N^0 d'ordre :24/2012 - M/M.T

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DES SCIENCES ET TECHNOLOGIE DE HOUARI BOUMEDIENNE
FACULTÉ DES MATHÉMATIQUES

MÉMOIRE
PRÉSENTÉ POUR L'OBTENTION DU DIPLÔME DE MAGISTER.

EN MATHÉMATIQUES
OPTION : PROBABILITÉ & STATISTIQUE

Par : Melle. SELLAMI Loundja

Analyse actuarielle de la durée de vie d'un contrat d'assurance

Soutenu publiquement, le 21/01/2012, devant le jury composé de :

Mr. Mustapha MOULAI	Professeur	à l'USTHB	Président.
Mr. Kamal BOUKHETALA	Professeur	à l'USTHB	Directeur de mémoire.
Mr. Rachid OUAFI	Maître de Conférences	à l'USTHB	Examinateur.
Mr. Abdelkader TATACHAK	Maître de Conférences	à l'USTHB	Examinateur.

Remerciements

Je tiens à exprimer ma profonde reconnaissance à mon directeur de thèse pour m'avoir témoigné de sa confiance en me proposant ce sujet et à le remercier pour ses orientations scientifiques judicieuses qu'il m'a accordées pour l'élaboration de ce mémoire.

Mes respects aux membres du jury qui me feront honneur d'apporter des critiques et appréciations sur mon travail.

Dédicaces

Je dédie ce mémoire, particulièrement, à deux personnes qui me sont très chères, et que leur présence dans ma vie m'a donné du courage et la volonté de réussir dans mes études :

Ma mère, l'émeraude de ma vie.

Et

Ma tante, la prune de mes yeux.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Généralement, l'analyse de survie est une collection de procédures statistiques pour l'analyse de données, pour laquelle la variable d'intérêt est le temps jusqu'à l'occurrence d'un certain événement. Cet événement peut être, par exemple, le décès d'un individu, la rechute d'un traitement, le changement de résidence, le retour au travail, le divorce, la panne d'une pièce ... etc.

Dans ce cas, on se réfère à la variable temps comme temps de survie parce qu'elle donne le temps qu'un individu survive durant une certaine période. On se réfère également à l'événement comme échec, parce que l'événement d'intérêt est habituellement, soit un décès, une incidence d'une maladie ou d'autres expériences individuelles négatives. Cependant, le temps de survie peut être le temps de retour au travail, dans lequel le cas d'échec est un événement positif.

Les données de survie sont très fréquemment incomplètes, que ce soit censures ou troncatures ; ce qui complique l'analyse de ces données. Par conséquent, nous avons recours à des modèles plus appropriées à ce type de problème, dites modèles de survie.

L'objectif de ce mémoire est d'étudier le phénomène de résiliation d'un contrat d'assurance automobile associé à un cout de sinistralité, en utilisant les outils d'inférence statistique de l'analyse de survie.Ce mémoire est articulé en cinq volets :

Chapitre1 : Le contrat d'assurance.
Chapitre2 : Concept de base d'analyse de survie.
Chapitre3 : Méthodes non-paramétriques.
Chapitre4 : Modèles de régression en analyse de survie.
Chapitre5 : Application de l'analyse de survie à l'assurance non-vie.

1

Le contrat d'assurance

Sommaire

1.1 Introduction

L'assurance est un service qui consiste à fournir une prestation prédéfinie, généralement financière, à un individu, une association ou une entreprise lors de la survenance d'un risque, en échange de la perception d'une cotisation ou prime. Par extension, l'assurance est le secteur économique qui regroupe les activités de conception, de production et commercialisation de ce type de service.

Au sens étroit, l'assurance est une technique qui repose sur un contrat, dit contrat d'assurance, en vertu duquel un assureur s'engage, en contre partie du versement régulier d'une prime, d'indemniser les sinistres dont l'assuré est victime.

Le contrat d'assurance n'a plus aujourd'hui une simple dimension individuelle même s'il demeure un contrat synallagmatique. Compte tenu du développement du secteur de l'assurance, le contrat d'assurance ne se résume plus à des rapports entre un assureur et un assuré, car il a acquis une dimension collective, par exemple, lorsqu'un même assureur garantit les membres d'un même groupe.

Dans ce chapitre, on tentera de mettre en clarté le concept d'un contrat d'assurance [14] qui est extrêmement répondu en pratique. En premier lieu, afin de se familiariser au vocabulaire d'assurance, on commencera par donner un petit lexique des termes de l'assurance, ainsi que les différentes distinctions existant dans une compagnie d'assurance. En second, on s'intéressera au contrat d'assurance comme étant un contrat d'adhésion, en explorant les points suivants : la signification d'un contrat d'assurance, la situation d'un contrat d'assurance en pratique, ses caractères, ses éléments et sa prise d'effet. On s'intéressera également aux obligations réciproques entre un assuré et un assureur et les conséquences qui en résultent dans le cas où l'un d'entres eux se désengage de ses obligations.

1.2 Lexique des termes de l'assurance

- **Adhésion**

  Signature par un sujet de droit (personne physique ou morale) d'un contrat avec tous les droits et les obligations que cet acte implique. Le terme d'adhésion étant synonyme de souscription (souscription du contrat d'assurance). Le contrat d'assurance constitue en effet un contrat d'adhésion.

- **Assurabilité**

  Pour qu'un risque soit assurable, il doit théoriquement répondre à trois conditions : être aléatoire, mesurable et compensable. Dans la plupart des cas, le caractère aléatoire du risque l'emporte sur les autres.

Ainsi, un risque dont la survenance est certaine ou quasi certaine ne sera généralement pas assuré par un assureur. Exemple : maison très vétuste en zone inondable.

- **Assurance IARD**

  Incendie, Accidents et Risques Divers. Ce terme désigne plus généralement l'ensemble des assurances dommages. Pour les particuliers, les principales assurances IARD sont l'assurance automobile, la multi-risque habitation et la responsabilité civile.

- **Assurance vie**

  On distingue deux types d'assurances vie :

  – *Assurance en cas de vie* : produit de capitalisation permettant de constituer une épargne et prévoyant le versement de cette épargne sous forme de capital ou de rente si l'assuré est en vie au terme du contrat ;

  – *Assurance en cas décès* : contrat prévoyant le versement d'un capital à un bénéficiaire désigné dans le contrat en cas de décès de l'assuré avant le terme du contrat.

- **Assuré**

  Personne garantie par un contrat d'assurance. L'assuré n'est pas toujours le souscripteur du contrat d'assurance, ni celui qui paie la cotisation. L'assuré est également appelé preneur d'assurance :

  – En assurance de responsabilité civile, le responsable à la qualité de l'assuré ;

  – En assurance vie, l'assuré est la personne dont le décès entraîne le versement du capital ou de la rente prévue dans le contrat ;

  – En automobile, l'assuré est le propriétaire du véhicule ainsi que toute personne conduisant le véhicule mais également toute personne transportée à titre gratuit ;

  – En habitation, l'assuré est le bénéficiaire du contrat ainsi que son conjoint. Il peut s'agir également de leurs descendants et ascendants vivant habituellement sous le même toit.

La définition de l'assuré peut varier d'une compagnie à l'autre et dépend de la nature du risque à assurer. Il convient de toujours vérifier dans les conditions générales et particulières du contrat la personne qui a la qualité de l'assuré.

- **Assureur**

  Terme générique désignant tout professionnel de l'assurance habilité à assurer des risques. Il peut s'agir d'un agent général, d'un courtier, d'une compagnie d'assurances, d'une mutuelle d'assurances, d'une société de réassurance, etc.

- **Bonus-Malus**

  Terme commun utilisé pour désigner le système du coefficient de réduction-majoration qui accorde aux assurés des réductions ou des majorations en fonction des sinistres qu'ils causent. C'est un élément essentiel de la tarification de l'assurance automobile : l'application de ce coefficient peut diminuer de moitié la cotisation (0.5 de bonus) ou l'augmenter considérablement jusqu'à trois fois et demi le tarif de base (3.5 de malus). Ce coefficient est une note qui est personnelle à l'assuré, reflétant son historique en tant que conducteur.

- **Déchéance**

  Sanction prise à l'encontre d'un assuré pour manquement aux obligations résultant du contrat d'assurance, en particulier le fait de ne pas avoir déclaré telle ou telle circonstance à l'assureur. La déchéance a des conséquences différentes selon les personnes se trouvant impliquées dans un sinistre.

- **Garantie**

  Au sens étroit, ce terme désigne dans un contrat d'assurance la protection accordée par l'assureur : chaque contrat indique alors les conditions précises permettant de bénéficier de cette prestation ainsi que ses limites. Au sens large, la garantie est la couverture qu'un assuré bénéficie en cas de sinistre dans la mesure où il a versé une prime.

- **Prime d'assurance ou Cotisation**

  La prime désigne la plupart du temps la somme d'argent versée par une personne à une autre à certaines occasions avec toutes les conséquences que cela implique pour les deux parties. La signification qui nous

intéresse est bien entendu celle retenue en droit des assurances.

Il est nécessaire de distinguer plusieurs catégories de primes qui ont vocation à s'imbriquer et qui représentent chacune une fraction de la dépense qui devra être exposée par l'assuré pour obtenir une garantie contre un risque. Pour désigner la prime, plusieurs appellations sont utilisées dans la pratique :

- **La prime simple :** La prime simple, " technique " ou " pure " correspond à la somme d'argent nécessaire à l'assureur pour constituer des réserves financières, elles-mêmes destinées à indemniser les sinistres. Il est usuel de distinguer les primes au sens strict des cotisations d'assurance. La prime est versée aux compagnies d'assurance classiques et la cotisation est versée aux entreprises d'assurances à caractères mutualiste. D'ailleurs, dans ce cas l'assurée porte le nom de " sociétaire " dans la mesure où il est considéré comme membre d'une institution mutualiste.

- **La prime facturée :** Le montant total de la prime que l'assurée doit acquitter résulte de l'addition de plusieurs étages pour diverses raisons, en particulier des raisons financières. L'assureur doit faire face à plusieurs types de dépenses qui ont pour conséquence d'augmenter le montant de la prime de base.

- **Résiliation**

Cessation définitive et anticipée du contrat. La plupart des contrats se renouvellent automatiquement. Ils ne prennent fin que si l'assuré ou la société d'assurances les résilie. Il ne suffit donc pas de cesser de payer la cotisation. Chaque partie doit respecter certains délais et certaines formes pour demander la résiliation. A défaut le contrat continue.

- **Risque**

Fait ou ensemble de faits contre lesquels l'assuré recherche une couverture. Cette dernière dépend des garanties prévues dans le contrat d'assurance. Une distinction est à opérer de ce point de vue entre les assurances " multirisques " qui couvrent plusieurs risques et les assurances " tout risque " qui sont censés couvrir toute espèce d'éventualité. En réalité, l'appellation " tout risque " est trompeuse dans la mesure où certains assureurs se limitent à une liste de garanties (dans le domaine de l'assurance automobile, il s'agit souvent de la responsabilité civile, des dommages au véhicule, de la garantie défense secours et l'individuelle accident).

- **Sinistre**

Le mécanisme de la réparation se met en marche dès lors qu'un sinistre vient de se produire : l'assuré doit le déclarer à l'assureur et ce dernier doit verser une indemnisation à l'assuré. Il est donc indispensable de s'entendre sur la signification de ce terme afin de savoir dans quels cas la technique de l'assurance peut ou non fonctionner.

Le mot sinistre trouve toute sa signification en droit des assurances, ce qui peut sembler curieux car ce droit n'est pas la seule matière à organiser une protection contre des risques. Le droit de la sécurité sociale par exemple organise une couverture contre des évènements comme les accidents du travail, la maladie ou l'invalidité, sans toutefois reprendre le mot sinistre. Il y a là une spécificité du droit des assurances.

On peut tenter de définir le mot sinistre de façon négative, en insistant sur ce qu'il n'est pas. De ce point de vue, il ne doit pas être confondu, en droit des assurances, avec le terme risque : en théorie, un sinistre est la conséquence logique d'un risque précis. La grêle par exemple n'est pas sinistre en elle-même. Elle ne représente qu'un risque contre lequel une assurance est possible. Les dégâts causés par la grêle (à des récoltes ...) constituent le sinistre qu'il faudra indemniser.

On peut tenter de définir positivement le terme de sinistre. Ce mot s'applique tout simplement à un évènement défavorable.

1.3 Qu'est ce qu'un contrat d'assurance

Le contrat d'assurance peut être défini comme un accord de volonté qui intervient entre un assureur et un assuré aux termes duquel le premier s'engage à garantir un ou plusieurs risques moyennant une prime ou cotisation payée par le second. Ce contrat est régi par le code civil. On peut distinguer selon la cotisation trois types de contrats :

- **Contrat à cotisations périodiques**

  Contrat pour lequel sont prévues plusieurs cotisations, dont le montant et la périodicité sont fixés au moment de la souscription. Le capital prévu au terme du contrat (hors participation aux bénéfices) est connu lors de la souscription.

- **Contrat à cotisation unique**

  Contrat pour lequel est prévue une seule cotisation, versée dans son intégralité au moment de la souscription. Le capital prévu au terme du contrat (hors participation aux bénéfices) est connu lors de la

souscription.

- **Contrat à versements libres**

  Contrat pour lequel il est possible d'effectuer plusieurs versements, dont le montant et la périodicité ne sont pas fixés au moment de la souscription. Le capital versé au terme du contrat sera égal au montant du capital constitué au cours de la vie du contrat, valorisé de l'intérêt technique et de la participation aux bénéfices.

Article $n^o 7$ de l'ordonnance $95-07$:

" Le contrat d'assurance est écrit. Il est rédigé en caractères apparents. Il doit contenir obligatoirement, outre les signatures des parties, les mentions ci-après :

- les noms et domiciles des parties contractantes ;
- la chose ou la personne assurée ;
- la nature des risques garantie ;
- la date de la souscription ;
- la date d'effet et la durée du contrat ;
- le montant de la garantie ;
- le montant de la prime ou cotisation d'assurance."

1.4 Les caractères d'un contrat d'assurance

Le contrat d'assurance est à :

1.4.1 Caractère consensuel

Il est réputé conclu dés le moment où intervient l'accord des parties (assureur-assuré).

1.4.2 Caractère aléatoire

Ce caractère est inhérent à la nature même de l'assurance et la définition du risque.

- **Le sinistre ne doit pas être réalisé lors de la souscription du contrat :** L'assureur ne répond que des "cas fortuit" ou des conséquences dommageables de la faute simple de l'assuré, à condition que celle-ci ne soit pas volontaire.

Le caractère aléatoire du contrat d'assurance s'oppose à ce qu'un assureur prenne en charge un sinistre que l'assuré savait déjà réalisé au moment de la souscription du contrat : notion de passé inconnu.

– **Le sinistre doit dépendre d'un cas fortuit :** Toutefois, l'assureur ne répond pas des pertes et des dommages provenant d'une faute intentionnelle ou dolosive de l'assuré (Absence d'aléa).

1.4.3 Caractère synallagmatique

Ce contrat met à la charge des parties (assureur-assuré), des obligations nécessairement réciproques :

a. Pour l'assureur

L'obligation de l'assureur consiste en l'exécution d'une prestation en cas de réalisation du risque assuré, laquelle peut prendre plusieurs formes telles le Paiement d'une indemnité ou d'un capital, l'organisation de la défense de son assuré (garantie défense et recours, protection juridique), la prestation d'assistance... etc.

b. Pour l'assuré

– **Obligation de déclaration du risque, ou de son aggravation :** Dans les contrats d'assurance de dommage, par exemple, l'assuré doit déclarer, en cours de contrat, les circonstances nouvelles qui ont pour conséquence soit d'aggraver les risques, soit d'en créer de nouveaux et rendent de ce fait inexactes ou caduques les réponses faites à l'assureur.

– **Obligation de payer la prime d'assurance :** Le contrat d'assurance met nécessairement à la charge de l'assuré le paiement d'une prime ou cotisation, proportionnée à l'importance et à la probabilité de réalisation du sinistre, aux époques convenues.

– **Obligation de respecter les conditions de garantie.**

– **Obligation de prendre des mesures conservatoires en cas de sinistre.**

– **Obligation de déclaration du sinistre :** L'assureur doit être averti le plus rapidement possible de la survenance d'un sinistre, de manière à lui permettre de prendre les mesures nécessaires pour en limiter les conséquences, ou exercer ses recours éventuels.

1.4.4 Caractère de bonne foi

La bonne foi doit précéder et accompagner toute la vie du contrat.

a. La bonne foi de l'assureur

- **Au moment de la souscription du contrat :** L'assureur est tenu d'une obligation de conseil tout au long de la vie du contrat, et notamment lors de la souscription du contrat. Il doit faire preuve de loyauté, en conseillant à son client des garanties adaptées, et en l'informant clairement sur les clauses et conditions du contrat.

- **A l' occasion du règlement du sinistre :** Il est fait appel à la notion de bonne foi pour sanctionner l'assureur qui se comporte de manière déloyale à l'égard de l'assuré, en refusant ou en retardant le règlement du sinistre.

b. La bonne foi de l'assuré

L'assuré doit répondre de bonne foi aux questions qui lui sont posées par l'assureur lors de la déclaration du risque, et doit déclarer les circonstances nouvelles d'aggravation de risque, faute de quoi il s'expose à la nullité du contrat, en cas de preuve de mauvaise foi de sa part.

L'assuré doit respecter les conditions de garantie prévues dans la police d'assurance (mesures de prévention, utilisation de moyens de protection ...) faute de quoi il s'expose à un non garanti.

Les conséquences de la faute intentionnelle de l'assuré sont légalement inassurables, ce qui est un principe d'ordre public.

L'assuré devra faire preuve de bonne foi dans la déclaration de sinistre (prise de mesures de sauvegardes, préservation des recours de l'assureur, respect du délai de déclaration du sinistre, accomplissement des formalités prévues au contrat ...), faute de quoi il s'expose à une déchéance de garantie. En cas de réalisation d'un sinistre frauduleux, l'assuré est non seulement déchu de la garantie, mais pourra être poursuivi pour le délit d'escroquerie à l'assurance.

1.5 Les éléments d'un contrat d'assurance

La souscription d'un contrat d'assurance est basée sur les éléments suivants :

1.5.1 La proposition d'assurance

Pour souscrire un contrat d'assurance, il est possible de s'adresser à un agent d'assurances, à un courtier d'assurances ou directement à une société d'assurances.

L'assureur sollicité remet au demandeur d'assurance (l'assuré), un imprimé présenté sous forme de questionnaire dit proposition d'assurance qui lui permettera par la suite d'évaluer les risques et de fixer la prime.

La proposition est la base pour la rédaction du contrat et la référence en cas de litige sur les déclarations initiales sur les risques. Elle n'engage, ni l'assuré, ni l'assureur :

– Après examen l'assureur peut refuser le risque ;

– Le proposant (l'assuré) peut refuser les offres de l'assureur.

L'assureur doit également remettre à l'assuré une fiche d'information sur les prix et les garanties et un exemplaire du projet de contrat et de ses annexes ou une notice d'information détaillée.

Remarques :

– Avant la conclusion d'un contrat comportant des garanties de responsabilité, l'assureur remet à l'assuré une fiche d'information décrivant le fonctionnement dans le temps des garanties déclenchées par le fait dommageable, le fonctionnement dans le temps des garanties déclenchées par la réclamation, ainsi que les conséquences de succession de contrats ayant des modes de déclenchements différents ;

– Lorsque l'assuré décide de remplir la proposition d'assurance, il ne doit pas donner d'informations inexactes, ni faire d'omissions, sous peine de ne pas être complètement indemnisé en cas de sinistre, ou de voir son contrat annulé ;

– La proposition remplie, l'assuré peut encore revenir sur sa décision, avant que l'assureur ait donné son accord ;

– Si la proposition d'assurance comporte la liste des garanties choisies et le montant de la prime, la signature de la personne intéressée l'engage, et le contrat est conclu dès que l'assureur donne son accord. Si l'assureur donne son accord, il doit remettre au futur assuré le contrat (conditions générales et particulières). Ce contrat doit également préciser le moment à partir duquel le risque est garanti et la durée de cette garantie ;

– L'assuré peut demander à être garanti provisoirement, par une note de couverture. L'assureur peut refuser ;

– L'assuré est engagé une fois que sa demande est acceptée, sauf s'il s'avère que le contrat définitif n'est pas conforme à la proposition (montant de la cotisation plus élevé, garanties différentes) ;

– Pour les contrats d'assurance vie et individuelle accident, les assureurs peuvent demander de remplir un questionnaire médical ou de se soumettre à un examen médical.

Article $n^o 8$ de l'ordonnance $95-07$:

" La proposition d'assurance n'engage l'assuré et l'assureur qu'après acceptation. La preuve de l'engagement des parties peut être établie soit par la police, soit par la note de couverture ou tout autre écrit signé de l'assureur. Est considérée comme acceptée, la proposition faite par lettre recommandée, de prolonger ou de remettre en vigueur un contrat suspendu ou de modifier un contrat sur l'étendue et le montant de la garantie, si l'assureur ne refuse pas cette proposition dans les vingt (20) jours après qu'elle lui soit parvenue. Les dispositions de cet alinéa ne s'appliquent pas aux assurances de personnes."

1.5.2 La police d'assurance

Dans la pratique, la police est synonyme de contrat d'assurance. Il s'agit d'un document contractuel qui régit les relations entre la compagnie d'assurance et l'assuré. La police d'assurance constitue la preuve du contrat. Toute police comporte au minimum deux parties :

a. Les conditions générales

Ce sont des documents pré-imprimés, commune à tous les assurés d'une même compagnie d'assurance pour chaque catégorie de risques. Elles reproduisent les dispositions communes à cette catégorie. Les conditions générales développent les thèmes suivants :

– les risques couverts (l'objet du contrat) ;
– les exclusions ;
– les obligations de l'assuré ;
– les dispositions relatives aux sinistres ;
– les règles de compétence et de prescription en cas de litiges.

L'assureur a l'obligation d'imprimer en caractères très apparents les paragraphes de la police édictant des nullités, des déchéances, ou des exclusions de garantie.

b. Les conditions particulières

Ce sont en partie pré-imprimés et, pour l'essentiel, dactylographiées ou, plus fréquemment émises par ordinateurs. Elles adaptent le contrat à la situation et au choix de chaque assuré. Les conditions particulières :

- personnalisent le risque ;
- prévalent sur les conditions générales et peuvent y déroger ;
- comportent obligatoirement les noms et domiciles des parties contractantes (y compris de l'intermédiaire), l'objet ou la personne assurée, la nature des risques garantis, la date d'effet et la durée du contrat, le montant de la garantie et les franchises et le montant de la prime.

c. Les conventions spéciales et les annexes

Ce sont des documents pré-imprimés qui précisent une garantie ou un point particulier. Les conventions spéciales prévalent sur les conditions générales, mais non sur les conditions particulières.

1.5.3 La note de couverture

La note de couverture est un document qui atteste de façon provisoire de l'existence d'une garantie à effet immédiat et pour une durée limitée. Elle est délivrée par l'assureur avant que le contrat lui-même soit établi. Par exemple, le conducteur d'un véhicule reçoit de son assureur une attestation provisoire d'assurance lorsqu'il vient d'acquérir celui-ci, dans l'attente du contrat définitif.

1.5.4 La durée du contrat

La durée du contrat est laissée au choix des parties contractantes. Le législateur intervient pour permettre aux partie, et notamment l'assuré, de se dégager périodiquement (résiliation du contrat à l'expiration d'un délai raisonnable).

Article $n^o 10$ de l'ordonnance $95-07$:

" La durée du contrat est fixée par les parties contractantes. Les conditions de résiliation sont régies par les dispositions afférentes à chaque catégorie d'assurance. Sous réserve des dispositions relatives aux assurances de personnes, l'assuré et l'assureur peuvent, dans les contrats à durée supérieure à trois (03) ans, demander la résiliation du contrat tous les trois (03) ans, moyennant un préavis de trois (03) mois."

1.6 L'assuré face à un sinistre

L'assuré a l'obligation d'informer l'assureur de la réalisation d'un sinistre. Deux situations peuvent se présenter : soit l'assuré se soumet à cette obligation, soit il s'y soustrait. Ce sont les deux cas de figure à reprendre, étant précisé que des sanctions attendent l'assuré qui manque à cette obligation.

