UNIVERSITE DE BLIDA 1

Faculté des Sciences

THESE DE DOCTORAT

Spécialité : Statistique

INFERENCE STATISTIQUE DANS LES MODELES DE DUREES

AVEC DONNEES MANQUANTES

Par

Amel MEZAOUER

Devant le jury :

M.Blidia	Professeur, Université de Blida1	Président
N.Oukid	Maître de conférences(A), Université de Blida1	Examinateur
G. Saidi	Maître de conférences(A), ENSSEA, Alger	Examinateur
A. Rassoul	Maître de conférences(A), ENSH, Blida	Examinateur
K . Boukhetala	Professeur, U.S.T.H.B, Alger	Directeur de thèse
J.F. Dupuy	Professeur des Universités, INSA de Rennes, France	Co-directeur de thèse

Blida, Avril 2016

Résumé : Dans ce travail, nous considérons en première partie, le problème de l'inférence statistique dans le modèle semi-paramétrique de transformation linéaire, en présence de données manquantes. Nous proposons une méthode d'estimation du paramètre d'intérêt. Nous établissons les propriétés asymptotiques de cet estimateur et nous en étudions les propriétés dans des échantillons de taille finie, au moyen de simulations.
Dans une seconde partie, nous proposons une nouvelle statistique du test du log-rank stratifié avec données manquantes et censure dépendante du groupe de traitement. Nous donnons sa distribution asymptotique sous l'hypothèse nulle d'égalité des groupes de traitement randomisés. Une étude numérique permet d'examiner les propriétés de ce test dans des échantillons finis.

Abstract: In this work, we consider in a first part, the problem of statistical inference in the semi-parametric linear transformation model, with missing data. We propose an estimation method for the parameter of interest. We establish the asymptotic properties of the proposed estimator, and we study its finite sample properties via simulations.
In the second part, we propose a new statistic in stratified log-rank test with missing data and censoring depending on the treatment group. We give its asymptotic distribution under the null hypothesis of equality of the randomized treatment groups. A numerical study is conducted to examine the finite-sample behavior of this test.

ملخص:في هذا العمل، نهتم في الجزء الأول بمشكلة الاستدلال الإحصائي في النموذج التحول الخطي في وجود بيانات ناقصة. نقترح طريقة تقدير العامل.نعطي الخصائص المقاربة للمقدر. ندرس خصائص المقدر في عينات منتهية.
في الجزء الثاني، نقترح احصائية جديدة لفحص log-rank طبقي في وجود بيانات ناقصة ونقص متعلق بفريق العلاج.نعطي التوزيع النهائي تحت الفرضية المنعدمة للتساوي في فرق العلاج العشوائية.نقوم بدراسة رقمية تتيح معرفة خصائص هذه الإحصائية في عينات منتهية.

REMERCIEMENTS

Mes chaleureux remerciements vont tout d'abord à mes directeurs de thèse le Professeur K. Boukhetala et le Professeur J.F. Dupuy pour avoir accepté de m'encadrer au long de ce travail de thèse. Je les remercie pour leur aide et leurs précieux conseils. Sans leurs encouragements et leur soutien, ce travail n'aurait pu aboutir.

Mes remerciements s'adressent ensuite aux membres du jury qui ont accepté d'examiner mon travail de recherche.

Je remercie également le Professeur Blidia de me faire l'honneur de présider le jury.

TABLE DES MATIERES

INTRODUCTION

L'analyse de survie recouvre des méthodes d'analyse statistique de durées jusqu'à l'occurrence d'un événement d'intérêt bien défini. Cet évènement est souvent appelé *décès* qui, selon le domaine d'application, peut représenter : la mort d'un patient, la rechute ou la guérison d'une maladie, la panne d'un équipement industriel, la fin d'une période de chômage, l'expiration d'une ligne téléphonique ou d'un contrat d'assurance…

Depuis l'introduction en 1972 par D. Cox du modèle à risques proportionnels, la littérature consacrée aux modèles de régression de durées a connu un essor remarquable. De nombreux modèles, ont été proposés et leurs différents aspects: inférence, validation, sélection, application en fiabilité et en analyse de survie ont été étudiés en détail, nourrissant une littérature foisonnante. Plusieurs ouvrages dressent un panorama de ce domaine de recherche en perpétuelle évolution. Parmi la très grande variété des modèles qui ont été développés au cours des trente dernières années, la classe des modèles semi-paramétriques de transformation linéaire offre l'avantage d'inclure plusieurs des modèles les plus utilisés dans les applications. Cette classe de modèles a suscité une littérature très abondante depuis vingt ans. Citons par exemple : Chen et al. (1995), Cheng et al.(1995), Fine et al. (1998), Fleming et Lin (2000), Slud et Vonta (2004), Ma et Kosorok (2005), Martinussen et Schieke (2006), Kong et al. (2004,2006), Kosorog et Son (2007), Dupuy (2008).

Notons T la durée aléatoire jusqu'à un instant de défaillance ou de panne et $Z = (Z^1, \dots, Z^p)^\intercal$ un vecteur de dimension p de variables explicatives (où $(.)^\intercal$ désigne la transposée). La classe des modèles semi-paramétriques de transformation linéaire exprime la relation entre T et Z sous la forme :

$$e(T) = -\beta_0^\intercal Z + \varepsilon \qquad (1)$$

où e est une fonction strictement croissante inconnue, $\beta_0 = (\beta^1, \dots, \beta^p)^\intercal$ est un vecteur de paramètres de régression inconnus (paramètres d'intérêt du modèle) et ε désigne un terme d'erreur aléatoire (indépendant de Z) dont la loi de probabilité est supposée connue (on notera F_ε sa fonction de répartition). Si $H(u) = exp(e(u))$ et h désigne la dérivée de H, alors la fonction de risque instantané conditionnelle $\lambda_Z(t) = \lim_{\delta \to 0} \frac{\mathbb{P}(t \leq T < t+\delta | T \geq t, Z)}{\delta}$ de T sachant Z peut s'écrire sous la forme

$$\lambda_Z(t) = \lambda_{e^\varepsilon}(e^{\beta_0^\intercal Z} H(t)) e^{\beta_0^\intercal Z} h(t) \qquad (2)$$

Où λ_{e^ε} désigne la fonction de risque instantané de $exp(\varepsilon)$. On déduit facilement de (2) des cas particuliers remarquables du modèle (1). Ainsi, si ε suit la loi des valeurs extrêmes (c'est-à dire: $F_\varepsilon(u) = 1 - \exp(-\exp(u))$), alors $\exp(\varepsilon)$ suit une loi exponentielle de paramètre 1, d'où $\lambda_{e^\varepsilon} = 1$ et (2) se réduit à $\lambda(t) = e^{-\beta_0^\intercal Z} h(t)$ qui est le modèle à risques proportionnels de Cox de fonction de risque de base instantané h. Si e suit la loi logistique (c'est à-dire : $F_\varepsilon(u) = \exp(u)/(1 + \exp(u))$), alors $\lambda_{e^\varepsilon} = (1+t)^{-1}$ et (2) se ramène à $\lambda(t) = h(t)/\left[H(t) + e^{-\beta_0^\intercal Z}\right]$ qui est le modèle à risques convergents.

Plusieurs méthodes ont été proposées pour estimer le paramètre β_0 dans la classe de modèles (1). Cheng et al. (1995) ont en particulier proposé des équations d'estimation simples à partir d'un échantillon d'observations indépendantes (X_i, Δ_i, Z_i) , $i = 1, \dots, n$ du triplet (X, Δ, Z), où $X = min(T, C)$ désigne la durée observée, C une censure aléatoire, Z un vecteur de variables explicatives, $\Delta = 1(T \leq C)$ et 1(.) désigne la fonction indicatrice. L'estimateur proposé est consistant et asymptotiquement gaussien.

Dans ce travail, nous adaptons ces équations d'estimation à une situation de données manquantes.

Supposons que l'on dispose d'un échantillon de n items. Pour chacun d'entre eux, on observe un vecteur de variables explicatives Z à l'instant $t = 0$ (début de l'étude) et l'on souhaite observer la durée jusqu'à la défaillance. On considère la situation où l'observation de la durée (éventuellement censurée) et de l'indicatrice de censure n'est possible que pour un sous-échantillon aléatoire de l'échantillon initial. Cette situation se rencontre en particulier en fiabilité lorsque des contraintes techniques inattendues viennent limiter les possibilités de recueil des données ou interrompre une partie d'un essai en cours. On dispose alors d'observations du triplet (X, Δ, Z) sur un sous-ensemble des n items tandis

que pour les autres items, on ne dispose que des observations de Z. Dans ce contexte de données manquantes, une solution simple pour estimer β_0 consiste à mener une analyse en "cas complets" ("CC" par la suite) c'est-à-dire à : i) retirer de l'échantillon les items i pour lesquels l'information (X_i, Δ_i) est manquante, ii) calculer l'estimateur de Cheng et al. (1995) sur les items restants. Cette solution entraîne néanmoins une perte d'information et n'est donc pas satisfaisante. Nous proposons donc dans ce travail une alternative basée sur le principe de la "pondération par probabilité inverse" ("Inverse Weighted Probability (IWP)".

Nous utilisons également ce principe dans un test du log-rank stratifié avec données manquantes et censure dépendante du groupe.

Le test du log-rank est un test non paramétrique qui est souvent utilisé pour comparer des groupes de traitements randomisés en présence de durées censurées. Le test du log-rank est un test non-paramétrique, il permet de tester l'hypothèse nulle H_0 d'égalité des fonctions de risque instantané dans les différents groupes. L'idée est de comparer l'estimateur de Nelson-Aalen du groupe spécifié à celui commun à tous les groupes et calculé sous H_0. Le test se généralise au cas de données stratifiées. Considérons un essai clinique où *n* patients sont randomisés dans *K* groupes de traitement. On souhaite comparer les distributions de survie de ces groupes tout en ajustant un facteur *S* à *L* modalités, appelées strates. On note $\lambda_{k,l}$ la fonction de risque instantané d'un patient du groupe k, appartenant à la strate l. Les hypothèses à tester peuvent être formulées ainsi :

$$H_0: \lambda_{1l} = \cdots = \lambda_{Kl}, \text{ pour tout } l = 1, \ldots, L$$

et H_a: il existe j et j' tels que $\lambda_{jl} \neq \lambda_{j'l}$ pour au moins un l

Lorsque la variable *S* n'est pas observée pour tous les individus de l'échantillon, une solution consiste à utiliser l'analyse en cas complet décrite ci-dessus.

Dans plusieurs applications, la censure peut dépendre du groupe de traitement.

Dans ce manuscrit, nous proposons une version modifiée du test du log-rank stratifié dans le cas de données manquantes et censure dépendante du groupe de traitement en combinant :

a) La régression par calibration qui consiste à remplacer chaque indicatrice d'appartenance aux strates $1(S = l)$ non observée, par son espérance conditionnelle sachant des covariables auxiliaires et,
b) le principe de IPCW (Inverse Probability of Censoring Weighted) qui consiste à pondérer chaque individu par l'inverse de la fonction de survie de la censure étant donné le groupe de l'individu.

Dans le premier chapitre de ce manuscrit, nous introduisons la notion de durée de survie et présentons les définitions des outils utilisés dans l'analyse des durées de vie.

Dans le chapitre 2, nous construisons des équations d'estimation adaptées au problème de données manquantes dans un modèle de transformation linéaire. Puis nous montrons la consistance de l'estimateur obtenu. Nous évaluons ensuite les propriétés de cet estimateur par simulations et nous le comparons à l'estimateur CC.

Dans le chapitre 3, nous présentons le test du log-rank et sa généralisation au cas stratifié avec données manquantes. Puis nous introduisons une nouvelle statistique de ce test dans le cas où la censure est dépendante du groupe de traitement. Nous donnons sa distribution asymptotique sous l'hypothèse nulle. Nous comparons cette nouvelle statistique à celle basée sur l'analyse en cas complet au moyen de simulations.

CHAPITRE 1

INTRODUCTION A L'ANALYSE DE SURVIE

1.1. Introduction

L'analyse de survie voit sa naissance au 17ème siècle sur les premières études de la mortalité en Angleterre. Elle recouvre des méthodes d'analyse statistique de durées jusqu'à l'occurrence d'un événement d'intérêt bien défini. Cet évènement est souvent appelé *décès* qui, selon le domaine d'application, peut représenter : la mort d'un patient, la rechute ou la guérison d'une maladie, la panne d'un équipement industriel, la fin d'une période de chômage, l'expiration d'une ligne téléphonique ou d'un contrat d'assurance…

L'analyse des données de survie a la particularité de ne concerner que des variables aléatoires positives.

Une durée de vie T est donc une variable aléatoire positive définie et continue sur un espace probabilisé $(\Omega, A, \mathbb{P})$. Elle représente le temps écoulé entre un point de départ et la survenue de l'événement d'intérêt.

La fonction de répartition de T, appelée *fonction de distribution cumulative*, est la probabilité pour que l'évènement d'intérêt se produise entre $t_0 = 0$ et t. Elle est définie pour $t \geq 0$ par

$$F(t) = \mathbb{P}(T \leq t).$$

Si F est dérivable, notons f la densité de probabilité de T.

La *fonction de survie,* notée $S(t)$ est la probabilité que l'évènement d'intérêt survienne après l'instant t. Elle est définie pour tout $t \geq 0$ par :

$$S(t) = \mathbb{P}(T > t)$$

$$= 1 - F(t)$$

Par définition, la fonction de survie S est décroissante, continue à droite avec $\lim_{t\to 0} S(t) = 1$ et $\lim_{t\to\infty} S(t) = 0$.

La fonction de risque instantané (ou taux de panne, taux de défaillance, taux de décès, etc.), est défini pour tout $t \geq 0$ par:

$$\lambda(t) = \frac{f(t)}{S(t)} = \lim_{h\to 0} \frac{\mathbb{P}(t \leq T < t + h | T \geq t)}{h}.$$

La quantité $\lambda(t)$ représente la probabilité pour que l'évènement d'intérêt survienne à l'instant t sachant qu'il ne s'est pas produit avant t.

La *fonction de risque cumulé* est définie, pour tout $t \in \mathbb{R}^+$ par :

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Les cinq fonctions définies ci-dessus : $F(t)$, $f(t)$, $S(t)$, $\lambda(t)$ et $\Lambda(t)$ caractérisent la loi de la durée de vie et elles sont équivalentes. En effet chacune de ces fonctions peut être obtenue à partir de l'une quelconque des autres. Par exemple

$$\Lambda(t) = -\ln(S(t))$$

1.2. Données censurées

Dans plusieurs études, les dates d'occurrence de l'évènement d'intérêt ne sont pas connues et donc elles ne sont pas observées lors du recueil de données de survie. Dans ce cas, on parle d'observations censurées et une variable de censure *C* est introduite dans l'analyse. Soient T une variable durée de vie et C une variable aléatoire de censure. On distingue trois types de modèles de censure : censure à droite, censure à gauche et censure par intervalles. Nous détaillons dans ce qui suit ces trois types de censure.

1.2.1 Censure à droite :

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. On distingue :

i) La censure non-aléatoire de type I : Pour des considérations de temps ou de coût, l'observateur décide de terminer l'étude à une date prédéterminée. Etant donné un nombre positif fixé c, on observe X_i tel que $X_i = T_i \wedge c$ ($\wedge$ indique le minimum) et $\Delta = 1(T_i \leq c)$ la fonction indicatrice de l'observation réelle de l'évènement d'intérêt, qui est égale à 1 si $T_i \leq c$ et 0 sinon. Ce type de censure est rencontré dans les applications industrielles, en testant par exemple la durée de vie de n composants identiques sur un intervalle d'observation fixé, ou en biologie en testant l'efficacité d'une molécule sur des souris et celles qui sont vivantes au bout d'un temps fixé seront sacrifiées.

ii) La censure de type II : Dans ce cas on décide d'observer les durées de vie de n individus et d'arrêter l'étude lorsque k d'entres eux sont décédés. On observe ainsi les variables : $X_{(1)} = T_{(1)}, \dots, X_{(k)} = T_{(k)}, X_{(k+1)} = T_{(k)}, \dots, X_{(n)} = T_{(k)}$ où $X_{(i)}$ et $T_{(i)}$ représentent les statistiques d'ordre des variables X_i et T_i. $T_{(k)}$ étant la date de censure. Ce modèle de censure est utilisé dans les études de fiabilité lorsqu'on veut tester la durée de bon fonctionnement de n équipements et le test est terminé lorsque k équipements tombent en panne.

iii) La censure de type III (censure aléatoire de type I) : Soient $C_1, C_2, \dots, C_n$ des variables aléatoires i.i.d. On observe les variables $X_i = T_i \wedge C_i$ et $\Delta = 1(T_i \leq C_i)$. La censure aléatoire est souvent rencontrée en pratique. En particulier dans les essais cliniques, elle survient lorsque le patient quitte l'étude (il sera considéré comme perdu de vue) ou lorsque l'étude se termine alors que certains patients n'ont pas subi l'évènement d'intérêt (le patient est dit ainsi exclu-vivant).

1.2.2 Censure à gauche :

La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant la date du début d'observation. Dans ce cas, on sait seulement que la date de l'événement est inférieure à une certaine date connue.

On observe Z tel que $Z = T \vee C$ ($\vee$ indique le maximum) et l'indicatrice $1(T \geq C)$.

L'exemple de censure à gauche le plus cité dans la littérature considère l'horaire où les babouins, qui passent la nuit dans les arbres, descendent pour aller manger. L'horaire de la descente de l'arbre est observé seulement si le babouin descend de l'arbre après l'arrivée des observateurs. Dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. On observe donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs.

1.2.3 Censure par intervalles :

Les deux cas précédents peuvent être combinés. Dans ce cas, la durée de vie appartient à un intervalle de temps (ie. $C_1 < T < C_2$).

Ce type de censure est rencontré dans les suivis périodiques des patients, on sait alors uniquement que l'événement s'est produit entre deux dates de visites. Ce type de censure peut également se produire dans les inspections périodiques des équipements industriels.

Parmi tous ces modèles de censure, la censure aléatoire de type I reste la plus employée en pratique en supposant l'indépendance de T_i et C_i (censure non-informative) qui constitue une hypothèse indispensable et très utile d'un point de vue mathématique.

D'autres types de censure et plusieurs exemples sont proposés dans Bagdonavicuis et Nikulin (2002), Klein et Moeschberger (1997) et Lee et Wang (2003).

1.3. Données manquantes

On parle de données manquantes lorsque les données ne sont pas disponibles pour tous les individus de l'étude. Une telle situation peut entrainer d'importantes erreurs dans l'estimation des paramètres et leur inférence statistique.

Les premiers travaux sur le problème des données manquantes sont apparus dans les années 20 et 30, suivis après des travaux de Afifi et Elashoff (1966), Hartley et Hocking (1971) et Rubin (1976) et une décennie plus tard avec le livre de Little et Rubin (1987).

Dans l'analyse de survie, les données manquantes surviennent pour différentes raisons. Par exemple, dans les études épidémiologiques, l'information sur le certificat de décès peut être manquante, ou les résultats de l'autopsie peuvent être non conclusifs. Dans de tels cas, il n'est pas possible de déterminer si le décès est dû à la cause d'intérêt ou à d'autres causes extérieures. Une étude sur la mortalité des jeunes personnes dans les Pays-Bas (Van Der Laan et McKeague (1988)) montre que 90% des cas étaient enregistrés comme « cause de décès inconnue ».

Supposons que pour chaque élément indépendant i des *n* éléments de l'étude, on collecte un ensemble de mesures Y_{ij}, $j = 1, \dots, n_i$ (n_i est le nombre de mesures pour l'élément i) regroupées dans un vecteur $Y_i = \left(Y_{i1}, \dots, Y_{in_i}\right)'$. On définit le vecteur R_i des indicateurs de données manquantes R_{ij}:

$$R_{ij} = \begin{cases} 1 & \text{si } Y_{ij} \text{ est observé} \\ 0 & \text{sinon} \end{cases}$$

Le vecteur Y_i est partitionné en deux sous vecteurs : le vecteur Y_i^0 qui contient les Y_{ij} pour lesquels $R_{ij} = 1$, et Y_i^m qui contient les autres composantes.

On introduit la terminologie suivante (Molenberghs et Kenward (2007)):

Données complètes : données qui correspondent au vecteur (Y_i, R_i) des mesures qu'on veut collecter sur tous les individus.

Données observées : données réellement observées sur les éléments de l'étude dont une partie est manquante. Elles correspondent au vecteur $(R_i, R_i Y_i)$.

Cas complet: données sur les individus de l'étude mais sans les données manquantes. Cela signifie que les individus avec des données manquantes sont écartés de l'analyse.

1.3.1. Mécanismes de données manquantes

Il est important de distinguer le mécanisme qui engendre les données manquantes. Rubin (1976) a introduit trois mécanismes :

a) Données manquantes complètement au hasard (MCAR : Missing Completety At Random) : les probabilités de ne pas observer des données ne dépendent ni des données observées ni des celles non observées.

b) Données manquantes au hasard (MAR : Missing At Random) : c'est le cas où la probabilité pour qu'une donnée soit manquante dépend uniquement des données observées.

c) Données manquantes non-ignorables (NMAR : No Missing At Random) : la probabilité d'avoir des données manquantes peut dépendre des données non observées.

Le dernier scénario est plus problématique du fait que le mécanisme de manque peut dépendre des données non observées. Des problèmes de non identifiabilité peuvent être rencontrés.

Plus de détails sur les mécanismes de données manquantes et exemples peuvent être trouvés dans Little et Rubin (1987), Heitjan et Rubin (1991), Jacobsen et Keiding (1995), Gill *et al*. (1997) et Tsiatis (2006).

1.3.2 Méthodes de traitement des données manquantes au hasard

Sous l'hypothèse de données manquantes au hazard (MAR), plusieurs méthodes peuvent être utilisées pour estimer les paramètres du modèle. Nous citons dans cette section la méthode des cas complets et la méthode de pondération par probabilité inverse.

a) La méthode des cas complets

Elle consiste à ne garder dans l'analyse que les cas qui ne comportent aucune donnée manquante. Cette méthode entraîne généralement une diminution, considérable parfois, de la taille de l'échantillon, entraînant une diminution de la puissance et les estimateurs obtenus ainsi seront biaisés.

b) La pondération par probabilité inverse (en Anglais : inverse weighted probability)

La pondération par probabilité inverse est une méthode qui consiste à utiliser des poids qui peuvent être calculés par l'inverse de la probabilité qu'une donnée soit manquante. Ainsi chaque donnée observée sera pondérée par l'inverse de cette probabilité.

1.4. Martingales

Soit un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, où Ω est un ensemble, $\mathcal{F}$ est une tribu contenue dans l'ensemble des parties de Ω et $\mathbb{P}$ une probabilité sur la tribu $\mathcal{F}$.

Définition 1.4.1- Un *processus stochastique* est une famille de variables aléatoires indexées par le temps $(X(t): t \geq 0)$.

Définition 1.4.2- Une *filtration* $\mathcal{F}_t$ est une famille croissante de sous-tribus de $\mathcal{F}$: pour tout $s \leq t, \mathcal{F}_s \subset \mathcal{F}_t$.

On peut voir une filtration comme l'information que l'on dispose concernant les individus jusqu'à l'instant *t* inclus. $\mathcal{F}_{t-}$ représente le passé jusqu'à un instant avant *t* (ie que *t* n'est pas inclus).

Définition 1.4.3- Soit $\mathcal{F}_t$ une filtration. On dit que $X(t)$ est un processus $\mathcal{F}_t$-*adapté* si pour tout $t \geq 0$, X_t est $\mathcal{F}_t$-mesurable.

Définition 1.4.4- Soit $M = M(t); t \geq 0$ un processus stochastique continu à droite et limité à gauche (*cadlag*) et $\mathcal{F}_t$une filtration. M est une *martingale* par rapport à la filtration $\mathcal{F}_t$ si :

i. M est $\mathcal{F}_t$-adaptée,
ii. $\mathbb{E}(|M(t)|) < \infty$ (M intégrable),
iii. $\mathbb{E}(M(t)|\mathcal{F}_s) = M(s)$ pour tout $s \leq t$

La condition (iii) est équivalente à :

$$\mathbb{E}(dM(t)|\mathcal{F}_{t-}) = 0, \forall\, t > 0$$

Où $dM(t) = M((t + dt)-) - M(t-)$

Propriétés :

- Une martingale a une moyenne constante : $\mathbb{E}(M(t)) = \mathbb{E}(M(0))$.
- Les martingales ont des incréments non corrélés : $cov(M(t) - M(s), M(v) - M(u)) = 0$ pour tout $0 \leq s \leq t \leq u \leq v$.

Si la martingale M satisfait $\mathbb{E}(M(t)|\mathcal{F}_s) \geq M(s)$ pour tout $s \leq t$ alors M est une *sous-martingale*. Une martingale est dite *de carré intégrable* si $sup_t \mathbb{E}(M(t)^2) < \infty$. Une *martingale locale* est un processus pour lequel il existe une suite de temps d'arrêt $\{\tau_n, n \geq 0\}$ tels que pour tout n, le processus $M^{\tau_n} = M(t \wedge \tau_n)$ soit une martingale. En plus si M^{τ_n} est une martingale de carré intégrable alors M est dite une *martingale locale de carré intégrable*.

Définition 1.4.5- Un processus $X(t)$ est *prévisible* si et seulement si $X(t)$ est $\mathcal{F}_{\tau-}$-mesurable pour tout temps d'arrêt τ (ie sa valeur à un temps *t* est connue juste avant *t*)

Définition 1.4.6- Soit X un processus adapté et cadlag. A est dit *compensateur* de X si A est prévisible, cadlag et tel que $X - A$ soit une martingale de moyenne nulle. Si le compensateur existe, il est unique.

1.5. Processus de comptage

Définition 1.5.1. Un *processus de comptage* $N(t), t \geq 0$ est un processus stochastique adapté à une filtration $\mathcal{F}_t$ tel que

- $N(0) = 0$, $\mathbb{P}(N(t) < \infty) = 1$
- $N(s) \leq N(t)$ pour $s < t$
- $dN(t) = N(t) - N(t-)$ est égal à 0 ou 1

Où $N(t_-)$ désigne $\lim_{h \to 0} N(t-h)$

Proposition 1.5.1 Soit $N(t)$ un processus de comptage. Il existe un processus prévisible $\Lambda(t)$, croissant, continu à droite et nul en zéro tel que :

$$M(t) = N(t) - \Lambda(t)$$

$\Lambda(t)$ est le compensateur de $N(t)$ et on a $\mathbb{E}\big(N(t)\big) = \mathbb{E}\big(\Lambda(t)\big)$.

Proposition 1.5.2. Si N est absolument continu alors son compensateur Λ a la forme :

$\Lambda(t) = \int_0^t \lambda(s)\, ds$ où $\lambda(t)$ est un processus prévisible appelé *processus d'intensité*.

1.6. L'estimateur de Nelson-Aalen et l'estimateur de Kaplan-Meier :

En présence de données censurées, particulièrement à droite, les durées de vie ne sont pas totalement observées et dans ce cas, il est nécessaire d'estimer la distribution des durées de vie ainsi que la fonction de risque cumulative.

Soit T une durée de vie avec fonction de survie $S(t) = P(T > t)$ et fonction de risque $\lambda(t)$. Soit C une variable de censure à droite. On dispose de n observations indépendantes (X_i, Δ_i) , où $X = T \wedge C$ et $\Delta = 1(T \leq C)$ est l'indicatrice de censure. Soit $N_i(t) = 1(X_i \leq t, \Delta_i = 1)$, $Y_i(t) = 1(X_i \geq t)$. Soit le processus de comptage $N(t) = \sum_{i=1}^n N_i(t)$, $Y(t) = \sum_{i=1}^n Y_i(t)$ et la martingale de carré intégrable $M(t) = N(t) - \int_0^t Y(s)\lambda(s)ds$.

Puisque $dM(t)$ est un processus de moyenne nulle alors on a l'équation d'estimation suivante :

$$Y(t)d\Lambda(t) = dN(t)$$

où $\Lambda(t) = \int_0^t \lambda(s)ds$

1.6.1 L'estimateur de Nelson-Aalen

L'estimateur de Nelson-Aalen de la fonction de risque cumulative Λ est obtenu comme solution de l'équation d'estimation :

$$Y(t)d\Lambda(t) = dN(t)$$

Il est donné par :

$$\widehat{\Lambda}(t) = \int_0^t \frac{J(s)}{Y(s)} dN(s)$$

Avec $J(s) = 1(Y(s) > 0)$ et la convention que $0/0 = 0$.

L'estimateur $\widehat{\Lambda}(t)$ a été introduit par Nelson (1969,1972) puis à été généralisé par Aalen (1978b) pour les modèles de processus de comptage.

Sous certaines conditions de régularité, $n^{1/2}\left(\widehat{\Lambda} - \Lambda\right)$ converge en distribution vers une martingale gaussienne sur $[0, \tau]$, de variance estimée d'une manière consistante par :

$$n\int_0^t \frac{J(s)}{Y^2(s)} dN(s)$$

1.6.2. L'estimateur de Kaplan-Meier

Appelé encore '' Produit-Limite'', il a été introduit par Kaplan et Meier (1958), il joue un rôle important dans les méthodes non paramétriques pour les durées de vie.

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t.

