F_n = 1 - \widehat{S}_n$ et $\Lambda(t) = \int_0^t \frac{df_x(u)}{S_X(u)}$.
Pour tout t tel que $I_L < t < T_R$,$\Lambda(t)$ peut s'écrire

$$\Lambda(t) = \int_0^t \frac{d\widetilde{f}(u)}{F_L(u) - H(u)}. \tag{2.44}$$

et peut donc s'estimer par

$$\Lambda_n(t) = \int_0^t \frac{d\widetilde{f}_n(u)}{G_n(u) - H_n(u)}. \tag{2.45}$$

L'utilisation de l'équation de Duhamel permet alors d'écrire pour tout $t \leq \theta < \min(T_R, T_X)$

$$|F_n(t) - F_X(t)| = (1 - F_X(t)) \left| \int_0^t \frac{1 - F_n(u^-)}{1 - F_X(u)} d(\Lambda_n - \Lambda)(u) \right|. \tag{2.46}$$

Posons $M_n(t) = \int_{I_X}^t d(\Lambda_n - \Lambda)(u)$ et intégrons par parties, nous obtenons

$$\begin{aligned}
|F_n(t) - F_X(t)| &\leq |\int_{I_x}^t \frac{1-F_n(u^-)}{1-F_X(u^-)} dM_n(u)| \\
&\leq |\frac{1-F_n(t)}{1-F_X(t)} M_n(t) - \int_{I_X}^t M_n(u) d\left(\frac{1-F_n(u)}{1-F_X(u)}\right)| \\
&\leq \frac{1}{1-F_X(\theta)} |M_n(t)| + |\int_{I_X}^t Mn(u) \frac{dF_n(u)}{1-F_X(u)} \\
&+ |\int_{I_X}^t M_n(u) \left(1 - F_n(u^-)\right) \frac{dF_X(u)}{(1-F_X(u))(1-F_X(u))}| \\
&\leq \frac{2(1-F_X(\theta))+1}{(1-F_X(\theta))^2} \sup_{I_X \leq u \leq \theta} |M_n(u)|.
\end{aligned} \tag{2.47}$$

Il reste donc à montrer que

$$\begin{aligned}
&\sup_{I_X \leq u \leq \theta} |\Lambda_n(t) - \Lambda(t)| = 0_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \\
|\Lambda_n(t) - \Lambda(t)| &\leq \int_{I_X}^t |\frac{1}{G_n(u) - H_n(u)} - \frac{1}{dF_L(u) - H(u)}| d(u) \\
&+ |\int_{I_X}^t \frac{1}{F_L(u) - H(u)} d(\widetilde{F}_n - \widetilde{F})(t)| \\
&=: B_{n,1}(t) + B_{n,2}(t).
\end{aligned} \tag{2.48}$$

Etudions ces deux termes.
– Puisque $F_L(u) - H(u) = F_L(u) S_R(u) S_X(u)$, nous avons

$$\begin{aligned}
B_{n,1}(t) &\leq \frac{\sup_{I_X \leq u \leq \theta} |F_L(u) - G_n(u) + H_n(u) - H(u)|}{F_L(I_X) S_R(\theta) S_X(\theta)} \int_{I_X}^t \frac{d\widetilde{F}_n}{G_n(u) - H_n(u)} \\
&\leq 2 \frac{\sup_{I_X \leq u \leq \theta} |F_L(u) - G_n(u) + H_n(u) - H(u)|}{F_L(I_X) S_R(\theta) S_X(\theta) \inf_{I_X \leq u \leq \theta} |G_n(u) - H_n(u)|}.
\end{aligned} \tag{2.49}$$

Puisque $I_R < I_X$, le théorème 1 de Bitouzé et al. (1999) permet d'avoir

$$\sup_{I_X < t} |G_n(t) - F_L(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{2.50}$$

Par ailleurs, l'application de l'inégalitéDKW(voir Dvoretzky et al. (1956)) donne

$$\sup_{t\in R}|H_n(t)-H(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right) \tag{2.51}$$

De plus, pour $\epsilon_0 \in]0, F_L(I_X)S_R(\theta)S_X(\theta)[$, nous avons

$$P\left(\inf_{I_X\leq\leq\theta}|G_n(u)-H_n(u)|<\epsilon_0\right) \leq P\left(\sup_{I_X\leq u\leq\theta}|F_L(u)-G_n(u)+H_n(u)-H(u)|>\epsilon\right) \tag{2.52}$$

où $\epsilon = F_L(I_X)S_R(\theta)S_X(\theta)-\epsilon_0$. Le terme à droite de l'inégalité est le terme général d'une série convergente. Nous pouvons donc affirmer que.

$$\sup_{I_X\leq t\leq\theta} B_{n,1}(t) = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{2.53}$$

– Intégrons encore par parties, il vient

$$\begin{aligned} B_{n,2}(t) &\leq \left|\tfrac{\widetilde{F}_n(t)-\widetilde{F}(t)}{F_L(t)-H(t)}\right| + \left|\tfrac{\widetilde{F}_n(I_X)-\widetilde{F}(I_X)}{F_L(I_X)-H(I_X)}\right| \\ &\quad + \left|\int_{I_X}^t \widetilde{F}_n(u)-\widetilde{F}(u)d\left(\tfrac{1}{F_L(u)-H(u)}\right)\right| \\ &\leq \tfrac{2}{F_L(I_X)S_R(\theta)S_X(\theta)} \sup_{I_X\leq u\leq\theta}\left|\widetilde{F}_n(u)-\widetilde{F}(u)\right| \\ &\quad + \left|\int_{I_X}^t \tfrac{\widetilde{F}_n(u)-\widetilde{F}(u)}{F_L(u)}d\left(\tfrac{1}{S_R(u)S_X(u)}\right)\right| \\ &\quad + \left|\int_{I_X}^t \tfrac{\widetilde{F}_n(u)-\widetilde{F}(u)}{S_R(u)S_X(u)}d\left(\tfrac{1}{F_L(u)}\right)\right| \\ &\leq D \sup_{I_X\leq u\leq\theta}\left|\widetilde{F}_n(u)-\widetilde{F}(u)\right|, \end{aligned} \tag{2.54}$$

où D est une constante déterministe.
Il reste à appliquer le théorème 1 - m de Kiefer (1961) pour obtenir

$$\sup_{I_X\leq u\leq\theta}\left|\widetilde{F}_n(u)-\widetilde{F}(u)\right| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right) \tag{2.55}$$

Chapter 3

Estimateurs non paramétriques de la fonction de régression

Dans ce chapitre Un estimateur noyau pour la fonction de régression est introduit Lorsque la variable de réponse est soumise à un contrôle mixte, Etant donné une covariable aléatoire X à valeurs dans R^d, une variable réponse Y positive et deux variables de censure R et L positives, nous nous plaçons dans le contexte de la censure mixte exposé au chapitre 1. Plus précisément, Y est censurée et nous disposons seulement d'observations du triplet (X, Z, A) où $Z = \max(\min(Y, R)L)$

$$A = \begin{cases} 0 & , siL < Y < R \\ 1 & , siL < R \leq Y \\ 2 & , si \min(Y, R) \leq L \end{cases} \qquad (3.1)$$

Notre but est de construire des estimateurs $r_n(x)$ de la fonction de régression $r(x) = E(Y\ X = x)$, qui minimisent l'erreur quadratique moyenne,

ce qui revient à faire tendre

$$\int |r(x) - r_n(x)|^2 \mu(dx) \tag{3.2}$$

vers zéro, où μ est la loi de probabilité de X. Il est bien connu que lorsque Y n'est pas censurée, il existe des estimateurs (estimateurs à noyau, à partition, des k plus proches voisins, des moindres carrés, spline de lissage) pour $\int |r(x)-r_n(x)|^2\mu(dx)$ converge en probabilité ou presque sûrement vers zéro. Les estimateurs à noyau, à p artition et des k plus proches voisins font partie d'une classe d'estimateurs appelés estimateurs à poids, sur les quels nous revenons ci-dessous.

