$N°$**D'ordre :** $31/2011-M/MT$

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene
Faculté de Mathématiques

Mémoire

Présenté pour l'obtention du diplôme de Magister

En : Mathématiques

Spécialité : Probabilités et Statistiques

Par : EL MAHDAOUI Assia

Thème

Modélisation actuarielle de la durée de vie d'un contrat d'assurance.

Soutenu publiquement le 10/07/2011, devant le jury composé de :

$M.$	MOULAI	Mustapha	Professeur	à l'USTHB	Président
$M.$	BOUKHETALA	Kamal	Professeur	à l'USTHB	Directeur de mémoire
$M.$	SADKI	Ourida	Maître de Conférence (A)	à l'USTHB	Examinatrice
$M.$	OUAFI	Rachid	Maître de Conférence (A)	à l'USTHB	Examinateur

Remerciements

Je souhaite tout d'abord exprimer ma reconnaissance à Monsieur K.BOUKHETALA pour la direction de ce mémoire et son soutien constant.J'ai beaucoup apprécié sa disponibilité et je le remercie pour ses remarques pertinentes.

Un grand merci également aux membres de mon jury, pour l'intérêt porté à mon travail et le temps qu'ils ont consacré à la lecture de ce présent manuscrit.

Je dédie ce modeste travail à Mohand Arab et sa famille ainsi à mon frère Lotfi, mes soeurs Iméne et Lynda et à tous mes proches.

Merci, enfin, à ma mère pour son soutient, pour sa patience, sa compréhension et ses encouragements tout au long de mon cycle universitaire.

Table des matières

Table des figures

Introduction générale

Comme toutes les entreprises présentes sur un marché concurrentiel, les sociétés d'assurances naissent, et l'on ne voit pas pourquoi elles ne pourraient pas mourir. Il est vrai que, dans le monde de l'assurance, les contrôles opérés sont assez strictes : la commission de contrôle, mais aussi les fédérations professionnelles sont là pour « veiller au grain ».

Auparavant les marchés d'assurances étaient réguliers, moins volatils et les rendements exigés par les actionnaires étaient de fait moins élevés.
Progressivement, la compétition s'est accrue, la sinistralité a augmenté avec une plus grande concentration des risques assurés, l'environnement juridique est devenu de plus en plus incertain ; pour cela la faillite des compagnies d' assurances n'est pas, désormais, un risque théorique. Il est donc nécessaire pour l'assureur d'avoir toujours à l'esprit les relations entre tous les risques de son portefeuille lorsqu'il offre différents produits.
Les causes de défaillance des entreprises d'assurances observées de par le monde sont les suivantes :

- le risque à la souscription : la société a mal souscrit, mal tarifié en dommages comme en vie ;
- la mauvaise gestion des actifs et leur inadéquation au passif ;
- une mauvaise réassurance, offrant une couverture mal calculée ;
- un mauvais calcul des provisions techniques ;
- mal préoccupation de la qualité des assurés.

Dans ce nouveau contexte, les assureurs sont désormais fortement incités à développer une gestion optimale de leurs fonds propres qui doit satisfaire des intérêts divergents.

Dans ce travail, nous nous intéressons à l'estimation des durées de vie moyennes de contrats dans une compagnie d'assurance automobile sur le marché Algérien en fonction des caractéristiques suivantes : l'age du conducteur et l'ancienneté de son permis, durée de mise en circulation de la voiture et le bonus - malus afin de mieux gérer la résiliation de contrats.
Pour cette finalité, tout en utilisant les logiciels EXCEL et EXCEL-STAT nous avons tout d'abord focalisé notre attention sur l'analyse de base de données sinistralités matérielles qui a servi de

plate- forme pour ce travail ; ceci nous a permit de trouver que l'age moyen d'un contrat est d'environ 8*ans* .
Dans une seconde étape,et sous le logiciel SPSS nous avons brièvement passé au modèle non-paramétrique par Kaplan-Meier qui nous a indiqué que l'age de circulation du véhicule et le genre du conducteur ainsi que l'ancienneté du permis ont un impact significatif sur le niveau de risque et par conséquent la résiliation du contrat ; contrairement à la variable bonus-malus qui n'influe pas sur cette durée.
Par suite, nous avons complété l'etude par un modèle semi-paramétrique intitulé modèle de COX appliquer sur les différentes variables exogènes pour pouvoir tester et quantifier les caractéristiques individuelles des variables sur le risque de résiliation afin d'ajuster le bon modèle.

Chapitre 1

Généralité et concepts de l'assurance

Sur le plan pratique, les sciences actuarielles concernent l'application des méthodes mathématiques et statistiques appliqué aux finances et aux assurances. Plus précisément, ces méthodes mathématiques et statistiques sont utilisées dans le but d'appréhender la notion de risque lié aux activités économiques.
C'est dans cette optique que les actuaires sont formés à l'utilisation de méthodes quantitatives permettant de mieux connaître les risques afin de les estimer et de limiter leur impact sur l'avenir économique et financier d'une personne, d'une entreprise ou d'un état.

La statistique a depuis sa naissance, bien entendu dépassé le cadre des probabilités et de leurs strictes application aux jeux de hasard. C'est une discipline scientifique qui s'est ouverte à tous les domaines de la médecine à la finance, la statistique a également fourni très tôt à l'actuaire des outils de tarification scientifiquement rigoureux.

Le risque est un aspect du travail de l'actuaire, il s'agit de le quantifier et surtout de proposer aux agents économiques des outils de prévention, à travers le calcul de primes d'assurances. Ces outils de prévention sont censés les mettre à l'abri des catastrophes que si elles sont financières à la base, ne sont souvent pas moins humaines ou sociales ensuite,c'est ainsi que l'élément financier est intrinsèquement lié à l'activité de l'assurance.

Les actuaires ont donc dès le début de leur activité fait appel à la mathématique financière élémentaire afin d'évaluer les réserves et les primes dans une optique temporelle.
Or, la statistique connaît de permanente évolution due notamment au dynamisme des chercheurs mais également aux progrès techniques qui mettent entre les mains des statisticiens des outils informatiques puissants.
La statistique se découvre donc de nouvelles facettes, de nouveaux champs d'application et continue de se voir comme un outil puissant applicable à de nombreux domaines ; non seulement

elle fournit des pistes directement exploitables en actuariat, mais elle pousse aussi beaucoup de développements nouveaux en finance.

1.1 Qu'est-ce que l'actuariat ?

1.1.1 Actuaire et actuariat

Si l'étymologie du mot "Actuaire" est latine (comptable, rédacteur des livres de comptes), ce terme n'apparaît qu'au $XVIII^{\circ}$ siècle, repris de l'anglais "actuary".
Le dictionnaire Larousse le définit ainsi :

Actuaire : "spécialiste qui fait des calculs statistiques pour les assurances".

Les mots "actuariat" (fonction d'actuaire) et "actuariel" (calcul effectué par des actuaires) se définissent par rapport à l'actuaire. De façon un peu plus large et moderne, les actuaires proposent :

"*spécialiste de l'analyse et du traitement des impacts financiers du risque*".

1.1.2 Petite histoire de l'actuariat

La profession actuarielle a vu le jour au milieu du $XVIII^{\circ}$ siècle au Royaume Uni, période d'essor simultané de l'assurance (création des premières compagnies d'assurances officielles par George I$^{\circ}$ en 1720) et de la statistique (travaux de Bernoulli, Galton, Gauss...). Mais l'actuariat, tel que nous le connaissons aujourd'hui, est né un peu plus tard avec la révolution industrielle, notamment les multiples formes d'assurances conçues pour réduire les risques.

Au milieu du XIX° siècle, un "Institut des actuaires" voyait le jour à Londres et une "Faculté des actuaires" à Edimbourg. Aux Etats Unis, l'Actuarial Society of America (ASA) devenait la structure de la profession de ce pays en 1889. Elle précédait de peu l'Institut des Actuaires Français (IAF) crée en 1890.

Aux Etats Unis, les actuaires étaient moins de 100 en 1889. Mais, dès 1900, l'Actuarial Society of America propose des examens et une certification. A la base, les actuaires travaillaient pour les compagnies d'assurances vie. En 1909, appara^une seconde association d'actuaires vie (AIA), puis une d'actuaires dommages (CAS) en 1914 et peu après, une quatrième spécifique aux mutuelles (FAA). Aujourd'hui subsistent deux organisations, une vie(SOA) et une dommages (CAS).

A partir des années 50, le développement des assurances maladie devient le principal facteur de développement, et révèle un intérêt croissant des pouvoirs publics pour l'actuariat.

Les années 70 constituent un tournant, du fait de la crise non prévue, et de ses conséquences la législation nord américaine accroît la responsabilité des actuaires pour la certification des réserves, rôle qui existait déjà dans le domaine des pensions depuis les années 60.

Les années 80 marquent la fin des certitudes. Les modèles considérés comme improbables sont largement dépassés. Il faut cependant proposer des produits de plus en plus attractifs du fait de la concurrence. L'actuariat continue de se développer.
Le début des années 90 est marqué par la chute de grandes sociétés d'assurance vie aux Etats Unis. Le défi de l'actuariat dans ce pays est aujourd'hui de rétablir la confiance et l'intégrité financière.

1.2 Risques

Dans le monde extrêmement complexe qui est le nôtre, les exemples abondent pour lesquels il est nécessaire de pouvoir quantifier « l'incertain ». Pour évaluer ce qu'il en coûte à la société de tolérer des activités potentiellement dangereuses et pour permettre aux citoyens d'effectuer des choix, en pleine connaissance de cause, dans la conduite de leur existence quotidienne.
Donnons trois exemples :

1. Les effets du tabagisme sur la santé sont bien connus : cancers des voies respiratoires, troubles cardio-vasculaires, etc... Ces effets peuvent varier d'un individu à l'autre, d'où l'aspect « incertain ». Si de nombreux individus développent un cancer qui peut être précoce, d'autres peuvent fort bien vivre jusqu'à un âge avancé sans connaître de problèmes particuliers et décéder pour d'autres raisons.

2. Le transport aérien a pris une importance grandissante ces dernières années. Les accidents d'avion sont peu fréquents (aspect incertain), en particulier au nombre d'avions en circulation dans le monde. La raison en est une grande maîtrise de la technologie, alliée à des procédures de contrôle très strictes du trafic aérien. Néanmoins, des accidents ont lieu, dont les conséquences sont importantes, faisant de nombreuses victimes.

3. L'automobile est source importante d'accidents avec pertes de vies humaines. Les causes d'accidents sont multiples mais souvent reliées à l'excès de vitesse. L'expérience montre que, même un conducteur prudent, peut être victime d'un accident mortel (ce qui contribue à l'aspect incertain) sans que sa responsabilité soit mise en cause.

Les trois activités évoquées ci-dessus comportent un risque.

Remarque Le risque nul n'existe pas. L'échelle des risques est logarithmique :un risque peut être réduit, sans jamais être annulé. Ainsi, par exemple, un risque minime parce que la probabilité (P) de l'événement en question est très faible , mais dont les conséquences seraient incalculables.

1.3 Assurance

Définition 1. *L'assurance est un service qui consiste à fournir une prestation pré-définie, généralement financière, à un individu, une association ou une entreprise lors de la survenant d'un risque, en échange d'une cotisation ou d'une prime. Par extension, l'assurance est le secteur économique qui regroupe les activités de conception, de production et commercialisation de ce type de service.*

1.4 Evolution de l'actuariat dans l'assurance

L'actuariat en assurance contribue à l'élaboration de l'offre commerciale par des études de faisabilité et de rentabilité pour le lancement de nouveaux produits, et les tarifications des nouveaux produits ou garanties.
Deux éléments font croître les besoins d'études actuarielles dans l'assurance :

– L'augmentation de la concurrence pousse à développer de nouveaux produits plus adaptés à un segment de clientèle et à proposer les tarifs les plus justes ce qui nécessite des analyses plus poussées et une prévision la plus proche possible de la réalité.

– Les systèmes d'informations sont beaucoup plus puissants aujourd'hui qu'ils ne l'étaient il y a dix ans et les informations disponibles sont plus nombreuses, elles sont surtout plus fiables et plus facilement accessibles.

L' assurance vie est le domaine où l'on trouve le plus d'actuaires. Même si ce secteur ne connaît pas un grand développement, il est probable que les activités d'actuariat vont encore se développer dans ce domaine à cause des évolutions de produits et de marchés, particulièrement importantes dans cette branche.
Les assurances dommages occupent proportionnellement peu d'actuaires et un nombre plus important de statisticiens, mais cela pourrait changer avec l'émergence de nouveaux risques plus complexes comme :

– les risques tempêtes,

– les risques rares ou nouveaux comme ceux liés à la pollution ou à l'internet (produit de couverture contre le risque de virus).

1.5 Contrat d'assurance

C'est le document essentiel, qui matérialise l'accord entre l'assuré et l'assureur.