Supposons qu'on est en présence de censure aléatoire à droite et que la variable aléatoire de censure C est indépendante de T.L'estimateur de Kaplan-Meier est défini par :

$$\hat{S}(t) = \prod_{s \le t} \left(1 - \Delta\hat{\Lambda}(s)\right) = \prod_{s \le t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right)$$

Où $\hat{\Lambda}(t)$ est l'estimateur de Nelson-Aalen de $\Lambda(t)$. L'estimateur peut être interprété comme le produit de probabilités conditionnelles. Soient $\tau_1, \dots, \tau_{N(t)}$ les dates de sauts de N sur $[0, t]$. On peut interpréter le facteur $(1 - \frac{1}{Y(\tau_k)})$ comme étant la probabilité de survivre sur l'intervalle $(\tau_k, \tau_{k+1}]$ sachant qu'on ait survécu sur $(0, \tau_k]$.

Les propriétés asymptotiques de l'estimateur de Kaplan-Meier sont obtenues des propriétés de l'estimateur de Nelson-Aalen. Une autre approche est basée sur la relation :

$$\frac{\hat{S}(t)}{S^*(t)} - 1 = -\int_0^t \frac{\hat{S}(s-)J(s)}{S^*(s)Y(s)}\, dM(s)$$

pour $t \in [0, \tau)$ où $S^*(t) = exp(-\Lambda^*(t))$ avec $\Lambda^*(t) = \int_0^t J(s)\lambda(s)ds$.

En se basant sur la formule précédente et sous certaines conditions, on peut montrer que $\hat{S}$ est uniformément consistant sur des intervalles compacts et que pour tout $t \in [0, \tau)$, $n^{1/2}\left(\hat{S} - S\right)$ converge en distribution vers $-U.S$, où U est une martingale gaussienne.

La variance de $\hat{S}(t)$ est estimée par :

$$\tilde{\Sigma}(t) = \hat{S}(t)^2 \int_0^t \bar{Y}(s)^2 dN(s)$$

Un autre estimateur consistant de la variance de $\hat{S}(t)$ est donnée par :

$$\hat{\Sigma}(t) = \hat{S}(t)^2 \int_0^t \{Y(s)(Y(s) - N(s))\}^{-1} dN(s)$$

Cette dernière formule est connue sous l'appellation '' la formule de Greenwood''.

1.7. Le modèle de regression de Cox

Le modèle de régression de Cox, appelé également modèle à risques proportionnels, a été introduit par Cox(1972). Il permet de relier la durée de survie d'un patient à plusieurs variables explicatives ou facteurs pronostiques. Il est donné par :

$$\lambda_{T^0|Z}(t) = Y(t)\lambda_0(t)\exp(Z^\intercal(t)\beta)$$

Où T^0 dénote la variable aléatoire durée de vie, $Y(t)$ est une indicatrice de risque, $Z(t) = (Z_1(t), \dots, Z_p(t))$ est un p-vecteur de covariables, $\lambda_{T^0|Z}$ est la fonction de risque instantanée sachant Z, β est un vecteur de paramètres de régression et λ_0 est une fonction définie sur $\mathbb{R}^+$ à valeurs positives appelée fonction de risque instantané de base, c'est le risque instantané d'un individu quand le vecteur des covariables Z est nul. En particulier, le modèle de Cox comprend le modèle exponentiel et le modèle de Weibull, obtenus avec les fonctions de risque instantané $\lambda_0 = \lambda$ et $\lambda_0(t) = \lambda\alpha t^{\alpha-1}$ respectivement.

Le modèle de Cox est un modèle semi-paramétrique, puisque dans la définition du risque instantané, le terme $\exp(Z^\intercal(t)\beta)$ est paramétrique et le risque de base $\lambda_0(t)$ est non-paramétrique.

Si les covariables sont indépendantes du temps t, alors le rapport des risques instantanés est constant par rapport à t pour deux individus avec des vecteurs de covariables Z_1 et Z_2 respectivement:

$$\frac{\lambda_{T^0|Z_1}(t)}{\lambda_{T^0|Z_2}(t)} = \frac{Y(t)\lambda_0(t)\exp(Z_1^\intercal\beta)}{Y(t)\lambda_0(t)\exp(Z_2^\intercal\beta)} = \exp(Z_1 - Z_2)^\intercal\beta$$

Le problème statistique consiste à estimer le paramètre de régression β et la fonction de risque cumulé de base $\Lambda_0(t) = \int_0^t \lambda(u)du$. L'estimation se fait avec la méthode de la vraisemblance partielle.

Soient n indépendantes copies de $(N_i(t), Y_i(t), Z_i(t))$, $i = 1, \dots, n$ observées sur un intervalle $[0, \tau], \tau < \infty$. L'estimateur du paramètre de régression β est obtenu en maximisant la fonction de vraisemblance partielle de Cox (Cox(1972, 1975)) :

$$L(\beta) = \prod_t \prod_i \left(\frac{\exp(Z_i^\intercal(t)\beta)}{S_0(t,\beta)} \right)^{\Delta N_i(t)}$$

Où

$$S_0(t,\beta) = \sum_{i=1}^{n} Y_i(t)\exp(Z_i^\intercal(t)\beta)$$

Définissons les dérivées d'ordre 1 et 2 de $S_0(t,\beta)$ par rapport à β :

$$S_1(t,\beta) = \sum_{i=1}^{n} Y_i(t)\exp(Z_i^\intercal(t)\beta)Z_i^\intercal(t)$$

$$S_2(t,\beta) = \sum_{i=1}^{n} Y_i(t)\exp(Z_i^\intercal(t)\beta)Z_i^\intercal(t)^{\otimes 2}$$

Avec $v^{\otimes 2} = vv^\intercal$ pour un vecteur v.

L'estimateur $\hat{\beta}$ de β est la solution de la fonction du score $U(\hat{\beta}) = 0$, où :

$$U(\beta) = \sum_{i=1}^{n} \int_0^\tau (Z_i(t) - E(t,\beta)dN_i(t))$$

Avec

$$E(t,\beta) = \frac{S_1(t,\beta)}{S_0(t,\beta)}$$

La résolution de l'équation du score précédente se fait numériquement en utilisant habituellement l'algorithme itératif de Newton-Raphson.

Si β est fixé alors un estimateur de $\Lambda_0(t)$ est donné par l'estimateur de Nelson-Aalen :

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{1}{S_0(s,\beta)} dN(s)$$

Où $N(t) = \sum_i N_i(t)$

Posons $I(\beta) = I(\tau, \beta)$ avec :

$$I(t,\beta) = \sum_{i=1}^{n} \int_0^t \left(\frac{S_2(s,\beta)}{S_0(s,\beta)} - E(s,\beta)^{\otimes^2} \right) dN_i(s)$$

Notons β_0 la vraie valeur de β. Certaines conditions de stabilité et de régularité asymptotiques décrites dans Andersen et Gill (1982), permettent de montrer la consistance de $\hat{\beta}$ et d'énoncer les théorèmes suivants qui établissent les propriétés asymptotiques des estimations $\hat{\beta}$ et $\widehat{\Lambda}_0$:

Théorème 1.7.1 : Lorsque $n \to \infty$

$$n^{-1/2} U(\beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

$$n^{-1/2} (\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma^{-1})$$

et Σ est estimée par $n^{-1} I(\hat{\beta})$

Théorème 1.7.2 : Lorsque $n \to \infty$

$$n^{-1/2} \left(\widehat{\Lambda}_0(t, \hat{\beta}) - \Lambda_0(t) \right) \xrightarrow{\mathcal{L}} U(t)$$

Où $U(t)$ est un processus gaussien de moyenne nulle et fonction covariance estimée par :

$$n\left(\int_0^t S_0(s,\hat{\beta})^{-2} dN(s) + \int_0^t E(s,\hat{\beta})^{\intercal} d\hat{\Lambda}_0(s,\hat{\beta}) \left(n^{-1} I(\hat{\beta})\right)^{-1} \int_0^t E(s,\hat{\beta})\, d\hat{\Lambda}_0(s,\hat{\beta})\right)$$

1.8. Modèle de transformation linéaire

La classe des modèles de transformation linéaire a fait l'objet de nombreux travaux, citons par exemple : Chen et al. (1995), Cheng et al.(1995), Fine et al. (1998), Fleming et Lin (2000), Slud et Vonta (2004), Ma et Kosorok (2005), Martinussen et Schieke (2006), Kong et al. (2004,2006), Kosorok et Son (2007), Dupuy (2008).

Soit *T* une durée de vie et soit *Z* un vecteur de *p* covariables. On considère le modèle semi-paramétrique de transformation linéaire:

$$h(T) = -\beta_0^{\intercal} Z + \varepsilon \qquad (1.8.1)$$

où h est une fonction croissante et inconnue, β_0 est un vecteur de paramètres inconnus et ε est une erreur aléatoire avec fonction de distribution connue F_ε. $v^{\intercal}$ désigne la transposée de v.

Soit $S(t|Z)$ la fonction de survie de *T* étant donné *Z* alors on a :

$$\begin{aligned} S_Z(t) &= \mathbb{p}(T > t|Z) \\ &= \mathbb{P}(h(T) > h(t)|Z) \\ &= \mathbb{P}\left(-\beta_0^{\intercal} Z + \varepsilon > h(t)\right) \\ &= \mathbb{P}(\varepsilon > h(t) + \beta_0^{\intercal} Z) \\ &= 1 - F_\varepsilon(h(t) + \beta_0^{\intercal} Z) \end{aligned}$$

La fonction de risque de T sachant Z est :

$$\lambda(t,Z) = -\frac{S_Z'(t)}{S_Z(t)}$$

$$= \frac{[S_\varepsilon(h(t) + \beta_0^\intercal Z)]'}{S_\varepsilon(h(t) + \beta_0^\intercal Z)}$$

$$= \left\{\frac{dh(t)}{dt}\right\} \lambda_\varepsilon(h(t) + \beta_0^\intercal Z)$$

Où $\lambda_\varepsilon(h(t) + \beta_0^\intercal Z)$ est la fonction de risque associé à ε.

Le modèle (1.8.1) peut être réécrit comme suit :

$$H(T) = e^{-\beta_0^\intercal Z} e^{\varepsilon}$$

Où $H(.) = \exp(h(.))$ est une fonction positive strictement croissante telle que $H(0) = 0$ et $\lim_{t\to\infty} H(t) = \infty$.

La fonction de survie s'écrit alors :

$$S_Z(t) = \mathbb{P}(e^{-\beta_0^\intercal Z} e^{\varepsilon} > H(t))$$

$$= \mathbb{P}(e^{\varepsilon} > H(t) e^{\beta_0^\intercal Z})$$

$$= S_\varepsilon\left(H(t) e^{\beta_0^\intercal Z}\right)$$

Où S_ε est la fonction de survie de ε.

Soit $\dot{H}$ la dérivée de H , alors la fonction de risque de T sachant Z peut être écrite sous la forme :

$$\lambda(t) = \dot{H}(t) e^{\beta_0^\intercal Z} \lambda_\varepsilon(e^{\beta_0^\intercal Z} H(t)) \qquad (1.8.2)$$

où λ_ε est la fonction de risque associée à exp(ε).

La classe des modèles de transformation linéaire est une classe plus générale qui contient en particulier le modèle à risques proportionnels de Cox et le modèle à risques convergents (proportional Odds model).

1.8.1. Modèle à risques proportionnels

Si ε suit une distribution à valeurs extrêmes, ie $F_\varepsilon(u) = 1 - \exp(-\exp(u))$ alors $\exp(\varepsilon)$ suit une loi exponentielle $(\lambda_0(t) = 1)$ et la fonction du risque (1.8.2) deviendra

$$\lambda(t) = \dot{H}(t)e^{\beta_0^\intercal Z}$$

et le modèle (1.8.1) se réduit à un modèle à risques proportionnels de Cox décrit dans (1.7) avec fonction de risque cumulé de base $H(t)$.

1.8.2. Modèle à risques convergents (proportional Odds model)

Dans le cas où ε suit une distribution logistique standard $F_\varepsilon(u) = \exp(u)/(1 + \exp(u))$, on obtient le modèle à risques convergents.

Les fonctions de survie et de risque sont données respectivement par :

$$S(t/Z) = \frac{1}{1 + H(t)\exp(\beta_0^\intercal Z)}$$

$$\lambda(t) = \frac{\dot{H}(t)}{\exp(-\beta_0^\intercal Z) + H(t)}$$

Le modèle à risques convergents suppose que le rapport des risques de deux individus avec covariables Z_1 et Z_2 respectivement, converge vers 1 quand t augmente :

$$RR(t) = \frac{\lambda(t, Z_2)}{\lambda(t, Z_1)} = \frac{\exp(-\beta_0^\intercal Z_1) + H(t)}{\exp(-\beta_0^\intercal Z_2) + H(t)}$$

Avec $RR(0) = \exp(\beta_0^\intercal(Z_2 - Z_1))$, et $\lim_{t\to\infty} RR(t) = 1$.

CHAPITRE 2

ESTIMATION DANS UN MODELE DE TRANSFORMATION LINEAIRE AVEC DONNEES MANQUANTES

2.1. Introduction

Depuis l'introduction en 1972 par D. Cox du modèle à risques proportionnels, la littérature consacrée aux modèles de régression de durées a connu un essor remarquable. De nombreux modèles, ont été proposés et leurs différents aspects : inférence, validation, sélection, application en fiabilité et en analyse de survie ont été étudiés en détail, nourrissant une littérature foisonnante. Plusieurs ouvrages dressent un panorama de ce domaine de recherche en perpétuelle évolution. Citons, par exemple, Andersen et al.(1993), Bagdonavicius et Nikulin (2002), Fleming et Harrington (1991), Klein et Moeschberger (1997), Lawless (2003), Martinussen et Scheike (2006), Meeker et Escobar (1998) qui s'adressent aussi bien au lecteur intéressé par les aspects théoriques de l'inférence statistique dans les modèles de durées qu'au praticien de la statistique à la recherche d'outils de modélisation et d'exemples concrets. Parmi la très grande variété des modèles qui ont été développés au cours des trente dernières années, la classe des modèles semi-paramétriques de transformation linéaire offre l'avantage d'inclure plusieurs des modèles les plus utilisés dans les applications. Cette classe de modèles a suscité une littérature très abondante depuis vingt ans. Citons par exemple : Chen et al. (1995), Cheng et al.(1995), Fine et al. (1998), Fleming et Lin (2000), Slud et Vonta (2004), Ma et

Kosorok (2005), Martinussen et Schieke (2006), Kong et al. (2004,2006), Kosorok et Song (2007), Dupuy (2008).

Notons T la durée aléatoire jusqu'à un instant de défaillance ou de panne et $Z = (Z^1, \dots, Z^p)^\intercal$ un vecteur de dimension p de variables explicatives (où $(.)^\intercal$ désigne la transposée). La classe des modèles semi-paramétriques de transformation linéaire exprime la relation entre T et Z sous la forme :

$$h(T) = -\beta_0^\intercal Z + \varepsilon \tag{2.1}$$

où e est une fonction strictement croissante inconnue, $\beta_0 = (\beta^1, \dots, \beta^p)^\intercal$ est un vecteur de paramètres de régression inconnus (paramètres d'intérêt du modèle) et ε désigne un terme d'erreur aléatoire (indépendant de Z) dont la loi de probabilité est supposée connue (on notera F_ε sa fonction de répartition). Cette classe contient en particulier le modèle à risques proportionnels de Cox et le modèle à risques convergents.

Plusieurs méthodes ont été proposées pour estimer le paramètre β_0 dans la classe de modèles (2.1).

Cheng et al. (1995) ont en particulier proposé des équations d'estimation simples à partir d'un échantillon d'observations indépendantes (X_i, Δ_i, Z_i) , $i = 1, \dots, n$ du triplet (X, Δ, Z), où $X = min(T, C)$ désigne la durée observée, C une censure aléatoire, Z un vecteur de variables explicatives, $\Delta = 1(T \leq C)$ et 1(.) désigne la fonction indicatrice. L'estimateur proposé est consistant et asymptotiquement gaussien mais il n'est pas plus efficace que les estimateurs déjà développés pour les deux cas particuliers du modèle (2.1) (Andersen et *al.* (1993), Murphy et *al.* (1997)). Sa simplicité, en revanche, en fait un point de départ intéressant pour construire de nouveaux estimateurs dans des situations moins standards que celle décrite par les données (X_i, Δ_i, Z_i), $i = 1, \dots, n$.

Ainsi, Kong et al. (2004, 2006) ont récemment adapté les équations d'estimation de Cheng et *al*. aux études cas-cohorte, Fine et *al.* (1998) les ont adaptées au cas où le support de la censure est inclus dans celui de la durées d'intérêt T.

Dans ce travail, nous adaptons ces équations d'estimation à une situation de données manquantes. Nous construisons des équations d'estimation puis nous montrons la consistance de l'estimateur obtenu. Nous évaluons ensuite les propriétés de cet estimateur par simulations et nous le comparons à l'estimateur CC. Enfin, nous illustrons la méthode proposée sur un jeu de données réelles.

2.2. Equations généralisées de Cheng et *al.*

Soit T_i la durée de vie pour un individu i, $i = 1, \dots, n$. Soit Z_i le vecteur de covariables pour l'individu i. Soit C_i la variable de censure de fonction de survie $G(t) = \mathbb{P}(C > t)$. On observe ainsi n copies indépendantes (X_i, Δ_i, Z_i) du vecteur aléatoire (X, Δ, Z) où $X = \min(T, C)$ et $\Delta = 1(T \leq C)$. On suppose que C_i indépendante de T_i conditionnellement à Z_i et que la censure est non-informative. Supposons également que ε_i sont indépendants et identiquement distribuées pour $i = 1, \dots, n$.

Soient les variables $1\big(T_i \geq T_j\big), (i \neq j = 1, \dots, n)$. Alors :

$$\mathbb{E}\big[1\big(T_i \geq T_j\big)|Z_i, Z_j\big] = \mathbb{P}\big[h(T_i) \geq h(T_j)|Z_i, Z_j\big]$$

$$= \mathbb{P}\big(-Z_i^{\intercal}\beta_0 + \varepsilon_i \geq -Z_j^{\intercal}\beta_0 + \varepsilon_j|Z_i, Z_j\big)$$
$$= \mathbb{P}\big(\varepsilon_i - \varepsilon_j \geq (Z_i - Z_j)^{\intercal}\beta_0|Z_i, Z_j\big)$$

$$= \int_{-\infty}^{+\infty} \int_{u+Z^{\intercal}{}_{ij}\beta_0}^{+\infty} f_{\varepsilon_i,\varepsilon_j}(v, u)\, dudv$$

$$= \int_{-\infty}^{+\infty} f_{\varepsilon_j}(u) \left(\int_{u+\beta_0^{\intercal}Z_{ij}}^{+\infty} f_{\varepsilon_i}(v) dv \right) du$$

$$= \int_{-\infty}^{+\infty} f_{\varepsilon_j}(u)\big(1 - F_{\varepsilon_i}(u + Z_{ij}^{\intercal}\beta_0)\big) du$$

$$= \int_{-\infty}^{+\infty} \Big(1 - F_{\varepsilon_i}\big(u + Z_{ij}^{\intercal}\beta_0\big)\Big)\, dF_{\varepsilon}(u)$$

$$:= \xi\left(Z_{ij}^{\intercal}\beta_0\right)$$

Où $Z_{ij} := Z_i - Z_j$

Les équations généralisées d'estimation proposées par Liang et Zeger (1986) permettent de faire l'inférence statistique de β_0 :

$$\tilde{\varphi}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega(Z_{ij}^{\intercal}\beta) Z_{ij}\left[1\left(T_i \geq T_j\right) - \xi(Z_{ij}^{\intercal}\beta)\right]$$

où ω est une fonction de poids.

En supposant que $\omega\left(Z_{ij}^{\intercal}\beta\right) = 1$, Cheng et al. (1995) proposent d'estimer β_0 dans le modèle (2.1) par la solution de l'équation d'estimation suivante :

$$\varphi(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega(Z_{ij}^{\intercal}\beta) Z_{ij}\left\{\frac{\Delta_j 1\left(X_i \geq X_j\right)}{\hat{G}^2\left(X_j\right)} - \xi\left(Z_{ij}^{\intercal}\beta\right)\right\} = 0 \qquad (2.2)$$

Où $\widehat{G}$ désigne l'estimateur de Kaplan-Meier de la fonction de survie *G*. Ils montrent que lorsque $\omega = 1$, l'équation (2.2) admet une solution unique.

Cheng et al. (1995) ont montré que sous des hypothèses de régularité, l'estimateur obtenu par la résolution de l'équation (2.2) est consistant et qu'il converge approximativement vers une distribution normale de moyenne nulle et de matrice de variance-covariance qui peut être estimée de manière consistante (la preuve des résultats de Cheng et *al.* (1995) est donnée en Annexe 2.B) .

2.3. Estimation par la méthode IPW en présence de données manquantes

Supposons que l'on dispose d'un échantillon de n items. Pour chacun d'entre eux, on observe un vecteur de variables explicatives Z à l'instant $t = 0$ (début de l'étude) et l'on souhaite observer la durée jusqu'à la défaillance. On considère la situation où l'observation de la durée (éventuellement censurée) et de l'indicatrice de censure n'est possible que pour

un sous-échantillon aléatoire de l'échantillon initial. Cette situation se rencontre en particulier en fiabilité lorsque des contraintes techniques inattendues viennent limiter les possibilités de recueil des données ou interrompre une partie d'un essai en cours. On dispose alors d'observations du triplet (X, Δ, Z) sur un sous-ensemble des n items tandis que pour les autres items, on ne dispose que des observations de Z. Dans ce contexte de données manquantes, une solution simple pour estimer β_0 consiste à mener une analyse en "cas complets" ("CC" par la suite) c'est-à-dire à : i) retirer de l'échantillon les items i pour lesquels l'information (X_i, Δ_i) est manquante, ii) calculer l'estimateur de Cheng et al. (1995) sur les items restants. Cette solution entraîne néanmoins une perte d'information et conduit à des estimateurs inefficaces et biaisés (Little et Rubin (1987)).

Nous proposons donc dans ce travail une alternative basée sur le principe de la "pondération par probabilité inverse (Inverse Weighted Probability (IWP), dont l'article de Seaman et White (2013) dresse un panorama récent.

Dans ce contexte, les données sont constituées de n vecteurs indépendants $(R_i, X_i R_i, \Delta_i R_i, Z_i)$ où $R_i = 1$ si (X_i, Δ_i) est observé et $R_i = 0$ sinon. Nous supposons que les données sont manquantes au hasard (Tsiatis (2006)):

A. La probabilité d'observer le couple (X, Δ) ne dépend pas de (X, Δ) mais peut dépendre de Z. Autrement dit, $\mathbb{P}(R = r|X, \Delta, Z) = \mathbb{P}(R = r|Z), r = 0,1.$

Dans la suite, nous notons $\eta_i = \mathbb{P}(R_i = 1|Z_i)$ la probabilité d'observer le vecteur complet (X, Δ, Z) pour l'item *i* et nous l'appelons probabilité de sélection.

L'intuition à la base du principe de la pondération par probabilité inverse est la suivante. Soit un item i, dont la probabilité de fournir une observation complète (X_i, Δ_i, Z_i) est égale à η_i et tel que $R_i = 1$. Alors cet item peut être considéré comme le représentant observé d'un groupe de taille $\frac{1}{\eta_i}$ d'items similaires mais non observés. La pondération par probabilité inverse consiste donc à pondérer chaque item i tel que $R_i = 1$ par $\frac{1}{\eta_i}$ afin de tenir compte de la contribution d'items similaires mais incomplètement observés. Les probabilités de sélection $\eta_i$étant généralement inconnues, il est nécessaire de les estimer.

Sous l'hypothèse A, elles peuvent être estimées (paramétriquement, non/semi-paramétriquement) à partir des observations $(R_i, Z_i), i = 1, \dots, n$.

Par exemple, si l'on suppose que les η_i suivent un modèle de régression logistique (ie $ogit(\eta_i) = logit\big(\eta_i(\gamma_0)\big) = \gamma_0^{\intercal} Z_i$), un estimateur $\hat{\gamma}_n$ de $\gamma_0 \in \mathbb{R}^p$ est obtenu en maximisant la vraisemblance $\prod_{i=1}^{n} {\eta_i}^{R_i} \{1 - \eta_i\}^{1-R_i}$. Si ce modèle est correct, η_i peut être estimée de manière consistante par $\widehat{\eta_i} = \eta_i(\hat{\gamma}_n)$ (voir Fahrmeir et Kaufmann (1985)).

Nous proposons d'estimer β_0 par la solution $\hat{\beta}_n$ de l'équation d'estimation

$$\varphi_n(\beta) := \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} Z_{ij} \left\{ \frac{\Delta_j 1\big(X_i \geq X_j\big) R_i R_j}{G^2\big(X_j\big)\eta_i \eta_j} - \xi\big(Z_{ij}^{\intercal}\beta\big) \right\} = 0$$

$$:= \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \varphi_{ij}\,(\beta) \qquad (2.3)$$

En effet, on a :

$$\mathbb{E}\left[\frac{\Delta_j 1\big(X_i \geq X_j\big) R_i R_j}{G^2\big(X_j\big)\eta_i \eta_j} |Z_i, Z_j \right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\Delta_j 1\big(X_i \geq X_j\big) R_i R_j}{G^2\big(X_j\big)\eta_i \eta_j} |T_j, Z_i, Z_j \right] |Z_i, Z_j \right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{1(T_j \leq C_j) 1\big(C_i \wedge T_i \geq T_j\big) R_i R_j}{G^2\big(T_j\big)\eta_i \eta_j} |T_j, Z_i, Z_j \right] |Z_i, Z_j \right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{1(T_j \leq C_j) 1\big(C_i \geq T_j\big) 1(T_i \geq T_j) R_i R_j}{G^2\big(T_j\big)\eta_i \eta_j} |T_j, Z_i, Z_j \right] |Z_i, Z_j \right]$$

$$= \mathbb{E}\left[\frac{1}{G^2\big(T_j\big)\eta_i \eta_j} \mathbb{E}\big[1(T_j \leq C_j) 1\big(C_i \geq T_j\big) 1(T_i \geq T_j) R_i R_j |T_j, Z_i, Z_j \big] |Z_i, Z_j \right]$$

$$= \mathbb{E}\big[1\big(T_i \geq T_j\big) |Z_i, Z_j \big] = \xi\big(Z_{ij}^{\intercal}\beta_0\big)$$

Où β_0 est la vraie valeur de β.

Pour que la solution de $\varphi_n(\beta) = 0$ soit un estimateur de β_0, nous montrons que $\mathbb{E}\big(\varphi_{ij}(\beta_0)\big) = 0$:

$$\mathbb{E}\left[Z_{ij}\frac{\Delta_j 1\big(X_i \geq X_j\big)R_iR_j}{G^2\big(X_j\big)\eta_i\eta_j}|Z_i, Z_j\right] = \mathbb{E}\left[\mathbb{E}\left[Z_{ij}\frac{\Delta_j 1\big(X_i \geq X_j\big)R_iR_j}{G^2\big(X_j\big)\eta_i\eta_j}|T_j, Z_i, Z_j\right]|Z_i, Z_j\right]$$

$$= \mathbb{E}\left[Z_{ij}\mathbb{E}\left[\frac{\Delta_j 1\big(X_i \geq X_j\big)R_iR_j}{G^2\big(X_j\big)\eta_i\eta_j}|T_j, Z_i, Z_j\right]|Z_i, Z_j\right]$$

$$= Z_{ij}\mathbb{E}\big[1(T_i \geq T_j)|Z_i, Z_j\big]$$

$$= Z_{ij}\xi\big(Z_{ij}^{\intercal}\beta_0\big)$$
$$\Longrightarrow \mathbb{E}\big(\varphi_{ij}(\beta_0)|Z_i, Z_j\big) = 0$$

D'où il est naturel d'estimer β_0 par la solution $\hat{\beta}$ de $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\varphi_{ij}\,(\beta) = 0$.

G peut être estimé par $\hat{G}$ l'estimateur de Kaplan-Meier (Andersen *et al*. (1993)). En remplaçant η_i, η_j et G par leurs estimateurs, nous obtiendrons une version approchée de l'équation (2.3) qui permet d'estimer le paramètre β_0 :

$$\Psi_{\mathrm{n}}(\beta) = \sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} Z_{ij}\left\{\frac{\Delta_j 1\big(X_i \geq X_j\big)R_iR_j}{\hat{G}^2\big(X_j\big)\hat{\eta}_i\hat{\eta}_j} - \xi\big(Z_{ij}^{\intercal}\beta\big)\right\} = 0 \qquad (2.4)$$

La solution $\hat{\beta}_n$de l'équation (2.4), est une approximation de l'estimateur théorique $\hat{\beta}$.

Remarque 2.1 :

Nous avons supposé l'indépendance de C et Z. Néanmoins, comme le mentionnent Cheng et al. (1995), cette hypothèse peut aisément être relâchée en remplaçant $\hat{G}$ dans l'équation (2.3) par un estimateur de la fonction de survie conditionnelle de C sachant Z (par exemple, un estimateur de type noyau si Z est continue, voir Dabrowska (1992)).

Remarque 2.2 :

La démarche décrite ci-dessus peut être généralisée à une modélisation semi paramétrique (par exemple : $logit(\eta_i) = \gamma_0^\intercal Z_i + e(\theta_0^\intercal Z_i)$) ou non paramétrique (par exemple : $logit(\eta_i) = e(\theta_0^\intercal Z_i)$) des η_i ou e désigne une fonction inconnue.

