UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLES DE SURVIE AVEC UN POINT DE RUPTURE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

YASSIR RABHI

JANVIER 2006

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.01-2006). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Avant de commencer, je tiens à remercier Madame Brenda MacGibbon, ma directrice de recherche. Ses conseils judicieux et son encouragement ont été précieux. Sa disponibilité et le temps qu'elle a consacré à répondre à mes questions ont facilité la rédaction de ce mémoire, sans oublier son soutien financier.

Mes sincères remerciements à Monsieur Omar Cherkaoui pour m'avoir aidé à réaliser ce travail, et à mon professeur Glenn Shorrock pour la formation académique qu'il m'a apportée à travers ses cours de statistiques.

Enfin, je tiens à remercier mes parents, ma famille et mon ami Hocine qui m'ont grandement soutenu et encouragé constamment à réaliser mes rêves d'accomplir mes études supérieures.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Matthews et Farewell (1982, 1985) ont proposé le modèle de risque instantané constant avec un point de rupture pour modéliser les temps de rechute après le traitement « remission induction ». Dans ce mémoire, des études théoriques et simulées des statistiques de score et du test des rapports de vraisemblances sont présentées, afin de tester l'existence d'un temps de rupture dans le modèle de risque instantané constant. Les méthodes d'estimation du point de rupture et des autres paramètres du modèle sont aussi discutées dans la première partie de ce travail, avec une application du modèle sur des données réelles.

Le modèle des risques proportionnels avec un point de rupture de Liang *et al.*(1990) est discuté dans la deuxième moitié du mémoire comme un cas spécial du modèle de Cox avec covariables dépendantes du temps. Les caractéristiques de ce modèle et quelques difficultés d'inférence liées à celui-ci sont également présentées. À titre d'exemple, le modèle est appliqué à une série de données recueillies dans un groupe de patients qui suivent deux traitements différents du cancer du poumon.

Mots clés : Temps de rupture, modèle de risque constant avec un point de rupture, modèle de Cox avec un point de rupture, statistique de score, statistique du test des rapports de vraisemblances.

INTRODUCTION

L'analyse de la survie est née au vingtième siècle, et a connu un développement important dans la seconde moitié du siècle. Les développements dans ce domaine, qui ont eu le plus profond impact sur les essais cliniques, sont la méthode de Kaplan-Meier (1958) pour l'estimation de la fonction de survie, la statistique du log-rank (Mantel, 1966) pour comparer deux distributions de survie, et le modèle des risques proportionnels (Cox, 1972) pour quantifier les effets de covariables sur le temps de survie.

La théorie des martingales pour processus de comptage, élaborée par Aalen (1975), offre un cadre unifié pour l'étude des propriétés des statistiques d'analyse de la survie, aussi bien pour les petits que pour les grands échantillons. Des progrès significatifs ont été réalisés, et on peut espérer de nouveaux développements dans plusieurs domaines, comme le modèle à temps accéléré, les données de survie multivariées, les données censurées par intervalle, les protocoles de traitement dynamiques et l'inférence causale, la modélisation jointe de données longitudinales et de données de survie, et les méthodes bayésiennes.

Il arrive fréquemment lors d'essais cliniques que l'un des objectifs soit d'étudier l'impact d'une nouvelle thérapie sur le taux de risque des patients, et de savoir si celle-ci produit un écart significatif dans le taux de risque après le début du traitement avec risque constant. Il serait donc intéressant d'identifier le point de rupture de ce taux de risque, et de déterminer par conséquent l'intervalle du temps dans lequel le traitement était assez efficace.

Le test d'une rupture dans une séquence de variables $T_1, T_2, ..., T_n$ en un point inconnu où la distribution change de forme de $F(t, \theta_1)$ pour $T_1, ..., T_m$, à $F(t, \theta_2)$ pour $T_{m+1}, ..., T_n$, $(m < n)$, a reçu un intérêt considérable de la part de nombreux chercheurs, notamment

Hinkley (1970), Hawkins (1977), Hinkley et Hinkley (1970), Worsley (1986) et James, James et Siegmund (1987).
Ce mémoire considère ce point en détails pour une forme particulière du problème avec point de rupture : celle qui teste des données de survie pour un risque instantané constant contre l'alternative d'un risque avec une rupture en un temps inconnu. Une autre forme de problème est discutée, celle-ci met le doigt sur les covariables dépendantes du temps avec un seul temps de rupture dans le modèle des risques proportionnels de Cox.

Matthews et Farewell (1982,1985) étaient les premiers à considérer le problème du test pour un point de rupture dans la fonction de risque qui n'est pas un indice des observations. En utilisant les techniques dc simulation, ils ont estimé les valeurs critiques du test des rapports de vraisemblances pour un point de rupture. Worsley (1988) élimine la singularité de la fonction de vraisemblance (en censurant la dernière observation), et utilise des formules exactes pour trouver les valeurs critiques du test des rapports de vraisemblances. Ceci va être le sujet du deuxième chapitre de ce mémoire.

Matthews, Farewell et Pyke (1985) proposent, parallèlement au travail de Davies (1977), une autre alternative pour tester le point de rupture dans la fonction de survie instantanée basée sur le processus de la statistique du score.
Liang, Self et Liu (1990) s'en sont inspirés pour développer le modèle des risques proportionnels de Cox pour un cas spécial des covariables dépendantes du temps, où celles-ci ne changent leurs valeurs qu'une seule fois en fonction de temps, et utilisent la statistique du score pour tester le point de rupture dans la fonction de risque instantané. Tout ceci sera discuté dans les chapitre 3 et 4 de ce travail.

Une introduction de certains concepts de base de l'analyse de la survie sera donnée dans le chapitre 1, et sera accompagnée d'un aperçu sur les trois approches de l'analyse de la survie : paramétrique, semi-paramétrique et non paramétrique.

CHAPITRE I

INTRODUCTION À L'ANALYSE DE LA SURVIE

Les données de survie se distinguent par une discipline statistique particulière. Ce premier chapitre essaie d'une part, de décrire ce que sont les données de survie et les données censurées, et d'autre part, de donner un aperçu des modèles de survie paramétriques et semi paramétriques, et des méthodes non paramétriques.

1.1 Données de survie et censure

1.1.1 Données de survie

Les données de survie représentent le temps écoulé entre le début d'une observation et l'arrivée d'un événement. Le cas d'événement le plus simple est le décès ; cependant, le terme « donnée de survie » couvre d'autres événements, comme l'apparition d'une maladie ou d'une épidémie. Dans l'industrie, il peut s'agir du bris d'une machine, ou en économie, du temps écoulé pour qu'une personne accepte un travail.

Dans plusieurs cas, l'événement est la transition d'un état à un autre. Par exemple, le décès est la transition de l'état « vivant » vers l'état « mort ». L'apparition d'une maladie est la transition de l'état « en santé » vers l'état « malade ». La figure (1.1) illustre ces deux exemples.

Selon le contexte, les termes décès, événement, échec ou transition peuvent être utilisés pour désigner l'événement constaté, et plus précisément ce qui se passe au temps de la réponse. Dans certain cas, l'aspect intéressant est la transition correspondant à l'inci-

dence d'une maladie, et dans d'autres cas, l'aspect intéressant est l'état de la disparition d'une maladie (Hougaard, 1999).

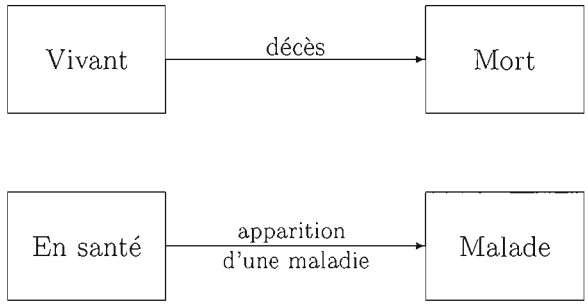

Figure 1.1 Modèles de temps de survie.

1.1.2 Données censurées

Une caractéristique importante de l'analyse de la survie est la présence des données censurées. Cette caractéristique, source de difficulté, a nécessité le développement de techniques alternatives à l'inférence usuelle.

Les données censurées sont des observations pour lesquelles la valeur exacte d'un événement n'est pas toujours connue. Cependant, nous disposons tout de même d'une information partielle permettant de fixer une borne inférieure (censure à droite) ou une borne supérieure (censure à gauche).

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

Il existe trois catégories de censure qu'on nomme censure à droite, censure à gauche et censure par intervalle (lorsqu'on connait la borne supérieure et la borne inférieure d'un événement).

À l'intérieur de ces trois catégories, il existe différents types de censure :

1. Censure de type I, si le temps de censure est fixé par le chercheur comme étant la fin de l'étude.
2. Censure de type II, se caractérise par le fait que l'étude cesse aussitôt qu'a eu lieu un nombre d'événements prédéterminé par l'expérimentateur.
3. Censure aléatoire, lorsque le moment de censure n'est plus sous le contrôle du chercheur et/ou que le temps d'entrée varie aléatoirement.

(Klein et Moeschberger, 2003).

La figure (1.2) illustre les situations de censure de type I et de censure aléatoire. Un rond vide indique une censure et une croix indique un événement. Dans le premier graphique (celui de gauche), les censures de types I sont déterminées par la fin de l'étude, tandis que les censures du deuxième graphique varient aléatoirement et peuvent surgir avant la fin de l'étude (les censures A et E dans le graphique de droite).

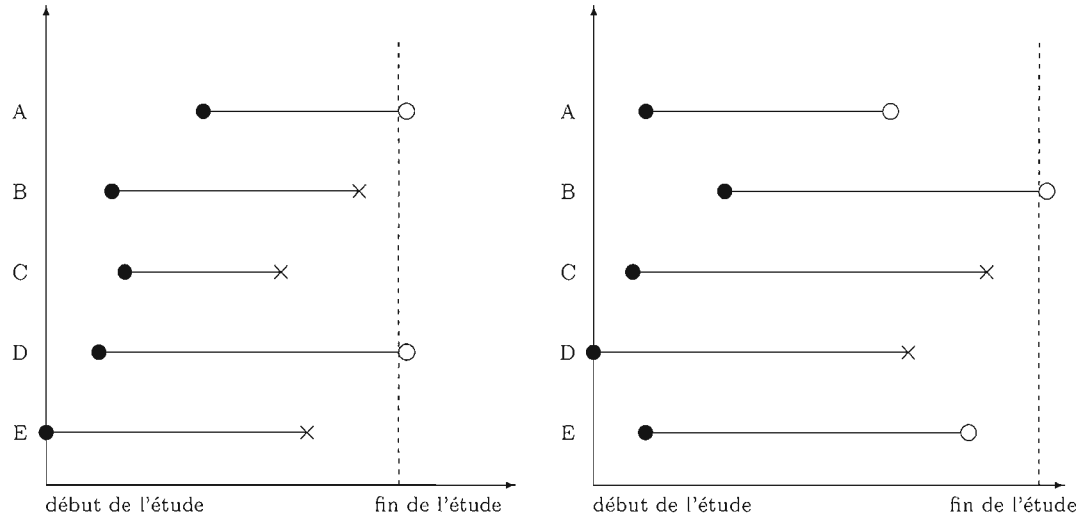

Figure 1.2 Illustration de censure de type I (à gauche) et de censure aléatoire (à droite).

1.2 La fonction de risque et la fonction de survie

Le développement de la méthodologie de l'analyse de la survie a connu un énorme progrès, et les premiers efforts ont été concentrés de façon prédominante sur l'estimation de la fonction de survie $S(t)$.

Dans l'analyse des données de survie (censurées) provenant d'études médicales, la fonction de risque est très utile. Elle contient de l'information sur le changement de risque comme fonction de temps. Cette quantité fondamentale de l'analyse de la survie est définie par :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(\, t \leq T < t + \Delta t \mid \, T \geq t)}{\Delta t} \tag{1.1}$$

où T est une variable aléatoire non négative représentant le temps de survie d'un individu. Ce condititionnement successif fait en sorte que la fonction de risque est le concept le plus pertinent, car il décrit la probabilité qu'un décès (événement) ait lieu dans un petit intervalle de temps, sachant que l'individu est vivant au temps t.

La fonction de survie représente la probabilité qu'un individu ait une durée de vie supérieure à t, et peut s'exprimer par :

$$S(t) = P(\, T > t \,) \tag{1.2}$$

Si T est une variable aléatoire continue, la relation entre $\lambda(t)$ et $S(t)$ peut être formulée par :

$$S(t) = e^{-\int_0^t \lambda(u)du} \tag{1.3}$$

La fonction $\Lambda(t) = \int_0^t \lambda(u)du$ est connue sous le nom de fonction de risque cumulé.

1.3 Estimateur non paramétrique de Kaplan-Meier

Les méthodes non paramétriques sont souvent préférables pour estimer la fonction de survie $S(t)$. Elles prennent mieux en compte la censure et la troncature, et elles donnent une meilleure adéquation. Kaplan et Meier (1958) ont proposé un estimateur très efficace de $S(t)$, nommé l'estimateur produit limite.

Si on considère $t_1 < t_2 < ... < t_D$, les temps de survie distincts de n individus, où au temps t_i, il y a d_i événements et plus que Y_i individus suceptibles de subir un événement, l'estimateur de la fonction de survie proposé par Kaplan-Meier est donné par :

$$\widehat{S}(t) = \begin{cases} 1 & \text{si } t < t_1 \\ \prod_{t_i<t}\left[1 - \frac{d_i}{Y_i}\right] & \text{si } t \geq t_1 \end{cases} \tag{1.4}$$

La quantité $\frac{d_i}{Y_i}$ estime la valeur de la fonction de risque $\lambda(t)$ pour $t = t_i$.

L'estimateur produit limite est une fonction en escaliers qui effectue des sauts à chaque événement au temps t_i. La grandeur du saut ne dépend pas uniquement du nombre d'événements au temps t_i mais aussi du nombre de censures à ce temps là.

La figure (1.3) représente un exemple de l'estimateur Kaplan-Meier de la fonction de survie du temps de la récurrence du cancer du poumon, à partir des données « Essais cliniques sur le cancer du poumon » décrites dans le chapitre 4.

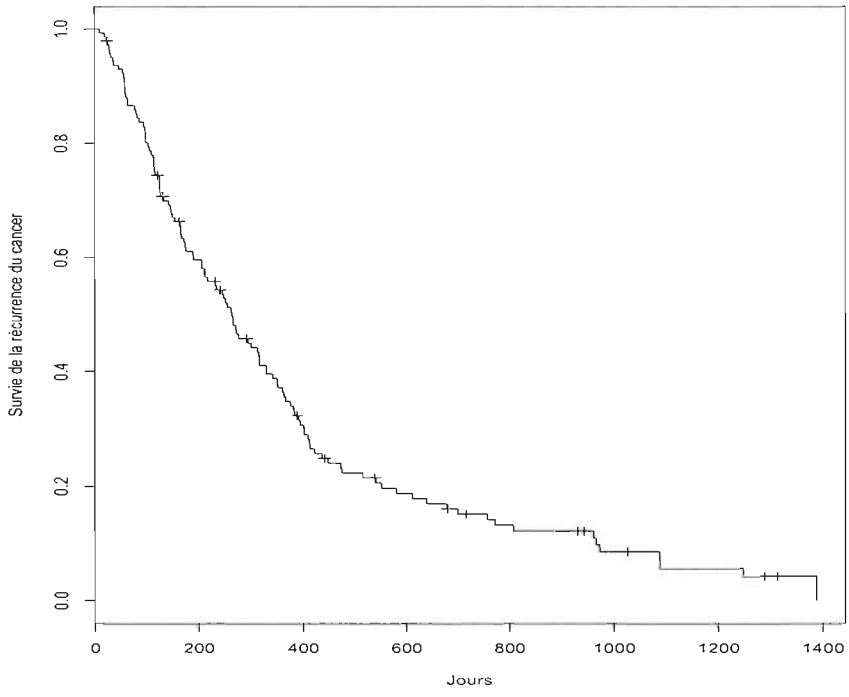

Figure 1.3 Courbe de Kaplan-Meier de la fonction de survie $S(t)$.

1.4 Modèles de régression

1.4.1 Modèle semi-paramétrique de Cox

Les facteurs explicatifs ou les covariables sont fréquemment disponibles dans une telle étude, et nous sommes intéressés de savoir les effets de ces facteurs sur le changement du taux de risque à travers le temps.
Soit $z = (z_1, ..., z_p)$ un vecteur de covariables d'un individu ; ce vecteur serait composé de variables identifiant le traitement suivi, et de facteurs tels que l'âge d'un individu, sa pression systolique, etc.. Le modèle de régression le plus connu est le modèle des risques proportionnels, qui spécifie le risque d'un individu avec covariable z par :

$$\lambda(t|z) = \lambda_0(t)e^{\beta' z} \tag{1.5}$$

où $\beta = (\beta_1, ..., \beta_p)^t$ est le vecteur des paramètres de régression.
Ainsi, la fonction de risque est le produit d'un terme dependant du temps, $\lambda_0(t)$, nommé risque instantané de base, et d'un terme qui ne dépend que des covariables z, $e^{\beta' z}$.

Cox (1972) a suggéré une procédure qui permet d'estimer les paramètres β sans faire de supposition sur la nature exacte de $\lambda_0(t)$. Ainsi, toute l'analyse se concentre sur les effets des covariables, d'où l'attribution du nom de modèle semi-paramétrique.

Dans son article, Cox (1975) introduit la vraisemblance partielle basée sur des données qui ne demandent pas d'information sur $\lambda_0(t)$. Il écarte spécifiquement les temps de survie (ou d'événement) observés et le nombre d'événements à ces temps là. En supposant que les censures sont indépendantes et non informatives, Cox écarte aussi les temps de censures et l'identité des individus associés à ces temps de censures.
La fonction de vraisemblance partielle est donc basée sur l'identité des individus susceptibles de subir un événement à chaque temps de survie (non censuré), le nombre des événements et l'identité des individus en risque à ce temps là étant connus.

cette fonction de vraisemblance prend la forme

$$L(\beta) = \prod_{i \in D} \frac{e^{\beta' Z_{(i)}}}{\sum_{j \in R_i} e^{\beta' Z_j}} \tag{1.6}$$

où D représente l'ensemble des indices des temps de survie (événements) observés, $Z_{(i)}$ le vecteur des covariables pour les individus qui ont échoué au $i^{ème}$ temps de survie $t_{(i)}$, et R_i l'ensemble des individus qui sont en risque d'échouer au temps de survie $t_{(i)}$ (Hougaard, 1999).

1.4.2 Modèles paramétriques

Si la forme de $\lambda_0(t)$ est précisée dans le modèle (1.5), on dira qu'il s'agit d'un modèle paramétrique. Parmi les modèles paramétriques, on mentionne le modèle dont le temps t suit une loi exponentielle caractérisée par un risque instantané constant (*i.e.* $\lambda_0(t) = \lambda$, avec λ une constante), et le modèle Weibull, caractérisé par $\lambda_0(t) = \alpha\,\lambda\,t^{\alpha-1}$, avec $\alpha > 0$.