1.5.1 Les parties au contrat

L'assureur

En général, l'assureur est la partie au contrat qui s'engage à garantir l'assuré contre les risques prévus au contrat, et à payer la prestation indemnitaire ou forfaitaire en cas de sinistre.
Dans la loi des assurances le terme « entreprise d'assurances » désigne aussi «un assureur» est utilisé aussi pour désigner cette partie principale au contrat.
L'entreprise d'assurance est définie par : « une entreprise qui se fonde, s'organise et fonctionne selon les dispositions de la présente loi et d'autres textes juridiques relatifs aux opérations d'assurance et de réassurance ».
Ceci implique que pour pouvoir effectuer des opérations d'assurance , l'entreprise doit, d'une part, être obligatoirement constituée sous l'une des formes prévues par la loi des assurances et d'autre part, faire l'objet d'un agrément par le ministère des finances.

L'assuré

Dans l'assurance , l'assuré est généralement un armateur, ou parfois leur créanciers hypothécaires désireux d'assurer des intérêts financiers ,ou des objets etc. Mais l'armateur reste la personne qui doit payer la prime d'assurance.

1.6 La classification des assurances

1.6.1 Assurances de dommages

Assurance de chose

Indemnisation dans le patrimoine de l'assuré, ou dans l'actif de l'entreprise , soient :
des pertes matérielles directes tel que : les incendies, catastrophes naturelles et tous risques chantier....ou bien des pertes immatérielles, dites indirectes par exemple assurance annulation de voyage ou annulation de spectacle ou de manifestation sportive.

Assurance de responsabilité

L'assurance de responsabilité couvre les dommages causés au tiers (étranger, inconnu) ; elle garantit l'assuré contre les recours exercés contre lui par des tiers cherchant sa responsabilité. Il est possible de faire garantir par un assureur l'amputation faite à son patrimoine par une dette de responsabilité :

– pour les particuliers
  Assurance automobile(près de 50% des encaissements).
– pour les entreprises
  Assurance responsabilité civile exploitation,assurance responsabilité civile (RC) après travaux ou après livraison, elle couvre les conséquences en raison des dommages causés aux tiers du fait des activités des entreprises.

1.6.2 Le principe indemnitaire

Définition 2. *Article du Code des Assurances : L'assurance relative aux biens est un contrat d'indemnité : l'indemnité due par l'assureur à l'assuré ne peut dépasser le montant de la valeur de la chose assurée au moment du sinistre.*
Il peut être stipulé que l'assuré reste obligatoirement son propre assureur pour une somme ou une quantité déterminée, ou qu'il supporte une déduction fixée d'avance sur l'indemnité du sinistre.

Toute forme d'assurance est de remplacer ce qui a été perdu,l'assurance ne vise pas à ce que l'assuré tire profit d'un sinistre mais simplement à ce que sa situation ne soit pas aggravée par le sinistre.
L'assurance de dommage est soumise au principe indemnitaire, selon lequel l'assurance ne peut être une source de bénéfice ou d'enrichissement pour l'assuré mais une indemnité compensatrice d'un préjudice, sinon cela n'encouragerait pas l'assuré à favoriser la réalisation du risque, de plus en assurance, l'assureur n'a que peu de moyen pour contrôler les agissements de l'assuré.

Le principe indemnitaire est défini dans deux situation : le cas de majoration de la valeur assurée et le cas sous-assurance.

1. **Le cas de majoration de la valeur assurée**
   Il y a une majoration de la valeur assurée lorsque la somme assuré est supérieure à la valeur de l'objet assuré. Cet excès peut résulter, soit d'un contrat unique (sur-assurance), soit de plusieurs contrats (assurance cumulative).

   – La sur-assurance :
     Le contrat de sur-assurance est un contrat dans lequel la somme assurée est supérieure à la valeur vénale de la chose assurée (la valeur financière estimée) au moment de la souscription du contrat.

- L'assurance cumulative :
  Le cumul d'assurance est une situation dans laquelle le souscripteur a assuré auprès de plus d'une entreprise d'assurance, par plusieurs polices pour un même intérêt, avec les mêmes conditions de garantie et contre un même risque.

2. **Le cas de sous-assurance**
   A l'inverse du cas précédent, le contrat de sous-assurance est un contrat dans lequel la somme assurée est inférieure à la valeur vénale de la chose assurée et l'indemnité ne pourra dépasser la valeur réelle de la chose assurée quelque soit l'intention.

1.7 Concepts importants d'assurance

Voici la définition de certains termes importants d'assurance :

Assurance Engagement donné par contrat, par un assureur à un assuré, de garantir en cas de survenance d'un événement incertain affectant sa personne, ses biens.
Cette garantie est donné contre le paiement d'une cotisation.

Assurance permanente Assurance vie qui protège l'assuré durant sa vie , à condition que les primes soient payées.

Assurance soins de longue durée Assurance qui prend en charge les soins personnels et certains soins médicaux de la personne qui est incapable d'accomplir les actes de la vie quotidienne. Les soins peuvent être dispensés à domicile ou dans un établissement de soins de longue durée.

Assurance temporaire Assurance vie qui n'est en vigueur que pendant une période déterminée.

Assurance vie entière Assurance vie permanente dont la prime est fixée au moment de la souscription et n'augmente jamais.

Assurance vie universelle Assurance vie permanente qui comporte également une composante de placements à l'abri de l'impôt.

Attestation d'assurance Document écrit remis par l'assureur à l'assuré précisent qu'une assurance à été souscrite au profit de celui-ci.

Contrat d'assurance Document qui constate l'engagement réciproque de l'assureur et de l'assuré (souscripteur) ; ce document est composé au moins des conditions générales et des conditions particulières. On parle aussi de *police d'assurance*.

Conditions générales Document qui regroupe l'ensemble des dispositions communes à tous les assurés pour un type de contrat.Il décrit les garanties proposées ainsi que les obligations de l'assuré et de l'assureur.

Conditions particulières Document complétant les conditions générales qui précise la situation et les choix de l'assuré (risque souscrit, renseignements concernant l'assuré, garanties choisis, cotisation, durée du contrat...).

Garantie Couverture d'un risque par l'assureur en contre partie d'une cotisation.

Risque événement incertain (qui n'a pas encore eu lieu) affectant une personne,ses biens.

Souscripteur Personne physique ou morale qui conclut un contrat d'assurance avec l'assureur.

Échéance Date à laquelle le contrat d'assurance prend fin ou se reconduit automatiquement.

Sinistre Réalisation de l'événement incertain, créant des dommages.

Responsabilité civile (RC) Obligation légale pour toute personne de réparer les dommages causés.

Bénéficiaire Personne physique ou morale qui a le droit de recevoir les sommes dues au titre de l'assurance.

Conseiller autorisé en assurance Personne qui est autorisée à représenter l'assuré ou l'assureur pour traiter de questions d'assurance.

Franchise Somme fixe qu'un assuré doit débourser avant que l'assureur ne verse des indemnités à l'égard d'un sinistre.

Invalidité Incapacité de travailler en raison d'une maladie ou d'une blessure.

Montant d'assurance Montant de la protection garantie par la police d'assurance.

Police Document juridique qui constate l'existence et les conditions de l'assurance.

Cotisation Versement effectué par le souscripteur ou l'adhérent en contre partie des garanties accordées par l'assureur.
Pour les contrats d'assurances autres que les contrats d'assurance vie, le non paiement de la cotisation entraine de la déchéance de garantie (c'est à dire la fin de cette garantie) ; la cotisation pour l'assuré est également appelé *prime*

Assurance automobile L'assurance automobile a pour objectif premier de garantir le conducteur d'un véhicule automobile contre les conséquences de dommages matériels ou corporels causés par son véhicule à des tiers.
C'est une assurance obligatoire.

1.8 Le système de prévention Bonus-Malus

Réduction (bonus) ou Majoration (malus) du montant de la cotisation de base en assurance automobile ; le niveau du bonus ou malus dépend du nombre d'années d'assurance du conducteur et de sa responsabilité dans des accidents.

1.8.1 Principe

Le bonus-malus est un système de calcul pour l'assurance auto qui permet à l'assureur de faire varier le montant de la cotisation annuelle selon la quantité de sinistres déclarés au cours de l'année précédente.
Ainsi, si l'assuré n'avez eu aucun accident, il bénéfice d'une réduction du montant de sa cotisation, appelée *Bonus*. Par contre, s'il avait eu un ou plusieurs accidents, le montant de sa cotisation sera augmenté, il s'agit du *Malus*.
Les règles applicables à ce système de bonus malus sont déterminées par un arrêté ministériel,elles sont donc les mêmes pour tous les assurés, et quelle que soit la société d'assurance.

1.8.2 Fonctionnement

Lorsque vous vous assurez, sans bonus ni malus, vous êtes redevable d'une cotisation de base, dépendant des tarifs de votre assureur. C'est *la cotisation de référence*.
Puis, chaque année l'assureur examine votre situation pour calculer votre cotisation. Pour cela, il prend en compte tous les sinistres intervenus au cours d'une période d'un an précédant de deux mois l'échéance annuelle du contrat.

1.8.3 Règles principales

- Le bonus maximal est obtenu après 13 années consécutives sans accident et est plafonné à 50% (soit un coefficient de 0,50).

- Un malus sera appliqué si l'assuré est responsable (partiellement ou totalement) d'un accident. S'il est entièrement responsable le coefficient est majoré de 25%,par contre si la responsabilité est partielle , le coefficient est majoré de 12,5%.

- Le malus maximal est plafonné à 3,5%.

Conclusion

Dans ce premier chapitre, nous avons introduit les concepts de base de l'actuariat et de l'assurance dans un but qu'il soit une source d'information pour être consultés de façon indépendante selon les besoins.

Chapitre 2

Analyse de survie

2.1 Notions préliminaires

L'analyse des durées de vie réfère aux méthodes employées pour l'étude du temps jusqu'à l'occurrence d'un certain événement, comme un décès ou l'apparition de symptômes. Nous appelons *temps de survie* les données qui mesurent le délai entre le début de l'étude et la manifestation de l'événement.

Les méthodes standards d'analyse statistique sont in-appropriées pour les données de survie. En effet, ces données présentent plusieurs particularités, notamment une distribution non symétrique.
Pour cette raison, de nouveaux modèles de distribution doivent être adoptés. De plus, les temps de survie sont fréquemment censurés.

Par définition, le temps de survie (Kalbf'Leisch.J.D Ross.L.P,(2002)) d'un individu est dit censuré, lorsque sa valeur exacte n'est pas observée ; seules une des bornes supérieures ou inférieures pour cette valeur est disponible. La censure peut se manifester pour différentes raisons : l'événement d'intérêt n'est pas survenu au moment de l'analyse, un sujet peut être perdu de vue avant d'avoir expérimenté l'événement d'intérêt, un événement concurrent peut être survenu avant l'événement d'intérêt, un sujet peut être exclu de l'étude sans avoir expérimenté l'événement d'intérêt.

Il existe plusieurs types de censure dont la censure à droite, la censure à gauche et la censure par intervalle.
La censure à droite est la forme de censure la plus commune dans les études médicales.
Par exemple, lors d'un essai clinique pour tester la survie de souris à un nouveau virus, certaines souris decedent entre le moment de l'injection du virus et la date de la fin de l'étude, tandis que

d'autres survivront jusqu'à la fin.
La durée de vie des souris non décédées à la fin de l'étude sera censurée à droite.

2.2 La durée de survie

Définition 3. *Le terme de durée de survie est employé de manière générale pour désigner le temps qui s'écoule jusqu'à la survenue d'un événement particulier qui n'est pas forcément la mort : il peut s'agir par exemple d'une rechute et la durée de survie sera dans ce cas, un délai de rémission ou de la guérison et la durée de survie représente le délai allant jusqu'à la guérison.*

2.3 Fonctions liées à la survie

La distribution des temps de survie est généralement caractérisée par trois fonctions : la fonction de survie, la fonction de densité et la fonction de risque. Ces trois fonctions sont en fait inter-reliées car la connaissance d'une seule est suffisante pour dériver les autres.
Les sections suivantes présentent ces trois fonctions liées à la survie.

2.3.1 La fonction de survie

Distribution continue

Nous noterons T $(\mathrm{T} \geq 0)$, la variable aléatoire qui représente la durée de vie d'une unité dans une expérience.

La fonction de survie, noté $\mathrm{S}(t)$, est définie comme la probabilité qu'un individu survive au-delà du temps t, c'est à dire qu'il expérimente l'événement après le temps t :

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

Lorsque T est une variable aléatoire continue, la fonction de survie est le complément de la fonction de répartition, car

$$S(t) = 1 - P(un\ individu\ echoue\ avant\ le\ temps\ t) = 1 - F(t),$$

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

La fonction de densité $f(t)$, peut être calculer à partir de la fonction de survie ou de la fonction de répartition :

$$f(t) = \frac{dF(t)}{dt} = \frac{d\{1 - S(t)\}}{dt} = -\frac{dS(t)}{dt}$$

De plus , si la fonction de densité est connue alors la fonction de répartition et la fonction de survie peuvent être trouvé en intégrant la fonction de densité $f(t)$:
$F(t) = \int_0^t f(x)dx$ et de même la fonction de survie $S(t) = \int_t^\infty f(x)dx$.

La fonction de survie S(t) possède trois caractéristiques importantes :

1. $S(t)$ est une fonction monotone décroissante ;
2. $S(t) = 1$ pour $t = 0$;
3. $S(t) = 0$ pour $t = \infty$.

La représentation graphique de la survie est appelée *courbe de survie*, une fonction monotone non croissante de la probabilité de survie en fonction du temps.