Des méthodes de type vraisemblance locale peuvent alors être utilisées pour estimer les η_i (voir Carroll et *al.* (1997) par exemple). Dans la suite, nous nous plaçons dans un cadre paramétrique et supposons que les η_i suivent le modèle logistique $logit(\eta_i) = \gamma_0^\intercal Z_i$ où $\gamma_0 \in \mathbb{R}^p$, $\eta_i := \eta_i(\gamma_0)$ et $\eta_i(\gamma) \coloneqq \mathbb{P}(R_i = 1|Z_i; \gamma)$.

2.4. Propriétés asymptotiques

Nous établissons la consistance de la suite d'estimateurs $(\hat{\beta}_n)$. Les conditions de régularité suivantes seront utiles :

B. Il existe des compacts $\mathcal{B}$ et $\mathcal{C}$ de $\mathbb{R}^p$ tels que $\beta_0 \in \mathcal{B}$ et $\gamma_0 \in \mathcal{G}$.

C. Les covariables sont bornées, c'est à dire $|Z^k| \leq c_1$ pour $k = 1, \dots, p$.

D. Il existe des constantes strictement positives c_2 et c_3 telles que $G(\tau) > c_2$ et $inf_{\gamma \in \mathcal{G}} \mathbb{P}(R = 1|Z = z; \gamma) > c_3$ pour tout z.

E. La fonction de répartition $F_\varepsilon(.)$ de ε est de classe $\mathcal{C}^1$.

F. La matrice $\mathbb{E}[Z_{12} Z_{12}^\intercal \dot{\xi}(Z_{12}^\intercal \beta_0)]$ est définie positive.

Nous énonçons le théorème suivant :

Théorème 2.4.1: (Mezaouer et *al*.(2014))

Sous les conditions A-F,

a) la suite d'estimateurs $(\hat{\beta}_n)$ *converge en probabilité vers* β_0 *lorsque n tend vers l'infini et,*

b) $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$.

Preuve du théorème 2.4.1 :

Pour monter le résultat (a) du théorème 2.4.1, nous vérifions les conditions du théorème de Foutz (Foutz(1977)). Ce résultat est basé sur le théorème d'inversion locale (voir Rudin(1964)), qui spécifie les conditions suffisantes pour l'existence et la consistance de la solution d'une équation d'estimation du type $S_n(\theta) = 0$ où θ est le paramètre à estimer. Une généralisation du résultat de Foutz est donnée dans Strawderman et Tsiatis (1996) dans le cas où la dimension de θ croît avec la taille de l'échantillon.

Conditions du théorème de Foutz :

1. $n^{-2}\Psi_n(\beta_0)$ converge en probabilité vers 0 lorsque n tend vers l'infini
2. $\partial\Psi_n(\beta)/\partial\beta^{\intercal}$ existe et est continue sur un voisinage de β_0
3. $-n^2\,\partial\Psi_n(\beta)/\partial\beta^{\intercal}$ converge uniformément en probabilité vers une fonction $A(\beta)$ sur un voisinage de β_0 et $A(\beta_0)$ est définie positive.

Preuve de la condition 1 :

Pour montrer que la condition 1 est vérifiée, nous allons écrire $n^{-2}\Psi_n(\beta)$ sous frome de somme de trois termes dont les deux derniers convergent en probabilité vers 0 et le premier converge vers $\Psi(\beta)$, puis nous montrons que $\Psi(\beta_0) = 0$.

Posons $U_{ij} = \Delta_j 1(X_i \geq X_j) R_i R_j$ et décomposons $n^{-2}\Psi_n(\beta)$ sous la forme suivante :

$$n^{-2}\Psi_n(\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}\left\{\frac{U_{ij}}{\hat{G}^2(X_j)\hat{\eta}_i\hat{\eta}_j} - \frac{U_{ij}}{\hat{G}^2(X_j)\eta_i\eta_j} + \frac{U_{ij}}{\hat{G}^2(X_j)\eta_i\eta_j} - \frac{U_{ij}}{G^2(X_j)\eta_i\eta_j} + \frac{U_{ij}}{G^2(X_j)\eta_i\eta_j} - \xi(Z_{ij}^{\intercal}\beta)\right\}$$

$$= \frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}\left\{\frac{U_{ij}}{G^2(X_j)\eta_i\eta_j} - \xi(Z_{ij}^{\intercal}\beta)\right\}$$

$$+ \frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}U_{ij}}{\eta_i\eta_j}\left\{\frac{1}{\widehat{G}^2(X_j)} - \frac{1}{G^2(X_j)}\right\}$$

$$+\frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}U_{ij}}{\widehat{G}^2(X_j)}\left\{\frac{1}{\hat{\eta}_i\hat{\eta}_j}-\frac{1}{\eta_i\eta_j}\right\}$$

$$=\frac{(n-1)}{n}\left[\Psi_{n,1}(\beta)+\Psi_{n,2}+\Psi_{n,3}\right]$$

Considérons le premier terme $\Psi_{n,1}(\beta)$. Sous les conditions A-F, on a pour tout $k=1,\dots,p$ et $\beta\in\mathcal{B}$:

$$\mathbb{E}\left[\left|Z_{12}^k\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}-\xi(Z_{12}^{\intercal}\beta)\right\}\right|\right]\leq 2c_1\left(\frac{1}{c_2^2c_3^2}+1\right)<\infty \qquad (2.5)$$

Ce terme $\Psi_{n,1}(\beta)$ est une U-statistique et d'après la loi des grand nombres pour les U-statistiques (Hoeffding (1961)), lorsque n tend vers l'infini :

$$\Psi_{n,1}(\beta)\xrightarrow{P}\mathbb{E}\left[\left|Z_{12}^k\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}-\xi(Z_{12}^{\intercal}\beta)\right\}\right|\right]$$

Où $\xrightarrow{P}$ désigne la convergence en probabilité.

Soit $\Psi_{n,2}^k$ la *k*-ème composante de $\Psi_{n,2}$ est donnée par :

$$\Psi_{n,2}^k=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left\{\frac{1}{\hat{G}^2(X_j)}-\frac{1}{G^2(X_j)}\right\}$$

$$=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left\{\frac{1}{\hat{G}^2(X_j)}-\frac{1}{G^2(X_j)}\right\}$$

$$=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left[\frac{G(X_j)-\hat{G}(X_j)}{G^3(X_j)}\right]\left[\frac{G(X_j)\left(G(X_j)+\hat{G}(X_j)\right)}{2\hat{G}^2(X_j)}\right]$$

$$=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{Z_{ij}^kU_{ij}}{\eta_i\eta_j}\left[\frac{G(X_j)-\hat{G}(X_j)}{G^3(X_j)}\right]\left[1+\frac{G(X_j)\left(G(X_j)+\hat{G}(X_j)\right)}{2\hat{G}^2(X_j)}-1\right]$$

$$= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{Z_{ij}^{k} U_{ij}}{\eta_i \eta_j} \left[\frac{G(X_j) - \hat{G}(X_j)}{G^3(X_j)} \right] [1 + o_p(1)]$$

Sous les hypothèses C et D, on a :

$$\left| \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{Z_{ij}^{k} U_{ij}}{\eta_i \eta_j} \left[\frac{G(X_j) - \hat{G}(X_j)}{G^3(X_j)} \right] \right| \leq \frac{2c_1}{c_2^3 c_3^2} sup_{u\in[0,\tau]} |\hat{G}(u) - G(u)| \qquad (2.6)$$

$$= O_p(1)$$

D'où

$$\Psi_{n,2}^{k} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{Z_{ij}^{k} U_{ij}}{\eta_i \eta_j} \left[\frac{G(X_j) - \hat{G}(X_j)}{G^3(X_j)} \right] + o_p(1)$$

L'inégalité (2.6) et la convergence uniforme en probabilité de l'estimateur de Kaplan-Meier (Fleming et Harrington (1991)) entraînent $\Psi_{n,2}^{k} \xrightarrow{p} 0$ lorsque n tend vers l'infini.

Considérons maintenant la *k*-ème composante de $\Psi_{n,3}$ pour $k = 1, \dots, p$.

Sous les hypothèses B,C et D :

$$\left|\Psi_{n,3}^{k}\right| = \left| \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{Z_{ij} U_{ij}}{\widehat{G}^2(X_j)} \left\{ \frac{1}{\hat{\eta}_i \hat{\eta}_j} - \frac{1}{\eta_i \eta_j} \right\} \right|$$

$$\leq \frac{1}{n(n-1)} \frac{c_1}{c_3^4} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \frac{1}{\widehat{G}^2(\tau)} \left| \eta_i \eta_j - \hat{\eta}_i \hat{\eta}_j \right| \qquad (2.7)$$

Par un développement de $\hat{\eta}_i := \mathbb{P}(R_i = 1 | Z_i; \hat{\gamma}_n)$ et de $\hat{\eta}_j := \mathbb{P}(R_j = 1 | Z_j; \hat{\gamma}_n)$ au voisinage de γ_0 nous avons :

$$\eta_i \eta_j - \hat{\eta}_i \hat{\eta}_j = (\hat{\gamma}_n - \gamma_0)^{\intercal} \left(\eta_i \dot{\eta}_j(\tilde{\gamma}_n) + \eta_j \dot{\eta}_i(\tilde{\gamma}_n) \right) + (\hat{\gamma}_n - \gamma_0)^{\intercal} \dot{\eta}_i(\tilde{\gamma}_n) (\hat{\gamma}_n - \gamma_0)^{\intercal} \dot{\eta}_j(\tilde{\gamma}_n)$$

Où $\dot{\eta}_i(\gamma) = \partial \eta_i(\gamma) / \partial \gamma$ et $\tilde{\gamma}_n \xrightarrow{p} \gamma_0$ lorsque n tend vers l'infini.

En utilisant les inégalités triangulaire et de Cauchy-Schwarz, nous obtenons :

$$\left|\eta_i\eta_j - \hat{\eta}_i\hat{\eta}_j\right| \leq \|\hat{\gamma}_n - \gamma_0\|\left\|\eta_i\dot{\eta}_j(\tilde{\gamma}_n) + \eta_j\dot{\eta}_i(\tilde{\gamma}_n)\right\| + \|\hat{\gamma}_n - \gamma_0\|^2\|\dot{\eta}_i(\tilde{\gamma}_n)\|\left\|\dot{\eta}_j(\tilde{\gamma}_n)\right\|$$

Où $\|.\|$ désigne la norme euclidienne sur $\mathbb{R}^p$.

Sous les conditions B et C, il existe une constante $c_4 < \infty$ telle que $\left\|\dot{\eta}_j(\tilde{\gamma}_n)\right\| < c_4$ pour tout $i = 1, \dots, n$, et donc :

$$\left|\eta_i\eta_j - \hat{\eta}_i\hat{\eta}_j\right| \leq 2c_4\|\hat{\gamma}_n - \gamma_0\| + c_4^2\|\hat{\gamma}_n - \gamma_0\|^2$$

En remplaçant dans (2.7), on obtient :

$$\left|\Psi_{n,3}^k\right| \leq \frac{c_1}{c_3^4}\frac{1}{G^2(\tau) + o_p(1)}(2c_4\|\hat{\gamma}_n - \gamma_0\| + c_4^2\|\hat{\gamma}_n - \gamma_0\|^2) \qquad (2.8)$$

Et donc $\Psi_{n,3}^k \xrightarrow{p} 0$ lorsque n tend vers l'infini.

Finalement, pour tout $\beta \in \mathcal{B}$ et lorsque n tend vers l'infini, $n^{-2}\Psi_n(\beta)$ converge en probabilité vers :

$$\Psi(\beta) \coloneqq \mathbb{E}\left[Z_{12}\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2} - \xi(Z_{12}^{\intercal}\beta)\right\}\right]$$

Nous montrons maintenant que $\Psi(\beta_0) = 0$.

Sous l'hypothèse A et par indépendance de C et Z, on a :

$$\mathbb{E}\left[Z_{12}\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}\right] = \mathbb{E}\left[\mathbb{E}\left[Z_{12}\frac{1(T_2 \leq C_2)1(X_1 \geq X_2)R_1R_2}{G^2(X_2)\eta_1\eta_2}\right]|Z_1, Z_2, T_2\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[Z_{12}\frac{1(T_2 \leq C_2)1(T_1 \geq T_2)1(C_1 \geq T_2)R_1R_2}{G^2(T_2)\eta_1\eta_2}\right]|Z_1, Z_2, T_2\right]$$

$$= \mathbb{E}\left[Z_{12}\frac{G(T_2)G(T_2)}{G^2(T_2)}\mathbb{E}[1(T_1 \geq T_2)|Z_1, Z_2, T_2]\right]$$

$$= \mathbb{E}[Z_{12}1(T_1 \geq T_2)]$$

$$= \mathbb{E}[Z_{12}\mathbb{E}[1(T_1 \geq T_2)|Z_1, Z_2]]$$

Or $\mathbb{E}[1(T_1 \geq T_2)|Z_1, Z_2] = \xi(Z_{12}^{\intercal}\beta_0)$ (paragraphe 2.2) et donc :

$$\mathbb{E}\left[Z_{12}\frac{U_{12}}{G^2(X_2)\eta_1\eta_2}\right] = \mathbb{E}[Z_{12}\xi(Z_{12}^{\intercal}\beta_0)]$$

Et donc $\Psi(\beta_0) = 0$. Ainsi :

$$n^2\Psi_n(\beta_0) \xrightarrow{p} 0$$

Preuve de la condition 2 :

Notons que :

$$\partial\Psi_n(\beta)/\partial\beta^{\intercal} = -\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}Z_{ij}^{\intercal}\dot{\xi}(Z_{ij}^{\intercal}\beta)$$

Sous la condition E, $\partial\Psi_n(\beta)/\partial\beta^{\intercal}$ est continue sur $\mathcal{B}$.

Preuve de la condition 3 :

Pour tout $\beta \in \mathcal{B}$, on a

$$-n^2\,\partial\Psi_n(\beta)/\partial\beta^{\intercal} = \frac{(n-1)}{n}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} Z_{ij}Z_{ij}^{T}\dot{\xi}(Z_{ij}^{\intercal}\beta)$$

qui est une U-statistique et d'après la loi des grands nombres pour les U-statistiques (Hoeffding(1961)), lorsque n tend vers l'infini :

$$-n^2\,\partial\Psi_n(\beta)/\partial\beta^{\intercal} \xrightarrow{p} A(\beta) = \int_{z_1,z_2} z_{12}z_{12}^{\intercal}\dot{\xi}(z_{12}^{\intercal}\beta)dF_z(z_1)dF_z(z_2)$$

avec F_z désigne la fonction de répartition de Z.

La convergence uniforme en probabilité sur $\mathcal{B}$ de $-n^2\,\partial\Psi_n(\beta)/\partial\beta^\intercal$ vers $A(\beta)$ est obtenue par application de la loi des grands nombres uniforme pour les U-statistiques (corollaire 4.1 de Newey (1991)). D'après la condition F, $A(\beta_0)$ est définie positive.

Les trois conditions du théorème de Foutz (Foutz(1977)) sont vérifiées donc $\hat{\beta}_n$ converge en probabilité vers β_0 lorsque n tend vers l'infini.

Pour monter le résultat (b) du théorème 2.4.1, nous considérons le développement de $\Psi_n(\beta)$ au voisinage de β_0 :

$$\Psi_n(\hat{\beta}_n) = \Psi_n(\beta_0) + \left(\partial\Psi_n(\tilde{\beta}_n)/\partial\beta^\intercal\right)(\hat{\beta}_n - \beta_0)$$

où $\tilde{\beta}_n \xrightarrow{p} \beta_0$ lorsque n tend vers l'infini. Donc :

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left(-n^2\,\partial\Psi_n(\tilde{\beta}_n)/\partial\beta^\intercal\right)^{-1} n^{-3/2}\Psi_n(\beta_0).$$

Or

$$\begin{aligned}\left|-n^2\,\partial\Psi_n(\tilde{\beta}_n)/\partial\beta^\intercal - A(\beta_0)\right| &\leq \left|-n^2\,\partial\Psi_n(\tilde{\beta}_n)/\partial\beta^\intercal - A(\tilde{\beta}_n)\right| + \left|A(\tilde{\beta}_n) - A(\beta_0)\right| \\ &\leq sup_{\beta\in\mathcal{B}}\left|-n^2\,\partial\Psi_n(\tilde{\beta}_n)/\partial\beta^\intercal - A(\tilde{\beta}_n)\right| + \left|A(\tilde{\beta}_n) - A(\beta_0)\right| \\ &\leq o_p(1) + o_p(1)\end{aligned}$$

Par la condition F, $A(\beta_0)$ est inversible et on obtient donc :

$$\left(-n^2\,\partial\Psi_n(\tilde{\beta}_n)/\partial\beta^\intercal\right)^{-1} \xrightarrow{p} A(\beta_0)^{-1}$$

Ainsi $\sqrt{n}(\hat{\beta}_n - \beta_0)$ et $A(\beta_0)^{-1}n^{-3/2}\Psi_n(\beta_0)$ convergent vers la même limite. Puisque :

$$n^{-3/2}\Psi_n(\beta_0) = \frac{(n-1)}{n}\left[\sqrt{n}\Psi_{n,1}(\beta_0) + \sqrt{n}\Psi_{n,2} + \sqrt{n}\Psi_{n,3}\right]$$

et d'après le théorème central limite pour les U-statistiques (Kowalski et Tu(2007)) alors $\sqrt{n}\Psi_{n,1}(\beta_0)$ converge en loi vers un vecteur gaussien centré de matrice de variance-

covariance $\mathbb{E}[\ell(\vartheta_1,\vartheta_2)\ell(\vartheta_1,\vartheta_2')^\intercal]$, où $\ell(\vartheta_1,\vartheta_2) = Z_{12}\left\{\frac{U_{12}}{G^2(X_2)\eta_1\eta_2} - \xi(Z_{12}^\intercal\beta_0)\right\}$ et $\vartheta_1,\vartheta_2,\vartheta_2'$ sont des répliques indépendantes de $\vartheta_i \coloneqq (X_iR_i, \Delta_iR_i, R_i, Z_i)$.

On montre à partir de (2.6) et (2.8) que $\sqrt{n}\Psi_{n,2} = O_p(1)$ et $\sqrt{n}\Psi_{n,3} = O_p(1)$.

Finalement $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = O_p(1)$.

Remarque 2.3 :

Les termes $\sqrt{n}\Psi_{n,2}$ et $\sqrt{n}\Psi_{n,3}$ apparaissent lorsque G et η_i sont inconnus et doivent être estimés. D'après la démonstration ci-dessus, si G et η_i étaient connus, $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ convergerait en loi vers un vecteur gaussien centré de variance

$$A(\beta_0)^{-1}\mathbb{E}[\ell(\vartheta_1,\vartheta_2)\ell(\vartheta_1,\vartheta_2')^\intercal]A(\beta_0)^{-1}.$$

Si G et η_i sont correctement estimés, il est raisonnable de penser que $\hat{\beta}_n$ sera également asymptotiquement gaussien (ce point sera examiné dans l'étude de simulation de la section 2.6).

2.5. Estimation de la distribution conditionnelle de T sachant Z

Notons $S_z(t) \coloneqq \mathbb{P}(T > t|Z)$ la fonction de survie conditionnelle de T sachant Z. On montre $S_Z(t) = 1 - F_\varepsilon(\beta_0^\intercal Z + h(t))$ (chapitre 1)

Pour estimer $S_Z(.)$ et donc estimer $h(.)$, nous proposons le lemme suivant :

Lemme 2.1:

Sous l'hypothèse A et les hypothèses d'indépendance conditionnelle de T et C et

d'indépendance de C et Z, on a :

$$\mathbb{E}\left[\frac{R1(X>t)}{\eta G(t)} - \left(1 - F_\varepsilon(\beta_0^\intercal Z + h(t))\right)\right] = 0$$

Démonstration

$$\mathbb{E}\left[\frac{R1(X>t)}{\eta G(t)}|Z\right] = \frac{1}{\eta G(t)}\mathbb{E}[R1(X>t)|Z]$$

$$= \frac{1}{\eta G(t)}\mathbb{E}[R|Z]\mathbb{E}[1(X>t)|Z]$$

$$= \frac{1}{G(t)}\mathbb{E}[1(C>t|Z)]\mathbb{E}[1(T>t)|Z]$$

$$= \mathbb{P}[T>t|Z]$$

$$= \quad 1 - F_\varepsilon(\beta_0^\intercal Z + h(t))$$

En prenant l'espérance de chaque côté, on obtient le résultat.

Au vu de ce lemme, un estimateur naturel de $h(t)$ peut être construit comme la solution $\hat{h}(t)$ de l'équation d'estimation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i 1(X_i>t)}{\hat{\eta}_i \hat{G}(t)} - \left(1 - F_\varepsilon\left(\hat{\beta}_n^\intercal Z_i + h(t)\right)\right) = 0$$

Un estimateur de $S_{z^*}(t)$ est :

$$\hat{S}_{z^*}(t) = 1 - F_\varepsilon\left(\hat{\beta}_n^\intercal z^* + \hat{h}(t)\right)$$

2.6. Etude de simulation :

Nous considérons le modèle de transformation linéaire $e(T) = -\beta_0{}^\intercal Z + \varepsilon$, où ε suit une distribution à valeur extrême (on obtient ainsi un modèle de Cox à risques proportionnels). Nous considérons le cas où le modèle contient une seule covariable Z qui suit une loi

normale centrée réduite. Les valeurs $ln(1.5) \approx 0.405$ et 0 sont considérées pour β_0. La variable de censure *C* est simulée suivant une loi exponentielle dont le paramètre $\lambda > 0$ est choisi de telle manière d'avoir 15% et 30% de données censurées. La fonction de survie *G* de la censure est estimée par l'estimateur de Kaplan-Meier. Les indicatrices de données manquantes $R_i, i = 1, \dots, n$ sont simulées suivant une loi de Bernoulli de paramètre $\mathbb{P}(R = 1|Z = z) = exp(\theta_0 + \theta_1 Z) / \big(1 + exp(\theta_0 + \theta_1 Z)\big)$ où θ_0 et θ_1 sont choisis pour produire des pourcentages de 15% et 30% de données manquantes. Un modèle de régression logistique est posé pour (R_i, Z_i), qui permet d'estimer les probabilités $\eta_i = \mathbb{P}(R_i = 1|Z_i)$ par $\hat{\eta}_i = exp(\hat{\gamma}_0 + \hat{\gamma}_1 Z) / \big(1 + exp(\hat{\gamma}_0 + \hat{\gamma}_1 Z)\big)$.

Pour des tailles d'échantillon faible $n = 75$ et modérée $n = 150$, deux méthodes d'estimation sont utilisées: la méthode proposée dans ce travail, et la méthode dite "cas complet"(CC), qui consiste à retirer de l'analyse les observations *i* pour lesquelles le couple (X_i, Δ_i) est manquant. Le cas où tous les couples (X_i, Δ_i) sont observés est également considéré. Ce cas (idéal dans les applications) fournira en effet la valeur de référence à laquelle nous comparerons les estimateurs $\hat{\beta}_n$ et CC (noté $\hat{\beta}_{n,CC}$). Dans ce cas, β_0 est estimé en utilisant les équations de Cheng (1995) avec $\omega(.) = 1$.

Pour chaque combinaison des différents paramètres de l'étude (taille d'échantillon, pourcentages de durées censurées et de données manquantes), nous simulons N = 500 échantillons. Pour chaque échantillon simulé $j, j = 1, \dots, N$, nous calculons les estimateurs $\hat{\beta}_n^{(j)}$et $\hat{\beta}_{n,CC}^{(j)}$ de β_0. Les valeurs moyennes $N^{-1} \sum_{j=1}^{N} \hat{\beta}_n^{(j)}$ et $N^{-1} \sum_{j=1}^{N} \hat{\beta}_{n,CC}^{(j)}$ et variances empiriques de ces estimations sont données dans la Table 2.1 (lignes intitulées "moyenne" et "variance"). Pour $\beta_0 \neq 0$ (respectivement $\beta_0 = 0$), nous obtenons la puissance empirique (respectivement le niveau empirique) du test de Wald à un niveau de signification de 5%. Les résultats sont donnés dans Table 2.1 :

β_0	n	%censure	%dm	0	15		30	
			estimateur	DC	$\hat{\beta}_n$	$\hat{\beta}_{n,CC}$	$\hat{\beta}_n$	$\hat{\beta}_{n,CC}$
0.405	75	15	moyenne	0.386	0.393	0.705	0.382	0.825
			moyenne($\widehat{e.t}$)	1.193	1.355	2.712	1.481	4.719
			variance	0.019	0.027	0.152	0.031	0.259
			puissance	0.803	0.672	0.582	0.631	0.141
		30	moyenne	0.363	0.366	0.636	0.354	0.865
			moyenne($\widehat{e.t}$)	1.168	1.312	2.372	1.494	5.174
			variance	0.021	0.032	0.135	0.041	0.358
			puissance	0.756	0.689	0.653	0.545	0.116
	150	15	moyenne	0.382	0.379	0.639	0.383	0.834
			moyenne($\widehat{e.t}$)	1.218	1.382	2.411	1.591	4.564
			variance	0.011	0.016	0.062	0.026	0.129
			puissance	0.968	0.901	0.909	0.794	0.473
		30	moyenne	0.367	0.367	0.634	0.369	0.841
			moyenne($\widehat{e.t}$)	1.199	1.348	2.386	1.551	4.633
			variance	0.011	0.018	0.079	0.028	0.150
			puissance	0.956	0.874	0.885	0.764	0.501
0	75	15	moyenne	-0.003	0.002	0.003	0.010	0.015
			moyenne($\widehat{e.t}$)	1.125	1.351	1.697	1.583	2.476
			variance	0.019	0.026	0.064	0.031	0.154
			niveau	0.028	0.038	0.038	0.037	0 .015
		30	moyenne	0.002	0.003	0.002	-0.002	0.004
			moyenne($\widehat{e.t}$)	1.137	1.367	1.726	1.636	2.962
			variance	0.022	0.030	0.077	0.047	0.293
			niveau	0.048	0.056	0.047	0.072	0.016
	150	15	moyenne	-0.002	0.000	0.000	0.000	-0.007
			moyenne($\widehat{e.t}$)	1.143	1.369	1.673	1.609	2.644
			variance	0.009	0.012	0.030	0.017	0.105
			niveau	0.032	0.028	0.032	0.052	0.041
		30	moyenne	-0.002	-0.004	-0.007	-0.008	-0.028
			moyenne($\widehat{e.t}$)	1.138	1.382	1.704	1.649	2.737
			variance	0.009	0.015	0.038	0.019	0.129
			niveau	0.024	0.040	0 .044	0.052	0.028

Note : DC pour Données Complètes

Table 2.1 : Résultats des simulations

Au vu de ces résultats, l'estimateur $\hat{\beta}_n$ semble nettement supérieur à l'estimateur CC en terme de biais comme de précision. Si les performances des deux estimateurs se dégradent lorsque le pourcentage de données manquantes augmente, $\hat{\beta}_n$ semble plus robuste à cette augmentation que $\hat{\beta}_{n,CC}$.

Nous avons montré dans la section 2.4 que $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$. Comme mentionné dans la remarque 2.3, on peut de plus s'attendre à ce que $\sqrt{n}(\hat{\beta}_n - \beta_0)$ suive asymptotiquement une loi normale. La démonstration théorique de cette intuition reste un problème ouvert. Néanmoins, notre étude de simulation peut fournir des indications utiles sur ce point. Nous construisons les histogrammes et les diagrammes quantiles-quantiles (Q-Q plots) de $\hat{\beta}_n^{(j)}$et de l'estimateur $\hat{\beta}_{n,CC}^{(j)}$ pour $j = 1, \dots, N$. Ils sont donnés dans les Figures 2.1-2.8.

Il ressort de ces graphiques qu'une approximation gaussienne de la loi de $\hat{\beta}_n$ semble raisonnable, y compris lorsque la taille n de l'échantillon est relativement faible. Cette approximation semble en revanche plus discutable pour la loi de l'estimateur CC.

Pour chaque échantillon *j*, nous obtenons enfin des estimations $\widehat{e.t}(\hat{\beta}_n^j)$et $\widehat{e.t}(\hat{\beta}_{n,CC}^j)$ de l'écart-type asymptotique de $\sqrt{n}(\hat{\beta}_n^j - \beta_0)$ par la méthode de bootstrap et $\sqrt{n}(\hat{\beta}_{n,CC}^j - \beta_0)$ en utilisant l'estimateur proposé dans Cheng et *al*. (1995). Les moyennes $\overline{e.t}(\hat{\beta}_n) \coloneqq N^{-1}\sum_{j=1}^N \widehat{e.t}(\hat{\beta}_n^j)$ et $\overline{e.t}(\hat{\beta}_{n,CC}) \coloneqq N^{-1}\sum_{j=1}^N \widehat{e.t}(\hat{\beta}_{n,CC}^j)$ de ces estimations sont fournies dans la Table 1 (lignes intitulées "moyenne $\widehat{e.t}$"). Notons qu'asymptotiquement, la variance empirique de $\hat{\beta}_n$ (respectivement $\hat{\beta}_{n,CC}$) devrait être proche de $\left(\overline{e.t}(\hat{\beta}_n)\right)^2/n$ (respectivement $\left(\overline{e.t}(\hat{\beta}_{n,CC})\right)^2/n$. Ceci est presque toujours vérifié pour $\hat{\beta}_n$. Ce n'est en revanche pas le cas pour $\hat{\beta}_{n,CC}$ ce qui indique que l'estimateur de la variance asymptotique proposé dans Cheng et *al*. (1995) et adapté à la méthode CC n'est ici pas satisfaisant (en raison sans doute de la faible taille d'échantillon résultant de l'approche CC).