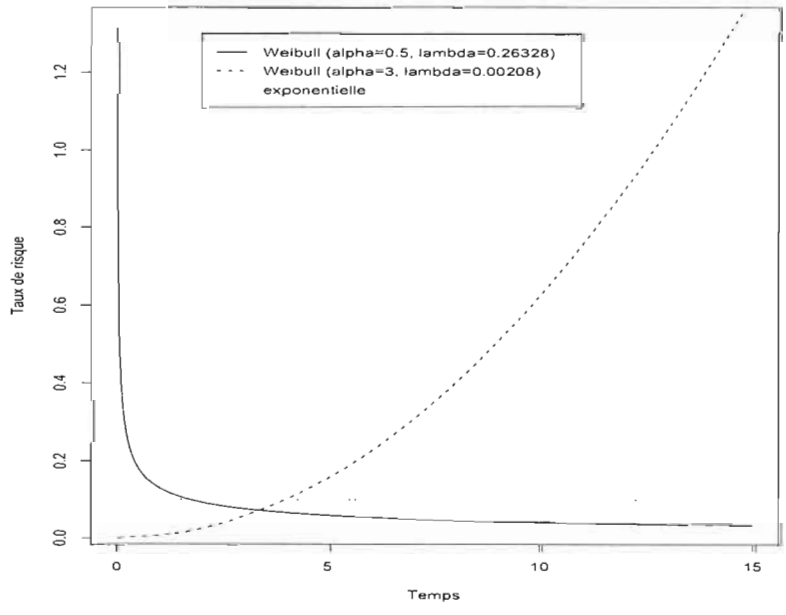

Figure 1.4 Fonctions de risque de la loi exponentielle, et de la loi Weibull pour différentes valeurs de α et λ.

1.4.3 Modèle de temps du survie accéléré

En dépit de la grande popularité et de la polyvalence du modèle de Cox, il y a des bonnes raisons d'explorer des modèles alternatifs. Premièrement, l'hypothèse de la proportionnalité du modèle peut être non satisfaite pour certaines applications. Deuxièmement, il serait intéressant de trouver des modèles alternatifs qui caractérisent d'une façon différente l'association entre les covariables et le temps de survie.

Une approche pour modéliser l'effet des covariables est analogue à la régression linéaire classique. Dans cette approche, le logarithme du temps de survie $log(T)$ est modélisé par l'expression :

$$log(T) = \beta' z + \epsilon \tag{1.7}$$

où β est le vecteur de paramètres de régression inconnus, et ϵ la variable d'erreur qui est indépendante du vecteur des covariables z.
Une transformation exponentielle de (1.7) mène à :

$$T = e^{\beta' z} T_0 \tag{1.8}$$

avec $T_0 = e^{\epsilon}$. Cette expression montre que le rôle de z est d'accélérer ou de ralentir le temps de survie. Ainsi, on dit qu'il s'agit d'un modèle de temps de survie accéléré.
Le modèle (1.7) peut être spécifié par la fonction de risque :

$$\lambda(t|z) = \lambda_0(t\, e^{-\beta' z})\, e^{-\beta' z} \tag{1.9}$$

où $\lambda_0(t)$ est la fonction de risque de T_0.
Un exemple de $\lambda_0(t)$ est le risque instantané de base correspondant à la distribution de Weibull, *i.e.* $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$. (Klein et Moeschberger, 2003).

1.5 Processus de comptage et analyse de la survie

Au milieu des années 1970, Aalen présente sa théorie des martingales pour processus de comptage qui offre un cadre unifié pour les méthodes statistiques de l'analyse de la

survie. Dans son travail, l'approche du processus de comptage utilise la représentation intégrale pour les statistiques des données censurées qui fournit une forme simple et unifiée des estimateurs, des statistiques de test, et des méthodes de régression. Ces méthodes utilisant les martingales permettent d'obtenir de simples expressions pour des statistiques compliquées, pour des distributions asymptotiques dc statistiques de test, et pour des estimateurs.

Dans l'approche du processus de comptage, la $i^{ème}$ observation est représentée par le couple $\{N_i(t), Y_i(t)\}$ $(t > 0)$, avec :

$$N_i(t) = I(X_i \leq t, \delta_i = 1) \quad et \quad Y_i(t) = I(X_i \geq t) \tag{1.10}$$

où X_i est le minimum entre le temps de survie et le temps de censure, et δ_i l'indicateur de l'observation ($\delta_i = 1$ si la donnée est non censurée et $\delta_i = 0$ sinon).

Le processus continu à droite $N(t)$ est connu simplement sous le nom de processus de comptage, du fait qu'il compte le nombre des événements observés jusqu'au temps t ; et le processus continu à gauche $Y(t)$ est connu sous le nom de processus de risque, indiquant si un individu est susceptible de subir un événement au temps t.

Une illustration importante de l'approche du processus de comptage se résume dans l'étude des propriétés de l'estimateur de Nelson-Aalen $\hat{\Lambda}(t)$ de $\Lambda(t)$ (le risque cumulé). Le risque cumulé dans une région où il existe au moins une observation est :

$$\Lambda^*(t) = \int_0^t J(s)\lambda(s)ds \tag{1.11}$$

où $J(t) = I[\bar{Y}(t) > 0]$ est l'indicateur pour lequel au moins une observation est encore à risque. Avec $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$ et n la taille de l'échantillon.

Avec cette notation, $\hat{\Lambda}(t) = \int_0^t \frac{J(s)}{\bar{Y}(s)} d\bar{N}(s)$ estime $\Lambda^*(t)$, où $\bar{N}(t) = \sum_{i=1}^n N_i(t)$.
En effet,

$$\begin{aligned} \hat{\Lambda}(t) - \Lambda^*(t) &= \int_0^t \frac{J(s)}{\bar{Y}(s)}(d\bar{N}(s) - \bar{Y}(s)\lambda(s)ds) \\ &= \sum_{i=1}^n \int_0^t \frac{J(s)}{\bar{Y}(s)} dM_i(s) \end{aligned} \tag{1.12}$$

où la martingale $M_i(s) = N_i(s) - \int_0^s Y_i(u)\lambda(u)du$ (spécifique du $i^{ème}$ individu) a une espérance égale à zéro.

La martingale $M_i(s)$ représente la différence entre le nombre d'événements observé, *i.e.* $N_i(s)$, et le nombre d'événements prédit par le modèle pour le $i^{ème}$ individu.

En remarquant que $\hat{\Lambda}(t) - \Lambda(t) = [\hat{\Lambda}(t) - \Lambda^*(t)] + [\Lambda^*(t) - \Lambda(t)]$, on a :

$$\begin{aligned} E[\hat{\Lambda}(t) - \Lambda(t)] &= E[\Lambda^*(t) - \Lambda(t)] \\ &= -\int_0^t Pr(\bar{Y}(s) = 0)\lambda(s)ds \\ &\longrightarrow 0 \ (lorsque \ n \longrightarrow \infty) \end{aligned} \tag{1.13}$$

$\hat{\Lambda}(t)$ est un estimateur biaisé de $\Lambda(t)$.

(Pour plus de détails, voir Therneau et Grambsch (2000), page 27).

Cette approche basée sur les martingales permet de faire des développements élégants des propriétés des estimateurs Nelson-Aalen et Kaplan-Meier.

(Fleming et Harrington (1991), et Therneau et Grambsch (2000)).

CHAPITRE II

MODÈLE DE RISQUE INSTANTANÉ AVEC UN POINT DE RUPTURE

Une question que se posent fréquemment les chercheurs en cancerologie est de savoir à partir de quand on peut juger qu'une nouvelle thérapie produit un écart significatif dans le taux de risque après l'essai d'un traitement avec risque constant. En fait, cet écart est significatif s'il est accompagné par une réduction substantielle dans le taux de risque.

Selon Matthews et Farewell (1982), la réponse repose essentiellement sur le test de rapport de vraisemblance (T.R.V) de l'hypothèse d'un risque instantané constant contre l'alternative d'un risque constant avec un simple point de rupture. L'approche de Matthews et Farewell (1982, 1985) à ce problème est présentée dans ce chapitre.

2.1 Modèle avec un point de rupture

Le modèle de survie le plus simple est le modèle de risque instantané constant, qui correspond à une fonction de survie de loi exponentielle. Ce modèle suppose que le risque d'un événement ne change pas en fonction du temps. Cependant, il est important d'identifier les changements dans le risque instantané. De simples tendances de changement ont l'avantage d'être simples à décrire et à comprendre. Un des modèles les plus simples du changement de risque instantané est le modèle de risque constant avec un point de rupture.

Soit T le temps à un événement particulier, par exemple le temps écoulé jusqu'à la rechute après le traitement par « remission induction » pour les patients de la leucémie. Matthews et Farewell (1982) ont considéré un modèle pour la distribution de T spécifié par la fonction de risque :

$$\lambda(t) = \begin{cases} \lambda & \text{si } t \leq \tau \\ (1-\epsilon)\lambda & \text{si } t > \tau \end{cases} \tag{2.1}$$

Ce modèle a trois paramètres, λ, ϵ $(0 \leq \epsilon < 1)$, et le point de rupture τ, qui sont tous inconnus. En d'autres termes, le taux de risque est constant jusqu'à un point inconnu τ, après quoi il prend une nouvelle valeur constante.

Nous sommes intéressés essentiellement par le test de l'hypothèse nulle sans point de rupture, $H_0 : \epsilon = 0$, contre l'hypothèse alternative de l'existence d'une rupture, $H_a : \epsilon \neq 0$.

Remarque 2.1

Matthews et Farewell ont noté que l'hypothèse $H_0 : \epsilon = 0$ *versus* $H_a : \epsilon \neq 0$ *est équivalente à l'hypothèse* $H_0 : \tau = 0$ *contre l'alternative* $H_a : \tau \neq 0$.
On a utilisé la notation de Matthews et al.(1985) pour le modèle (2.1).

2.2 Inférence sur le modèle

Soient $T_1, T_2, ..., T_n$ les temps de survie pour n individus. Ces variables aléatoires positives sont supposées être indépendantes et identiquement distribuées (i.i.d), et leur fonction de densité existe et est définie selon le modèle avec point de rupture par :

$$f(t) = \begin{cases} \lambda exp(-\lambda t) & \text{si } t \leq \tau \\ \rho\lambda exp(-\lambda\tau - \rho\lambda(t-\tau)) & \text{si } t > \tau \end{cases} \tag{2.2}$$

où $\rho = 1 - \epsilon$, et où λ, ϵ et τ sont les paramètres du modèle (2.1).
Soient λ_1 et λ_2 les paramètres définis par $\lambda_1 = \lambda$ et $\lambda_2 = (1-\epsilon)\lambda$, la fonction de densité

(2.2) peut donc s'écrire sous la forme :

$$f(\ t\ |\lambda_1\ \lambda_2, \tau) = \begin{cases} \lambda_1\, exp(-\lambda_1 t) & \text{si } t \leq \tau \\ \lambda_2\, exp(-\lambda_2(t-\tau) - \lambda_1 \tau) & \text{si } t > \tau \end{cases} \tag{2.3}$$

$F(\ t\ |\lambda_1\ \lambda_2, \tau)$ va représenter ici la fonction de répartition correspondante.

Ces n individus vont être l'objet de censures randomisées à droite, représentées par $Y_1, Y_2, ..., Y_n$, variables aléatoires i.i.d, dont la fonction de répartition et la densité sont respectivement $G(y) = P(Y_i \leq y)$ et $g(y)$.
Notons que les Y_i et les paramètres associés à leur distribution sont supposés indépendants des temps de survie T_i et de leurs paramètres associés.

Pour chaque sujet, le minimum entre le temps de survie et le temps censuré va être observé, ainsi que l'indicateur de l'événement.
Soit $X_i = min(T_i, Y_i)$

et $\quad \delta_i = \begin{cases} 1 & \text{si } T_i \leq Y_i \\ 0 & \text{si } T_i > Y_i \end{cases}$

Remarque 2.2

L'hypothèse $H_0 : \epsilon = 0$ *versus* $H_a : \epsilon \neq 0$ *est maintenant équivalente à l'hypothèse* $H_0 : \lambda_1 = \lambda_2$ *contre l'alternative* $H_a : \lambda_1 \neq \lambda_2$.

2.2.1 Fonction de vraisemblance

Notons par $(x_1, \delta_1), ..., (x_n, \delta_n)$ les observations obtenues. Selon MacGibbon et Groshen (2003), pour des données indépendantes censurées à droite, la fonction de vraisemblance devient :

$$\begin{aligned} L(\lambda_1\ \lambda_2, \tau) &= \prod_{i=1}^{n} [(1 - F(x_i|\lambda_1\ \lambda_2, \tau))\, g(x_i)]^{1-\delta_i} [(1 - G(x_i))\, f(x_i|\lambda_1\ \lambda_2, \tau)]^{\delta_i} \\ &= \prod_{i=1}^{n} [1 - F(x_i|\lambda_1\ \lambda_2, \tau)] \left[\frac{f(x_i|\lambda_1\ \lambda_2, \tau)}{1 - F(x_i|\lambda_1\ \lambda_2, \tau)}\right]^{\delta_i} [(g(x_i))^{1-\delta_i} (1 - G(x_i))^{\delta_i}]. \end{aligned}$$

Le logarithme de cette fonction de vraisemblance est donné par,

$$l(\lambda_1\,\lambda_2,\tau) = \sum_{i=1}^{n} log[1 - F(x_i|\lambda_1\,\lambda_2,\tau)] + \sum_{i=1}^{n} \delta_i\, log\left[\frac{f(x_i|\lambda_1\,\lambda_2,\tau)}{1 - F(x_i|\lambda_1\,\lambda_2,\tau)}\right] + C(\underline{x},\underline{\delta}) \quad (2.4)$$

où $C(\underline{x},\underline{\delta})$ représente la sommation de tous les termes censurés mais pas des paramètres des temps de survie T_i, *i.e.* λ_1, λ_2 et τ.

$$C(\underline{x},\underline{\delta}) = \sum_{i=1}^{n}[(1-\delta_i)\,log(g(x_i)) + \delta_i\, log(1 - G(x_i))]$$

Pour le modèle avec un point de rupture (2.1) dont la fonction de densité est définie par (2.3), l'équation (2.4) deviendra :

$$l(\lambda_1\,\lambda_2,\tau) = \sum_{i:x_i\le\tau} \lambda_1\, x_i - \sum_{i:x_i>\tau} [\lambda_2\,(x_i - \tau) + \lambda_1\tau] + \sum_{i:x_i\le\tau} \delta_i\, log(\lambda_1) + \sum_{i:x_i>\tau} \delta_i\, log(\lambda_2) + C(\underline{x},\underline{\delta}) \quad (2.5)$$

2.2.2 Estimateurs du maximum de vraisemblance sous les hypothèses H_0 et H_a

Sous l'hypothèse nulle H_0 : $\lambda_1 = \lambda_2$ (pas de point rupture), il n'y aura qu'un paramètre $\lambda = \lambda_1 = \lambda_2$. D'après l'équation (2.5), l'estimateur du maximum de vraisemblance de λ est :

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i} \quad (2.6)$$

Le maximum du logarithme de la fonction de vraisemblance sous H_0 sera donc :

$$\begin{aligned} l(\widehat{\lambda}) &= -\sum_{i=1}^{n} \widehat{\lambda}\, x_i + \sum_{i=1}^{n} \delta_i\, log(\widehat{\lambda}) + C(\underline{x},\underline{\delta}) \\ &= -\sum_{i=1}^{n} \delta_i + log\left[\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}\right] + C(\underline{x},\underline{\delta}) \\ &= -d + d\,log\left[\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}\right] + C(\underline{x},\underline{\delta}) \end{aligned}$$

où d est le nombre total des événements non censurés, *i.e.* $d = \sum_{i=1}^{n} \delta_i$.

Si le point de rupture est connu, *i.e.* $\tau \neq 0$, alors les estimateurs du maximum de vraisemblance de λ_1 et λ_2 sous l'alternative $H_a : \lambda_1 \neq \lambda_2$, sont :

$$\begin{aligned} \widehat{\lambda_1} &= \frac{\sum_{i:x_i \leq \tau} \delta_i}{\sum_{i:x_i \leq \tau} x_i + \sum_{i:x_i > \tau} \tau} \\ \widehat{\lambda_2} &= \frac{\sum_{i:x_i > \tau} \delta_i}{\sum_{i:x_i > \tau} (x_i - \tau)} \end{aligned} \tag{2.7}$$

En substituant $\widehat{\lambda_1}$ et $\widehat{\lambda_1}$ à λ_1 et λ_2 (resp) dans l'équation (2.5), MacGibbon et Groshen (2003) ont obtenu le logarithme de la fonction « profile » de vraisemblance :

$$\begin{aligned} l(\tau) &= -\widehat{\lambda_1}(\tau)\left(\sum_{i:x_i \leq \tau} x_i + \sum_{i:x_i > \tau} \tau\right) - \widehat{\lambda_2}(\tau)\left(\sum_{i:x_i > \tau} (x_i - \tau)\right) + \sum_{i:x_i \leq \tau} log(\widehat{\lambda_1}(\tau)) \\ &\quad + \left(\sum_{i=1}^{n} \delta_i - \sum_{i:x_i \leq \tau} \delta_i\right) log(\widehat{\lambda_2}(\tau)) + C(\underline{x}, \underline{\delta}) \\ &= -d + d(\tau) log(\widehat{\lambda_1}(\tau)) + (d - d(\tau)) log(\widehat{\lambda_2}(\tau)) + C(\underline{x}, \underline{\delta}) \end{aligned}$$

C'est-à-dire :

$$\begin{aligned} l(\tau) &= -d + d(\tau)\left[log(d(\tau)) - log(\sum_{i:x_i \leq \tau} x_i + \sum_{i:x_i > \tau} \tau)\right] \\ &\quad + (d - d(\tau))\left[log(d - d(\tau)) - log(\sum_{i:x_i > \tau} (x_i - \tau))\right] + C(\underline{x}, \underline{\delta}) \end{aligned} \tag{2.8}$$

avec $d(\tau) = \sum_{i:x_i \leq \tau} \delta_i$ étant le nombre des événements non censurés avant ou au temps τ.

Si maintenant, $z_1 \leq z_2 ... \leq z_k$ représentent les temps de survie ordonnés, l'estimateur du maximum de vraisemblance du triplet (λ_1,λ_2,τ) est la valeur de τ et les estimateurs $\widehat{\lambda_1}(t)$ et $\widehat{\lambda_2}(t)$ associés qui maximise $l(\tau)$.

La fonction $l(\tau)$ est continue sauf aux points z_i, et elle est convexe dans l'intervalle ouvert $]z_i, z_{i+1}[$ (Voir appendices de MacGibbon et Groshen (2003)), et il est donc suffisant d'évaluer $l(z_i^-)$ et $l(z_i^+)$, pour $i = 1, 2, .., k$.