1.6.1 La déclaration du sinistre par l'assuré

L'obligation de déclaration est générale. L'assuré est tenu de déclarer les événements qui mettent en jeu une garantie dès qu'il en a connaissance. De plus, et en fonction des risques couverts, il fournit un état estimatif détaillé des dommages subis par exemple par son véhicule assuré ou son habitation. L'assuré doit se soumettre aux recommandations voire aux instructions que l'assureur juge nécessaire à la protection de ses intérêts.

- **L'obligation de déclaration :** L'assuré est tenu de déclarer le sinistre à l'assureur dès qu'il en a connaissance. Il va de soi que seul doit être déclaré le sinistre qui va entrainer la garantie de l'assureur. S'il s'agit d'un sinistre sans rapport avec la garantie prévue au contrat, l'assureur répondra qu'il n'est pas concerné par cet événement. La déclaration est effectuée directement auprès de l'assureur ou d'un intermédiaire relevant de la compagnie.

- **Les délais de déclaration :** Ces délais sont importants. La seule difficulté en ce domaine réside dans le fait qu'il n'existe pas un délai uniforme mais des délais différents en fonction des sinistres : trois (03) jours en cas de vol, quatre (04) jours pour les dégâts causés par la grêle et 24 heures en cas de mortalité de bétail… Ces différents délais présentent tout de même un point commun : qu'ils soient longs ou courts, ils sont conçus pour laisser à l'assuré le temps de réagir après qu'un sinistre se soit produit.

1.6.2 La sanction principale : la déchéance de garantie

Les sanctions prévues n'ont pas toutes la même intensité et les mêmes conséquences. La sanction principale est la déchéance de garantie, les autres peuvent être considérées comme complémentaires.

Il arrive parfois que l'assuré décide de surestimer le sinistre afin de percevoir une indemnité plus élevée ou alors que sa déclaration soit intervenue de façon trop tardive. Dans les faits, la notion de déchéance a parfois du mal à être distinguée de celle d'exclusion. En toute logique, la première se définit comme la perte d'un droit qui avait été reconnu à l'assuré (le droit de bénéficier d'une garantie). La seconde correspond au refus de l'assureur de prendre en charge un risque dès la conclusion du contrat.

La déchéance est une *sanction conventionnelle*. Il faut donc s'en remettre à la manière dont la clause de déchéance a été rédigée dans le contrat d'assurance. Les déchéances doivent figurer dans le contrat en termes apparents et la clause doit être claire et précise. L'assureur doit le cas échéant établir que le retard de l'assuré lui

a causé un préjudice. La déchéance s'applique à la garantie. L'assuré est privé de garantie s'agissant du sinistre qui vient de se produire. Le contrat d'assurance, quant à lui, demeure valable pour l'avenir.

1.7 Diverses évolutions du risque

Dans la pratique, le risque n'est pas figé et peut varier d'intensité en fonction de divers paramètres. En général, l'assuré doit en cours d'exécution du contrat informer la compagnie des éléments nouveaux qui ont pour conséquences de rendre inexacte ou caduque les renseignements donnés lors de la souscription du contrat. Il y a logiquement deux situations à distinguer selon que le risque s'aggrave ou s'amoindrit.

1.7.1 Le cas de l'aggravation du risque

Le terme aggravation est empreint d'une signification particulière en droit des assurances compte tenu de la diversité des situations pratiques et de la diversité des risques. Il nous semble donc utile de réserver quelques développements à la notion d'aggravation, avant de s'intéresser à ses conséquences : sa signification en droit des assurances n'est pas la même que dans d'autres matières.

Le mot " aggravation " peut s'appliquer aux circonstances nouvelles qui ont pour conséquences d'augmenter la probabilité de réalisation du risque ou d'accroitre la gravité de celui-ci. En théorie, il faut distinguer l'aggravation d'un risque existant de l'apparition d'un risque nouveau. Dans le cadre de l'assurance habitation par exemple, le fait d'acquérir un objet de valeur et de l'intégrer dans le mobilier du logement peut constituer une aggravation : en effet, en cas d'incendie ou de vol, la disparition de l'objet de valeur peut représenter un cout supplémentaire pour l'assureur.

L'obligation de déclaration s'impose à l'assuré puisque les renseignements donnés dans le questionnaire initial se trouvent modifiés. L'assureur doit être informé de l'aggravation afin qu'il puisse apprécier cette dernière et en tirer toutes les conséquences quant à l'étendue de ses engagements. La poursuite du contrat d'assurance ainsi que le montant de la prime en dépendent.

Les conséquences de l'aggravation sont diverses. En premier lieu, l'assureur a la faculté de résilier le contrat d'assurance en raison de l'aggravation, par exemple si le risque excède les limites de son activité ou de ses possibilités financières. En second lieu, l'assureur peut décider de poursuivre l'exécution du contrat. Dans cette hypothèse une distinction supplémentaire peut être opérée. Tantôt l'assureur décide d'une augmentation du montant de la prime, ou d'une surprime, tantôt il décide d'exclure de la garantie la circonstance nouvelle dont il vient d'être informé.

1.7.2 Le cas de la diminution du risque

En théorie, on peut entendre la notion de diminution de plusieurs façons. Il peut s'agir de la circonstance nouvelle dont la conséquence est de réduire l'intensité ou la probabilité de réalisation d'un risque. On peut ajouter la disparition d'une circonstance qui aggravait le risque initial. Selon le code des assurances, l'assuré peut prétendre à une diminution du montant de la prime. En cas de refus de l'assureur, l'assuré peut dénoncer le contrat.

1.8 Le versement de la prime

L'assuré est débiteur de la prime en vertu du contrat d'assurance. Logiquement, deux hypothèses sont à envisager selon que l'assuré exécute ou n'exécute pas son obligation. Dans le premier cas, il faut s'interroger sur les modalités de paiement de la prime ; dans le second, il faut envisager les sanctions qui peuvent être prises contre un assuré défaillant.

1.8.1 L'obligation de l'assuré de verser la prime

En matière d'assurance comme bien d'autres domaines, l'essentiel est que la prime soit payée et cela au bon moment. Le reste est secondaire : qu'elle soit payée par chèque, virement bancaire ou selon un autre procédé est sans incidence sur la garantie accordée par l'assureur. La seule certitude que souhaitent obtenir les assureurs est d'être payé à la date convenue et de recevoir le montant qu'il leur revient. Le lieu et le procédé de paiement ne sont pas des éléments déterminant.

La date du paiement dépend de ce qui a été convenu par les parties et de la branche d'assurance. En général, la prime est versée en début de contrat. C'est souvent le cas en assurance automobile : dans ce cas l'assureur calcule une prime annuelle qui doit être réglée dès les premiers jours. La prime peut, dans d'autres cas, être versée tous les mois, par exemple en assurance vie.

1.8.2 Les conséquences en cas de non paiement

- **La procédure de sanction :** En pratique, l'assureur fait parvenir à l'assuré un avis d'échéance par lequel il l'informe que la prime est exigible. A défaut de paiement, l'assureur lui adresse une lettre recommandée dans laquelle il rappelle le montant de la prime, l'obligation de paiement et les sanctions encourues. La mise en demeure intervient en principe dans les dix jours qui suivent la date d'échéance. L'assuré dispose d'un délai de 30 jours à compter de la mise en demeure pour régler la prime. Pendant ce délai, la garantie est maintenue ; au delà, elle est suspendue. En cas de sinistre, aucune indemnité n'est alors due à l'assuré.

- **La suspension de la couverture :** C'est la garantie qui est suspendue et non le contrat d'assurance. Ainsi, l'assuré continue à rester débiteur de la prime alors que l'assureur est déchargé de son obligation de ga-

rantie. Si un sinistre se produit pendant la période de suspension, il n'est tenu de verser aucune indemnité à l'assuré. La garantie retrouve son plein effet lorsque l'assuré verse la prime.

– **La résiliation du contrat pour l'assureur :** Il s'agit d'une *sanction radicale*. Elle ne peut intervenir qu'au bout de dix jours après l'expiration des trente jours suivant la mise en demeure. L'assureur fait parvenir à l'assuré un avis de résiliation du contrat d'assurance. Ce dernier cesse alors d'exister.

1.9 Conclusion

Etant à caractère synallagmatique, le contrat d'assurance met à la charge des parties assuré-assureur, des obligations nécessairement réciproques. L'assuré doit verser la prime à son délai, prévenir l'assureur de la survenance du sinistre dès qu'il en connaissance, déclarer les circonstances nouvelles d'aggravation du risque assuré et respecter les conditions de la garantie prévue dans la police d'assurance. Quant à l'assureur, il est tenu d'accorder une prestation à l'assuré à la date convenue, en cas de réalisation du sinistre.

Dans le cas où l'un d'entre eux se désengage de ses obligations que ce soit à l'initiative de l'assuré ou celle de l'assureur, des sanctions sont prévues. Il s'agit de deux types de sanctions : une sanction conventionnelle telle la suspension de la couverture (la garantie) et non pas du contrat et une sanction radicale qui n'est rien d'autre que la résiliation du contrat.

Les compagnies d'assurances s'intéressent de plus en plus à l'étude du phénomène de résiliation des contrats, en faisant appel à des méthodes statistiques plus efficaces et adéquates à ce type de problème, telles les méthodes de l'analyse de survie.

2

Concept de base d'analyse de survie

Sommaire

2.1 Introduction

L'analyse des données de survie possède deux particularités intrinsèques, d'une part, celle-ci concerne que des variables aléatoires positives et d'autre part, la présence de données censurées.

Dans le but de présenter les concepts de base de l'analyse de survie, nous commençons d'abord par décrire ce que sont les données de survie et la censure. Nous rappelons ensuite le principe de l'analyse des données de survie et les différentes fonctions utilisées dans cette analyse, ainsi que les tests d'adéquation employées pour déterminer si un échantillon de temps de survie suit une distribution de probabilité connue. Enfin, nous donnons les différentes distributions paramétriques utilisées pour modéliser les durées.

2.2 Données de survie et censure

2.2.1 Données de survie

Les données de survie [3] représentent le temps écoulé entre le début d'une observation et la survenue d'un certain évènement ; où ce temps est connu sous le nom de temps de survie. Cet évènement peut être le décès d'un patient après une intervention chirurgicale, l'apparition d'une maladie ou d'une épidémie, le divorce, la naissance d'un enfant, la reprise d'un emploi, la panne d'une machine, le changement d'une résidence ... etc.

Dans plusieurs cas, l'évènement d'intérêt 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 2.1 illustre ces deux exemples.

Selon le contexte, les termes décès, évènement, échec ou transition peuvent être utilisés pour désigner l'évènement 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'incidence d'une maladie, et dans d'autres cas, l'aspect intéressant est l'état de la disparition d'une maladie (Hougaard, 1999).

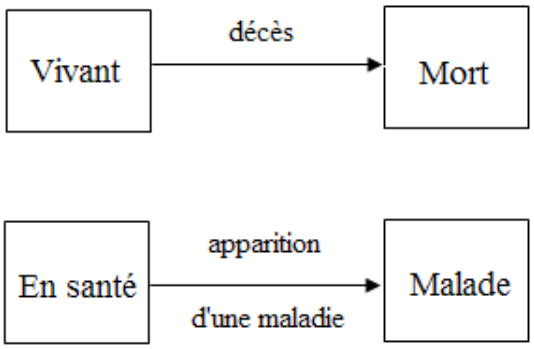

FIG. 2.1 – Exemples illustratifs de temps de survie.

2.2.2 Censure

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

Les données censurées [8] sont des observations pour lesquelles la valeur exacte d'un évènement 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 (si l'évènement d'intérêt considéré est le décès).

Il existe trois (03) 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 évènement) ; par exemple, si on souhaite étudier l'âge d'apprentissage de la lecture (notée T) dans une classe d'élèves de première année de cours primaire, trois cas peuvent se produire :

– Si un élève sait déjà lire en entrant au cours primaire, T n'est pas observé. On sait seulement que T est inférieur à l'âge de cet élève à l'entrée au cours primaire. Il s'agit donc d'une *censure à gauche* ;

– Si un élève achève sa première année de cours primaire sans savoir lire, on n'observe pas non plus T. On sait seulement que T est supérieur à l'âge de cet élève à la fin de la première année. Alors, on dit qu'il

s'agit d'une *censure à droite* ;

– Par contre, si l'élève apprend à lire en cours d'année, la durée T est bien observée et la donnée est dite *complète*.

À l'intérieur de ces trois (03) catégories, il existe différents types de censure :

– *Censure de type I* : Si le temps de censure est fixé par le chercheur comme étant la fin de l'étude ;

– *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 ;

– *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).

Considérons par exemple une étude de la survie de patients qui ont été soumis à une transplantation du cœur et qui sont suivis après l'opération pendant une période de 52 semaines. Dans ce cas, le temps origine est représenté par le moment de la transplantation et l'événement d'intérêt est le décès.

La figure 2.2 illustre les situations de censure, à travers la représentation de survie de cinq patients. Un cercle plain indique un événement observé ; un cercle vide représente un événement non observé ; un carré représente une censure fixe. Une ligne continue représente une période pendant laquelle les sujets sont observés être soumis au risque de connaître l'événement ; une ligne pointillée, une période pendant laquelle un sujet reste soumis au risque, sans qu'il ne soit observé.

La première observation est non-censurée ; le deuxième sujet est censuré car il est encore vivant à la fin des 52 semaines de l'étude (censure fixe) ; le troisième patient sort de l'étude, et donc la durée correspondante est censurée, 20 semaines après la transplantation, par exemple parce qu'il déménage et il est suivi par d'autres médecins (censure aléatoire). Les deux censures considérées représentent des cas de censure à droite. Il est toutefois intéressant de remarquer la possibilité d'une censure à gauche qui se vérifie quand un sujet entre dans l'étude un certain temps après le début de l'étude (*late entry*). C'est le cas des patients 4 et 5, dont le premier connaît l'événement avant la fin de l'étude, alors que le deuxième est déjà à une censure a droite (censure par intervalle).

Dans les méthodes d'analyse de survie, tous les sujets qui sont encore observés en t sont considérés comme "soumis au risque" à un certain temps t, et donc qui n'ont pas été censurés avant t. Or, pour que les estimateurs des temps ou des probabilités de survie soient non biaisées, il faut assumer que les sujets observés en t soient représentatifs de tous les sujets, même de ceux qui sont sortis de l'étude avant t. Ceci équivaut à as-

sumer que le mécanisme de censure est indépendant du temps. On parle dans ce cas de *censure non-informative*.

Désormais, nous s'intéressons à la censure à droite, non informative.

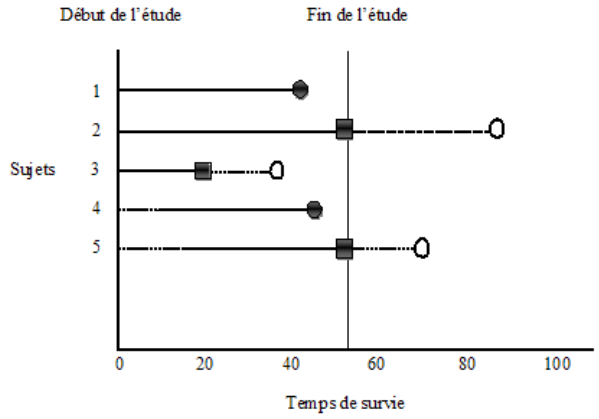

FIG. 2.2 – Survie de patients soumis à une transplantation du cœur.

2.3 Principe de l'analyse de survie

L'analyse de survie est un domaine de la statistique qui a pour objet l'étude de la survenue au cours du temps d'un évènement, comme par exemple le décès, et ceci en présence de données censurées. Ce type d'analyse est largement utilisé en épidémiologie clinique. Il permet la description de la survie d'un groupe de patients mais aussi la comparaison de la survie de deux ou plusieurs groupes de patients afin d'étudier les facteurs pronostiques, c'est-à-dire les facteurs susceptibles d'expliquer la survenue du décès.

Récemment, l'analyse de survie a été appliquée à un large éventail de domaines des sciences sociales. Ces applications incluent l'étude du taux d'échec des nouvelles entreprises canadiennes (Baldwin et coll., 2000), de la durée des grèves (Greene, 1993), de la durée du chômage (Kiefer, 1988), du roulement et de la mobilité des entreprises (Caves, 1998), de la survie des nouvelles entreprises (Audretsch et Mahmood, 1995) et de la durée des cycles économiques (Abderrezak, 1997).

2.4 Fonctions de survie et de hasard

Nous considérons une variable aléatoire positive T s'interprétant comme le temps de survie, par exemple d'un individu. De façon à simplifier la présentation, nous supposons que la distribution de la variable T est continue, de densité de probabilité f et de fonction de répartition F.

La loi des temps de survie peut aussi être caractérisée par l'intermédiaire d'autres fonctions, faciles à interpréter dans ce contexte et qui de plus s'introduisent dans divers calculs. Parmi ces fonctions les plus importantes : la *fonction de survie* et la *fonction de hasard* [10].

La fonction de survie,

$$S(t) = P(T \geq t) = \int_t^{+\infty} f(u)du = 1 - F(t) \tag{2.1}$$

donne la probabilité que le temps de survie d'un individu dépasse t, c'est-à-dire la probabilité que l'individu soit toujours vivant après t unités de temps. C'est une fonction monotone décroissante, satisfaisant les conditions limites :

$$S(t) = \begin{cases} 1, & \text{si } t = 0 \\ 0, & \text{si } t = \infty \end{cases}$$

Pour décrire l'évolution de la survie au cours du temps, Berkson (1942) a recommandé une présentation graphique de $S(t)$, appelée *courbe de survie*.

La fonction de hasard (ou fonction de risque),

$$\lim_{\Delta t \to 0} P(T \in [t, t + \Delta t[/ T \geq t)$$

est la probabilité qu'un l'individu décède à l'instant t sachant qu'il était encore vivant avant cet instant.

Il est possible d'établir des liens entre la fonction de hasard, la fonction densité de probabilité et la fonction de survie :

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t) \tag{2.2}$$

La fonction de hasard caractérise la loi de T, du fait de la relation :

$$S(t) = \exp\left[- \int_0^t h(u)du \right] \tag{2.3}$$

Cette formule conduit d'ailleurs à introduire le hasard cumulé définit par :

$$H(t) = \int_0^t h(u)du = -\log S(t) \tag{2.4}$$

2.5 Tests d'ajustement

Les tests d'ajustement seront employés dans le but de vérifier si un échantillon $(t_1, t_2, \cdots, t_n)$ provient ou non de la variable aléatoire T de distribution connue F_0.

Etant donnée $F(t)$ la fonction de répartition de la variable échantillonnée, le problème revient alors à tester les deux hypothèses suivantes :

$$H_0 : F(t) = F_0(t) \quad contre \quad H_1 : F(t) \neq F_0(t)$$

Les deux tests les plus classiques sont : *test du Chi-deux* et *test de Kolmogorov- Smirnov*. Ces tests sont fondés sur des statistiques qui dépendent des fonctions de répartitions respectivement, empirique F_n et théorique F assimilable à des distances ou à des pseudo-distances entre les lois de probabilités.

On possède le nombre de réalisations $n_i \ \ (i = \overline{1,m})$ de m éventualités au cours de n expériences identiques indépendantes, c'est-à-dire les fréquences empiriques. Ainsi, on observe une variable T sur n individus et l'espace des observations a été divisé en m catégories.

Notons $p_1, p_2, \cdots, p_m$ les probabilités de chaque éventualité calculées à partir d'une distribution théorique de T donnée, parfaitement spécifiées, de fonction de répartition connue F.

2.5.1 Test du Chi-deux

La répartition de l'échantillon de taille n sur les m éventualités suit une loi multinomiale dont les paramètres sont $n, p_1, p_2, \cdots, p_m$.

Soit F_n la distribution empirique de T observée à partir de l'échantillon. On appelle *distance du Chi-deux* entre la loi théorique et la loi empirique observée, la quantité donnée par :

$$D(F_n, F) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \tag{2.5}$$

La statistique $D(F_n, F)$ suit asymptotiquement la loi du Chi-deux à $(m-1)$ degrés de liberté.

Soit $\hat{p}_i = \frac{n_i}{n}$ la proportion empirique, on a alors :

$$D(F_n, F) = D(\hat{p}, p) = \sum_{i=1}^{m} \frac{(\hat{p}_i - p_i)^2}{p_i} \tag{2.6}$$

Le problème consiste alors à tester les deux hypothèses suivantes :

$$H_0 : \texttt{la loi de } T\texttt{est } F \quad contre \quad H_1 : \texttt{la loi de } T\texttt{n'est pas } F$$

La réponse à ce problème est définie par la région critique qui est donnée par :

$$\left\{ \ W = (n_1, n_2, \cdots, n_m) / D(\hat{p}, p) > c \quad \text{tels que } P_{H_0}(W) = \alpha. \right.$$

Si le nombre de paramètre à estimer est k, alors le nombre de degrés de liberté du Chi-deux sera $(m-k-1)$.

2.5.2 Test de Kolmogorov-Smirnov

Soit T une variable aléatoire réelle de loi inconnue, de fonction de répartition F et soit $(t_1, t_2, \cdots, t_n)$ un échantillon de T.

On désire ajuster une loi inconnue à une loi donnée de fonction de répartition F_0 au vu de la fonction de répartition empirique F_n.

Il s'agit donc de tester le jeu d'hypothèses suivantes :

$$H_0 : F = F_0 \quad contre \quad H_1 : F \neq F_0$$

Pour cela, on définit la statistique :

$$K_n = \sup |F_n(t) - F(t)| \tag{2.7}$$

Théorème 2.1 *Sous l'hypothèse de H_0, la statistique $n^{\frac{1}{2}} K_n$ converge asymptotiquement vers la loi de fonction de répartition R définie par :*

$$R(t) = 1 - \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 t^2)$$

Les conclusions du test sont classiques, on acceptera H_0 si la statistique K_n prend des valeurs faibles, d'où la région critique sera :

$$\left\{ W = (t_1, t_2, \cdots, t_n)/K_n > c \quad \text{tels que } P_{H_0}(W) = \alpha. \right.$$

2.6 Différentes distributions de survie

Parfois, il peut être intéressant de spécifier une forme paramétrique de la distribution des temps de survie [7], de façon à pouvoir résumer l'information relative à cette variable à l'aide d'un petit nombre de paramètres. En principe toute distribution de variable aléatoire positive peut être utilisée pour représenter les durées ; les plus couramment utilisées sont : la distribution exponentielle, la distribution de Weibull, la distribution de valeur extrême, la distribution Log-normale et la distribution Log-logistique. Certain d'autres modèles paramétriques peuvent être trouvés dans Kalbfleisch et Prentice (2002) et Lawless (2003).

2.6.1 Distribution exponentielle

La distribution la plus importante et la plus simple dans les études de survie est la distribution exponentielle. En 1940, les chercheurs ont commencé à choisir cette distribution pour décrire la durée de vie des systèmes électronique.

La distribution exponentielle est caractérisée par une fonction de hasard constante, indépendante du temps t :

$$h(t) = \lambda, \quad t \geq 0 \tag{2.8}$$

où $\lambda > 0$. La fonction densité de probabilité et la fonction de survie sont définies respectivement par

$$f(t) = \lambda e^{(-\lambda t)} \tag{2.9}$$

et

$$S(t) = e^{(-\lambda t)} \tag{2.10}$$

La distribution où $\lambda = 1$ est appelée la *distribution exponentielle standard* ; sa fonction densité de probabilité est montrée dans la figure 2.3.

Dans une étude de nouveaux traitements anti-cancer appliqués sur un ensemble d'animaux atteints de leucémie, Zelen (1966) a utilisé la distribution exponentielle pour modéliser le temps de survie de ces animaux [8].