Pour chaque échantillon j ($j = 1, \dots, N$), la normalité asymptotique de l'estimateur de Cheng et al. (1995) permet de calculer une statistique de test de type Wald pour tester

$H_0: \beta_0 = 0$ contre $H_1: \beta_0 \neq 0$ à partir des cas complets. Une règle de décision de niveau asymptotique 5% consiste alors à rejeter H_0 si $\left|\hat{\beta}^j_{n,CC} / \widehat{e.t}(\hat{\beta}^j_{n,CC})\right| > 1.96$.

Lorsque $\beta_0 = 0.405$, on calcule à partir des N échantillons simulés la puissance empirique du test de Wald basé sur $\hat{\beta}_{n,CC}$. Lorsque $\beta_0 = 0$, on peut calculer le niveau empirique de ce test.

Nous n'avons pas démontré la normalité asymptotique de $\hat{\beta}_n$ mais encouragés par les résultats des Figures 2.1-2.8, nous calculons les puissance et niveau empiriques du test de Wald de H_0 basé sur $\hat{\beta}_n$. L'ensemble de ces résultats est donné dans la Table 1 (lignes "puissance" et "niveau"). Le niveau empirique du test basé sur $\hat{\beta}_n$ (respectivement $\hat{\beta}_{n,CC}$) varie entre 2.8% et 5.6% (respectivement 1.5% et 7.2%). On constate donc de meilleures performances pour le test basé sur $\hat{\beta}_n$. Les puissances des deux tests diminuent lorsque la proportion de données manquantes augmente. Mais cette diminution est beaucoup plus marquée pour le test basé sur $\hat{\beta}_{n,CC}$, le test construit sur $\hat{\beta}_n$ conservant une puissance relativement élevée (en comparaison du cas sans données manquantes).

L'ensemble de ces résultats indiquent que l'estimateur proposé améliore de façon notable l'estimateur CC, qui constitue pour l'instant la seule solution disponible pour estimer les paramètres du modèle de transformation linéaire en présence de données manquantes. L'estimateur $\hat{\beta}_n$ fournit des estimations pertinentes y compris lorsque la taille n de l'échantillon est faible (de l'ordre de quelques dizaines) et/ou la proportion de données manquantes (du couple (X, Δ)) est relativement élevée.

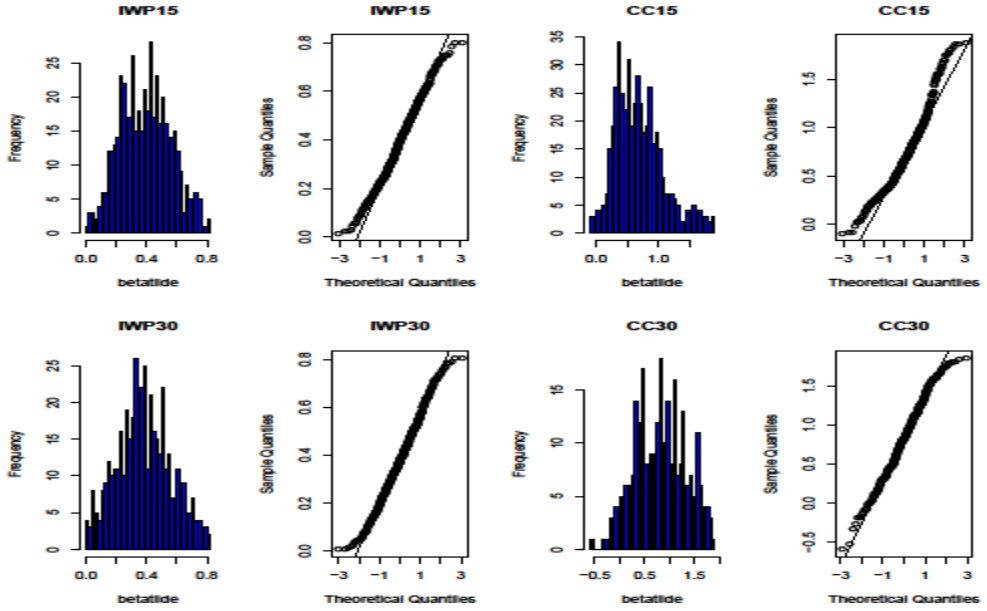

Figure2.1.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$, *j=1,..,N, pour15% et 30% de données manquantes, n=75,* $\beta_0 = 0.405$ *et 15% de censure.*

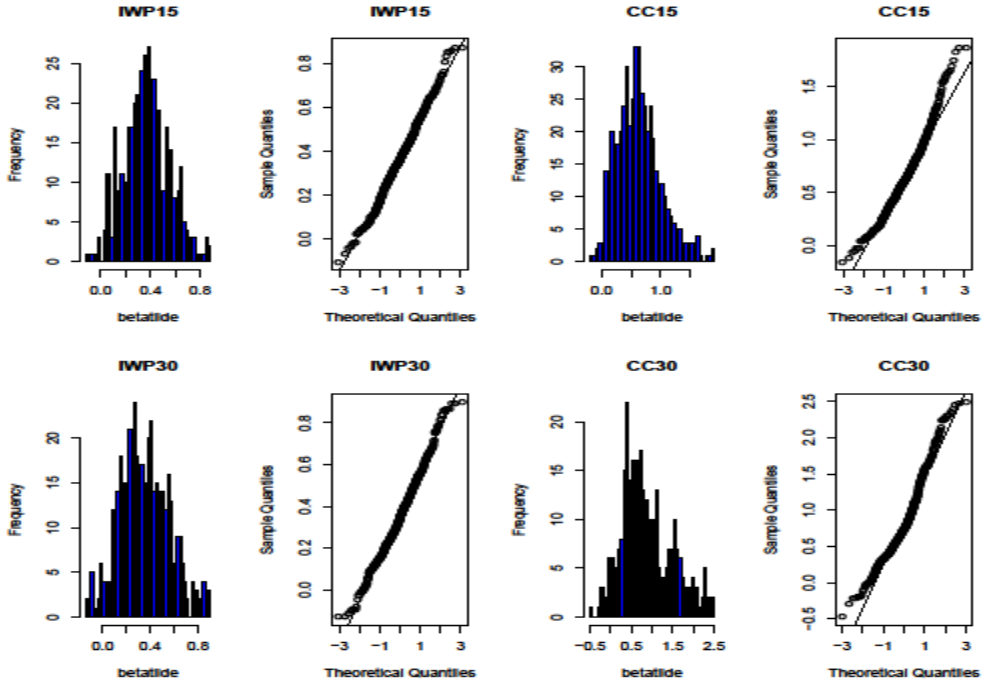

Figure2.2.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$, *j=1,..,N, n=75,* $\beta_0 = 0.405$ *et 30% de censure.*

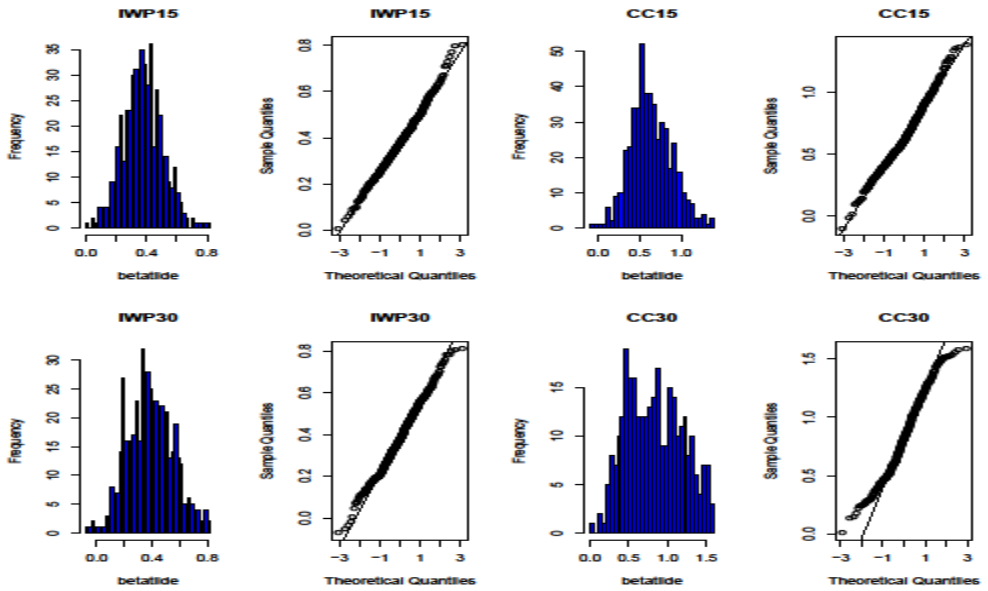

Figure2.3.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$*, j=1,..,N, n=150,* $\beta_0 = 0.405$ *et 15% de censure.*

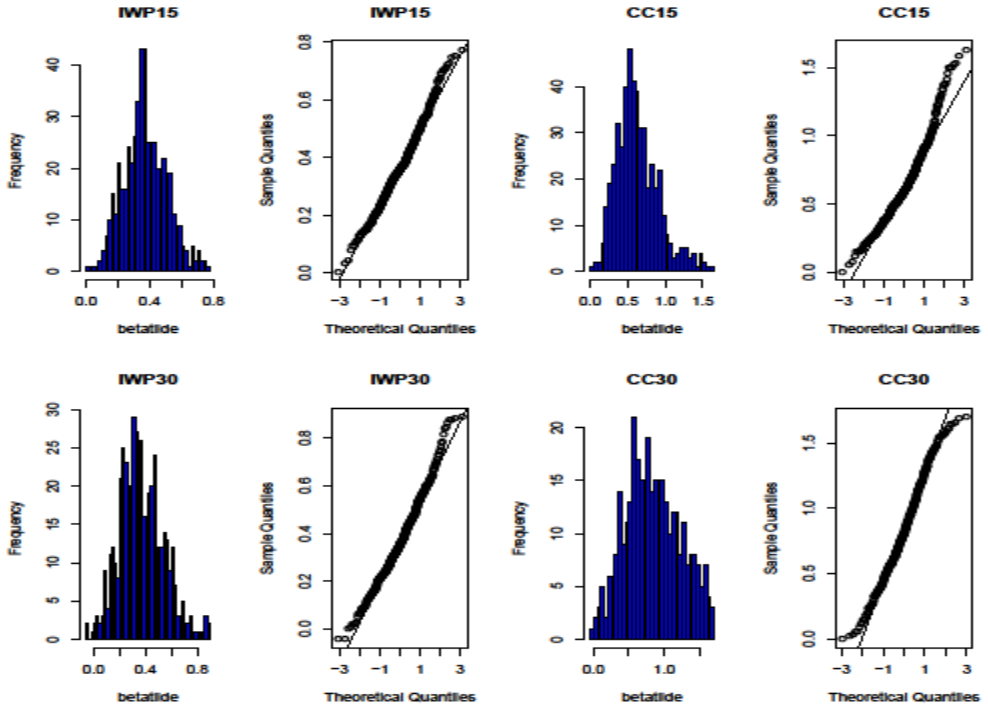

Figure2.4.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$ *j=1,..,N, n=150,* $\beta_0 = 0.405$ *et 30% de censure.*

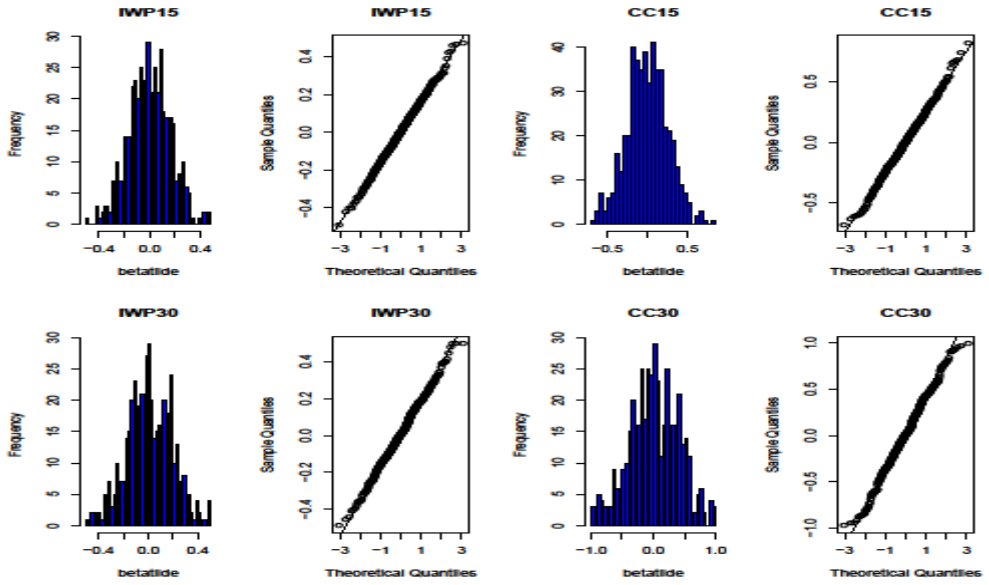

Figure2.5.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$*, j=1,..,N, n=75,* $\beta_0 = 0$ *et 15% de censure.*

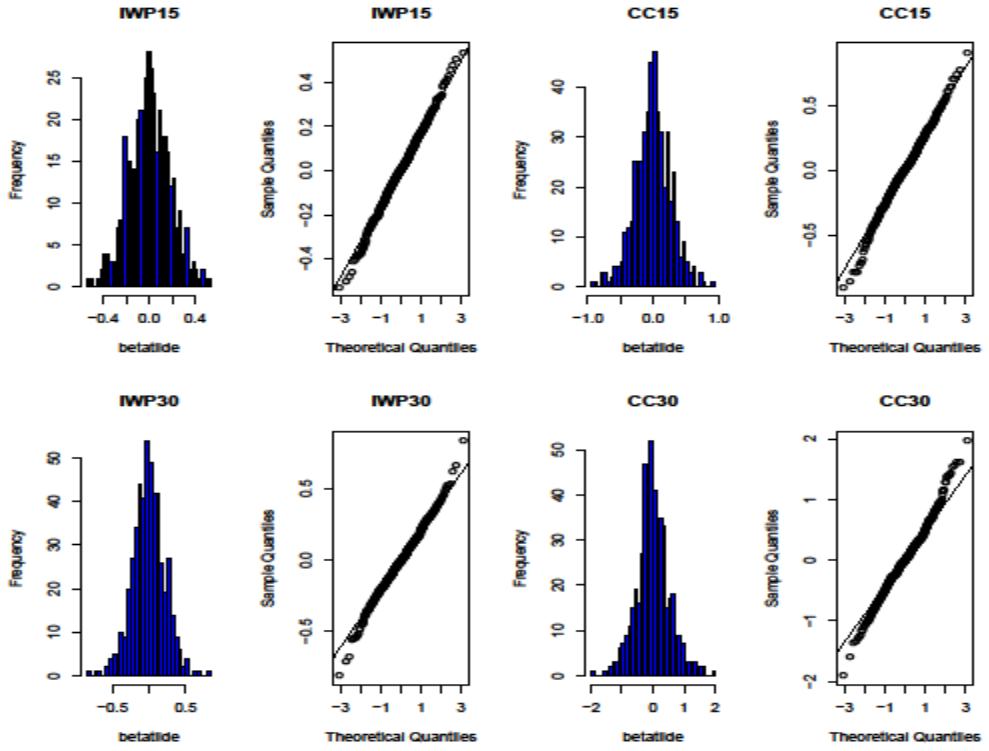

Figure2.6.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$*, j=1,..,N, n=75,* $\beta_0 = 0$ *et 30% de censure.*

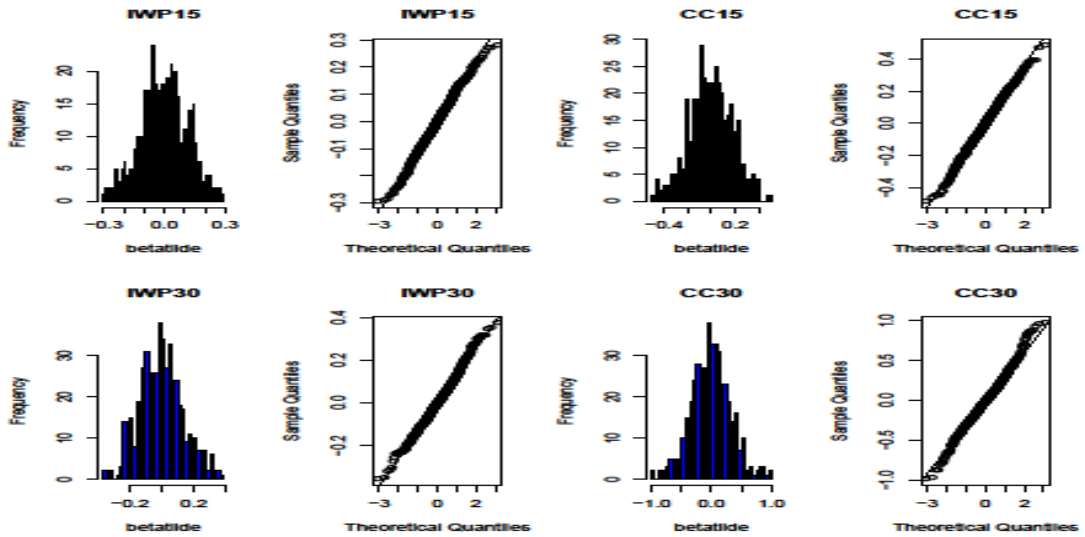

Figure2.7.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$*, j=1,..,N, n=150,* $\beta_0 = 0$ *et 15% de censure.*

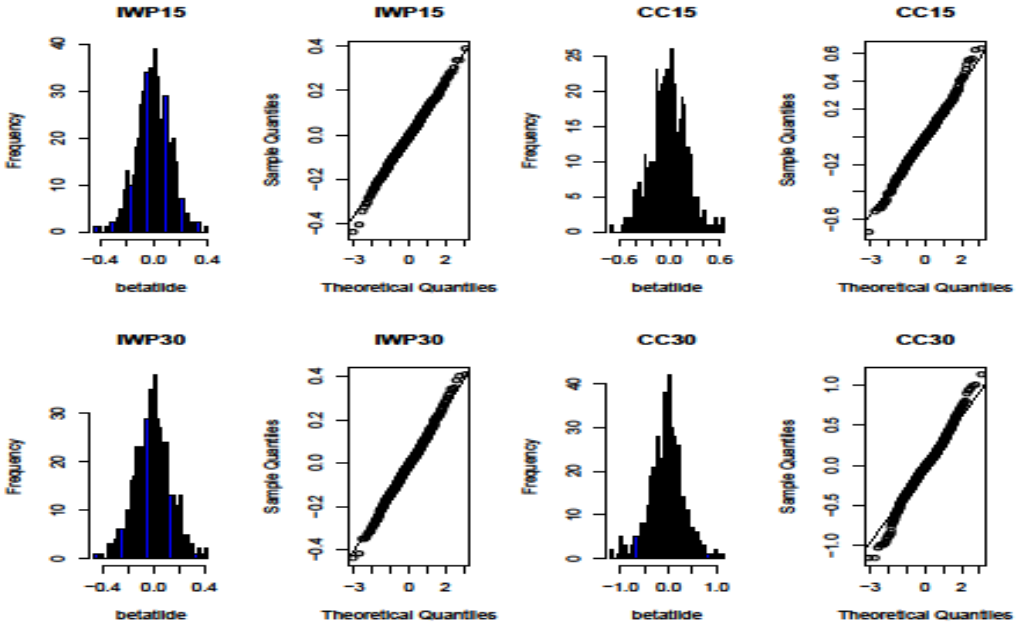

Figure2.8.*Histogrammes et QQ-plots des* $\hat{\beta}_n^j$ *et* $\hat{\beta}_{n,CC}^j$ *j=1,..,N, n=150,* $\beta_0 = 0$ *et 30% de censure.*

2.7. Exemple

Nous illustrons la méthode proposée sur un jeu de données réelles constitué de n = 142 brins de kevlar soumis chacun à un stress donné (ici la suspension d'une charge accrochée au brin). On dispose pour 108 brins seulement de la charge Z qui leur est appliquée, de la durée X jusqu'à la rupture du brin ou censure et de l'indicatrice de censure Δ. Pour les 34 autres brins, on dispose seulement de la charge qui leur est appliquée. La censure intervient soit à la fin de l'essai (au bout de 11h et 24 minutes environ) soit en raison d'un détachement de la charge sans rupture du brin (du à un défaut d'attache de la charge). Dans ce dernier cas, la censure n'est pas liée à la valeur de la charge et peut être considérée comme indépendante de Z. Nous estimons le paramètre β_0 (paramètre de régression associé à la charge Z) dans le modèle (2.1) où ε suit une loi des valeurs extrêmes. Nous obtenons $\hat{\beta}_n$ = 0.760(0.085) et $\hat{\beta}_{n,CC}$ = 0.677(0.090) (les nombres entre parenthèses sont les écarts-type estimés). Si les deux estimateurs s'accordent sur l'existence d'un effet de la charge sur le risque de rupture des brins, on note néanmoins un écart relatif de 11% dans l'appréciation quantitative de cet effet. Au vu des résultats de la section 2.6, un crédit plus

important pourra être donné à l'estimation $\hat{\beta}_n$, dont on pourra penser qu'elle reflète de manière moins biaisée que $\hat{\beta}_{n,CC}$ l'influence de la charge sur le risque de rupture.

2.8. Conclusion et discussion

Dans cet article, nous avons adapté à un problème de données manquantes les équations d'estimation proposées par Cheng et al. (1995) pour estimer le paramètre de régression d'un modèle de transformation linéaire. L'estimateur proposé repose sur le principe de la pondération par probabilité inverse. Cet estimateur est consistant et l'étude de simulation que nous avons menée suggère que sa distribution peut être approchée par une loi normale lorsque la taille de l'échantillon est suffisamment grande. De plus, cet estimateur améliore de façon significative la seule solution disponible à ce jour pour implémenter les équations d'estimation de Cheng et al. dans un contexte de données manquantes, à savoir la méthode "cas complets". Au-delà de ces résultats encourageants, plusieurs problèmes restent ouverts. Citons en particulier l'étude de la robustesse de l'estimateur proposé à une

mauvaise définition du modèle de $\mathbb{P}(R_i = 1|Z_i)$ et la construction d'estimateurs robustes de β_0. Les techniques exposées dans Tsiatis (2006) devraient pouvoir être utilisées avec profit pour résoudre cette question.

ANNEXE 2.A : *U*-STATISTIQUES

Soit $X_1, \cdots, X_m$ un échantillon de m variables indépendantes et identiquement distribuées et soit h une fonction symétrique par rapport à ses arguments qui satisfait :

$$E(|h(X_1, \cdots, X_m)|) < \infty$$

Soit θ un paramètre à estimer tel que :

$$\theta = E\big(h(X_1, \cdots, X_m)\big)$$

Alors :

$$U_n(h) = \binom{n}{m}^{-1} \sum_{(i_1,\ldots,i_m)\epsilon C_m^n} h\big(X_{i_1}, \ldots, X_{i_m}\big)$$

est un estimateur sans biais de θ où $C_m^n = \{(i_1, \ldots, i_m); i < i_1 < \cdots < i_m < n\}$ représente l'ensemble de toutes les combinaisons distinctes de m indices $(i_1, \ldots, i_m)$ choisis parmi $\{1,2, \ldots, n\}$

Définition : une *U*-statistique de noyau h et d'ordre m est définie par :

$$U_n = U_n(h) = \binom{n}{m}^{-1} \sum_{(i_1,\ldots,i_m)\epsilon C_m^n} h\big(X_{i_1}, \ldots, X_{i_m}\big)$$

Si la fonction h n'est pas symétrique, il est possible de la ramener à une fonction symétrique :

$$h(x_1, \ldots, x_m) = \frac{1}{m!} \sum_{\pi\epsilon\Pi_m} f(x_{\pi_1}, \ldots, x_{\pi_m})$$

où la somme est prise sur toutes les permutations d'un vecteur de longueur m.

Exemple :

Soit :

$$f(x_1, x_2) = x_1^2 - x_1 x_2$$

Le noyau symétrique qui correspond à cette fonction est

$$h(x_1, x_2) = \frac{1}{2}[f(x_1, x_2) + f(x_2, x_1)]$$

$$= \frac{(x_1 - x_2)^2}{2}$$

On obtient alors la *U*-statistique :

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} \frac{(X_i - X_j)^2}{2}$$

$$= \frac{1}{(n-1)} \sum_i (X_i - \bar{X})^2$$

La décomposition de Hoeffding :

On définit les espérances conditionnelles pour c =*1,2,…,m* :

$$h_c(x_1, \dots, x_c) = \mathbb{E}[h(x_1, \dots, x_c, X_{c+1}, \dots, X_m)]$$

$$h_0 = \theta$$

Avec :

$$\mathbb{E}[h_c(X_1, \dots, X_c)] = \theta$$

Les variances sont définies par :

$$\sigma_c^2 = var\{h_c(X_1, \dots, X_c)\}$$

$$\sigma_0^2 = 0$$

Soit :

$$\tilde{h}_1(x_1) = h_1(x_1)$$

Et pour $c = 2, \dots, m$

$$\tilde{h}_c(x_1, \dots, x_c) = h_c(x_1, \dots, x_c) - \sum_{j=1}^{c-1} \sum_{c,j} \tilde{h}_j \left(x_{i_1}, \dots, x_{i_j}\right) - \theta$$

Si $\tilde{H}_j, j = 1, \dots, m$ est la *U*-statistique basée sur le noyau $\tilde{h}_j$ alors la *H*-décomposition de U_n est donnée par :

$$U_n = \theta + \sum_{j=1}^{m} \binom{m}{j} \tilde{H}_j$$

Théorème 2.A.1:

Pour $j < j'$:

$$cov\left(\tilde{H}_j, \tilde{H}_{j'}\right) = 0$$

<u>La distribution asymptotique de U_n :</u>

$$var(U_n) = \binom{n}{m}^{-2} \sum_{i \epsilon C_m^n} \sum_{j \epsilon C_m^n} cov\left[h\left(X_{i_1}, \dots, X_{i_m}\right), h\left(X_{j_1}, \dots, X_{j_m}\right)\right]$$

$$= \binom{n}{m}^{-2} \sum_{i \epsilon C_m^n} \sum_{j \epsilon C_m^n} \sigma_c^2$$

$$= \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2$$

Où *c* est le nombre d'entiers communs entre $(i_1, \dots, i_m)$ et $(j_1, \dots, j_m)$

Théorème 2.A.2:

Si $\sigma_1^2 > 0$ alors :

$$\sqrt{n}(U_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, m^2\sigma_1^2)$$

ANNEXE 2.B : Preuve des résultats de Cheng et *al*.(1995)

Considérons les équations d'estimations de Cheng et *al*.(1995) :

$$U(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left(Z_{12}'\beta\right) Z_{ij} \left\{ \frac{\Delta_j 1\left(X_i \geq X_j\right)}{\hat{G}^2\left(X_j\right)} - \xi\left(Z_{ij}^{\intercal}\beta\right) \right\} = 0 \qquad (2.B.1)$$

Nous montrons tout d'abord la consistance de l'estimateur de Cheng et *al*.(1995), et ensuite la normalité asymptotique de cet estimateur.

2.B.1 : Consistance de $\hat{\beta}_n$

Soient $z_{12} = z_1 - z_2$ et H la fonction de répartition de Z.