Il est nécessaire d'ajouter la restriction $\tau \leq z_{k-1}$ pour éviter que $l(\tau)$ soit non bornée, comme cela a été noté par Nguyen *et al.*(1984) et Yao (1986) ; plus d'explications seront

données dans la section 3 de ce chapitre.

En l'absence de censure, Yao (1986) et Pham et Nguyen (1990) ont montré que l'estimateur de vraisemblance maximale de τ est convergent.

MacGibbon et Groshen (2003) ont trouvé que l'estimateur du maximum de vraisemblance de (λ_1,λ_2,τ) sera le triplet ($\widehat{\lambda_1}(z_j^+)$,$\widehat{\lambda_1}(z_j^+)$,$z_j$) ou ($\widehat{\lambda_1}(z_j^-)$,$\widehat{\lambda_1}(z_j^-)$,$z_j$) , qui maximise $l(\tau)$, où :

$$\widehat{\lambda_1}(z_j^-) = \frac{\sum_{i:x_i<z_j} \delta_i}{\sum_{i:x_i<z_j} x_i + \sum_{i:x_i\geq z_j} z_j}$$

$$\widehat{\lambda_1}(z_j^+) = \frac{\sum_{i:x_i\leq z_j} \delta_i}{\sum_{i:x_i\leq z_j} x_i + \sum_{i:x_i>z_j} z_j}$$

$$\widehat{\lambda_2}(z_j^-) = \frac{\sum_{i:x_i\geq z_j} \delta_i}{\sum_{i:x_i\geq z_j} (x_i - z_j)}$$

$$\widehat{\lambda_2}(z_j^+) = \frac{\sum_{i:x_i>z_j} \delta_i}{\sum_{i:x_i>z_j} (x_i - z_j)}$$

2.2.3 La statistique du test des rapports de vraisemblances

Selon MacGibbon et Groshen (2003), la statistique du test des rapports de vraisemblances pour le test de l'hypothèse nulle H_0 : $\lambda_1 = \lambda_2$ (sans point de rupture τ) est donnée par sa transformation $-2\,log$, dont l'expression est :

$$\Delta = -2\,[\,l(\widehat{\lambda}) - l(\widehat{\lambda}_1(\tau), \widehat{\lambda}_2(\tau), \tau)] \tag{2.9}$$

Pour chaque z_j, on peut évaluer les limites à droite et à gauche du logarithme de la fonction « profile » de vraisemblance $l(\tau)$ donnée par (2.8).

Du fait que $\sum_{i:x_i\leq z_j} \delta_i = j$, pour $\tau = z_j^+$ on a :

$$\begin{aligned} l(z_j^+) = & -d + j\left[log(j) - log(\sum_{i:x_i\leq z_j} x_i + \sum_{i:x_i>z_j} z_j)\right] \\ & +(d-j)\left[log(d-j) - log(\sum_{i:x_i>z_j} (x_i - z_j))\right] + C(\underline{x}, \underline{\delta}) \end{aligned}$$

Et pour $\tau = z_j^-$:

$$\begin{aligned} l(z_j^-) &= -d + (j-1)\Big[log(j-1) - log(\sum_{i:x_i<z_j} x_i + \sum_{i:x_i\geq z_j} z_j)\Big] \\ &\quad +(d-j+1)\Big[log(d-j+1) - log(\sum_{i:x_i\geq z_j}(x_i - z_j))\Big] + C(\underline{x},\underline{\delta}) \end{aligned}$$

dans ce cas, $\sum_{i:x_i<z_j} \delta_i = j-1$.

Remarque 2.3

MacGibbon et Groshen (2003) ont remarqué que :

$$\begin{aligned} \sum_{i:x_i\leq z_j} x_i + \sum_{i:x_i>z_j} z_j &= \sum_{i:x_i<z_j} x_i + \sum_{i:x_i\geq z_j} z_j \\ &= \sum_{i=1}^{n} min(x_i, z_j) \end{aligned}$$

La statistique Δ (2.9) évaluée à $\tau = z_j^+$ sera exprimée par :

$$\begin{aligned} \Delta_j^+ &= -2[l(\widehat{\lambda}) - l(\widehat{\lambda}_1(z_j^+), \widehat{\lambda}_2(z_j^+), z_j)] \\ &= -2\Big[j[log(j) - log(\sum_{i=1}^{n} min(x_i, z_j))] + (d-j)[log(d-j) - log(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} min(x_i, z_j))] \\ &\quad +d\,log(\sum_{i=1}^{n} x_i) - d\,log(d)\Big] \\ &= -2\Big[j[log(j) - log(\sum_{i=1}^{n} min(x_i, z_j)) + log(\sum_{i=1}^{n} x_i)] \\ &\quad +(d-j)[log(d-j) - log(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} min(x_i, z_j)) + log(\sum_{i=1}^{n} x_i)] - d\,log(d)\Big] \\ &= -2\Big[j[log(j) - log(U_j)] + (d-j)[log(d-j) - log(1-U_j)] - d\,log(d)\Big] \end{aligned}$$

donc,

$$\Delta_j^+ = -2\Big[j\,log\Big(\frac{j}{U_j}\Big) + (d-j)log\Big(\frac{d-j}{1-U_j}\Big) - d\,log(d)\Big] \tag{2.10}$$

avec $U_j = \frac{\sum_{i=1}^{n} min(x_i, z_j)}{\sum_{i=1}^{n} x_i}$ *(comme il a été noté par Worsley(1988)).*

En suivant le même procédé, l'expression de la statistique Δ *évaluée à* $\tau = z_j^-$ *va être :*

$$\Delta_j^- = -2\left[(j-1)log\left(\frac{j-1}{U_j}\right) + (d-j+1)log\left(\frac{d-j+1}{1-U_j}\right) - dlog(d)\right] \qquad (2.11)$$

Et avec la convention $0\,log(0) = 0$, la statistique du test des rapports de vraisemblances Δ est donnée par :

$$\Delta = max_{j=1,..,k}[max(\Delta_j^+, \Delta_j^-)] \qquad (2.12)$$

2.3 Propriétés de la statistique du test des rapports de vraisemblances

2.3.1 Distribution de la statistique Δ pour des données non censurées

Nguyen *et al.*(1984) et Yao (1986) ont noté que la fonction « profile » de vraisemblance (2.7) est non bornée si la dernière observation $x_{(n)}$ est non censurée, *i.e.* $x_{(n)} = z_k$, et quand $\widehat{\lambda}_1$ est fixé et $\tau \nearrow x_{(n)}$, $\widehat{\lambda}_2(\tau)$ tend vers l'infini. Pour le cas où il n'y a que des événements, le point de rupture peut difficilement être estimé s'il se trouve à droite de toutes les valeurs de l'échantillon. Donc, on doit supposer que le point de rupture τ se trouve quelque part à gauche de la plus grande observation ($\tau \leq z_{k-1}$).

Pour des échantillons de temps de survie (non censurés) de taille $10, 20, .., 100, 200$, on a généré 10000 réplications à partir de la distribution exponentielle de paramètre 1, tout en calculant Δ.

Les temps de survie sont donc générés à partir de l'expression suivante :

$$x = \frac{-1}{\lambda}\, log(1-u)$$

où u est une suite de nombres aléatoires générés par une uniforme(0,1), et $\lambda = 1$.

Les quantiles 90,95,99 de la distribution de la statistique Δ (où le point de rupture $\tau \leq x_{(n-1)}$) trouvés par simulation sont données dans la table(2.1).

Tableau 2.1 Quantiles de la statistique du test des rapports de vraisemblances Δ, avec $\tau \leq x_{(n-1)}$.

Taille de	Quantiles		
l'échantillon	90	95	99
10	10.286	12.692	17.942
20	10.447	12.720	18.055
30	10.610	13.069	18.380
40	10.681	12.923	18.670
50	10.792	13.005	18.362
60	10.683	12.980	18.386
70	10.841	13.090	18.672
80	10.806	12.800	18.686
90	10.945	13.280	18.981
100	11.038	13.136	18.761
200	11.349	13.397	18.472

Les quantiles de la table (2.1) ne semblent pas converger vers une limite finie avec la taille croissante de l'échantillon. Des comportements similaires ont été notés pour d'autres statistiques entrainant le point de rupture, voir Worsley (1986).

En pratique, on pense qu'il est commode que le point de rupture τ appartienne à un ensemble d'observations plus restreint que l'ensemble $x_{(1)}, ..., x_{(n-1)}$, comme cela a été noté par Matthews *et al.*(1985) et Worsley (1988).

Par exemple, on peut contraindre le point de rupture τ à être entre le $p^{ième}$ quantile de l'échantillon et le $(1-p)^{ième}$ quantile de celui-ci, pour des valeurs $p = 0.1$, $p = 0.2$.

Plus précisement, si n est la taille de l'échantillon et $m = [pn]$, $\Delta(\tau)$ est maximisée pour $\tau \in \{x^-_{m+1}, x^+_{m+1}, ..., x^-_{n-m}, x^+_{n+m}\}$.

Quelques quantiles de la statistique du test des rapports de vraisemblances Δ pour les deux cas $p = 0.1$ et $p = 0.2$ sont reportés dans les deux tables (2.2) et (2.3).

Tableau 2.2 Quantiles de la statistique Δ, avec $p^{ième}quantile < \tau < (1-p)^{ième}quantile$ et p = 0.1 .

Taille de	Quantiles		
l'échantillon	90	95	99
10	9.942	12.352	18.061
20	8.562	10.550	14.727
30	8.272	10.162	14.344
40	8.020	9.721	13.777
50	7.963	9.582	13.569
60	7.917	9.657	13.178
70	7.809	9.450	13.040
80	7.662	9.168	12.794
90	7.799	9.466	13.517
100	7.733	9.274	12.687
200	7.773	9.365	12.770

Tableau 2.3 Quantiles de la statistique Δ, avec $p^{ième}quantile < \tau < (1-p)^{ième}quantile$ et p = 0.2 .

Taille de	Quantiles		
l'échantillon	90	95	99
10	7.687	9.710	14.380
20	7.260	8.900	12.930
30	6.892	8.488	12.334
40	6.850	8.497	12.312
50	6.798	8.390	11.908
60	6.964	8.603	12.190
70	6.805	8.442	12.074
80	6.958	8.463	11.853
90	6.826	8.509	12.117
100	6.707	8.200	11.451
200	6.795	8.361	12.040

On voit bien que les quantiles des tables (2.2) et (2.3) sont plus petits que ceux de la table (2.1), et de plus, ils semblent décroître vers une limite finie avec l'accroissement de la taille de l'échantillon. Ceci montre que la distribution nulle de la statistique du test des rapports de vraisemblances Δ dépend fortement de l'intervalle où se trouve le point de rupture τ.

2.3.2 Distribution de Δ pour des données censurées

Matthews et Farewell (1982) ont suggéré qu'un nombre modéré de censures dans l'échantillon des observations, a un faible impact sur la distribution nulle de la statistique de test de rapport de vraisemblance Δ pour des données non censurées. Cette conclusion est basée sur des résultats de simulation pour des censures de type I ; les observations qui

dépassent une valeur fixe, sont censurées à cette valeur.

Worsley (1988) a noté que cet impact reste faible malgré la présence des censures de type II. Avec ce type de censure, pour une valeur fixe de $k < n$, toutes les observations qui sont supérieures à la $k^{ième}$ plus grande observation sont censurées à cette valeur.

Pour d'autre formes de censures non informatives, Barndorff-Nielsen et Cox (1984) ont montré que la distribution asymptotique de la statistique du test des rapports de vraisemblances (pour des données non censurées) reste inchangée en général, et l'utilisation de la distribution nulle reste valide pour un nombre modéré de censures non informatives.

2.4 Application

2.4.1 Description des données d'étude

Matthews et Farewell (1982) ont considéré l'exemple des temps de rechute de 84 patients atteints de la leucémie non-lymphoblastique aiguë, qui ont été traités par un protocole commun dans des hôpitaux universitaires et des institutions privées dans le nord-ouest du pacifique. Les patients ont été randomisés selon le protocole expérimental et parmi 33 observations censurées, 24 ont été censurées à 182 jours.
Les données ordonnées de cette étude sont présentées dans la table (2.4).

2.4.2 Résultats

Il y a 51 temps de survie qui sont inférieurs à la plus grande observation censurée, par conséquent la fonction « profile » de vraisemblance (2.8) est bornée. Ainsi, on n'a pas besoin de censurer artificiellement la dernière observation.

La valeur de la statistique du test des rapports de vraisemblances Δ, calculée à partir des équations (2.10), (2.11) et (2.12), est 14.504. Elle est observée juste après la $49^{ème}$ observation non censurée, *i.e.* $\tau = 697$.

Tableau 2.4 Les durées du traitement (remission induction) pour les 84 patients atteints par la leucémie non-lymphoblastique.

Observations non censurés (51)							
24	46	57	57	64	65	82	89
90	90	111	117	128	143	148	152
166	171	186	191	197	209	223	230
247	249	254	258	264	269	270	273
284	294	304	304	332	341	393	395
487	510	516	518	518	534	608	642
697	955	1160					
Observations censurés (33)							
68	119	182	182	182	182	182	182
182	182	182	182	182	182	182	182
182	182	182	182	182	182	182	182
182	182	583	1310	1538	1634	1908	1996
2057							

Les estimateurs de maximum de vraisemblance de λ_1 et λ_2 sont respectivement $\widehat{\lambda}_1 = \widehat{\lambda}_1(z_{49}^+) = 0.00208$ et $\widehat{\lambda}_2 = \widehat{\lambda}_2(z_{49}^+) = 0.00028$.

À partir de la table (2.1) des simulations (de la statistique du test des rapports de vraisemblances Δ), on peut voir que la p-valeur du test de l'hypothèse $H_0 : \tau = 0$ contre l'alternative $H_0 : \tau \neq 0$ est significative au niveau 0.05, *i.e.* la valeur $\tau = 697$ est significative à un niveau de 0.05.

En résumé, pour les données de survie censurées indiquées, le modèle avec un point de rupture spécifié par la fonction de risque, est :

$$\lambda(t) = \begin{cases} 0.00208 & \text{si } t \leq 697 \\ 0.00028 & \text{si } t > 697 \end{cases} \tag{2.13}$$

La figure (2.1) montre les deux courbes des estimations Kaplan-Meier et du modèle avec un point de rupture (2.12) du logarithme de la fonction de survie ($log(S(t))$) correspondant au traitement « remission induction ».

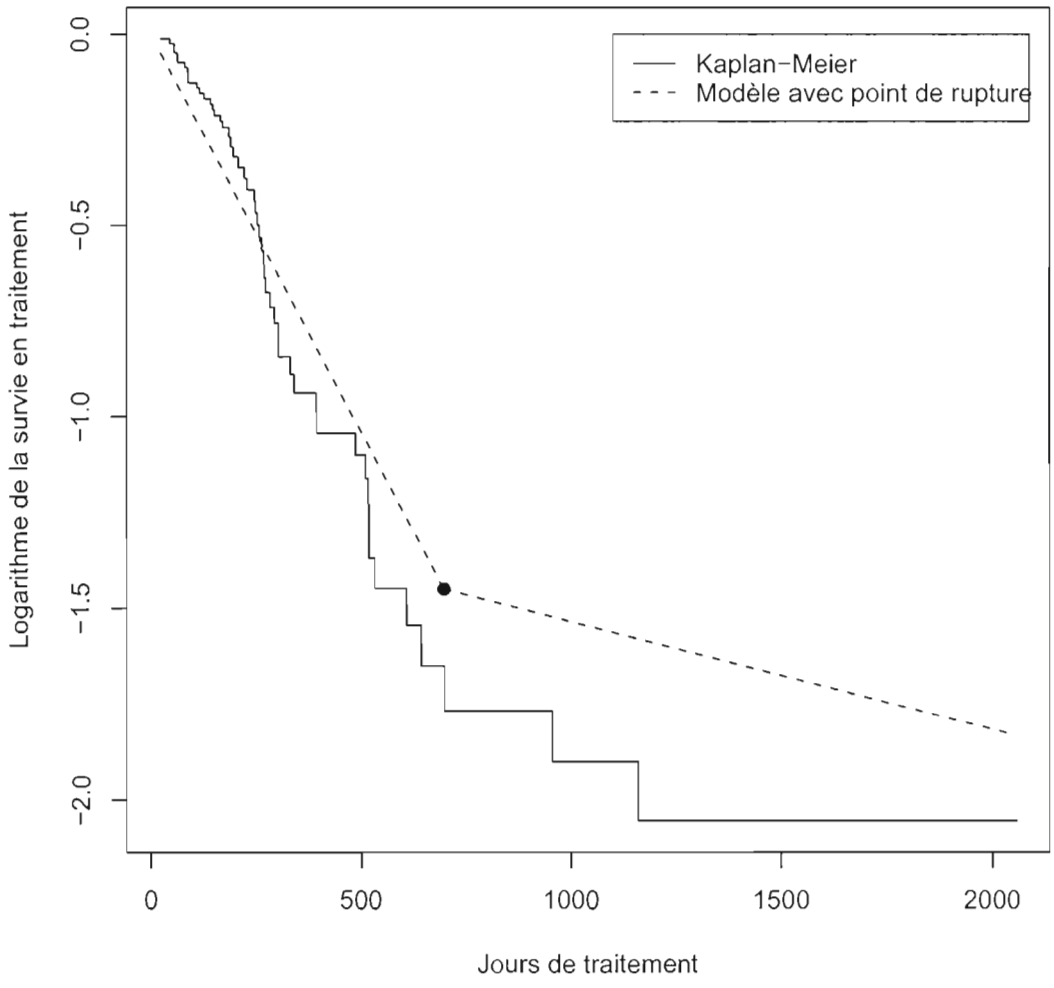

Figure 2.1 Logarithme de l'estimateur de la survie en traitement, basé sur l'estimateur Kaplan-Meier et sur le modèle avec un point de rupture. Le point noir correspond au point de rupture $\tau = 697$.

CHAPITRE III

PROCESSUS DE LA STATISTIQUE DE SCORE POUR TESTER LE MODÈLE AVEC UN POINT DE RUPTURE

3.1 Introduction

Matthews et Farewell (1982) ont considéré le problème de tester l'hypothèse « sans point de rupture » τ, en examinant par simulation la distribution de la statistique du test des rapports de vraisemblances $\Delta(\tau)$ donnée par (2.8). Cependant, pour des observations non censurées, cette statistique est non bornée quand τ se rapproche de la dernière observation. Matthews et Farewell (1985) résolvent le problème de la singularité de $\Delta(\tau)$ en remarquant que toutes les données sont effectivement discrètes, et que donc la fonction de vraisemblance doit être calculée à partir des fonctions de probabilité plutôt qu'à partir des densités. Par conséquent, la fonction de vraisemblance dans (2.8) sera toujours bornée par 1, et le problème de la singularité de $\Delta(\tau)$ ne survient pas.