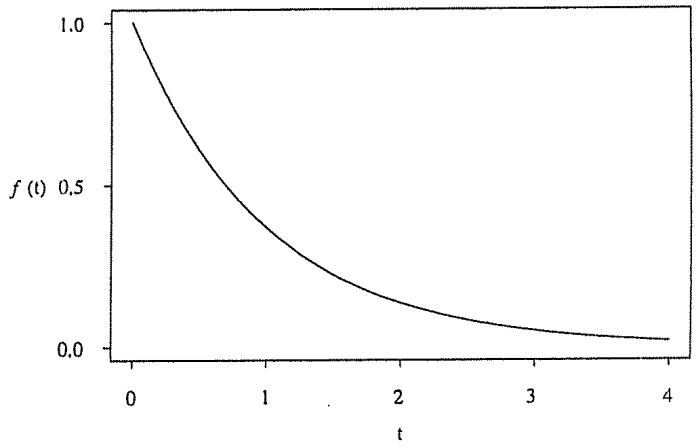

FIG. 2.3 – Fonction densité de probabilité de la distribution exponentielle standard.

2.6.2 Distribution de Weibull

La distribution de Weibull est très flexible pour modéliser des temps de survie. La fonction de hasard de cette distribution est donnée par

$$h(t) = \lambda \gamma t^{\gamma - 1} \tag{2.11}$$

où $\lambda > 0$ et $\gamma > 0$ sont les paramètres. La fonction densité de probabilité et la fonction de survie sont

$$f(t) = \lambda \gamma t^{\gamma - 1} \exp(-\lambda t^{\gamma}), \quad t > 0 \tag{2.12}$$

et

$$S(t) = \exp(-\lambda t^{\gamma}), \quad t > 0 \tag{2.13}$$

La fonction de hasard de la distribution Weibull est monotone croissante si $\gamma > 1$, décroissante si $\gamma < 1$, et constante similaire à celle de la distribution exponentielle pour $\gamma = 1$. La figure 2.4 montre certaines fonctions densité de probabilités et les fonctions de hasard correspondantes pour $\lambda = 1$ et différentes valeurs de γ.

Pike (1966) a utilisé la distribution Weibull pour décrire le temps de survie, en jours, de 40 rats atteints de cancer [8].

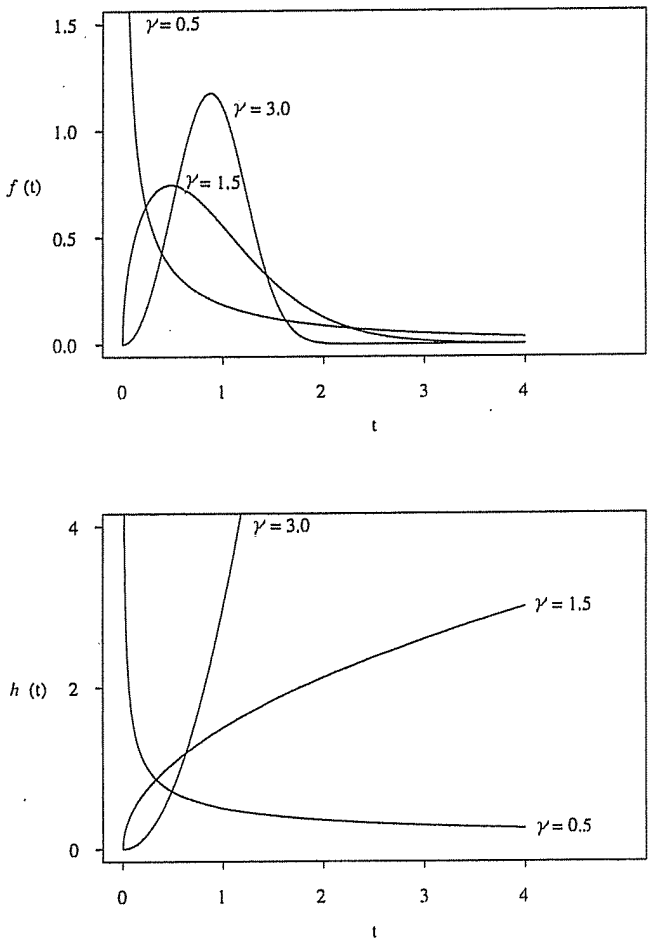

FIG. 2.4 – Fonctions densité de probabilité et fonctions de hasard de la distribution de Weibull pour $\lambda = 1$ et $\gamma = 0.5, 1.5, 3.0$

2.6.3 Distribution valeur extrême

La fonction densité de probabilité et la fonction de survie pour la distribution valeur extrême sont respectivement,

$$f(y) = b^{-1} \exp\left[\frac{y-\mu}{b} - \exp\left(\frac{y-\mu}{b}\right)\right], \quad -\infty < y < +\infty \tag{2.14}$$

$$S(t) = \exp\left[-\exp\left(\frac{y-\mu}{b}\right)\right], \quad -\infty < y < +\infty \tag{2.15}$$

où $b > 0$ et $(-\infty < \mu < +\infty)$ sont les paramètres. On peut facilement déduire que si T a une distribution Weibull avec la fonction densité de probabilité (2.12), alors $\log T$ suit une distribution valeur extrême avec $b = \gamma^{-1}$ et $\mu = -\log\lambda$.

La distribution valeur extrême avec $\mu = 0$ et $b = 1$ est dite *distribution valeur extrême standard*. Un graphe de sa fonction densité de probabilité est donné dans la figure 2.5.

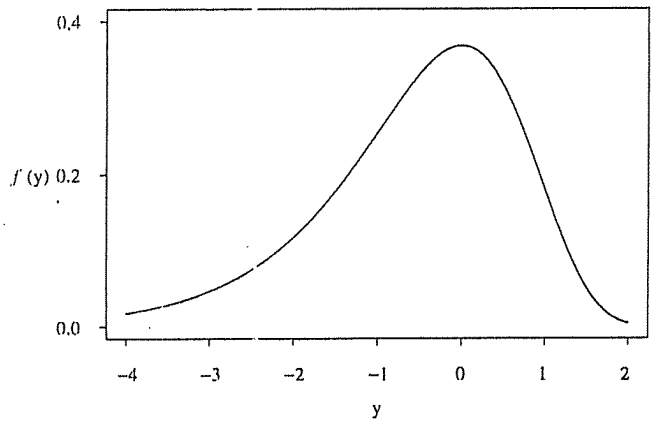

FIG. 2.5 – Fonction densité de probabilité de la distribution valeur extrême standard.

2.6.4 Distribution Log-normale

La variable temps de survie T est dite distribuée selon une Log-normale si $Y = \log T$ a une distribution normale. Sa fonction densité de probabilité s'écrit comme suit :

$$f(t) = \frac{1}{s\,t\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log t - m}{s}\right)^2\right], \quad t > 0 \tag{2.16}$$

où les paramètres m et s désignent respectivement la moyenne et l'écart-type du logarithme.

La fonction de survie de cette distribution est donnée par

$$S(t) = 1 - \Phi\left(\frac{\log t - m}{s}\right) \tag{2.17}$$

où Φ est la fonction de répartition de la loi normale standard.

La fonction de hasard est donnée par la relation suivante $h(t) = f(t)/S(t)$. Cette fonction est non monotone et uni-modale (croissante puis décroissante) ; $lim_{t\nearrow 0}\ h(t) = 0$ et $lim_{t\searrow \infty}\ h(t) = 0$. La figure 2.6 des fonctions densité de probabilité, et des fonctions de hasard pour $m = 0$ et différentes valeurs de s.

En 1960, Feinleib et MacMahon ont appliqué la distribution Log-normale pour modéliser le temps de survie, en mois, de 234 patients de sexe masculins, atteints d'une leucémie lymphocytique chronique [8].

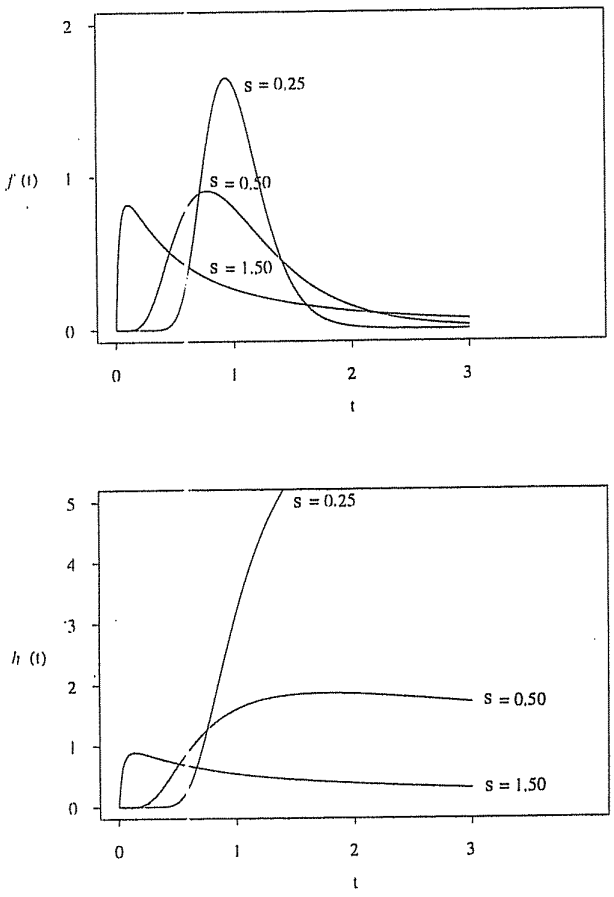

FIG. 2.6 – Fonctions densité de probabilités et fonctions de hasard de la distribution Log-normale pour $m = 0$, et $s = 0.25,\ 0.5,\ 1.5$

2.6.5 Distribution Log-logistique

La distribution Log-logistique a une fonction densité de probabilité de la forme :

$$f(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{[1+(t/\alpha)^{\beta}]^2}, \quad t > 0 \tag{2.18}$$

où $\alpha > 0$ et $\beta > 0$ sont les paramètres. La fonction de survie et la fonction de hasard sont respectivement,

$$S(t) = [1+(t/\alpha)^{\beta}]^{-1} \tag{2.19}$$

$$h(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{[1+(t/\alpha)^{\beta}]} \tag{2.20}$$

Le nom de la distribution log-logistique dérive du fait que $Y = logT$ a une distribution logistique, de fonction densité de probabilité

$$f(y) = \frac{b^{-1}\exp[(y-\mu)/b]}{(1+\exp[(y-\mu)/b])^2}, \quad -\infty < y < \infty \tag{2.21}$$

où $\mu = \log\alpha$ et $b = \beta^{-1}$, avec $-\infty < \mu < \infty$ et $b > 0$. La figure 2.7 montre les fonctions densité de probabilité et les fonctions de hasard de T pour $\mu = 0$ et différentes valeurs de b.

Si $\beta > 1$, la fonction de hasard de la distribution Log-logistique est similaire à celle de la distribution Log-normale ; $h(t)$ croit de 0 vers un maximum et puis décroit vers 0. Si $\beta \leq 1$, la fonction de hasard est monotone décroissante.

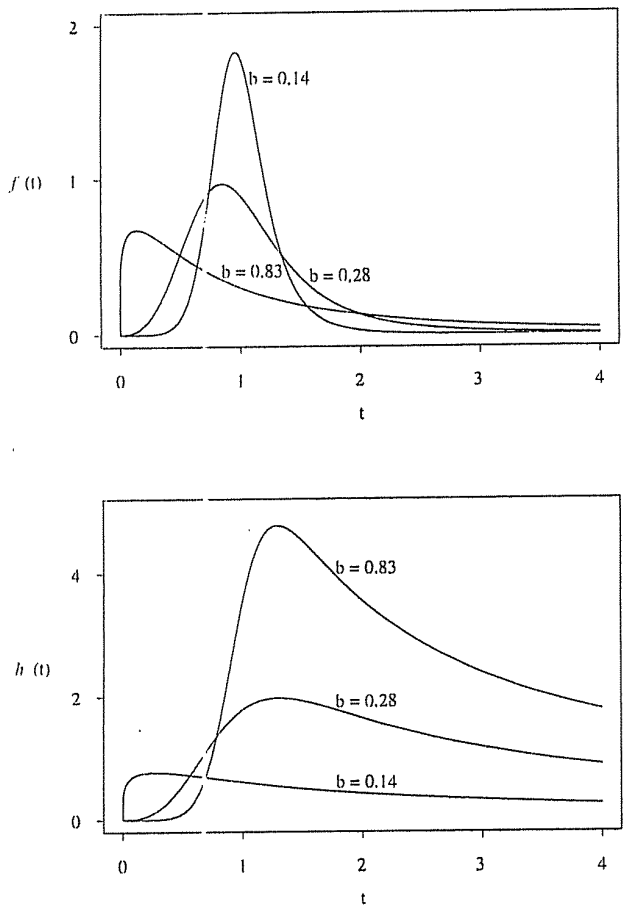

FIG. 2.7 – Fonctions densité de probabilité et fonctions de hasard de la distribution Log-logistique pour $\mu = 0$ et $b = 0.14, 0.28, 0.83$

2.7 Conclusion

La donnée temps de survie qui mesure le temps écoulé jusqu'à la survenue d'un certain évènement d'intérêt (défaillance d'une machine, décès, divorce, résiliation d'un contrat ... etc.) est considéré comme une variable aléatoire positive dont la distribution est généralement caractérisée par trois (03) fonctions : fonction de survie, fonction densité de probabilité et fonction de hasard. Ces fonctions sont mathématiquement équivalentes ; car si l'une d'entre elles est donnée, les deux autres peuvent être déterminées.

En pratique, ces trois fonctions sont introduites pour illustrer les différents aspects de données. Le problème de base en analyse des données de survie est d'évaluer ces fonctions, en raison de la présence de censures. Pour cela, des procédures non paramétriques sont utilisées notamment pour l'estimation de la fonction de survie, ainsi décrites dans le chapitre suivant.

3

Méthodes non-paramétriques

Sommaire

3.1 Introduction

Une première étape dans l'analyse des données de survie est de présenter des synthèses numériques et graphiques des temps de survie pour des individus dans un groupe particulier. Les données de survie sont pratiquement résumées à travers des estimations de la fonction de survie et la fonction de hasard. Dans ce chapitre, nous s'intéressons aux estimations de la fonction de survie en utilisant des méthodes de survie non-paramétriques [6][9], car elles n'exigent pas une hypothèse spécifique à la distribution des durées, à savoir : estimation table de survie, estimation Kaplan-Meier et estimation Nelson-Aalen.

Une fois que, la fonction de survie est estimée, l'erreur standard de cette dernière peut être déterminée, ainsi que l'intervalle de confiance à 95% pour la vraie valeur de la fonction de survie.

En outre, une procédure non - paramétrique pour la comparaison des courbes de survie sera également décrite dans ce chapitre ; il s'agit du test Log-rank.

3.2 Estimation de la fonction de survie

Considérons d'abord un échantillon de temps de survie, où les observations sont non censurées. La fonction de survie $S(t)$, définie dans l'équation (2.1), est la probabilité qu'un individu survive pour un temps supérieur ou égale à t. Cette fonction peut être estimée par la fonction de survie empirique, donnée par

$$\hat{S}(t) = \frac{\texttt{Nombre d'individus ayant des temps de survie} \geq t}{\texttt{Nombre d'individus dans l'ensemble des données}} \tag{3.1}$$

De même, $\hat{S}(t) = 1 - \hat{F}(t)$, où $\hat{F}(t)$ est la fonction de distribution empirique, qui est, le rapport entre le nombre total d'individus qui sont vivants à l'instant t et le nombre d'individus dans l'étude. Notons que la fonction de survie empirique est égale à une unité pour des valeurs de t précédant le premier temps de décès, et zéro après le dernier temps de décès.

La fonction de survie estimée $\hat{S}(t)$ est supposée être constante entre deux temps de décès adjacents, et ainsi le tracé de $\hat{S}(t)$ en fonction du temps, est une fonction en escalier. La fonction décroit immédiatement à chaque temps de survie observé.

Exemple 3.1 Une complication dans la gestion de patients ayant une tumeur d'os maline, ou osteosarcoma, est que la tumeur se propage souvent aux poumons. Cette metastasis pulmonaire est une menace pour la survie des patients. Dans une étude concernant le traitement de cette maladie, Burdette et Gehan (1970) ont présentés les temps de survie suivants, en mois, de 11 patients de sexe masculin :

11 13 13 13 13 13 14 14 15 15 17

En utilisant l'équation (3.1), les valeurs estimées de la fonction de survie respectivement aux instants 11, 13, 14, 15 et 17 mois sont 1.000, 0.909, 0.455, 0.273, et 0.091. La valeur estimée de la fonction de survie est une unité à partir du temps d'origine jusqu'à 11 mois, et zéro après 17 mois. Un graphe de la fonction de survie estimée est donné dans la figure 3.1.

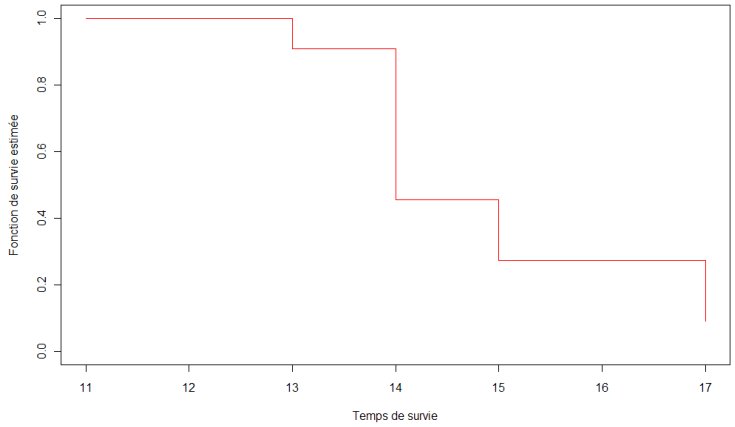

FIG. 3.1 – Fonction de survie estimée pour les données de l'exemple 3.1.

La méthode d'estimation de la fonction de survie illustrée dans l'exemple ci-dessus ne peut pas être employée, lorsqu'il y'a des observations censurées. Cela dû au fait que la méthode ne permet pas l'information fournie par un individu dont le temps de survie est censuré avant t, à être utilisée dans le calcul de la fonction de survie estimée à cet instant. Des procédures non-paramétriques pour estimer $S(t)$, qui peuvent être utilisées en présence des temps de survie censurés, sont décrites dans les sections suivantes.

3.2.1 Estimation table de survie

L'estimation table de survie ou l'estimation actuarielle de la fonction de survie est obtenue par une première division de la période d'observation en une série d'intervalles de temps. Ces intervalles ne sont pas nécessairement de même longueur, bien qu'ils le soient habituellement. Le nombre d'intervalles utilisés dépendra du nombre d'individus dans l'étude, mais il est très fréquemment situé entre 5 et 15.

Supposons que le j-ème intervalle $I_j,\ j=\overline{1,m}$, s'étend du l'instant $t_j^{'}$ jusqu'à $t_{j+1}^{'}$ et notons, respectivement, d_j et c_j, le nombre de décès et le nombre des temps de survie censurés durant cette intervalle de temps. En outre, soit n_j le nombre d'individus qui sont encore vivant, et donc exposés au risque de décès, au début de l'intervalle I_j. Posons l'hypothèse que le processus de censures est tel que, les temps de survie censurées se produit uniformément sur tout l'intervalle, ainsi que le nombre moyen d'individus à risque durant cet intervalle

est

$$n_j^{'} = n_j - \frac{c_j}{2} \tag{3.2}$$

Cette hypothèse est connue sous le nom de l'hypothèse actuarielle. Dans l'intervalle I_j, la probabilité de décès peut être estimée par $\frac{d_j}{n_j^{'}}$, ainsi que la probabilité de survie correspondante est $\frac{n_j^{'}-d_j}{n_j^{'}}$.

Maintenant, considérons la probabilité qu'un individu survive au-delà de $t_k^{'},\ k=\overline{1,m}$, c'est-à-dire jusqu'à un certain instant après le début du k-ième intervalle. Ce sera le produit des probabilités qu'un individu survive au-delà du début du k-ième intervalle et durant chacun des $(k-1)$-ième intervalles précédents, et donc l'estimation table de survie de la fonction de survie est donnée par

$$S^*(t) = \prod_{j=1}^{k} \frac{n_j^{'} - d_j}{n_j^{'}} \tag{3.3}$$

Pour $t_k^{'} < t < t_{k+1}^{'},\ k=\overline{1,m}$. La probabilité estimée de survivre jusqu'au début du premier intervalle $t_1^{'}$ est une unité, tandis que la probabilité estimée de survivre au-delà de $t_{m+1}^{'}$ est zéro. Une estimation graphique de la fonction de survie sera alors, une fonction en escalier avec des valeurs constantes de la fonction dans chaque intervalle de temps.

Exemple 3.2 Pour illustrer le calcul de l'estimation table de survie, considérons les données sur les temps de survie de 48 patients de myeloma multiple, ainsi présentées dans le tableau 3.1.

Myeloma multiple est une maladie maline caractérisée par l'accumulation des cellules anormales de plasma, un type de cellule globule blanche, dans la moelle. La prolifération des cellules anormales de plasma dans l'os provoque une douleur et une destruction du tissu d'os. En outre, les patients de myeloma multiple, éprouvent l'anémie, les hémorragies, les infections récurrentes et la faiblesse. Le but d'une étude effectuée au Centre Médical d'Université de West Virginia, USA, était d'examiner la relation entre les valeurs de certaines variables exogènes ou covariables et le temps de survie des patients. Dans l'étude, la variable d'intérêt est le temps, en mois, entre le diagnostic et le décès du patient atteint de myeloma multiple.

Les données dans le tableau 3.1, obtenues par Krall, Uthoff et Harley (1975), sont reliées aux 48 patients, âgés entre 50 et 80 ans. Certains de ces patients, n'était pas décédés en fin d'étude, et ainsi ces individus contribuent des temps de survie censurés à droite. Le codage du statut de survie d'un individu, dans le tableau, est tel que, zéro désigne une observation censurée et une unité un décès de myeloma multiple. Au moment du diagnostic, les valeurs d'un nombre de variables explicatives étaient enregistrées pour chaque patient. Ceux-ci inclus :

- L'âge du patient en années ;
- Le sexe (0 = masculin, 1 = féminin) ;

– Les niveaux d'azote d'urée de sang (Bun) ;
– Le sérum de Calcium (Ca) ;
– Hémoglobine (Hb) ;
– Le pourcentage de cellules de plasma dans la moelle (Pcells) ;
– Une variable indicatrice (Protein), qui note si le Protéine Bence-Jones était présent ou non dans l'urine (0 = absent, 1 = présent).

Le but principal d'une analyse de ces données, est d'étudier l'effet des facteurs de risques Bun, Ca, Hb, Pcells et Protein sur le temps de survie des patients de myeloma multiple.

Dans cette illustration de calcul de l'estimation table de survie, l'information rassemblée sur les variables exogènes pour chaque individu sera ignorée.

Les temps de survie sont d'abord groupés pour donner le nombre de patients décédés d_j, et le nombre de ceux qui sont censurés, c_j, dans chacune des cinq première années de l'étude. Le nombre d'individus confrontés au risque de décès au début de chacun de ces intervalles, n_j, est alors calculé, en parallèle avec le nombre d'individus à risque ajusté $n_j^{'}$. Finalement, la probabilité de survie durant chaque intervalle est estimée, à partir duquel la fonction de survie estimée est obtenue en utilisant l'équation (3.3). Les calculs sont donnés dans le tableau 3.2, dans lequel la période de temps est enregistrée en mois, et l'intervalle qui commence en $t_k^{'}$ et se termine juste avant $t_k^{'}$, pour $k = 1, 2, \cdots, m$, est noté par $t_{k^-}^{'}$.