2.3.2 La fonction de risque (the hazard function)

Cas absolument continu

La fonction de risque est aussi appelée taux de panne, taux de décès conditionnel ou force de mortalité c'est la probabilité de décès dans un intervalle par unité de temps parmi les sujets encore vivants au début de l'intervalle.
Elle est définie par :

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T \leq t + \Delta t / T \geq t]}{\Delta t}.$$

Cette expression peut être modifiée pour obtenir une autre forme :

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \{\frac{P(t \leq T \leq t + \Delta t)}{\Delta t . P(T \geq t)}\} && (2.1) \\ &= \lim_{\Delta t \to 0} \{\frac{F(t + \Delta t) - F(t)}{\Delta t . P(T \geq t)}\} && (2.2) \\ &= \frac{f(t)}{S(t).} && (2.3) \end{aligned}$$

La fonction de survie peut être exprimer par la fonction de risque

$$h(t) = -\frac{d}{dt}\{logS(t)\},$$

on integrant $logS(t) = -\int_0^t h(x)dx$, on aura :

$$S(t) = exp\{-\int_0^t h(x)dx\}$$

La fonction de risque cumulé est définie par :

$$H(t) = \int_0^t h(x)dx.$$

Et la fonction de survie peut s'écrire :

$$S(t) = \exp\{-H(t)\}.$$

2.4 Les censures

Pour bien comprendre voici un exemple [17] :

Exemple considérons une étude du $(6M-P)$ une thérapie d'entretien pour des enfants pour une remise de la leucémie aiguë.
Cette étude a pour but de comparer les durées de rémission des sujets atteints de la leucémie selon qu'ils ont reçu du $(6M-P)$ ou un placebo pour le groupe témoin , c'est le traitement habituel.

6 M-P	6,	6,	6,	6^+,	7,	9^+,	10,	10^+,	11^+,	13,	16,	17^+,
	19^+,	20^+,	22,	23,	25^+,	32^+,	32^+,	34^+,	35^+.			
Placebo	1,	1,	2,	2,	3,	4,	4,	5,	5,	8,	8,	8,
	8,	11,	11,	12,	12,	15,	17,	22,	23.			

FIG. 2.1 – Données d'étude

Les quarante-deux patients de thérapie d'induction partager on deux groupes de nombres égaux.

Tous les chiffres suivis du signe $+$ correspondent à des patients qui ont été perdus de vue à la date considérée ils sont donc exclus vivants de l'étude , on sait seulement que leurs durée de survie est supérieure à celle indiquée.
Par exemple, le quatrième patient traité, par $(6M-P)$ a eu une durée de rémission supérieure à 6 semaines,On dira que les perdus de vue ont été *censurés*.

Une telle survie serait *censurée à droite* parce que sur un graphique le temps de rechute se trouverait quelque part à la droite de la période de censure.

En effet, si l'on se contentait d'éliminer les observations incomplètes c'est à dire les 12 patients censurés du groupe traité par le $(6M-P)$ on perdrait beaucoup d'information,car on ne tiendrait pas compte des patients qui ont justement les durées de rémission les plus longues.

2.4.1 Les trois types de censure [30]

Censure fixe (non aléatoire de type I)

Au lieu d'observer toutes les variables $X_1,\dots,X_n$, on n'observe que la variable X_i si X_i est inférieur ou égal à une durée fixé C noté $X_i \leq C$.

On note : $T_i = X_i \wedge C$ où le signe $\wedge$ signifie :
($a \wedge b = min(a,b)$, la plus petite des deux valeurs a et b).

La censure gauche indique le fait que nous ne pouvons pas déterminer la date initiale de l'événement étudié, par exemple la date de l'acquisition d'un produit pour personnes renseignées sur leur achat. Nous pouvons avoir des informations pour lesquelles il y a un point de fin (fin d'observation) , mais aucunes informations sur quand le sujet a été exposé la première fois au risque(début de l'étude).

Définition 4. *Censure à gauche*
On a une Censure à gauche si le temps de l'événement t est connu pour être inférieur ou égal à un certain point de coupure, T. On d'autres mots, des données sont censurées à gauche si la date initiale de l'événement est inconnue.

D'autre part si la variable X_i est supérieur à la durée fixé C noté $X_i > C$; il s'agira d'une censure à droite.

Définition 5. *Censure à droite*
Une observation serait censurée à droite si le temps t de l'événement est supérieur ou égal à un certain point de coupure,T .
On d'autres termes, censure à droite signifie que l'événement d'intérêt ne se produit pas au cours d'une période d'observation de longueur T(soit jamais ou bien après T).

Censure d'attente (type II)

On décide d'observer les durées de survie des n patients jusqu'à ce que r d'entre eux soient décédés et d'arrêter l'étude à ce moment là.
Si on ordonne les durées de survie $X_1,\dots,X_n$, tel que $X_{(1)}$ soit la plus petite et $X_{(i)}$ la i^{eme} durée de survie etc.....

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

On dit que les $X_{(i)}$ sont les statistiques d'ordre des X_i. La date de censure est alors $X_{(r)}, r^{eme}$ panne et on observe :

$$\begin{aligned} T_{(1)} &= X_{(1)} \\ T_{(2)} &= X_{(2)} \\ T_{(r)} &= X_{(r)} \\ T_{(r+1)} &= X_{(r)} \\ \ldots\ldots\ldots & \\ T_{(n)} &= X_{(r)} \end{aligned}$$

Désormais, nous nous plaçerons dans le cadre d'un mécanisme de censure à droite.

Censure aléatoire

Associons à chaque individu i son temps de survie X_i et aussi un temps de censure C_i.
On n'observera évidemment que le plus petit des deux, c'est à dire :

$$T_i = X_i \wedge C_i.$$

Mais on peut supposer que tout comme les $X_i$et les C_i sont indépendantes et équi-distribuées (i.i.d) de fonction de répartition G.
La censure aléatoire lors d'une étude peut avoir plusieurs causes comme :

1. Perte de vue : l'individu quitte l'étude, par exemple le patient decide de se faire soigner ailleurs et on ne le revois plus ;

2. Arrêt de l'étude : l'étude se termine alors que certain individus sont toujours vivants.

Exemple

La figure (2.2)représente la survie de 3 patients. Le premier est entré au début de l'étude et il est mort à la date $X_1 = 6$. Le deuxième était toujours vivant à la fin de l'étude, il est donc censuré à $t = 10$. Le dernier patient à été perdu de vue avant la fin de l'étude, il a était censuré à $t = 7$

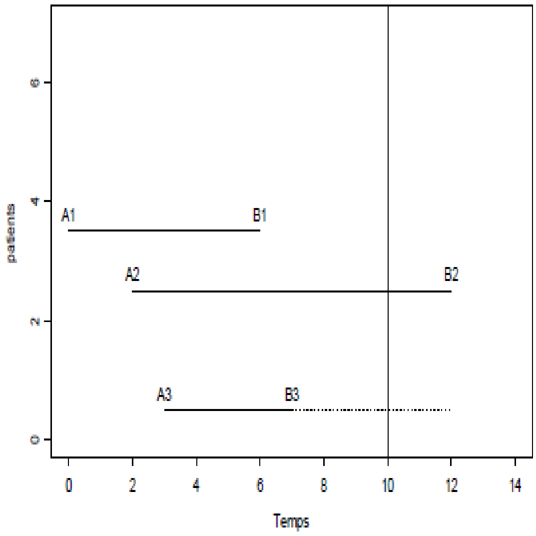


FIG. 2.2 – Exemple : 3 patients

Conclusion

L'analyse de survie est devenue un outil important dans l'étude des durées de vie dans différent domaines tel que la médecine,l'informatique et l'assurance.
De ce fait, nous avons présenté les différentes notions de base de la survie pour pouvoir les utiliser dans la suite de ce travail.

Chapitre 3

Les modèles paramétriques

Les modèles paramétriques sont de distribution en probabilité connu durant un temps de survie à la différence des modèles semi paramétriques (aller au chapitre 5) qui ne sont pas limité a une forme spécifique et définit sur un temps de survie inconnu, mais reste qu'ils ont une application bien répondu en pratique.

3.1 Modèles de fonction de risque

Soit un modèle pour un temps de survie de fonction de survie et de risque respectivement données par :

$$S(t) = 1 - \int_0^t f(u)du \tag{3.1}$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\Big[\log S(t)\Big] \tag{3.2}$$

Tel que $f(t)$ est la densité de probabilité du temps de survie.

$$S(t) = \exp\{-H(t)\} \tag{3.3}$$

de l'équation (3.2)

$$f(t) = h(t).S(t) = -\frac{dS}{dt} \tag{3.4}$$

Avec $H(t) = \int_0^t h(u)du$

3.2 Risque instantané constant

La distribution exponentielle C'est l'unique distribution continue qui admet un risque contant et égale à un paramètre λ.
Dans ce modèle la fonction de risque est : $h(t) = \lambda$ pour $0 \leq t < \infty$, le paramètre λ est une constante positive à estimer.

De l'équation (3.2), la fonction de survie

$$S(t) = \exp\left\{-\int_0^t \lambda.du\right\}$$

$$S(t) = \exp(-\lambda t) \tag{3.5}$$

et de densité

$$f(t) = \lambda\exp(-\lambda t), \quad 0 \leq t < \infty,$$

la moyenne est $E(T) = \frac{1}{\lambda} = \mu$
Le $p^{\text{éme}}$ quantile temps de survie est donné par :

$$t(p) = \frac{1}{\lambda}\log\left(\frac{100}{100-p}\right) \tag{3.6}$$

Par conséquent le temps médian de l'exponentielle

$$\begin{aligned} t(50) &= \frac{1}{\lambda}\log\left(\frac{100}{100-50}\right) \\ &= \frac{1}{\lambda}\log 2 \end{aligned}$$

et

$$\begin{aligned} S(t(50)) &= \exp\{-\lambda t(50)\} = \exp\{-\lambda.\frac{1}{\lambda}\log 2\} \\ &= \frac{1}{2} \end{aligned}$$

La distribution du $p^{\text{éme}}$ temps de survie est

$$S(t(50)) = 1 - \frac{p}{100} \tag{3.7}$$

Le graphe de la fonction de risque pour trois valeurs de λ, $\lambda = 1,\ 0.1,\ 0.01$ est le suivant :

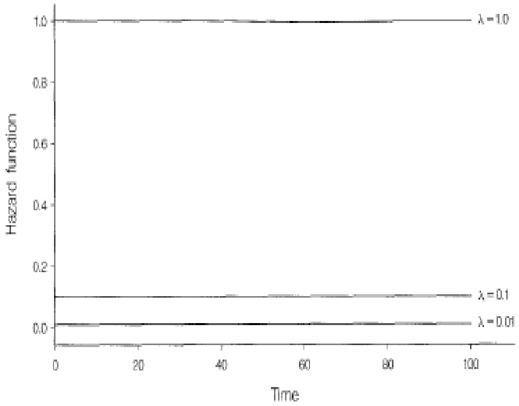


FIG. 3.1 – fonction de risque d'une exponentielle

Les moyennes de la distribution exponentielle pour $\lambda = 1,\ \ 0.1,\ \ 0.01$ sont respectivement $1,\ \ 10,\ \ 100$ et le temps médian de survie est respectivement $0.69,\ \ 6.93,\ \ 69.31$.

Le graphe de la fonction densité est si dessous :

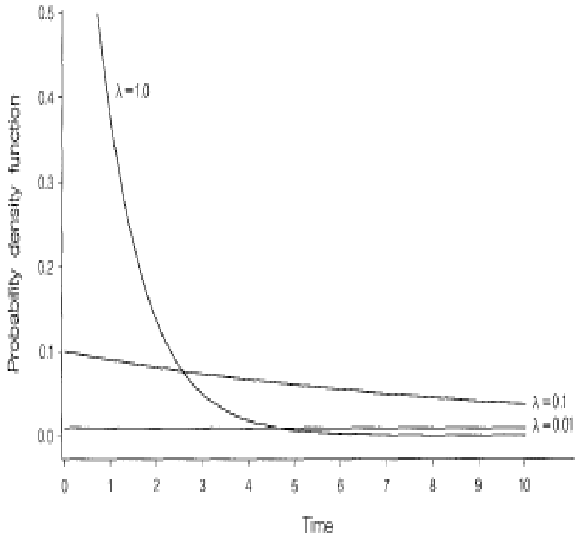


FIG. 3.2 – fonction de densité d'une exponentielle

3.3 Risque instantané monotone

3.3.1 La distribution de Weibull $W(\lambda, \gamma)$

C'est la généralisation de la loi exponentielle (obtenu dans le cas particulier $\lambda = 1$),la forme général de la fonction de risque est :

$$h(t) = \lambda\gamma . t^{\gamma-1}, \quad 0 \leq t < \infty \tag{3.8}$$

cette fonction dépend de deux paramètres λ et γ, si $\gamma = 1$ alors $h(t) = \lambda$ c'est une distribution exponentielle ;et pour d'autres valeurs de γ la fonction de risque est monotone (croissante ou décroissante)

1. Si $\gamma > 1$, le risque instantané est croissant de 0 à ∞.

2. Si $0 < \gamma < 1$, le risque instantané est décroissant de ∞ à 0.

Remarque 1. *[9] γ est un paramètre de forme(Shape parameter)*
λest un paramètre d'échelle (Scale parameter)