3.1 Estimateurs à poids

Une classe bien connue et très utilisée d'estimateurs non-paramétriques est la classe des estimateurs dits à poids qui s'écrivent. $r_n(x) = \sum_{i=1} nw_{n,i}(x) \cdot Y_i$

où les poids $W_{n,i}(x) = W_{n,i}(x, X_1, ..., X_n) \in R$ dépendent de $X_1, \cdots X_n$. Habituellement les poids sont positifs, leur somme ne dépasse pas 1 et $W_{n,i}(x)$ est petit si x est (éloigné) de X_n. L'idée d'estimer la régression de cette façon vient du raisonnement suivant : si plusieurs observations sont disponibles, on peut estimer $E(Y/X = x_0)$ par la moyenne des Y_i pour lesquelles les X_i correspondantes sont "assez proches" de x_0.

3.1.1 Cas des données censurées 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'ètude ou être toujours vivant à la n 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 $(X_1, Z_1, \delta_1) \cdots (X_n, Z_n, \delta_n)$ où $Z_i = Y_i \wedge C_i$ et $\delta_i = I_{Y_i \leq C_i}$ (δ_i est l'indicatrice de censure). Nous nous proposons d'estimer $r_x = E(Y/X = x)$, les estimateurs donnés au chapitre précédent ne peuvent plus être utilisés puisque Y n'est pas toujours observé.

3.1.2 Cas de la censure mixte

Lorsque Y est soumise à une censure mixte, nous estimons pour toute fonction $h : R^d \cdot R \longrightarrow R$, $Eh(X,Y)$ par

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1_{\{A_i=0\}} Z_i}{S_n Z_i F_n Z_i} \tag{3.3}$$

où S_n et F_n sont respectivement les estimateurs de S_R et F_L dont nous rappelons les expressions à la section 3.2.2. Ceci est motivé par le fait que

$$\begin{aligned} E\left(\frac{1_{\{A=0\}} h(X,Z)}{S_R(Z)F_L(Z)}\right) &= E\left(\frac{E\left(\frac{1_{\{A_i=0\}} h(X,Y)}{S_R(Z)F_L(Z)}\right)}{(X,Y)}\right) \\ &= E\left(\frac{h(X,Y)E(1_{A_i=0})}{S_R(Y)F_L(Y)}/(X,Y)\right) \\ &= E\left(h(X,Y)\right) \end{aligned} \tag{3.4}$$

car

$$E\left(1_{\{A_i=0\}}/(X,Y)\right) = S_R(Y)F_L(Y) \tag{3.5}$$

si les hypothèses $H_{1,1}$, $H_{1,2}$, $H_{1,3}$ ci-dessous sont satisfaites. En effet Pour tout B dans (X,Y) (tribu engendrée par le couple (X,Y)) il existe un borélien C tel que $B = (X,Y)^{-1}(C)$. L'indépendance de (X,Y) et (L,R) permet d'ecrire

$$\begin{aligned} \int_B (1_{A=0})dp &= \int_B (1_{L<Y<R})dp \\ &= \int_{c\cdot R_+^2} (1_{l<y<r})dp_{(X,Y,L,R)} \\ &= \int_{c\cdot R_+^2} (1_{l<y<r})dp_{(X,Y)} \otimes dp_{(L,R)} \end{aligned} \tag{3.6}$$

Maintenant par le théorème de Fubini et l'indépendance de R et L, nous obtenons

$$\begin{array}{c} \int_B (1_{A=0}) dp = \int_c (\int_{R_+^2} (1_{l<y<r}) dp_{L,R}) dp_{(X,Y)} \\ = \int_c (\int_{R_+} (1_{l<y}) dp_L \cdot \int_{R_+} (1_{y<r}) dp_R) dp_{(X,Y)} \\ = \int_c (F_L(Y) S_R(Y)) dp_{(X,Y)} \\ = \int_B F_L(Y_1) S_R(Y_1) dp \end{array} \tag{3.7}$$

car F est continue. De plus, $F_L(Y)S_R(Y)$ étant clairement mesurable par rapport à $\sigma(X, Y)$, le résultat en découle.
La suite de ce chapitre est vouée à l'étude des estimateurs à poids que nous définirons pour Y dans ce cadre de censure.

3.2 Hypothèses et estimation

3.2.1 Hypothèses

Comme dans la section 2.3, nous notons pour toute variable aléatoire U, F_u (resp. S_u) sa fonction de répartition (resp. survie, $S_u = 1 - F_u$). $T_u = \sup\{t : F_u(t) < 1\}$ et
$I_u = \inf\{t : F_u(t) = 0\}$ dénotent les points terminaux du support de U.

Soit H_1 l'hypothèse comprenant les cinq conditions suivantes
-$H_{1,1}$: Y,R et L sont indépendantes,
-$H_{1,2}$: (R, L) et (X, Y) sont indépendantes,
-$H_{1,3}$:$\exists\ T < T_R$ et $I > I_L$ tels que
$\forall n \in N$,$\forall i (1 \leq i \leq n)$, $A_i = 0 \implies I \leq Z_i \leq T$ a,s.
- $H_{1,4}$: F_L est continue sur $]0, \infty[$,
-$H_{1,5}$: $T_R \vee T_L < \infty$.
Nous aurons aussi besoin de l'hypothèse d'identifiabilité suivante
-H_2 : $I_Y \leq I_L < I_R$ et $T_R \leq T_Y$.
Remarquons que puisque $I_L < Z_i < T_R$ dès que $A_i = 0$, l'hypothèse $H_{1,3}$ semble ne pas être trop restrictive.