Dans ce chapitre 3, seront développés certains résultats de Matthews, Farewell et Pyke (1985) et Davies (1977), qui portent sur les statistiques du score normalisée et du score partielle, dérivés du modèle avec un point de rupture (2.1) :

1. On donnera la démonstration de la convergence du processus de score en fonction de τ (point de rupture) dérivé de la statistique du score normalisée sous l'hypothèse que λ est connu, et on identifiera le processus Ornstein-Uhlenbeck comme étant la limite de cette convergence, pour λ_1 connu, paramètre du modèle (2.1). Les grandes lignes de cette preuve suivent l'approche de Matthews *et al.*(1985), mais

plus de détails seront ajoutés dans la démonstration.

2. On présentera le niveau de signification asymptotique (à partir du processus Ornstein-Uhlenbeck) pour tester l'hypothèse nulle $H_0 : \tau = 0$ (sans point de rupture).
3. Une extension des résultats trouvés sera donnée pour le cas où le paramètre λ_1 est inconnu.

3.2 Le processus de la statistique du score et sa limite

On adopte la notation de Matthews *et al.*(1985) pour le modèle avec un point de rupture, spécifié par la fonction de risque :

$$h(t) = \begin{cases} \lambda & \text{si } t < \tau \\ (1-\epsilon)\lambda & \text{si } t \geq \tau \end{cases} \tag{3.1}$$

dont les paramètres ϵ et τ satisfont : $0 \leq \epsilon < 1$ et $\tau \geq 0$.
La densité correspondante est donnée par :

$$f(t) = \begin{cases} \lambda exp(-\lambda t) & \text{si } t < \tau \\ \rho\lambda exp(-\lambda\tau - \rho\lambda(t-\tau)) & \text{si } t \geq \tau \end{cases} \tag{3.2}$$

où $\rho = 1 - \epsilon$.

Remarque 3.1

L'hypothèse nulle $H_0 : \tau = 0$ (sans point de rupture) est équivalente à l'hypothèse $H_0 : \epsilon = 0$.

3.2.1 Statistique du score normalisée

Définition 3.1

Soient $T_1, T_2, ..., T_n$ des variables aléatoires indépendantes et identiquement distribuées selon la densité (3.2) avec fonction de vraisemblance L. Quand λ est connu la statistique du score normalisée est :

$$\left(\frac{d\,log\,L}{d\epsilon}\right)\left[E\left(-\frac{d^2 log\,L}{d\epsilon^2}\right)\right]^{-\frac{1}{2}}.$$

Matthews *et al.*(1985) ont noté que la statistique du score normalisée évaluée à $\epsilon = 0$, pour τ fixé, peut s'écrire sous la forme :

$$Z_n(\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} e^{\frac{\lambda\tau}{2}} (\lambda T_i - \lambda\tau - 1) H(T_i - \tau) \tag{3.3}$$

avec $H(T_i - \tau) = 1$ si $T_i \geq \tau$, et $H(T_i - \tau) = 0$ sinon.

Il suffit de remarquer, à partir du logarithme de la fonction de vraisemblance, que :

$$\begin{aligned} log\,L(\epsilon, \underline{T}) &= \sum_{i=1}^{n} log[f(T_i)] \\ &= \sum_{i:T_i<\tau} [log(\lambda) - \lambda T_i] + \sum_{i:T_i\geq\tau} [log((1-\epsilon)\lambda) - \lambda\tau - \lambda(1-\varepsilon)(T_i - \tau)], \end{aligned} \tag{3.4}$$

et on a alors :

$$\begin{aligned} \frac{d\,log\,L}{d\epsilon} &= \sum_{i=1}^{n}\left[-\frac{1}{1-\epsilon} - \lambda\tau + \lambda T_i\right] H(T_i - \tau), \\ \frac{d^2 log\,L}{d\epsilon^2} &= \sum_{i=1}^{n} \frac{-H(T_i - \tau)}{(1-\epsilon)^2}, \end{aligned}$$

et par conséquent :

$$\begin{aligned} \left.\frac{d\,log\,L}{d\epsilon}\right|_{\epsilon=0} &= \sum_{i=1}^{n}\left[-1 - \lambda\tau + \lambda T_i\right] H(T_i - \tau), \qquad (3.5) \\ E\left[-\frac{d^2 log\,L}{d\epsilon^2}\right] &= \sum_{i=1}^{n} E\left[\frac{H(T_i - \tau)}{(1-\epsilon)^2}\right] \\ &= \sum_{i=1}^{n} \int_{\tau}^{\infty} \frac{1}{(1-\epsilon)^2} f(T_i)\, dT_i. \end{aligned}$$

Sachant que $T_i \geq \tau$ et $\rho = 1$ pour ϵ évalué à 0, alors :

$$E\left[-\frac{d^2 log\, L}{d\epsilon^2}\right]\Bigg|_{\epsilon=0} = \sum_{i=1}^{n} \int_{\tau}^{\infty} \lambda\, e^{-\lambda T_i}\, dT_i$$
$$= n e^{-\lambda\tau}. \tag{3.6}$$

Donc, à partir de (3.5) et (3.6), la statistique du score normalisée évaluée à $\epsilon = 0$, peut s'écrire sous la forme (3.3).

3.2.2 La convergence faible des $Z_n(\tau)$

Nous suiverons ici l'approche de Matthews *et al.*(1985).
La statistique du score normalisée sous l'hypothèse nulle $H_0 : \epsilon = 0$, avec τ fixé, est donnée par le processus (3.3) :

$$Z_n(\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} e^{\frac{\lambda\tau}{2}} (\lambda T_i - \lambda\tau - 1) H(T_i - \tau)$$

et d'après le théorème de la limite centrale (pour $T_1, ..., T_n$ i.i.d), Z_n a une distribution asymptotique normale $AN(0, 1)$ pour chaque valeur de τ, avec :

$$cov(Z_n(\tau_1), Z_n(\tau_2)) = exp\left[\frac{-\lambda \mid \tau_1 - \tau_2 \mid}{2}\right]$$

Il est facile de démontrer la dernière formule, et de prouver que $E[Z_n(\tau)] = 0$ et que $var[Z_n(\tau)] = 1$ (Voir les preuves dans l'appendice A).

Maintenant, en suivant l'approche de Matthews *et al.*(1985), on cherche la limite de Z_n. Soit F_n la distribution empirique des observations indépendantes $T_1, ..., T_n$ qui suivent la loi exponentielle avec moyenne $1/\lambda$. Le processus Z_n peut s'exprimer par :

$$Z_n(\tau) = (n e^{\lambda\tau})^{-\frac{1}{2}} \int_{\tau}^{\infty} [\lambda(x - \tau) - 1]\, dF_n(x) \tag{3.7}$$

Il suffit de remarquer que :

$$\int g(x) dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} \int g(x)\, dI(T_i \leq x)$$
$$= \frac{1}{n} \sum_{i=1}^{n} g(T_i)$$

où $g(x) = [\lambda(x-\tau)-1]H(x-\tau)$ et $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq x)$.

Notons par $U_n^F(x) = n^{\frac{1}{2}}[F_n(x) - F(x)]$, le processus empirique basé sur les observations $T_1, ..., T_n$, dont la fonction de répartition est F.

Alors, en remarquant que l'intégrale dans (3.7) est égale à zéro si F_n est remplacé par F, la formule (3.7) peut être réécrite comme :

$$Z_n(\tau) = e^{\frac{\lambda\tau}{2}} \int_{\tau}^{\infty} [\lambda(x-\tau)-1]\, dU_n^F(x) \tag{3.8}$$

Une intégration par partie du terme de droite de (3.8), donne :

$$Z_n(\tau) = -e^{\frac{\lambda\tau}{2}} \left[\int_{\tau}^{\infty} \lambda\, U_n^F(x)dx - U_n^F(\tau) \right] \tag{3.9}$$

$U_n(u) = n^{\frac{1}{2}}[F_n(u) - u]$ désigne le processus uniforme empirique basé sur n variables aléatoires indépendantes $uniforme(0,1)$. Étant donné que la distribution de $F^{-1}(1-U)$ est F si U est uniforme en $[0,1]$, une représentation possible de U_n^F en terme du processus U_n est $U_n^F \overset{loi}{=} U_n o(1-F)$ (pour plus de détails, voir Pyke (1972)).

En introduisant le changement de variable $u = 1 - F(x) = e^{-\lambda x}$ et $t = 1 - F(\tau) = e^{-\lambda\tau}$ dans (3.9), nous obtenons :

$$Z_n^*(t) = Z_n(\tau) = t^{-\frac{1}{2}} \left[U_n(t) - \int_0^t u^{-1} U_n(u)\, du \right] \tag{3.10}$$

avec $t \in [0,1]$,et $Z_n^*(t) = Z_n\left(-\frac{1}{\lambda} log(t)\right)$.

Soit $U = \{U(u) : 0 \leq u \leq 1\}$ un pont Brownien. Le processus U_n converge vers U, sous la métrique pondérée ρ_q définie par $\rho_q(f,g) = Sup_{0\leq u\leq 1}\left|\frac{f(u)-g(u)}{u^{\beta}}\right|$, avec $0 < \beta < 1/2$, (voir Pyke et Shorack (1968) et O'Reilly (1974)).

Par conséquent, si :

$$Z^*(t) = t^{-\frac{1}{2}} \left[U(t) - \int_0^t u^{-1} U(u)\, du \right], \tag{3.11}$$

alors, pour $t_0 > 0$, on a :

$$\begin{aligned} Sup_{t\in[t_0,1]}|Z_n^*(t) - Z^*(t)| &= Sup_{t\in[t_0,1]}\left\{t^{-\frac{1}{2}}\left|\int_0^t u^{-1}(U_n(u) - U(u))\,du\right|\right\} \\ &\leq \rho_q(U_n, U)\, Sup_{t\in[t_0,1]}\left\{t^{-\frac{1}{2}}\int_0^t u^{-1+\beta}\,du\right\} \\ &= \rho_q(U_n, U)\,\beta^{-1}\,t_0^{\beta-\frac{1}{2}} \\ &\xrightarrow{p.s} 0 \end{aligned}$$

Ceci montre que $Z_n^* \longrightarrow Z^*$, sous la métrique ρ définie par $\rho(f,g) = Sup_{0\leq u\leq 1}|f(u) - g(u)|$, et que les processus U_n et U sont restreints à l'intervalle $[t_0, 1]$.

3.2.3 Identification de la limite Z^*

En posant $W(t) = U(t) - \int_0^t u^{-1}U(u)\,du$ et $U(t) = B(t) - t\,B(1)$, où $\{B(t), t \geq 0\}$ est le mouvement brownian centré réduit, on a :

$$W(t) = B(t) - \int_0^t u^{-1}B(u)\,du \tag{3.12}$$

et :

$$Z^*(t) = t^{-\frac{1}{2}}\,W(t) \tag{3.13}$$

W(t) est gaussien, continu de moyenne zéro et d'incrémentations indépendantes, avec $Cov(W(s), W(t)) = s$, pour $0 \leq s \leq t$, (voir Matthews *et al.*(1985)). Par conséquent, W est un mouvement brownien centré réduit, (Billingsley, 1968, théorème 19.1).

Alors, de (3.12) et (3.13), on déduit que Z^* est la normalisation de B, et que donc la limite $Z^*(t)$ de $Z_n^*(t) = Z_n(\tau)$ est identifiée par le processus Ornstein-Uhlenbeck avec : $Cov(Z^*(t_1), Z^*(t_2)) = (\frac{t_1}{t_2})^{\frac{1}{2}}$, pour $0 \leq t_1 \leq t_2$.
En effet :

$$\begin{aligned} Cov(Z^*(t_1), Z^*(t_2)) &= Cov(t_1^{-1/2}W(t_1)\,,\, t_2^{-1/2}W(t_2)) \\ &= t_1^{-1/2}\,t_2^{-1/2}\,Cov(W(t_1)\,,\,W(t_2)) \\ &= t_1^{-1/2}\,t_2^{-1/2}\,t_1 \;=\; \left(\frac{t_1}{t_2}\right)^{1/2} \end{aligned}$$

Et puisque $t = e^{-\lambda\tau}$, la fonction de covariance de $Z_n(\tau)$ est $exp\{-\frac{\lambda}{2}|\tau_1 - \tau_2|\}$ (pour plus de détails, voir Matthews *et al.*(1985)).
Ce qui complète la preuve du théorème suivant.

Théorème 3.1

Pour $\tau < \infty$, le processus des statistiques de score (3.3) $\{Z_n(\tau) : 0 < \tau \leq \tau_0\}$ basé sur les variables aléatoires $T_1, ..., T_n$ i.i.d selon la loi exponentielle avec paramètre λ, converge faiblement vers le processus Ornstein-Uhlenbeck $\{Z(\tau) : 0 < \tau \leq \tau_0\}$ avec moyenne zéro et fonction de covariance $exp[-\frac{\lambda}{2}|\tau_1 - \tau_2|]$.
Cette convergence est sous la topologie de Skorohod définie dans Billingsley (1968).

3.3 Test de l'hypothèse nulle $H_0 : \epsilon = 0$

À partir du processus $Z_n(\tau)$ défini dans l'équation (3.3), pour le cas λ connu, et parallèlement au travail de Davies (1977), Matthews *et al.*(1985) proposent un test qui rejette l'hypothèse nulle $H_0 : \epsilon = 0$ (sans point de rupture) pour des valeurs larges de la statistique :

$$M_n(\tau_l, \tau_u) = Sup_{\tau_l \leq \tau \leq \tau_u} Z_n(\tau) \tag{3.14}$$

avec $[\tau_l, \tau_u]$ étant l'intervalle qui recouvre les valeurs de τ (selon Mathews *et al.*(1985), le choix de τ_l et τ_u peut être arbitraire dans les applications).
Par le théorème (3.1), on remarque que :

$$M_n(\tau_l, \tau_u) \longrightarrow M(\tau_l, \tau_u) = Sup_{\tau_l \leq \tau \leq \tau_u} Z(\tau)$$

où $Z(\tau) = Z^*(t)$ est la limite Ornstein-Uhlenbeck (3.13) du processus $Z_n(\tau)$.
Selon Matthews *et al.*(1985), il suffit de prendre $\tau_l = 0$ et d'utiliser simplement l'écriture $M(\tau_u)$ (par le fait que les temps de survie sont positifs).

Pour $c > 0$, Matthews *et al.*(1985) introduisent la probabilité :

$$P(M(\tau_u) \geq c) = 1 - P(\tau_u, c) \tag{3.15}$$

Cette probabilité est introduite pour trouver le niveau de signification asymptotique de l'hypothèse nulle $H_0 : \epsilon = 0$, où $P(\tau_u, c)$ exprime la probabilité que le processus Ornstein-Uhlenbeck ne dépasse pas la valeur c dans un intervalle de longeur τ_u.

Mandl (1962) fournit des tables pour calculer cette probabilité et pour approximer le niveau de signification asymptotique du test de l'hypothèse $H_0 : \epsilon = 0$ (sans point de rupture) (voir Mathews *et al.*(1985) et Mandl (1962) pour plus de détails).

3.4 Le cas de λ inconnu

3.4.1 La statistique du score partielle

La statistique du score normalisée spécifiée par (3.3) n'est plus valide pour tester l'hypothèse $H_0 : \epsilon = 0$, (sans point de rupture) dans le cas où λ est inconnu.
Une alternative appropriée est la statistique du score partielle.

Définition 3.2

Soient $T_1, T_2, ..., T_n$ des variables aléatoires i.i.d selon la densité (3.2) avec fonction de vraisemblance L, la statistique du score partielle évaluée à (λ_0, ϵ_0) est :

$$\left(\frac{d\log L}{d\epsilon}\right)\left[E\left(-\frac{d^2\log L}{d\epsilon^2}\right) - E\left(-\frac{d^2\log L}{d\epsilon\, d\lambda}\right)E\left(-\frac{d^2\log L}{d\lambda^2}\right)^{-1}E\left(-\frac{d^2\log L}{d\lambda\, d\epsilon}\right)\right]^{-\frac{1}{2}}\Bigg|_{\lambda=\lambda_0,\,\epsilon=\epsilon_0} \tag{3.16}$$

(voir Lawless (1982), p 524).

Sous l'hypothèse $H_0 : \epsilon = 0$, la statistique (3.16) évaluée à $\lambda = \hat{\lambda}_n$, peut s'écrire sous la forme :

$$\widehat{Z}_n(\tau) = \left[n e^{-\lambda\tau}(1 - e^{-\lambda\tau})\right]^{-\frac{1}{2}} \frac{d\log L}{d\epsilon}\Bigg|_{\epsilon=\epsilon_0,\lambda=\hat{\lambda}_n} \tag{3.17}$$

avec $\hat{\lambda}_n = n/\sum_{i=1}^{n} T_i$ étant l'estimateur de maximum de vraisemblance de λ sous $H_0 : \epsilon = 0$.

En effet, pour démontrer (3.17), à partir de l'expression (3.4) :

$$log\, L(\epsilon, \underline{T}) = \sum_{i:T_i<\tau} [log(\lambda) - \lambda T_i] + \sum_{i:T_i\geq\tau} [log((1-\epsilon)\lambda) - \lambda\tau - \lambda(1-\varepsilon)(T_i-\tau)]$$

on a :

$$\begin{aligned}
\frac{dlogL}{d\lambda} &= \sum_{i:T_i<\tau}\left[\frac{1}{\lambda} - T_i\right] + \sum_{i:T_i\geq\tau}\left[\frac{1}{\lambda} - \tau - (1-\epsilon)(T_i-\tau)\right] \\
\frac{d^2logL}{d\lambda^2} &= \sum_{i=1}^{n} -\frac{1}{\lambda^2} = -\frac{n}{\lambda^2} \\
\frac{d^2logL}{d\lambda\, d\epsilon} &= \frac{d^2logL}{d\epsilon\, \lambda} = \sum_{i=1}^{n}(T_i-\tau)H(T_i-\tau)
\end{aligned}$$

et par conséquent :

$$\begin{aligned}
E\left[\frac{d^2logL}{d\lambda^2}\right] &= \frac{n}{\lambda^2} \qquad (3.18)\\
E\left[\frac{d^2logL}{d\epsilon\, d\lambda}\right] &= \sum_{i=1}^{n} E[-(T_i-\tau)H(T_i-\tau)] = \sum_{i=1}^{n}\int_{\tau}^{\infty} -(T_i-\tau)f(T_i)\, dT_i
\end{aligned}$$

avec f la densité définie par (3.2).

Sachant que pour ϵ évaluée à zéro, $T_i \geq \tau$ et $\rho = 0$, alors :

$$\begin{aligned}
E\left[\frac{d^2logL}{d\epsilon\, d\lambda}\right]\Big|_{\epsilon=0} &= \sum_{i=1}^{n}\int_{\tau}^{\infty} -(T_i-\tau)\lambda\, e^{-\lambda T_i} dT_i \\
&= \sum_{i=1}^{n} -\frac{e^{-\lambda\tau}}{\lambda} = -\frac{n\, e^{-\lambda\tau}}{\lambda} \qquad (3.19)
\end{aligned}$$

En remarquant que $E\left[-\frac{d^2logL}{d\epsilon^2}\right]\Big|_{\epsilon=0} = n\, e^{-\lambda\tau}$, et à partir de (3.18) et (3.19), la statistique du score partielle évaluée à $(\epsilon = 0, \lambda = \hat{\lambda}_n)$ peut s'écrire sous la forme (3.17).