La forme générale de la fonction de risque pour les différents valeurs de γ :

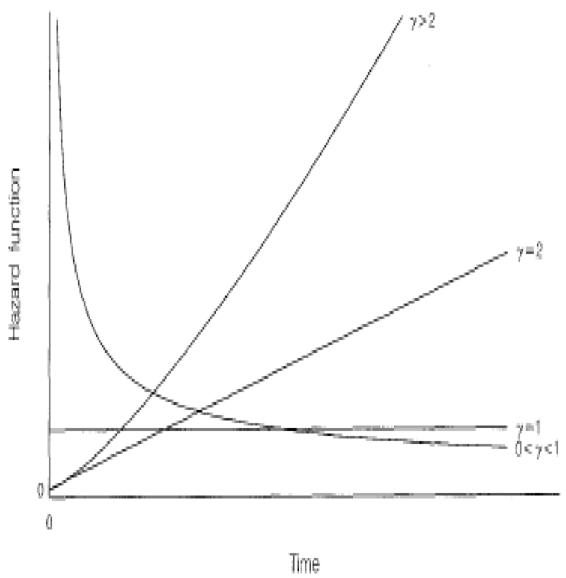


FIG. 3.3 – fonction de risque de weibull

La fonction de survie est :

$$S(t) = \exp\left\{ -\int_0^t \lambda\gamma u^{\gamma-1} du \right\} = \exp(-\lambda t^\gamma), \tag{3.9}$$

la densité

$$f(t) = \lambda \gamma t^{\gamma-1}.\exp(-\lambda t^{\gamma}), \text{ pour } 0 \leq t < \infty \tag{3.10}$$

La moyenne de $W(\lambda, \gamma)$ est donnée par

$$E(T) = \lambda^{\frac{-1}{\gamma}}.\Gamma(\gamma^{-1} + 1)$$

$\Gamma(x)$ est la fonction gamma définit :

$$\Gamma(x) = \int_0^{\infty} u^{x-1}.e^{-u} du,$$

le $p^{\text{éme}}$ temps de weibull $W(\lambda, \gamma)$ est :

$$t(p) = \left[\frac{1}{\lambda} \log\left(\frac{100}{100-p}\right)\right]^{\frac{1}{\gamma}} \tag{3.11}$$

Par conséquent, le temps médian est :

$$t(50) = \left\{\frac{1}{\lambda} \log 2\right\}^{\frac{1}{\gamma}}$$

et

$$\begin{aligned} S(t(50)) &= \exp\{-\lambda.\left(\frac{1}{\lambda} \log 2\right)^{\frac{1}{\gamma}\gamma}\} \\ &= \exp\{-\log 2\} \\ &= \frac{1}{2} \end{aligned}$$

3.3.2 La loi gamma

La moyenne d'une loi gamma est $(\frac{\varphi}{\lambda})$, la variance $(\frac{\varphi}{\lambda^2})$ et une densité égale à

$$f(t) = \frac{\lambda^{\varphi}.t^{\varphi-1}.e^{-\lambda t}}{\Gamma(\varphi)}, \quad 0 \leq t < \infty, \ \lambda > 0 \ \ \varphi > 0; \tag{3.12}$$

de fonction de survie donnée par :

$$S(t) = 1 - \Gamma_{\lambda t}(\varphi) \tag{3.13}$$

tel que $\Gamma_{\lambda t}(\varphi)$ est la fonction gamma incomplète égale à :

$$\Gamma_{\lambda t}(\varphi) = \frac{1}{\Gamma(\varphi)} \int_0^{\lambda t} \mu^{\varphi-1}.e^{-u}.du$$

En remplaçant (3.12) et(3.13)dans

$$h(t) = \frac{f(t)}{S(t)},$$

on aura la fonction risque d'une gamma.

1. Si $\varphi > 1$,la fonction risque $h(t)$ croit.
2. Si $\varphi < 1$,la fonction risque $h(t)$ décroît.

La généralisation d'une gamma est une extension de gamma, on inclus le paramètre θ ([17],[9]) ou $\theta > 0$ et définit par :

$$f(t) = \frac{\theta\lambda^{\varphi\theta}.t^{\varphi\theta-1}.\exp\{-(\lambda t)^{\theta}\}}{\Gamma(\varphi)}, \quad 0 \leq t < \infty \tag{3.14}$$

La fonction de survie de cette distribution est définit par :

$$S(t) = 1 - \Gamma_{(\lambda t)^{\theta}}(\varphi) \tag{3.15}$$

Quand :

1. $\varphi = 1$, la fonction de distribution est une Weibull.
2. $\theta = 1$, la fonction de distribution est une Gamma.
3. $\varphi \to \infty$, la densité est une Log-normal.

3.4 Risque instantané en forme de cloche

3.4.1 La log-normale $LN(\mu, \sigma^2)$

Cette loi est définit par la variable aléatoire positive de densité de probabilité :

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}}.t^{-1}.\exp\{-(\log t - \mu)^2/2\sigma^2\}, \quad \text{pour } 0 \leq t < \infty, \ \sigma > 0 \tag{3.16}$$

et de fonction de survie :

$$S(t) = 1 - \Phi\Big(\frac{\log t - \mu}{\sigma}\Big) \tag{3.17}$$

Ou $\Phi(.)$ est la f.r normal définit par :

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}.\int_{-\infty}^{z} \exp\left(\frac{-u^2}{2}\right) du$$

Le $p^{\text{éme}}$ quantile est

$$t(p) = \exp\left\{\sigma.\Phi^{-1}(p/100) + \mu\right\}$$

3.4.2 La distribution log-logistique

Les fonctions de risque et de survie sont respectivement les suivantes :

$$h(t) = \frac{e^{\theta}.\kappa.t^{\kappa-1}}{1+e^{\theta}.t^{\kappa}}, \;\; 0 \le t < \infty, \;\; \kappa > 0 \tag{3.18}$$

$$S(t) = \{1+e^{\theta}.t^{\kappa}\}^{-1} \tag{3.19}$$

et une densité de paramètre θ,κ donné par

$$f(t) = \frac{e^{\theta}.\kappa.t^{\kappa-1}}{(1+e^{\theta}.t^{\kappa})^2} \tag{3.20}$$

Le $p^{\text{éme}}$ quantile de la distribution log-logistique

$$t(p) = \left(\frac{p.e^{-\theta}}{100-p}\right)^{\frac{1}{\kappa}}$$

La distribution de la fonction de risque d'une log-logistique pour $\kappa = 0.5, 2.0 et 5.0$ est si dessous, les valeurs correspondantes de θ pour cette distribution sont respectivement $-1.5, -6.0 et -15.0$

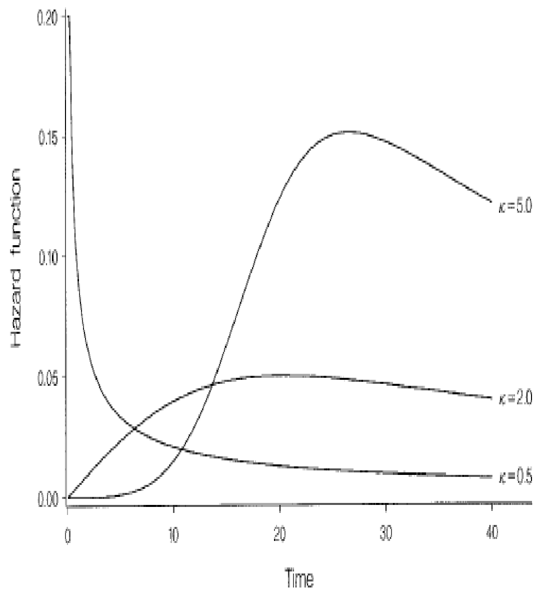


FIG. 3.4 – fonction de risque d'une log-logistique pour $\kappa = 0.5, 2.0 et 5.0$

Conclusion

L'approche paramétrique stipule que les intensités de transition appartiennent à une classe particulière de fonctions, qui dépendent d'un nombre fini de paramètres. L'avantage de cette approche est la facilité attendue de la phase d'estimation des paramètres. L'inconvénient est l'inadéquation pouvant exister entre le modèle retenu et le phénomène étudié.
Maintenant nous aborderons dans la quatrième partie les modèles non paramétriques.

Chapitre 4

Les modèles non paramétriques

A la différence d'un modèle paramétrique qui spécifie la loi de chaque observation par la valeur d'un paramètre, un modèle non paramétrique laisse beaucoup plus de souplesse à la forme et à la nature possible des lois des observations. Bien que les modèles paramètriques soient particulièrement appréciés pour leurs simplicités et la précision des descriptions offertes, ils sont peux envisageables.
Pendant très longtemps, le développement des méthodes statistiques et l'étude de leurs propriétés ont été fondés essentiellement sur la normalité de la famille de lois. Or la validité d'un tel modèle n'est pas toujours assurée.
En effet, différents facteurs comme des erreurs d'expérimentation, des erreurs de mesures ou encore des perturbations aléatoires non prises en compte dans le modèle peuvent rendre un modèle paramétrique inexact.

4.1 Estimation de la fonction de survie

Supposons qu'on a un simple temps de survie et sans censure,la fonction de survie $S(t)$ définie par :

$$S(t) = P(T \geq t) = 1 - F(t),$$

est la probabilité qu'un individu survie pour un temps supérieur ou égale à t.

Cette fonction peut être estimée et donner par :

$$\hat{S}(t) = \frac{\text{nombre d'individus avec un temps de survie} \geq t}{\text{nombre d'individus dans l'ensemble des données}}; \qquad (4.1)$$

nommée *la fonction de survie empirique* et équivalente à écrire :

$$\hat{S}(t) = 1 - \hat{F}(t);$$

c'est le rapport du nombre de personnes vivants jusqu'à l'instant t sur le nombre d'individus dans l'étude tel que $\hat{F}(t)$ est *la fonction de distribution empirique*.

Notons que la fonction de survie empirique est égale à 1 pour valeurs de t inférieure au 1er décès.

Le graphe de $\hat{S}(t)$ en fonction de t est un diagramme en escalier, constant entre deux moments adjacents, tel que la fonction décroît immédiatement après chaque temps de survie.

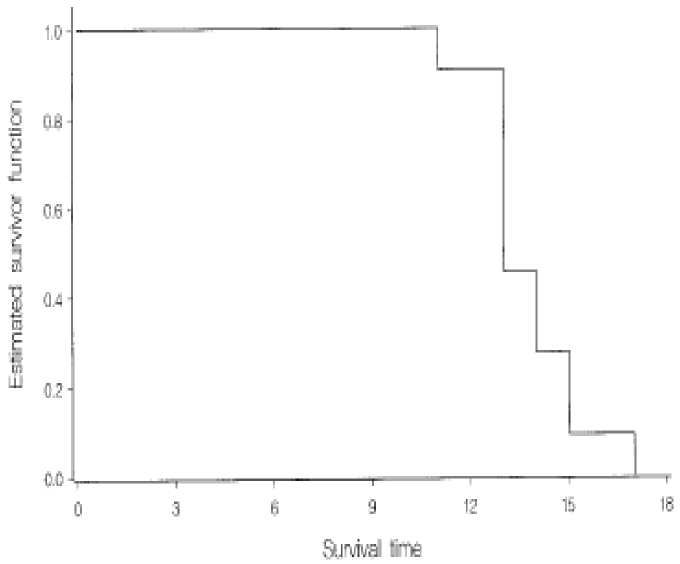


FIG. 4.1 – Diagramme de la fonction de survie estimé

Cette méthode d'estimation de la fonction de survie ne peut malheureusement être employée lorsque nous sommes en présence de temps de survie censurés,car cette fonction n'apporte pas d'information sur les temps de survies censurés.
Par ailleurs,Il existe des méthodes non paramétriques pour l'estimation de $S(t)$ qui sont utilisées en présence de la censure du temps de survie.

4.1.1 La méthode actuarielle

Cette approche consiste en premier lieu à diviser la période d'observation en une série d'intervalles de temps tel que le nombre d'intervalle dépend du nombre d'individus dans l'étude, mais souvent choisit entre 5 et 15 intervalles et à estimer la proportion de survie conditionnelle pour chaque intervalle.

Supposons que le $j^{\text{éme}}$ intervalle choisi parmi m intervalles, $j = \overline{1,m}$ et compris entre t'_j et t'_{j+1}, Notons par :
d_j : le nombre de mort (décès) ;
c_j : le nombre de survie censurés pendant cette période ;
n_j : le nombre d'individu en vie ou sous le risque de mourir au début du $j^{\text{éme}}$intervalle.