3.2.2 Estimation

Notons $\{W_j, 1 \leq j \leq M\}$ les valeurs distinctes de $\{Z_i, 1 < i \leq n\}$, rangées par ordre croissant. Posons
$D_{kj} = \sum_{i=1}^{n} 1_{\{Z_i = W_j, A_i = k\}}$

et $N_j = \sum_{i=1}^{n} 1_{\{Z_i \leq Wj\}}$

L'estimateur produit-limite de S_R donné dans Patilea et Rolin (2006) est

$$\begin{array}{c} \widehat{S}_n(W_j) = \prod_{i \leq l \leq j} \left\{1 - \frac{D_{1l}}{U_{l-1} - N_{l-1}}\right\} \\ U_{j-1} = n \prod_{j \leq l \leq M} \left\{1 - \frac{D_{2l}}{N_l}\right\} \end{array} \tag{3.8}$$

Patilea et Rolin (2006) ont montré que sous les hypothèses H_2 et $H_{1,1}$

$$\lim_{n \to +\infty} \sup_{t \in R^+} |\widehat{S}_n(t) - S_R(t)| = 0.p.s. \tag{3.9}$$

Par inversion du temps, nous obtenons l'estimateur inversé de celui de Kaplan- Meier, qui est un estimateur de F_L (censure à gauche) et il est donné par

$$\widehat{F}_n(W_j) = \prod_{j<l\leq M} \left\{1 - \frac{1_{\{A_l=2\}}}{l}\right\} \tag{3.10}$$

En adaptant le théorème, de type Glivenko-Cantelli, de Stute et Wang (1993), nous obtenons

$$lim_{n\to+\infty} \sup_{l_L<t} |\widehat{F}_n(t) - F_L(t)| = 0 p.s. \tag{3.11}$$

si F_L est continue ce qui est supposé dans l'hypothèse $H_{1,4}$. Nous pouvons déduire de l'hypothèse $H_{1,3}$ que

$$\widehat{S}_R(T) > 0 \quad et \widehat{F}_L(I) > 0 \tag{3.12}$$

En vertu de (3.2), nous proposons comme estimateurs poids de $r(x)$

$$r_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_n(Z_i)F_n(Z_i)} \tag{3.13}$$

En appliquant le lemme 1.1 à l'estimateur de S_R, on voit que k_0
est le plus petit entier naturel k
tel que $\widehat{S}_n(Z_k) = 0$ si
$D_{1k_0} \neq 0, D_{0k_0} = 0$ et $\forall j D_{1j} = D_{0j} = 0$
ce qui implique que $\widehat{S}_n(Z_i) \neq 0$ dès que $A_i = 0$. Nous pouvons aussi remarquer que
$\widehat{F}_n(Z_i) \neq 0$ dans l'expression de $r_n(x)$ donnée en (3.10) dès que $A_i = 0$.

3.3 Convergence de r_n

Théorème 3.1. Si les poids $W_{n,i}$ sont positifs, si la somme des poids est au plus égale à un et si pour toutes les lois de (X, Y) avec $|Y|$ bornée presque sûrement

$$\lim_{n\to\infty} \int_{R^d} |\sum_{i=1}^n W_{n,i} Y_i - r(x)|^2 \mu(dx) = 0 p.s \tag{3.14}$$

alors sous les hypothèses H_1 et H_2, les estimateurs r_n satisfont à

$$\lim_{n\to\infty} \int_{R^d} |\sum_{i=1}^n r_n(x) - r(x)|^2 \mu(dx) = 0 p.s \tag{3.15}$$

Démonstration. Introduisons les quantités

$$\begin{array}{c} \widehat{r}_n(x) = \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i) F_n(Z_i)}. \\ et \\ \widehat{r}_n(x) = \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i) F_L(Z_i)} \end{array} \tag{3.16}$$

Puisque

$$\begin{array}{c} \int_{R^d} |r_n(x) - r(x)|^2 \mu(dx) \leq 2 \int_{R^d} |r_n(x) - \widehat{r}_n(x)|^2 \mu(dx) \\ +4 \int_{R^d} |\widehat{r}_n(x) - \bar{r}_n(x)|^2 \mu(dx) \\ +4 \int_{R^d} |\bar{r}_n(x) - r_n(x)|^2 \mu(dx) \end{array} \tag{3.17}$$

la preuve est divisée en trois étapes.
Dans la première étape nous montrons

$$\int_{R^d} |r_n(x) - \widehat{r}(x)|^2 \mu(dx) \to 0 p.s. \tag{3.18}$$

Utilisant (3.8), (3.9), l'hypothèse H_2 et le fait que les poids soient positifs et que leur somme soit au plus égale à un, nous trouvons pour n suffisamment grand

$$\begin{aligned} \mid \bar{r}_n(x) - \widehat{r}_n(x) &\leq \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} Z_i \frac{\widehat{S}_n(Z_i)}{S_R(Z_i)\widehat{S}_n(Z_i)\widehat{F}_n(Z_i)} \\ &\leq T_L \vee T_R \frac{\sup_{t\in R} |S_n(t) - \widehat{S}_R(t)|}{S_R(T)\widehat{S}_n(T)F_n(I)} \end{aligned} \tag{3.19}$$

Donc (3.12) a lieu en vertu de (3.5) et de l'hypothèse $H_{1,5}$.
Dans la deuxième étape nous montrons que

$$\int_{R^d} |\bar{r}_n(x) - \widehat{r}_n(x)|^2 \mu(dx) \to 0 p.s. \tag{3.20}$$

dont la preuve se fait de la même manière que la précédente mais en utilisant
(3.7) au lieu de (3.5).
Finalement, dans la dernière étape il reste à prouver que

$$\int_{R^d} |\bar{r}_n(x) - r_n(x)|^2 \mu(dx) \to 0 p.s. \tag{3.21}$$

Nous pouvons déduire des hypothèses $H_{1,3}$ et H_2

$$\begin{aligned} 0 &\leq 1_{\{A_1=0\}} \frac{Z_1}{S_R(Z_1)F_L(Z_1)} \\ &\leq \frac{T_L \vee T_R}{S_R(T)F_L(I)} p.s. \end{aligned} \tag{3.22}$$

D'autre part, la relation (3.3) nous permet d'écrire

$$\begin{aligned} & E\left(\frac{1_{\{A_1=0\}}}{S_R(Z_1)F_L(Z_1)} / X_1\right) \\ &= E\left[\frac{Y_1}{S_R(Y_1)F_L(Y_1)} E(1_{\{A_1=0\}}) / (X_1.Y_1) / X_1)\right] \\ &= r_{(X_1)} \end{aligned} \tag{3.23}$$

Nous pouvons donc appliquer l'hypothèses (3.11) pour aboutir à (3.13).

3.4 Application aux différents estimateurs à poids

Si dans l'expression de r_n

$$a)\mathrm{W}_{n,i}(x)=\begin{cases} \frac{1}{k_n} & ,si\mathrm{X}_i \text{ est parmi les } k_n \text{ plus proches voisin de } x, \\ 0 & ,sinon\mathrm{k}_n \end{cases} \quad (3.24)$$

est un paramètre de l'estimation. Nous obtenons l'estimateur des k_n plus proches voisins noté par $r_{n,1}$. Le théorème 3.1 et l'article de Devroye et al. (1994), nous permettent d'énoncer le résultat suivant.