D'après la loi des grands nombres, on a:

$$\frac{1}{n^2} U_n^{\intercal}(\beta)(\beta - \beta_0) \xrightarrow{\mathcal{P}} \mathbb{E}\left(\omega(Z_{12}^{\intercal}\beta) Z_{12}^{\intercal}(\beta - \beta_0) \left\{ \frac{\Delta_2 1(X_1 \geq X_2)}{G^2(X_2)} - \xi(Z^{\intercal}{}_{12}\beta) \right\} \right)$$

$$\mathbb{E}\left(\omega(Z^{\intercal}{}_{12}\beta) Z^{\intercal}{}_{12}(\beta - \beta_0) \left\{ \frac{\Delta_2 1(X_1 \geq X_2)}{G^2(X_2)} - \xi(Z^{\intercal}{}_{12}\beta) \right\} \right)$$

$$= \mathbb{E}\left(\omega(Z^{\intercal}{}_{12}\beta) Z^{\intercal}{}_{12}(\beta - \beta_0) \frac{\Delta_2 1(X_1 \geq X_2)}{G^2(X_2)} \right) - \mathbb{E}\left(\omega(Z^{\intercal}{}_{12}\beta) Z^{\intercal}{}_{12}(\beta - \beta_0) \xi(Z^{\intercal}{}_{12}\beta) \right)$$

Et

$$\mathbb{E}\left(\omega(Z^{\intercal}{}_{12}\beta) Z^{\intercal}{}_{12}(\beta - \beta_0) \frac{\Delta_2 1(X_1 \geq X_2)}{G^2(X_2)} \right)$$

$$= \mathbb{E}\left(\mathbb{E}\left[\omega(Z^{\intercal}{}_{12}\beta) Z^{\intercal}{}_{12}(\beta - \beta_0) \frac{\Delta_2 1(X_1 \geq X_2)}{G^2(X_2)} \right] | Z_1, Z_2 \right)$$

$$= \mathbb{E}\left(\omega(Z^{\intercal}{}_{12}\beta)Z^{\intercal}{}_{12}(\beta-\beta_0)\mathbb{E}\left[\frac{\Delta_2 1(X_1 \geq X_2)}{G^2(X_2)}\right]|Z_1, Z_2\right)$$

$$= \mathbb{E}\big(\omega(Z^{\intercal}{}_{12}\beta)Z^{\intercal}{}_{12}(\beta-\beta_0)\xi(Z^{\intercal}{}_{12}\beta)\big)$$

Et donc

$$\mathbb{E}\left(\omega(Z^{\intercal}{}_{12}\beta)Z^{\intercal}{}_{12}(\beta-\beta_0)\left\{\frac{\Delta_2 1(X_1 \geq X_2)}{G^2(X_2)} - \xi(Z^{\intercal}{}_{12}\beta)\right\}\right)$$

$$= \mathbb{E}[\omega(Z^{\intercal}{}_{12}\beta)Z^{\intercal}{}_{12}(\beta-\beta_0)\{\xi(Z^{\intercal}{}_{12}\beta) - \xi(Z^{\intercal}{}_{12}\beta_0)\}]$$

$$= \int_{z_1,z_2} \omega(z^{\intercal}{}_{12}\beta)z^{\intercal}{}_{12}(\beta-\beta_0)\left\{\xi\left(z^{\intercal}{}_{12}{}_{12}\beta\right) - \xi(z^{\intercal}{}_{12}\beta_0)\right\} dH(z_1)dH(z_2)$$

Ce qui signifie que $\frac{1}{n^2}U_n^{\intercal}(\beta)(\beta-\beta_0)$ converge vers : $\int_{z_1,z_2} \omega(z_{12}^{\intercal}\beta)z_{12}^{\intercal}(\beta-\beta_0)\{\xi(z^{\intercal}{}_{12}\beta) - \xi(z^{\intercal}{}_{12}\beta_0)\}\, dH(z_1)dH(z_2)$

Cette limite est nulle seulement si $\beta = \beta_0$ d'après la théorie des Z-estimateurs (Van der Vaart (1998)). On en déduit que $\hat{\beta}_n$ converge en probabilité vers β_0 .

2.B.2 : Normalité asymptotique de $\hat{\beta}_n$

Le développement de $U_n(\beta)$ en série de Taylor autour de β_0 donne :

$U_n(\beta) = U_n(\beta_0) + (\beta-\beta_0)\dot{U}_n(\beta^*)$ avec $\beta^* \in \left[\hat{\beta}_n, \beta_0\right]$ et $\dot{U}$ représente la dérivée de U.

On a alors

$$U_n\big(\hat{\beta}_n\big) = U_n(\beta_0) + \big(\hat{\beta}-\beta_0\big)\dot{U}_n(\beta^*)$$

$$0 = \sqrt{n}U_n(\beta_0) + \sqrt{n}(\beta-\beta_0)\dot{U}_n(\beta^*)$$

$$\sqrt{n}\big(\hat{\beta}_n-\beta_0\big) = n^{-\frac{3}{2}}U_n(\beta_0)\left[-\frac{1}{n^2}\dot{U}_n(\beta^*)\right]^{-1}$$

$$= n^{-\frac{3}{2}} U_n(\beta_0) \left[-\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Z^{\intercal}{}_{ij} \dot{\omega}\left(Z^{\intercal}{}_{ij}\beta^*\right) Z_{ij} \left\{ \frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)} - \xi\left(Z^{\intercal}{}_{ij}\beta^*\right) \right\} \right]$$
$$+ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega(Z^{\intercal}{}_{ij}\beta^*)\, \dot{\xi}\left(Z^{\intercal}{}_{ij}\beta^*\right) Z_{ij} Z^{\intercal}{}_{ij}$$

On a

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Z^{\intercal}{}_{ij} \dot{\omega}\left(Z^{\intercal}{}_{ij}\beta^*\right) Z_{ij} \left\{ \frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)} - \xi\left(Z^{\intercal}{}_{ij}\beta^*\right) \right\}$$
$$\xrightarrow{\mathcal{P}} \mathbb{E}\left\{ Z^{\intercal}{}_{ij} \dot{\omega}\left(Z^{\intercal}{}_{ij}\beta^*\right) Z_{ij} \left\{ \frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)} - \xi\left(Z^{\intercal}{}_{ij}\beta^*\right) \right\} = 0 \right\}$$

Et

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega(Z^{\intercal}{}_{ij}\beta^*)\, \dot{\xi}\left(Z^{\intercal}{}_{ij}\beta^*\right) Z_{ij} Z^{\intercal}{}_{ij} \xrightarrow{\mathcal{P}} \mathbb{E}\left(\omega(Z^{\intercal}{}_{ij}\beta^*) \dot{\xi}\left(Z^{\intercal}{}_{ij}\beta^*\right) Z_{ij} Z^{\intercal}{}_{ij}\right) = \Lambda^{-1}$$

Alors, on a

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = n^{-\frac{3}{2}} U_n(\beta_0) \Lambda^{-1} + n^{-\frac{3}{2}} U_n(\beta_0) \times o_p(1)$$

Pour montrer la normalité asymptotique de $\hat{\beta}_n$, il suffit de montrer la convergence de $n^{-\frac{3}{2}} U_n(\beta_0)$ vers une loi normale.

2.B.2.1: Normalité asymptotique de $n^{-\frac{3}{2}} U_n(\beta_0)$

On peut remarquer que $n^{-\frac{3}{2}} U_n(\beta_0)$ est une *U*-statistique qui, suivant le théorème central limite pour les *U*-statistiques, est asymptotiquement normale avec moyenne nulle et matrice de variance-covariance Σ (Kowalski et Tu (2007)). On en déduit :

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

La matrice Σ est estimée par $\hat{\Sigma} = \hat{\Lambda}^{-1} \hat{\Gamma} \hat{\Lambda}'^{-1}$ où

$$\Lambda^{-1} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega({Z^{\intercal}}_{ij}\hat{\beta}_n)\,\dot{\xi}\left({Z^{\intercal}}_{ij}\hat{\beta}_n\right)Z_{ij}{Z^{\intercal}}_{ij}$$

et $\hat{\Gamma}$ est l'estimateur de Γ la variance de la loi limite de $n^{-\frac{3}{2}}U_n(\beta_0)$. Pour déterminer $\hat{\Gamma}$, une approximation de $n^{-\frac{3}{2}}U_n(\beta_0)$ est donnée dans ce qui suit.

2.B.2.2 : Approximation de $n^{-\frac{3}{2}}U_n(\beta_0)$

Ecrivons le terme $U_n(\beta_0)$ sous la forme suivante :

$$U_n(\beta_0) = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left({Z^{\intercal}}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{\hat{G}^2(X_j)} - \xi\left({Z^{\intercal}}_{ij}\beta_0\right)\right\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left({Z^{\intercal}}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{\hat{G}^2(X_j)} - \xi\left({Z^{\intercal}}_{ij}\beta_0\right)\right\}$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left({Z^{\intercal}}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} - \xi\left({Z^{\intercal}}_{ij}\beta\right)\right\}$$

$$- \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left({Z^{\intercal}}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} - \xi\left({Z^{\intercal}}_{ij}\beta_0\right)\right\}$$

$$= \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left({Z^{\intercal}}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} - \xi\left({Z^{\intercal}}_{ij}\beta_0\right)\right\}}_{A}$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left({Z^{\intercal}}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{\hat{G}^2(X_j)}\right\}$$

$$- \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left({Z^{\intercal}}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)}\right\}$$

$$= A + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left(Z^{\intercal}{}_{ij}\beta_0\right) Z_{ij}\Delta_j 1\left(X_i \geq X_j\right)\left\{\frac{1}{\hat{G}^2\left(X_j\right)} - \frac{1}{G^2\left(X_j\right)}\right\}$$

$$= A + \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left(Z^{\intercal}{}_{ij}\beta_0\right) Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)}\left\{\frac{G^2\left(X_j\right) - \hat{G}^2\left(X_j\right)}{\hat{G}^2\left(X_j\right)}\right\}$$

$$= A + 2\sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left(Z^{\intercal}{}_{ij}\beta_0\right) Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)}\frac{G\left(X_j\right) - \hat{G}\left(X_j\right)}{G\left(X_j\right)}\left(\frac{1}{2}\right)\frac{G\left(X_j\right)\left(G\left(X_j\right) + \hat{G}\left(X_j\right)\right)}{\hat{G}^2\left(X_j\right)}$$

$$= A + 2\sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left(Z^{\intercal}{}_{ij}\beta_0\right) Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)}\frac{G\left(X_j\right) - \hat{G}\left(X_j\right)}{G\left(X_j\right)}\left\{1 + \frac{G\left(X_j\right)\left(G\left(X_j\right) + \hat{G}\left(X_j\right)\right)}{2\hat{G}^2\left(X_j\right)} - 1\right\}$$

$$= A + 2\sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left(Z^{\intercal}{}_{ij}\beta_0\right) Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)}\frac{G\left(X_j\right) - \hat{G}\left(X_j\right)}{G\left(X_j\right)}\left\{1 + o_p(1)\right\}$$

$$= A + 2B\left(1 + o_p(1)\right)$$

avec

$$B = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega\left(Z^{\intercal}{}_{ij}\beta_0\right) Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)}\frac{G\left(X_j\right) - \hat{G}\left(X_j\right)}{G\left(X_j\right)}$$

On peut écrire donc

$$n^{-\frac{3}{2}}U_n(\beta_0) = n^{-\frac{3}{2}}A + 2n^{-\frac{3}{2}}B + 2n^{-\frac{3}{2}}B \times o_p(1)$$

Le dernier terme $2n^{-\frac{3}{2}}B \times o_p(1)$ peut s'écrire :

$$2n^{-\frac{3}{2}}B \times o_p(1) = 2\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\Delta_j\left(G\left(X_j\right) - \hat{G}\left(X_j\right)\right)}{G^3\left(X_j\right)}\frac{1}{n}\sum_{i=1}^{n}\omega\left(Z^{\intercal}{}_{ij}\beta\right) Z_{ij}1\left(X_i \geq X_j\right) o_p(1)$$

Sous la condition C, on a

$$\frac{1}{n}\sum_{i=1}^{n}\omega\left(Z^{\intercal}{}_{ij}\beta_0\right)Z_{ij}1\left(X_i \geq X_j\right) \xrightarrow{\mathcal{P}} \mathbb{E}\left[\omega\left(Z^{\intercal}{}_{ij}\beta_0\right)Z_{ij}1\left(X_i \geq X_j\right)\right] = O_p(1)$$

Et donc :

$$2n^{-\frac{3}{2}}B \times o_p(1) = 2\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\Delta_j\left(G\left(X_j\right) - \hat{G}\left(X_j\right)\right)}{G^3\left(X_j\right)}o_p(1)$$

$$= 2\frac{1}{n}\sum_{j=1}^{n}\frac{\Delta_j\sqrt{n}\left(G\left(X_j\right) - \hat{G}\left(X_j\right)\right)}{G^3\left(X_j\right)}o_p(1)$$

$$= 2\frac{1}{n}\sum_{j=1}^{n}\frac{\Delta_j}{G^3\left(X_j\right)}O_p(1)o_p(1)$$

$$= 2O_p(1)O_p(1)o_p(1) = o_p(1)$$

Car $\sqrt{n}\left(G\left(X_j\right) - \hat{G}\left(X_j\right)\right)$ converge vers un processus gaussien donc $\sqrt{n}\left(G\left(X_j\right) - \hat{G}\left(X_j\right)\right) = O_p(1)$.

Pour déterminer le terme $2n^{-\frac{3}{2}}B$, on considère la représentation de $(\hat{G} - G)/G$ (Gill,1980) :

$$\frac{G\left(X_j\right) - \hat{G}(X_j)}{G\left(X_j\right)} = \int_0^{X_j}\frac{\hat{G}(s_-)}{G(s)}\frac{J_n(s)}{\bar{Y}_n(s)}d(\bar{N}_n^c(s) - \bar{Y}_n(s)\Lambda_c(s))$$

Avec

$$\bar{Y}_n(s) = \sum_{i=1}^{n}Y_i(s) = \sum_{i=1}^{n}1(X_i \geq s)$$

$$J_n(s) = 1(\bar{Y}_n(s) > 0)$$

$$\bar{N}_n^c(s) = \sum_{i=1}^{n} N_i^c(s) = \sum_{i=1}^{n} 1(X_i \leq s, \Delta_i = 0)$$

On pose $M_k(s) = N_k^c(s) - \int_0^s Y_k(u) d\Lambda_c(u)$ qui est une martingale centrée. Alors :

$$\frac{G(X_j) - \hat{G}(X_j)}{G(X_j)} = \sum_{k=1}^{n} \int_0^{X_j} \frac{\hat{G}(s_-)}{G(s)} \frac{J_k(s)}{\bar{Y}_k(s)} d(N_k^c(s) - Y_k(s)\Lambda_c(s))$$

$$= \sum_{k=1}^{n} \int_0^{\infty} 1(X_j \geq s) \frac{\hat{G}(s_-)}{G(s)} \frac{J_k(s)}{\bar{Y}_k(s)} dM_k(s)$$

En remplaçant dans l'expression précédente de $2n^{-\frac{3}{2}}B$, on obtient :

$$2n^{-\frac{3}{2}}B = 2\frac{1}{n}\frac{1}{\sqrt{n}} \sum_{i=1}^{n}\sum_{j=1}^{n} \omega(Z^{\intercal}{}_{ij}\beta_0) Z_{ij} \frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} \left(\sum_{k=1}^{n} \int_0^{\infty} 1(X_j \geq s) \frac{\hat{G}(s_-)}{G(s)} \frac{J_k(s)}{\bar{Y}_k(s)} dM_k(s) \right)$$

$$= 2\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \int_0^{\infty} \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \omega(Z^{\intercal}{}_{ij}\beta_0) Z_{ij} \frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} 1(X_j \geq s) \frac{\hat{G}(s_-)}{G(s)} \frac{J_k(s)}{\frac{1}{n}\bar{Y}_k(s)} dM_k(s)$$

$$= 2\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \int_0^{\infty} \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{q(s)}{\pi(s)} + o_p(1) \right) dM_k(s)$$

Avec

$$\pi(s) = \lim_{n\to\infty} \frac{1}{n} \bar{Y}_k(s)$$

$$q(s) = \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \omega(Z^{\intercal}{}_{ij}\beta_0) Z_{ij} \frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} 1(X_j \geq s)$$

Donc

$$2n^{-\frac{3}{2}}B = 2\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\int_0^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{q(s)}{\pi(s)}dM_k(s) + 2\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\int_0^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}dM_k(s)\,o_p(1)$$

Puisque $\mathbb{E}\left(\int_0^{\infty} dM_1(s)\right) = 0$ (car $\int_0^{\infty} dM_1(s)$ est une martingale centrée), on a alors :

$$2\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\int_0^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}dM_k(s)\,o_p(1) = 2\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}\int_0^{\infty}dM_k(s) - \mathbb{E}\left(\int_0^{\infty}dM_1(s)\right)\right)o_p(1)$$

$$= O_p(1)o_p(1) = o_p(1)$$

Où $O_p(1)$ vient du théorème central limite.

Ce qui donne à la fin l'approximation de $2n^{-\frac{3}{2}}B$:

$$2n^{-\frac{3}{2}}B = n^{-\frac{3}{2}}A + 2n^{-\frac{1}{2}}\sum_{k=1}^{n}\int_0^{\infty}\frac{q(s)}{\pi(s)}dM_k(s) + o_p(1)$$

$$= n^{-\frac{3}{2}}A + T + o_p(1)$$

Avec : $T = 2n^{-\frac{1}{2}}\sum_{k=1}^{n}\int_0^{\infty}\frac{q(s)}{\pi(s)}dM_k(s)$

et

$$A = \sum_{i=1}^{n}\sum_{j=1}^{n}\omega\left(Z^{\intercal}{}_{ij}\beta_0\right)Z_{ij}\left\{\frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)} - \xi\left(Z^{\intercal}{}_{ij}\beta_0\right)\right\}$$

$$\equiv \sum_{i=1}^{n}\sum_{j=1}^{n}h\left(Z_i, Z_j\right)$$

Où $\mathbb{E}\left[h\left(Z_i, Z_j\right)\right] = \mathbb{E}\left(\mathbb{E}\left[h\left(Z_i, Z_j\right)\right]|Z_i, Z_j\right)$

$$= \mathbb{E}\left[0\right] = 0$$

2.B.3 : détermination de Γ

Variance asymptotique de $n^{-\frac{3}{2}}A$: Il existe un noyau $\bar{h}(Z_i, Z_j)$ symétrique (il reste invariant en permutant ses arguments) de telle manière que $n^{-\frac{3}{2}}A$ soit une *U*-statistique de noyau $\bar{h}$. En effet, posons

$$\bar{h}(Z_i, Z_j) = \frac{1}{2}\Big(h(Z_i, Z_j) + h(Z_j, Z_i)\Big)$$

$$= \frac{1}{2}\Big(\omega({Z^\intercal}_{ij}\beta_0)Z_{ij}e_{ij}(\beta_0) - \omega({Z^\intercal}_{ji}\beta_0)Z_{ji}e_{ji}(\beta_0)\Big)$$

$$= \frac{1}{2}\Big(\omega({Z^\intercal}_{ij}\beta_0)Z_{ij}e_{ij}(\beta_0) - \omega({Z^\intercal}_{ji}\beta_0)Z_{ij}e_{ji}(\beta_0)\Big)$$

$$= \frac{1}{2}\Big(\omega({Z^\intercal}_{ij}\beta_0)e_{ij}(\beta_0) - \omega({Z^\intercal}_{ji}\beta_0)e_{ji}(\beta_0)\Big)Z_{ij}$$

Où

$$e_{ij}(\beta_0) = \frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} - \xi({Z^\intercal}_{ij}\beta_0)$$

Notons que $\mathbb{E}\bar{h}(Z_i, Z_j) = 0$

On écrit $n^{-\frac{3}{2}}A$ sous la forme

$$n^{-\frac{3}{2}}A = n^{-\frac{3}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{h}(Z_i, Z_j)$$

$$= n^{-\frac{3}{2}}\sum_{i=1}^{n}\left(\sum_{j<i}^{n}\bar{h}(Z_i, Z_j) + \sum_{i=j}^{n}\bar{h}(Z_i, Z_j) + \sum_{j>i}^{n}\bar{h}(Z_i, Z_j)\right)$$

$$= n^{-\frac{3}{2}}\sum_{i=1}^{n}\sum_{j>i}^{n}\bar{h}(Z_i, Z_j)$$

$$= \frac{(n-1)}{\sqrt{n}} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i}^{n} \bar{h}(Z_i, Z_j)$$

$$= \frac{(n-1)}{n} \sqrt{n} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i}^{n} \bar{h}(Z_i, Z_j)$$

$$= \frac{(n-1)}{n} \sqrt{n}\, S_n$$

S_n est une *U*-statistique de noyau $\bar{h}$ (Annexe 2.A), qui d'après le théorème limite central est asymptotiquement gaussienne (Kowalski et Tu (2007)) avec variance asymptotique Γ_1:

$$\sqrt{n}(S_n - 0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma_1)$$

Où

$$\Gamma_1 = 4var\left[\mathbb{E}\left(\bar{h}(Z_1, Z_2)|Z_1\right)\right]$$

$$= 4\mathbb{E}\left[\mathbb{E}\left(\bar{h}(Z_1, Z_2)|Z_1\right)\right]^{\otimes 2} - 4\left(\underbrace{\mathbb{E}\left[\mathbb{E}\left(\bar{h}(Z_1, Z_2)|Z_1\right)\right]}_{0}\right)^{\otimes 2}$$

$$= 4\mathbb{E}\left[\left\{\mathbb{E}\left(\bar{h}(Z_1, Z_2)|Z_1\right)\right\}\left\{\mathbb{E}\left(\bar{h}(Z_1, Z_2)|Z_1\right)\right\}^{\intercal}\right]$$

$$= 4\mathbb{E}\left[\left\{\mathbb{E}\left(\bar{h}(Z_1, Z_2)|Z_1\right)\right\}\left\{\mathbb{E}\left(\bar{h}(Z_1, Z_3)|Z_1\right)\right\}^{\intercal}\right]$$

$$= 4\mathbb{E}\left[\mathbb{E}\left\{\bar{h}(Z_1, Z_2)\bar{h}^{\intercal}(Z_1, Z_3)|Z_1\right\}\right]$$

$$= 4\mathbb{E}\left[\bar{h}(Z_1, Z_2)\bar{h}^{\intercal}(Z_1, Z_3)\right]$$

qui peut être estimée par :

$$\frac{4}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\bar{h}(Z_i,Z_j)\bar{h}^{\intercal}(Z_i,Z_k)$$

$$=\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\{\omega({Z^{\intercal}}_{ij}\beta_0)e_{ij}(\beta_0)$$

$$-\omega({Z^{\intercal}}_{ji}\beta_0)e_{ji}(\beta_0)\}\{\omega(Z_{ik}^{\intercal}\beta_0)e_{ik}(\beta_0)\quad-\omega(Z_{ki}^{\intercal}\beta_0)e_{ki}(\beta_0)\}Z_{ij}Z_{ik}^{\intercal}$$

Variance asymptotique de T : Remarquons que dans l'expression de $T = 2n^{-\frac{1}{2}}\sum_{k=1}^{n}\int_0^{\infty}\frac{q(t)}{\pi(t)}dM_k(t)$, $\frac{q(t)}{\pi(t)}$ est un processus prévisible et $\int_0^{\infty}\frac{q(t)}{\pi(s)}dM_k(t)$ est une martingale (Martinussen et Scheike (2006)), et on a

$$T = 2n^{-\frac{1}{2}}\sum_{k=1}^{n}\int_0^{\infty}\frac{q(t)}{\pi(t)}dM_k(t)$$

$$=\sqrt{n}\left(\frac{2}{n}\sum_{k=1}^{n}\int_0^{\infty}\frac{q(t)}{\pi(t)}dM_k(t)\right)$$

En appliquant le théorème central limite pour les martingales (Rebolledo (1980)), on a

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}2\int_0^{\infty}\frac{q(t)}{\pi(t)}dM_k(t)-0\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Gamma_2)$$

Où

Γ_2 est donnée par :

$$\Gamma_2 = var\left[2\int_0^{\infty}\frac{q(t)}{\pi(t)}dM_k(t)\right]$$

$$=4\int_0^{\infty}\frac{q(t)q^{\intercal}(t)}{\pi(t)\pi^{\intercal}(t)}var[dM_k(t)]$$

$$=4\int_0^{\infty}\frac{q(t)q^{\intercal}(t)}{\pi(t)\pi^{\intercal}(t)}\mathbb{E}\left[dM_k(t)^2|\mathcal{F}_{t_-}\right]$$

$$= 4\int_0^{\infty} \frac{q(t)q^{\intercal}(t)}{\pi(t)\pi^{\intercal}(t)} \bar{Y}_k(t) d\Lambda_G(t)$$

$$\approx 4\int_0^{\infty} \frac{q(t)q^{\intercal}(t)}{\pi(t)} d\Lambda_G(t)$$

Covariance entre $n^{-\frac{3}{2}}A$ *et* T : on remarque que

$$\mathbb{E}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)}(1-\Delta_j)\frac{q^{\intercal}(X_i)}{\pi(X_i)}\right\} = \mathbb{E}\left[\mathbb{E}\left\{\frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)}(1-\Delta_j)\frac{q^{\intercal}(X_i)}{\pi(X_i)}\right\}|C_i, T_i, T_j\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left\{\frac{\Delta_j 1(C_i \geq X_j)}{G^2(X_j)} 1(T_i > C_i)\frac{q^{\intercal}(C_i)}{\pi(C_i)}\right\}|C_i, T_i, T_j\right]$$

$$= \mathbb{E}\left\{\int_0^{\infty} \frac{\Delta_j 1(t \geq X_j)}{G^2(X_j)} 1(T_i > t)\frac{q^{\intercal}(t)}{\pi(t)} dF_c\right\}$$

$$= \mathbb{E}\left\{\int_0^{\infty} \frac{\Delta_j 1(t \geq X_j)}{G^2(X_j)} 1(T_i > t)\frac{q^{\intercal}(t)}{\pi(t)} G(t) d\Lambda_G\right\}$$

$$= \mathbb{E}\left\{\int_0^{\infty} \frac{\Delta_j 1(X_i \geq X_j)}{G^2(X_j)} 1(X_j \leq t) 1(X_i \geq t)\frac{q^{\intercal}(t)}{\pi(t)} d\Lambda_G\right\}$$

Donc

$$cov\left\{n^{-\frac{3}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega(Z^{\intercal}{}_{ij}\beta_0)Z_{ij}e_{ij}(\beta_0), 2n^{-\frac{1}{2}}\sum_{k=1}^{n}\int_0^{\infty}\frac{q(t)}{\pi(t)}dM_k(t)\,|Z_i, Z_j\right\}$$

$$= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\int_0^{\infty}\omega(Z^{\intercal}{}_{ij}\beta_0)Z_{ij}e_{ij}(\beta_0)\frac{q^{\intercal}(t)}{\pi(t)}\sum_{k=1}^{n}dM_k(t)\,|Z_i, Z_j\right]$$

$$\approx \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\int_0^{\infty}\underbrace{\omega(Z^{\intercal}{}_{ij}\beta_0)Z_{ij}\frac{\Delta_j 1(X_i \geq X_j)R_iR_j}{G^2(X_j)\eta_i\eta_j}}_{H}\frac{q^{\intercal}(t)}{\pi(t)}d\{M_i(t)+M_j(t)\}\,|Z_i, Z_j\right]$$

$$\approx \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\int_0^{\infty} H\frac{q^{\intercal\prime}(t)}{\pi(t)}d\left\{N_i(t)-\int_0^{t}1(X_i\geq s)d\Lambda_G(s)+N_j(t)\right.\right.$$
$$\left.\left.-\int_0^{t}1\big(X_j\geq s\big)d\Lambda_G(s)\right\}|Z_i,Z_j\right]$$

$$\approx \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}d\big[N_i(t)+N_j(t)\big]\right.$$
$$\left.-\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}d\int_0^{t}\Big(1(X_i\geq s)+1\big(X_j\geq s\big)\Big)d\Lambda_G(s)\,|Z_i,Z_j\right]$$

$$\approx \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[H(1-\Delta_i)\frac{q^{\intercal}(X_i)}{\pi(X_i)}-\int_0^{\infty} H\frac{q^{\intercal}(X_i)}{\pi(X_i)}d\big[N_i(t)+N_j(t)\big]\right.$$
$$\left.-\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}\Big(1(X_i\geq t)+1\big(X_j\geq t\big)\Big)d\Lambda_G(t)|Z_i,Z_j\right]$$

$$\approx \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\{\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}1\big(X_j\leq t\big)1(X_i\geq t)d\Lambda_G-\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}\Big(1(X_i\geq t)\right.$$
$$\left.+1\big(X_j\geq t\big)\Big)d\Lambda_G(t)|Z_i,Z_j\right\}$$

$$\approx \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\{\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}\Big(1-1\big(X_j>t\big)\Big)1(X_i\geq t)d\Lambda_G-\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}\Big(1(X_i\geq t)\right.$$
$$\left.+1\big(X_j\geq t\big)\Big)d\Lambda_G(t)|Z_i,Z_j\right\}$$

$$\approx -\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\{\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}1\big(X_j>t\big)1(X_i\geq t)d\Lambda_G|Z_i,Z_j\right\}$$
$$-\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\{\int_0^{\infty} H\frac{q^{\intercal}(t)}{\pi(t)}\Big(1(X_i\geq t)+1\big(X_j\geq t\big)\Big)d\Lambda_G(t)|Z_i,Z_j\right\}$$

$$\approx -\frac{4}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\{\int_0^{\infty}\omega\big(Z^{\intercal}{}_{ij}\beta_0\big)Z_{ij}\frac{\Delta_j 1\big(X_i\geq X_j\big)}{G^2\big(X_j\big)}\frac{q^{\intercal}(t)}{\pi(t)}1\big(X_j\geq t\big)d\Lambda_G|Z_i,Z_j\right\}$$

$$\approx -4\int_0^{\infty} \mathbb{E}\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega\left(Z^{\intercal}{}_{ij}\beta_0\right)Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{G^2\left(X_j\right)}\frac{q'(t)}{\pi(t)}1\left(X_j \geq t\right)d\Lambda_G|Z_i, Z_j\right\}$$

$$\approx -4\int_0^{\infty}\frac{q(t)q^{\intercal}(t)}{\pi(t)}d\Lambda_G$$

On a donc :

$$\Gamma = \lim_{n\to\infty}\left[\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\left\{\omega\left(Z^{\intercal}{}_{ij}\beta_0\right)e_{ij}(\beta_0) - \omega\left(Z_{ji}^{\intercal}\beta_0\right)e_{ji}(\beta_0)\right\}\left\{\omega(Z_{ik}^{\intercal}\beta_0)e_{ik}(\beta_0)\right.\right.$$
$$\left.\left. - \omega(Z_{ki}^{\intercal}\beta_0)e_{ki}(\beta_0)\right\}Z_{ij}Z_{ik}^{\intercal} - 4\int_0^{\infty}\frac{q(t)q^{\intercal}(t)}{\pi(t)}d\Lambda_G\right]$$

Cette expression de Γ peut être estimée en remplaçant β_0, G et η_i par $\hat{\beta}_0$, $\hat{G}$ et $\hat{\eta}_i$, et en remplaçant Λ_G par son estimateur de Nelson-Aalen (Andersen (1993)).