Le nombre de censure est un processus avec une distribution uniforme sur le $j^{\text{éme}}$ intervalle et le nombre de sujets à risque sur tout l'intervalle est donc donné par :

$$n'_j = n_j - \frac{c_j}{2}. \tag{4.2}$$

La probabilité de mort est estimée par $\frac{d_j}{n'_j}$ et celle de survie pour qu'un individu survie au delà de t'_k , $k = \overline{1,m}$ sera donné par $\frac{n'_j - d_j}{n'_j}$;

Alors l'estimation de la fonction de survie sera le produit des probabilités de survies [9]

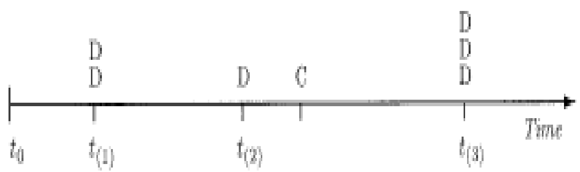
$$S^*(t) = \prod_{j=1}^{k} \left(\frac{n'_j - d_j}{n'_j}\right) \;\; , \;\; pour \;\; t'_k \leq t < t'_{k+1} / \;\; k = \overline{1,m}. \tag{4.3}$$

4.1.2 Estimateur de Kaplan-Meier [12]

Cet estimateur de la fonction de survie a été proposé par Kaplan et Meier (1958).
La méthode de Kaplan-Meier repose sur les mêmes principes que la méthode actuarielle, la première étape de cette analyse est de séparer les données de survie censurés.Cependant, chaque intervalle construit est conçu pour contenir une seul fois le temps de mort.

Pour illustrer l'idée, supposons que $t_{(1)}, t_{(2)}, t_{(3)}$ sont trois temps de survie observés tel que : $t_{(1)} < t_{(2)} < t_{(3)}$ et C est la censure du temps de survie entre $t_{(2)}$ et $t_{(3)}$.
Les intervalles construit commencent par $t_{(1)}, t_{(2)}, t_{(3)}$ et contiennent un seul temps de mort.

La situation est si dessous tel que :
D représente une mort (un décès), C est la censure.



$[t_0,t_1[$ n'inclut pas un temps de mort.
$I_1=[t_1,t_2[$ cet intervalle à un seul temps de mort en t_1.
$I_2=[t_2,t_3[$ une mort à t_2 et une censure C.
$I_3=[t_3,\infty[$ qui contient le plus temps de survie.

Généralisation :
Pour n individus observés sur $t_{(1)},t_{(2)},.....,t_{(n)}$ quelque observations sont censurées à droite.
On a r temps de morts parmi n individus tel que $r\leq n$ et $t_{(1)}<t_{(2)}<.....<t_{(n)}$.
Les intervalles construites sont de forme : $[t_j-\delta,t_j[$, δ un temps infinitésimal qui contient un seul décès.

La probabilité de décès sur cet intervalle est estimée par $\frac{d_j}{n_j}$, par conséquent la probabilité de survie est $\frac{n_j-d_j}{n_j}$ tel que :

- n_j : nombre d'individus vivants juste avant l'instant t_j y compris ceux qui sont sur le point de mourir.
- d_j : nombre d'individus décédés en se temps.

Supposons maintenant que les décès se produisent indépendamment l'un de l'autre ; donc la fonction de survie estimé par Kaplan-Meier sur l'intervalle t_k et t_{k+1} pour $k=1,..,n$ est :

$$\hat{S}(t)=\prod_{j=1}^{k}\frac{n_j-d_j}{n_j} \quad , \quad t_k\leq t<t_{k+1} \quad et \quad k=\overline{1,r} \tag{4.4}$$

Avec

$$\hat{S}(t)=\begin{cases} 1 & \text{pour } t\leq t_1 \\ 0 & \text{pour tout } t_{(r+1)} \end{cases}$$

Si le plus grand temps d'observations est un temps censuré t^*, $\hat{S}(t)$ est non définit pour $t>t^*$.
D'autre part si le plus grand temps de survie observé t_r est une observation non censuré, $n_r=d_r$ alors $\hat{S}(t)$ est égale à zéro pour $t\geq t_r$.
Le diagramme de la fonction de survie estimé par Kaplan-Meier est un diagramme en escalier.

l'éstimateur de Kaplan-Meier est une valeur limite de l'actuarielle pour l'équation (4.3) car le nombre d'intervalle tend vers ∞ et leur largeur tend vers 0, pour cela l'éstimateur Kaplan-Meier est connu sous le non de *produit-limite* de la fonction de survie.

S'il n'y a pas de censure donc : $(n_j-d_j)=n_{j+1}$, $j=\overline{1,k}$ et l'équation (4.4) sera :

$$\hat{S}(t)=\frac{n_2}{n_1}\times\frac{n_3}{n_2}\times....\times\frac{n_{k+1}}{n_k}$$

$$\hat{S}(t)=\frac{n_{k+1}}{n_1} \quad , \quad k=\overline{1,r-1}, \tag{4.5}$$

Avec :

$$\hat{S}(t) = \begin{cases} 1 & \text{pour } t < t_1 \\ 0 & \text{pour } t \geq t_r \end{cases}$$

et

- n_1 : nombre d'individus en risque juste avant le 1^{er} décès.
- n_{k+1} nombre d'individus avec un temps de survie supérieur ou égale à $t_{(k+1)}$.

Par conséquent, dans l'absence de censure,$\hat{S}(t)$ est simplement la fonction de survie estimée en (4.1).
En conclusion Kaplan-Meier est une généralisation de la fonction de survie empirique qui s'adapte à des observations censurées.

4.1.3 Estimateur de Nelson-Aalen [29] [9]

La fonction de survie estimer est donnée par :

$$\tilde{S}(t) = \prod_{j=1}^{k} \exp(\frac{-d_j}{n_j}) \tag{4.6}$$

L'estimateur Kaplan-Meier de la fonction de survie peut être vu comme une approximation à l'estimateur Nelson-Aalen et voici la démonstration :

Démonstration Pour cela nous employons le résultat suivant :

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

l'approximation est égale à $(1-x)$ si x est petit on aura alors :

$$\exp(\frac{-d_j}{n_j}) \simeq 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j} \ , \ d_j < n_j$$

L'estimateur Nelson-Aalen de la fonction de survie est plus grand que l'estimateur Kaplan-Meier car :

$$e^{-x} > 1 - x \ , \ \forall \text{ la valeur de x}$$

4.2 L'erreur standard des estimateurs de la fonction de survie

L'interprétation de n'importe quel quantité à estimer indique une erreur standard, elle est définie par la racine carré de la variance estimée.

L'estimateur Kaplan-Meier est le plus important et le plus utilisé pour l'estimation de la fonction de survie, la dérivation de l'erreur standard de $\tilde{S}(t)$ est présenté si dessous :

4.2.1 L'erreur standard de l'estimateur Kaplan-Meier

l'estimateur Kaplan-Meier pour toute valeurs de t dans l'intervalle t_k à t_{k+1} est donné par :

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j \ , \ k = 1, ..., r \tag{4.7}$$

ou $\hat{p}_j = \frac{n_j - d_j}{n_j}$ est la probabilité qu'un individu survie jusqu'à l'instant t_j, $j = \overline{1, r}$

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

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

Le nombre d'individus vivants au début de l'intervalle en t_j suivent une loi Binomial de paramètres n_j et p_j, $(n_j - d_j) \rightsquigarrow B(n_j, p_j)$, la variance sera :

$$Var(n_j - d_j) = n_j p_j (1 - p_j),$$

sachant que $\hat{p}_j = \frac{n_j - d_j}{n_j}$;

$$\begin{aligned} Var(\hat{p}_j) &= Var(\frac{n_j - d_j}{n_j}) \\ &= \frac{1}{n_j^2} Var(n_j - d_j) = \frac{n_j p_j (1 - p_j)}{n_j^2} \\ &= \frac{p_j (1 - p_j)}{n_j} \end{aligned}$$

La variance de $\hat{p}_j$ est estimé par

$$\hat{p}_j (1 - \hat{p}_j) / n_j \tag{4.8}$$

et pour obtenir l'estimation de la variance de $log(\hat{p}_j)$, on se sert d'un résultat :approximation de la variance d'une fonction de variable aléatoire [9] ; selon ce résultat D-méthode la variance d'une fonction $g(X)$ tel que X est une variable aléatoire

$$Var\{g(X)\} \approx \{\frac{dg(X)}{dX}\}^2 . Var(X) \tag{4.9}$$

(Série approximative de Taylor).
Par conséquent l'approximation de la variance de $\log \hat{p}_j$ sera :

$$\begin{aligned} Var\{\log(\hat{p}_j)\} &\approx \{\frac{1}{\hat{p}_j}\}^2 . Var(\hat{p}_j) \\ &\approx \frac{1}{\hat{p}_j^{\,2}} . \frac{\hat{p}_j (1 - \hat{p}_j)}{n_j} \end{aligned}$$

$$Var\{\log(\hat{p}_j)\} \approx \frac{(1-\hat{p}_j)}{n_j.\hat{p}_j} \tag{4.10}$$

on a : $\hat{p}_j = \frac{n_j - d_j}{n_j}$ alors $1-\hat{p}_j = \frac{d_j}{n_j}$.

Remplaçant

$$\begin{aligned}\frac{(1-\hat{p}_j)}{n_j.\hat{p}_j} &= \frac{\frac{d_j}{n_j}}{\frac{n_j-d_j}{n_j}.n_j} \\ &= \frac{d_j}{n_j(n_j-d_j)}\end{aligned}$$

dans l'équation (4.10) :

$$Var\{\log(\hat{p}_j)\} = \frac{d_j}{n_j(n_j-d_j)} \tag{4.11}$$

ainsi

$$Var\{\log(\hat{S}(t))\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)} \tag{4.12}$$

Appliquons le résultat (4.9) à (4.12) on aura :

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

et

$$Var\{\hat{S}(t)\} \approx [\hat{S}(t)]^2.Var\{\log(\hat{S}(t))\}$$

Donc :

$$Var\{\hat{S}(t)\} \approx [\hat{S}(t)]^2.\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)} \tag{4.13}$$

Enfin, l'erreur standard de l'estimateur de Kaplan-Meier définit par la racine carré de la variance estimée et donnée par :

$$Se\{\hat{S}(t)\} \approx [\hat{S}(t)].[\sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}]^{\frac{1}{2}} \quad , \quad \text{pour} \ \ t_k < t < t_{k+1} \tag{4.14}$$

Ce résultat est nommé *formule de Greenwood's*[14].

Cas sans censure

S'il n'y a pas de censure dans le temps de survie, $n_j - d_j = n_{j+1}$ et l'expression (4.11) deviendra :

$$Var\{\log(\hat{p}_j)\} = \frac{n_j - n_{j+1}}{n_j.n_{j+1}}$$

$$\sum_{j=1}^{k} \frac{n_j - n_{j+1}}{n_j.n_{j+1}} = \sum_{j=1}^{k} (\frac{1}{n_{j+1}} - \frac{1}{n_j}) = \frac{n_1 - n_{k+1}}{n_1.n_{k+1}}$$

qui est égale à

$$\frac{1-\hat{S}(t)}{n_1.\hat{S}(t)} \text{ tel que } \hat{S}(t) = \frac{n_{k+1}}{n_1}$$

dans l'absence de censure.

Donc :

$$\begin{aligned} Var\{\hat{S}(t)\} &\approx [\hat{S}(t)]^2 [\frac{1-\hat{S}(t)}{n_1.\hat{S}(t)}] \\ &\approx \frac{\hat{S}(t)[1-\hat{S}(t)]}{n_1} \end{aligned}$$

4.2.2 L'erreur standard de la méthode actuarielle et Nelson-Aalen

L'estimateur de la méthode actuarielle pour la fonction de survie est similaire à celui de Nelson-Aalen, l'erreur standard de ces deux estimateurs s'obtient de la même manière.
L'erreur standard de la méthode actuarielle est donnée par :

$$Se\{S^*(t)\} \approx S^*(t). \sum_{j=1}^{k} \{\frac{d_j}{n'_j.(n'_j - d_j)}\}^{\frac{1}{2}} \tag{4.15}$$

et l'erreur standard de Nelson-Aalen est :

$$Se\{\tilde{S}(t)\} \approx \tilde{S}(t).\{\sum_{j=1}^{k} \frac{d_j}{n_j^2}\}^{\frac{1}{2}} \tag{4.16}$$

4.3 Estimation de la fonction de risque

Soit un échantillon de survie , deux méthodes sont décrites si dessous pour estimer la fonction de risque :

4.3.1 La méthode actuarielle pour la fonction de risque

Supposons que le temps de survie est groupé en m intervalles.
L'estimation du risque moyen de mort par unité de temps (sur chaque intervalle) est le nombre

de décès observés par le temps moyen des survécus sur cet intervalle, cette dernière quantité est le nombre moyen de personnes en risque sur l'intervalle multiplier par la longueur de l'intervalle.

$$RM = \frac{\text{nombre de décès sur l'intervalle}}{\text{temps moyen des survécus sur l'intervalle}} \tag{4.17}$$

Le temps moyen de survie sur l'intervalle est :

$$\frac{n_j^{'} - d_j}{2}.\tau_j \tag{4.18}$$

tel que :
d_j : nombre de décès au $j^{\text{éme}}$ intervalle, $j = \overline{1,m}$
$n_j^{'}$: nombre moyen de personne en risque sur l'intervalle , $n_j^{'} = n_j - \frac{c_j}{2}$
τ_j : longueur de l'intervalle.