3.4.2 La convergence du processus $\hat{Z}_n(\tau)$

D'après l'équation (3.3), $Z_n(\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} e^{\frac{\lambda\tau}{2}}(\lambda T_i - \lambda\tau - 1)H(T_i-\tau)\big|_{\epsilon=0}$, on peut écrire (3.17) sous la forme :

$$\hat{Z}_n(\tau) = (1-e^{-\lambda\tau})^{-\frac{1}{2}} Z_n(\tau,\lambda)\big|_{\lambda=\hat{\lambda}_n}, \qquad (3.20)$$

où $Z_n(\tau,\lambda) = Z_n(\tau)$. Mais cette nouvelle notation indique la présence du paramètre inconnu λ dans sa forme fonctionnelle.

Selon l'équation (3.7), $Z_n(\tau,\hat{\lambda}_n) = (ne^{\hat{\lambda}_n\tau})^{\frac{1}{2}} \int_\tau^\infty [\hat{\lambda}_n(x-\tau)-1]\, dF_n(x)$. En suivant les étapes de (3.7) à (3.9), on a :

$$\begin{aligned}
\int_\tau^\infty [\hat{\lambda}_n(x-\tau)-1]\, dF_n(x) &= \int_\tau^\infty [\lambda(x-\tau)-1]\, dF_n(x) + (\hat{\lambda}_n-\lambda)\int_\tau^\infty (x-\tau)\, dF_n(x) \\
&= (n\, e^{\lambda\tau})^{-\frac{1}{2}}\, Z_n(\tau,\lambda) + \frac{\hat{\lambda}_n-\lambda}{\lambda}\int_\tau^\infty [\lambda(x-\tau)-1]\, dF_n(x) \\
&\quad + \frac{\hat{\lambda}_n-\lambda}{\lambda}\int_\tau^\infty dF_n(x) \\
&= (n\, e^{\lambda\tau})^{-\frac{1}{2}}\, Z_n(\tau,\lambda) + \frac{\hat{\lambda}_n-\lambda}{\lambda}\left[-n^{-\frac{1}{2}}\,\lambda\int_\tau^\infty U_n^F(x)\, dx - n^{-\frac{1}{2}}\, U_n^F(\tau)\right] \\
&\quad + \frac{\hat{\lambda}_n-\lambda}{\lambda}\, F_n(\tau).
\end{aligned}$$

et sachant que $n^{-\frac{1}{2}}\, U_n^F(\tau) = F_n(\tau) - F(\tau)$, on a :

$$\begin{aligned}
\int_\tau^\infty [\hat{\lambda}_n(x-\tau)-1]\, dF_n(x) &= (n\, e^{\lambda\tau})^{-\frac{1}{2}}\, Z_n(\tau,\lambda) - (\hat{\lambda}_n-\lambda)\int_\tau^\infty n^{-\frac{1}{2}}\, U_n^F(x)\, dx \\
&\quad + \frac{\hat{\lambda}_n-\lambda}{\lambda}(1-F(\tau)).
\end{aligned}$$

avec la remarque :

$$\begin{aligned}
-\int_0^\infty U_n^F(x)\, dx &= -n^{\frac{1}{2}}\int_0^\infty \Big[F_n(x) - F(x))\, dx\Big] \\
&= \frac{-n^{\frac{1}{2}}}{n}\sum_{i=1}^n \int_0^\infty \Big[I(T_i \le x) - (1-e^{-\lambda x})\Big]\, dx \\
&= n^{-\frac{1}{2}}\sum_{i=1}^n \int_0^\infty \Big[I(T_i > x) - e^{-\lambda x}\Big]\, dx \\
&= n^{-\frac{1}{2}}\left[\sum_{i=1}^n T_i - \frac{n}{\lambda}\right] \\
&= n^{\frac{1}{2}}\left(\frac{1}{\hat{\lambda}_n} - \frac{1}{\lambda}\right)
\end{aligned}$$

on a le résultat :

$$\begin{aligned}\int_\tau^\infty [\hat{\lambda}_n(x-\tau)-1]\,dF_n(x) &= (n\,e^{\lambda\tau})^{-\frac{1}{2}}\,Z_n(\tau,\lambda)-(\hat{\lambda}_n-\lambda)\int_\tau^\infty n^{-\frac{1}{2}}\,U_n^F(x)\,dx\\ &\quad+\hat{\lambda}_n\Big(\frac{1}{\lambda}-\frac{1}{\hat{\lambda}_n}\Big)(1-F(\tau))\\ &= (n\,e^{\lambda\tau})^{-\frac{1}{2}}\,Z_n(\tau,\lambda)-(\hat{\lambda}_n-\lambda)\int_\tau^\infty n^{-\frac{1}{2}}\,U_n^F(x)\,dx\\ &\quad+\hat{\lambda}_n(1-F(\tau))\int_0^\infty n^{-\frac{1}{2}}\,U_n^F(x)\,dx\end{aligned}$$

et par conséquent :

$$\begin{aligned}\hat{Z}_n(\tau) &= (1-e^{-\hat{\lambda}_n\tau})^{-\frac{1}{2}}Z_n(\tau,\hat{\lambda}_n)\\ &= (1-e^{-\hat{\lambda}_n\tau})^{-\frac{1}{2}}e^{\frac{(\hat{\lambda}_n-\lambda)\tau}{2}}\Big[Z_n(\tau,\lambda)+(\hat{\lambda}_n-\lambda)\int_\tau^\infty U_n^F(x)\,dx\\ &\quad+\hat{\lambda}_n\,e^{\frac{-\lambda\tau}{2}}\int_0^\infty U_n^F(x)\,dx\Big]\end{aligned}$$

D'après (3.9), $\int_\tau^\infty U_n^F(x)\,dx=-\lambda^{-1}e^{\frac{-\lambda\tau}{2}}(Z_n(\tau,\lambda)-e^{\frac{\lambda\tau}{2}}U_n^F(\tau))$, donc :

$$\begin{aligned}\hat{Z}_n(\tau) &= (1-e^{-\hat{\lambda}_n\tau})^{-\frac{1}{2}}e^{\frac{(\hat{\lambda}_n-\lambda)\tau}{2}}\Big\{Z_n(\tau,\lambda)+\hat{\lambda}_n e^{\frac{-\lambda\tau}{2}}\int_0^\infty U_n^F(x)\,dx\\ &\quad+(\hat{\lambda}_n-\lambda)\lambda^{-1}\Big[Z_n(\tau,\lambda)-e^{\frac{\lambda\tau}{2}}U_n^F(\tau)\Big]\Big\}\end{aligned}\tag{3.21}$$

Soit $\hat{Z}_n^*(t)=\hat{Z}_n(\tau)$, avec $u=1-F(x)=e^{-\lambda x}$ et $t=1-F(\tau)=e^{-\lambda\tau}$.
En appliquant (3.11) et (3.12) sur (3.21), $\hat{Z}_n^*$ converge uniformément dans l'intervalle $[t_0,t_1]$ (Matthews *et al.*(1985)) avec probabilité 1 vers :

$$\begin{aligned}\hat{Z}^*(t) &= (1-t)^{\frac{-1}{2}}\Big\{Z^*(t)+\sqrt{t}\int_0^1 U(u)\,du\Big\}\\ &= (t(1-t))^{\frac{-1}{2}}\Big\{U(t)-\int_0^t u^{-1}U(u)\,du+t\int_0^1 u^{-1}U(u)\,du\Big\}\\ &= \frac{W(t)-t\,W(1)}{[t(1-t)]^{\frac{1}{2}}}\end{aligned}\tag{3.22}$$

avec W le mouvement Brownien défini dans (3.12), et $0<t_0\le t_1<1$.

Cette preuve est suffisante pour montrer que $\hat{Z}_n^*\longrightarrow\hat{Z}^*$ sous la topologie de Skorohod de l'espace métrique $D([t_0,t_1])$ (Matthews *et al.* (1985)).

3.4.3 Niveau de signification asymptotique

Dans le cas de λ inconnu, Matthews *et al.*(1985) ont noté qu'un test approprié pour l'hypothèse « sans point de rupture » est un test qui rejette $H_0 : \epsilon = 0$ pour des valeurs larges de la statistique :

$$Sup_{\tau \in [\tau_l, \tau_u]} \hat{Z}(\tau)$$

avec $\hat{Z}(\tau) = \hat{Z}^*(t)$, et $[\tau_l, \tau_u]$ l'intervalle qui couvre les valeurs de τ.

Pour évaluer le niveau de signification asymptotique donné par

$$P\left[Sup_{\tau \in [\tau_l, \tau_u]} \hat{Z}(\tau) > c\right] \quad , \qquad c > 0 \tag{3.23}$$

Mandl (1962) et Matthews *et al.*(1985) proposent d'utiliser la probabilité

$$1 - P\left(Z(s) < c \ ; \ 0 \leq s \leq log\left\{\frac{t_1(1-t_0)}{t_0(1-t_1)}\right\}\right) = 1 - P\left(log\left\{\frac{t_1(1-t_0)}{t_0(1-t_1)}\right\}; c\right) \tag{3.24}$$

comme approximation de niveau de signification de l'hypothèse « sans point de rupture » $H_0 : \tau = 0$.

La probabilité $P\left(log\left\{\frac{t_1(1-t_0)}{t_0(1-t_1)}\right\}; c\right)$ représente la probabilité qu'un processus Ornstein-Uhlenbeck ne dépasse pas la valeur c dans un intervalle de temps de longueur $log\left\{\frac{t_1(1-t_0)}{t_0(1-t_1)}\right\}$, qu'on peut calculer à partir des tables de Mandl (1962).
Notons que $t_0 = 1 - F(\tau_l) = e^{-\lambda \tau_l}$ et $t_1 = 1 - F(\tau_u) = e^{-\lambda \tau_u}$

CHAPITRE IV

MODÈLE DES RISQUES PROPORTIONNELS AVEC UN POINT DE RUPTURE

4.1 Introduction

Le modèle des risques proportionnels (Cox, 1972) a reçu ces dernières années un remarquable intérêt pour analyser les données de temps de survie. La nature semi-paramétrique du modèle a permis de quantifier l'effet des covariables comme l'âge ou la pression sanguine, sur le risque de la survie d'un individu. Cependant, l'hypothèse des risques instantanés proportionnels du modèle peut ne pas satisfaire certains cas de données. Voir par exemple Lagakos et Schoenfeld (1984), Struthers et Kalbfleisch (1986), Lin et Wei (1989).
Quelques alternatives du modèle de Cox existent dans la littérature, notamment le modèle de temps de survie accéléré (Cox et Oakes, 1984), ou encore le modèle de régression linéaire des rangs de Prentice (1978).

Dans ce chapitre, on présentera le modèle des risques proportionnels avec un point de rupture introduit par Liang, Self et Liu (1990). Ce modèle peut être vu comme un cas spécial du modèle de Cox avec des covariables dépendantes du temps et évoque l'hypothèse que l'introduction de covariables dépendantes du temps dans le modèle de Cox va dans le sens de l'hypothèse de proportionnalité des risques instantanés du modèle.
Une première étape dans l'étude de l'ajustement du modèle de Cox avec un point de rupture est de tester la nécéssité d'un tel changement ou d'une telle rupture.

Dans la section 2, on présente la définition de liang *et al.*(1990) du modèle avec certaines de ses caractéristiques. La statistique du score partielle sera présentée dans la section 3 pour tester l'hypothèse nulle $H_0 : \tau = 0$ (sans point de rupture), et quelques difficultés d'inférence seront brièvement décrites et discutées. À la fin de ce chapitre, une application du modèle sur une étude de différents traitements du cancer du poumon sera donnée.

4.2 Le modèle de Cox avec un point de rupture

Tout d'abord, il faut mentionner que le modèle des risques proportionnels avec un point de rupture est une extension du modèle proposé par Matthews et Farewell (1982, 1985), et que Liang, Self et Liu (1990) l'ont introduit et illustré dans le contexte d'une étude épidémiologique sur l'hypertension dans un groupe d'étudiants de médecine de race blanche, entre les années 1948 et 1964 à l'université John Hopkins.
Leur modèle est défini par :

$$\lambda(t \mid z, x) = \lambda_0(t)\, exp[(\beta + \theta\, I(t \leq \tau))z + \gamma' x] \tag{4.1}$$

où $\lambda_0(t)$ est la fonction de risque instantané de base, indépendante des covariables, z, le facteur de risque relié à différentes magnitudes du temps t (par exemple l'âge), τ, le point de rupture, et x, le vecteur de covariables indépendantes du temps t.

Alors, $e^{\beta z}$ est le risque relatif de l'apparition d'une maladie (ou un autre événement) sur un individu âgé de plus de τ avec facteur de risque z par rapport à un individu du même âge avec facteur $z = 0$. D'autre part, $e^{(\beta+\theta)z}$ est le risque relatif des mêmes individus, mais avec des âges inférieurs à τ.
Ce modèle possède les deux caractéristiques suivantes :

1. Différents paramètres correspondant aux risques relatifs (β et θ) peuvent différer selon l'apparition précoce ou tardive de la maladie étudiée (apparition de la maladie avant ou après l'âge τ).
2. L'introduction d'un paramètre supplémentaire évite de devoir préciser dès le début, le point de rupture τ auquel les risques relatifs changent de valeurs.

Dans le but de dériver une alternative du modèle de Cox dans le cas où les risques ne sont pas proportionnels, Breslow, Edler et Berger (1984) proposent un modèle sous la forme (4.1) avec $x = 0$, z dichotomique et $I(t \leq 0)$ remplacé par une fonction lisse et connue du temps $g(t)$, le choix de $g(t)$ étant arbitraire (par exemple log(t)).

4.3 Test de l'hypothèse $H_0 : \theta = 0$, sans point de rupture

4.3.1 Statistique de test

Soient $0 < t_{(1)} < ... < t_{(k)}$ les temps de survie distincts de k individus, et $R(t)$ l'ensemble des sujets à risque juste avant t.
La fonction de vraisemblance partielle est donnée par :

$$L(\beta, \gamma, \theta, \tau) = \prod_{i=1}^{k} \frac{exp[(\beta + \theta\, I(t_{(i)} \leq \tau))Z_{(i)} + \gamma' x_{(i)}]}{\left\{\sum_{j \in R(t_{(i)})} exp[(\beta + \theta\, I(t_{(i)} \leq \tau))Z_{ij} + \gamma' x_{ij}]\right\}^{d_i}} \tag{4.2}$$

où $Z_{(i)} = \sum_{j \in R(t_{(i)})} Z_{ij}$ est la somme des facteurs Z_{ij} (covariable dépendante du temps t de l'individu j) pour tous les individus qui ont subi une maladie (ou autre événement) au temps $t_{(i)}$, $x_{(i)} = \sum_{j \in R(t_{(i)})} x_{ij}$ est la somme des facteurs x_{ij} (covariable indépendante du temps t de l'individu j) pour les individus subissant une maladie au temps $t_{(i)}$, et d_i est le nombre d'événements surgissant au temps $t_{(i)}$.

Notre intérêt se porte sur l'hypothèse $H_0 : \theta = 0$, sans point de rupture, qui revient à dire que l'effet du facteur de risque Z est le même sur les individus qui ont subi une maladie précoce ou tardive.

Une remarque importante à signaler, est que l'approximation Chi-deux de la distribution asymptotique des statistiques de test standard comme du test des rapports de vraisemblances ou du score, ne s'applique pas ici pour deux raisons : la première est que la fonction de vraisemblance $L(\beta, \gamma, \theta, \tau)$ n'est pas une fonction lisse de τ. Deuxièmement, le paramètre de nuisance τ est présent seulement dans le cas où $\theta \neq 0$, et donc, le nombre de degrés de liberté pour l'approximation Chi-deux (en supposant que cette approximation est appropriée à notre cas) n'est pas précis.

En s'inspirant du travail de Matthews *et al.*(1985), Liang *et al.*(1990) proposent d'utiliser comme statistique de test :

$$M = Sup_{\tau \in [a,b]}[S(\tau)] \tag{4.3}$$

avec

$$S(\tau) = \left(\frac{d\,log\,L}{d\,\theta}\right)\left[-\frac{d^2 log\,L}{d\theta^2} - \left(-\frac{d^2 log\,L}{d\,\theta\,d\delta}\right)' \left(-\frac{d^2 log\,L}{d\,\delta^2}\right)^{-1}\left(-\frac{d^2 log\,L}{d\,\theta\,d\delta}\right)\right]^{-\frac{1}{2}}_{\delta=\hat{\delta},\,\theta=0} \tag{4.4}$$

où $\hat{\delta} = (\hat{\beta}, \hat{\gamma})$ est l'estimateur du maximum de vraisemblance de $\delta = (\beta, \gamma)$ sous l'hypothèse nulle $H_0 : \theta = 0$, et l'intervalle $[a, b]$ couvre les valeurs de τ (a et b peuvent être choisis arbitrairement, Liang *et al.*(1990)).

Selon Liang *et al.*(1990), $S(\tau)$ peut être interprétée comme la statistique du score partielle pour tester $H_0 : \theta = 0$, qu'on peut calculer au cas où τ est fixé et connu.

En l'absence des covariables x du modèle (4.1) et avec la transformation $v(\tau)$ de τ donnée par :

$$v = v(\tau) = \frac{(-d^2 logL/d\theta^2)(\beta, \gamma, 0, \tau)}{(-d^2 logL/d\theta^2)(\beta, \gamma, 0, \infty)} \tag{4.5}$$

$S(.)$ converge vers le processus Ornstein-Uhlenbeck quand $k \longrightarrow \infty$ (voir l'appendice de Liang *et al.*(1990) et Prentice et Self (1983)). Le processus est de moyenne 0, de variance 1, et de corrélation $e^{|v_2 - v_1|}$.

Ainsi, la probabilité :

$$1 - P\left(log\left\{\frac{v(b)(1-v(a))}{v(a)(1-v(b))}\right\}; M\right) \tag{4.6}$$

permet d'approximer le niveau de signification de test $H_0 : \theta = 0$ (discuté dans la section (3.5.1)), tout en utilisant les tables de Mandl (1962).

Si les covariables x sont présentes dans le modèle (4.1), le même résultat est obtenu sous la condition que :

$$B(\tau) = I^{-1}_{\beta\beta}(\tau) I_{\beta\gamma}(\tau) = \lim_{k \to \infty}\left[\frac{d^2 logL}{d\beta^2}(\beta, \gamma, 0, \tau)\right]^{-1}\left[\frac{d^2 logL}{d\beta d\gamma}(\beta, \gamma, 0, \tau)\right] \tag{4.7}$$

soit indépendante de τ. Plus de détails dans Liang *et al.*(1990) et Prentice et Self (1983).