Soit $\widehat{\Lambda}_{\mathrm{G}}(t)$ l'estimateur de Nelson-Aalen de la fonction de risque cumulé $\Lambda_{\mathrm{G}}(t)$:

$$\widehat{\Lambda}_{\mathrm{G}}(t) = \int_0^t \frac{d\bar{N}_n^c(s)}{\bar{Y}_n(s)} = \sum_{l=1}^{n}\frac{dN_l^c(t)}{\bar{Y}_n(t)} = \sum_{l=1}^{n}\frac{1(X_l \leq t)(1-\Delta_l)}{\sum_{j=1}^{n}1(X_j \geq t)}$$

On peut écrire donc :

$$\int_0^{\infty}\frac{q(t)q^{\intercal}(t)}{\pi(t)}d\widehat{\Lambda}_G = \int_0^{\infty}\frac{q(t)q^{\intercal}(t)}{\pi(t)}\sum_{l=1}^{n}\frac{dN_l^c(t)}{\bar{Y}_n(t)}$$

$$= \sum_{l=1}^{n}\frac{q(X_l)q^{\intercal}(X_l)}{\pi(X_l)\bar{Y}_n(X_l)}(1-\Delta_l)R_l$$

$$= \sum_{l=1}^{n}\frac{(1-\Delta_l)}{\frac{1}{n}\left(\bar{Y}_n(X_l)\right)^2}\left\{\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega\left(Z^{\intercal}{}_{ij}\hat{\beta}_n\right)Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{\hat{G}^2\left(X_j\right)}1\left(X_j \geq X_l\right)\right\}^{\otimes 2}$$

$$= \frac{1}{n^3}\sum_{l=1}^{n}\frac{(1-\Delta_l)}{(\sum_k 1(X_k \geq X_l))^2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega\left(Z^{\intercal}{}_{ij}\hat{\beta}_n\right)Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{\hat{G}^2\left(X_j\right)}1\left(X_j \geq X_l\right)\right\}^{\otimes 2}$$

La matrice Γ peut être donc estimée par :

$$\hat{\Gamma} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq j}^{n}\left\{\omega\left(Z^{\intercal}{}_{ij}\hat{\beta}_n\right)\hat{e}_{ij}\left(\hat{\beta}_n\right) - \omega\left(Z^{\intercal}{}_{ji}\hat{\beta}_n\right)\hat{e}_{ji}\left(\hat{\beta}_n\right)\right\}\left\{\omega\left(Z^{\intercal}{}_{ik}\hat{\beta}_n\right)\hat{e}_{ik}\left(\hat{\beta}_n\right)\right.$$
$$\left. - \omega\left(Z^{\intercal}{}_{ki}\hat{\beta}_n\right)\hat{e}_{ki}\left(\hat{\beta}_n\right)\right\}Z_{ij}Z^{\intercal}{}_{ik}$$

$$-\frac{4}{n^3}\sum_{l=1}^{n}\frac{(1-\Delta_l)}{(\sum_k 1(X_k \geq X_l))^2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega\left(Z^{\intercal}{}_{ij}\widehat{\beta}_n\right)Z_{ij}\frac{\Delta_j 1\left(X_i \geq X_j\right)}{\hat{G}^2\left(X_j\right)}1(X_j \geq X_l)\right\}^{\otimes 2}$$

Où :

$$\hat{e}_{ij}\left(\hat{\beta}_n\right) = \frac{\Delta_j 1\left(X_i \geq X_j\right)}{\hat{G}^2\left(X_j\right)} - \xi\left(Z^{\intercal}{}_{ij}\hat{\beta}_n\right)$$

et $v^{\otimes 2} = vv^{\intercal}$ pour un vecteur v.■

Annexe 2.C : PROGRAMME DE SIMULATION

Nous présentons une partie des programmes utilisés dans l'étude de simulation décrite dans la section 2.6. Cette partie correspond à 15% de données manquantes et 15% de censure

```
n=75
Z=matrix(nrow=n,ncol=p) #matrice des covariables
Z[,1]=rnorm(n)  # covariable 1
T=-exp(-Z[,1]*b0)*log(runif(n,0,1))/lambda
l=.167     # pourcentage de censure  15%
C=rexp(n,l)
matrZ=Z ; vectT=T ; vectC=C
X=pmin(T,C) ; deltac=as.integer(C<=T)
if(deltac[which.max(X)]==1)
{X[which.max(X)]=T[which.max(X)];deltac[which.max(X)]=0}
delta=as.integer(T==X)
Gkm=survfit(Surv(X,deltac)~1)  #estimateur de Kaplan-Meier pour G la survie de la
censure

######Calcul pour Donnees Completes DC#

v=NULL
for (j in 1:n){if (sum(as.integer(Gkm$time<=X[j]))!=0)
{v[j]=delta[j]/(Gkm$surv[sum(as.integer(Gkm$time<=X[j]))])^2} else {v[j]=1} }
vectV=v
d=matrix(rep(v),1,ncol=n,nrow=n)
II=matrix(nrow=n,ncol=n)  # matrice des indicatrices
X2=matrix(rep(X),1,ncol=n,nrow=n) ; X1=t(X2)
II=t(matrix(as.integer(X1>=X2),nrow=n,byrow=T))
m=d*II ; matrM=m
Z1=t(matrix(rep(Z[,1]),1,nrow=n,ncol=n))
A1=Z1-t(Z1)
indexi=rep(1:n,each=n); indexj=rep(1:n,n)
s3=function(b){#la somme des Zij * xi
xi=function(i,j){                  # fonction xi(.)
h=function (u){h=exp(-exp(u+crossprod(Z[i,]-Z[j,],b)))*exp(u-exp(u))
h}
xi=integrate(h,-Inf,+Inf)$value
xi}#fin fonction xi
ft=matrix(mapply( xi,indexi,indexj),nrow=n,byrow=T)
sp=sum(A1*ft)
sp}#fin de s3
Udc=function(b){ #fonction d estimation
r=colSums(Z*rowSums(m))-colSums(Z*colSums(m))-s3(b)
r     }#fin de Udc
DC[k,]=dfsane(par=.5, fn=Udc)$par
```

```
#calcul  ecart type pour DC
indi=rep(1:n,each=n*n);indj=rep(rep(1:n,each=n),n);inds=rep(1:n,n*n)
terme= function(i,j,s){if(s!=j){
t=(e(i,j)-e(j,i))*(e(i,s)-e(s,i))*tcrossprod(Z[i,]-Z[j,],Z[i,]-Z[s,])
} else {t=matrix(0,nrow=p,ncol=p)}
t}#fin de terme
terme_array=array(mapply(terme,indi,indj,inds),c(p,p,n*n*n))
terme1=sum_array(terme_array,sum)
R=rep(1,n)
terme2ft=function(s){ sums=function(i,j){sums=t(Z[i,]-
Z[j,])*m[i,j]*as.integer(X[j]>=X[s])    }#fin de sums
carre_array=array(mapply(sums, indexi,indexj),c(p,p,n*n))
carre_sum=sum_array(carre_array,sum)
terme2ft=(R[s]*(1-delta[s])/(colSums(R*II)[s])^2)*tcrossprod(carre_sum)
terme2ft}#fin de terme2ft
terme2_array=array(mapply(terme2ft,1:n),c(p,p,n*n))
terme2=sum_array(terme2_array,sum)/n
vargamma=(1/n^3)*terme1-(4/n^3)*terme2      #calcul de variance de gamma
#calcul lambda inverse pour DC
derivxi=function (i,j){ #derivee de xi
prime=function(u){prime=exp((u+crossprod(Z[i,]-Z[j,],DC[k,]))- exp(u+crossprod(Z[i,]-
Z[j,],DC[k,])) )*exp(u-exp(u))   }
xiprime=-integrate(prime,-Inf,+Inf)$value
xiprime}#fin derivxi
inverse=function(i,j){inverse=tcrossprod(Z[i,]-Z[j,] )* derivxi(i,j)  }#fin inverse
lambda_array=array(mapply(inverse,indexi,indexj),c(p,p,n*n))
invlambda=ginv((1/n^2)*sum_array(lambda_array,sum))
sigma=invlambda%*%vargamma%*%invlambda
seDC[k,]=sqrt(sigma)

#######debut calcul pour IWP 15% de données manquantes#####

a=2 ;   g=1        #pourcentage de données manquantes de 15%
p=exp(a+g*Z[,1])/(1+exp(a+g*Z[,1]))
```

```
R=rbinom(n,1,p)
gammac=glm(R~Z[,1], family="binomial")$coefficient
muchap=exp(gammac[1]+Z[,1]*gammac[2])/(1+exp(gammac[1]+Z[,1]*gammac[2]))
X=pmin(T,C)

# equations généralisées IWP
m=m*matrix(rep(R/muchap),1,ncol=n,nrow=n)
Z1=t(matrix(rep(Z[,1]),1,nrow=n,ncol=n))
A1=Z1-t(Z1)
indexi=rep(1:n,each=n); indexj=rep(1:n,n)
Uiwp=function(b){ #fonction d estimation
r=colSums((R/muchap)*Z*rowSums(m))-colSums(Z*colSums(m*(R/muchap)))-s3(b)
r       }#fin de Uiwp
IWP15[k,]=dfsane(par=.5, fn=Uiwp)$par

#calcul écart type pour iwp15%
terme_array=array(mapply(terme,indi,indj,inds),c(p,p,n*n*n))
terme1=sum_array(terme_array,sum)
terme2_array=array(mapply(terme2ft,1:n),c(p,p,n*n))
terme2=sum_array(terme2_array,sum)/n
vargamma=(1/n^3)*terme1-(4/n^3)*terme2

#calcul lambda inverse pour IWP15
derivxi=function (i,j){ #derivee de xi
prime=function(u){prime=exp((u+crossprod(Z[i,]-Z[j,],IWP15[k,]))-
exp(u+crossprod(Z[i,]-Z[j,],IWP15[k,])) )*exp(u-exp(u))}
xiprime=-integrate(prime,-Inf,+Inf)$value
xiprime  }#fin derivxi
lambda_array=array(mapply(inverse,indexi,indexj),c(p,p,n*n))
invlambda=ginv((1/n^2)*sum_array(lambda_array,sum))
sigma=invlambda%*%vargamma%*%invlambda
seIWP15[k,]=sqrt(sigma)

#######debut calcul pour CC 15% (sans données manquantes)#####

T=T[R==1] ;C=C[R==1] ;X=pmin(T,C)
n=length(X)
Z1aux=NULL ;Z1aux=Z[,1]
Z=matrix(nrow=n,ncol=p) ;Z[,1]=Z1aux[R==1]
muchap=muchap[R==1] ;muchap=c(rep(1,n))
v=v[R==1] ; R=R[R==1]
d=matrix(rep(v),1,ncol=n,nrow=n)
X2=matrix(rep(X),1,ncol=n,nrow=n)
X1=t(X2)
II=t(matrix(as.integer(X1>=X2),nrow=N,byrow=T))
m=d*II*matrix(rep(R/muchap),1,ncol=n,nrow=n)
```

```
Z1=t(matrix(rep(Z[,1]),1,nrow=n,ncol=n))
A1=Z1-t(Z1)
indexi=rep(1:n,each=n); indexj=rep(1:n,n)
CC15[k,]=dfsane(par=.5, fn=Uiwp)$par

#calcul variance CC15%

indi=rep(1:n,each=n*n); indj=rep(rep(1:n,each=n),n); inds=rep(1:n,n*n)
e=function(i,j){ #eijchap de betachap
e=m[i,j]-xi(i,j)}#fin e
terme_array=array(mapply(terme,indi,indj,inds),c(p,p,n*n*n))
terme1=sum_array(terme_array,sum)
terme2_array=array(mapply(terme2ft,1:n),c(p,p,n*n))
terme2=sum_array(terme2_array,sum)/n
vargamma=(1/n^3)*terme1-(4/n^3)*terme2

#calcul lambda inverse pour CC15

derivxi=function (i,j){ #derivee de xi
prime=function(u){prime=exp((u+crossprod(Z[i,]-Z[j,],CC15[k,]))- exp(u+crossprod(Z[i,]-
Z[j,],CC15[k,])) )*exp(u-exp(u))}
xiprime=-integrate(prime,-Inf,+Inf)$value
xiprime}#fin derivxi
lambda_array=array(mapply(inverse,indexi,indexj),c(p,p,n*n))
invlambda=ginv((1/n^2)*sum_array(lambda_array,sum))
sigma=invlambda%*%vargamma%*%invlambda
seCC15[k,]=sqrt(sigma)
```

CHAPITRE 3

TEST DU LOG-RANK AVEC STRATES MANQUANTES ET CENSURE DEPENDANTE

3.1. Introduction

Le test du log-rank est souvent utilisé pour comparer des groupes de traitement randomisés en présence de censure. Le test du log-rank est un test non-paramétrique, il permet de tester l'hypothèse nulle H_0 d'égalité des fonctions de risque dans les différents groupes. L'idée du test est de comparer l'estimateur de Nelson-Aalen du groupe spécifié à celui calculé sous H_0, commun à tous les groupes. Ce test peut être généralisé au test du log-rank stratifié.

Dans ce chapitre, nous présentons le test du log-rank et sa généralisation au cas stratifié, puis nous introduisons une nouvelle version modifiée de ce test dans le cas où les strates sont manquantes et la censure est dépendante du groupe de traitement. Nous établissons les propriétés asymptotiques de la statistique du test obtenue sous l'hypothèse nulle. Nous la comparons ensuite à la statistique du test basé sur l'analyse en cas complet à l'aide d'une étude de simulation.

3.2. Comparaison des groupes de survie

Considérons que dans un essai clinique, *n* individus sont randomisés dans *K* différents groupes de traitement. Supposons que l'on veut étudier l'effet d'une ou plusieurs covariables sur la survie dans ces *K* groupes.

Soit $\lambda_k(t)$ la fonction de risque instantané d'un individu dans le groupe $k, k = 1, \dots, K$. Le problème revient à tester l'hypothèse $H_0: \lambda_1 = \lambda_2 = \cdots = \lambda_K$.

Soient $N_{ik}(t)$ et $Y_{ik}(t)$ le processus de comptage individuel et l'indicateur de risque respectivement.

Soit n_k le nombre d'individus dans le groupe *k*, et soit $\sum_{k=1}^{K} n_k = n$ le nombre total d'individus supposés indépendants.

Dans le groupe *k*, l'estimateur de Nelson-Aalen de la fonction de risque cumulé $\Lambda_k(t) = \int_0^t \lambda_k(s)ds$ est donné par:

$$\widehat{\Lambda}_k(t) = \int_0^t \frac{dN_k(s)}{Y_k(s)}$$

Où $N_k(t) = \sum_{i=1}^{n_k} N_{ik}(t)$ représente la somme des processus de comptage individuels dans le groupe *k* et $Y_k(t) = \sum_{i=1}^{n_k} Y_{ik}(t)$ représente le nombre des individus à risque à l'instant *t* dans le groupe *k*.

L'estimateur de Nelson-Aalen, sous H_0 de la fonction de risque commune à tous les groupes $\Lambda(t) = \int_0^t \lambda(s)ds$ est donné par :

$$\widehat{\Lambda}(t) = \int_0^t \frac{dN_{\cdot}(s)}{Y_{\cdot}(s)}$$

Où $N_{\cdot}(t) = \sum_{k=1}^{K} N_k(t)$ et $Y_{\cdot}(t) = \sum_{k=1}^{K} Y_k(t)$

Soit $\widetilde{\Lambda}_k(t) = \int_0^t J_k(s)d\widehat{\Lambda}(s)$

$$= \int_0^t J_k(s)\frac{dN_{\cdot}(s)}{Y_{\cdot}(s)}$$

Avec $J_k(t) = I(Y_k(t) > 0)$

On a : $\widehat{\Lambda}_k(t) - \widetilde{\Lambda}_k(t) = \int_0^t \frac{dN_k(s)}{Y_k(s)} - \int_0^t J_k(s) \frac{dN_\cdot(s)}{Y_\cdot(s)}$

Soit $M_k(t) = N_k(t) - \int_0^t Y_k(s)\lambda(s)ds$ une martingale de carré intégrable. On peut donc écrire :

$$\widehat{\Lambda}_k(t) - \widetilde{\Lambda}_k(t) = \int_0^t \frac{J_k(s)dM_k(s)}{Y_k(s)} ds + \int_0^t \frac{J_k(s)Y_k(s)\lambda_k(s)}{Y_k(s)} ds - \int_0^t \frac{J_k(s)dM_\cdot(s)}{Y_\cdot(s)} ds - \int_0^t \frac{J_k(s)Y_\cdot(s)\lambda(s)}{Y_\cdot(s)}$$

$$= \underbrace{\int_0^t \frac{J_k(s)Y_k(s)\lambda(s)}{Y_k(s)} ds - \int_0^t \frac{J_k(s)Y_\cdot(s)\lambda(s)}{Y_\cdot(s)}}_{=0} + \int_0^t \frac{J_k(s)dM_k(s)}{Y_k(s)} ds - \int_0^t \frac{J_k(s)dM_\cdot(s)}{Y_\cdot(s)} ds$$

$$= \int_0^t \frac{J_k(s)dM_k(s)}{Y_k(s)} ds - \int_0^t \frac{J_k(s)d\sum_{j=1}^K M_j(s)}{Y_\cdot(s)} ds$$

$$= \int_0^t \frac{J_k(s)dM_k(s)}{Y_k(s)} ds - \sum_{j=1}^K \int_0^t \frac{J_k(s)dM_j(s)}{Y_\cdot(s)} ds$$

$$= \int_0^t \frac{J_k(s)\sum_{j=1}^K \delta_{jk}(s)dM_j(s)}{Y_k(s)} ds - \sum_{j=1}^K \int_0^t \frac{J_k(s)dM_j(s)}{Y_\cdot(s)} ds$$

$$= \sum_{j=1}^K \int_0^t J_k(s) \left[\frac{\delta_{jk}}{Y_k(s)} - \frac{1}{Y_\cdot(s)}\right] dM_j(s)$$

$$= \int_0^t \left[\frac{dN_k(s)}{Y_k(s)} - \frac{dN_\cdot(s)}{Y_\cdot(s)}\right]$$

Où $\delta_{jk} = I(j = k)$, $\sum_{j=1}^K \delta_{jk} = 1$ et on a $\lambda_k = \lambda$ sous H_0.

On pose :

$$Z_k = Z_k(\tau) = \int_0^\tau w_k(t) d\left(\widehat{\Lambda}_k - \widetilde{\Lambda}_k\right)(t)$$

Où $w_k(.)$ est une fonction de poids. Si $w_k(t) = Y_k(t)w(t)$ alors :

$$Z_k = \int_0^\tau Y_k(t)w(t)\left[\frac{dN_k(t)}{Y_k(t)} - \frac{dN_.(t)}{Y_.(t)}\right]$$

$$= \int_0^\tau w(t)\left[dN_k(t) - \frac{Y_k(t)dN_.(t)}{Y_.(t)}\right]$$

On a $\sum_{k=1}^K Z_k = 0$. En effet,

$$\sum_{k=1}^K Z_k = \int_0^\tau w(t)\left[\sum_{k=1}^K dN_k(t) - \frac{\sum_{k=1}^K Y_k(t)}{Y_.(t)} dN_.(t)\right] = 0$$

On pose $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_{K-1} \end{pmatrix}$ alors sous certaines conditions de régularité (voir Anderson et al.(1993)), $n^{1/2}Z$ est asymptotiquement gaussien de moyenne nulle et de matrice variance-covariance Σ, d'éléments Σ_{kl}, estimée d'une manière consistante par :

$$\hat{\Sigma}_{kl} = n\int_0^\tau w_k(t)\, w_l(t) J_k(t) J_l(t) \frac{1}{Y_k(t)Y_.(t)}\left(\delta_{kl} - \frac{Y_k(t)}{Y_.(t)}\right) Y_.(t)\lambda(t)dt$$

On pose $Q = nZ'\hat{\Sigma}^{-1}Z$ et sous H_0, Q suit asymptotiquement une loi $\chi^2(K-1)$.

Différents tests sont construits pour différents choix de la fonction $w(t)$. En particulier, le test du log-rank est construit en choisissant $w(t) = I(Y_.(t) > 0)$ et donc :

$$Z_k = \int_0^\tau I(Y_.(t) > 0)\left[dN_k(t) - \frac{Y_k(t)}{Y_.(t)} dN_.(t)\right]$$

Si $Y_.(t) > 0, \forall t \in [0,\tau]$, $Y_i(t) = 1(T_i \geq t)$, et si $1(G_i = k)$ est l'indicatrice que l'individu *i* appartient au groupe *k*, on a :

$$Z_k = \int_0^\tau \left[\sum_{i=1}^n 1(G_i = k)dN_i(t) - \frac{\sum_{j=1}^n 1(G_j = k)Y_j(t)}{\sum_{j=1}^n Y_j(t)} \sum_{i=1}^n dN_i(t)\right]$$

$$= \sum_{i=1}^{n} \int_0^{\tau} \left[1(G_i = k) - \frac{\sum_{j=1}^{n} 1\big(G_j = k\big) Y_j(t)}{\sum_{j=1}^{n} Y_j(t)} \right] dN_i(t)$$

3.3. Test du log-rank stratifié :

Considérons un essai clinique randomisé où *n* patients sont affectés aléatoirement à *K*groupes de traitement, en ajustant à un facteur *S* à *L* modalités, appelées strates (stades de la maladie par exemple).

Soit $\lambda_{kl}(t)$ la fonction de risque d'un individu dans le groupe *k* et la strate *l*, $k = 1, \ldots, K$ et $l = 1, \ldots, L$.

On veut tester l'hypothèse que la survie est identique pour tous les individus de la même strate:

$$H_0 : \lambda_{1l} = \cdots = \lambda_{Kl}, \text{ pour tout } l = 1, \ldots, L$$

et H_a: il existe *j* et *j'* tels que $\lambda_{jl} \neq \lambda_{j'l}$ pour au moins un *l*

Ce qui revient à comparer l'estimateur de Nelson-Aalen de chaque strate à celui calculé sous H_0 (Martinussen et Scheike (2006))

Soit:

$$Z_{kl} = Z_{kl}(\tau) = \int_0^{\tau} w_{kl}(t) d\big(\widehat{\Lambda}_{kl} - \widetilde{\Lambda}_l\big)(t)$$

Où $w_{kl}(.)$ est une fonction de poids, et

$$\widehat{\Lambda}_{kl}(t) = \int_0^{t} \frac{dN_{kl}(s)}{Y_{kl}(s)}$$

$$\widetilde{\Lambda}_l(t) = \int_0^{t} J_{kl}(s) \frac{dN_{.l}(s)}{Y_{.l}(s)}$$

$J_{kl}(t) = I(Y_{kl}(t) > 0),\ N_{.l}(t) = \sum_{k=1}^{K} N_{kl}(t)$ et $Y_{.l}(t) = \sum_{k=1}^{K} Y_{kl}(t)$

Avec $N_{kl}(t)$ représente la somme des processus de comptage dans le groupe *k* et la strate *l*, et $Y_{kl}(t)$ représente le nombre des individus à risque à l'instant *t* dans le groupe *k* et la strate *l*.

On pose $Z_l = \begin{pmatrix} Z_{1l} \\ \vdots \\ Z_{K-1,l} \end{pmatrix}$ et soit $\hat{\Sigma}_l$ l'estimateur de la variance de Z_l. Alors sous H_0 et pour *l* fixé, ${Z_l}'{\hat{\Sigma}_l}^{-1} Z_l$ suit asymptotiquement une loi $\chi^2(K-1)$. Pour combiner l'information sur toutes les strates, on pose $Z = \sum_{l=1}^{L} Z_l$ et $\hat{\Sigma} = \sum_{l=1}^{L} \hat{\Sigma}_l$. Ainsi, une classe de tests stratifiés pour H_0 est donnée par $Z'\hat{\Sigma}^{-1}Z$ qui suit asymptotiquement une loi $\chi^2(K-1)$.

Le test du log-rank stratifié est obtenu si $w_{kl}(t) = Y_{kl}(t)1(Y_{.l}(t) > 0)$. Soit $1(S_i = l)$ l'indicatrice d'appartenance à la strate *l*, on a donc :

$$Z_{kl} = \sum_{i=1}^{n} \int_0^{\tau} J_{kl}(t) \left[1(G_i = k)1(S_i = l) - \frac{\sum_{j=1}^{n} Y_j(t)1(G_j = k)1(S_j = l)}{\sum_{j=1}^{n} Y_j(t)1(S_j = l)} 1(S_i = l) \right] dN_i(t)$$

On pose $Z_k = \sum_{l=1}^{L} Z_{kl}$

$$Z_k = \sum_{l=1}^{L} \sum_{i=1}^{n} \int_0^{\tau} J_{kl}(t) \left[1(G_i = k)1(S_i = l) - \frac{\sum_{j=1}^{n} Y_j(t)1(G_j = k)1(S_j = l)}{\sum_{j=1}^{n} Y_j(t)1(S_j = l)} 1(S_i = l) \right] dN_i(t)$$

$$= \sum_{i=1}^{n} \int_0^{\tau} \sum_{l=1}^{L} J_{kl}(t) \left[1(G_i = k)1(S_i = l) - \frac{\sum_{j=1}^{n} Y_j(t)1(G_j = k)1(S_j = l)}{\sum_{j=1}^{n} Y_j(t)1(S_j = l)} 1(S_i = l) \right] dN_i(t)$$

$$Z_k = \sum_{i=1}^{n} \int_0^{\tau} \left\{ 1(G_i = k) - \sum_{l=1}^{L} 1(S_i = l)\mathcal{E}_{n,k}(t,l) \right\} dN_i(t) \qquad (3.1)$$

Où

$$\mathcal{E}_{n,k}(t,l) = \frac{\sum_{j=1}^{n} Y_j(t) 1\big(G_j = k\big) 1\big(S_j = l\big)}{\sum_{j=1}^{n} Y_j(t) 1\big(S_j = l\big)}$$

En posant $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_{K-1} \end{pmatrix}$ alors $Z^T \hat{\Sigma}^{-1} \mathrm{Z}$ suit asymptotiquement une loi $\chi^2(K-1)$ sous l'hypothèse nulle H_0.

3.4. Test du log-rank stratifié avec strates manquantes et censure dépendante

Nous considérons le problème d'implémentation d'un test du log-rank stratifié lorsque les strates ne sont pas observées pour quelques individus, et lorsque la censure est dépendante du groupe de traitement. Cette situation est rencontrée, par exemple, lorsque la censure est causée par l'effet secondaire d'un traitement. Les patients du groupe qui supportent mal les effets secondaires du traitement auront plus tendance à quitter l'étude, et donc le taux de censure sera plus élevé dans ce groupe que dans les autres.

Considérons un échantillon de *n* individus indépendants observées sur un intervalle $[0, \tau]$. Supposons que la strate *S* n'est pas observée pour un sous ensemble d'individus. Soit *R* l'indicatrice qui est égale à *1* si la strate *S* est observée et *0* sinon.

Supposons également que l'on observe des covariables auxiliaires $W \in \mathbb{R}^p$ pour tous les individus de l'étude. Ces covariables fournissent une information partielle sur les patients.

Soient $T_1^0, \dots, T_n^0$ les dates de décès observées sur les *K* groupes sur un intervalle fini $[0, \tau], \tau < \infty$. Soient $C_1, \dots, C_n$ les dates de censure à droite. Les observations consistent donc en *n* couples indépendants $(T_i, \Delta_i, G_i, W_i, R_i, R_i S_i)$, $i = 1, \dots, n$, où $T_i = \min(T_i^0, C_i)$ et $\Delta_i = 1(T_i^0 \leq C_i)$. Supposons que T^0 et *C* sont indépendants étant donnés *G*, *S*, *W* et *R* et

que *C* est indépendant de *S* et de *W* étant donné *G*. Supposons également que *G* est indépendant de *S* et *W*. Supposons qu'on est en présence d'un mécanisme de manque aléatoire (Missing At Random) ce qui signifie que *R* est indépendant de *S* étant donné *W*. Notons par C1 cet ensemble de conditions.

Nous proposons une version modifiée du test du log-rank stratifié pour H_0 qui consiste à :

a) remplacer dans (3.1) chaque indicatrice $1(S_i = l)$ non observée par l'espérance conditionnelle $\mathbb{E}[1(S = l)|W_i] = \mathbb{P}(S = l|W_i)$ (régression par calibration, voir Thurston et *al.* (2003), Weller et *al.* (2007), Dupuy & Leconte (2008,2009)),
b) pondérer chaque individu par l'inverse de la fonction de survie de la censure étant donné le groupe de l'individu $\mathbb{P}(C \geq t|G)$ (le principe de Inverse probability of censoring weighted (ICPW), voir Robins & Finkelstein (2000), Yoshida et *al.* (2007), Cain & Cole (2009)).