Corollaire 3.1. Si $k_n \to \infty, \frac{k_n}{n} \to 0$ et si les hypothèses H_1 et H_2 sont satisfaites alors

$$\lim_{n\to\infty} \int_{R^d} |r_{n,1}(x) - r(x)|^2 \, \mu(dx) = 0 p.s. \quad (3.24)$$

si $||X - x||$ est absolument continue pour tout $x \in R^d$.

b) $W_{n,i}(x) = \sum_J \frac{1_{\left\{A_{n,J}\right\}}(X_i)}{\sum_{k=1}^n 1_{\left\{A_{n,J}\right\}}(X_k)} 1_{\left\{A_{n,J}\right\}}(x)$, où $(A_{n,j})_j$ est une partition de R^d et
1_A représente la fonction indicatrice de l'ensemble A, nous obtenons l'estimateur à partition noté par $r_{n,2}$. D'après le théorème 3.1 et l'article de Devroye et Györfi (1983) nous obtenons le résultat suivant.

3.5 Choix du paramètre de lissage

Par construction l'estimateur à noyau dépend de deux paramètres : Le noyau K et le paramètre de lissage h. Dans la pratique, il faut décider du choix à faire pour ces deux paramètres. Comme d'habitude le noyau n'a pas une grande influence sur l'estimateur, par contre le choix de h_n est essentiel. Tous les résultats de convergence pour lesquels des vitesses de convergence sont précisées mettent en évidence le rôle du paramètre de lissage. Pour la convergence presque complète, nous savons que

$$r_n(x) - r(x) = O(h_n^l) + O_{a.co}\left(\sqrt{\frac{\log n}{nh_n}}\right) \tag{3.24}$$

Théoriquement, pour choisir h, il suffit de minimiser le 2^e membre de (4.6). La vitesse de convergence presque complète est la même que lorsque Y est complétement observée (voir Ferraty et Vieu (2002)). La condition à imposer à h pour atteindre asymptotiquement le minimum est

$$h = C\left(\frac{n}{\log n}\right)^{-\frac{l}{2l+1}} \quad 0 < C < \infty \tag{3.24}$$

Ceci implique que la vitesse optimale est

$$r_n(x) - r(x) = O_{p.co}\left(\left(\frac{n}{\log n}\right)^{-\frac{l}{2l+1}}\right) \tag{3.24}$$

Il en est de même pour la vitesse de convergence presque complète uniforme et nous avons

$$\sup_{x\in S} r_n(x) - r(x) = O_{p.co}\left(\left(\frac{n}{\log n}\right)^{-\frac{l}{2l+1}}\right). \tag{3.24}$$

Notons que la vitesse de convergence presque complète est la même pour les données complètes et pour les données soumises à une censure mixte. En ce qui concerne les données censurées à droite, Guessoum et Ould Saïd (2008) ont obtenu la même vitesse mais pour une convergence presque sûre.

Chapter 4

Vitesse de Convergence presque complète de l'estimateur à noyau de la fonction de régression

Au chapitre 3, nous avons introduit un estimateur de régression par noyau En présence d'un contrôle mixte, la quadrature intégrale s'est avérée erronée Entre cette estimation et la régression, il converge presque certainement vers 0. Dans ce chapitre, nous montrons une convergence presque parfaite La précision et la cohérence de cet estimateur, ainsi que la limite de vitesse La convergence que nous appelons $r_n$est l'estimateur de régression.

$$r_n(x) = \sum_{i=1}^{n} W_{i,n}(x) 1_{\{A_i=0\}} \frac{Z_i}{\widehat{S}_n(Z_i)\widehat{F}_n(Z_i)}$$

où $W_{n,i}$ est donné par

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

4.1 Hypothèses

h_1 : Les fonctions de répartition des variables Y , R et
L sont continues, $I_Y \leq I_L < I_R$ et$T_R < T_Y$.
$h_{2,1}$: r et f sont ' fois continuement dérivables dans un voisinage de x.
$h_{2,2}$: $f(x) > 0$,
$h_{2,3}$: $\lim_{n\to+\infty} h_n = 0$, $\lim_{n\to+\infty} \frac{nh_n}{\log n} = +\infty$.
$h_{2,4}$: K est bornée, intégrable, à support compact et $\int k(t)dt = 1$.

$h_{2,5}$: $\int t^j k(t)dt = 0$,$\forall\ j = 1, \cdots, l-1$ et $0 < \left|\int t^l k(t)\right| < \infty$.
Soit S un sous-ensemble compact de R.
$h'_{2,1}$: r et f sont ' fois continûment dérivables autour de S.
$h'_{2,2}$:$\exists \theta > 0, \inf_{x \in S} f(x) > \theta$,
$h'_{2,6}$:$\exists \beta > 0, \exists C < \infty$,$\forall x \in S$:$|k(x) - k(y)| \leq C\,|x-y|^{\beta}$
Nous terminons la liste des hypothèses par les conditions introduites au chapitre 3 en vue de traiter notre type de censure.
$H_{1,2}$: (R, L) et (X, Y) sont indépendants.
$H_{1,3}$:$\exists\ T < T_R$ et $I > I_L$ tels que $\forall\ n \in N, \forall\ i\,(1 \leq i \leq n)$,$A_i = 0 \implies I \leq Z_i \leq T$ p.s.

4.2 Convergence presque complète ponctuelle

Le théorème suivant donne le taux de convergence presque complète ponctuelle de r_n.

Théorème 4.1. Sous les hypothèses h_1, $h_{2,1}$–$h_{2,5}$, $H_{1,2}$ et $H_{1,3}$, nous avons.

$$r_n(x) - r(x) = O\left(h_n^l\right) + O_{a.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. Posons

$$\begin{aligned}
\widehat{r}_n(x) &= \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i)F_L(Z_i)} \\
&= \frac{\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)\frac{1_{A_i=0}Z_i}{S_R(Z_i)F_L(Z_i)}}{\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \\
&= \frac{\widehat{r}_{n,N}(x)}{f_n(x)}, \\
f_n(x) &= \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right),
\end{aligned}$$

est l'estimateur à noyau de f, puis effectuons la décomposition suivante

$$r_n(x) - r(x) = r_n(x) - \widehat{r}_n(x) + \frac{\widehat{r}_{n,N} - g(x)}{f_n(x)} + \frac{f(x) - f_n(x)}{f_n(x)} r(x)$$

où $g = rf$. La démonstration du théorème est alors une conséquence di recte des lemmes suivants.

Lemme 4.1. Sous les hypothèses h_1, $h_{2,3}$ et $H_{1,3}$, nous avons

$$r_n(x) - \widehat{r}_n(x) = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right)$$

Démonstration. Les définitions de $r_n(x)$, de $\widehat{r}_n(x)$ et l' hypothèse $H_{1,3}$ nous permettent d'écrire

$$\begin{aligned}
&|r_n(x) - \widehat{r}_n(x)| \\
&\leq \left| \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{\widehat{S}_n(Z_i)\widehat{F}_n(Z_i)} - \sum_{i=1}^n W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i)F_L(Z_i)} \right| \\
&\leq T \sum_{i=1}^n W_{n,i}(x) 1_{A_i=0} \\
&\times \left| \frac{S_R(Z_i)F_L(Z_i) - \widehat{S}_n(Z_i)\widehat{F}_n(Z_i) + \widehat{S}_n(Z_i)\widehat{F}_n(Z_i) - \widehat{S}_n(Z_i)\widehat{F}_n(Z_i)}{\widehat{S}_n(T)\widehat{F}_n(I)S_R(T)F_L(I)} \right| \\
&\leq T \frac{\sup_{t>I_L} \left|\widehat{F}_n(t) - F_L(t)\right| + \sup_{t\in R} \left|\widehat{S}_n(t) - S_R(t)\right|}{\widehat{S}_n(T)\widehat{F}_n(I)F_L(I)S_R(T)} \sum_{i=1}^n W_{n,i}(x)
\end{aligned}$$