4.3.2 Estimation par intervalle des paramètres β, θ et γ

Liang, Self et Liu (1990) proposent de tester l'hypoythèse $H_0 : \theta = 0$ par le biais des estimations par intervalles pour β, θ et γ. Ils suggèrent de faire l'inférence sur ces paramètres conditionnellement à $\hat{\tau}$, estimateur de τ, afin d'éviter une possible variation de $\hat{\tau}$.

Plusieurs raisons justifient cette approche, en premier lieu, éviter la difficulté statistique de l'analyse de problème créée dans le cas où la vraisemblance n'est pas nécessairement une fonction lisse de τ. La deuxième raison importante concerne l'interprétation de θ dépendemment de la valeur de τ. Bien que $e^{\theta(\tau)}$ puisse être interprété comme le rapport de deux risques relatifs en dépit de τ, l'interprétation de $e^{\theta(\tau)}$ dépend de τ dans le contexte du modèle (4.1). Par exemple, si $\theta(\tau_1) = \theta(\tau_2)$ pour $\tau_1 \neq \tau_2$, l'interprétation de $\theta(\tau_1)$ est différente de $\theta(\tau_2)$ dans le contexte du modèle (4.1), du fait que ce ne soit pas la même rupture de temps.

4.4 Application

4.4.1 Description de l'étude

Dans l'étude Lad, Rubinstein, Sadeghi *et al.*(1988), des essais cliniques ont été menés auprès de 172 patients atteints du cancer du poumon. L'étude a été planifiée pour évaluer le temps écoulé entre le début d'un traitement et la récurrence du cancer. Les patients ont été randomisés selon le protocole expérimental entre les années 1979 et 1985 ; 78 patients ont suivi une radiothérapie (groupe 1), et 64 patients ont subi le même traitement plus des doses de C.A.P (cytoxan, doxorubicin et platinum) (groupe 0). Les données ont été publiées dans Steven Piantadosi (1997).

Notons qu'on a pris ici un échantillon de 142 patients parmi les 172 de l'étude, pour illustrer le test proposé dans ce chapitre.

Nous nous sommes intéressés à savoir si le risque de la récurrence de la maladie est le même entre les deux groupes de patients, ou bien, si à un moment donné (point de rupture qu'on doit identifier) le risque change de valeur entre ces deux groupes.

Les données ordonnées de cette étude sont présentées dans le tableau (4.1).

Tableau 4.1 Les temps de la récurrence du cancer du poumon (en jours).

Radiothérapie plus C.A.P											
9	22	35	53	76	81	94	97	103	114	115	121*
126	147	154	162*	167	190	205	211	211	217	231*	255
262	264	266	271	277	292*	295	313	316	342	342*	352
361	367	376	382	384	392	402	403	410	414	441*	448
515	538*	540	551	580	640	680*	716*	772	807	931*	943*
966	1089	1248	1314								
Radiothérapie											
18	23*	25	27	28	30	36	45	55	56	57	57
57	59	62	63	638	79	85	96	97	97	105	109
113	114	115*	116	125	125	125	127	131*	133	142	146
149	164	165	165	172	174	175	189	206	232	234	241*
245	248	248*	252	266	273	301	317	317	330	330	351
364	388*	395	413	423	436	473	475	612	679	700	757
961	973	1026*	1087	1289*	1388						

(*) observation censurée.

4.4.2 Résultats

La figure (4.1) montre la courbe Kaplan-Meier de la distribution du temps de la récurrence du cancer selon deux niveaux de la covariable « traitement » (radiothérapie et radiothérapie plus C.A.P). D'une part, la croissance du risque de la récurrence de la maladie chez le groupe 1 traité par la radiothérapie, est plus rapide que celle du groupe 0 traité par radiothérapie + C.A.P. D'autre part, la figure indique que, approximativement juste après la $200^{ème}$ journée, les deux courbes Kaplan-Meier représentant les deux groupes sont presque parallèles. Ceci suggère que le traitement, mesuré par la covariable Z, peut

être un facteur de risque pour un temps précoce (un temps de survie inférieur à 200 jours), mais ne l'est pas pour des temps tardifs de la récurrence du cancer (des temps de survie supérieurs à 200 jours).

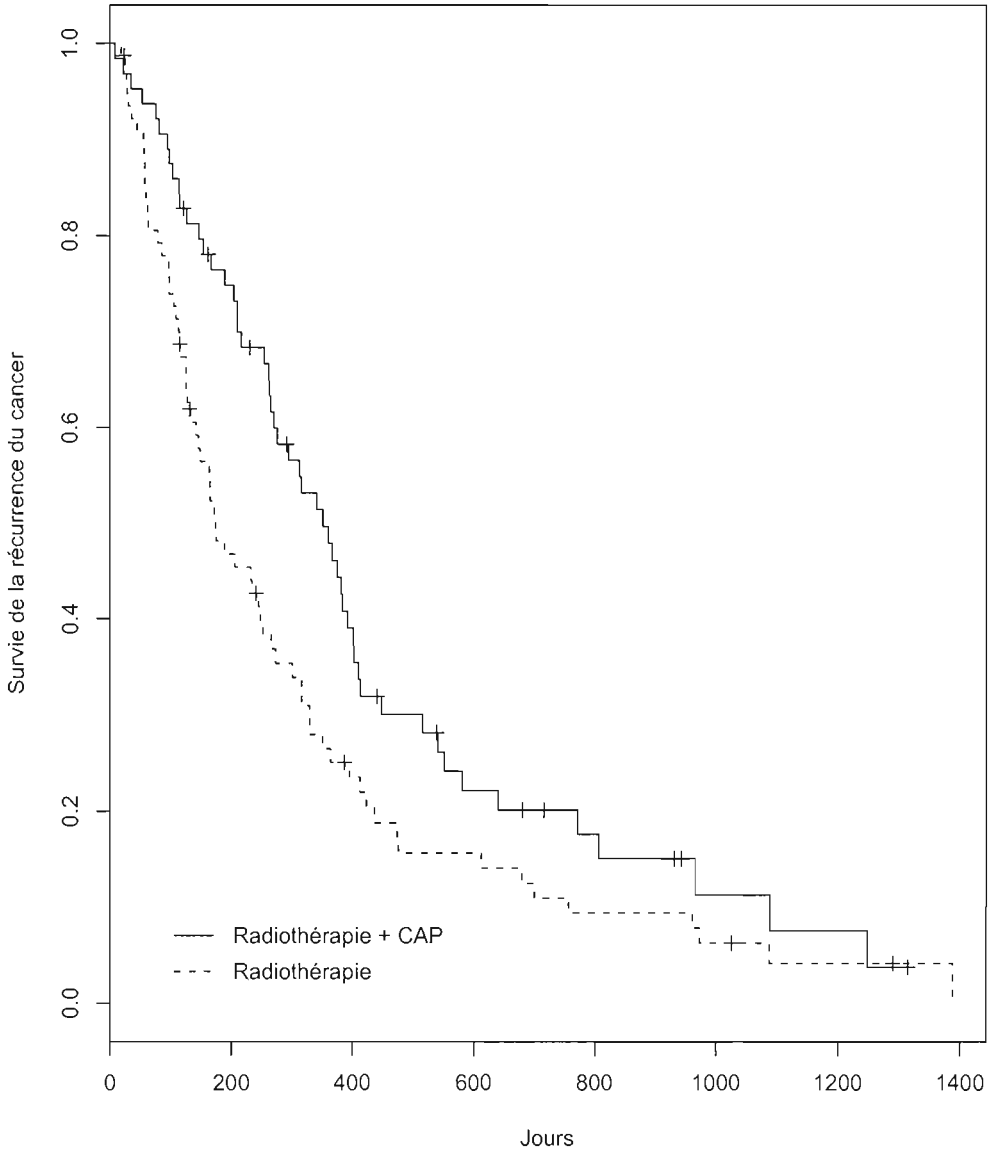

Figure 4.1 Courbes de Kaplan-Meier de la fonction de survie du temps de la récurrence du cancer pour les deux traitements, radiothérapie et radiothérapie + C.A.P .

Notre intérêt se porte maintenant sur l'hypothèse $H_0 : \theta = 0$, *i.e.* le risque relatif de la récurrence du cancer pour les deux groupes qui ont subi deux traitements différents (mesuré par la covariable Z) est le même à travers le temps.

Davies (1977) suggère pour le test de l'hypothèse bilatérale $H_0 : \theta = 0$ *vs* $H_a : \theta \neq 0$ d'utiliser la statistique du test :

$$M = Sup_{\tau \in [a,b]} |S(\tau)| \tag{4.8}$$

où a et b sont choisis arbitrairement. Une estimation du maximum de vraisemblance du coefficient de la covariable Z donne la valeur 0.43.
Pour $a = 109$ et $b = 966$, on observe :

$$M = Sup_{\tau \in [109,966]} |S(\tau)| = |S(966)| = 3.67$$

Pour calculer le niveau de signification asymptotique de ce test, nous utilisons l'expression (4.5) pour transformer a et b, et nous obtenons $v(a) = 0.2461$ et $v(b) = 0.9672$.

Par conséquent :

$$\begin{aligned} P(M \geq 3.67 \,|H_0) &= 2 \times P(S(966) \geq 3.67 \,|H_0) \\ &\simeq 2 \times \left\{ 1 - P\left[log\left(\frac{0.9672(1-0.2461)}{0.2461(1-0.9672)} \right); 3.67 \right] \right\} \\ &= 2 \times [1 - P(4.506\ ;\ 3.6)] \end{aligned}$$

avec $P(t; A)$ représentant la probabilité qu'un processus Ornstein-Uhlenbeck ne dépasse pas la valeur A dans un intervalle de longueur t.

À partir de la table 1 de Mandl (1962), nous observons une approximation de 0.006 de la p-valeur de H_0. Donc, l'existence d'un point de rupture est significative, et la réaction des tumeurs envers les deux traitements dépend fortement de ce point. Un traitement peut être efficace avant le temps de rupture (206 jours), alors que l'autre peut avoir un effet positif après ce temps là.

Nous allons maintenant estimer le risque relatif entre les deux groupes (groupe 0 et groupe 1). La figure (4.2) montre la courbe du logarithme de la fonction « profile »de vraisemblance partielle $l(\hat{\beta}(\tau), \hat{\theta}(\tau), \tau)$, en fonction de τ, où $\hat{\beta}(\tau)$ et $\hat{\theta}(\tau)$ représentent les estimateurs de maximum de vraisemblance de β et θ (resp) pour τ fixé. Notons que le paramètre γ discuté dans les sections précédentes, ne figure pas ici, à cause de l'absence des covariables x. Ce graphique montre que la fonction « profile » de vraisemblance est maximisée à $\hat{\tau} = 206$.

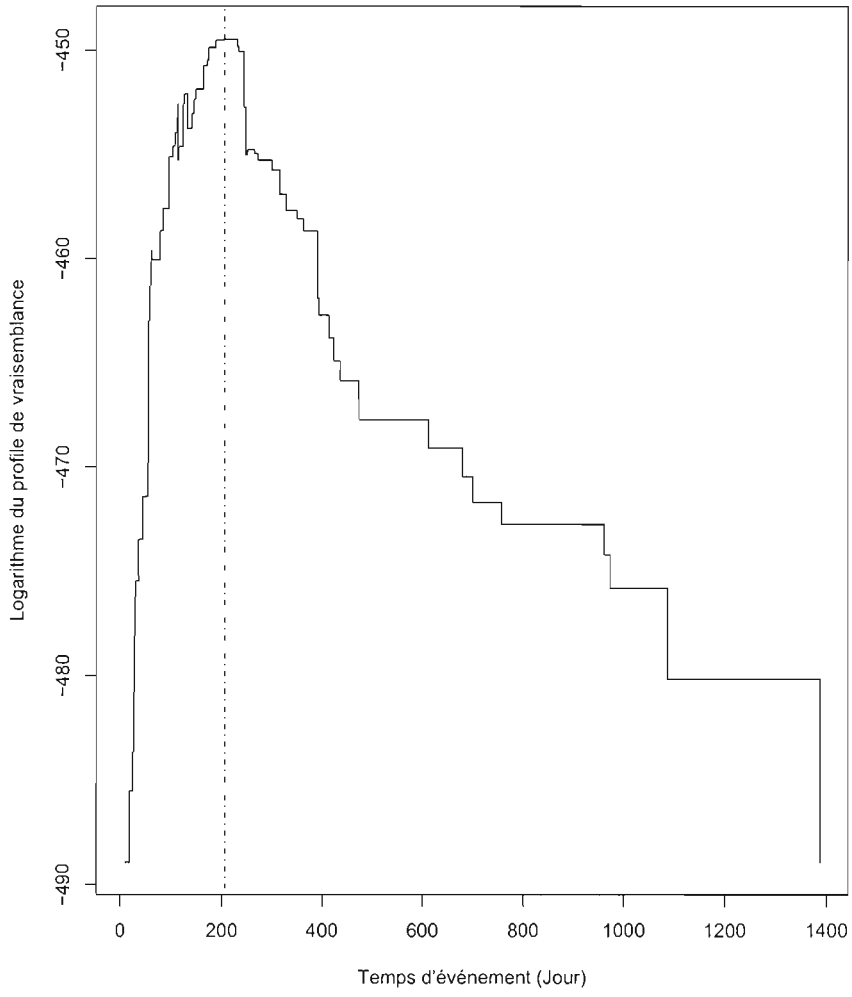

Figure 4.2 Graphique du logarithme du « profile » de vraisemblance partielle *vs* τ.

Les estimations des coefficients de régression du modèle de Cox (4.1) avec un point de rupture τ fixé à 206 sont reportées dans la table (4.2) avec leurs écarts-types.

Tableau 4.2 Les estimations des coefficients de régression avec les écarts-types du modèle (4.1) utilisant les données de temps de la récurrence du cancer du poumon.

Facteur de risque	coefficient	e^{coef}	écart-type	p-valeur
Traitement (Z)	-0.338	0.713	0.238	0.16
$Z.I(t \leq \tau)$	2.885	17.898	0.358	6.7 e-16

Pour une durée de temps de traitement inférieure à 206 jours, le taux de risque (4.1) du groupe 1 (radiothérapie) est : $\lambda(t|z=1) = \lambda_0(t)exp[(-0.338 + 2.885) \times 1]$, et celui du groupe 0 (radiothérapie + C.A.P) est : $\lambda(t|z=0) = \lambda_0(t)exp[(-0.338 + 2.885) \times 0]$.

Donc, le risque relatif de la réapparition de la maladie du cancer est estimé par $e^{(-0.338+2.885)} = 12.76$. Le risque de la récurrence du cancer chez le groupe 1 est douze fois plus grande que chez le groupe 0, traité par la radiothérapie plus les doses de C.A.P.

Les taux de risque des deux groupes correspondant à des durées de traitement dépassant 206 jours, sont :
groupe 1 : $\lambda(t|z=1) = \lambda_0(t)exp[(-0.338 + 0) \times 1]$
groupe 0 : $\lambda(t|z=0) = \lambda_0(t)exp[(-0.338 + 0) \times 0]$.

Alors, $e^{-0.338} = 0.72$ est le risque relatif de la récurrence des tumeurs. Le groupe de patients traités par la radiothérapie seule (groupe 1) semble maintenant montrer moins de risque de réapparition de la maladie. Cependant, ce risque relatif reste proche de la valeur 1, avec une p-valeur non significative (0.16), et donc, les deux traitements ont presque le même effet sur la maladie pour des durées de traitement assez longues.

On conclut donc, que le risque relatif entre les deux groupes qui ont suivi des traitements différents, depend fortement du point de rupture $\tau = 206$; le risque relatif change sa valeur de $e^{(-0.338+2.885)}$ avant le temps de rupture 206, à $e^{-0.338}$ après ce temps là.

Le traitement mesuré par la covariable Z peut donc être un facteur de risque de la récurrence du cancer du poumon pour un temps précoce (avant 206 jours), mais ne l'est pas pour un temps tardif (après 206 jours) .
On peut dire aussi que la dose de C.A.P ajoutée à la radiothérapie, a pu retarder la réapparition de la maladie pendant les premiers 206 jours du traitement, par rapport à la radiothérapie seule.

Il est intéréssant à la fin de cette analyse d'examiner comment l'inférence sur θ dépend de l'estimateur du point de rupture, $\hat{\tau}$. La figure (4.3) montre la région de confiance (à 95%) pour θ et τ. Parmi toutes les valeurs possibles de τ dans $[109, 966]$, les intervalles de confiance (basés sur l'approximation normale de $\hat{\theta}$) correspondant à θ semblent toujours exclure la valeur zéro, ce qui confirme le rejet de l'hypothèse nulle $H_0 : \theta = 0$ « sans point de rupture ».

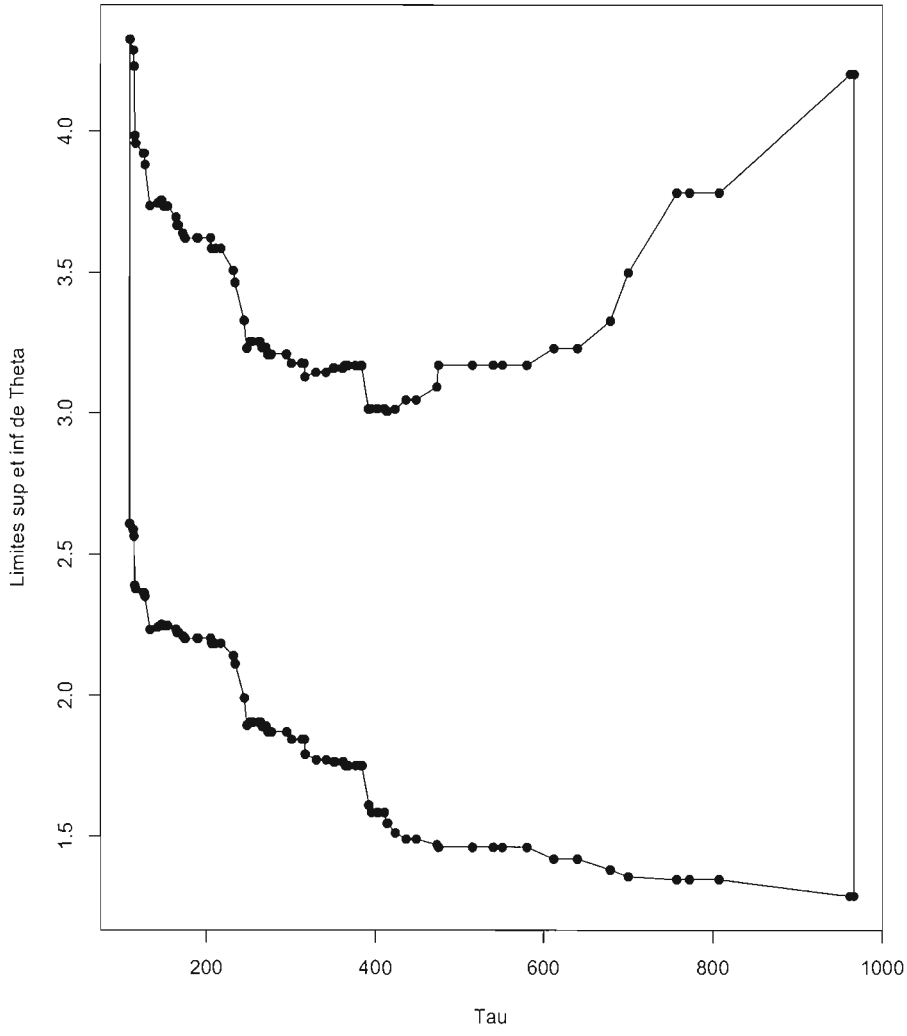

Figure 4.3 Région de confiance à 95% de (τ, θ)

CONCLUSION

Une des raisons qui expliquent la motivation de certains chercheurs à étudier le modèle de risque instantané (de Cox, constant) avec un point de rupture est qu'il est simple à décrire, à développer, et à comprendre. Il est intéressant d'identifier des modèles de survie simples expliquant le changement de risque chez une population de patients après l'essai d'un nouveau traitement.