La statistique du test que nous proposons est basée sur la version modifiée $\tilde{Z}_k$de (3.1) :

$$\tilde{Z}_k = \sum_{i=1}^{n} \int_0^{\tau} \mu(G_i, t) \left[G_i^{(k)} - \sum_{l=1}^{L} D_i^l \tilde{\mathcal{E}}_{k,l}(t) \right] dN_i(t) \quad , k = 1, \dots, K \tag{3.2}$$

Où

$$\mu(G_i, t) = \frac{1}{\mathbb{P}(C \geq t|G_i)}$$

$$G_i^k = 1(G_i = k)$$

$D_i^l := R_i 1(S_i = 1) + (1 - R_i)\mathbb{P}(S = l|W_i)$ pour $i = 1, \dots, n$ et $l = 1, \dots, L$,

$$\tilde{\mathcal{E}}_{k,l}^{(n)}(t) = \frac{\tilde{S}_{k,l}^{(n)}(t)}{\tilde{S}_l^{(n)}(t)}$$

$$\tilde{S}_{k,l}^{(n)}(t) = n^{-1} \sum_{i=1}^{n} Y_i(t) G_i^{(k)} D_i^l \mu(G_i, t)$$

$$\tilde{S}_l^{(n)}(t) = n^{-1}\sum_{i=1}^{n} Y_i(t)D_i^l\,\mu(G_i,t)$$

Définissons, pour $i = 1,\dots,n$ et $k = 1,\dots,K$:

$$V_{i,k} = \int_0^\tau \mu(G_i,t)\left\{G_i^k - \sum_{l=1}^{L} D_i^l\tilde{\mathcal{E}}_{k,l}^{(n)}(t)\right\}dN_i(t) - \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{L}\int_0^\tau \frac{Y_i(t)D_i^l D_j^l\mu(G_j,t)}{\tilde{S}_l(t)}\Big\{G_i^k - \tilde{\mathcal{E}}_{k,l}^{(n)}(t)\Big\}dN_j(t) \qquad (3.3)$$

Posons $\mathbb{V}_i = \left(V_{i,1},\dots,V_{i,K-1}\right)^\intercal$. Et pour $k = 1,\dots,K$, $l = 1,\dots,L$:

$\tilde{s}_{k,l}(t) = \mathbb{E}[Y(t)G^k D^l\mu(G,t)]$

$\tilde{s}_l(t) = \mathbb{E}[Y(t)D^l\mu(G,t)]$

$\tilde{e}_{k,l}(t) = \dfrac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}$

Les conditions de régularité suivantes sont nécessaires pour énoncer notre résultat :

C2 : Il existe une constante positive c_0 telle que $\mathbb{P}(C \geq \tau|G) > c_0$ pour $G \in \{1,\dots,K\}$, et la fonction de survie $\mathbb{P}(C \geq t|G)$ est continue sur $[0,\tau]$.

C3 : Pour tout $k = 1,\dots,K$ et $l = 1,\dots,L$, $sup_{t\in[0,\tau]}\lambda_{k,l}(t) < c_1$ où c_1 est une constante finie positive.

C4 : Il existe une constante positive c_2 telle que $inf_{t\in[0,\tau]}\tilde{s}_l(t) < c_2$ pour $l = 1,\dots,L$.

Le théorème suivant donne le résultat concernant la nouvelle statistique du test :

$$\widetilde{U} \coloneqq \left(\tilde{Z}_1,\dots,\tilde{Z}_{K-1}\right)\hat{\Sigma}^{-1}\left(\tilde{Z}_1,\dots,\tilde{Z}_{K-1}\right)^\intercal$$

Où : $\hat{\Sigma} = \sum_{i=1}^{n}\mathbb{V}_i\,\mathbb{V}_i^\intercal$.

Théorème 3.1 : (Mezaouer et al. (2013))

Sous les conditions C1-C4, et sous H_0, $\widetilde{U}$ converge en loi vers une loi χ^2à K-1 degrés de liberté.

En se basant sur ce théorème, H_0 est rejetée si $\widetilde{U} \geq \chi^2_{1-\alpha}(K-1)$, où $\chi^2_{1-\alpha}(K-1)$ est le quantile d'ordre $1-\alpha$ de la loi de $\chi^2(K-1)$.

Preuve du théorème 3.1 :

Pour montrer la distribution asymptotique de la statistique $\widetilde{U}$ sous H_0, nous montrons tout d'abord que $\tilde{Z}_k$ est asymptotiquement linéaire, ensuite nous utilisons un théorème central limite pour la somme de *n* termes i.i.d.

Soit $\mathcal{F}_{t,i} = \sigma\{N_i(s), (1-\Delta_i)1(T_i \leq s), G_i, S_i, W_i : 0 \leq s \leq t\}$ la σ-algèbre engendrée par les dates des évènements d'intérêt et les dates de censure de l'individu *i* sur $[0,t]$, et par le groupe, la strate et les informations auxiliaires de cet individu. L'intensité du processus de comptage $N_i(t)$ est donnée par :

$$Y_i(t)\lambda_i(t) = Y_i(t)\sum_{k=1}^{K}\sum_{l=1}^{L}\lambda_{k,l}(t)G_i^k 1(S_i = l)$$

Si S_i est manquante alors l'information pour l'individu *i* est représentée par la plus petite σ-algèbre $\mathcal{G}_{t,i} = \sigma\{N_i(s), (1-\Delta_i)1(T_i \leq s), G_i, W_i : 0 \leq s \leq t\} \subseteq \mathcal{F}_{t,i}$.

L'intensité de $N_i(t)$ par rapport à $\mathcal{G}_{t,i}$ est $Y_i(t)\gamma_i(t) \coloneqq \mathbb{E}\big[Y_i(t)\lambda_i(t)|\mathcal{G}_{t-,i}\big]$, où $\gamma_i(t) = \sum_{k=1}^{K}\sum_{l=1}^{L}\lambda_{k,l}(t)G_i^k\mathbb{E}\big[1(S_i = l)|\mathcal{G}_{t-,i}\big]$.

Soit $\mathcal{H}_{t,i} = (\mathcal{F}_{t,i})^{R_i}(\mathcal{G}_{t,i})^{1-R_i}$ la filtration observée, alors le processus $N_i(t)$ a l'intensité $Y_i(t)\zeta_i(t) \coloneqq Y_i(t)[\lambda_i(t)R_i + \gamma_i(t)(1-R_i)]$ par rapport à la filtration $(\mathcal{H}_{t,i})_{t\geq 0}$. Notons $\kappa_l(t) = \mathbb{E}[Y(t)\zeta(t)D^l\mu(G,t)], l = 1, \ldots, L$.

Le lemme suivant permet d'établir l'approximation de $n^{-\frac{1}{2}}\tilde{Z}_k$:

Lemme 3.1

Pour tout $i = 1, \dots, n$ *et* $= 1, \dots, K$ *, soit*

$$Q_{i,k} = \int_0^\tau \sum_{l=1}^{L} D_i^l \mu(G_i, t) \left(G_i^k - \tilde{e}_{k,l}(t) \right) \left[dN_i(t) - Y_i(t) \frac{\kappa_l(t)}{\tilde{s}_l(t)} dt \right]$$

Sous les conditions C1-C4, $n^{-\frac{1}{2}} \tilde{Z}_k = n^{-\frac{1}{2}} \sum_{i=1}^n Q_{i,k} + o_p(1)$.

De plus, si H_0 *est vraie alors* $\mathbb{E}(Q_{i,k}) = 0$.

La preuve du Lemme 3.1 est donnée en Annexe 3.A.

Il s'en suit du Lemme 3.1, de la version multi-variée du théorème limite central, et du théorème de Slustky que sous H_0, $n^{-\frac{1}{2}} (\tilde{Z}_1, \dots, \tilde{Z}_{K-1})^\intercal$ converge en loi lorsque n tend vers ∞, vers un vecteur gaussien de longueur *K*-1, de moyenne nulle et de matrice de variance-covariance $\Sigma = \mathbb{E}[\mathbb{Q}_1 \mathbb{Q}_1^\intercal]$, où $\mathbb{Q}_i = (Q_{i,1}, \dots, Q_{i,K-1})^\intercal$. En conséquence, sous H_0 $n^{-\frac{1}{2}} (\tilde{Z}_1, \dots, \tilde{Z}_{K-1}) \Sigma^{-1} (\tilde{Z}_1, \dots, \tilde{Z}_{K-1})^\intercal \xrightarrow{\mathcal{L}} \chi^2(K-1)$ lorsque n tend vers l'infini.

Un estimateur consistant de Σ est

$$n^{-1} \hat{\Sigma} := n^{-1} \sum_{i=1}^{n} \mathbb{V}_i \, \mathbb{V}_i^\intercal$$

où $\mathbb{V}_i = (V_{i,1}, \dots, V_{i,K-1})^\intercal$ et $V_{i,k}$ est donné par (3.3).

En effet, on a $n^{-1} \hat{\Sigma} = n^{-1} \sum_{i=1}^n (\mathbb{V}_i \mathbb{V}_i^\intercal - \mathbb{Q}_i \mathbb{Q}_i^\intercal) + n^{-1} \sum_{i=1}^n \mathbb{Q}_i \mathbb{Q}_i^\intercal$ donc il suffit de montrer que $V_{i,k} - Q_{i,k} \xrightarrow{\mathcal{P}} 0$ lorsque n tend vers l'infini. Les arguments et les calculs de démonstration sont similaires à ceux donnés dans la preuve du lemme 3.2 (Annexe 3.B). Il s'en suit du théorème de Slutsky, que lorsque n tend vers l'infini

$$n^{-\frac{1}{2}} (\tilde{Z}_1, \dots, \tilde{Z}_{K-1}) \Sigma^{-1} (\tilde{Z}_1, \dots, \tilde{Z}_{K-1})^\intercal \xrightarrow{\mathcal{L}} \chi^2(K-1)$$

3.5. Etude de simulation

Nous considérons le cas de deux groupes de traitements $K = 2$ et deux strates $L = 2$. Dans chaque groupe et chaque strate, les durées T_i^0 sont générées à partir d'une distribution Weibull $W(\alpha, \lambda)$ avec un taux de risque $\lambda(t) = \alpha\lambda t^{\alpha-1}$. Les durées de vie de la strate 1 dans le groupe 1, sont générées à partir d'une $W(\alpha_1, \lambda_1)$ et celles de la strate 1 dans le groupe 2 d'une $W(\alpha_1, \lambda_1 r_1)$ où r_1 est le rapport des risques de deux patients dans la strate 1 des groupes 1 et 2 respectivement.

Les durées de vie de la strate 2 dans le groupe 1, sont générées à partir d'une $W(\alpha_2, \lambda_2)$ et celles de la strate 2 dans le groupe 2 à partir d'une $W(\alpha_2, \lambda_2 r_2)$ où r_2 est le rapport des risques de deux patients dans la strate 2 des groupes 1 et 2 respectivement.

Nous utilisons $\alpha_1 = .5$, $\alpha_2 = .75$, $\lambda_1 = .75$ et $\lambda_2 = 1.5$. Trois cas sont considérés : (a) $(r_1, r_2) = (1,1)$, (b) $(r_1, r_2) = (1.5,1.5)$, (c) $(r_1, r_2) = (1.25,2)$. Le cas (a) correspond au cas dit « nul » où il n y a pas de différence entre les groupes de traitement dans chaque strate. Les cas (b) et (c) correspondent aux cas avec différence entre les groupes.

Dans chaque cas, les temps de censure sont générés à partir de lois exponentielles de paramètres θ_1 dans le groupe 1 et θ_2 dans le groupe 2, avec θ_1 et θ_2 choisis de manière à avoir des pourcentages de censure c_1 dans le groupe 1 et c_2 dans le groupe 2 (en prenant $\theta_1 \neq \theta_2$ cela nous assure que la censure dépend du groupe).

Soient n_1 et n_2 les tailles d'échantillon dans les groupes 1 et 2 respectivement (avec $n = n_1 + n_2$). Nous considérons plusieurs valeurs pour $(n_1, n_2) = (50,50)$, $(n_1, n_2) = (100,100)$ et $(n_1, n_2) = (150,150)$. La variable auxiliaire W est une bidimensionnelle ($W = (W_1, W_2)'$) , avec W_1 (respectivement W_2) générée à partir d'une distribution uniforme sur $[-1,1]$ (respectivement de $\mathcal{N}(0, .5)$). Nous exprimons la relation entre S et W par un modèle de régression logistique :

$$\mathbb{P}(S = 1|W) = \frac{exp(b_0 + b_1 W_1 + b_2 W_2^2)}{1 + exp(b_0 + b_1 W_1 + b_2 W_2^2)}$$

Où $(b_0, b_1, , b_2)$ sont choisis de telle manière que dans chaque groupe de traitement, chaque strate contient approximativement la moitié des patients. Nous considérons les pourcentages 20% et 40% de données manquantes. La variable R est simulée suivant une

loi de Bernoulli de paramètre choisi de manière à obtenir les différents pourcentages de données manquantes considérés. Nous résumons les différents schémas considérés dans les simulations dans Table 3.1.

Paramètre	Valeurs	Description
(r_1, r_2)	(1,1), (1.5,1.5), (1.25,2)	Rapport des risques
(c_1, c_2)	(5,20), (20,50), (30,20), (40,10)	Pourcentages de censure
n	100, 200, 300	Taille de l'échantillon total

Table 3.1 : Paramètres et valeurs incluses dans les simulations.

En pratique, les fonctions des poids $\mu(G_i, .)$ et/ou les probabilités conditionnelles $\mathbb{P}(S_i = l|W_i) = \mathbb{E}[1(S_i = l|W_i)]$ peuvent être connues ou complètement inconnues. Dans ce dernier cas, elles seront estimées et remplacées par leurs estimateurs dans la statistique du test $\widetilde{U}$ (la statistique avec les estimateurs sera notée $\widehat{U}$ dans la suite). Nous estimons $\mathbb{P}(S_i = l|W_i)$ par une régression logistique locale (voir Loader (1999) pour plus de détails), en utilisant le package locfit du logiciel R . Nous utilisons les estimateurs non paramétriques de Kaplan-Meier dans chaque groupe pour estimer les fonctions de survie de la censure $\mathbb{P}(C \geq t|G_i)$. Pour 1000 répétitions nous obtenons le niveau (cas (a)) et la puissance (cas(b) et (c)) de la statistique estimée $\widehat{U}$ à un niveau de signification de 0.05. Pour la comparaison, nous incluons les résultats de la statistique U_{cc} du test du log-rank stratifié basé sur les cas complet (*i.e.* seulement les individus avec des strates connues). Les résultats sont donnés dans Table 3.2.

	Pourcentages de censure (c_1, c_2)			
	(5,20)	(20,50)	(30,20)	(40,10)

n	(r_1, r_2)	$\widehat{U}$	U_{cc}	$\widehat{U}$	U_{cc}	$\widehat{U}$	U_{cc}	$\widehat{U}$	U_{cc}
100	(1,1)	.067	.045	.070	.067	.057	.045	.061	.057
	(1.5,1.5)	.354	.228	.272	.204	.401	.258	.360	.221
	(1.25,2)	.349	.227	.303	.220	.361	.231	.314	.216
200	(1,1)	.074	.055	.065	.051	.067	.049	.051	.047
	(1.5,1.5)	.692	.440	.565	.369	.651	.381	.574	.423
	(1.25,2)	.592	.386	.605	.418	.603	.403	.465	.401
300	(1,1)	.060	.046	.076	.048	.078	.052	.073	.048
	(1.5,1.5)	.823	.596	.762	.472	.822	.566	.698	.568
	(1.25,2)	.772	.535	.813	.527	.746	.522	.611	.583

Table 3.2 : Résultats des simulations pour 40% de données manquantes.

3.6. Discussion

Les résultats obtenus montrent que la statistique proposée $\widehat{U}$ est meilleure que la statistique basée sur le cas complet. Le niveau empirique de $\widehat{U}$ tend à dépasser faiblement 0.05. Ceci est dû au fait que $\mathbb{P}(S_i = l|W_i)$ et $\mu(G_i, .)$ sont remplacés par leurs estimateurs, provoquant ainsi une faible déviation de la distribution asymptotique de $\widehat{U}$ sous H_0 , de la distribution $\chi^2(K-1)$. Dans les cas (b) et (c), les puissances de $\widehat{U}$ sont supérieures de celles de U_{cc} pour les différentes tailles d'échantillon et différents pourcentages de censure. En particulier, $\widehat{U}$ maintient une forte puissance quand la censure dépend fortement du groupe de traitement (20% dans le groupe 1, 50% dans le groupe 2), tandis que la puissance de U_{cc} décroit.

Annexe 3.A. : Preuve du Lemme 3.1

On a :

$$n^{-\frac{1}{2}}\tilde{Z}_k = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\tau \mu(G_i,t)G_i^k\, dN_i(t) - n^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{l=1}^{L}\int_0^\tau \mu(G_i,t)\, D_i^l\tilde{\mathcal{E}}_{k,l}^{(n)}(t)dN_i(t)$$

$$= A_{1,k}^{(n)} - A_{2,k}^{(n)}$$

Et

$$A_{2,k}^{(n)} = n^{\frac{1}{2}}\sum_{l=1}^{L}\left[\int_0^\tau \tilde{e}_{k,l}(t)\left\{n^{-1}\sum_{i=1}^{n}\mu(G_i,t)D_i^l dN_i(t) - \kappa_l(t)dt\right\}\right.$$
$$\left. + \int_0^\tau v_{k,l}^{(n)}(t)\left(n^{-1}\sum_{i=1}^{n}\mu(G_i,t)D_i^l dN_i(t) - \kappa_l(t)dt\right) + \int_0^\tau \tilde{\mathcal{E}}_{k,l}^{(n)}(t)\,\kappa_l(t)dt\right]$$

Où $v_{k,l}^{(n)}(t) = \tilde{\mathcal{E}}_{k,l}^{(n)}(t) - \tilde{e}_{k,l}(t)$

Laissons le premier terme de $A_{2,k}^{(n)}$ inchangé. Les second et troisième termes satisfont respectivement les deux lemmes suivants. Leurs preuves sont données dans Annexe 3.B

Lemme 3.2. Sous les conditions C1-C4,

$$n^{\frac{1}{2}}\sum_{l=1}^{L}\int_0^\tau v_{k,l}^{(n)}(t)\left[\frac{1}{n}\sum_{i=1}^{n}\mu(G_i,t)D_i^l dN_i(t) - \kappa_l(t)dt\right] \xrightarrow{\mathcal{P}} 0, n\to\infty$$

Lemme 3.3. Sous les conditions C1-C4 et quand $n\to\infty$

$$n^{\frac{1}{2}}\sum_{l=1}^{L}\int_0^\tau \tilde{\mathcal{E}}_{k,l}^{(n)}(t)\,\kappa_l(t)dt = n^{\frac{1}{2}}\sum_{l=1}^{L}\int_0^\tau\left\{\tilde{e}_{k,l}(t) + \frac{\tilde{S}_{k,l}^{(n)}(t)}{\tilde{s}_l(t)} - \frac{\tilde{s}_{k,l}(t)\tilde{S}_l^{(n)}(t)}{{\tilde{s}_l}^2(t)}\right\}\kappa_l(t)dt + o_p(1)$$

En utilisant ces deux lemmes, on peut réécrire $n^{-\frac{1}{2}}\tilde{Z}_k$ sous la forme suivante :

$$n^{-\frac{1}{2}}\tilde{Z}_k = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\tau \mu(G_i,t)G_i^k dN_i(t)$$
$$- n^{\frac{1}{2}}\sum_{l=1}^{L}\int_0^\tau \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\left\{n^{-1}\sum_{i=1}^{n}\mu(G_i,t)D_i^l dN_i(t) - \kappa_l(t)dt\right\}$$
$$- n^{\frac{1}{2}}\sum_{l=1}^{L}\int_0^\tau \frac{1}{\tilde{s}_l(t)}\left[\tilde{S}_{k,l}^{(n)}(t) - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\left(\tilde{S}_{k,l}^{(n)}(t) - \tilde{s}_l(t)\right)\right]\kappa_l(t)dt$$
$$- n^{\frac{1}{2}}\int_0^\tau \sum_{l=1}^{L} o_p(\frac{1}{n^2})\,\kappa_l(t)dt + o_p(1)$$

$$n^{-\frac{1}{2}}\tilde{Z}_k = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\tau \mu(G_i,t)G_i^k dN_i(t) - \sum_{l=1}^{L} B_l + o_p(1)$$

Où :

$$B_l = n^{\frac{1}{2}}\int_0^\tau \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\left(n^{-1}\sum_{i=1}^{n}\mu(G_i,t)D_i^l dN_i(t) - \kappa_l(t)dt\right)$$
$$+ n^{\frac{1}{2}}\int_0^\tau \frac{1}{\tilde{s}_l(t)}\left[\tilde{S}_{k,l}^{(n)}(t) - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\left(\tilde{S}_{k,l}^{(n)}(t) - \tilde{s}_l(t)\right)\right]\kappa_l(t)dt$$

$$= n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\tau \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\,\mu(G_i,t)D_i^l dN_i(t) - n^{\frac{1}{2}}\int_0^\tau \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\kappa_l(t)dt$$
$$+ n^{\frac{1}{2}}\int_0^\tau \frac{1}{\tilde{s}_l(t)}\left[\tilde{S}_{k,l}^{(n)}(t) - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\tilde{S}_{k,l}(t) + \tilde{s}_{k,l}(t)\right]\kappa_l(t)dt$$

$$= n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\tau \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\mu(G_i,t)D_i^l dN_i(t)$$
$$+ n^{\frac{1}{2}}\int_0^\tau \frac{1}{\tilde{s}_l(t)}\left[\frac{1}{n}\sum_{j=1}^{n}Y_j(t)G_j^k D_j^l\mu(G_j,t) - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)}\frac{1}{n}\sum_{j=1}^{n}Y_j(t)D_j^l\mu(G_j,t)\right]\kappa_l(t)dt$$

$$= n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{\tau} \left[\frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \mu(G_i, t) D_i^l dN_i(t) + \frac{1}{\tilde{s}_l(t)} Y_i(t) G_i^k D_i^l \mu(G_i, t) \kappa_l(t) dt \right.$$
$$\left. - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} Y_i(t) D_i^l \mu(G_i, t) \kappa_l(t) dt \right]$$

$$B_l = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{\tau} \left[\frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \mu(G_i, t) D_i^l dN_i(t) + \frac{Y_i(t) D_i^l \mu(G_i, t)}{\tilde{s}_l(t)} \left(G_i^k - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \right) \kappa_l(t) dt \right]$$

En remplaçant dans l'expression de $n^{-\frac{1}{2}} \tilde{Z}_k$, on obtient :

$$n^{-\frac{1}{2}} \tilde{Z}_k = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left\{ \int_0^{\tau} \left(G_i^k - \sum_{l=1}^{L} \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} D_i^l \right) \mu(G_i, t) dN_i(t) \right.$$
$$\left. - \sum_{l=1}^{L} \int_0^{\tau} \frac{Y_i(t) \mu(G_i, t) D_i^l}{\tilde{s}_l(t)} \left(G_i^k - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \right) \kappa_l(t) dt \right\} + o_p(1)$$

$$= n^{-\frac{1}{2}} \sum_{i=1}^{n} Q_{i,k} + o_p(1)$$

Où

$$Q_{i,k} = \int_0^{\tau} \left(G_i^k - \sum_{l=1}^{L} \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} D_i^l \right) \mu(G_i, t) dN_i(t)$$
$$- \sum_{l=1}^{L} \int_0^{\tau} \frac{Y_i(t) \mu(G_i, t) D_i^l}{\tilde{s}_l(t)} \left(G_i^k - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \right) \kappa_l(t) dt$$

$$Q_{i,k} = \int_0^{\tau} \left[\sum_{l=1}^{L} D_i^l \left(G_i^k - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \right) \mu(G_i, t) dN_i(t) \right.$$
$$\left. - \sum_{l=1}^{L} D_i^l \left(\frac{Y_i(t) \mu(G_i, t)}{\tilde{s}_l(t)} \left(G_i^k - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \right) \kappa_l(t) \right) dt \right]$$

$$= \int_0^{\tau} \sum_{l=1}^{L} D_i^l \left(G_i^k - \frac{\tilde{s}_{k,l}(t)}{\tilde{s}_l(t)} \right) \left[\mu(G_i, t) dN_i(t) - \frac{Y_i(t) \mu(G_i, t)}{\tilde{s}_l(t)} \kappa_l(t) dt \right]$$

$$Q_{i,k} = \int_0^\tau \sum_{l=1}^{L} D_i^l \mu(G_i, t) \left(G_i^k - \tilde{e}_{k,l}(t) \right) \left[dN_i(t) - Y_i(t) \frac{\kappa_l(t)}{\tilde{s}_l(t)} dt \right]$$

Calculons maintenant $\mathbb{E}Q_{i,k}$ sous H_0. Pour $i = 1, \dots, n$, $k = 1, \dots, K$ et $t \in [0, \tau]$, $\mu(G, t) G_i^k$ est $\mathcal{H}_{t,i}$-mesurable, donc le processus $\left(\int_0^t \mu(G_i, s) G_i^k dM_i(s) \right)_{t \geq 0}$ est une martingale de moyenne nulle. Notons que $\sum_{l=1}^{L} D_i^l = 1$, il s'ensuit donc que :

$$\mathbb{E}\left[\int_0^\tau \sum_{l=1}^{L} D_i^l \mu(G_i, t) G_i^k dN_i(t) \right] = \mathbb{E}\left[\int_0^\tau \mu(G_i, t) G_i^k dN_i(t) \right]$$

$$= \int_0^\tau \mathbb{E}\left[\mu(G_i, t) G_i^k Y_i(t) \zeta_i(t) \right] dt$$

$$\mathbb{E} \int_0^\tau \sum_{l=1}^{L} D_i^l \, \mu(G_i, t) \tilde{e}_{k,l}(t) dN_i(t) = \sum_{l=1}^{L} \int_0^\tau \tilde{e}_{k,l}(t) \kappa_l(t) dt,$$

$$\mathbb{E}\left[\int_0^\tau \sum_{l=1}^{L} D_i^l \mu(G_i, t) G_i^k Y_i(t) \frac{\kappa_l(t)}{\tilde{s}_l(t)} dt \right] = \sum_{l=1}^{L} \int_0^\tau \tilde{e}_{k,l}(t) \kappa_l(t) dt,$$

Et

$$\mathbb{E}\left[\int_0^\tau \sum_{l=1}^{L} D_i^l \, \mu(G_i, t) \tilde{e}_{k,l} Y_i(t) \frac{\kappa_l(t)}{\tilde{s}_l(t)} dt \right] = \sum_{l=1}^{L} \int_0^\tau \tilde{e}_{k,l}(t) \kappa_l(t) dt$$

Donc

$$\mathbb{E}\left[Q_{i,k} \right] = \int_0^\tau \mathbb{E}\left[\mu(G_i, t) G_i^k Y_i(t) \zeta_i(t) \right] dt - \sum_{l=1}^{L} \int_0^\tau \tilde{e}_{k,l}(t) \kappa_l(t) dt$$

$$:= B_{1,k} - B_{2,k}$$

Nous montrons que sous H_0, $B_{2,k} = \int_0^\tau \mathbb{E}[G^k] \, \mathbb{E}[Y(t) \zeta(t) \mu(G, t)] dt$. Sous la condition *C1* et puisque la distribution de T^0 ne dépend pas de *G* sous H_0 alors nous pouvons remarquer:

$$\tilde{s}_{k,l}(t) = \mathbb{E}[Y(t)G^k D^l \mu(G,t)]$$

$$= \mathbb{E}\Big[\mathbb{E}[Y(t)G^k D^l \mu(G,t)|G,S,W,R]\Big]$$

$$= \mathbb{E}\Big[G^k D^l \mu(G,t)\mathbb{E}[Y(t)|G,S,W,R]\Big]$$

$$=_{H_0} \mathbb{E}\Big[G^k D^l \mu(G,t)\mathbb{E}[1(T^0 \geq t)|S]\mathbb{E}[1(C \geq t)|G]\Big]$$

$$= \mathbb{E}\Big[G^k D^l \mathbb{E}[1(T^0 \geq t)|S]\Big]$$

$$= \mathbb{E}[G^k]\mathbb{E}\Big[D^l \mathbb{E}[1(T^0 \geq t)|S]\Big]$$

Nous pouvons écrire :

$$\tilde{s}_{k,l}(t) = \mathbb{E}[G^k]\mathbb{E}\Big[D^l \mathbb{E}[1(T^0 \geq t)|S]\mu(G,t)\mathbb{E}[1(C \geq t)|G]\Big]$$

$$= \mathbb{E}[G^k]\mathbb{E}\Big[D^l \mu(G,t)\mathbb{E}[Y(t)|G,S,W,R]\Big]$$

$$= \mathbb{E}[G^k]\tilde{s}_l(t).$$

Donc sous H_0,

$$B_{2,k} = \sum_{l=1}^{L} \int_0^{\tau} \mathbb{E}[G^k]\kappa_l(t)dt$$

$$= \int_0^{\tau} \mathbb{E}\,[G^k]\mathbb{E}[Y(t)\zeta(t)\mu(G,t)]dt$$

De la même manière, nous pouvons montrer que sous *H0* :

$$B_{1,k} = \int_0^{\tau} \mathbb{E}\,[G^k]\mathbb{E}[Y(t)\zeta(t)\mu(G,t)]dt$$

Et donc $\mathbb{E}\big[Q_{i,k}\big] = 0.$■

ANNEXE 3.B : Preuves des Lemme 3.2 et Lemme 3.3

Preuve du Lemme 3.2

Soit $l \in \{1, \dots, L\}$. Posons

$$n^{\frac{1}{2}} \int_0^{\tau} v_{k,l}^{(n)}(t) \left[\frac{1}{n} \sum_{i=1}^{n} \mu(G_i, t) D_i^l \, dN_i(t) - \kappa_l(t) dt \right] = C_{1,k,l}^{(n)} + C_{2,k,l}^{(n)}$$