En tenant compte de h1; nous pouvons appliquer la relation 2.16 et le théorème 1 dans Bitouzé et al. (1999) . De plus $\sum_{i=1}^n W_{n,i}(x) = 1$, nous pouvons

$$r_n(x) - \widehat{r}_n(x) = O_{p.co}\left(\sqrt{\frac{\log n}{n}}\right).$$

De plus, puisque $\lim_{n\to\infty} = 0$ (hypothèse $h_{2,3}$), nous déduisons que donc conclure que

$$r_n(x) - \widehat{r}_n(x) = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Lemme 4.2. Sous les hypothèses $h_{2,1}$, $h_{2,4}$, $h_{2,5}$ et $H_{1,2}$, nous obtenons

$$E\widehat{r}_{n,N}(x) - g(x) = O\left(h_n^l\right)$$

Démonstration. Puisque (X_i, Z_i, A_i) sont i.i.d., nous avons

$$\begin{aligned}
&E\widehat{r}_{N,n}(x) - g(x) \\
&= E\left[\frac{1}{h_n} K\left(\frac{x - X_1}{h_n}\right) E\left(\frac{1_{\{A_1=0\}} Y_1}{F_L(Y_1)S_R(Y_1)} / X_1\right) - g(x)\right] \\
&= \int \frac{1}{h_n} K\left(\frac{x-u}{h_n}\right) E\left\{\frac{1_{\{A_1=0\}} Y_1}{F_L(Y_1)S_R(Y_1)} / X_1 = u\right\} f(u)du - g(x).
\end{aligned}$$

L'utilisation de la relation (3.3) permet d'écrire

$$\begin{aligned}
E\left[\frac{1_{\{A_1=0\}} Y_1}{F_L(Y_1)S_R(Y_1)} / X_1\right] &= E\left[\frac{Y_1}{F_L(Y_1)S_R(Y_1)} E(1_{\{A_1=0\}} / (X_1, Y_1)) / X_1\right] \\
&= r(X_1)
\end{aligned}$$

Donc, et puisque $\int K(t)dt = 1$, nous obtenons

$$E\widehat{r}_{n,N}(x) - g(x) = \int \frac{1}{h_n} K\left(\frac{x-u}{h_n}\right) g(u)du - g(x)$$
$$= \int (g(x - zh_n) - g(x))K(z)dz.$$

Les hypothèses $h_{2,1}$, $h_{2,4}$ et $h_{2,5}$ et le développement de Taylor au voisinage de x permettent d'écrire.

$$E\widehat{r}_{n,N}(x) - g(x) = (-1)^l h_n^l \int z^l K(z) \frac{g^l(x)dz}{l} + o(h^l).$$

4.3 Convergence presque complète uniforme

Théorème 4.2. Sous les hypothèses h_1, $h_{2,3} - h_{2,5}$, $H_{1,2}$,$H_{1,3}$,$h'_{2,1}$, $h'_{2,2}$ et$h'_{2,6}$ nous avons

$$\sup_{x \in S} |r_n(x) - r(x)| = O(h_n^l) + O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. Nous suivons la même démarche que pour la démonstration du théorème 4.1. L'utilisation de la relation (4.2), la décomposition suivants permettent de déduire directement le résultat du théorème.

Lemme 4.3. Sous les hypothèses h_1, $h_{2,3}$ et $H_{1,3}$, nous avons

$$\sup_{x\in S}|r_n(x)-\widehat{r}_n(x)| = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. La démonstration est la même que celle du lemme 4.1

Lemme 4.4. Sous les hypothèses $h_{2,4}$,$h_{2,5}$, $H_{1,2}$ et$h_{2,1}^{'}$, nous avons

$$\sup_{x\in S}|E\widehat{r}_{n,N}(x)-g(x)| = O(h_n^l)$$

Démonstration. En tenant compte de l'hypothèse $h_{2,1}^{'}$, la démonstration se fait de la même manière que celle du lemme 4.2.

Lemme 4.5. Sous les hypothèses $h_{2,3}$, $h_{2,4}$, $H_{1,3}$,$h_{2,1}^{'}$ et$h_{2,6}^{'}$, nous avons

$$\sup_{x\in S}|E\widehat{r}_{N,n}(x)-\widehat{r}_{N,n}(x)| = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. Dans cette preuve $C^{'}$ désigne une constante générique. La compacité de S nous permet d'écrire $S\subset U_{k=1}^{z_n}]t_k-l_n,t_k+l_n[$ où l_n et z_n vérifient

$$l_n=(C_{z_n}^{'})^{-1}etz_n\sim(C_n^{'})^{\frac{\beta+1}{\beta}}$$

En posant $t_x=\arg\min_{t\in t_1,t_2,\cdots,t_{z_n}}|x-t|$, nous obtenons

$$\begin{aligned}\sup_{x\in S}|E\widehat{r}_{N,n}(x)-\widehat{r}_{N,n}(x)| \leq & \sup_{x\in S}|\widehat{r}_{N,n}(x)-\widehat{r}_{N,n}(t_x)|\\ &+\sup_{x\in S}|E\widehat{r}_{N,n}(x)-E\widehat{r}_{N,n}(t_x)|\\ &+\sup_{x\in S}|E\widehat{r}_{N,n}(t_x)-\widehat{r}_{N,n}(t_x)|\end{aligned}$$

En utilisant les hypothèses $H_{1,3}$ et $h_{2,6}^{'}$ nous obtenons

$$\sup_{x\in S}|\widehat{r}_{N,n}(x)-\widehat{r}_{N,n}(t_x)|\leq C^{'}\frac{l_n^{\beta}}{h^{\beta+1}n}.$$

et nous pouvons déduire que

$$\sup_{x\in S}|E\widehat{r}_{N,n}(x)-E\widehat{r}_{N,n}(t_x)|\leq C^{'}\frac{l_n^{\beta}}{h^{\beta+1}n}.$$

Maintenant, en vertu de (4.4) et de l'hypothèse $H_{1,3}$ nous pouvons voir que $\frac{l_n^\beta}{h^{\beta+1}n}\sqrt{nh_n} \to 0$, nous en déduisons que pour tout $\epsilon > 0$ et pour n suffisamment grand

$$P\left(\frac{l_n^\beta}{h^{\beta+1}n}\sqrt{\frac{nh_n}{\log n}} > \epsilon\right) = 0.$$