Dans une partie de ce mémoire, on a conduit une étude théorique (MacGibbon et Groshen (2003), Matthews *et al.* (1985)) menée par des simulations pour étudier les propriétés des statistiques de test standard, *i.e.* la statistique du test des rapports de vraisemblances et la statistique du score (normalisée et partiel), dans le but de tester l'existence d'un point de rupture dans un modèle de risque instantané constant.
Les résultats de simulation ont montré que le comportement de la statistique du test des rapports de vraisemblances dépend fortement de l'intervalle qui recouvre les valeurs du point de rupture τ, et qu'un intervalle d'événements plus restreint couvrant τ implique une convergence de la statistique du test des rapports de vraisemblances.
L'étude théorique de Matthews *et al.* (1985) discutée dans la troisième section, a mis le point sur la convergence des deux statistiques de score, partielle et normalisée, et nous a indiqué une alternative à la statistique du test des rapports de vraisemblances (non bornée dans le cas où la dernière observation d'un échantillon est non censurée) pour tester l'hypothèse $H_0 : \tau = 0$ « sans point de rupture ».

Dans la deuxième moitié de ce travail, on a discuté du modèle des risques proportionnels avec un point de rupture de Liang *et al.* (1990). Au lieu d'utiliser d'autres alternatives pour modéliser les données de survie quand les tests d'ajustement indiquent que le modéle de Cox n'est pas adéquat, le modéle des risques proportionnels avec un point de rupture peut être utilisé, le point de rupture correspondant au point qui maximise

la fonction de vraisemblance.

Le modèle de Cox avec un point de rupture a des avantages et des inconvénients. L'un des avantages est qu'il est plus simple d'introduire un paramètre de plus que d'introduire par exemple la covariable z^2, de plus il est aisé d'interpréter ses résultats (le risque relatif change sa valeur de $e^{(\beta+\theta)}$ à e^{β} pour un temps de survie dépassant τ, avec une augmentation d'une unité de la covariable z). Un autre avantage est que l'alternative du modèle avec un point de rupture reste dans le cadre du modèle de régression de Cox, et que les développements et les calculs sont simples à effectuer avec le modèle de risques proportionnels avec un point de rupture.

Enfin, il faut ajouter que même si Matthews *et al.* (1985)) ont démontré la convergence de la statistique de score pour le modèle de risque instantané constant avec un point de rupture (3.1), il faut étudier les propriétés de convergence de la statistique $M = Sup_{\tau\in[a,b]}|S(\tau)|$, qui dépend de la statistique de score, pour le modèle de Cox avec un point de rupture (4.1).

Pour des échantillons de temps de survie (non censurés) de taille $10, 20, .., 100, 200$, on a généré 10000 réplications à partir de la distribution correspondante au modèle de Cox, tout en calculant la valeur de M.

Dans le tableau (4.3), les résultats de simulation montrent une légère croissance des quantiles 90, 95 et 99 de la distribution de M (les résultats ont été obtenus en utilisant notre programme de simulation, voir annexe 2). On se demande si l'ajout de certain contraintes sur la statistique M, comme on l'a fait pour la statistique du test des rapports de vraisemblances dans la section (2.3.1), pourra améliorer la convergence de M. Cette question sera le sujet de nos prochaines recherches.

Tableau 4.3 Quantiles de la statistique M, définie dans (4.3)

Taille de	Quantiles		
l'échantillon	90	95	99
10	2.189	2.419	3.005
20	2.413	2.679	3.339
30	2.494	2.757	3.301
40	2.583	2.836	3.380
50	2.635	2.891	3.401
60	2.704	2.957	3.489
70	2.746	2.980	3.419
80	2.812	3.039	3.527
90	2.834	3.065	3.485
100	2.884	3.111	3.552
200	3.264	3.521	3.958

APPENDICE A

PREUVES

A.1 Démonstration de $E[Z_n(\tau)] = 0$.

Sous l'hypothèse nulle $H_0 : \epsilon = 0$, sans point de rupture, on donne la preuve de $E[Z_n(\tau)] = 0$, où $Z_n(\tau) = n^{-\frac{1}{2}} \sum_{i=1}^{n} e^{\frac{\lambda\tau}{2}} (\lambda T_i - \lambda\tau - 1) H(T_i - \tau)$ est le processus défini par (3.3).

Notons que les variables aléatoires $T_1, T_2, ..., T_n$ sont indépendantes et identiquement distribuées selon la densité exponentielle de moyenne $1/\lambda$ sous l'hypothèse $H_0 : \epsilon = 0$.

$$
\begin{aligned}
E[Z_n(\tau)] &= E\left[n^{-\frac{1}{2}} \sum_{i=1}^{n} e^{\frac{\lambda\tau}{2}} (\lambda T_i - \lambda\tau - 1) H(T_i - \tau) \right] \\
&= n^{-\frac{1}{2}} e^{\frac{\lambda\tau}{2}} \sum_{i=1}^{n} E[(\lambda(T_i - \tau) - 1) H(T_i - \tau)] \\
&= n^{-\frac{1}{2}} e^{\frac{\lambda\tau}{2}} \sum_{i=1}^{n} \int_{\tau}^{\infty} [\lambda (T_i - \tau) - 1] \lambda e^{-\lambda T_i} \, dT_i \\
&= n^{-\frac{1}{2}} e^{\frac{\lambda\tau}{2}} \sum_{i=1}^{n} \left[(\lambda\tau - 1) \int_{\tau}^{\infty} -\lambda e^{-\lambda T_i} \, dT_i - \lambda \int_{\tau}^{\infty} T_i \, (-\lambda e^{-\lambda T_i}) \, dT_i \right] \\
&= n^{-\frac{1}{2}} e^{\frac{\lambda\tau}{2}} \sum_{i=1}^{n} \left[-(\lambda\tau - 1) e^{-\lambda\tau} + (\lambda\tau - 1) e^{-\lambda\tau} \right] = 0.
\end{aligned}
$$

A.2 Preuve de $Var[Z_n(\tau)] = 1$.

Toujours sous l'hypothèse nulle $H_0 : \epsilon = 0$, sans point de rupture, on donne la preuve de $Var[Z_n(\tau)] = 1$.

Sachant que les variables aléatoires $T_1, T_2, ..., T_n$ sont indépendantes, alors :

$$\begin{aligned} Var[Z_n(\tau)] &= Var\left[n^{-\frac{1}{2}} \sum_{i=1}^{n} e^{\frac{\lambda\tau}{2}} (\lambda T_i - \lambda\tau - 1) H(T_i - \tau)\right] \\ &= \frac{e^{\lambda\tau}}{n} \sum_{i=1}^{n} Var[(\lambda T_i - \lambda\tau - 1) H(T_i - \tau)] \\ &= \frac{e^{\lambda\tau}}{n} \sum_{i=1}^{n} \left\{ E[(\lambda T_i - \lambda\tau - 1)^2 H(T_i - \tau)^2] - E[(\lambda T_i - \lambda\tau - 1) H(T_i - \tau)]^2 \right\}. \end{aligned}$$

D'après la preuve de $E[Z_n(\tau)] = 0$, $E[(\lambda T_i - \lambda\tau - 1) H(T_i - \tau)] = 0$, donc :

$$\begin{aligned} Var[Z_n(\tau)] &= \frac{e^{\lambda\tau}}{n} \sum_{i=1}^{n} E[(\lambda T_i - \lambda\tau - 1)^2 H(T_i - \tau)^2] \\ &= \frac{e^{\lambda\tau}}{n} \sum_{i=1}^{n} \int_{\tau}^{\infty} (\lambda T_i - \lambda\tau - 1)^2 \, \lambda \, e^{-\lambda T_i} \, dT_i, (sous\, H_0 : \epsilon = 0) \\ &= \frac{e^{-1}}{n} \sum_{i=1}^{n} \int_{\tau}^{\infty} (\lambda T_i - \lambda\tau - 1)^2 \, \lambda \, e^{-(\lambda T_i - \lambda\tau - 1)} \, dT_i. \end{aligned}$$

avec le changement de variable $u = \lambda T_i - \lambda\tau - 1$, $du = \lambda \, dT_i$ et $-1 \leq u < \infty$, on a :

$$\begin{aligned} Var[Z_n(\tau)] &= \frac{e^{-1}}{n} \sum_{i=1}^{n} \int_{-1}^{\infty} u^2 e^{-u} du \\ &= e^{-1} \int_{-1}^{\infty} u^2 e^{-u} du \\ &= e^{-1} \times e = 1. \end{aligned}$$

(en utilisant une double intégration par parties).

A.3 Preuve de $Cov[Z_n(\tau_1)\,,\, Z_n(\tau_2)] = exp\left[\frac{-\lambda|\tau_1 - \tau_2|}{2}\right]$

Soient g_{τ_1} et g_{τ_1} deux fonctions définies par $g_{\tau_1}(t) = (\lambda t - \lambda\tau_1 - 1) H(T_i - \tau_1)$ et $g_{\tau_2}(t) = (\lambda t - \lambda\tau_2 - 1) H(T_i - \tau_2)$. Pour des variables aléatoires indépendantes $T_1, T_2, ..., T_n$, distribuées selon la fonction de densité $\lambda \, e^{-\lambda t}$ sous $H_0 : \epsilon = 0$, $Cov(g_{\tau_1}(T_i)\,,\, g_{\tau_2}(T_j)) = 0$ si $i \neq j$.

Supposons que $\tau_1 > \tau_2$.

$$\begin{aligned} Cov[Z_n(\tau_1)\,,\, Z_n(\tau_2)] &= Cov\Big[\sum_{i=1}^{n} n^{-\frac{1}{2}}\, e^{\frac{\lambda\tau_1}{2}}\, g_{\tau_1}(T_i)\,,\, \sum_{j=1}^{n} n^{-\frac{1}{2}}\, e^{\frac{\lambda\tau_2}{2}}\, g_{\tau_2}(T_j)\Big] \\ &= \frac{e^{\frac{\lambda}{2}(\tau_1+\tau_2)}}{n}\, Cov\Big[\sum_{i=1}^{n} g_{\tau_1}(T_i)\,,\, \sum_{j=1}^{n} g_{\tau_2}(T_j)\Big] \\ &= \frac{e^{\frac{\lambda}{2}(\tau_1+\tau_2)}}{n} \sum_{i=1}^{n} Cov[g_{\tau_1}(T_i)\,,\, g_{\tau_2}(T_i)]. \end{aligned}$$

D'après la preuve de $E[Z_n(\tau)] = 0$, $E[g_{\tau_j}(T_i)] = E[(\lambda T_i - \lambda\tau_j - 1)H(T_i - \tau_j)] = 0$ pour j = 1, 2. D'autre part, $H(T_i - \tau_1)\, H(T_i - \tau_2) = H(T_i - \tau_1)$ si $\tau_1 > \tau_2$, donc :

$$\begin{aligned} Cov[Z_n(\tau_1)\,,\, Z_n(\tau_2)] &= \frac{e^{\frac{\lambda}{2}(\tau_1+\tau_2)}}{n} \sum_{i=1}^{n} E[g_{\tau_1}(T_i)\, g_{\tau_2}(T_i)] \\ &= e^{\frac{\lambda}{2}(\tau_1+\tau_2)} \int_{\tau_1}^{\infty} (\lambda t - \lambda\tau_1 - 1)(\lambda t - \lambda\tau_2 - 1)\, \lambda\, e^{-\lambda t}\, dt, \\ &= e^{\frac{\lambda}{2}(\tau_1+\tau_2)} \Big[-\lambda^2 \int_{\tau_1}^{\infty} t^2\, (-\lambda e^{-\lambda t})\, dt \;+\; \lambda(\lambda\tau_1 \;+\; \lambda\tau_2 \;+\; 2) \\ &\quad \int_{\tau_1}^{\infty} t\, (-\lambda e^{-\lambda t})\, dt \;-\; (\lambda\tau_1 + 1)(\lambda\tau_2 + 1) \int_{\tau_1}^{\infty} -\lambda e^{-\lambda t}\, dt \Big]. \end{aligned}$$

en utilisant l'intégration par parties, on trouve :

$$\begin{aligned} Cov[Z_n(\tau_1)\,,\, Z_n(\tau_2)] &= e^{\frac{\lambda}{2}(\tau_1+\tau_2)} \Big[(\lambda^2\tau_1^2 \;+\; 2\lambda\tau_1 \;+\; 2)\, e^{-\lambda\tau_1} \;+\; (\lambda\tau_1 \;+\; \lambda\tau_2 \;+\; 2) \\ &\quad (-\lambda\tau_1 \;-\; 1)\, e^{-\lambda\tau_1} \;+\; (\lambda\tau_1 \;+\; 1)(\lambda\tau_2 \;+\; 1)\, e^{-\lambda\tau_1} \Big] \\ &= exp\Big[\frac{-\lambda(\tau_1 - \tau_2)}{2}\Big]. \end{aligned}$$

Pour le cas $\tau_1 < \tau_2$, on trouve de façon analogue $Cov[Z_n(\tau_1)\,,\, Z_n(\tau_2)] = exp\Big[\frac{-\lambda(\tau_2-\tau_1)}{2}\Big]$, et par conséquent :

$$Cov[Z_n(\tau_1)\,,\, Z_n(\tau_2)] = exp\Big[\frac{-\lambda|\tau_1 - \tau_2|}{2}\Big].$$

APPENDICE B

PROGRAMMES

B.1 Simulation des données non censurées sans restriction

Ce programme en R simule les valeurs de la statistique du rapport de vraisemblance calculées à partir d'un échantillon de données de survie de taille n.
La fonction *simulation*(n) de variable n retourne la valeur simulée de la statistique du rapport de vraisemblance.

```
simulation ←function(n){
        statistic ← 0
        # les temps de survie simulés selon la loi exponentielle(λ = 1)
        t ← log(1 - runif(n))
        # z, vecteur des temps de survie dans l'ordre croissant
        z ← sort(t)
        # calculer la valeur de U_i définie dans la remarque (2.3)
        U ← rep(NA , n)
        for(i in 1 : n){
                U[i] ← 0
                for(j in 1 : n){
                        U[i] ← U[i] + min(t[j], z[i])/ sum(t)
                        }
                }
```

```
# vecteur des statistiques Δ_i^+ définies dans la remarque (2.3)
Lp ← rep(NA , n-1)
# vecteur des statistiques Δ_i^- définies dans la remarque (2.3)
Lm ← rep(NA , n-1)

# calculer les valeurs de Lp[i] = Δ_i^+ et Lm[i] = Δ_i^-
for(i in 1 : n-1){
    if(i == 1){
      Lm[i] ← 2 * ( (n-i+1) * log( (n-i+1)/(1-U[i]) ) - n * log(n) )
      Lp[i] ← 2 * ( i * log(i/U[i]) + (n-i) * log( (n-i)/(1-U[i]) ) - n * log(n) )
      }
    else {
      Lm[i] ← 2 * ( (i-1) * log((i-1)/U[i]) + (n-i+1) * log( (n-i+1)/(1-U[i]) )
                - n * log(n) )
      Lp[i] ← 2 * ( i * log(i/U[i]) + (n-i) * log( (n-i)/(1-U[i]) ) - n * log(n) )
      }
    }

# la statistique du rapport de vraisemblance Δ, maximum de Lm et Lp,
# définie dans la remarque (2.3)
statistic ← max( Lp , Lm)
# retourner la valeur de la statistique du rapport de vraisemblance
statistic
}
# Fin de la fonction simulation(n)
```

```
# Tailles différentes de l'échantillon des temps de survie
sample ← c(10,20,30,40,50,60,70,80,90,100,200)
# Matrice des quantiles 90, 95 et 99 pour chaque échantillon
Quantiles ← matrix(NA , nrow=11 , ncol=3)

k ← 1
for(n in sample){
        statistics ← rep(NA , 10000)
        # simuler 10000 valeurs de la statistique du rapport de vraisemblance
        # à partir d'un échantillon de n temps de survie
        for(i in 1 : 10000){
                        statistics[i] ← simulation(n)
                        }
        # extraire les quantiles 90, 95, 99 des 10000 valeurs simulées
        Quantiles[k,] ← quantile(statistics , probs = c(90,95,99)/100)
        k ← k + 1
        }
```

B.2 Simulation des données non censurées avec la restriction : p-quantile $< \tau <$ (1-p)-quantile

Ici, le programme simule les valeurs de la statistique du rapport de vraisemblance à partir d'un échantillon de données de survie de taille n, avec la restriction p-quantile $< \tau <$ (1-p)-quantile et τ le point de rupture.
La fonction *simulation*(n , p) retourne la valeur simulée de la statistique du rapport de vraisemblance.

```
# n est la taille de l'échantillon
# p est la proportion de restriction
```

```
simulation ←function(n , p){
        statistic ← 0
        # les temps de survie simulés selon la loi exponentielle(λ = 1)
        t ← log(1 - runif(n))
        # z, vecteur des temps de survie dans l'ordre croissant
        z ← sort(t)

        U ← rep(NA , n)
        # calculer la valeur de U_i définie dans la remarque (2.3)
        for(i in 1 : n){
                U[i] ← 0
                for(j in 1 : n){
                        U[i] ← U[i] + min(t[j], z[i])/ sum(t)
                        }
                }
        # p-quantile de l'échantillon
        k ← round(p * n)
        # vecteur restreint des statistiques Δ_i^+ définies dans la remarque (2.3)
        Lp ← rep(NA , n- 2*k)
        # vecteur restreint des statistiques Δ_i^- définies dans la remarque (2.3)
        Lm ← rep(NA , n- 2*k)
        # calculer les valeurs de Lp[τ] et Lm[τ] pour τ = z_{k+1}^-, z_{k+1}^+, ..., z_{n-k}^-, z_{n+k}^+
        j ← 1
        for(i in (k+1) : (n-k)){
            Lm[i] ← 2 * ( (i-1) * log((i-1)/U[i]) + (n-i+1) * log( (n-i+1)/(1-U[i]) )
                    - n * log(n) )
            Lp[i] ← 2 * ( i * log(i/U[i]) + (n-i) * log( (n-i)/(1-U[i]) ) - n * log(n) )
            j ← j + 1
            }
```

```
        # la statistique du rapport de vraisemblance Δ, maximum de Lm et Lp,
        # définic dans la remarque (2.3)
        statistic ← max( Lp , Lm)
        # retourner la valeur de la statistique du rapport de vraisemblance
        statistic
        }
# Fin de la fonction simulation(n)

# Tailles différentes de l'échantillon des temps de survie
sample ← c(10,20,30,40,50,60,70,80,90,100,200)
# Matrice des quantiles 90, 95 et 99 pour chaque échantillon
Quantiles ← matrix(NA , nrow=11 , ncol=3)

k ← 1
for(n in sample){
        # simuler 10000 valeurs de la statistique du rapport de vraisemblance
        # à partir d'un échantillon de n temps de survie
        statistics ← rep(NA , 10000)
        for(i in 1 : 10000){
                        statistics[i] ← simulation(n , 0.1)
                        }
        # extraire les quantiles 90, 95, 99 des 10000 valeurs simulées
        Quantiles[k,] ← quantile(statistics , probs = c(90,95,99)/100)
        k ← k + 1
        }
```