Patient	Temps de survie	Status	*Age*	*Sexe*	*Bun*	*Ca*	*Hb*	*Pcells*	*Protein*
1	13	1	66	1	25	10	14,6	18	1
2	52	0	66	1	13	11	12	100	0
3	6	1	53	2	15	13	11,4	33	1
4	40	1	69	1	10	10	10,2	30	1
5	10	1	65	1	20	10	13,2	66	0
6	7	0	57	2	12	8	9,9	45	0
7	66	1	52	1	21	10	12,8	11	1
8	10	0	60	1	41	9	14	70	1
9	10	1	70	1	37	12	7,5	47	0
10	14	1	70	1	40	11	10,6	27	0
11	16	1	68	1	39	10	11,2	41	0
12	4	1	50	2	172	9	10,1	46	1
13	65	1	59	1	28	9	6,6	66	0
14	5	1	60	1	13	10	9,7	25	0
15	11	0	66	2	25	9	8,8	23	0
16	10	1	51	2	12	9	9,6	80	0
17	15	0	55	1	14	9	13	8	0
18	5	1	67	2	26	8	10,4	49	0
19	76	0	60	1	12	12	14	9	0
20	56	0	66	1	18	11	12,5	90	0
21	88	1	63	1	21	9	14	42	1
22	24	1	67	1	10	10	12,4	44	0
23	51	1	60	2	10	10	10,1	45	1
24	4	1	74	1	48	9	6,5	54	0
25	40	0	72	1	57	9	12,8	28	1
26	8	1	55	1	53	12	8,2	55	0
27	18	1	51	1	12	15	14,4	100	0
28	5	1	70	2	130	8	10,2	23	0
29	16	1	53	1	17	9	10	28	0
30	50	1	74	1	37	13	7,7	11	1
31	40	1	70	2	14	9	5	22	0
32	1	1	67	1	165	10	9,4	90	0
33	36	1	63	1	40	9	11	16	1
34	5	1	77	1	23	8	9	29	0
35	10	1	61	1	13	10	14	19	0
36	91	1	58	2	27	11	11	26	1
37	18	0	69	2	21	10	10,8	33	0
38	1	1	57	1	20	9	5,1	100	1
39	18	0	59	2	21	10	13	100	0
40	6	1	61	2	11	10	5,1	100	0
41	1	1	75	1	56	12	11,3	18	0
42	23	1	56	2	20	9	14,6	3	0
43	15	1	62	2	21	10	8,8	5	0
44	18	1	60	2	18	9	7,5	85	1
45	12	0	71	2	46	9	4,9	62	0
46	12	1	60	2	6	10	5,5	25	0
47	17	1	65	2	28	8	7,5	8	0
48	3	0	59	1	90	10	10,2	6	1

TAB. 3.1 – Temps de survie de 48 patients de myeloma multiple.

Intervalle	Période de temps	d_j	c_j	n_j	$n_j^{'}$	$\frac{n_j^{'}-d_j}{n_j^{'}}$	$S^*(t)$
1	0_	16	4	48	46.0	0.652	0.652
2	12_	10	4	28	26.0	0.615	0.401
3	24_	1	0	14	14.0	0.929	0.373
4	36_	3	1	13	12.5	0.760	0.283
5	48_	2	2	9	8.0	0.750	0.212
6	60_	4	1	5	4.5	0.111	0.024

TAB. 3.2 – Estimation table de survie de la fonction de survie pour les données de l'exemple 3.2.

Un graphe de l'estimation table de survie de la fonction survie est montré dans la figure 3.2.

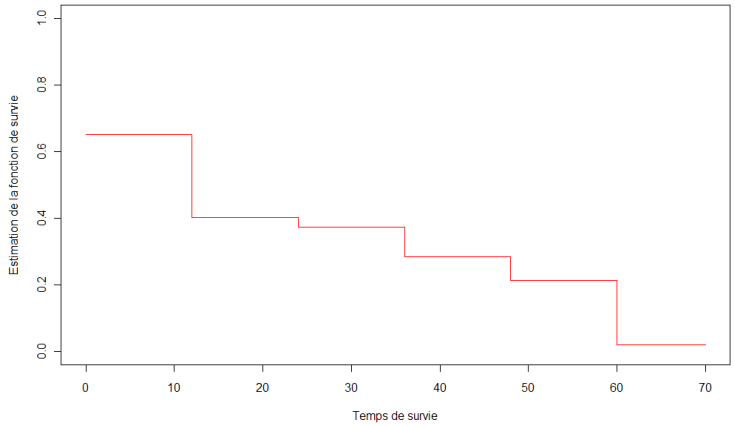

FIG. 3.2 – Estimation table de survie de la fonction de survie.

La forme de la fonction de survie estimée, obtenue avec cette méthode est sensible au choix des intervalles, employés dans sa construction. D'une autre part, l'estimation table de survie est particulièrement appropriée aux situations dans lesquelles les temps de décès réels sont inconnues, et l'unique information disponible est le nombre de décès et le nombre des observations censurées, qui se produisent dans une série d'intervalles de temps consécutives. Lorsque les temps de survie réels sont connues, l'estimation table de survie peut être employée, comme dans l'exemple 3.2, mais le groupement des temps de survie peut avoir comme conséquences, la perte d'une certaines information, particulièrement, lorsque le nombre des patients est petit, par exemple, inférieur à 30.

3.2.2 Estimation Kaplan-Meier

La première étape dans l'analyse des données de survie censurées non-groupées est de construire l'estimation Kaplan-Meier de la fonction de survie. Pour cela, une série d'intervalles de temps est construite, comme pour l'estimation table de survie. Cependant, chacun de ces intervalles est conçu, tel que, chaque intervalle contient un temps de décès où ce dernier se produit au début de l'intervalle.

Comme une illustration, supposons que $t_{(1)}, t_{(2)}$ et $t_{(3)}$, sont trois temps de survie observées, arrangées selon l'ordre croissant, tels que, $t_{(1)} < t_{(2)} < t_{(3)}$, et que C est un temps de survie censurées situé entre $t_{(2)}$ et $t_{(3)}$. Alors, les intervalles construit débutent respectivement en $t_{(1)}, t_{(2)}$ et $t_{(3)}$, et chaque intervalle inclus un temps de décès, bien qu'il pourrait y avoir plus qu'un individu, qui décède en un temps de décès quelconque. Notons qu'il n'existe pas d'intervalle qui débute en un temps censuré C. La situation est illustrée schématiquement dans Figure 3.3, dans laquelle D représente un décès et C un temps de survie censuré. Notons que, deux individus décèdent en $t_{(1)}$, un décède en $t_{(2)}$, et trois décèdent en $t_{(3)}$.

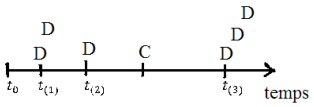

FIG. 3.3 – Construction des intervalles correspondant à l'estimation Kaplan-Meier.

Le temps d'origine est noté par $t_{(0)}$, et ainsi, il y a une période initiale qui débute en $t_{(0)}$ et se termine juste avant $t_{(1)}$, le temps du premier décès. Cela signifie que, l'intervalle qui s'étend de $t_{(0)}$ à $t_{(1)}$, n'inclura pas un temps de décès. Le premier intervalle construit débute en $t_{(1)}$ et se termine juste avant $t_{(2)}$, le second temps de décès ; cet intervalle inclus un seul temps de décès en $t_{(1)}$. Le second intervalle débute en $t_{(2)}$ et se termine juste avant $t_{(3)}$, et inclus le temps de décès en $t_{(2)}$ et le temps censuré C, et ainsi de suite...

Dans le cas général, supposons qu'il y a n individus avec des temps de survie observés $t_{(1)}, t_{(2)}, \cdots, t_{(n)}$. Nous supposons également, qu'il y a r temps de décès parmi les individus, où $r \leq n$. Après l'arrangement de ces temps de décès selon l'ordre croissant, le j-ème est noté t_j, pour $j = \overline{1,r}$, et ainsi les r temps de décès ordonnées sont $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Le nombre d'individus qui sont encore vivants juste avant $t_{(j)}$ et exposé au risque de décès à cet instant, sera noté n_j, et d_j désignera le nombre de décès en ce temps.

Il arrive souvent que les temps de survie censurés se produisent simultanément, avec un ou plusieurs décès. Dans ce cas, le temps de survie censuré est réalisé juste après le temps de décès lors du calcul des valeurs de n_j.

Sous l'hypothèse que les décès des individus dans l'échantillon se produisent de façons indépendantes. Alors, la fonction de survie estimée en un temps t quelconque, dans le k-ème intervalle de temps construit, de $t_{(k)}$ à $t_{(k+1)}$, $k = \overline{1,r}$, où $t_{(r+1)}$ est infini, sera la probabilité estimée de survivre au-delà de $t_{(k)}$. Il s'agit de la probabilité de survivre durant l'intervalle de $t_{(k)}$ à $t_{(k+1)}$ et tous les intervalles précédents, d'où l'estimation de Kaplan-Meier de la fonction de survie donnée par

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right). \tag{3.4}$$

Pour $t_{(k)} \leq t < t_{(k+1)}$, $k = \overline{1,r}$, avec $\hat{S}(t) = 1$ pour $t < t_{(1)}$, et où $t_{(r+1)}$ est infini. Si la plus grande observation est un temps de survie censuré, noté t^*, $\hat{S}(t)$ est non-définie pour $t > t^*$. D'une autre part, si le plus grand temps de survie observé, $t_{(r)}$, est une observation non-censurée, $n_r = d_r$, et ainsi $\hat{S}(t)$ est zéro pour $t \geq t_{(r)}$. Le graphe de l'estimation Kaplan-Meier de la fonction de survie est une fonction en escalier, dans laquelle les probabilités de survie estimées sont constantes entre deux temps de décès adjacents et décroit à chaque temps de décès.

L'équation (3.4) montre que, comme pour l'estimation table de survie de la fonction de survie dans l'équation (3.3), l'estimation Kaplan-Meier est exprimée sous la forme d'un produit d'une série de probabilités estimées.

En réalité, l'estimation Kaplan-Meier est la valeur limite de l'estimation table de survie dans l'équation (3.3), lorsque le nombre d'intervalles tend vers l'infini et leurs largeur tend vers zéro. Pour cette raison, l'estimation Kaplan-Meier est aussi connue sous le nom de l'estimation produit limite de la fonction de survie.

En absence de censure, $\hat{S}(t)$ est simplement la fonction de survie empirique définie dans l'équation (3.1). L'estimation Kaplan-Meier est donc une généralisation de la fonction de survie empirique, qui s'adapte aux observations censurées.

La méthode de Kaplan-Meier repose sur les mêmes principes que celle de la méthode actuarielle, à une différence que les intervalles de temps sont déterminés a posteriori par les temps de décès observés.

Exemple 3.3 Pour favoriser la recherche des méthodes d'analyse du saignement menstruel des femmes dans des essais contraceptifs, l'Organisation Mondiale de la Santé (World Health Organisation) a mis en disponibilité les données des essais thérapeutiques, impliquant un nombre de différents types de contraceptifs (WHO, 1987). Une partie de cet ensemble de données est reliée au temps, dans lequel une femme commence à utiliser une méthode particulière jusqu'à la discontinuation est enregistrée une fois connue. Les données dans le tableau 3.3 se réfère au nombre de jours, du début de l'utilisation d'un type particulier intrauterine device (IUD), jusqu'à la discontinuation, en raison des problèmes du saignement menstruel. Les données sont présentées pour 18 femmes, âgées entre 18 et 35 ans et qui ont expérimentées deux grossesses précédentes. Les temps de discontinuation censurées sont marquées avec "*".

Pour ces données, la fonction de survie, $S(t)$, représente la probabilité qu'une femme arrête d'utiliser le dispositif contraceptif après un certain temps t. L'estimation Kaplan-Meier de la fonction de survie est obtenue en utilisant l'équation (3.4), et les calculs exigés sont données dans le tableau 3.4. La fonction de survie estimée est tracée dans la figure 3.4.

10	13*	18*	19	23*	30	36	38*	54*
56*	59	75	93	97	104*	107	107*	107*

TAB. 3.3 – Temps, en jours, à la discontinuation d'utilisation d'un IUD.

Intervalle de temps	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
10_	18	1	0.944	0.944
19_	15	1	0.933	0.881
30_	13	1	0.923	0.814
36_	12	1	0.917	0.746
59_	8	1	0.875	0.653
75_	7	1	0.857	0.559
93_	6	1	0.833	0.466
97_	5	1	0.800	0.373
107	3	1	0.667	0.249

TAB. 3.4 – Estimation Kaplan-Meier de la fonction de survie pour les données de l'exemple 3.3.

Notons que, lorsque le plus grand temps d'arrêt de 107 jours est censuré, $\hat{S}(t)$ n'est pas définie au-delà de $t = 107$.

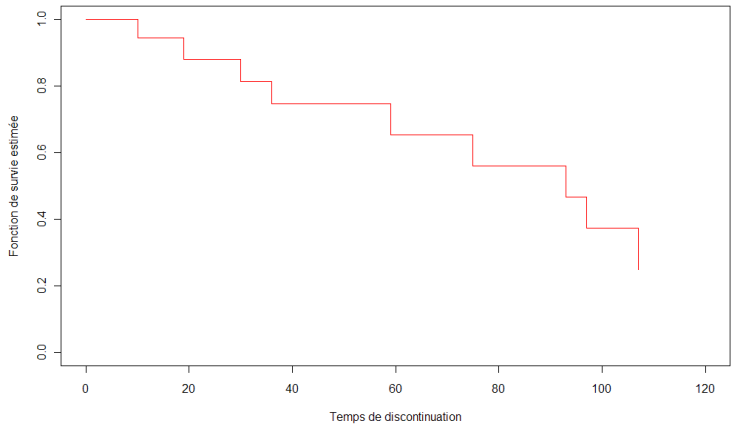

FIG. 3.4 – Estimation Kaplan-Meier de la fonction de survie pour les données de l'exemple 3.3.

3.2.3 Estimation Nelson-Aalen

Une estimation alternative de la fonction de survie, est l'estimation Nelson-Aalen donnée par

$$\widetilde{S}(t) = \prod_{j=1}^{k} \left(-\frac{d_j}{n_j} \right). \tag{3.5}$$

L'estimation Kaplan-Meier de la fonction de survie peut être considérée comme une approximation de

l'estimation Nelson-Aalen. Pour montrer ceci, nous utilisons le résultat suivant

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots$$

Qui est approximativement égale à $1 - x$ pour x petit.

Par analogie, $\exp(-d_j/n_j) \approx 1 - (d_j/n_j) = (n_j - d_j)/n_j$, à condition que d_j soit relativement petit par rapport à n_j.

Par conséquent, l'estimation Kaplan-Meier, $\hat{S}(t)$, dans l'équation (3.4), approxime l'estimation Nelson-Aalen, dans l'équation (3.5).

L'estimation Nelson-Aalen de la fonction de survie est toujours plus grande que celle de Kaplan-Meier, pour tout t donné, avec $e^{-x} \geq 1 - x$, et pour toutes les valeurs de x.

Exemple 3.4 Les valeurs montrées dans le tableau 3.4, qui donnent l'estimation Kaplan-Meier de la fonction de survie pour les données sur le temps de discontinuation d'utilisation d'un IUD, peuvent être utilisé pour calculer l'estimation Nelson-Aalen. Cette estimation est montrée dans le tableau 3.5.

Intervalle de temps	$\exp(-d_j/n_j)$	$\widetilde{S}(t)$
0_	1.000	1.000
10_	0.946	0.946
19_	0.935	0.885
30_	0.926	0.819
36_	0.920	0.754
59_	0.882	0.665
75_	0.867	0.579
93_	0.846	0.488
97_	0.819	0.400
107	0.716	0.286

TAB. 3.5 – Estimation Nelson-Aalen de la fonction de survie pour les données de l'exemple 3.3.

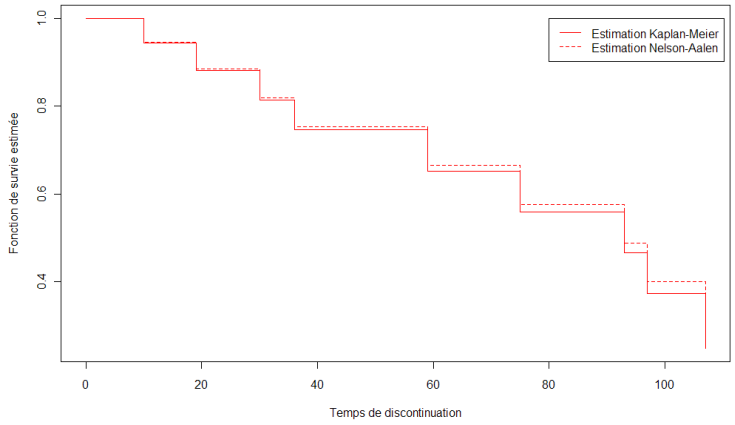

FIG. 3.5 – Estimation Nelson-Aalen et estimation Kaplan-Meier de la fonction de survie pour les données de l'exemple 3.3.

3.3 Erreur standard de la fonction de survie estimée

Dans cette section, nous donnons l'erreur standard des estimations de la fonction de survie. Comme la méthode de Kaplan-Meier est la plus employée dans l'estimation de la fonction de survie, la dérivation de l'erreur standard de $\hat{S}(t)$ sera présentée en détail.

3.3.1 Erreur standard de l'estimation Kaplan-Meier

L'estimation Kaplan-Meier de la fonction de survie pour toute valeur de t dans l'intervalle, qui s'étend de $t_{(k)}$ à $t_{(k+1)}$, peut s'écrire sous la forme

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j,$$

Pour $j = 1, 2, \cdots, r$, où $\hat{p}_j = (n_j - d_j)/n_j$ est la probabilité estimée qu'un individu soit encore vivant durant l'intervalle de temps, qui débute en $t_{(j)}$, $j = \overline{1,r}$.

Prenons le logarithme,

$$\log \hat{S}(t) = \sum_{j=1}^{k} \log \hat{p}_j$$

et ainsi la variance de $\log \hat{S}(t)$ est donnée par

$$V\{\log \hat{S}(t)\} = \sum_{j=1}^{k} V\{\log \hat{p}_j\} \tag{3.6}$$

Le nombre d'individus qui sont vivants durant l'intervalle, débutant en $t_{(j)}$, peut être supposé, ayant une distribution binomiale avec les paramètres n_j et p_j, où p_j est la vraie probabilité de survivre durant cet intervalle. Le nombre de survivants observés est $n_j - d_j$, et en utilisant le résultat que la variance d'une variable aléatoire binomiale avec les paramètres n, p est $np(1-p)$, la variance de $n_j - d_j$ est donnée par

$$V(n_j - d_j) = n_j p_j (1 - p_j)$$

Comme $\hat{p}_j = (n_j - d_j)/n_j$, la variance de $\hat{p}_j$ est $V(n_j - d_j)/n_j^2$, qui est égale à $p_j(1-p_j)/n_j$. La variance de $\hat{p}_j$ peut être donc estimée par

$$\hat{p}_j(1-\hat{p}_j)/n_j \tag{3.7}$$

Dans le but d'obtenir la variance de $\log \hat{p}_j$, nous utilisons un résultat général pour l'approximation de la variance d'une fonction à une variable aléatoire. D'après ce résultat, la variance d'une fonction $g(X)$ de la variable aléatoire X est donnée par

$$V\{g(X)\} = \left\{ \frac{dg(X)}{dX} \right\}^2 V(X) \tag{3.8}$$

Ceci est connu comme l'approximation des séries de Taylor de la variance d'une fonction à une variable aléatoire. En utilisant l'équation (3.8), la variance approximée de $\log \hat{p}_j$ est $V(\hat{p}_j)/\hat{p}_j^2$, et en utilisant l'expression (3.7), l'approximation de la variance estimée de $\log \hat{p}_j$ est $(1-\hat{p}_j)/n_j\hat{p}_j$, ce qui est équivalent à

$$\frac{d_j}{n_j(n_j - d_j)} \tag{3.9}$$

À partir de l'équation (3.6),

$$V\{\log \hat{S}(t)\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \tag{3.10}$$

Une seconde application du résultat dans l'équation (3.8) donne

$$V\{\log \hat{S}(t)\} \approx \frac{1}{[\hat{S}(t)]^2} . V\{\hat{S}(t)\},$$

Ainsi que,

$$V\{\hat{S}(t)\} \approx [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \tag{3.11}$$

Finalement, l'erreur standard de l'estimation Kaplan-Meier de la fonction de survie est donnée par

$$se\{\hat{S}(t)\} \approx \hat{S}(t) . \left\{ \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \right\}^{\frac{1}{2}} \tag{3.12}$$

pour $t_{(k)} \leq t < t_{(k+1)}$. Ce résultat est connu sous le nom de *Formule de Greenwood*.

3.3.2 Erreur standard des estimations table de survie et Nelson-Aalen

L'estimation table de survie de la fonction de survie est similaire, dans la forme, a l'estimation Kaplan-Meier, et ainsi l'erreur standard de cet estimateur est obtenu de façon similaire. L'erreur standard de l'estimation table de survie est donnée par

$$se\{S^*(t)\} \approx S^*(t).\left\{\sum_{j=1}^{k}\frac{d_j}{n_j^{'}(n_j^{'}-d_j)}\right\}^{\frac{1}{2}} \tag{3.13}$$

L'erreur standard de l'estimateur Nelson-Aalen est

$$se\{\widetilde{S}(t)\} \approx \widetilde{S}(t).\left\{\sum_{j=1}^{k}\frac{d_j}{n_j^2}\right\}^{\frac{1}{2}} \tag{3.14}$$

3.3.3 Intervalles de confiance pour des valeurs de la fonction de survie

Une fois que, l'erreur standard d'une estimation de la fonction de survie est calculée, un intervalle de confiance pour la valeur correspondante de la fonction de survie à un temps t donné, peut être déterminé.

Un intervalle de confiance pour la vraie valeur de la fonction de survie $S(t)$ est obtenu, en supposant que la valeur estimée de la fonction de survie en t est asymptotiquement distribuée selon une loi normale, de moyenne $\hat{S}(t)$ et de variance estimée donnée par l'équation (3.11).

D'où, l'intervalle de confiance de niveau asymptotique $1-\alpha\ \ (0<\alpha<1)$ pour la fonction de survie est :

$$[\widehat{S}(t)-U_{1-\frac{\alpha}{2}}se\{\hat{S}(t)\},\widehat{S}(t)+U_{1-\frac{\alpha}{2}}se\{\hat{S}(t)\}]$$

où $se\{\hat{S}(t)\}$ est l'erreur standard de la fonction de survie estimée, déterminée par l'équation (3.12) et $U_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $N(0,1)$. Cet intervalle peut être superposé au graphe de la fonction de survie estimée, ainsi montrer dans la figure 3.6.

Exemple 3.5 L'erreur standard de la fonction de survie estimée $\hat{S}(t)$, et les bornes inférieure et supérieure d'un intervalle de confiance à 95% pour la vraie valeur de $S(t)$, associées aux données de l'exemple 3.3 sur les temps de discontinuation de l'utilisation d'un IUD, sont données dans le tableau 3.6. Un graphe de la fonction de survie estimée, avec les bornes de confiance à 95%, sont présentés dans la figure 3.6.

Intervalle de temps	$\hat{S}(t)$	$se\{\hat{S}(t)\}$	Intervalle de confiance à 95%
0_	1.000	0.000	
10_	0.944	0.054	$(0.839, 1.000)$
19_	0.881	0.079	$(0.727, 1.000)$
30_	0.814	0.098	$(0.622, 1.000)$
36_	0.746	0.111	$(0.529, 0.963)$
59_	0.653	0.130	$(0.397, 0.908)$
75_	0.559	0.141	$(0.283, 0.836)$
93_	0.466	0.145	$(0.182, 0.751)$
97_	0.373	0.143	$(0.093, 0.653)$
107	0.249	0.139	$(0.000, 0.522)$

TAB. 3.6 – Erreur standard de $\hat{S}(t)$ et intervalles de confiance à 95% pour $S(t)$ associées aux données de l'exemple 3.3.

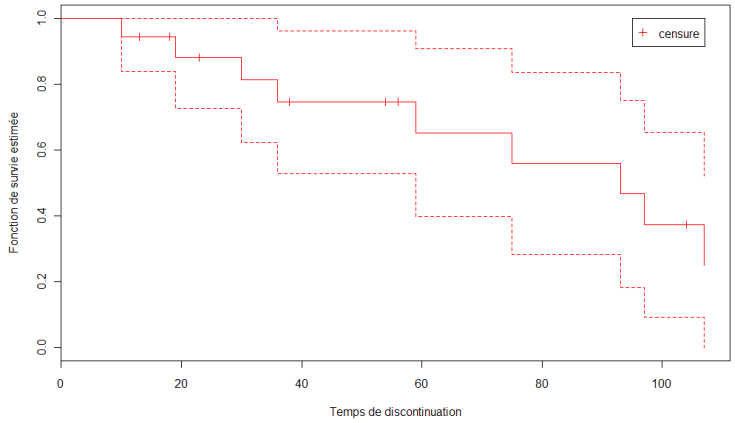

FIG. 3.6 – Fonction de survie estimée et les bornes de confiance à 95% pour $S(t)$.