Finalement, l'utilisation de la relation (4.4) et (4.5) permet d'obtenir

$$\begin{aligned} & P\left[\sup_{x\in S}|E\widehat{r}_{N,n}(t_x) - \widehat{r}_{N,n}(t_x)| > \epsilon_0\sqrt{\frac{nh_n}{\log n}}\right] \\ \leq\ & P\left[\max_{k\in 1,\cdots,z_n}|E\widehat{r}_{N,n}(t_k) - \widehat{r}_{N,n}(t_k)| > \epsilon_0\sqrt{\frac{nh_n}{\log n}}\right] \\ & \leq 2z_n \exp^{\frac{-n\epsilon_0^2 h_n}{4C}} \leq (C'n)^{\frac{\beta+1}{\beta}-\frac{\epsilon_0^2}{4c}}. \end{aligned}$$

Le choix d'un ϵ_0 suffisamment grand permet d' avoir

$$\sum_1^n P\left[\sup_{x\in S}|E\widehat{r}_{N,n}(t_x) - \widehat{r}_{N,n}(t_x)| > \epsilon_0\sqrt{\frac{nh_n}{\log n}}\right] < \infty.$$

4.4 Méthode de noyau

L'estimation du noyau n'est pas valable Estimation non paramétrique de la densité d'une variable aléatoire. Cette méthode permet d'obtenir une densité continue et en ce sens forme une généralisation, ou en d'autres termes,

le remplacement de la fonction indicatrice utilisée dans le graphe par une fonction Continu (noyau) pour que la somme des fonctions continues reste continue.

En cas réel

Soit $K : R \to R$ integrable telle que $\int_{-\infty}^{+\infty} k(u)du = 1$. Alors K est appelé noyau. Pour tout $h_n > 0$ petit et$n \in N^*$, on peut définir $x \in R$ par :

$$\begin{aligned}\widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}k\left(\frac{x-X_i}{h_n}\right)\\ &= \frac{1}{nh_n}\sum_{i=1}^{n}k\left(\frac{x-X_i}{h_n}\right).\end{aligned}$$

estimateur à noyau de f:

On a $\int \widehat{f}_n(x)dx = 1$ et si $K > 0$ alors$\widehat{f}_n$ est une densité. Alors on dit que K est le noyau de cet estimateur et Le paramétre $h_n > 0$ est appelé paramétre de lissage (on note$h_n = h$).

Remarque

L'estimateur à noyau est une densité quelle que soient les valeur des observation $X_1, \cdots, X_n$, et on dit symétrique si, pour tout u dans son ensemble de définition, $K(u) = K(-u)$.

On commence par remarque que la densité est la dérivée de la fontion de répartition, ce qui permet d'ecrire pour tout x :

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

Donc pour un $h > 0$, on peut penser á estimer $f(x)$ par :

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{2h}\left(F_n(x+h) - F_n(x-h)\right)\\ &= \frac{1}{2nh}\sum_{i=1}^{n}1_{]x-h,x+h]}(X_i).\end{aligned}$$

4.4.1 L'estimateur de $\widehat{f}$

Pour obtenir quelque chose de plus lisse, on peut remarquer que :

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{2nh}\sum_{i=1}^{n} 1_{]x-h_n,x+h_n]}(X_i) \\ &= \frac{1}{nh_n}\sum_{i=1}^{n} \frac{1}{2}1_{\{x-h_n<X_i\leq x+h_n\}} \\ &= \frac{1}{nh_n}\sum_{i=1}^{n} \frac{1}{2}1_{[-1,1[}\left(\frac{x-X_i}{h_i}\right) \\ &= \frac{1}{nh_n}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h_n}\right).\end{aligned}$$

avec $k = \frac{1}{2}1_{[-1,+1[}(u)$

Chapter 5

Simulation

5.0.1 Estimateur de Kaplan-Meier de la fonction de la survie

Frrireich,en1963,afait un essai thérapeutique ayant pour but de comparer les durées de rémission en semains ,de sujets atteint de leucémie selon qu'ils ont reçu ou non du 6-MP. durée de rémmission,en semaine ,selon le traitement

6-MP	6	6	6	6+	7	9+	10	10+	11+	13	16
		17+	19+	20+	22	23	25+	32+	32+	34+	35+
placebo	1	1	2	2	3	4	4	5	5	8	8
		8	8	11	11	12	12	15	17	22	23

le signe +correspond à des patients quiont quitte l'etude àla date considerée. dans l'analyse de survie on tient comptede toutes les observations censurées ou non,en effet dans les problémes d'estimations statistiques si on élimine les observation censurées du groupe traité par le 6-MP (12patients) on pert de l'information puisque on ne tient pas compte des patients ayant des durées de rémission plus longues. l'estimateur empirique pour le groupe traite par un placebo (pas de censurée) donne le tableau saivant:

Semaine i	Nombre de rémissions àla Semaine i	$\widehat{S}_{placebo}(t)$
0	21	1
1	19	0.0047
2	17	0.0095
3	16	0.76
4	14	0.66
5	12	0.57
8	8	0.38
11	6	0.26
12	4	0.19
15	3	0.14
17	2	0.09
22	1	0.05
23	0	0

l'estimateur de kaplan-Meier de la fonction de survie S de groupe de 21 malades traité par le traitement 6-MP donne le tableau suivant:

Temps t_i	n_i	d_i	$\widehat{S}_{6-MP}(t_i)$
0	21	0	1
6	21	3	(1-3/21)*1-0.857
7	17	1	(1-1/17)*0.857-0.807
10	15	1	(1-1/15)*0.807-0.753
13	12	1	(1-1/12)*0.753-0.690
16	11	1	(1-1/11)*0.690-0.627
22	7	1	(1-1/7)*0.627-0.538
23	6	1	(1-1/6)*0.538-0.448

et si on compare ces chiffres aux valeurs critiques afférentes aux seuil de risque usuels de distribution de KHI-DEUX à un degré de liberté on conclut que lavantage de traitement est significatif.

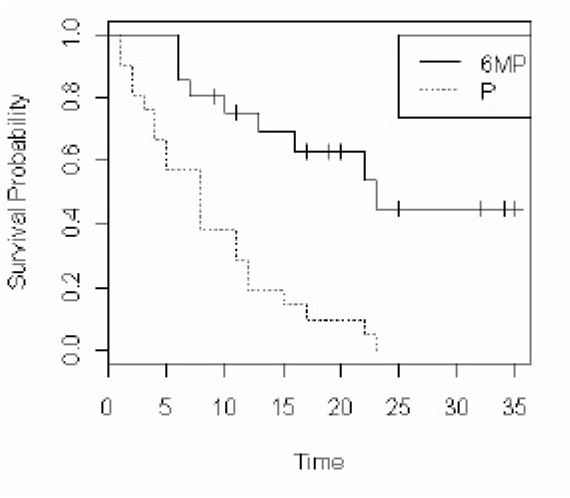

Figure 5.1: Les deux courbes de Survie

Code R:

$X = c(6, 6, 6, 6, 7, 9, 10, 10, 11, 13, 16, 17, 19, 20, 22, 23, 25, 32, 32, 34, 35, 1, 1, 2, 2, 3, 4, 4, 5, 5,$
$C = c(1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$
t=$c(rep("6MP", 21), rep("P", 21))$
f=data.frame(X, C, t)
library(survival)
s=survfit(Surv(X, C) t, data=f)
plot(s, lty=$c(1, 3)$, xlab="Time", ylab="Survival Probability")legend(10, 1.0, $c("6MP", "P")$, lty=$c(1, 3)$)

5.0.2 Estimateur à noyau de la régression

Nous terminons ce mémoire par un travail de simulation de l'estimateur à noyau de la fonction de régression aussi bien pour des données complètes que pour des données censurées à droite ou à gauche et pour diférents modèles.