L'éstimateur de la fonction de risque est :

$$h^*(t) = \frac{d_j}{(n_j^{'} - d_j/2).\tau_j}, \;\; t_j^{'} \leq t < t_{j+1}^{'} \tag{4.19}$$

L'erreur standard de cet estimateur est donné par Gehan (1969)

$$Se\{h^*(t)\} = \frac{h^*(t).\sqrt{1 - [h^*(t)\tau_j/2]}}{\sqrt{d_j}} \tag{4.20}$$

4.3.2 L'estimateur de Kaplan-Meier

Pour estimer la fonction de risque pour des données de survies, il suffit de prendre le rapport entre le nombre de décès sur le nombre d'individus en risque à ce moment là.
Si la fonction de risque est constante entre deux temps consécutifs de décès le risque par unité de temps est trouvé en divisant sur la longueur de l'intervalle.
La fonction de risque estimé entre t_j et t_{j+1} est : [9]

$$\hat{h}(t) = \frac{d_j}{n_j.\tau_j} \tag{4.21}$$

tel que :
d_j : nombre de décès a t_j, $\;\; j = \overline{1,r}$.
n_j : nombre de personnes en risque à t_j.
$\tau_j = t_{j+1} - t_j$

Remarque L'équation (4.21) ne peut être utilisée pour estimer la fonction de risque sur l'intervalle $[t_j,+\infty[$ tel que t_j est le dernier moment de décès car cet intervalle est ouvert (τ_j est inconnu)et sa longueur est infini.

$\hat{h}(t)$ est l'estimation de la fonction de risque par unité de temps sur le $j^{\text{éme}}$ intervalle.

$$\hat{h}(t)=\frac{d_j}{n_j.\tau_j} \implies \frac{d_j}{n_j}=\hat{h}(t).\tau_j$$

$\frac{d_j}{n_j}$ est la probabilité qu'un individu meurt sur cet intervalle alors la probabilité de survie est $1-\frac{d_j}{n_j}$.

L'erreur standard de la fonction de risque $Se\{\hat{h}(t)\}=?$ tel que $\hat{h}(t)=\frac{d_j}{n_j}$ donc :

$$Var\{\hat{h}(t)\}=Var\{\frac{d_j}{n_j}\}=\frac{1}{n_j^2}.Var(d_j)$$

$d_j \rightsquigarrow B(n_j,p_j)$ alors $Var(d_j)=n_j.p_j.(1-p_j)$

$$\begin{aligned}
Var\{\hat{h}(t)\} &= \frac{1}{n_j^2}n_j\hat{p}_j(1-\hat{p}_j) \\
&= \frac{\hat{p}_j(1-\hat{p}_j)}{n_j} \\
&= \Big(\frac{d_j}{n_j}\Big)\Big(\frac{n_j-d_j}{n_j}\Big)\Big(\frac{1}{n_j}\Big) \\
&= \frac{d_j}{n_j}\Big(\frac{n_j-d_j}{n_jd_j}\Big)\frac{d_j}{n_j} \\
&= \Big(\frac{d_j}{n_j}\Big)^2.\Big(\frac{n_j-d_j}{n_jd_j}\Big)=[\hat{h}(t)]^2.\Big(\frac{n_j-d_j}{n_jd_j}\Big)
\end{aligned}$$

$$Var\{\hat{h}(t)\}=[\hat{h}(t)]^2.\Big(\frac{n_j-d_j}{n_jd_j}\Big) \tag{4.22}$$

et

$$Se\{\hat{h}(t)\}=[\hat{h}(t)]\sqrt{\Big(\frac{n_j-d_j}{n_jd_j}\Big)} \tag{4.23}$$

4.4 Comparaison de courbes de survie estimées par Kaplan-Meier

les tests mis en oeuvre en vue de comparer deux ou plusieurs sous-populations sont les tests du Log-Rank, de Breslow et de Tarone-Ware. Ces trois tests statistiques sont très voisins les uns

des autres.
L'idée générale consiste à comparer la distribution d'occurrence des événements dans chacune des sous-populations sous l'hypothèse nulle H_o que les distributions des fonctions de séjour sont semblables dans chacune des sous-populations.Chaque test statistique s'appuie sur le calcul d'une grandeur U qui correspond au temps t_i de la somme de chaque écart entre le nombre d'échéances observées (d_i) et le nombre d'échéances (décès) attendues (e_i), multiplié par une pondération w_i :

$$U = \sum_i w_i(d_i - e_i) \tag{4.24}$$

Le nombre e_i correspond au nombre d'échéances qui seraient observées au temps t_i si les deux distributions étaient strictement identiques Les trois tests diffèrent entre eux par la pondération w_i qui est prise en compte. Dans le cas du test du Log-Rank, tous les poids sont égaux à 1, ce qui revient à dire que le même poids est donné à chaque événement.
Pour le test de Breslow, le poids correspond à la population soumise au risque en t_i $(w_i = n_i)$ alors qu'il correspond à sa racine carrée $(w_i = \sqrt{n_i})$ lorsqu'il s'agit du test de Tarone Ware. Dans le cas d'une comparaison entre deux sous-populations, les trois statistiques obtenues doivent être comparées à un χ_2 à 1 degré de liberté.

Conclusion

Ce chapitre est consacré à la mise en oeuvre des méthodes non-paramétriques de l'analyse de survie qui visent à répondre à un objectif d'exploration des données. Ces explorations consistent à observer la distribution de l'événement étudié au cours du temps, ainsi que de poser ou de tester des hypothèses quant aux différences de distribution entre des sous-populations.

Chapitre 5

Modèles semi-paramétriques

5.1 Modélisation des données de survie

Les modèles semi-paramétriques sont des modèles fondés sur les hypothèses de risque proportionnel et de loi de distribution inconnu sur le temps de survie.
Le modèle de base pour ces données est *le modèle proportionnel de risque* (Proportional hazard model), ce modèle était proposé par Cox en 1972 connu également sous le non *Modèle de régression de Cox*.

5.1.1 Qu'est ce qu'un modèle de Cox ?

Le modèle de Cox est employé lorsqu'on cherche à évaluer l'effet de certains facteurs,appelés covariables sur la durée de survie.

Un modèle de Cox est une technique statistique bien reconnue pour explorer le rapport entre la survie d'un patient ou d'un contrat et plusieurs variables explicatives.
Ce modèle s'applique à une situation où l'on étudie le délai de survie d'un événement (durée de vie/temps de survie).

Pour résoudre le problème de la durée et des facteurs explicatifs, Cox propose une analyse de régression non pas sur la caractéristique actuelle (au moment de l'enquête) de l'individu, mais sur les différentes valeurs prises par cette caractéristique, durant chaque période (année,jour,mois) de son existence jusqu'au moment de l'enquête. En quelque sorte, chaque année vécue par chaque personne enquêtée constitue une tranche d'observation.

Particularités du modèle

1. Notion dichotomique (survie/décès).
2. Notion quantitative (durée d'attente avant le décès).
3. Mauvaise adaptation de la distribution gaussienne (distribution exponentielle ou de Weibull).

5.1.2 Modèle pour comparaison de deux groupes

Supposons que deux groupes de patients soumis à deux types de traitements différents,un groupe suit le traitement habituel et l'autre un nouveau traitement.
Notons donc :
– $h_S(t)$: risque de mort sous le traitement standard.
– $h_N(t)$: risque de décès sous le nouveau traitement.
Le modèle proportionnel de risque est donné par :

$$h_N(t) = \psi . h_S(t) \Longrightarrow \psi = \frac{h_N(t)}{h_S(t)},$$

tel que ψ est une constante nommé *risque relative* ou *rapport de risque* (Hazard ratio).
• Si $\psi < 1$ le risque de mort à l'instant t est plus petit chez l'individu sous le nouveau traitement. le nouveau traitement est meilleur que le standard.
• Si $\psi > 1$ le risque de décès à t sous le nouveau traitement est plus grand ; le traitement habituel est mieux.

Soit des données de survie sur n individus, on note :
$h_i(t)$: la fonction de risque du $i^{\text{éme}}$ patient ; $i = \overline{1,n}$.
$h_0(t)$ la fonction de risque pour un individu sous le traitement standard.

Alors, la fonction de risque pour un individu sous le nouveau traitement est $\psi h_0(t)$ (le risque relative ψ ne peut être négative).

On prend :

$$\psi = \exp(\beta) \quad \Longrightarrow \quad \beta = \log \psi, \ \ \beta \in [-\infty, +\infty[$$

Les valeurs positives de β sont obtenues quand le rapport de risque ψ est supérieure à 1 c'est à dire le nouveau traitement est moins efficace au standard.
Soit X une variable indicatrice tel que :

$$X = \begin{cases} 0 & \text{si patient est sous traitement standard} \\ 1 & \text{sinon} \end{cases}$$

si on pose x_i est la $i^{\text{éme}}$ valeur de X pour $i = \overline{1,n}$ alors :

La fonction de risque de cet individu sera donnée par :

$$h_i(t) = e^{\beta x_i} h_0(t), \tag{5.1}$$

où

$x_i = 1$ *si* le patient est sous le nouveau traitement et
$x_i = 0$ sinon.

5.2 Modèle généralisé du risque proportionnel

Le cadre est le suivant :
Soient $2n$ variables $t_1, t_2, ..., t_n$ et $c_1, c_2, ..., c_n$ qui sont respectivement les durées de survie (variable d'intérêt) et les durées de censure des n individus supposés indépendants .
Notons que la censure peut être aléatoire, il est souhaitable que c et t soient indépendantes ; sur chacun des individus une variable $x_i = (x_{i_1}, ..., x_{i_p})$,cette variable x est généralement appelée «*covariable*» ou aussi «*variable exogène*».

La fonction de risque du $i^{\text{éme}}$ individu est :

$$h_i(t) = \psi(x_i) h_0(t) \tag{5.2}$$

$h_0(t)$ la fonction de risque de base.
$\psi(x_i) = \exp(\eta_i)$ est la fonction des valeurs des variables explicatives et η_i est une combinaison linéaire (une matrice)des p composantes du vecteur x_i ;

$$\eta_i = \beta_1 x_{1i} + ... + \beta_p x_{pi} = \sum_{j=1}^{p} \beta_j x_{ji}.$$

Le modèle de "risque proportionnel" ou modèle de Cox suppose que :

$$h_i(t) = h_0(t) \exp(\beta' x). \tag{5.3}$$

$$h_i(t) = h_0(t) \exp(\beta_1 x_{1i} + ... + \beta_p x_{pi}) \tag{5.4}$$

$\beta = (\beta_1, \beta_2, ..., \beta_p)$: le vecteur des coefficients de la régression, il s'agit d'estimer ces coefficients pour évaluer l'impact de chacun des facteurs sur la durée étudiées.
x : le vecteur des variables exogènes.
$h_0(t)$: la fonction de hasard de base ou le risque instantané de base, qu'il faut estimer elle aussi.

Remarque 2. *Le rapport des risques instantanés de deux sujets i et j avec les covariables fixes x_i et x_j vaut :*

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\beta' x_i)}{h_0(t)\exp(\beta' x_j)} \tag{5.5}$$

$$= \frac{\exp(\beta' x_i)}{\exp(\beta' x_j)} \tag{5.6}$$

$$= \exp(\beta' x_i - \beta' x_j) \tag{5.7}$$

C'est à dire que le rapport des fonctions de hasard pour deux individus de caractéristiques x_i et x_j ne dépend pas de t,mais seulement de x_i et x_j.
*Ce rapport est dit «**hazard ratio**».*

Remarque 3. *On appelle aussi ce modèle modèle à hasards proportionnels (PH).Cependant, le modèle (PH) est en fait plus général que le modèle de COX car le facteur multiplicatifs n'est pas nécessairement une exponentielle d'une fonction linéaire des covariables*

$$h(t/X = x, \beta) = h_0(t)g(x, \beta) \tag{5.8}$$

Dans ce modèle, g est une fonction spécifiée de la covariable x et du paramètre β.

5.3 Composante linéaire du modèle de risque proportionnel

La fonction de risque dépend de deux types de variables, des variables aléatoires et des facteurs.

La variable aléatoire est une variable numérique souvent continue comme l'age,la tension...., d'autre part, un facteur est une variable qui prend un certain nombre de valeurs par exemple : le sex(homme,femme).

Alors les variables et les facteurs sont incorporées dans la composante linéaire du modèle de risque proportionnel.

5.3.1 Les variables aléatoires

Pour illustrer l'idée, prenons la situation oú la fonction de risque dépend de deux variables aléatoires X_1 et X_2, la valeur de ces deux variables aléatoires pour le $i^{\text{éme}}$ individu sont x_{1i}, x_{2i} respectivement.

$$h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i})h_0(t) \tag{5.9}$$

est le modèle de risque proportionnel pour le $i^{\text{éme}}$ individu pour $i = \overline{1,n}$.

5.3.2 Les facteurs

Supposons que la fonction de risque dépend d'un seul facteur noté A tel que A est composé de a valeurs c'est à dire $A = (1, 2...., a)$.
La fonction de risque d'un individu est :

$$h_j(t) = \exp(\alpha_j).h_0(t) \tag{5.10}$$

α_j : la résiliation du facteur A au $j^{éme}$ niveau.
$(\alpha_1, \alpha_2, ..., \alpha_j, ..., \alpha_a)$: l'effet principal (the main effects) de A.
$h_0(t)$:fonction de risque de base.