B.3 Région de confiance de $(\tau\,,\,\theta)$

Ce programme en R trace la région de confiance de $(\tau\,,\,\theta)$ dont le graphique se trouve à la fin du chapitre 4.

```
# Vecteur des estimations du paramètre theta correspondant aux valeurs possibles de tau
theta.vect ← rep(NA , 78)
# Vecteur des écarts-types des estimations de theta
sd.vect ← rep(NA , 78)
# tau.vect, ensemble des valeurs de tau, point de rupture
i ← 1
for(tau in tau.vect){
                    # treat est le facteur de risque (traitement)
                    # Z0, facteur de risque relié aux temps de survie inférieurs à tau
                    Z0 ← treat * I(temps.survie ≤ tau)

                    # temps.survie est le vecteur des temps de récurrence du cancer
                    # statue est le vecteur indicateur de censure
                    cox.surv ← coxph( Surv(temps.survie , statue) ~ Z0 + treat)

                    # theta.vect[i] est l'estimation de theta correspondante à la valeur de tau
                    theta.vect[i] ← coef(cox.surv)[1]
                    # sd.vect[i], l'écart-type de theta.vect[i]
                    sd.vect[i] ← sqrt(cox.surv$var[1,1])
                    i ← i + 1
                    }
# Vecteur des bornes supérieures θ̂ + 1.96 × σ̂_θ̂ de l'intervalle de confiance de θ̂
borne.sup ← theta.vect + 1.96 * sd.vect
# Vecteur des bornes inférieures θ̂ − 1.96 × σ̂_θ̂ de l'intervalle de confiance de θ̂
borne.inf ← theta.vect - 1.96 * sd.vect
```

```
# x, vecteur des première et dernière valeurs possibles de tau
# y, vecteur composé de la borne supérieure et la borne inférieure des intervalles de confiance
# de theta correspondant (resp) à tau.vect[1] et tau.vect[78]
x ← c(tau.vect[1] , tau.vect[78])
y ← c(borne.sup[1] , borne.inf[78])
plot(y ~ x , xlab = "Tau" , ylab = "Limites sup et inf de Theta")

# Graphique des points (tau , borne.sup) et (tau , borne.inf)
for(i in 1 : 78){
                points(tau.vect[i] , borne.inf[i] , type = "p" , pch =19)
                points(tau.vect[i] , borne.sup[i] , type = "p" , pch =19)
                }

# Tracer les segments reliant (tau , borne.sup) et (tau , borne.inf)
for(i in 1 : 77){
                segments(tau.vect[i] , borne.inf[i] , tau.vect[i+1] , borne.inf[i+1])
                segments(tau.vect[i] , borne.sup[i] , tau.vect[i+1] , borne.sup[i+1])
                }
segments(tau.vect[1] , borne.inf[1] , tau.vect[1] , borne.sup[1])
segments(tau.vect[78] , borne.inf[78] , tau.vect[78] , borne.sup[78])
```

B.4 Simulation de la statistique M

B.4.1 Formule de la statistique $S(\tau)$

Soient $t_1, t_2, ..., t_n$ un échantillon de n temps de survie. La fonction de vraisemblance partielle correspondant au modèle de Cox avec un point de rupture (4.1), est donnée par :

$$L(\beta, \theta, \tau) = \prod_{i=1}^{n} \left\{ \frac{exp[(\beta + \theta\, I(t_i \leq \tau))\, Z_i]}{\sum_{j=1}^{n} exp[(\beta + \theta\, I(t_i \leq \tau))\, Z_j]\, I(t_j \geq t_i)} \right\} \qquad \text{(B.1)}$$

La fonction de vraisemblance (B.1) est équivalente à celle définie dans (4.2), pour des temps de survie non censurés, et dans le cas où $\gamma = 0$.

La statistique $S(\tau)$ (définie dans (4.4)) est donnée par :

$$S(\tau) = \left(\frac{d\,log\,L}{d\,\theta}\right)\left[-\frac{d^2log\,L}{d\,\theta^2} - \left(-\frac{d^2log\,L}{d\,\theta\,d\beta}\right)\left(-\frac{d^2log\,L}{d\,\beta^2}\right)^{-1}\left(-\frac{d^2log\,L}{d\,\theta\,d\beta}\right)\right]^{-\frac{1}{2}}_{\beta=\hat{\beta},\,\theta=0} \tag{B.2}$$

(où $\gamma = \beta$, et $\hat{\beta}$ est l'estimateur de maximum de vraisemblance de β).

Avec

$$\left.\frac{d\,log\,L}{d\,\theta}\right|_{\beta=\hat{\beta},\,\theta=0} = \sum_{i:t_i\leq\tau}\left[z_i - \frac{S_1(t_i,\hat{\beta})}{S_0(t_i,\hat{\beta})}\right] \tag{B.3}$$

$$\left.-\frac{d^2log\,L}{d\,\theta^2}\right|_{\beta=\hat{\beta},\,\theta=0} = \sum_{i:t_i\leq\tau}\left[\frac{S_2(t_i,\hat{\beta})}{S_0(t_i,\hat{\beta})} - \frac{S_1(t_i,\hat{\beta})^2}{S_0(t_i,\hat{\beta})^2}\right] \tag{B.4}$$

$$\left.-\frac{d^2log\,L}{d\,\beta^2}\right|_{\beta=\hat{\beta},\,\theta=0} = \sum_{i=1}^{n}\left[\frac{S_2(t_i,\hat{\beta})}{S_0(t_i,\hat{\beta})} - \frac{S_1(t_i,\hat{\beta})^2}{S_0(t_i,\hat{\beta})^2}\right] \tag{B.5}$$

$$\left.-\frac{d^2log\,L}{d\,\theta\,d\beta}\right|_{\beta=\hat{\beta},\,\theta=0} = \sum_{i:t_i\leq\tau}\left[\frac{S_2(t_i,\hat{\beta})}{S_0(t_i,\hat{\beta})} - \frac{S_1(t_i,\hat{\beta})^2}{S_0(t_i,\hat{\beta})^2}\right] \tag{B.6}$$

et où $S_k(t_i,\,\hat{\beta}) = \sum_{i=1}^{n} z_j^k\; exp(\hat{\beta}\,z_j)\,I(t_j \geq t_i)$.

B.4.2 Programme de simulation de la statistique M

D'après l'equation (B.2), la statistique $S(\tau)$ ne dépend pas de la fonction de risque de base $\lambda_0(t)$ du modèle (4.1). Alors, cette fonction de risque n'a aucun effet sur la distribution de la statistique $M = Sup_{\tau\in[a,b]}|S(\tau)|$, et donc, on peut supposer que $\lambda_0(t)$ est constante (par exemple $\lambda_0(t) = 1$) dans les simulations de M.

Pour des échantillons de temps de survie (non censurés) de taille 10, 20,..., 100, 200, on génére 10000 réplications de la statistique M, sous l'hypothèse H_0 : $\theta = 0$ (où θ le paramètre du modèle (4.1)), et en supposant que les valeurs de la covariable z sont pour moitié égales à 0, et pour l'autre moitié égales à 1.

Les temps de survie sont générés à partir de l'expression suivante :

$$t = -\frac{log(\,1\,-\,u\,)}{exp(z)}$$

où u une suite de nombres aléatoires générés par une uniforme(0,1).

La fonction S calcule le vecteur des valeur de la fonction $S_k(t_i, \hat{\beta})$ correspondant aux valeurs de t_i, définie ci-dessus.

```
S ← function(k, t, z, beta){
        S.vect ← rep(NA, length(t))
        for( i in 1 :length(t)){
                S.vect[i] ← sum((z^ k)*exp(beta*z)* I(t >= t[i]))
                }
        # retourne le vecteur des valeurs de S_k(t_i, β̂).
        S.vect
        }
```

La fonction d.theta calcule le vecteur des valeurs de $d\,log\,L/\,d\,\theta$ correspondant aux valeurs de τ, défini par (B.3).

```
# time, vecteur des temps de survie t, dans l'ordre croissant.
# S_0 et S_1 les valeurs de la fonction S pour k = 0 et k = 1, définie ci-dessus.
d.theta ← function(t, time, z, beta, S_0, S_1){
        vect ← rep(NA, length(t))
        i ← 1
        for(tau in time){
                vect[i] ← sum((z - S_1/S_0)* I(t <= tau))
                i ← i+ 1
                }
        # retourne le vecteur des valeurs de d log L/ d θ.
        vect
        }
```

La fonction d2.theta calcule le vecteur des valeurs de $-d^2 log\, L/\, d\theta^2$ correspondant aux valeurs de τ, défini par (B.4).

time, vecteur des temps de survie t, dans l'ordre croissant.

S_0 et S_1 et S_2 les valeurs de la fonction S pour $k = 0, 1$ et 2, définie ci-dessus.

```
d2.theta ← function(t, time, z, beta, S0, S1, S2){
        vect ← rep(NA, length(t))
        i ← 1
        for(tau in time){
                vect[i] ← sum((S2/S0 - (S1/S0)^ 2)* I(t <= tau))
                i ← i+ 1
                }
        # retourne le vecteur des valeurs de d^2 log L/ d theta^2.
        vect
        }
```

La fonction dtheta.dbeta calcule le vecteur des valeurs de $-d^2 log\, L/\, d\theta\, d\beta$ correspondant aux valeurs de τ, défini par (B.4).

time, vecteur des temps de survie t, dans l'ordre croissant.

S_0 et S_1 et S_2 les valeurs de la fonction S pour $k = 0, 1$ et 2, définie ci-dessus.

```
dtheta.dbeta ← function(t, time, z, beta, S0, S1, S2){
        vect ← rep(NA, length(t))
        i ← 1
        for(tau in time){
                vect[i] ← sum((S2/S0 - (S1/S0)^ 2)* I(t <= tau))
                i ← i+ 1
                }
        # retourne le vecteur des valeurs de d^2 log L/ d theta^2.
        vect
        }
```

Ce programme en R simule les valeurs de la statistique M calculées à partir d'un échantillon de données de survie de taille n.
La fonction *simulation*(n) de variable n retourne la valeur simulée de la statistique M.

```
simulation ← function( n ){
        # statistics, vecteur des valeurs S(τ), qui correspondent aux valeurs de τ.
        statistics ← rep(NA, n)
        # z est le vecteur des covariables de régression.
        z ← c( rep(1, n/2), rep(0, n/2) )
        # status est le vecteur indicateur de censure.
        status ← rep(1, n)
        # t, vecteur des temps de survie.
        t ← - log(1 - runif(n) )/ exp(z)

        reg ← coxph( Surv( t, status)~ z )
        # beta est l'estimateur de maximum de vraisemblance de β du
        # modèle (4.1), sous l'hypothèse H0 : θ = 0.
        beta ← reg$coef
        # time, vecteur des temps de survie t, dans l'ordre croissant.
        time ← sort( t )
        S0 ← S(0, t, z, beta)
        S1 ← S(1, t, z, beta)
        S2 ← S(2, t, z, beta)

        # d2.beta est la valeur de -d²log L / dβ² définie dans (B.5).
        d2.beta ← sum( S2/S0 - (S1/S0)^ 2)
        nomin ← d.theta(t, time, z, beta, S0, S1)
        denom ←d2.theta(t, time, z, beta, S0, S1, S2) -
                (dtheta.dbeta(t, time, z, beta, S0, S1, S2)^ 2)/ d2.beta
        statistics ← abs( nomin / sqrt( denom ) )
```

```
# La statistique M initialisé à -1.
M ← -1
for(i in 1 : n){
   if( is.finite(statistics[i]) && M < statistics[i]){
                                                M ← statistics[i]
                                                }
   }
# retourne la valeur calculée de la statistique M.
M
}
```

RÉFÉRENCES

Aalen, O. O. (1975). « Statistical inference for a family of counting processes ». Thèse de doctorat, University of California, Berkeley.

Barndorff-Nielsen, O.E. and Cox, D.R. (1984). « The effect of sampling rules on likelihood statistics ». *International Staistical Review* , vol. 52, p. 309-326.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Breslow, N.E., Edler, L. ,and Berger, J. (1984). « A two-sample censored-data rank test for acceleration ». *Biometrics* , vol. 40, p. 1049-1062.

Cox, D. R. (1972). « Regression models and life tables (with discussion) ». *Journal of the Royal Statistical Society Series B*, vol. 34, p. 187-220.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London : Chapman and Hall.

Davies, R.B. (1977). « Hypothesis testing when a nuissance parameter is present only under the alternative ». *Biometrika*, vol. 64, p. 247-254.

Fleming, T. R. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Hawkins, D.M. (1977). « Testing a sequence of observations for a shift in location ». *Journal of the American Statistical association*, vol. 72, p.180-186.

Henderson, R. (1990). « A problem with the likelihood ratio test for a change-point hazard rate model ». *Biometrika*, vol. 77, p. 833-843.

Hinkley, D.V. (1970). « Inference about the change-point in a sequence of random variables ». *Biometrika*, vol. 57, p. 1-17.

Hinkley, D.V. and E.A. Hinkley, (1970). « Inference about the change-point in a sequence of binomial random variables ». *Biometrika*, vol. 57, p. 477-488.

Hougaard Philip (1999). « Fundamentals of Survival Data ». *Biometrics* , vol. 55, p. 13-22.

James, B., James, K.L. and Siegmund, D. (1987). « Tests for a change-point ». *Biometrika* , vol. 74, p. 71-83.

Kaplan, E.L. and Meier, P. (1958). « Nonparametric estimator from incomplete observations ». *Journal of the American Statistical Association* , vol. 53, p. 457-481.

Klein, J.P. et Moeschberger, M.L. (2003). *Survival Analysis, Techniques for Censored and Truncated Data, second edition*. Springer-Varlag, New York.

Lad, T., Rubinstein, L., Sadeghi, A. et al. (1988). « The benefit of adjuvant treatment for resected locally advanced non-small cell lung cancer ». *Journal of Clinical Oncology*, vol. 6, p. 9-17.

Lagakos, S.W. and Schoenfeld, D.A. (1984). « Properties of proportional-hazards score tests under misspecified regression models ». *Biometrics* , vol. 40, p. 1037-1048.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Liang, K.Y., S.G. Self and Liu, X. (1990). « The Cox proportional hazards model with change point : an epidemiologic application ». *Biometrics*, vol. 46, p. 783-793.

Lin, D. Y. and Wei, L.J. (1989). « The robust inference for the Cox proportional hazards model ». *Journal of the American Statistical Association*, vol. 84, p. 1074-1078.

Loader, C.R. (1991). « Inference for a hazard rate change point ». *Biometrika* , vol. 78, p. 749-757.

Matthews, D. E. and Farewell, V. T. (1982). « On testing for a constant hazard against a change-point alternative ». *Biometrics*, vol. 38, p. 463-468.

——— (1985). « On a singularity in the likelihood for a change-point hazard rate model ». *Biometrika*, vol. 72, p. 703-704.

Matthews, D. E., V. T. Farewell and R. Pyke (1985). « Asymptotic score-statistic processes and tests for constant hazard against a change-point alternative ». *Annals of Statistics*, vol. 13, p. 583-591.

Mantel, N. (1966). « Evaluation of survival data and two new rank order statistics arising in its consideration ». *Cancer Chemotherapy Report*, vol. 50, p. 163-170.

MacGibbon, B., and Groshen, S. (2003). « Confidence intervals in the exponential hazard rate model with censoring ». International Conference on Recent Advances in Survey Sampling, A celebration of the work of J.N.K Rao. p. 147-160.

Mandl, P. (1962). « On the distribution of the time which the Uhlenbeck process requires to exceed a boundary (in Czechoslovakian with Russian and German summaries) ». *Applied Mathematics*, vol. 7, p. 141-148.

Nguyen, H.T., G.S. Rogers, and E.A. Walker, (1984). « Estimation in change-point hazard rate models ». *Biometrika* , vol. 71, p. 299-304.

O'Reilly, N.E. (1974). « On the weak convergence of empirical processes in sup-norm metrics », *Annals of Probability*, vol. 2, p. 642-651.

Piantadosi Steven (1997). *Clinical Trials : a Methodologic Perspective.* Wiley series in probability and statistics, New York.

Prentice, R.L. (1978). « Linear rank tests with right censored data ». *Biometrika*, vol. 65, p. 167-179 .

Prentice, R.L. and Self, S.G. (1983). « Asymptotic distribution theory for Cox-type regression models with general relative risk form ». *Annals of Statistics*, vol. 11, p. 804-813.

Pham, T.D. and H.T. Nguyen (1990). « Strong consistency of the maximum likelihood estimators in the change-point hazard model ». *Statistics*, vol. 21, p. 203-216.

Pyke, R. (1972). « Empirical Processes ». In Jeffery-Williams Lectures : 1968-1972. Canadian Mathematical Congress, Montreal.

Pyke, R. and Shorack (1968). « Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems ». *Annals of Mathematical statistics* , vol. 39, p. 755-771.

Struthers, C.A. and Kalbfleisch, J.D. (1986). « Misspecified proportional hazard models ». *Biometrika* , vol. 73, p. 363-369.

Therneau, T.M. and Grambsch, P.M. (2000). *Modeling Survival Data : Extending the Cox Model.* springer-Verlag, New York.

Worsley, K.J. (1986). « Confidence region and tests for a change-point in a sequence of exponential family random variables ». *Biometrika*, vol. 73, p. 91-104.

Worsley, K.J. (1988). « Exact percentage points of the likelihood ratio test for a change-point hazard rate model ». *Biometrics*, vol. 44, p. 259-263.

Yao, Y.C. (1986). « Maximum likelihood estimation in hazard rate models with change-point ». *Biometrika*, vol. 44, p. 259-266.