Nous choisissons le noyau gaussien:

$$k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

et la fenétre donnée par

$$h_n = \left(\log\left(\log(n)\right)/n\right)^{1/5}$$

Modèle non linéaire

Nous traitons le cas suivant

Cas parabolique :$y_i = \frac{3}{2}X_i^2 - \frac{1}{2} + \epsilon_i$.

où X_i etϵ_i sont deux suites de variables i.i.d. de loi $N(0,1)$:

La variable de censure est aussi régie par la loi $N(0;1)$ dans le cas parabolique,

Les résultats de la simulation sont présentés dans les pages suivantes et montrent la bonne performance des estimateurs étudiés aussi bien dans le cas de données complètes que censurées à droite ou à gauche.

La censure ne semble pas influer sur les résultats obtenus. Ceci est conforme aux résultats théoriques qui s'équivalent que les données soient censurées ou pas.

Données complètes

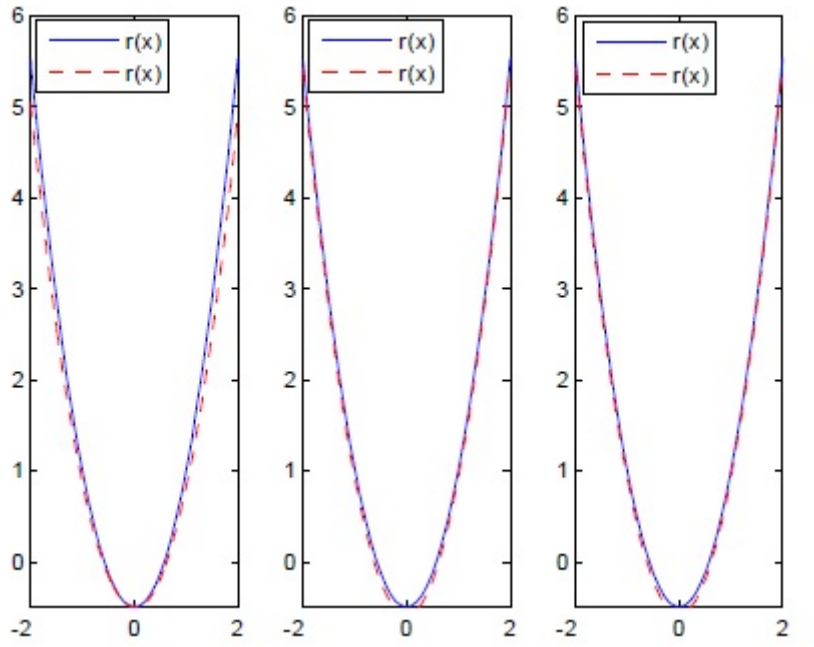

Figure 5.2: Cas parabolique avec n=100, 500, 1000 respectivement.

Données censurées à droite

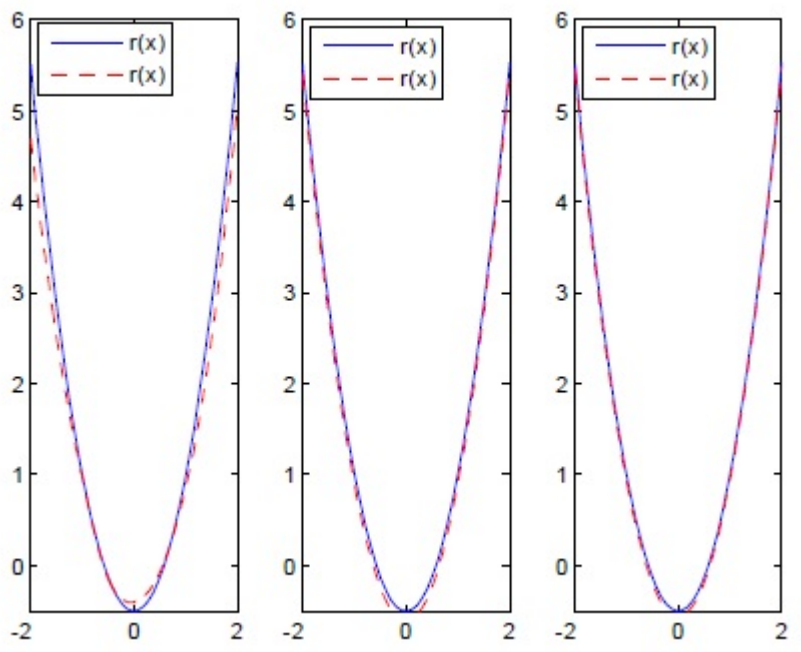

Figure 5.3: Cas parabolique avec n=100, 500, 1000 respectivement.

Données censurées à gauche

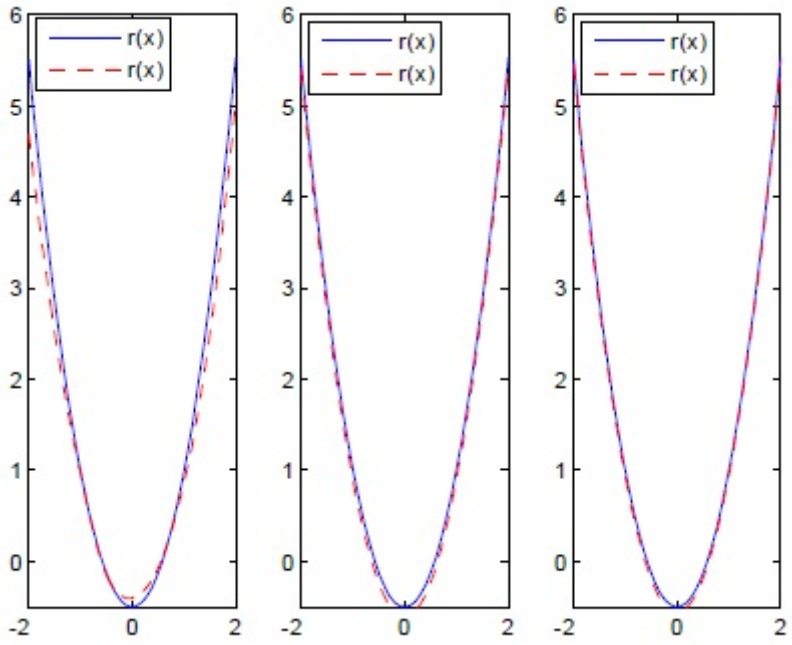

Figure 5.4: Cas parabolique avec n=100, 500, 1000 respectivement.