Remarque 4. *Si $\alpha_1 = 0$ les α_j sont définies sur $(a-1)$ variables explicatives, $X_2, X_3, ..., X_a$ avec les coefficients $\alpha_2, \alpha_3, ..., \alpha_a$. On d'autre terme α_j est remplacé par $\alpha_2 x_2 + \alpha_3 x_3 + ... + \alpha_a x_a$*

$$h(t) = \exp(\alpha_2 x_2 + \alpha_3 x_3 + ... + \alpha_a x_a).h_0(t) \tag{5.11}$$

tel que x_j est la j^{eme} valeur de X_j d'un individu de A au j^{eme} niveau comme si dessous :

Level of A	X_2	X_3	. . .	X_a
1	0	0	. . .	0
2	1	0	. . .	0
3	0	1	. . .	0
. . .	. . .	. . .	. . .	. . .
a	0	0	. . .	1

5.4 Ajustement du modèle de risque proportionnel

La fonction de survie est donnée par :

$$S(t,x) = 1 - F(t,x) = [S_0(x)]^{e^{(\beta' x)}} \tag{5.12}$$

ou F est la fonction de distribution cumulative,
S_0 fonction de survie de base.

Le taux de risque en t est :

$$h(t,x) = -\frac{d}{dt}\ln(S(t,x)) = h_0(t)\exp(\beta' x) \tag{5.13}$$

avec : h_0 est la fonction de hasard de base.

Le modèle statistique est nommé : **modèle de régression à risque proportionnel** "*proportional hazard regression model*" ou **modèle de régression de Cox** "*Cox regression model*" car pour n'importe deux valeurs x_1 et x_2 le hasard ratio est :

$$\frac{h_1(t)}{h_2(t)} = \exp(\beta' x_1 - \beta' x_2) \tag{5.14}$$

indépendant du temps t.

Notons par t_{x_1} et t_{x_2} deux temps de vie indépendants avec x_1 et x_2 paramètres de régression on a alors :

$$Pr(t_{x_1} < t_{x_2}) = \frac{\exp(\beta x_1)}{\exp(\beta x_1) + \exp(\beta x_2)} \tag{5.15}$$

car

$$\begin{aligned}
Pr(t_{x_1} < t_{x_2}) &= \int_0^\infty Pr(t < t_{x_2}).f(t,x)dt \\
&= \int_0^\infty S(t,x_2).h(t,x)S(t,x)dt \\
&= \int_0^\infty [S_0(t)]^{\exp(\beta x_1)}.h_0(t)e^{(\beta \tilde{x})}.[S_0(t)]^{\exp(\beta x_2)}dt \\
&= \exp(\beta \tilde{x}).\int_0^\infty h_0(t).[S_0(t)]^{\exp(\beta x_1)+\exp(\beta x_2)}dt
\end{aligned}$$

et pour tout réel ψ et une fonction de hasard $h(t) = \psi h_0(t)$ avec la fonction survie $S(t) = [s_0(t)]^\psi$

$$\begin{aligned}
\int_0^\infty h_0(t)[s_0(t)]^\psi dt &= \frac{1}{\psi}\int_0^\infty \psi h_0(t)[s_0(t)]^\psi dt \\
&= \frac{1}{\psi}\int_0^\infty h(t)S(t)dt \\
&= \frac{1}{\psi}\int_0^\infty f(t)dt = \frac{1}{\psi}.
\end{aligned}$$

et sous certaines conditions sur les paramètres de regression on peut construire *la fonction de vraisemblance partielle*

$$\prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\Sigma_{j\in R_i}\exp(\beta' X_j)}\right)^{1-c_i} = \prod_{\{i;c_i=0\}} \frac{\exp(\beta' x_i)}{\Sigma_{j\in R_i}\exp(\beta' x_j)} \tag{5.16}$$

Donc à partir d'un échantillon ordonné $(y_{(1)},....,y_{(n)})$, la fonction de vraisemblance partielle de Cox (s'il n'y a pas de données censurées) :

$$L(y_{(1)},....,y_{(n)};\beta)=\prod_{\{i;c_i=0\}}\frac{\exp(\beta' x_i)}{\Sigma_{j\in R_{y_i}}\exp(\beta' x_j)} \tag{5.17}$$

Tel que :

R_i **:** l'ensemble de risque pour toute les observations y_i non censuré ;

x **:** le vecteur des variables exogènes (covariables)de dimension p, de l'individu qui a eu un événement.

Et on reconnaît chacun des termes du produit qui forme la vraisemblance partielle de Cox.
Cox propose de traiter cette vraisemblance partielle comme une vraisemblance exacte, c'est-à-dire que l'estimateur de Cox s'obtient en maximisant la fonction de vraisemblance partielle $L(y_{(1)},....,y_{(n)};\beta)$ alors :

$$\log L(y_{(1)},....,y_{(n)};\beta)=\sum_{i=1}^{n}[\beta' x_i-\log(\sum_{j\in R_{y_i}}e^{\beta' x_j})] \tag{5.18}$$

5.4.1 Pourquoi la fonction de vraisemblance ?

Soit :

$$P(\text{individu de la variable } x_j \text{ meurt à } t_j \text{ /un décès à } t_j) \tag{5.19}$$

sachant que : $P(A/B)=P(A\cap B).P(B)$ donc l'équation (5.19)sera :

$$\begin{aligned}
\frac{P(\text{individu de la variable } x_j \text{ meurt à } t_j)}{P(\text{un mort à } t_j)} &= \frac{P(\text{individu de la variable } x_j \text{ meurt à } t_j)}{\sum_{l\in R(t_j)}P(\text{individu } l \text{ meurt à } t_j)}\\
&= \frac{P(\text{individu } x_j \text{ meurt entre } [t_j,t_j+\delta t])/\delta t}{\sum_{l\in R(t_j)}P(\text{individu } l \text{ meurt entre } [t_j,t_j+\delta t])/\delta t}\\
\text{quand } \delta t\longrightarrow 0 \text{ alors :}&\\
&= \frac{P(\text{individu de la variable } x_j \text{ meurt à } t_j)}{P(\text{un mort à } t_j)}\\
&= \frac{(\text{risque de mourir en } t_j \text{ pour l'individu de } x_j)}{\sum_{l\in R(t_j)}(\text{risque de mourir en} t_j \text{pour individu } l)}\\
&= \frac{h_i(t_j)}{\sum_{l\in R(t_j)}h_l(t_j)}=\frac{\exp(\beta' x_j)}{\sum_{l\in R(t_j)}\exp(\beta' x_l)}
\end{aligned}$$

C'est la fonction de vraisemblance partielle car il n'y a pas de données censurées.
L'estimateur du maximum de vraisemblance de β dans le modèle de risque proportionnel est trouvé en maximisant le log de la fonction de vraisemblance.Cette maximisation se fait par la *procédure de Newton-Raphson*.

5.4.2 La procédure de Newton-Raphson[6]

Le modèle de données de survie censurés sont habituellement adaptés à la procédure de Newton-Raphson pour maximiser la fonction de vraisemblance.

Soit
$u(\beta)$: un vecteur de dimension $(p \times 1)$ qui contient les valeurs de la première dérivé de log vraisemblance (The vector of efficient scores).
$I(\beta)$: matrice des valeurs de la $2^{\text{éme}}$ dérivé de log vraisemblance et de dimension$(p \times p)$ nommé matrice d'information observé (The observed information matrix).

Le (j,k) élément de la matrice $I(\beta)$ est :

$$\frac{-\partial^2 \log L(\beta)}{\partial\beta_j.\partial\beta_k}$$

La procédure de Newton-Raphson pour estimer le vecteur du paramètre β en $s+1$ termes est :

$$\hat{\beta}_{s+1} = \hat{\beta}_s + I^{-1}(\hat{\beta}_s).u(\hat{\beta}_s), \text{ pour } s = 0,1,.. \tag{5.20}$$

$I^{-1}(\hat{\beta}_s)$: inverse de la matrice d'information.

La procédure commence en $\hat{\beta}_0$, et on s'arrête dés que le log de vraisemblance est petit.Quand le procédé iteratif converge la matrice de variance covariance de paramètre estime est approximativement proche de l'inverse de variance covariance ie $I^{-1}(\hat{\beta}_s)$, la racine carré des éléments diagonals de cette matrice sont l'erreur standard des paramètres $\beta_1, \beta_2, ..., \beta_p$ estimés.

5.5 Estimation de la fonction de risque et de survie

Supposons que les composantes linéaire du modèle de risque proportionnel contient p variables explicatives $(X_1, X_2, ..., X_p)$ et les coefficients estimé $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$.
La fonction risque estimé par le $i^{\text{éme}}$ individu est :

$$\hat{h}_i(t) = \exp(\hat{\beta}' x_i).\hat{h}_0(t) \tag{5.21}$$

x_i : vecteur du $i^{\text{éme}}$ individu qui contient les valeurs des variables explicatives,
$\hat{\beta}$: vecteur des paramètres estimer,
$\hat{h}_0(t)$: fonction de risque de base estimer.

La fonction de risque pour un individu peut être estimer une fois que $h_0(t)$ est calculé.

Kalbfleisch et Prentice(1973) ont utilisé le maximum de vraisemblance pour estimer la fonction de risque de base $h_0(t)$.
Supposons avoir r temps de mort tel que : $t_1 < t_2 < ... < t_r$ et d_j décès, n_j individus en risque en t_j.Alors la fonction de risque de base estimée à t_j est donnée par

$$\hat{h}_0(t_j) = 1 - \hat{\xi}_j \tag{5.22}$$

ou : $\hat{\xi}_j$ est la solution de l'équation

$$\sum_{l \in D(t_j)} \frac{\exp(\hat{\beta}' x_l)}{1 - \hat{\xi}_j^{\exp(\hat{\beta}' x_l)}} = \sum_{l \in R(t_j)} \exp(\hat{\beta}' x_l) \;\text{ pour }\; j = \overline{1, r} \tag{5.23}$$

$D(t_j)$: l'ensemble de tout les individus décédés en t_j.
$R(t_j)$:l'ensemble de tout les individus en risque en t_j.

5.5.1 Pour un seul décès

L'estimation de $h_0(t)$ par cette façon est très complexe, sauf dans un cas particulier ou il n'y a pas de mort ie $d_j = 1$ pour $j = \overline{1, r}$, alors le coté gauche de l'équation (5.23) deviendra :

$$\frac{\exp(\hat{\beta}' x_l)}{1 - \hat{\xi}_j^{\exp(\hat{\beta}' x_l)}} = \sum_{l \in R(t_j)} \exp(\hat{\beta}' x_l) \tag{5.24}$$

de (5.24)

$$\begin{aligned} 1 - \hat{\xi}_j^{\exp(\hat{\beta}' x_l)} &= \frac{\exp(\hat{\beta}' x_l)}{\sum_{l \in R(t_j)} \exp(\hat{\beta}' x_l)} \\ \hat{\xi}_j^{\exp(\hat{\beta}' x_l)} &= 1 - \frac{\exp(\hat{\beta}' x_l)}{\sum_{l \in R(t_j)} \exp(\hat{\beta}' x_l)} \end{aligned}$$

Alors :

$$\hat{\xi}_j = \left[1 - \frac{\exp(\hat{\beta}' x_l)}{\sum_{l \in R(t_j)} \exp(\hat{\beta}' x_l)} \right]^{\exp(-\hat{\beta}' x_l)} \tag{5.25}$$

5.5.2 Pour plusieurs morts

Quand d_j est différent de 1 c'est a dire on a plusieurs morts l'équation (5.23) ne peut être résolu explicitement et un arrangement est exigé.

Nous supposons que le risque de mort est constant entre deux temps adjacents de décès, alors la fonction de risque sur cet intervalle de temps est obtenu en divisant $h_0(t)$ de l'équation (5.22)par la longueur de l'intervalle.

$$\hat{h}_0(t) = \frac{1-\hat{\xi}_j}{t_{j+1}-t_j} \text{ pour } t_j \leq t < t_{j+1} \;\; / \;\; j = \overline{1, r-1} \tag{5.26}$$

avec : $\hat{h}_0(t) = 0$ pour $t < t_1$.
La quantité $\hat{\xi}_j$ peut être considéré comme une estimation de probabilité qu'un individu survie sur l'intervalle t_j et t_{j+1}, et l'estimation de la fonction de survie initiale est :

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j \text{ pour } t_k \leq t < t_{k+1} \text{ et } k = \overline{1, r-1} \tag{5.27}$$

$$\hat{S}_0(t) = \begin{cases} 1 & \text{si } t < t_1 \\ 0 & \text{pour } t \geq t_r \end{cases}$$

La fonction initiale de risque cumulative est donnée par $H_0(t) = -\log S_0(t)$ et son estimation

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = -\sum_{j=1}^{k} \log \hat{\xi}_j \tag{5.28}$$

pour $t_k \leq t < t_{k+1}$ et $k = \overline{1, r-1}$.

avec $\hat{H}_0(t) = 0$ pour $t < t_1$

En particulier, la fonction de risque est estimée par :

$$\hat{h}_i(t) = \exp(\hat{\beta}' x_i).\hat{h}_0(t) \tag{5.29}$$

on intègre (5.29), on obtient

$$\begin{aligned} \int_0^t \hat{h}_i(u)du &= \exp(\hat{\beta}' x_i).\int_0^t \hat{h}_0(u)du \\ \hat{H}_i(t) &= \exp(\hat{\beta}' x_i)\hat{H}_0(t) \end{aligned}$$

Sachant que : $\hat{S}_i(t) = \exp\Big(-\hat{H}_i(t)\Big)$

$$\begin{aligned} \hat{S}_i(t) &= \exp\Big[-\exp(\hat{\beta}' x_i).\hat{H}_0(t)\Big] \\ &= \Big[\hat{S}_0(t)\Big]^{\exp(\hat{\beta}' x_i)} \end{aligned}$$

$$\hat{S}_i(t) = \Big[\hat{S}_0(t)\Big]^{\exp(\hat{\beta}' x_i)} \text{pour } t_k \leq t < t_{k+1} \tag{5.30}$$

par conséquent la fonction de risque cumulative est

$$\hat{H}_i(t) = -\log\left[\hat{S}_i(t)\right]$$

$$\hat{H}_i(t) = -\log\left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}' x_i)} \tag{5.31}$$

5.5.3 Pas de variables exogènes

S'il n'y a pas de variables exogènes, l'équation (5.23) est égale à :

$$\frac{d_j}{1-\hat{\xi}_i} = n_j \text{ tel que } \hat{\xi}_i = \frac{n_j - d_j}{n_j}$$

l'estimation de la fonction de survie est

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\xi}_j = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right), \tag{5.32}$$

c'est l'estimateur de Kaplan-Meier.