Où :

$$C_{1,k,l}^{(n)} = n^{\frac{1}{2}} \int_0^{\tau} v_{k,l}^{(n)}(t) \left[\frac{1}{n} \sum_{i=1}^{n} \mu(G_i, t) D_i^l \, dN_i(t) - \frac{1}{n} \sum_{i=1}^{n} \mu(G_i, t) D_i^l Y_i(t) \zeta_i(t) dt \right]$$

Et

$$C_{2,k,l}^{(n)} = \int_0^{\tau} v_{k,l}^{(n)}(t) \, n^{\frac{1}{2}} \left[\frac{1}{n} \sum_{i=1}^{n} \mu(G_i, t) D_i^l Y_i(t) \zeta_i(t) dt - \kappa_l(t) dt \right]$$

Dans ce qui suit, nous montrons que $C_{1,k,l}^{(n)} \xrightarrow{\mathcal{P}} 0$ quand $n \to \infty$. Notons que $C_{1,k,l}^{(n)}$ a la forme :

$$C_{1,k,l}^{(n)} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{\tau} H_{i,k,l}^{(n)}(t) \, dM_i(t)$$

Où $M_i(t) = N_i(t) - \int_0^{\tau} Y_i(s) \zeta_i(s) ds$ et $H_{i,k,l}^{(n)}(t) = v_{k,l}^{(n)}(t) \mu(G_i, t) D_i^l$ est un processus prévisible par rapport à $\mathcal{H}_t = \bigvee_{i=1}^{n} \mathcal{H}_{t,i}$. Le processus $H_{i,k,l}^{(n)}(t)$ est borné sur $[0, \tau]$ puisque $\left| H_{i,k,l}^{(n)}(t) \right| \leq \left| v_{k,l}^{(n)}(t) \right| . \frac{1}{c_0} \leq \frac{2}{c_0}$

On définit le processus $\left(C_{1,k,l}^{(n)}(t) \right)_{t \geq 0}$ par :

$$C_{1,k,l}^{(n)}(t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{t} \int_0^{\tau} H_{i,k,l}^{(n)}(s) \, dM_i(s)$$

Alors $\left(C_{1,k,l}^{(n)}(t)\right)_{t\geq 0}$ est une martingale et $C_{1,k,l}^{(n)} = C_{1,k,l}^{(n)}(\tau)$. La variation $< C_{1,k,l}^{(n)} > (t)$de $C_{1,k,l}^{(n)}(t)$ est donnée par :

$$< C_{1,k,l}^{(n)} > (t) = \int_0^t \frac{1}{n}\sum_{i=1}^{n}\left\{H_{i,k,l}^{(n)}(s)\right\}^2 Y_i(s)\zeta_i(s)ds = \int_0^t X^{(n)}(s)ds$$

Pour tour $s \in [0,\pi]$ et $n \geq 1$:

$$\left|X^{(n)}(s)\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\left\{H_{i,k,l}^{(n)}(s)\right\}^2 Y_i(s)\zeta_i(s)\right|$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\left\{H_{i,k,l}^{(n)}(s)\right\}^2 \zeta_i(s)$$

$$\leq \frac{4}{c_0^2}c_1$$

Par la proposition II.5.3 de Andersen *et al* (1993), $< C_{1,k,l}^{(n)} > (t) \xrightarrow{\mathcal{P}} 0$ quand $n \to \infty$. Soit pour tout $\varepsilon > 0$ et $t \in [0,\tau]$:

$$C_{1,k,l,\varepsilon}^{(n)}(t) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^t \left\{H_{i,k,l}^{(n)}(s)\right\}^2 1\left(n^{-\frac{1}{2}}\left|H_{i,k,l}^{(n)}(s)\right| \geq \varepsilon\right) dM_i\,(s)$$

Alors $< C_{1,k,l,\varepsilon}^{(n)} > (t)$ satisfait ce qui suit:

$$C_{1,k,l,\varepsilon}^{(n)}(t) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^t \left\{H_{i,k,l}^{(n)}(s)\right\}^2 1\left(n^{-\frac{1}{2}}\left|H_{i,k,l}^{(n)}(s)\right| \geq \varepsilon\right) Y_i(s)\zeta_i(s)\,ds$$

$$\leq < C_{1,k,l}^{(n)} > (t)$$

Il en résulte que $< C_{1,k,l}^{(n)} > (t) \xrightarrow{\mathcal{P}} 0$ quand $n \to \infty$. D'après le théorème 5.1.1 de Fleming and Harrington(1991), $C_{1,k,l}^{(n)}(t) \xrightarrow{\mathcal{L}} 0$ quand $n \to \infty$ pour tout $t \in [0,\tau]$, et donc $C_{1,k,l}^{(n)}(t) \xrightarrow{\mathcal{P}} 0$ quand $n \to \infty$. En particulier, si $t = \tau$, $C_{1,k,l}^{(n)} \xrightarrow{\mathcal{L}} 0$ quand $n \to \infty$.

Nous montrons également que $C_{2,k,l}^{(n)} \xrightarrow{\mathcal{P}} 0$ quand $n \to \infty$. On a

$$\sup_{t \in [0,\tau]} \left| v_{k,l}^{(n)}(t).n^{-\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^{n} \mu(G_i,t) D_i^l Y_i(t) \zeta_i(t) - \kappa_l(t) \right) \right|$$

$$\leq \sup_{t \in [0,\tau]} \left| v_{k,l}^{(n)}(t) \right| . \sup_{t \in [0,\tau]} \left| n^{-\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^{n} \mu(G_i,t) D_i^l Y_i(t) \zeta_i(t) - \kappa_l(t) \right) \right| \qquad (3.4)$$

$\tilde{\mathcal{E}}_{k,l}^{(n)}(t)$ est uniformément consistant pour $\tilde{e}_{k,l}(t)$. Pour $n \to \infty$, il s'en suit :

$$\sup_{t \in [0,\tau]} \left| v_{k,l}^{(n)}(t) \right| \xrightarrow{\mathcal{P}} 0 \qquad (3.5)$$

Donc, les classes $\{Y(t), t \in [0,\tau]\}$ et $\{\mu(G,t), t \in [0,\tau]\}$ sont Donsker (d'après le Lemme 4.1 dans Kosorok (2008)), et il est de même pour$\{\mu(G,t)Y(t), t \in [0,\tau]\}$ puisque le produit de classes Donsker est Donsker (corrollaire 9 .32 dans Kosorok (2008)) et $\{D^l\}$ est Donsker. La classe $\{\mu(G,t)D^l Y(t), t \in [0,\tau]\}$ est donc Donsker et le processus $G_n(.) \coloneqq n^{\frac{1}{2}} \left(n^{-1} \sum_{i=1}^{n} \mu(G_i,.) D_i^l Y_i(.) \zeta_i(.) - \kappa_l(.) \right)$ converge faiblement vers un processus gaussien G de moyenne nulle. D'après continuous mapping theorem , on a $\sup_{t \in [0,\tau]} |G_n(t)|$ converge faiblement vers $\sup_{t \in [0,\tau]} |G(t)|$ et donc $\sup_{t \in [0,\tau]} |G_n(t)| = O_p(1)$. En combinant ce résultat et résultat (3.5) dans (3.4), on obtient :

$$\sup_{t \in [0,\tau]} \left| v_{k,l}^{(n)}(t).n^{-\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^{n} \mu(G_i,t) D_i^l Y_i(t) \zeta_i(t) - \kappa_l(t) \right) \right| \xrightarrow{\mathcal{P}} 0$$

quand $n \to \infty$. Ceci implique que $C_{2,k,l}^{(n)} \xrightarrow{\mathcal{P}} 0$ quand $n \to \infty$.

Puisque pour tout $l \in \{1, \dots, L\}$, $C_{1,k,l}^{(n)} + C_{2,k,l}^{(n)} \xrightarrow{\mathcal{P}} 0$ lorsque $n \to \infty$ alors

$$\sum_{l=1}^{L} C_{1,k,l}^{(n)} + C_{2,k,l}^{(n)} \xrightarrow{\mathcal{P}} 0.$$

Ceci conclue la preuve du Lemme 3.2

Preuve du Lemme 3.3

Décomposons :

$$n^{\frac{1}{2}} \sum_{l=1}^{L} \int_0^{\tau} \tilde{\mathcal{E}}_{k,l}^{(n)}(t)\, \kappa_l(t) dt$$

$$= n^{\frac{1}{2}} \sum_{l=1}^{L} \int_0^{\tau} \left\{ \tilde{e}_{k,l}(t) + \frac{\tilde{S}_{k,l}^{(n)}(t)}{\tilde{s}_l(t)} - \frac{\tilde{s}_{k,l}(t)\tilde{S}_l^{(n)}(t)}{\tilde{s}_l(t)^2} \right\} \kappa_l(t) dt$$

$$+ \sum_{l=1}^{L} \int_0^{\tau} \left\{ \frac{\tilde{S}_{k,l}^{(n)}(t)}{\tilde{s}_l(t)} - \frac{\tilde{s}_{k,l}(t)\tilde{S}_l^{(n)}(t)}{\tilde{s}_l(t)^2} \right\} . n^{\frac{1}{2}} \left(\frac{\tilde{s}_l(t)}{\tilde{S}_l^{(n)}(t)} - 1 \right) \kappa_l(t) dt \qquad (3.6)$$

En utilisant les mêmes arguments que dans la preuve du Lemme 4.2, nous pourrons montrer que la deuxième somme de la partie droite de (3.6) converge en probabilité vers 0 lorsque n tend vers l'infini.

ANNEXE 3.C : Programme de simulation

Cette partie du programme correspond à des pourcentages de censure de 5% et 20% dans les deux groupes respectivement, 20% de données manquantes et n=100.

```
r1=1            # rapport du risque instantanne entre 2 ind de (S1,G1) et (S1,G2)
r2=1            # rapport du risque instantanne entre 2 ind de (S2,G1) et (S2,G2)

th1=.03      #pourcentage de censure dans G1 (5%)
th2=.195    # pourcentage de censure dans G2 (20%)

# 1) Initialisation des données

P=2   #nombre de covariables
L=2   #nombre de strates
K=2    #nombre de groupes
n=100  #effectif total
n1=n/2 ;  n2=n/2   #effectif du groupe 1  et du groupe 2
l1=.75   # lambda ds strate 1
l2=1.5   # lambda ds strate 2
a1=.5    #alpha1
a2=.75   #alpha2
b0=0 ;  b1=1.5 ; b2=.01#à choisir pour avoir 1/2 des individus dans chque strate
groupe=c(rep(1,n1),rep(2,n2)) #groupes 1 et 2
indG =matrix(nrow= n,ncol=K)       #indicatrice d'appart aux groupes
indG[,1]= as.integer(groupe==1); indG[,2]=1-indG[,1]
h=1

###################################20% données manquantes

a=.85 ; g=16    # 20% données manquantes
w=matrix(nrow=n,ncol=P)        # covariables
w[,1]=runif(n,-1,1);w[,2]=rnorm(n,0,.5)
pr=exp(a*w[,1]+g*w[,2]^2)/(1+exp(a*w[,1]+g*w[,2]^2))
R=rbinom(n,1,pr)

#####   estimation de p(S=l/Z)
Z=matrix(nrow=n,ncol=P)        # covariables pour modele logistique
Z[,1]=runif(n,-1,1);Z[,2]=rnorm(n,0,.5)
prS=exp(b0+b1*Z[,1]+b2*Z[,2]^2)/(1+exp(b0+b1*Z[,1]+b2*Z[,2]^2))
S=rbinom(n,1,prS)
pSchap=matrix(nrow=n,ncol=L)
pSchap[,1]=predict(locfit.raw(Z[R==1],S[R==1],cens=0,family="binomial",link="logit"),
Z) #prob(S=1/Z)
pSchap[,2]=1-pSchap[,1]
indS=matrix(nrow=n,ncol=L)   #indicatrice d'appart aux strates
```

```
indS[,1]=as.integer(S==1);indS[,2]=1-indS[,1]
D=matrix(nrow=n,ncol=L)   #matrice des poids Dil
D=R*indS+(1-R)*pSchap
nS1=S[groupe==1]
n11=length(nS1[nS1==1])    #effectif de strate 1 du groupe 1
n21=n1-n11              #effectif de strate 2 du groupe 1
nS2=S[groupe==2]
n12=length(nS2[nS2==1])    #effectif de strate 1 du groupe 2
n22=n2-n12              #effectif de strate 2 du groupe 2
X11=(-log(runif(n11))/l1  )^(1/a1)         #durées de vie de S1 de G1
X21=(-log(runif(n21))/l2  )^(1/a2)         #durées de vie de S2 de G1
X12=(-log(runif(n12))/(l1*r1)  )^(1/a1)        #durees de vie de S1 de G2
X22=(-log(runif(n22))/(l2*r2 ) )^(1/a2)        #durees de vie de S2 de G2
X1=c(X11,X21) ;  X2=c(X12,X22)                      #pour G2
C1= rexp(n1,th1) ;   C2= rexp(n2,th2)                #censure pour G1 et G2
C=c(C1,C2)
T1=pmin(X1,C1) ; T2=pmin(X2,C2)
T=c(T1,T2)
delta1=as.integer(T1==X1) ; delta2=as.integer(T2==X2) # cree les indicatrices delta pour
G1 et G2
delta=c(delta1,delta2)

######calcul avec mu estime

########estimateur de KM de p(C>=T/G)
deltac1=as.integer(C1<=X1)
if(deltac1[which.max(T1)]==1)
{T1[which.max(T1)]=X1[which.max(T1)];deltac1[which.max(T1)]=0}
deltac2=as.integer(C2<=X2)

if(deltac2[which.max(T2)]==1)
{T2[which.max(T2)]=X2[which.max(T2)];deltac2[which.max(T2)]=0}

pkm1=survfit(Surv(T1,deltac1)~1)  #estimateur de KM pour p(C>=T/G)
pkm2=survfit(Surv(T2,deltac2)~1)
mu=matrix(nrow=n,ncol=n)

for(i in 1:n1){for (j in 1:n){if (groupe[j]==1)  {if sum(as.integer(pkm1$time<=T1[i]))!=0)
{mu[i,j]=1/pkm1$surv[sum(as.integer(pkm1$time<=T1[i]))]} else {mu[i,j]=1}}  else {if
(sum(as.integer(pkm2$time<=T2[i]))!=0)
{mu[i,j]=1/pkm2$surv[sum(as.integer(pkm2$time<=T2[i]))]} else {mu[i,j]=1}} }  }

n1p=n1+1

for (i in n1p:n){for (j in 1:n){if (groupe[j]==1)  {if (sum(as.integer(pkm1$time<=T1[i-
n1]))!=0) {mu[i,j]=1/pkm1$surv[sum(as.integer(pkm1$time<=T1[i-n1]))]} else
{mu[i,j]=1}}  else {if (sum(as.integer(pkm2$time<=T2[i-n1]))!=0)
{mu[i,j]=1/pkm2$surv[sum(as.integer(pkm2$time<=T2[i-n1]))]} else {mu[i,j]=1}} }  }
```

```
MatrMu=matrix(nrow=n,ncol=n) ;MatrMu=mu

#2) Fonction qui calcule Sn,k (t,l) Numerateur de E*n,k(t,l)

Snk=function(i,l,k){ value=0

for (j in 1:n){value=value+D[j,l]*mu[i,j]*(T[j]>=T[i])*(indG[j,k]==1)
value=value/n
value}

#3) Fonction qui calcule Snpoint (t,l) Denominateur de E*n,k(t,l)

Snpoint=function(i,l){value=0

for (j in 1:n){value=value+(T[j]>=T[i])*D[j,l]*mu[i,j]}
value=value/n
value}

#4) Fonction qui calcule E*n,k(t,l)

Enk=function(num,denom){#num est la valeur de Sn,k(t,l), denom est la valeur de
Snpoint(t,l)

if (denom==0){Enk=0}else{Enk=num/denom}
Enk}#fin fonction Enk

#4p) Fonction qui calcule Ztilde :

Ztilde=function(k){value=0
for (i in 1:n) {som=0
for(l in 1:L){som=som+D[i,l]*Enk(Snk(i,l,k),Snpoint(i,l))}
value=value+(indG[i,k]-som)*delta[i]*mu[i,i]    }

value}

#5) Calcul de la fonction Qchap(k,i):

Qkichap=function(k,i){

#calcul du premier terme de la différence

som1=0
for (l in 1:L){epsi=Enk(Snk(i,l,k),Snpoint(i,l))
if(epsi!=0){som1=som1+epsi*D[i,l]}}

#calcul du deuxieme terme de la différence
som2=0;value=0
for (j in 1:n){som3=0
```

```
for (l in 1:L){Snpointj=Snpoint(j,l)
epsi=Enk(Snk(j,l,k),Snpointj)
if(Snpointj==0 || epsi!=0){som3=0
break}
if (T[i]>=T[j]){den=mu[i,i]*mu[i,j]*D[i,l]*D[j,l]*delta[j]*(indG[i,k]-epsi)
som3=som3+den/Snpointj}}
som2=som2+som3}
value=mu[i,i]*delta[i]*(indG[i,k]-som1)-som2/n
value}

#5p) Calcul de la fonction Sigmachap:

Sigmachap=function(n){ value=0
for (i in 1:n){value=value+Qkichap(K-1,i)^2}
value}

#6) Calcul de la fonction U:

U=function(n){return(Ztilde(K-1)^2/Sigmachap(n))}
#calcul avec mu estm pour 20% de donnees mqtes
u20[h]=U(n)
pu20[h]=as.integer(u20[h]>qchisq(0.95,1))

#calcul pour cas complet

LRS=survdiff(Surv(T[R==1],delta[R==1])~groupe[R==1]+strata(S[R==1]))
puCC20[h]=as.integer(LRS$chisq>qchisq(0.95,1))
```

CONCLUSION GENERALE

Dans ce travail, nous nous sommes intéressés au problème de l'inférence statistique dans des modèles de durées de vie avec données manquantes.

Dans ce contexte, nous avons adapté le principe de "pondération par probabilité inverse" (Inverse Weighted Probability (IWP)).

Dans une première partie, nous avons proposé une méthode d'estimation dans un modèle de transformation linéaire, en présence de données manquantes. Cette méthode permet d'inclure tous les individus dans l'analyse. Nous avons montré les bonnes propriétés asymptotiques de l'estimateur proposé.

L'implémentation informatique de la méthode proposée a été mise en œuvre avec le logiciel statistique R. Les propriétés numériques de l'estimateur proposé ont ensuite pu être évaluées à l'aide d'études de simulation. Nous avons notamment comparé cet estimateur avec l'estimateur obtenu par la méthode dite en "cas complet" et avons montré sa supériorité sous différentes conditions de taille d'échantillon, pourcentage de censure et de données manquantes.

Nous avons supposé que les données sont manquantes au hasard (Missing At Random (MAR)). Le cas où le mécanisme de données manquantes est non-ignorable peut constituer une intéressante future piste de recherche.

On pourra également s'intéresser à la robustesse des résultats obtenus, lorsque le modèle $\mathbb{P}(R_i = 1|Z_i)$ n'est pas correctement spécifié.

Dans une seconde partie, nous nous sommes intéressés au test du log-rank stratifié lorsque la variable renseignant l'appartenance aux strates est manquante au hasard (MAR) et lorsque la censure dépend du groupe de traitement.

En combinant la méthode de pondération par l'inverse de probabilité de censure et le principe de régression par calibration, nous avons proposé une nouvelle statistique et établi sa distribution asymptotique sous l'hypothèse nulle d'égalité des distributions de survie. Une étude de simulation nous a permis de montrer que cette nouvelle statistique est plus appropriée comparée à celle du test du log-rank stratifié basé sur l'analyse en cas complet. L'extension du test proposé peut être examinée dans le cas de données manquantes non-ignorables. On peut étudier également le cas où les strates sont manquantes dépendamment du groupe de traitement. Pour appliquer le principe de pondération par l'inverse de probabilité de censure, nous avons choisi une fonction de poids $\mu(G_i, t) = h\big(\mathbb{P}(C \geq t|G_i)\big)$ et $h(x) = {}^{1}/_{x}$. D'autres alternatives de la statistique du test peuvent être obtenues en spécifiant d'autre formes pour la fonction h. La recherche de la meilleure fonction conduisant à la plus efficace procédure du test, constitue une question intéressante.

REFERENCES

1. Aalen, O. O., ”Nonparametric inference for a family of counting processes”. Annals of Statistics, 6, (1978b),701-726.
2. Afifi, A. Elashoff, R., ”Missing observations in multivariate statistics I: Review of the literature”, Journal of the American Statistical Association, 61, (1966), 595-604.
3. Andersen, P. K. , Borgan, Ø., Gill, R. D. et Keiding, N. ”Statistical Models based on Counting Processes ”, New York: Springer, (1993).
4. Bagdonavicius, V. et Nikulin, M. S., ”Accelerated Life Models. Modeling and Statistical Analysis”, Chapman & Hall, (2002).
5. Cain, L.E., Cole, S.R.,’’ Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. Statistics in Medicine, 28, (2009), 1725-1738.
6. Carroll, R. J., Fan, J., Gijbels, I. et Wand, M. P., ’’Generalized partially linear single-index models’’. Journal of the American Statistical Association, 92,(1997),477-489.
7. Chen, K., Jin, Z. et Ying, Z., ’’Semiparametric Analysis of Transformation Models with Censored Data’’,Biometrika, 89, (2002),659-668.
8. Cheng, S.G., Wei, I. J. et Ying, Z., ” Analysis of transformation Models with Censored Data”, Biometrika, Vol. 82, No.4, (1995), 835-845.
9. Cox, D.R. ”Regression models and life tables (with discussion)”, J. Roy. Statist. Soc. Ser. B34, (1972), 187-220.
10. Cox, D. R. ’’Partial likelihood’’. Biometrika, 62, (1975), 269-276.
11. Dabrowska, D. M. ’’ Bandwidth conditional Kaplan-Meier estimate variable’’, Scandinavian Journal of Statistics 19,(1992), 351-361.
12. Dupuy, J.-F., ”Transformation models for failure time data: an overview of some recent developments”, Proceedings of the Second International Conference on Accelerated Life Testing in Reliability and Quality Control, (2008), 43-47, Bordeaux, France.

13. Dupuy, J.-F. et Leconte, E., ”Cox regression with missing values of a covariate having a non-proportional effect on hazard of failure”, Mathematical methods in survival analysis, reliability and quality of life (Applied Stochastic Methods series, (2008), 133-150, ISTE: London.
14. Dupuy, J.-F. et Leconte, E., ”A study of regression calibration in a partially observed stratified Cox model ’’. Journal of Statistical Planning and Inference, 139, (2009), 317-328.
15. Fahrmeir L. et Kaufmann H., Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics, 13, (1985), 342-368.
16. Fine, J.P. Ying, Z., et Wei, L.J., ’’On the linear transformation model for censored data’’, Biometrika,85,(1998),980-986.
17. Fleming, T. R. et Harrington, D.P., ’’Counting processes and survival analysis’’, Wiley & Sons: New York, (1991).
18. Fleming, T. R. et Lin, D. Y., ”Survival analysis in clinical trials: past developments and future directions”, Biometrics, 56, (2000), 971-983.
19. Foutz, R. V.,’’ On the unique consistent solution to the likelihood equations’’, Journal of the American Statistical Association, 72, (1977), 147-148.
20. Gill, R. D.,”Censoring and Stochastic Integrals”, Mathematical Center Tracts, No.124, Amsterdam: Mathematical Centrum, (1980).
21. Gill, R. D. Van der Laan, M. J., et Robins, J. M., ” Coarsening at random”, in 1st Seattle symposium in biostatistics: survival analysis, (1997), 255-294. (eds. D. Y. Lin), Springer.
22. Hartley, H. O. et Hocking, R., ”The analysis of incomplete data”, Biometrics, 27, (1971), 783-808.
23. Hoeffding, W. ’’The strong law of large numbers for U-statistics’’. Institute of Statistics Mimeo Series No. 302, University of North Carolina, Chapel Hill, N. C., 1961.
24. Jacobsen, M. et Keiding,N., ”Coarsening at Random in General Sample Spaces and Random Censoring in Continuous Time”, The Annals of Statistics, 23, No. 3 (1995), 774-786.

25. Klein, J. P. et Moeschberger, M. L., '' Analysis: Techniques for Censored and Truncated Data'', Survival Analysis, Springer: New York, (1997).
26. Kong, L., Cai, J., et Sen, P.K.,''Weighted estimating equations for semiparametric transformation models with censored data from case-cohort design'', Biometrika,91,(2004),305-319.
27. Kong, L., Cai, J., et Sen, P.K., ''Asymptotic results for fitting semiparametric transformation models to failure time data from case-cohort studies", Statistica Sinica, 16, (2006),135-151.
28. Kosorok, M. R. ''Introduction to empirical processes and semi parametric inference '', Springer, New York (2008).
29. Kosorok, M. R. et Song, R., ''Inference under right censoring for transformation models with a change-point based on a covariate threshold'', Annals of Statistics, 35, (2007), 957-989.
30. Kowalski, J., Tu et Xin M.,'' Modern Applied U-Statistics'', Wiley-Interscience, (2007).
31. Lawless, J. F. ''Statistical models and methods for lifetime data''. Wiley Series in Probability and Statistics.Wiley, Hoboken, second édition,(2003).
32. Lee, E. T. et Wang, J. W., '' Statistical Methods for Survival Data Analysis'', Third Edition Wiley-Interscience, (2003).
33. Little, R.J. et Rubin, B.D., '' Statistical Analysis with Missing Data'', John Wiley, New York, (1987).
34. Loader, C. ''Local Regression and Likelihood'', Springer: New York ,(1999).
35. Ma, S. et Kosorok, M. R. ''Penalized log-likelihood estimation for partly linear transformation models with current status data'', The Annals of Statistics, 33, (2005), 2256-2290.
36. Martinussen, T. et Scheike, T. H., "Dynamic Regression Models for Survival Data", Springer: New York, (2006).
37. Meeker, W. Q. et Escobar, L.A., "Statistical Methods for Reliability Data", Wiley, New York, (1998).
38. Mezaouer, A., Dupuy, J.-F et Boukhetala, K.," A nonparametric test for comparing treatments with missing data and dependent censoring". In : Statistical Models and

Methods for Reliability and Survival Analysis. (Eds, Couallier V. et al.), 297-310, ISTE-Wiley, (2013).

39. Mezaouer, A., Dupuy, J.-F et Boukhetala, K., ”Estimation dans le modèle de transformation linéaire avec données manquantes ’’, Journal de la Société Française de Statistique, vol.155,N°3 (2014), 120-134.
40. Molenberghs, G. et Kenward , M.,G., ” Missing Data in Clinical Studies”, statistics in practice, Wiley,(2007).
41. Murphy, S., Rossini A. et van der Vaart A. ,” Maximum likelihood estimation in the proportional odds model”, Journal of the American Statistical Association (1997), 968-976.
42. Nelson, W., ”Hazard plotting for incomplete failure data” J. Qual. Technol., 1, (1969), 27-52.
43. Nelson, W., ” Theory and applications of hazard plotting for censored failure data”, Technometrics 14, (1972), 945-965.
44. Newey, W. K., ’’Uniform convergence in probability and stochastic equicontinuity’’. Econometrica, 59,(1991),1161-1167.
45. Pons, O., ”Estimation in the Cox model with missing covariate data”, J. Nonparametric Statist., 14, (2002), 223-247.
46. Robolledo, R., ’’Central limit theorems for local martingales’’, Z. Wahrsch. Verw. Geb., 51, (1980), 269-286.
47. Robins, J.M. et Finkelstein, D.M., ’’ Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests’’. Biometrics, 56, (2000), 779-788.
48. Rubin, D.B., ”Inference and missing data”. Biometrika, 63, (1976), 581-592.
49. Rudin, W. ’’ Principles of Mathematical Analysis’’. McGraw-Hill, New York, (1964).
50. Seaman, S. R., et White, I. R., ’’Review of inverse probability weighting for dealing with missing data’’. Statistical Methods in Medical Research, 22, (2013), 278-295.
51. Slud, E. V. et Vonta, F., ”Consistency of the NPML estimator in the right-censored transformation model”, Scandinavian Journal of Statistics, 31, (2004), 21-41.

52. Strawderman, R. L. et Tsiatis, A. A., '' On consistency in parameter spaces of expanding dimension : an application of the inverse function theorem'', Statistica Sinica, 6, (1996), 917-923.
53. Thurston, S. W., Spiegelman, D. et Ruppert, D., '' Equivalence of regression calibration methods in main study/external validation study designs'', Journal of Statistical Planning and Inference, 113, (2003),527-539.
54. Tsiatis, A., "Semiparametic Theory and Missing Data", Springer, (2006).
55. Van der Laan, M.J. et McKeague, I.W., "Efficient estimation from right-censored data when failure indicators are missing at random", The Annals of Statistics, 26, (1998),164-182.
56. Van der Vaart, A. W., " Asymptotics Statistics", Cambridge University Press, (1998).
57. Weller, E.A., Milton, D.K., Eisen, E.A., et Spiegelman, D., ''Regression calibration for logistic regression with multiple surrogates for one exposure'', Journal of Statistical Planning and Inference, 137, (2007), 449-461.
58. Yoshida, M., Matsuyama, Y. et Ohashi, Y., ''Estimation of treatment effect adjusting for dependent censoring using the IPCW method: an application to a large primary prevention study for coronary events (MEGA study)'', Clinical Trials, 4, (2007), 318-328.