3.4 Comparaison de fonctions de survie : test Log-rank

Si on désire comparer l'évolution de deux ou plusieurs groupes d'individus, on pourrait comparer les pourcentages de décès survenant dans chacun de ces groupes ; ou encore comparer les taux de survie à un instant donné. Ces solutions ne permettent pas de tenir compte des moments auxquels les décès se produisent. Le test qui permet de tenir compte respectivement du nombre et du temps de décès est le *test Log-rank*.

Supposons qu'il y a deux groupes et notons $t_{(j)}$ le j-ème temps de décès ordonné dans les deux groupes combinés.

Construisons un tableau de contingence pour le j-ème temps de décès, $j = \overline{1,r}$ (voir le tableau 3.7).

Groupes	Nombre de décès en $t_{(j)}$	Nombre de survivants au-delà de $t_{(j)}$	Nombre d'individus à risque en $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

TAB. 3.7 – Tableau de contingence correspondant au j-ème temps de décès.

où :

d_{ij} est le nombre d'individus décédés à l'instant $t_{(j)}$ dans le groupe i ;

n_{ij} est le nombre d'individus à risque au j-ème temps de décès dans le groupe i ;

Par conséquent, il y'a en total $d_j = d_{1j} + d_{2j}$ décès et $n_j = n_{1j} + n_{2j}$ individus à risque, à l'instant $t_{(j)}$.

Sous l'hypothèse nulle que les fonctions de survie sont identiques dans les deux groupes, le nombre attendu de décès en $t_{(j)}$ dans le groupe i, $i = \overline{1,2}$ est estimé par

$$\hat{e}_{ij} = \frac{n_{ij} d_j}{n_j} \quad pour \quad j = \overline{1,r} \tag{3.15}$$

La variance de $\hat{e}_{ij}$ peut être estimée par

$$\hat{V}(\hat{e}_{ij}) = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \tag{3.16}$$

Pour chaque temps de décès $t_{(1)}, t_{(2)}, \cdots, t_{(r)}$, un tableau de contingence est construit, de façon similaire au tableau 3.7.

Le test Log-rank est basé sur la statistique suivante

$$W_L = \frac{(\sum_{j=1}^{r} d_{1j} - \sum_{j=1}^{r} \hat{e}_{1j})^2}{\sum_{j=1}^{r} \hat{V}(\hat{e}_{ij})} \tag{3.17}$$

Cette statistique est distribuée selon une Khi deux à un degré de liberté, sous l'hypothèse nulle. Une illustration de ce test est présentée ci-dessous dans l'exemple 3.6.

Ce test peut se généralisé également pour comparer plusieurs groupes de données de survie ($i \geq 2$).

Exemple 3.6 Il s'agit de la comparaison de deux traitements contre un cancer de sein : un essai thérapeutique a été mené chez des femmes atteintes d'un cancer de sein, assignés aléatoirement à deux groupes, l'un traité par A, l'autre traité par B. Les temps de survie, en mois, selon le traitement sont donnés dans le tableau 3.8.

Groupe A	Groupe B	
23	5	68
47	8	71
69	10	76*
70*	13	105*
71*	18	107*
100*	24	109*
101*	26	113
148	26	116*
181	31	118
198*	35	143
208*	40	154*
212*	41	162*
224*	48	188*
	50	212*
	59	217*
	61	225*

TAB. 3.8 – Temps de survie de femmes atteintes d'un cancer de sein.

L'estimation de Kaplan-Meier de la fonction de survie, pour chacun des deux groupes, est tracée dans la figure 3.7. Notons que dans cette figure, les estimations se prolongent au temps de la plus grande observation censurée dans chaque groupe.

La figure 3.7 montre que la fonction de survie estimée pour les femmes atteintes d'un cancer de sein, traitées avec A est toujours plus élevée que celle des femmes traitées avec B.

En utilisant le *test Log-rank*, nous examinons l'hypothèse nulle qu'il n'y a pas de différence dans l'expérience de survie des deux groupes. Les calcules exigés sont présentés dans le tableau 3.9.

Nous commençons d'abord par ordonner les temps de décès observés dans les deux groupes. Ces temps sont donnés dans la colonne1 du tableau 3.9.

Le nombre de décès et le nombre de femmes à risque, à chaque temps de décès sont calculés pour chaque groupe. Ces valeurs sont d_{1j}, n_{1j}, d_{2j} et n_{2j} données dans les colonnes 2 à 5 du tableau 3.9. Les colonnes 6 et 7 contiennent le nombre total de décès et de femmes à risque dans les deux groupes, calculés à chaque temps de décès. Les deux dernières colonnes donnent les valeurs de $\hat{e}_{1j}$ et $\hat{V}(\hat{e}_{1j})$, calculées respectivement, à partir des équations (3.15) et (3.16).

Ainsi, la statistique du test Log-rank prenant la valeur $W_L = 3.515$ confirme une différence significative entre les fonctions de survie des deux groupes avec une p-value de 0.061 environ.

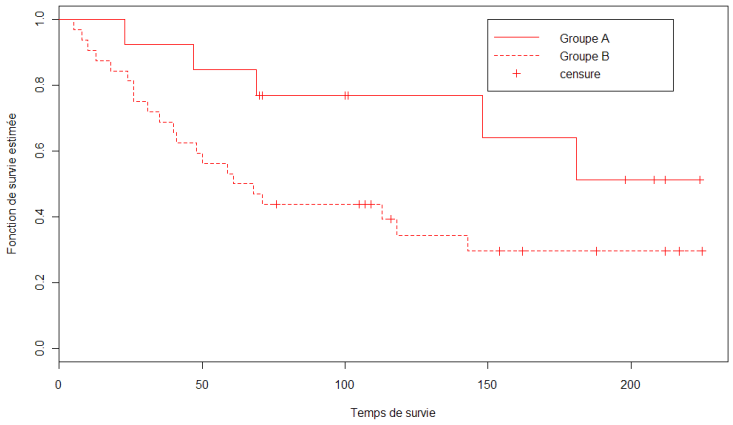

FIG. 3.7 – Comparaison de fonctions de survie correspondant à l'exemple 3.6.

Temps de décès	d_{1j}	n_{1j}	d_{2j}	n_{2j}	d_j	n_j	$\hat{e}_{ij}$	$\hat{V}(\hat{e}_{ij})$
5	0	13	1	32	1	45	0.2889	0.2054
8	0	13	1	31	1	44	0.2955	0.2082
10	0	13	1	30	1	43	0.3023	0.2109
13	0	13	1	29	1	42	0.3095	0.2137
18	0	13	1	28	1	41	0.3171	0.2165
23	1	13	0	27	1	40	0.3250	0.2194
24	0	12	1	27	1	39	0.3077	0.2130
26	0	12	2	26	2	38	0.6316	0.4205
31	0	12	1	24	1	36	0.3333	0.2222
35	0	12	1	23	1	35	0.3429	0.2253
40	0	12	1	22	1	34	0.3529	0.2284
41	0	12	1	21	1	33	0.3636	0.2314
47	1	12	0	20	1	32	0.3750	0.2344
48	0	11	1	20	1	31	0.3548	0.2289
50	0	11	1	19	1	30	0.3667	0.2322
59	0	11	1	18	1	29	0.3793	0.2354
61	0	11	1	17	1	28	0.3929	0.2385
68	0	11	1	16	1	27	0.4074	0.2414
69	1	11	0	15	1	26	0.4231	0.2441
71	0	9	1	15	1	24	0.3750	0.2344
113	0	6	1	10	1	16	0.3750	0.2344
118	0	6	1	8	1	14	0.4286	0.2449
143	0	6	1	7	1	13	0.4615	0.2485
148	1	6	0	6	1	12	0.5000	0.2500
181	1	5	0	4	1	9	0.5556	0.2469
Total	5						9.5652	5.9289

TAB. 3.9 – Calcul de la statistique Log-rank pour les données de l'exemple 3.6.

3.5 Conclusion

Dans le cas où les temps de survie observés sont non censurés, la fonction de survie empirique est employée pour évaluer la fonction de survie. Cependant, il existe des approches non-paramétriques pour estimer cette dernière qui peuvent être utilisées en présence de censures à savoir, estimation table de survie, estimation Kaplan-Meier et estimation Nelson-Aalen.

La méthode de Kaplan-Meier repose sur les mêmes principes que celle de la méthode table de survie en divisant la période d'observation en une série d'intervalles de temps, à une différence que ces intervalles sont déterminés a posteriori par les temps de décès observés. En outre, elle est la valeur limite de l'estimation table de survie lorsque le nombre d'intervalles tend vers l'infini et leur largeur tend vers zéro et elle peut être considérée comme une approximation de l'estimation Nelson-Aalen.

4

Modèles de régression en analyse de survie

Sommaire

4.1 Introduction

Les techniques non paramétriques, ainsi décrites dans le chapitre 3, peuvent être utiles lorsqu'il s'agit de l'analyse d'un échantillon simple de données de survie, ou bien pour la comparaison de deux ou plusieurs groupes de temps de survie. Cependant, dans la majorité des études, que ce soit médicale, économique, ou autres... ils peuvent avoir des informations supplémentaires enregistrées pour chaque individu. Ces informations sont structurées sous forme de variables exogènes.

L'objectif de ce chapitre est d'explorer la relation entre l'expérience de survie et les variables exogènes.

Dans ce contexte, plusieurs approches basées sur la modélisation statistiques, peuvent être utilisées : modèle de Cox [4], modèle à hasards proportionnels paramétrique [5][13] et le modèle temps de survie accéléré [11].

4.2 Le modèle semi-paramétrique : modèle de Cox

4.2.1 Présentation générale

Le modèle le plus utilisé pour évaluer l'effet de variables exogènes ou covariables telles, le sexe, l'âge, le nombre des cigarettes fumées par jour...etc. sur la survie d'un individu est dit modèle à hasards proportionnels de Cox ou tout simplement modèle de Cox.

Ce modèle permet la prise en compte simultanée de plusieurs variables pour expliquer la survenue d'un événement, sans avoir à exiger une forme particulière à la distribution des temps de survie, d'où le nom modèle semi-paramétrique. Il exprime la fonction de hasard d'un individu i ayant un vecteur de p variables exogènes $x_i = (x_{1i}, x_{2i}, \cdots, x_{pi})$ sous une forme multiplicative

$$h_i(t) = \exp(\beta' x_i) h_0(t) \tag{4.1}$$

C'est-à-dire le produit d'une fonction de hasard de base $h_0(t)$ correspondant à la fonction de hasard des individus pour lesquels toutes les variables x_i sont nulles, et d'une fonction de régression explicitée paramétriquement, $\exp(\beta' x_i)$, où $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$ est le vecteur des coefficients de régression inconnus.

Le modèle de Cox sous-tend deux hypothèses :

– **Hypothèse log-linéarité des variables :**

Il existe une relation log-linéaire entre fonction de hasard et covariables :

$$\log \frac{h_i(t)}{h_0(t)} = (\beta' x_i)$$

– **Hypothèse de proportionnalité des hasards :**

Le rapport des fonctions de hasard pour deux (02) individus i et j de caractéristiques x_i et x_j est indépendant du temps et ne dépend que de x_i et x_j :

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp(\beta' x_i)}{\exp(\beta' x_j)} = \exp(\beta'(x_i - x_j)) = K$$

C'est-à-dire que les fonctions de hasard des deux (02) individus i et j sont proportionnelles, et que leur rapport de proportionnalité ne dépend pas du temps. C'est le rapport de hasard (noté HR).

4.2.2 Estimation des β-paramètres du modèle de Cox

Les β-coefficients du modèle de Cox peuvent être estimés, en utilisant la méthode du maximum de vraisemblance.

Supposons que les données de survie sont disponibles pour n individus dont r temps de décès distinct et $n-r$ temps de survie censurés à droite. Nous supposons également qu'il y a un seul individu qui décède à chaque temps de décès.

Les r temps de décès ordonnées seront notés par $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ ainsi que $t_{(j)}$ est le j-ème temps de décès ordonnées.

Considérons la probabilité que le i-ème individu décède en $t_{(j)}$, sachant que ce temps appartient à l'ensemble des r temps de décès observés $t_{(1)}, t_{(2)}, \cdots, t_{(r)}$. Si le vecteur des variables exogènes associé à l'individu qui décède en $t_{(j)}$ est noté par $x_{(j)}$, cette probabilité est

$$\texttt{Prob (l'individu de variables } x_{(j)} \texttt{ décède en } t_{(j)} \texttt{/qu'il y a eu un seul décès en } t_{(j)}) \quad (4.2)$$

Cette probabilité peut s'écrire aussi comme

$$\frac{\texttt{P(l'individu de variables } x_{(j)} \texttt{ décède en } t_{(j)})}{\texttt{P(qu'il y a eu un seul décès en } t_{(j)})} \quad (4.3)$$

Sachant que les temps de décès sont indépendant, le dénominateur de cette expression est la somme des probabilités de décès en $t_{(j)}$, pour tout les individus qui sont exposés au risque de décéder en ce temps. Si ces individus sont indexés par l, avec $R(t_{(j)}$ exprimant l'ensemble des individus à risque en $t_{(j)}$, c'est-à-dire ceux qui sont encore vivant juste avant ce temps, l'expression 4.3 devient

$$\frac{\texttt{P(l'individu de variables } x_{(j)} \texttt{ décède en } t_{(j)})}{\sum_{l \in R(t_{(j)})} \texttt{P(l'individu } l \texttt{décède en } t_{(j)})} \quad (4.4)$$

Maintenant, les probabilités de décéder en $t_{(j)}$ dans l'expression (4.4) sont remplacées par les probabilités de décéder durant l'intervalle $[t_{(j)},\ t_{(j)}+\triangle t[$, et en divisant le numérateur et dénominateur de l'expression (4.4) par $\triangle t$, nous obtenons

$$\frac{\texttt{P(l'individu de variables } x_{(j)} \texttt{ décède durant l'intervalle } [t_{(j)},\ t_{(j)}+\triangle t[)/\triangle t}{\sum_{l\in R(t_{(j)})} \texttt{P(l'individu } l \texttt{ décède durant l'intervalle } [t_{(j)},\ t_{(j)}+\triangle t[)/\triangle t}$$

La valeur limite de cette expression quand $\triangle t \to 0$ est donc le rapport des probabilités dans l'expression (4.4).

Cette limite est aussi le rapport des hasards de décéder en $t_{(j)}$

$$\frac{\texttt{Hasard de décéder en } t(j) \texttt{ pour l'individu de variables } x_{(j)}}{\sum_{l\in R(t_{(j)})} \texttt{ Hasard de décéder en } t_{(j)} \texttt{ pour l'individu } l}$$

Si le ième individu décède en $t_{(j)}$, la fonction de hasard dans le numérateur de cette expression peut s'écrire comme $h_i(t_{(j)})$. De même, le dénominateur est la somme des hasards de décéder en $t_{(j)}$.

Par conséquent, la probabilité conditionnelle dans l'expression (4.2) devient

$$\frac{h_i(t_{(j)})}{\sum_{l\in R(t_{(j)})} h_l(t_{(j)})}$$

En utilisant l'équation (4.1), cette expression peut être exprimée par

$$\frac{\exp(\beta' x_{(j)})}{\sum_{l\in R(t_{(j)})} \exp(\beta' x_l)}$$

Pour estimer les β-paramètres du modèle de Cox, on calcul la fonction de vraisemblance partielle donnée par

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' x_{(j)})}{\sum_{l\in R(t_{(j)})} \exp(\beta' x_l)} \tag{4.5}$$

Supposons que les données se composent de n temps de survie, $t_1, t_2, \cdots, t_n$, et que δ_i est l'indicatrice d'évènement, qui prend zéro si le i-ème temps de survie t_i est censuré à droite, et une unité sinon. La fonction de vraisemblance partielle dans l'équation (4.5) peut s'écrire aussi sous la forme

$$\prod_{j=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum_{l\in R(t_i)} \exp(\beta' x_l)} \right)^{\delta_i}$$

où $R(t_i)$ est l'ensemble des individus à risque en t_i.

La fonction log-vraisemblance correspondante est données par

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i (\beta' x_i - \log \sum_{l \in R(t_i)} \exp(\beta' x_l)) \tag{4.6}$$

Pour déterminer les estimations du maximum de vraisemblance des β-coefficients de ce modèle, on procède à une maximisation de cette fonction en employant des méthodes numériques. Cette maximisation est généralement accomplie par la méthode de Newton-Raphson.

Exemple 4.1 Afin de mieux comprendre la structure de la vraisemblance partielle, considérons un échantillon de données de survie constituées de cinq individus, numérotées de 1 à 5. Les données de survie sont illustrées dans la figure 4.1.

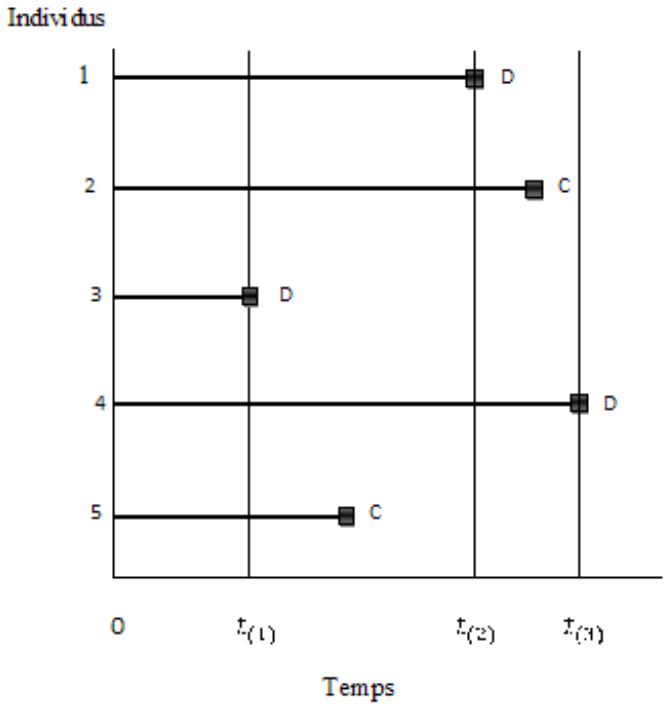

FIG. 4.1 – Temps de survie de cinq individus.

Les temps de survie des individus 2 et 5 seront considérés comme censurés à droite, et les trois temps de décès ordonnés sont notés $t_{(1)} < t_{(2)} < t_{(3)}$. Alors, $t_{(1)}$ est le temps de décès de l'individu 3, $t_{(2)}$ est celui de l'individu 1, et $t_{(3)}$ est celui de l'individu 4.

L'ensemble à risque pour chacun des trois temps de décès ordonnés est constitué des individus qui sont encore vivant juste avant, chaque temps de décès.

Par conséquent, l'ensemble à risque $R(t_{(1)})$ est constitué du total des cinq individus, l'ensemble à risque $R(t_{(2)})$ se compose des individus $1,2$ et 4, tandis que l'ensemble à risque $R(t_{(3)})$ inclut seulement l'individu 4.

Maintenant, nous pouvons écrire $\psi(i) = \exp(\beta' x_i)$, pour le score à risque du i-ème individu, où x_i est le vecteur des covariables associé à cet individu.

Les numérateurs de la fonction de vraisemblance partielle en $t_{(1)}, t_{(2)}$ et $t_{(3)}$, sont respectivement, $\psi(3), \psi(1)$ et $\psi(4)$, sachant que les individus $3, 1$ et 4 décèdent respectivement aux trois temps de décès ordonnées. La fonction de vraisemblance partielle en ces trois temps de décès est alors

$$\frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \times \frac{\psi(1)}{\psi(1)+\psi(2)+\psi(4)} \times \frac{\psi(4)}{\psi(4)}$$

4.2.3 Estimation des fonctions de survie et de hasard

Les β-paramètres du modèle à hasards proportionnels de Cox étant maintenant estimées, une estimation de la fonction de hasard pour le i-ème individu, $i = 1, 2, \cdots, n$, peut être donnée par

$$\hat{h}_i(t) = \exp(\hat{\beta}' x_i) h_0(t) \tag{4.7}$$

où x_i est le vecteur des valeurs des p covariables, associé au i-ème individu, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)$ est le vecteur des coefficients estimés, et $h_0(t)$ est la fonction de hasard de base estimée.

En utilisant cette équation, la fonction de hasard pour un individu peut être estimée une fois qu'une estimation de $h_0(t)$ est déterminée. La relation entre les fonctions de hasard, hasard cumulatives, et survie peut donc être employée pour donner des estimations de la fonction de hasard cumulative et la fonction de survie.

Une estimation de la fonction de hasard de base a été dérivée par Kalbfleisch et Prentice (1973) en utilisant une approche basée sur la méthode du maximum de vraisemblance. Supposons qu'il y a r temps de décès distinct, rangés selon l'ordre croissant, $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, et qu'il y a d_j décès et n_j individus à risque en $t_{(j)}$. La fonction de hasard de base estimée en $t_{(j)}$ est alors donnée par

$$\hat{h}_0(t_{(j)}) = 1 - \hat{\zeta}_j \tag{4.8}$$

où $\hat{\zeta}_j$ est la solution de l'équation

$$\sum_{l \in D(t_{(j)})} \frac{\exp(\hat{\beta}' x_l)}{1 - \hat{\zeta}_j^{\exp(\hat{\beta}' x_l)}} \quad \texttt{Pour} \quad j = \overline{1, r} \tag{4.9}$$

où $D(t_{(j)}$ est l'ensemble de tout les d_j individus qui décède au j-ème temps de décès $t_{(j)}$ et $R(t_{(j)})$ est l'ensemble de tout les n_j individus à risque en ce temps.

Les estimations des β-coefficients qui constituent le vecteur $\hat{\beta}$, sont ceux qui maximisent la fonction de vraisemblance dans l'équation (4.5). La détermination de cette estimation de $h_0(t)$ est complexe.

Dans le cas où il n'y a pas d'aequo des temps de décès, c'est-à-dire $d_j = 1$ pour $j = \overline{1,r}$, l'équation (4.9) peut être donc résolue pour donner

$$\hat{\zeta}_j = \left(1 - \frac{\exp(\hat{\beta}' x_{(j)})}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l)}\right)^{\exp(\hat{\beta}' x_{(j)})}$$

où $x_{(j)}$ est le vecteur des variables exogènes correspondant à l'individu qui décède en $t_{(j)}$.

Supposons que le hasard de décès est constant entre des temps de décès adjacents. Une estimation appropriée de la fonction de hasard de base dans cet intervalle est alors obtenant en divisant le hasard estimé dans l'équation (4.8) par la longueur de l'intervalle, pour donner la fonction en escalier

$$\hat{h}_0(t) = \frac{1 - \hat{\zeta}_j}{t_{(j+1)} - t_{(j)}} \tag{4.10}$$

Pour $t_{(j)} \leq t < t_{(j+1)},\ j = 1, 2, \cdots, (r-1)$, avec $\hat{h}_0(t) = 0$ pour $t < t_{(1)}$.

La quantité $\hat{\zeta}_j$ peut être considérée comme une estimation de la probabilité qu'un individu survive durant l'intervalle qui s'étend de $t_{(j)}$ à $t_{(j+1)}$.

La fonction de survie de base peut être donc estimée par

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\zeta}_j \tag{4.11}$$

Pour $t_{(k)} \leq t < t_{(k+1)},\ j = 1, 2, \cdots, (r-1)$, et ainsi cette estimation est aussi une fonction en escalier. La valeur estimée de la fonction de survie de base est une unité pour $t < t_{(1)}$, et zéro pour $t \geq t_{(r)}$, à moins qu'il y ait des temps de survie censurés plus grands que $t_{(r)}$. Si c'est le cas, $\hat{S}_0(t)$ sera égale à $\hat{S}_0(t_{(r)})$ jusqu'au plus grand temps censuré, mais la fonction de survie estimée ne sera pas définie au-delà de ce temps.