Code R :

rn(list=ls(all=TRUE))
$n = 100$
$X = rnorm(n, 0, 1)$
$E = rnorm(n, 0, 1)$
$c = rnorm(n, 0, 1)$
$z = \min(x, c)$
$Y = 3/2 * x^2 - 1/2 + E$
$K = function(t)(1/\sqrt{}(2 * pi)) * \exp(-0.5 * t^2)$
$h = (\log(\log(n))/n)^{1/5}$
$s = 100$
$a = \min(X)$
$b = \max(X)$
$x = seq(a, b, length = s)$
$V = numeric(n)$
$fn = numeric(s)$
for$(jin1 : s)$
for$(iin1 : n)V[i] = K((x[j] - X[i])/h)$
$fn[j] = \sum(V)/(n \cdot h)$
$L = numeric(n)$
$Tn = numeric(s)$
for$(jin1 : s)$
for$(iin1 : n)L[i] = S(z[i]) * F(z[i])$
$Tn[j] = z[i]/L[i]$
$W = numeric(n)$
$Hn = numeric(s)$
for$(jin1 : s)$
for$(iin1 : n)W[i] = K((x[j] - X[i])/h) * Tn[j]$
$Hn[j] = \sum(W)/(n * h)$
$Rn = Hn/fn$
op=par$(mfrow = c(1, 3))$
$plot(x, Rn, xlab = "x", ylab = "Rn(x)", main = "n = 50", type = l, col = 4, lwd = 2)$
lines$(x, 3/2 * x^2 - 1/2 + E, lwd = 2)$

Conclusion et Perspectives

Nous nous sommes intéressés à des estimateurs de la fonction de survie, et donnée des vitesses de convergence, dans un contexte de censure mixte et pour des données indépendantes.
Nous avons aussi d'une loi fonctionnelle du logarithme itèrè , qui nous ont permis d'ètablir des lois fortes pour des estimateurs à noyau .
Ceci constitue une extension, au cas des données soumises à la censure mixte, des résultats disponibles pour des données complétes ou censurées à droite.
Une premiére perspective de recherche serait l'extension de la loi fonctionnelle du logarithme itéré au cas uniforme, d'une maniére similaire .
Une question importante concernant l'estimation par la mèthode du noyau est le choix du paramétre de lissage h. Il serait donc intéressant de développer des méthodes qui permettent ce choix à partir des données. Pour justi

er théoriquement ce genre de méthode, il faudrait aussi établir la convergence uniforme par rapport au paramétre de lissage h. Une autre perspective est d'étudier ces mémes propriétés sous divers types de dépendances : α-mélange, φ-mélange association.
Le modéle de censure mixte est intéressant et réaliste lorsque nous le rapportons au domaine de la

abilité. Depuis son introduction par Patilea et Rolin (2006), un certain nombre de travaux le concernant ont vu le jour. Le modèle de censure double est un modèle semblable au modèle de censure mixte abordé dans ce mémoire dans le sens où on observe comme dans le cas de la censure mixte $\max(\min(X, R), L)$ mais où X, R et L ne sont pas indépendantes, mais $(L \leq R)$ presque sûrement. Il serait intéressant d'étudier les résultats acquis et en perspectives sous censure mixte, au cas de la censure double.
Enfin nous pensons que la réalisation de diverses applications sur des données régies par un modèle de censure mixte serait très intéressante.

Bibliographie

A. Földes et L. Rejtő : A LIL type result for the product limit estimator. Probability Theory and Related Fields, 56(1) :75–86, 1981a.
F. FERRATY et P. VIEU : Statistique fonctionnelle : Modèles nonparamétriques de régression. Notes de cours de DEA, Université Paul Sabatier, Toulouse. France, 2002.
F. FERRATY et P. VIEU : Nonparametric functional data analysis : Theory and practice. Springer, 2006.
K. Kebabi, I. Laroussi et F. Messaci : Least squares estimators of the regression function with twice censored data. Statistics Probability Letters, 81 (11) :1588–1593, 2011.
K. Kebabi et F. Messaci : Rate of the almost complete convergence of a kernel regression estimate with twice censored data. Statistics Probability Letters, 82(11) :1908–1913, 2012.
K.kebabi.Estimation non-paramétrique de la fonction de régression : cas d'un modèle de censure mixte (2014).
Z. GUESSOUM et E. OULD SAÏD : Central limit theorem for the kernel estimator of the regression function for censored time series. Journal of Nonparametric Statistics, 24(2):379–397, 2012.
P. HALL : Laws of the iterated logarithm for nonparametric density estimators. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 56 (1):47–61, 1981.
D. BITOUZÉ, B. LAURENT et P. MASSART : A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. Ann. Inst. Henri Poincaré, Probab. Stat., 35(6):735–763, 1999. ISSN 0246-0203.
K.-L. CHUNG : An estimate concerning the Kolmogoroff limit distribution. Transactions of the American Mathematical Society, 67(1):36–50, September 1949.
M. KOHLER, K. MÁTHÉ et M. PINTÉR : Prediction from randomly right censored data. J. Multivariate Anal., 80:73–100, 2002.
G. LUGOSI et K. ZEGER : Nonparametric estimation via empirical risk minimization. IEEE Trans. Inform. Theory, 41:677–687, 1995.

F. MESSACI : Local averaging estimates of the regression function with twice censored data. Statistics Probability Letters, 80(19-20):1508–1511, 2010.
F. MESSACI et N. NEMOUCHI : A law of the iterated logarithm for the product limit estimator with doubly censored data. Statistics Probability Letters, 81(8):1241–1244, 2011.
F. MESSACI et N. NEMOUCHI : Erratum to "a law of the iterated logarithm for the product limit estimator with doubly censored data" [Statist. Probab. Lett. 81 (2011) 1241–1244]. Statistics Probability Letters, 83(9):2142, 2013.
E. A. NADARAYA : Nonparametric estimation of probability densities and regression curves. Springer, 1989.

Résumé

Le but de ce mémoire est d'établir des résultats asymptotiques d'un estimateur à noyau de la fonction de régression mais pour une variable réponse soumise à une censure mixte. Il s'agit de la vitesse de convergence ponctuelle et uniforme et de la normalité asymptotique. Le mode de convergence utilisé est celui de la convergence presque complète. Cette notion de convergence presque complète entraîne la convergence presque sûre.
une loi du logarithme itéré de l'estimateur de Patilea et Rolin (2006) de la fonction de survie. Notons que cet estimateur intervient explicitement dans l'expression de l'estimateur à noyau de la régression qui fait l'objet de notre étude. Nous donnons aussi des illustrations de nos résultats sur des données simulées. Notre cadre de travail est celui de l'estimation non-paramétrique de la régression et des données censurées.

Mots clés : Régression non paramétrique, données censurées, estimateur à noyau, vitesse de convergence.

Abstract

The object of this memory is to establish asymptotic results of an estimator kernel of the regression function but for a response variable subject to censorship mixed. This is the point and uniform speed of convergence and asymptotic normality. The mode of convergence used is that of almost complete convergence. This notion of almost complete convergence results in almost sure convergence. a law of iterated logarithm of the Patilea and Rolin (2006) estimator of the function of survival. Note that this estimator intervenes explicitly in the expression the kernel estimator of the regression that is the subject of our study. We also give illustrations of our results on simulated data. Our framework is that of non-parametric estimation regression and censored data.

Key words : Nonparametric regression, censored data, kernel estimator, rate of convergence.