A partir de l'équation (2.4), la fonction de hasard cumulé de base est donnée par

$$H_0(t) = -\log S_0(t)$$

Et ainsi une estimation de cette fonction est

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = -\sum_{j=1}^{k} \log \hat{\zeta}_j \tag{4.12}$$

Pour $t_{(k)} \leq t < t_{(k+1)},\ j = 1, 2, \cdots, (r-1)$, avec $\hat{H}_0(t) = 0$ pour $t < t_{(1)}$.

Les estimations des fonctions de hasard, de survie et de hasard cumulé de base dans les équations (4.10), (4.11) et (4.12) peuvent être utilisées pour déterminer les estimations correspondantes pour un individu associé

au vecteur de covariables x_i.

En particulier, à partir de l'équation (4.7), la fonction de hasard est estimée par

$$\hat{h}_i(t) = \exp(\hat{\beta}' x_i)\hat{h}_0(t)$$

Ensuite, en intégrant les deux cotés de l'équation (4.7), nous obtenons

$$\int_0^t \hat{h}_i(u)du = \exp(\hat{\beta}' x_i) \int_0^t \hat{h}_0(u)du \tag{4.13}$$

Ainsi que la fonction de hasard cumulé estimée pour le i-ème individu est

$$\hat{H}_i(t) = \exp(\hat{\beta}' x_i)\hat{H}_0(t) \tag{4.14}$$

A partir de l'équation (4.14), nous pouvons donc déduire une estimation de la fonction de survie du i-ème individu, donnée par

$$\hat{S}_i(t) = \{\hat{S}_0(t)\}^{\exp(\hat{\beta}' x_i)} \tag{4.15}$$

Pour $t_{(k)} \leq t < t_{(k+1)},\ j = 1, 2, \cdots, (r-1)$.

– **Cas particulier : sans variables exogènes** [7]
En absence de variables exogènes, nous avons juste un échantillon simple de temps de survie, l'équation (4.9) devient

$$\frac{d_j}{1-\hat{\zeta}_j} = n_j$$

A partir de laquelle

$$\hat{\zeta}_j = \frac{n_j - d_j}{n_j}$$

D'après l'équation (4.8), la fonction de hasard de base estimée en $t_{(j)}$ est $1-\hat{\zeta}_j$, ce qui est égale à

$$\frac{d_j}{n_j}$$

A partir de l'équation (4.11), l'estimation correspondante de la fonction de survie est $\prod_{j=1}^{k} \hat{\zeta}_j$, c'est

$$\prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right)$$

Qui n'est rien autre que l'estimation Kaplan-Meier donnée par l'équation (3.4). Ceci montre que l'estimation de la fonction de survie donnée par l'équation (4.15) généralise l'estimation Kaplan- Meier au cas où la fonction de hasard dépend de variables exogènes.

– **Quelques approximations pour estimer les fonctions de base**

Dans le cas où il y a des aequo de temps de survie, la fonction de hasard de base estimée peut être seulement déterminée en utilisant une méthode itérative pour résoudre l'équation (4.9). Ce processus itératif peut être remplacé par une approximation du terme

$$\hat{\zeta}_j^{\exp(\hat{\beta}' x_l)}$$

de l'équation (4.9), qui peut s'écrire sous la forme

$$\exp\{e^{\hat{\beta}' x_l} \log \hat{\zeta}_j\} \approx 1 + e^{\hat{\beta}' x_l} \log \hat{\zeta}_j$$

Alors, $\hat{\zeta}_j$ peut être déterminé de l'équation (4.9) tel que

$$-\sum_{l \in D(t_{(j)})} \frac{1}{\log \hat{\zeta}_j} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l)$$

Par conséquent,

$$\frac{-d_j}{\log \hat{\zeta}_j} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l)$$

où d_j est le nombre de décès au j-ème temps de décès ordonnée, et ainsi

$$\hat{\zeta}_j = \exp\left(\frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l)}\right) \tag{4.16}$$

À partir de l'équation (4.11), une estimation de la fonction de survie, basée sur les valeurs de $\hat{\zeta}_j$, est donnée par

$$\hat{S}_0(t) = \prod_{j=1}^{k} \exp\left(-\frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l)}\right) \tag{4.17}$$

Pour $t_{(k)} \leq t < t_{(k+1)},\ j = 1, 2, \cdots, (r-1)$. L'estimation de la fonction de hasard cumulé de base, déduit à partir de l'équation (4.17) est

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l)} \tag{4.18}$$

Pour $t_{(k)} \leq t < t_{(k+1)},\ j = 1, 2, \cdots, (r-1)$.

Exemple 4.2 Des données sur les temps de survie de 48 patients atteint de myeloma multiple sont présentées dans le tableau 4.2 du chapitre3. Cette base de données contient également les valeurs de sept (07) autres variables, enregistrées pour chaque individu.

Par commodité, les valeurs de la variable désignant le sexe d'un patient ont été redéfinies en code binaire, zéro et une unité, respectivement, pour male et femelle. Les variables sont présentées comme suit :

– *Age* : Age du patient ;
– *Sexe* : Sexe du patient (0 = masculin, 1 = féminin) ;
– *Bun* : Azote d'urée de sang ;
– *Ca* : Sérum de Calcium ;
– *Hb* : Sérum d'Hémoglobine ;
– *Pcells* : Pourcentage de cellules de plasma ;
– *Protein* : Proteine Bence-Jones (0 = absent, 1 = présent).

Les estimations des β-coefficients du modèle de Cox sont données dans le tableau 4.1. Ainsi qu'une représentation graphique de la fonction de survie est montré dans la figure 4.2.

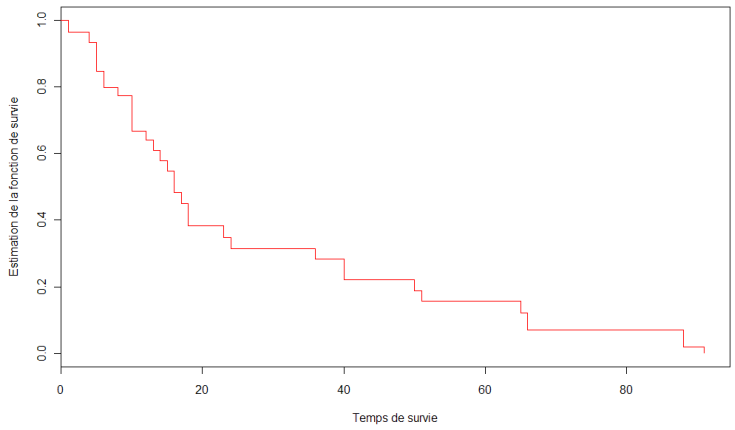

FIG. 4.2 – Estimation de la fonction de survie associée au modèle de Cox.

Variable	$\hat{\beta}$	$se(\hat{\beta})$
Age	−0.018	0.028
Sexe	−0.249	0.403
Bun	0.023	0.006
Ca	0.013	0.133
Hb	−0.133	0.069
Pcells	−0.001	0.007
Protein	−0.683	0.429

TAB. 4.1 – Estimation des β-coefficients du modèle de Cox correspondant aux données de l'exemple 3.2.

4.2.4 Extensions du modèle de Cox

- **Stratification :** Un modèle stratifié est un modèle où la fonction de hasard de base est propre à chaque strate. Il ne comporte cependant qu'un jeu de coefficients β qui sont donc supposées être les mêmes pour tous les groupes.

  La stratification se fait selon une variable supposée catégorielle. On repère ses valeurs par l'indice $s = 1, 2, \cdots$ Par exemple, une stratification selon le " *Sexe* " donnera lieu à deux (02) strates.

  Une fois qu'une variable de stratification est précisée, le modèle de Cox est estimé $[10]$ en admettant donc une fonction de hasard de base $h(0,s)(t)$ différente pour chaque strate s.

  L'hypothèse générale de proportionnalité est alors remplacée par :

  – Hasards proportionnels pour individus d'une même strate ;
  – Effets (β) des variables exogènes indépendants de la strate.

  La première hypothèse qui autorise des hasards non proportionnels entre individus de strates différentes, est un relâchement de l'hypothèse générale de proportionnalité. La seconde postule que les effets des covariables sur le rapport de hasard intra-strate sont les mêmes dans toutes les strates, ce qui permet d'estimer un seul jeu de coefficients pour l'ensemble des strates.

  Exemple 4.3 A titre d'illustration, reprenant l'exemple 4.2 en considérant l'estimation du modèle de Cox stratifié sur la variable " *Protéin* ". Les résultats obtenus sont décrit numériquement dans le tableau 4.2 et graphiquement par la figure 4.3.

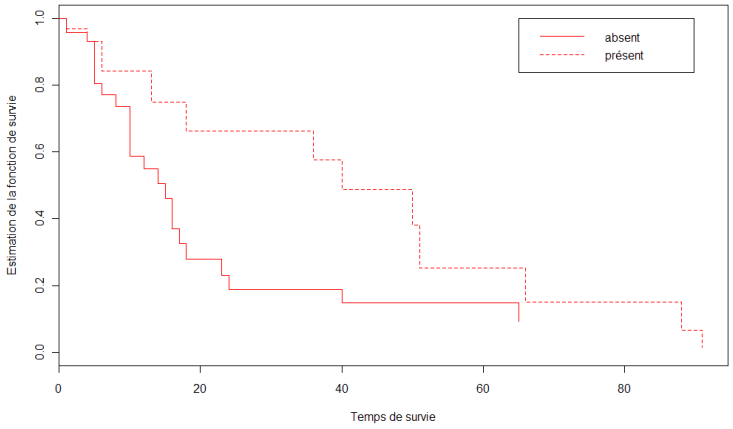

FIG. 4.3 – Estimation de la fonction de survie associée au modèle de Cox stratifié sur la variable " *Protéin* ".

Variable	$\hat{\beta}$	$se(\hat{\beta})$
Age	-0.009	0.029
Sexe	-0.396	0.410
Bun	0.021	0.006
Ca	0.005	0.134
Hb	-0.131	0.070
Pcells	-0.001	0.007

TAB. 4.2 – Estimation des coefficients du modèle de Cox, stratifié sur la variable " *Protéin* " de l'exemple 3.2.

- **Variables dépendantes du temps :** D'après le modèle semi paramétrique de Cox ainsi décrit par l'équation (4.1), la fonction de hasard du i-ème individu, peut s'écrire sous la forme

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\} h_0(t)$$

où x_{ji} est la j-ème variable exogène, $j = 1,2,\cdots,p$, associée au i-ème individu, $i = 1,2,\cdots,n$, et $h_0(t)$ est la fonction de hasard de base.

En généralisant ce modèle au cas où certaines de ces variables sont dépendantes du temps [10], $x_{ji}(t)$ désignera la jème variable exogènes enregistrée à l'instant t, pour le i-ème individu. Ce modèle devient alors

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}(t)\right\} h_0(t)$$

Il est important de noter que, dans le modèle ainsi donné, les valeurs des variables $x_{ji}(t)$ dépendent du temps, et ainsi le hasard relatif $h_i(t)/h_0(t)$ dépend également du temps.

4.3 Modèle à hasards proportionnels paramétrique

Si la forme de $h_0(t)$ est précisée dans le modèle (4.1), on dira qu'il s'agit d'un modèle à *hasards proportionnels paramétriques*.

4.3.1 Estimation d'un modèle paramétrique pour un échantillon simple

Les modèles paramétriques peuvent être ajustées à un ensemble de données de survie en utilisant la méthode du maximum de vraisemblance. Considérons d'abord le cas où les temps de survie ont été observés pour n individus, ainsi qu'il n'y a pas d'observations censurées. Si la fonction de densité de probabilité de la variable aléatoire associée au temps de survie est $f(t)$, la vraisemblance des n observations $t_1, t_2, \cdots, t_n$ est simplement le produit

$$\prod_{i=1}^{n} f(t_i)$$

Les estimations du maximum de vraisemblance des paramètres inconnus dans cette fonction sont ceux pour lesquels la fonction de vraisemblance est maximale. En pratique, il est plus commode d'utiliser le logarithme de la fonction de vraisemblance.

Considérons la situation où les données de survie incluent un ou plusieurs temps de survie censurées. Spécifiquement, supposons que r des n individus décèdent aux instants $t_1, t_2, \cdots, t_r$, et que les temps de survie des $n-r$ individus , $t_1^*, t_2^*, \cdots, t_{(n-r)}^*$, sont censurés à droite.

Les r temps de décès contribuent un terme de la forme

$$\prod_{j=1}^{r} f(t_j)$$

à la fonction de vraisemblance global. Evidemment, on ne peut pas ignorer l'information concernant l'expérience de survie des $n-r$ individus pour lesquels un temps de survie censuré a été enregistré.

Si un temps de survie est censuré en t^*, on dit que la durée de vie de l'individu est au moins t^*, et que la probabilité de cet évènement est $P(T \geq t^*)$, qui n'est autre que $S(t^*)$. Ainsi, chaque observation censurée contribuent un terme de cette forme à la vraisemblance des n observations. La fonction de vraisemblance totale est donc

$$\prod_{j=1}^{r} f(t_j) \prod_{l=1}^{n-r} S(t_l^*) \tag{4.19}$$

De façon plus compacte, supposons que les données sont enregistrées sous forme de n pairs d'observations, où la paire du i-ème individu est $(t_i, \delta_i),\ i = 1, 2, \cdots, n$. Dans cette notation, δ_i est une variable indicatrice qui prend la valeur zéro si le temps de survie t_i est censuré, et une unité sinon.

La fonction de vraisemblance peut s'écrire alors comme

$$\prod_{i=1}^{n}\{f(t_i)\}^{\delta_i}\{S(t_i)\}^{1-\delta_i} \tag{4.20}$$

Cette fonction, qui est équivalente à celle donnée dans l'expression (4.19), peut être donc maximisée par rapport aux paramètres inconnus dans les fonctions de densité et de survie.

Une expression alternative de la fonction de vraisemblance peut être obtenue en écrivant l'expression (4.20) sous la forme

$$\prod_{i=1}^{n}\left\{\frac{f(t_i)}{S(t_i)}\right\}^{\delta_i} S(t_i)$$

Ainsi que, cette expression devient

$$\prod_{i=1}^{n}\{h(t_i)\}^{\delta_i} S(t_i) \tag{4.21}$$

Cette version de la fonction de vraisemblance est particulièrement utile lorsque la forme de la fonction de densité de probabilité est complexe, ce qui est souvent le cas. Les estimations des paramètres inconnus dans cette fonction sont déterminées en maximisant le logarithme de cette fonction.

- **Distribution de Weibull**

Supposons que les temps de survie de n individus constituent un échantillon issu d'une distribution de Weibull, de paramètres λ et γ. Supposons également qu'il y a r décès parmi les n individus et $n-r$ temps de survie censurés à droite.
Les fonctions de densité de probabilité, de survie et de hasard de cette distribution sont données par

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma});\;\; S(t) = \exp(-\lambda t^{\gamma});\;\; h(t) = \lambda\gamma t^{\gamma-1}$$

D'après l'expression (4.20), la vraisemblance des n temps de survie est alors

$$\prod_{i=1}^{n}\{\lambda\gamma t_i^{\gamma-1}\exp(-\lambda t_i^{\gamma})\}^{\delta_i}\{\exp(-\lambda t_i^{\gamma})\}^{1-\delta_i}$$

où δ_i est égale à zéro si le i-ème temps de survie est censuré et une unité sinon.
De même, à partir de l'expression (4.21) la fonction de vraisemblance est

$$L(\lambda,\gamma) = \prod_{i=1}^{n}\{\lambda\gamma t_i^{\gamma-1}\}^{\delta_i}\exp(-\lambda t_i^{\gamma})$$

La fonction log-vraisemblance correspondante est donnée par

$$\log L(\lambda,\gamma) = \sum_{i=1}^{n}\delta_i\log(\lambda\gamma) + (\gamma-1)\sum_{i=1}^{n}\delta_i\log t_i - \lambda\sum_{i=1}^{n}t_i^{\gamma}$$

En notant que $\sum_{i=1}^{n} \delta_i = r$, la log-vraisemblance devient

$$\log L(\lambda,\gamma) = r\log(\lambda\gamma) + (\gamma-1)\sum_{i=1}^{n}\delta_i \log t_i - \lambda\sum_{i=1}^{n} t_i^{\gamma}$$

Les estimations du maximum de vraisemblance de λ et γ, notées par $\hat{\lambda}$ et $\hat{\gamma}$, sont solutions du système d'équation suivant

$$\frac{r}{\hat{\lambda}} - \sum_{i=1}^{n} t_i^{\hat{\gamma}} = 0 \tag{4.22}$$

et

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^{n}\delta_i \log t_i - \hat{\lambda}\sum_{i=1}^{n} t_i^{\hat{\gamma}}\log t_i = 0 \tag{4.23}$$

A partir de l'équation

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i^{\hat{\gamma}}} \tag{4.24}$$

Et en remplaçant $\hat{\lambda}$ par sa valeur dans l'équation (4.23), nous obtenons l'équation

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^{n}\delta_i \log t_i - \frac{r}{\sum_{i=1}^{n} t_i^{\hat{\gamma}}}\sum_{i=1}^{n} t_i^{\hat{\gamma}}\log t_i = 0 \tag{4.25}$$

C'est une équation non linéaire de $\hat{\gamma}$, qui peut être résolu en utilisant la méthode de Newton-Raphson.

4.3.2 Modèle à hasards proportionnels Weibull

Supposons que les p variables exogènes sont enregistrées pour chacun des n individus. D'après le modèle à hasards proportionnels paramétrique, la fonction de hasard du i-ème individu à l'instant t est

$$h_i(t) = \exp(\beta' x_i) h_0(t) \tag{4.26}$$

Pour $i = 1, 2, \cdots, n$. Bien que ce modèle a une apparence similaire à celle donnée dans l'équation (4.1), à une différence fondamentale, concernant la spécification de la fonction de hasard de base $h_0(t)$.

Dans le modèle de Cox, la forme de $h_0(t)$ est non spécifiée, tandis que, dans le modèle considérée dans cette section les temps de survie sont supposés ayant, par exemple, une distribution de Weibull, et ceci impose une forme paramétrique particulière pour $h_0(t)$.

Considérons un individu pour lequel toutes les valeurs des p covariables dans le modèle de l'équation (4.26) sont nulles. La fonction de hasard pour un tel individu est $h_0(t)$. Si le temps de survie de cet individu a une distribution de Weibull de paramètres λ et γ, alors leur fonction de hasard est telle que

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

En utilisant l'équation (4.26), la fonction de hasard pour le i-ème individu dans l'étude est alors donnée par

$$h_i(t) = \exp(\beta' x_i)\lambda\gamma t^{\gamma-1} \tag{4.27}$$

où $\exp(\beta' x_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$. La fonction de survie correspondante est

$$S_i(t) = \exp\{-\exp(\beta' x_i)\lambda t^{\gamma}\} \tag{4.28}$$

- **Estimation du modèle**

  Le modèle à hasards proportionnels Weibull est estimé en construisant la fonction de vraisemblance des n observations, et en maximisant cette fonction par rapport aux paramètres inconnus, $\beta_1, \beta_2, \cdots, \beta_p, \lambda$ et γ.

  Comme la fonction de hasard et la fonction de survie diffèrent pour chaque individu, la fonction de vraisemblance dans l'expression (4.21) peut s'écrire de la façon suivante

  $$\sum_{i=1}^{n} \{h_i(t_i)\}^{\delta_i} S_i(t_i) \tag{4.29}$$

  Le logarithme de cette fonction de vraisemblance est

  $$\sum_{i=1}^{n} \{\delta_i \log h_i(t_i) + \log S_i(t_i)\}$$

  En remplaçant $h_i(t_i)$ et $S_i(t_i)$ par leurs expressions données dans les équations (4.27) et (4.28), le log-vraisemblance devient

  $$\sum_{i=1}^{n} [\delta_i \{\beta' x_i + \log(\lambda\gamma) + (\gamma-1)\log t_i\} - \lambda \exp(\beta' x_i) t_i^{\gamma}]$$

  Qui peut s'écrire aussi

  $$\sum_{i=1}^{n} [\delta_i \{\beta' x_i + \log(\lambda\gamma) + \gamma \log t_i\} - \lambda \exp(\beta' x_i) t_i^{\gamma}] - \sum_{i=1}^{n} \delta_i \log t_i \tag{4.30}$$

  Le dernier terme de cette expression, $-\sum_{i=1}^{n} \delta_i \log t_i$, est indépendant des paramètres inconnus, et donc il peut être enlevé de la vraisemblance.
  La fonction log-vraisemblance résultante est alors

  $$\sum_{i=1}^{n} [\delta_i \{\beta' x_i + \log(\lambda\gamma) + \gamma \log t_i\} - \lambda \exp(\beta' x_i) t_i^{\gamma}] \tag{4.31}$$

  Supposons que les estimations des paramètres dans le modèle de l'équation (4.27) sont $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p, \hat{\lambda}$ et $\hat{\gamma}$. La fonction de survie estimée pour le i-ème individu est alors donnée par

  $$\hat{S}_i(t) = \exp\{-\exp(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi})\hat{\lambda} t^{\hat{\gamma}} \tag{4.32}$$

  Et la fonction de hasard estimée correspondante est

  $$\hat{h}_i(t) = \exp\{-\exp(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi})\hat{\lambda}\hat{\gamma} t^{\hat{\gamma}-1} \tag{4.33}$$

- **La forme log-linéaire du modèle**

  La fonction de survie du i-ème individu correspondant à un modèle à hasards proportionnels Weibull sous la forme log-linéaire, est exprimée par

$$S_i(t) = \exp\left\{-\exp\left(\frac{\log t - \mu - \alpha' x_i}{\sigma}\right)\right\} \quad pour \quad i = \overline{1,n}$$

  La relation entre ces deux représentations du modèle est telle que

$$\lambda = \exp(-\mu/\sigma);\ \ \gamma = \sigma^{-1};\ \ \beta_j = -\alpha_j/\sigma \quad pour \quad j = \overline{1,p}$$

  Les paramètres μ, σ sont nommés respectivement par *constante* (*intercept*) et *paramètre d'échelle* (*scale parameter*).

4.4 Modèle temps de survie accéléré

En dépit de la grande popularité et de la polyvalence du modèle de Cox, il y a de 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 différentes l'association entre les variables exogènes et le temps de survie. Il s'agit d'un modèle temps de survie accéléré.

4.4.1 La forme générale du modèle

D'après le modèle temps de survie accéléré général, la fonction de hasard du i-ème individu à l'instant t, est telle que

$$h_i(t) = e^{-\eta_i} h_0(t/e^{\eta_i}) \quad pour \quad i = \overline{1,n} \tag{4.34}$$

où $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ est la composante linéaire du modèle, dans lequel x_{ji} est la j-ème variable exogènes associée au i-ème individu.
La fonction de survie du ième individu est exprimée par

$$S_i(t) = S_0(t/e^{\eta_i}), \tag{4.35}$$

où $S_0(t)$ est la fonction de survie de base. Le terme $e^{-\eta_i}$ désigne le facteur d'accélération du i-ème individu.

4.4.2 La forme log-linéaire du modèle

Considérons un modèle log-linéaire pour la variable aléatoire T_i associée au temps de survie du i-ème individu, selon lequel

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\varepsilon_i \tag{4.36}$$

où $\alpha_1, \alpha_2, \cdots, \alpha_p$ sont les coefficients inconnus des p covariables et la quantité ε_i est une variable aléatoire supposée ayant une distribution de probabilité particulière (valeur extrême, normale, logistique$\cdots$).
D'après cette formulation, les α-paramètres reflètent l'effet de chaque variable exogène sur le temps de survie d'un individu.

Afin de montrer la relation entre cette représentation du modèle et celle donnée dans l'équation (4.34), considérons la fonction de survie de T_i. En utilisant l'équation (4.36), cette fonction est donnée par

$$S_i(t) = P(T_i \geq t) = P\{\exp(\mu + \alpha' x_i + \sigma\varepsilon_i) \geq t\},$$

où $\alpha' x_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$.
$S_i(t)$ peut s'écrire également sous la forme

$$S_i(t) = P\left\{\exp(\mu + \sigma\varepsilon_i) \geq \frac{t}{\exp(\alpha' x_i)}\right\}$$

Ainsi, la fonction de survie de base $S_0(t)$, désignant la fonction de survie d'un individu pour lequel les valeurs des p covariables sont nulles, est donnée par

$$S_0(t) = P\{\exp(\mu + \sigma\varepsilon_i) \geq t\}$$

Par conséquent,

$$S_i(t) = S_0\left(\frac{t}{\exp(\alpha' x_i)}\right) \tag{4.37}$$

Qui est la fonction de survie du i-ème individu correspondant à un modèle de temps de survie accéléré général.
À partir de l'équation (4.36),

$$S_i(t) = P(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\varepsilon_i \geq \log t) = P(\varepsilon_i \geq \frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}) \tag{4.38}$$

Ce résultat montre que la fonction de survie de T_i peut être déterminée à partir de la fonction de survie de la distribution de ε_i.
Si on note que $S_{\varepsilon_i}(t)$ est la fonction de survie de la variable aléatoire ε_i du modèle log-linéaire de l'équation (4.36), la fonction de survie du i-ème individu peut, à partir de l'équation (4.38), être exprimée par

$$S_i(t) = S_{\varepsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \tag{4.39}$$

La fonction de hasard cumulé de la distribution de T_i est donnée par $H_i(t) = -\log S_i(t)$, et à partir de l'équation (4.39),

$$H_i(t) = -\log S_{\varepsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) = H_{\varepsilon_i}(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}) \tag{4.40}$$

où $H_{\varepsilon_i}(\varepsilon) = -\log H_{\varepsilon_i}(\varepsilon)$ est la fonction de hasard cumulé de ε_i. La fonction de hasard correspondante est

$$h_i(t) = \frac{1}{\sigma t} h_{\varepsilon_i}\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \tag{4.41}$$

où $h_{\varepsilon_i}(\varepsilon)$ est la fonction de hasard de la distribution de ε_i.

4.4.3 Modèle temps de survie accéléré paramétrique

- **Modèle temps de survie accéléré Weibull**

Supposons que les temps de survie suivent une distribution de Weibull de paramètres λ et γ. La fonction de hasard de base est

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

À partir de l'équation (4.34), la fonction de hasard du i-ème individu est alors donnée par

$$h_i(t) = e^{-\eta_i}\lambda\gamma(e^{-\eta_i}t)^{\gamma-1} = (e^{-\eta_i})^{\gamma}\lambda\gamma t^{\gamma-1}$$

En termes de représentation log-linéaire du modèle temps de survie accéléré ainsi décrite dans l'équation (4.36), si T_i a une distribution Weibull, alors ε_i suit un type de distribution valeur extrême, nommé distribution de Gumbel, dont sa fonction de survie est donnée par

$$S_{\varepsilon_i}(\varepsilon) = \exp(-e^{\varepsilon})$$

Les fonctions de hasard cumulé et de hasard de cette distribution sont données respectivement par $H_{\varepsilon_i}(\varepsilon) = e^{\varepsilon}$ et $h_{\varepsilon_i}(\varepsilon) = e^{\varepsilon}$.
À partir de l'équation (4.39), nous pouvons déduire que la fonction de survie du ième individu correspondant à un modèle temps de survie accéléré Weibull, est donnée par

$$S_i(t) = \exp\left\{-\exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right)\right\} \tag{4.42}$$

Les fonctions de hasard cumulé et de hasard peuvent être déterminées à partir de la fonction de survie de l'équation (4.42), ou bien, à partir de $H_{\varepsilon_i}(\varepsilon)$ et $h_{\varepsilon_i}(\varepsilon)$, en utilisant les résultats générales des équations (4.40) et (4.41).
Ainsi, nous obtenons la fonction de hasard cumulé

$$H_i(t) = -\log S_i(t) = \exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right)$$

La fonction de hasard est alors donnée par

$$h_i(t) = \frac{1}{\sigma}\exp\left(\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \tag{4.43}$$

D'après le modèle à hasards proportionnels Weibull, la fonction de survie du i-ème individu est

$$S_i(t) = \exp\{-\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})\lambda t^{\gamma}\} \tag{4.44}$$

où λ et γ sont les paramètres de la fonction de hasard de base.
Il existe une relation directe entre l'équation (4.42) et l'équation (4.44), telle que

$$\lambda = \exp(-\mu/\sigma);\ \ \gamma = \sigma^{-1};\ \ \beta_j = -\alpha_j/\sigma, \quad pour \quad j = \overline{1,p}$$

4.4.4 Estimation du modèle temps de survie accéléré

Les modèles temps de survie accéléré sont estimés en utilisant la méthode du maximum de vraisemblance. A partir de l'expression (4.20), la vraisemblance des n observations, $t_1, t_2, \cdots, t_n$, est donnée par

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i}$$

où $f_i(t_i)$ et $S_i(t_i)$ sont les fonctions de densité et de survie du i-ème individu en t_i, et δ_i est l'indicatrice d'évènement de la i-ème observation, ainsi que, δ_i égale à une unité, si la i-ème observation est un décès et zéro si elle est censurée.
À partir de l'équation (4.39).

$$S_i(t_i) = S_{\varepsilon_i}(z_i)$$

où $z_i = (\log t_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi})/\sigma$, et sa dérivé par rapport à t_i donne

$$f_i(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}(z_i)$$

La fonction de vraisemblance peut être donc exprimée, en termes de fonctions de survie et de densité de ε_i, par

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} \{f_{\varepsilon_i}(z_i)\}^{\delta_i} \{S_{\varepsilon_i}(z_i)\}^{1-\delta_i}$$

La fonction log-vraisemblance est alors

$$\log L(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \{-\delta_i \log(\sigma t_i) + \delta_i \log f_{\varepsilon_i}(z_i) + (1-\delta_i) \log S_{\varepsilon_i}(z_i)\} \tag{4.45}$$

Les estimations du maximum de vraisemblance des $p+2$ paramètres inconnus, μ, σ et $\alpha_1, \alpha_2, \cdots, \alpha_p$ sont déterminés par une maximisation de la fonction log-vraisemblance sus citée, en utilisant la méthode itérative de Newton-Raphson.

Exemple 4.4 Reprenant les données de l'exemple 3.6 présentant les temps de survie de femmes atteintes d'un cancer du sein, randomisées en deux groupe, que ce soient traitées par A ou B.

Dans le but d'ajuster le modèle temps de survie accéléré à ces données, considérons à titre d'exemple le modèle temps de survie accéléré Weibull. Un modèle log-linéaire de la variable aléatoire T_i associée au temps de survie de la i-ème femme, est tel que

$$\log T_i = \mu + \alpha x_i + \sigma \varepsilon_i$$

où ε_i suit une distribution de Gumbel, μ, σ et α sont les paramètres, et x_i est la valeur d'une variable exogène associée au type de traitement, telle que $x_i = 0$ si la i-ème femme est traitée par A et $x_i = 1$ sinon.
Les estimations des paramètres de ce modèle sont $\hat{\mu} = 5.854$, $\hat{\sigma} = 1.067$ et $\hat{\alpha} = -0.997$.
Le facteur d'accélération, $e^{-\alpha x_i}$, est estimée par $e^{0.997} = 2.71$ pour une femme subissant le traitement B, c'est-à-dire que le temps de décès d'une femme traitée par B est accéléré par un facteur d'ordre 2.7
La fonction de survie estimée de la i-ème femme est donnée par

$$\hat{S}_i(t) = \exp\left\{-\exp\left(\frac{\log t - \hat{\mu} - \hat{\alpha} x_i}{\hat{\sigma}}\right)\right\}$$

et la fonction de hasard estimée est

$$\hat{h}_i(t) = \hat{\sigma}^{-1} t^{\hat{\sigma}^{-1}-1} \exp\left(\frac{-\hat{\mu} - \hat{\alpha} x_i}{\hat{\sigma}}\right)$$

Les graphes de ces fonctions selon le traitement sont montrés, respectivement, dans les figures 4.4 et 4.5.

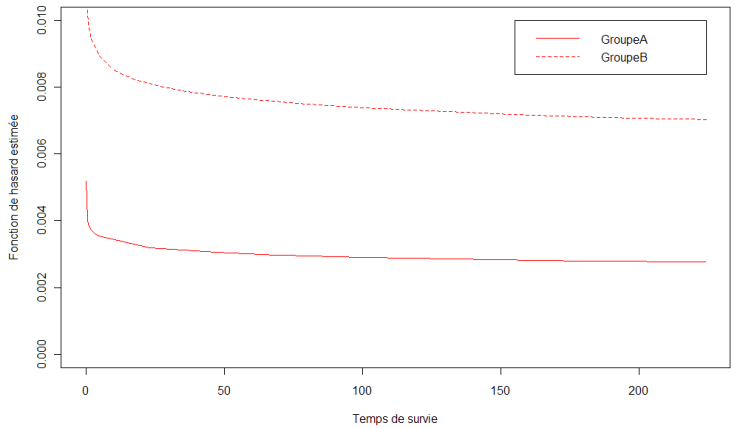

FIG. 4.4 – Fonctions de hasard estimées associées au modèle temps de survie accéléré Weibull.

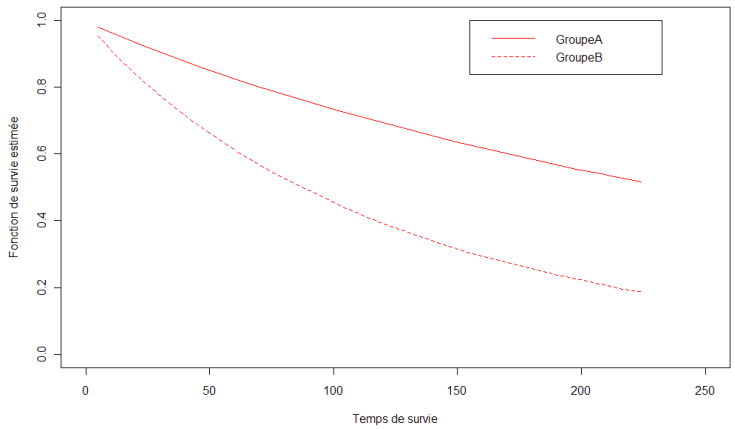

FIG. 4.5 – Fonctions de survie estimées associées au modèle temps de survie accéléré Weibull.

4.5 Conclusion

Pour évaluer l'effet de variables exogènes sur la durée de vie par exemple d'un individu, des modèles de régression de survie peuvent être employée à savoir, le modèle de Cox, le modèle à hasards proportionnels

paramétriques et le modèle temps de survie accéléré.

Le modèle de Cox permet d'exprimer notamment la fonction de hasard d'un individu sous une forme multiplicative d'une fonction de hasard de base non spécifiée et d'une fonction de régression de variables exogènes dont les coefficients sont inconnus et peuvent être estimés en utilisant la méthode de Newton-Raphson. Ce modèle est tributaire de deux (02) hypothèses cruciales ; il s'agit de la log-linéarité des variables et la proportionnalité des hasards.

Dans le cas où la fonction de hasard de base est paramétrique, les modèles à hasards proportionnelles paramétriques peuvent être utilisées. Cependant, si l'hypothèse de proportionnalité n'est pas vérifiée, il est commode d'utiliser les modèles temps de survie accéléré.

5

Application de l'analyse de survie à l'assurance non-vie

Sommaire

5.1 Introduction

Vu les changements auquel le marché de l'assurance automobile est subi, les assureurs sont désormais incités à développer des modèles optimaux de surveillance et de gestion de leur portefeuille afin de fidéliser les clients les plus rentables et de résilier certain contrats. Pour la réalisation de cet objectif, les compagnies d'assurances sont devenues de plus en plus motivées vis-à-vis de l'utilisation d'outils d'inférences statistiques plus développés, telles les modèles d'analyse de survie.

Ce chapitre sera consacré d'une part, à analyser le phénomène de résiliation d'un contrat d'assurance automobile en tenant compte du cout de sinistralité correspondant, et d'une autre part, à l'élaboration de modèles prédictifs pour la durée de vie de ces contrats [1][2][12]. Pour cela, nous introduisons les modèles de survie suivant : le modèle non-paramétrique de Kaplan-Meier, le modèle temps de survie accéléré paramétrique et le modèle semi-paramétrique de Cox.

5.2 Présentation des données

Notre étude est effectuée sur un mélange de données, simulées selon une certaine distribution de probabilité et réelles issues d'un portefeuille automobile géré par une agence d'une compagnie de taille significative sur le marché français d'assurance automobile. Pour des raisons de confidentialité, l'extrait de données réelles est non représentatif du portefeuille global. Néanmoins, cette sélection conserve des caractéristiques suffisamment adéquates pour permettre l'élaboration de notre analyse. Le fichier final à étudier comporte 1622 contrats Autos.

Concernant la partie de données simulées bien entendu en s'inspirant de la réalité, on suppose que pour chacun des contrats Auto considérés, il est enregistré un coût de sinistralité suite à la survenue d'un accident, qui sera simulé par la suite selon une Log-normale, de moyenne $m = 10.15$ et d'écart-type $s = 0.37$ (voir le Chapitre 2). En outre, on simule en supplément deux (02) autres variables exogènes : l'âge du permis selon une Bernoulli de $p_1 = 0.5$ et l'âge du conducteur selon une Bernoulli de $p_2 = 0.5$

Quant à la partie réelle, on suppose que pour chaque contrat, il est enregistré trois (03) autres variables exogènes à savoir le Bonus-Malus, l'âge du véhicule et la formule d'assurance.

- **Mécanisme de censure**

  L'application des modèles de survie étant tributaire de la caractéristique de censure et vue la structure de notre base de données, nous avons élaboré le mécanisme de censure suivant.

  La variable d'intérêt considérée est la durée de vie d'un contrat Auto (notée `DurVie`), définissant la diffé-

rence entre la date de résiliation et la date de création du contrat telle que :

- Si le coût de sinistralité correspondant à un contrat Auto dépasse un certain seuil fixe, la date résiliation du contrat est confondu avec la date d'accident, et donc nous considérons la différence entre la date de résiliation et la date de création ;

- Sinon, la date de résiliation n'est pas connue et donc nous considérons l'écart entre la date d'accident et la date création du contrat (Il s'agit alors d'une censure droite fixe).

La durée de vie du contrat étant indépendante de la survenue de l'évènement d'intérêt (la résiliation du contrat) ; on dit qu'il s'agit de censure non informative.

Généralement, le seuil de sinistralité fixé selon l'appréciation de l'utilisateur. Dans notre étude, le seuil considéré est le quantile d'ordre 0.3 de la distribution du coût de sinistralité, correspondant à 21077.72 DA (30% de censures).

- **Le codage des variables exogènes**

  Pour faciliter l'application des modèles de survie à l'aide du logiciel R sur la base de données considérées, nous procédons à un codage des cinq (05) variables exogènes :

1. L'âge du permis (noté `AgePerm`) :
   - `AgePerm`1 : L'âge du permis inférieur ou égal à 1 an ;
   - `AgePerm`2 : L'âge du permis strictement supérieur à 1 an.

2. L'âge du conducteur (noté `AgeCond`) :
   - `AgeCond`1 : L'âge du conducteur inférieur ou égal à 25 ans ;
   - `AgeCond`2 : L'âge du permis strictement supérieur à 25 ans.

3. Le Bonus - Malus (noté `CodeMalus`) :
   - `CodeMalus`1 : Le Bonus - Malus égal à 0.5 (50% de bonus) ;
   - `CodeMalus`2 : Le Bonus - Malus compris entre 0.5 et 0.7 (entre 30% et 50% de bonus) ;
   - `CodeMalus`3 : le Bonus - Malus strictement supérieur à 0.7

4. La formule d'assurance (notée `Code`) :
   - `Code`1 : Formule complète tous risques ;
   - `Code`2 : Formule RC et dommages hors tous risques ;

– `Code`3 : Formule RC seule.

5. L'âge du véhicule (noté `AgeVehic`), cette variable est définit comme la différence entre la date d'accident et la date de mise en circulation. Elle est regroupée en trois (03) classes :
   – `AgeVehic`1 : L'âge du véhicule inférieur ou égal à 4 ans ;
   – `AgeVehic`2 : L'âge du véhicule compris entre 4 ans et 8 ans (inclus) ;
   – `AgeVehic`3 : L'âge du véhicule strictement supérieur à 8 ans.

Les méthodes d'inférences statistiques classiques ne sont plus appropriées à l'ensemble de données considérées, en raison de la présence des censures. Afin d'estimer les lois décrivant la durée de vie des contrats Autos, particulièrement, la fonction de survie exprimant la probabilité de survie d'un contrat Auto, et éventuellement en présence des cinq (05) variables exogènes : l'âge du permis, l'âge du conducteur, le Bonus-Malus, la formule d'assurance et l'âge du véhicule, on propose d'employer les modèles de survie suivantes : modèle de Kaplan-Meier, modèle de Cox et modèle temps de survie.

5.3 Statistique descriptive

D'après le tableau 5.1, le portefeuille est composée de contrats avec 50% de bonus, de formules RC et dommages hors tous risques et de véhicules âgés de plus 8 ans. Il est également constitué de contrats dont l'assuré (conducteur du véhicule) est âgé de plus 25 ans, ayant un permis de conduire dépassant un 1 an.

Contrats	Résiliés	Censurés	**Total**
AgePerm1	555	230	785
AgePerm2	601	236	837
Total	1156	466	1622
AgeCond1	482	202	684
AgeCond2	674	264	938
Total	1156	466	1622
CodeMalus1	463	184	647
CodeMalus2	310	127	437
CodeMalus3	383	155	538
Total	1156	466	1622
Code1	414	165	579
Code2	457	187	644
Code3	285	114	399
Total	1156	466	1622
AgeVehic1	110	45	155
AgeVehic2	298	112	410
AgeVehic3	748	309	1057
Total	1156	466	1622

TAB. 5.1 – Les résultats d'une étude statistique exploratoire effectuée sur l'ensemble des données considérées.

Le tableau 5.2 donne quelques statistiques fondamentales concernant la variable " DurVie ", constituées respectivement de :

– La durée de vie moyenne des contrats ;

– La borne inférieure et supérieure d'un intervalle de confiance à 95% exprimées par :

$$\texttt{BorneInf} = \bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \quad et \quad \texttt{BorneSup} = \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

où n est la taille de l'échantillon considéré ;

– La durée de vie minimale et maximale (Min, Max) pour l'ensemble des données et pour chacune des cinq (05) variables exogènes considérées.

	Moyenne			BorneInf			BorneSup		
	Tout	Résil	Cens	Tout	Résil	Cens	Tout	Résil	Cens
Global	10.40	10.42	10.36	10.11	10.08	9.81	10.69	10.76	10.91
AgePerm1	10.53	10.48	10.65	10.12	10.01	9.85	10.94	10.95	11.45
AgePerm2	10.28	10.37	10.07	9.88	9.90	9.30	10.68	10.84	10.84
AgeCond1	10.71	10.72	10.67	10.27	10.20	9.81	11.15	11.24	11.53
AgeCond2	10.18	10.21	10.12	9.80	9.77	9.40	10.56	10.65	10.84
CodeMalus1	12.68	12.50	13.13	12.23	11.97	12.28	13.13	13.03	13.98
CodeMalus2	10.19	10.45	9.53	9.67	9.84	8.55	10.71	11.06	10.51
CodeMalus3	7.84	7.88	7.74	7.41	7.37	6.91	8.27	8.39	8.57
Code1	11.08	11.10	11.05	10.60	10.55	10.09	11.56	11.65	12.01
Code2	10.91	10.85	11.05	10.46	10.32	10.22	11.36	11.38	11.88
Code3	8.60	8.75	8.22	8.04	8.09	7.15	9.16	9.41	9.29
AgeVehic1	6.66	6.95	5.96	5.93	6.05	4.73	7.39	7.85	7.19
AgeVehic2	8.59	8.42	9.02	8.08	7.83	8.01	9.10	9.01	10.03
AgeVehic3	11.66	11.73	11.48	11.31	11.32	10.79	12.01	12.14	12.17

	Min			Max		
	Tout	Résil	Cens	Tout	Résil	Cens
Global	0.15	0.37	0.15	26.81	26.81	26.69
AgePerm1	0.33	0.85	0.33	26.81	26.81	26.69
AgePerm2	0.15	0.37	0.15	26.38	26.38	25.06
AgeCond1	0.15	0.41	0.15	26.69	26.38	26.69
AgeCond2	0.33	0.37	0.33	26.81	26.81	25.06
CodeMalus1	0.49	0.49	1.36	26.81	26.81	26.69
CodeMalus2	0.51	1.08	0.51	25.68	25.68	23.10
CodeMalus3	0.15	0.37	0.15	26.38	26.38	25.06
Code1	0.41	0.41	0.80	26.69	25.43	26.69
Code2	0.48	0.49	0.48	26.81	26.81	25.06
Code3	0.15	0.37	0.15	23.21	23.21	22.60
AgeVehic1	0.41	0.41	0.80	20.10	20.10	16.32
AgeVehic2	0.48	0.49	0.48	23.10	21.15	23.10
AgeVehic3	0.15	0.37	0.15	26.81	26.81	26.69

TAB. 5.2 – Statistiques fondamentales concernant la variable d'intérêt " DurVie ".

Ces statistiques ainsi présentées indiquent que la durée de vie moyenne des contrats est de l'ordre 10.4 ans

au global de ce portefeuille, avec des écarts plus ou moins important entre les différentes divisions étudiée, qui oscillent autour de 6 ans et 13 ans.

Cependant, ces statistiques ne nous donnent pas assez d'information concernant :

- Si les écarts sont significative ou pas ;
- Les distributions de probabilité caractérisant la variable d'intérêt " `DurVie` ", notamment, la fonction de survie.

En outre, les méthodes d'inférences statistiques classiques ne sont plus appropriées à l'ensemble de données considérées, en raison de la présence des censures. Afin d'estimer les lois décrivant la durée de vie des contrats Autos, particulièrement, la fonction de survie, et éventuellement en présence de variables exogènes, il est commode d'employer les modèles de l'analyse de survie ainsi décrits dans les chapitres précédents.

5.4 Approche non-paramétrique : estimateur de Kaplan-Meier

Sachant que les données considérées dans notre étude sont fractionnées en cinq groupes, liés aux cinq variables exogènes sus citées, l'approche non paramétrique de Kaplan-Meier est appliquée pour chacun de ces groupes. Supposons qu'il y a r temps de résiliation, $r \leq n \quad (n = 1622)$, arrangés selon l'ordre croissant, $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, où $t_{(j)}, \quad j = 1, 2, \cdots, r$, est le j-ème temps de résiliation d'un contrat Auto selon le mécanisme de censure ainsi décrit dans la section 5.2.

L'estimation de la fonction de survie par la méthode de Kaplan-Meier est donnée par l'équation (3.4), telle que, n_j est le nombre de contrats exposés au risque de résiliation à l'instant $t_{(j)}$, c'est-à-dire le nombre de contrats qui sont toujours juste avant $t_{(j)}$ (ni résiliés, ni censurés) et d_j est le nombre de résiliation en ce temps.

- L'âge du permis (noté `AgePerm`)

  La statistique Log-rank prenant la valeur 0.4, confirme qu'il n'y a pas de différence significative entre les deux (02) classes liées à la variable " `AgePerm` " avec une p-value de 0.519

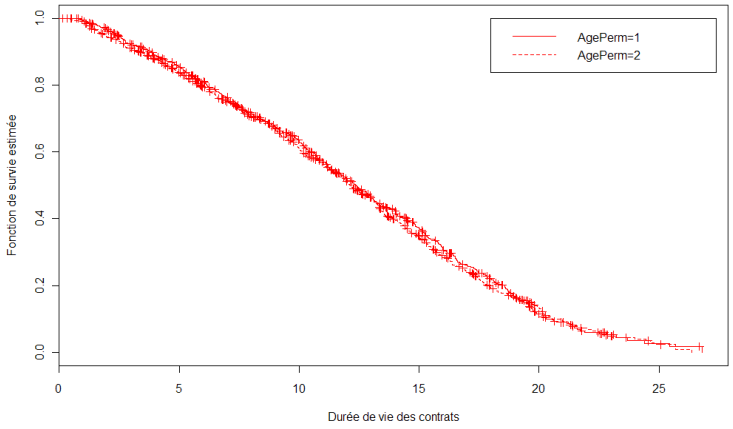

FIG. 5.1 – Estimation de Kaplan-Meier de la fonction de survie correspondant à la variable " `AgePerm` ".

- L'âge du conducteur (noté `AgeCond`)

  Vue la figure 5.2, nous constatons une différence faiblement significative entre les deux " `AgeCond` ". Ceci est confirmé par la valeur de la statistique Log-rank de 2.8 avec une p-value de 0.096 environ.

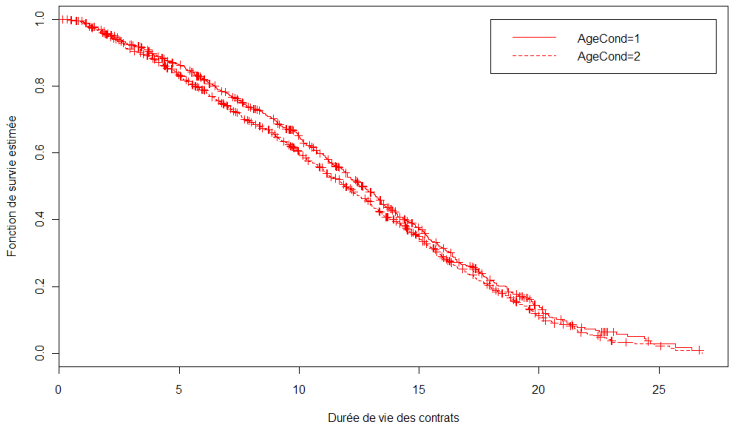

FIG. 5.2 – Estimation de Kaplan-Meier de la fonction de survie correspondant à la variable " `AgeCond` "

- Le Bonus - Malus (noté `CodeMalus`)

  La statistique Log-rank prenant la valeur 145 confirme une différence hautement significative entre les trois (03) classes de la variable " `CodeMalus` " avec une p-value nulle.

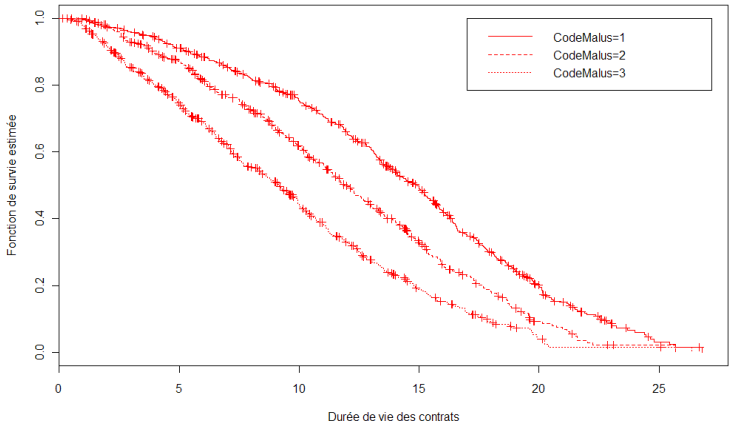

FIG. 5.3 – Estimation de Kaplan-Meier de la fonction de survie correspondant à la variable " `CodeMalus` ".

- La formule d'assurance (notée `Code`)

  La statistique Log-rank prenant la valeur 32.4 confirme une différence significative entre les différentes formule d'assurance avec une p-value de $9.17e-08$. D'après la figure 5.4, la classe correspondant au " `Code`3 " a plus de chance pour que la durée de vie des contrats soit plus faible que pour les deux autres classes.

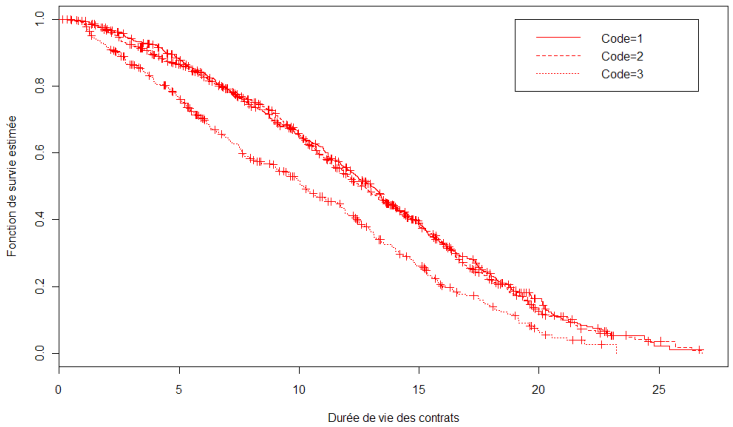

FIG. 5.4 – Estimation de Kaplan-Meier de la fonction de survie correspondant à la variable " `Code` ".

- L'âge du véhicule (noté `AgeVehic`)

  La figure 5.5 indique qu'il y a une différence hautement significative entre les trois " `AgeVehic` ". La statistique Log-rank prenant la valeur 141 confirme cette différence avec une p-value nulle.

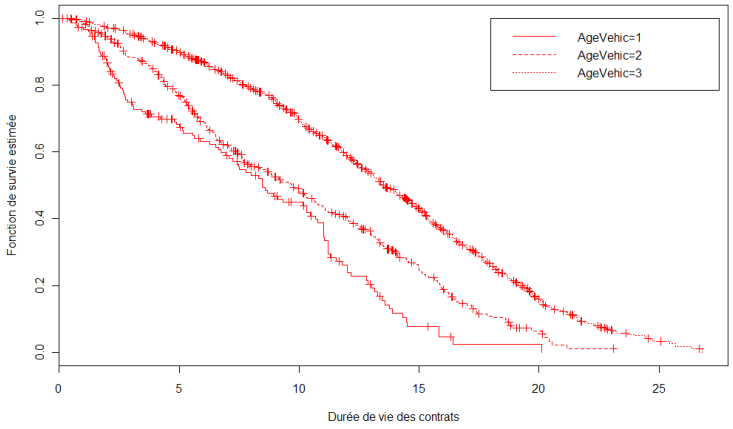

FIG. 5.5 – Estimation de Kaplan-Meier de la fonction de survie correspondant à la variable "`AgeVehic`".

D'après les figures ainsi présentées ci-dessus, nous constatons que les variables exogènes, telles, le Bonus -Malus, la formule d'assurance et l'âge du véhicule jouent un rôle dans la durée de vie des contrats Autos avec un effet plus ou moins important de l'âge du conducteur. Tandis que, l'âge du permis n'a pas d'effet significatif.

5.5 Approche paramétrique : modèle temps de survie accéléré

Dans le but de déterminer un modèle temps de survie accéléré pour la durée de vie des contrats Autos ainsi présenté dans le chapitre 4 où les α_i exprime l'impact des cinq (05) facteurs considérés sur cette derniere (durée) , nous proposons de tester les distributions de probabilités suivantes : la distribution de Weibull, la distribution Log-logistique et la distribution Log-normale, tout en faisant varier le nombre des variables exogènes ou covariables. Considérons les cas suivants :

- Le cas de cinq (05) variables exogènes (covariables)

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.565	0.094	$4.24e^{-164}$
`AgePerm`	-0.020	0.029	$4.84e^{-01}$
`AgeCond`	-0.055	0.029	$6.19e^{-02}$
`CodeMalus`	-0.168	0.017	$9.45e^{-23}$
`Code`	-0.174	0.022	$8.25e^{-16}$
`AgeVehic`	0.309	0.024	$1.44e-37$
$\log(scale)$	-0.715	0.024	$9.65e^{-200}$
Logvrais	-3866.1		

TAB. 5.3 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution de Weibull, en présence de cinq (05) covariables.

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.344	0.116	$4.03e^{-90}$
`AgePerm`	-0.059	0.035	$8.91e^{-02}$
`AgeCond`	-0.058	0.035	$1.02e^{-01}$
`CodeMalus`	-0.212	0.021	$2.91e^{-24}$
`Code`	-0.240	0.025	$2.57e^{-21}$
`AgeVehic`	0.418	0.031	$1.60e^{-42}$
$\log(scale)$	-0.965	0.024	$0.00e^{+00}$
Logvrais	-3931.9		

TAB. 5.4 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution Log-logistique, en présence de cinq (05) covariables.

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.312	0.122	$6.98e^{-80}$
`AgePerm`	-0.070	0.037	$5.88e^{-02}$
`AgeCond`	-0.075	0.038	$4.82e^{-02}$
`CodeMalus`	-0.209	0.022	$1.16e^{-20}$
`Code`	-0.283	0.027	$1.93e^{-25}$
`AgeVehic`	0.462	0.031	$2.03e^{-49}$
$\log(scale)$	-0.359	0.021	$1.09e^{-65}$
Logvrais	-3964.6		

TAB. 5.5 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution Log-normale, en présence de cinq (05) covariables.

Selon le critère de la vraisemblance maximale, le modèle temps de survie accéléré Weibull est le plus approprié pour la modélisation de la durée de vie des contrats Autos, avec une valeur du Logvrais d'ordre -3866.1

Les estimations des paramètres de ce modèle sont : $\mu(\texttt{Intercept}) = 2.565$;
$\sigma(\texttt{scale}) = 0.489$;
$\alpha_1(\texttt{AgePerm}) = -0.020$;
$\alpha_2(\texttt{AgeCond}) = -0.055$;
$\alpha_3(\texttt{CodeMalus}) = -0.168$;
$\alpha_4(\texttt{Code}) = -0.174$;
$\alpha_5(\texttt{AgeVehic}) = 0.309$

Dans le cas où on dispose de cinq (05) variables exogènes, l'impact des variables `CodeMalus`, `Code` et `AgeVehic` sur la durée de vie des contrats, est le plus significatif.

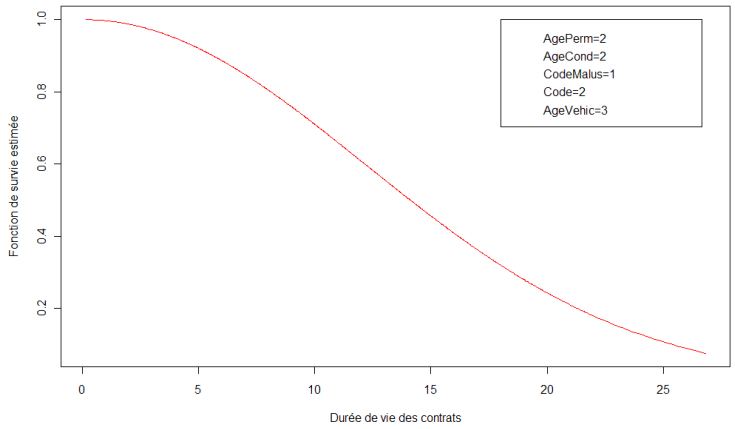

FIG. 5.6 – Estimation de la fonction de survie correspondant au modèle temps de survie accéléré Weibull (cas de cinq (05) covariables).

- Le cas de quatre (04) variables exogènes

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.534	0.083	$2.61e^{-207}$
`AgeCond`	-0.053	0.029	$6.70e-^{02}$
`CodeMalus`	-0.168	0.017	$8.87e^{-23}$
`Code`	-0.174	0.022	$6.61e^{-16}$
`AgeVehic`	0.309	0.024	$1.47e^{-37}$
$\log(scale)$	-0.715	0.024	$8.96e^{-200}$
Logvrais	-3866.4		

TAB. 5.6 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution de Weibull, en présence de quatre (04) covariables.

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.252	0.103	$2.16e^{-105}$
`AgeCond`	-0.054	0.035	$1.22e^{-01}$
`CodeMalus`	-0.211	0.021	$4.52e^{-24}$
`Code`	-0.240	0.025	$3.11e^{-21}$
`AgeVehic`	0.416	0.031	$3.87e^{-42}$
$\log(scale)$	-0.963	0.024	$0.00e^{+00}$
Logvrais	-3933.3		

TAB. 5.7 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution Log-logistique, en présence de quatre (04) covariables.

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.200	0.107	$7.54e^{-94}$
`AgeCond`	-0.071	0.038	$6.04e^{-02}$
`CodeMalus`	-0.208	0.022	$1.61e^{-20}$
`Code`	-0.284	0.027	$1.74e^{-25}$
`AgeVehic`	0.462	0.031	$3.62e^{-49}$
$\log(scale)$	-0.357	0.021	$3.60e^{-65}$
Logvrais	-3966.4		

TAB. 5.8 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution Log-normale, en présence de quatre (04) covariables.

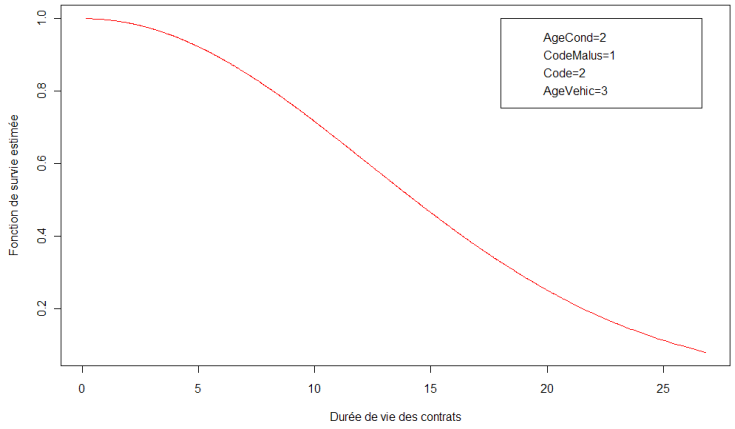

FIG. 5.7 – Estimation de la fonction de survie correspondant au modèle temps de survie accéléré Weibull (cas de quatre (04) covariables).

- Le cas de trois (03) variables exogènes

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.452	0.069	$2.31e^{-276}$
`CodeMalus`	-0.170	0.017	$3.49e^{-23}$
`Code`	-0.172	0.021	$1.22e^{-15}$
`AgeVehic`	0.308	0.024	$2.90e^{-37}$
$\log(scale)$	-0.714	0.024	$4.21e^{-199}$
Logvrais	-3868.1		

TAB. 5.9 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution de Weibull, en présence de trois (03) covariables.

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.168	0.088	$5.03e^{-134}$
`CodeMalus`	-0.213	0.021	$1.84e^{-24}$
`Code`	-0.239	0.025	$5.13e^{-21}$
`AgeVehic`	0.416	0.031	$5.50e^{-42}$
$\log(scale)$	-0.963	0.024	$0.00e^{+00}$
Logvrais	-3934.5		

TAB. 5.10 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution Log-logistique, en présence de trois (03) covariables.

	$\hat{\alpha}$	$se(\hat{\alpha})$	p-value
`Intercept`	2.092	0.090	$3.79e^{-119}$
`CodeMalus`	-0.210	0.022	$6.96e^{-21}$
`Code`	-0.282	0.027	$3.30e^{-25}$
`AgeVehic`	0.461	0.031	$7.32e^{-49}$
$\log(scale)$	-0.356	0.021	$1.11e^{-64}$
Logvrais	-3968.2		

TAB. 5.11 – Estimation des coefficients du modèle temps de survie accéléré avec une distribution Log-normale, en présence de trois (03) covariables.

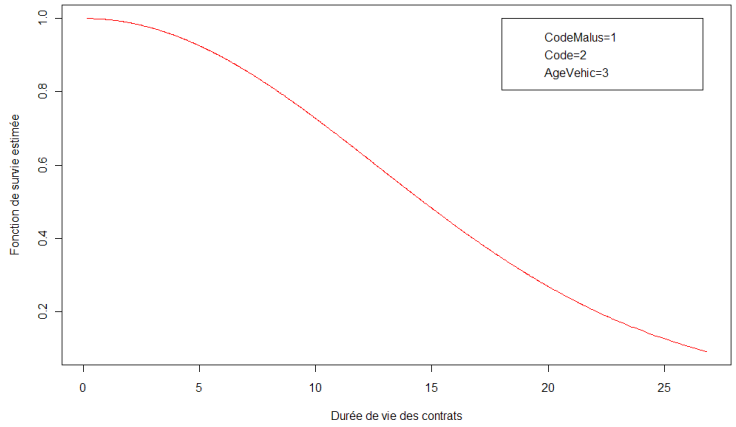

FIG. 5.8 – Estimation de la fonction de survie correspondant au modèle temps de survie accéléré Weibull (cas de trois (03) covariables).

Les résultats obtenus dans les trois (03) cas considérés montrent qu'en dépit de la variation du nombre des variables exogènes, le Bonus-Malus, la formule d'assurance et l'âge du véhicule, demeurent les plus influents sur la durée de vie des contrats Autos avec un effet faiblement significatif de l'âge du conducteur. Quant à l'âge du permis, il ne joue aucun rôle sur la variable d'intérêt.

5.6 Approche semi-paramétrique : modèle de Cox

Supposons qu'on observe la durée de vie de n contrats Autos $t_1, t_2, \cdots, t_n$, et que δ_i est l'indicatrice de l'occurrence de l'évènement d'intérêt, c'est-à-dire la résiliation du contrat, qui prend zéro si la ième durée de vie t_i, $i = 1, 2, \cdots, n$, est censurée à droite (le contrat est en vigueur et il n'a pas été résilié) et une unité sinon.

La fonction de vraisemblance partielle de Cox est :

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)} \right)^{\delta_i}$$

où $R(t_i)$ est l'ensemble des contrats à risque en t_i, x_i est le vecteur des cinq (05) variables exogènes au i-ème contrat et β est le vecteur des paramètres à estimer, exprimant l'effet de ces variables sur la durée de vie d'un contrat Auto.

Dans ce cas, le rapport $\frac{\exp(\beta' x_i)}{\sum_{l \in R(t_i)} \exp(\beta' x_l)}$ exprime la probabilité conditionnelle que le i-ème contrat soit résilié en t_i, sachant qu'il y a eu une seule résiliation en ce temps. La fonction log-vraisemblance correspondante est alors données par

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i (\beta' x_i - \log\{\exp(\beta' x_l)\})$$

Les β-paramètres de ce modèle sont estimés par la méthode itérative de Newton-Raphson.

- Le cas de cinq (05) variables exogènes

	$\hat{\beta}$	$se(\hat{\beta})$	p-value
`AgePerm`	0.025	0.059	0.670
`AgeCond`	0.122	0.060	0.041
`CodeMalus`	0.375	0.036	0.000
`Code`	0.383	0.044	0.000
`AgeVehic`	-0.693	0.051	0.000
Logvrais	-7239.40		

TAB. 5.12 – Estimations des coefficients du modèle de Cox correspondant aux cinq (05) variables exogènes considérées.

- Le cas de cinq (04) variables exogènes

	$\hat{\beta}$	$se(\hat{\beta})$	p-value
`AgeCond`	0.121	0.060	0.043
`CodeMalus`	0.375	0.036	0.000
`Code`	0.383	0.044	0.000
`AgeVehic`	-0.693	0.051	0.000
Logvrais	-7239.49		

TAB. 5.13 – Estimations des coefficients du modèle de Cox correspondant à quatre (04) variables exogènes.

- Le cas de cinq (03) variables exogènes

	$\hat{\beta}$	$se(\hat{\beta})$	p-value
`CodeMalus`	0.378	0.036	0
`Code`	0.378	0.044	0
`AgeVehic`	-0.689	0.050	0
Logvrais	-7241.55		

TAB. 5.14 – Estimations des coefficients du modèle de Cox correspondant à trois (03) variables exogènes.

Le modèle de Cox peut etre également utilisé en stratifiant sur une variable exogène, par exemple la variable " `Code` ", ainsi présenté dans la figure 5.9.

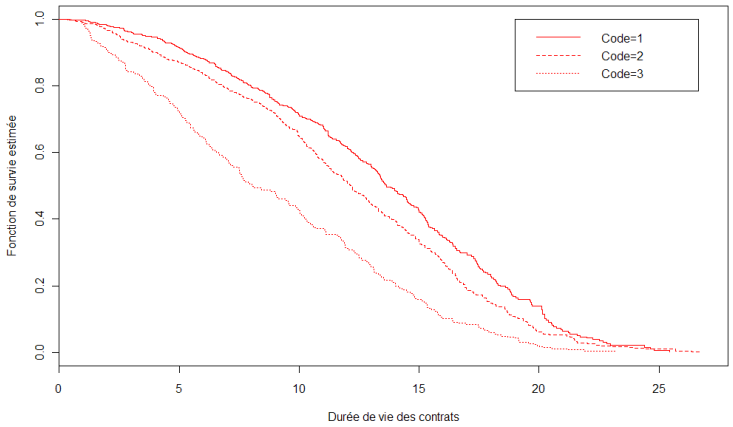

FIG. 5.9 – Estimation de la fonction de survie correspondant à un modèle de Cox stratifié sur la variable " `Code`".

5.7 Conclusion

En appliquant les modèles d'analyse de survie sur la base de données considérée, éventuellement en présence de variables exogènes enregistrées pour chaque contrat Auto, on constate que la variation du nombre des variables exogènes ne joue aucun rôle dans l'évaluation des facteurs qui peuvent influencer la durée de vie des contrats. De ce fait, le Bonus-Malus, la formule d'assurance et l'âge du véhicule demeurent à effet hautement significative avec un effet plus ou moins important de l'âge du conducteur.

Conclusion générale

L'utilisation des modèles survie en assurance non vie, particulièrement, automobile est susceptible de répondre à de nombreux enjeux. En effet, l'introduction de ces modèles nous a permis, selon un mécanisme de censure spécifique, d'évaluer l'impact des variables exogènes, telles le Bonus-Malus, la formule d'assurance, l'âge du véhicule, l'âge du conducteur et l'âge du permis, sur la durée de vie des contrats Autos associé à un cout de sinistralité et de sélectionner en outre l'ensemble des facteurs les plus influents sur la variable d'intérêt.

On a également constaté que la variation du nombre des variables exogènes n'a pas perturbé ou modifié les résultats obtenus par les trois (03) approches : non-paramétriques, paramétriques et semi-paramétriques, ainsi employées dans notre étude.

En conclusion, on peut dire que le phénomène de résiliation d'un contrat d'assurance automobile, est tributaires au global de trois (03) principaux facteurs à savoir, l'âge du véhicule, le Bonus-Malus et la formule d'assurance.

Bibliographie

[1] Boukhetala K, Marion J M (2010). *Apport des méthodes de durée de vie au domaine de l'assurance. Application aux contrats d'assurances automobiles*.

[2] Boukhetala K (2001). *Etude de données statistiques et tarification en assurance Responsabilité Civile Automobile* Projet CNA. Ministère des Finances.

[3] Collett D (2003). *Modelling Survival Data in Medical Research*, Second Edition. Chapman et Hall.

[4] Courilleau E, Marion J M (1999). *Comparaison de modèles d'estimation de la fonction de survie appliquée à des données routières*. Revue de statistique appliquée, tome 47, N.1, P. 81-97.

[5] Kalbfleisch J D, Prentice R L (2002). *The Statistical Analysis of Failure Time Data*, Second Edition. New York : Wiley and Sons, Inc.

[6] Kleinbaum D G, Klein M (2005). *Survival Analysis, A Self-Learning Text*, Second edition. Springer.

[7] Lawless J F (2003). *Statistical Models and Methods for Lifetime Data*, Second Edition. Wiley.

[8] Lee E T, Wang J W (2003). *Statistical Methods for Survival Data Analysis*, Third Edition. Wiley.

[9] Machin D, Cheung Y B (2006). *Survival Analysis*, Second Edition. Wiley.

[10] Marion J M (2008). Analyse des Durées de Vie et Applications Conférence, USTHB.

[11] Nikulin M, Bagdonavicius V (2002). *Accelerated Life Models, Modeling and Statistical Analysis*. Chapman et Hall.

[12] Perrigot R, Cliquet G, Mesbah M (2004). *Possible Applications of Survival Analysis in Franchising Research*, Int. Rev. of Retail, Distribution and Consumer Research, Vol.14, N.1, 129-143.

[13] Selvin S (2008). *Survival Analysis for Epidemiologic and Medical Research*, Practical Guide. Cambridge University Press, New York.

[14] Thierry T (2004). *Les assurances*. Publibook, Paris.