La fonction de risque :

$$\begin{aligned} \hat{h}_0(t) &= \frac{1-\hat{\xi}_i}{t_{j+1}-t_j} \\ &= \frac{\left(\frac{d_j}{n_j}\right)}{t_{j+1}-t_j} \end{aligned}$$

donc

$$\hat{h}_0(t) = \frac{d_j}{n_j(t_{j+1}-t_j)} \tag{5.33}$$

5.6 Quelque approximation pour estimation des fonctions de base

Supposons que les temps de survies ne sont pas indépendant ,la résolution de l'équation (5.23) se fera par la méthode itérative.Cette dernière peut être éviter en utilisant une approximation.

Le terme

$$\begin{aligned}\hat{\xi}_j^{\exp(\hat{\beta}' x_l)} &= \log\left(\hat{\xi}_j^{\exp(\hat{\beta}' x_l)}\right) = \exp(\hat{\beta}' x_l).\log(\hat{\xi}_j)\\ &= \exp\left(\hat{\beta}' x_l.\log(\hat{\xi}_j)\right)\\ &\approx 1+\exp(\hat{\beta}' x_l).\log\hat{\xi}_j\end{aligned}$$

donc

$$1-\hat{\xi}_j^{\exp(\hat{\beta}' x_l)} = -\exp(\hat{\beta}' x_l).\log(\hat{\xi}_j) \tag{5.34}$$

On note $1-\tilde{\xi}_j$ l'estimateur risque de base obtenus par cette approximation.
Par conséquent l'équation (5.23) sera :

$$\begin{aligned}-\sum_{l\in D(t_j)}\frac{1}{\log\tilde{\xi}_j} &= \sum_{l\in R(t_j)}\exp(\hat{\beta}' x_l)\\ -\frac{d_j}{\log\tilde{\xi}_j} &= \sum_{l\in R(t_j)}\exp(\hat{\beta}' x_l)\end{aligned}$$

$$\tilde{\xi}_j = \exp\left(\frac{-d_j}{\sum_{l\in R(t_j)}\exp(\hat{\beta}' x_l)}\right) \tag{5.35}$$

La fonction de survie basé sur l'estimateur$\tilde{\xi}_j$ est :

$$\tilde{S}_0(t) = \prod_{j=1}^{k}\exp\left(\frac{-d_j}{\sum_{l\in R(t_j)}\exp(\hat{\beta}' x_l)}\right) \tag{5.36}$$

$$\text{pour } \; t_k \leq t < t_{k+1}, k=\overline{1,r-1}$$

La fonction cumulative de risque de base estimée est :

$$\tilde{H}_0(t) = -\log\tilde{S}_0(t) = \sum_{j=1}^{k}\left(\frac{d_j}{\sum_{l\in R(t_j)}\exp(\hat{\beta}' x_l)}\right) \tag{5.37}$$

S'il n'y a pas de variables exogènes alors : $\sum_{l\in R(t_j)}\exp(\hat{\beta}' x_l) = n_j$ la fonction de survie et de risque de base sont respectivement :

$$\tilde{S}_0(t) = \prod_{j=1}^{k}\exp\left(\frac{-d_j}{n_j}\right) \tag{5.38}$$

$$\tilde{H}_0(t) = \sum_{j=1}^{k}\frac{-d_j}{n_j} \tag{5.39}$$

Une autre approximation pour l'equation (5.35) , noté ξ^* :

$$\xi_j^* = 1 - \frac{d_j}{\sum_{l\in R(t_j)} \exp(\hat{\beta}' x_l)} \tag{5.40}$$

tel que $\exp(x) = 1 + x + o(x)$
La fonction de risque de base estimée sur t_j et t_{j+1} est :

$$h_0^*(t) = \frac{d_j}{(t_{j+1} - t_j)\sum_{l\in R(t_j)} \exp(\hat{\beta}' x_l)} \tag{5.41}$$

la fonction de base de survie correspondante

$$S_0^*(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{\sum_{l\in R(t_j)} \exp(\hat{\beta}' x_l)}\right) \tag{5.42}$$

et la fonction de base de risque cumulative est

$$H_0^*(t) = \sum_{j=1}^{k} (t_{j+1} - t_j).h_0^*(t) \tag{5.43}$$

On remplaçant $H_0^*(t)$ et $S_0^*(t)$ dans les équations

$$H_i^*(t) = \exp(\hat{\beta}' x_i).H_0^*(t)$$

et

$$S_i^*(t) = \left(S_0^*(t)\right)^{\exp(\hat{\beta}' x_i)}$$

on aura estimer la fonction risque cumulative et la fonction de survie pour un individu du vecteur x_i.

5.7 Modèles stratifiés

Modèle de Cox stratifié est un modèle ou le risque initial $h_0(t)$ (Baseline Hazard)est propre à chaque strate. La stratification se fait selon la variable que l'on spécifie initialement, lorsqu'on précise une variable de stratification, le modèle de Cox est donc estimé en admettant un risque de référence $h_{0,s}(t)$ différent pour chaque strate s.

Conclusion

Les modèles semi-paramétriques permettent d'obtenir un compromis entre un modèle paramétrique, généralement trop restrictif et un modèle non-paramétrique où l'erreur d'estimation

devient vite trop grande en présence de variables explicatives de grande dimension.
Ils trouvent leur intérêt notamment dans les modèles de durée où l'on cherche à prendre en compte l'information apportée par des variables explicatives. Une spécificité de ces modèles tient à la présence fréquente de phénomènes de censures dans les observations.
Dans notre mémoire,on s'intéresse à la durée des contrats tout on censurons par rapport au coût d'un sinistre en faisant alors appel à des techniques adaptées à ce type de contexte.

Chapitre 6

Application

6.1 Présentation des données

On s'appuyant sur l'article (J.M.Marion et K.Boukhetala, A.Oulidi),nous nous intéressons à l'estimation des durées de vie moyennes de contrats d'assurance automobile et de determiner les facteurs qui induisent les accidents routiers.
Nos données sont issues d'un portefeuille prévenant d'une compagnie d'assurance de taille significative sur le marché Algérien d'assurance non vie.
C'est un fichier de sinistralités matérielles (le coût d'un accident), dont il existe plusieurs classes, chacune des classes est caractérisée par des caractéristiques telles : La vitesse (nombre de chevaux), la zone de circulation, l'ancienneté de la voiture. . .etc.

Pour notre étude on a choisi une classe de puissance 7 chevaux, une classe qui contient 5890 contrats. Mais après avoir éliminé quelque valeur aberrante le fichier final que nous étudions comporte 3569 données.
Ces contrats ont été crées entre 1990 et 2006, la variable d'intérêt est la durée de vie des contrats et nous allons censurer à 30% du coûts d'accident.

6.2 Statistique exploratoire

Pour notre étude statistique nous avons choisis quatre variables exogènes qui montrent les répartitions suivantes et il s'agit de :

1. L'âge en mise en circulation du véhicule (notée AMC), cette variable a été regroupée en quatre classes :

- $AMC1$ correspond à un âge de circulation inférieur ou égale à 6 ans ;
- $AMC2$ correspond à un âge de circulation compris entre 6 et 7ans (inclus) ;
- $AMC3$ une durée de circulation comprise entre 7 et 8 ans ;
- $AMC4$ correspond à une durée strictement supérieure à 8ans.

2. Le permis de conduite (notée Permis), cette variable est codée sous :
   - $Permis1$ correspond aux personnes ayant des anciens permis ;
   - $Permis2$ correspond aux personnes ayant de nouveau permis.
3. Age du conducteur, cette variable est regroupée en deux groupes et notée (Age C)
   - $AgeC1$ pour un conducteur âgé ;
   - $AgeC2$ pour un jeune conducteur .
4. Le Bonus-malus (noté BM), cette variable a été regroupé en trois classes :
   - $BM1$: un code Bonus-malus inférieur ou égale à 30%
   - $BM2$: un code Bonus-malus compris entre 30% et 50% de bonus ;
   - $BM3$:un code Bonus-malus supérieur à 50%.

Les tableaux si dessous représentent les répartitions des variables exogènes :

Contrats	Résiliés	Censurés	Total
AMC1	134	347	481
AMC2	360	748	1108
AMC3	560	1341	1901
AMC4	17	61	78
	1071	2497	3568

Contrats	Résiliés	Censurés	Total
Permis1	1059	2468	3527
Permis2	12	29	41
	1071	2497	3568

Contrats	Résiliés	Censurés	Total
AgeC1	1050	2400	3450
AgeC2	21	97	118
	1071	2497	3568

Contrats	Résiliés	Censurés	Total
BM1	1055	2463	3518
BM2	8	25	33
BM3	8	9	17
	1071	2497	3568

FIG. 6.1 – Répartitions des statistiques

A partir de ces statistiques, on remarque bien que le portefeuille étudié est constitué de contrats de 30% de bonus et de véhicules âgés entre 7 et 8 .Les tableaux qui suivent donne quelques statistiques concernant la variable d'intérêt durée vie dont la durée de vie moyenne , la borne inférieur et supérieur

	Moyenne			Borne inferieure			Borne supérieure		
	Tout	Résiliés	Censurés	Tout	Résiliés	Censurés	Tout	Résiliés	Censurés
Global	8.12	8.00	8.17	8.08	7.94	8.13	8.16	8.07	8.21
AMC1	8	7.995	7.99	7.992	7.98	7.97	8	8.005	8.004
AMC2	7.11	7.67	7.55	7.10	7.64	7.52	7.12	7.69	7.58
AMC3	7.5	7.52	7.45	7.47	7.36	7.31	7.50	7.67	7.59
AMC4	8.14	8.13	8.142	8.134	8.09	8.11	8.16	8.16	8.17
Permis1	8.13	8.01	8.18	8.09	7.94	8.13	8.16	8.07	8.22
Permis2	7.80	7.85	7.57	7.34	6.67	6.74	8.26	9.03	8.40
AgeC1	8.009	8.01	8.09	7.94	7.93	8.02	8.07	8.06	8.16
AgeC2	8.008	8	7.59	7.37	7.35	6.97	8.64	8.641	8.22
BM1	8.03	8.04	8.07	7.93	7.92	8.00	8.07	8.07	8.14
BM2	8.90	8.89	8.65	8.34	8.34	7.94	9.45	9.45	9.35
BM3	7.74	7.74	7.89	6.64	6.62	7.17	8.84	8.63	8.62

FIG. 6.2 – Tableau des statistiques (1)

ainsi la minimale et maximale pour chacune des variables exogènes.

	Min			Max		
	Tout	Résiliés	Censurés	Tout	Résiliés	Censurés
Global	1.86	3.98	1.86	10.62	10.34	10.62
AMC1	4.91	4.98	4.97	5	5	5
AMC2	6.08	6.08	6.08	7	7	7
AMC3	7.08	7.36	7.26	8	7.67	7.59
AMC4	8.08	8.08	8.08	8.25	8.25	8.25
Permis1	1.86	3.98	1.86	10.62	10.35	10.62
Permis2	4.03	4.03	5.85	10.01	9.97	9.90
AgeC1	3.98	3.98	3.99	10.34	10.34	10.62
AgeC2	5.13	5.13	5.85	9.97	9.97	9.69
BM1	3.98	3.98	3.99	10.34	10.34	10.62
BM2	7.94	7.72	7.18	9.63	9.35	9.87
BM3	6.34	6.34	7.24	9.38	9.38	8.87

FIG. 6.3 – Tableau des statistiques (2)

Ces statistiques nous indiquent une durée de vie moyenne des contrats de 8 ans au global sur ce portefeuille, et des différences plus au moins importantes entre les segments étudiés de 7.1 à 8.9 ans.

Ces statistiques nous donnent peu d'information sur les distributions des durées de vie ou sur la fonction de survie, pour cela nous proposons les méthodes suivantes d'estimation des durées de vie.

6.3 Estimation de Kaplan-Meier

En premier lieu est édité une table. Dans cette table, chaque ligne correspond à un individu et trié par ordre croissant de durée, d'autre part les colonnes de la table sont :

- *Time* : l'intervalle de temps considéré, dans notre cas le temps est en jours ;
- *Status* : l'indice de censure c_i ;
- *Cumulative Survival* : la fonction de survie $S(t)$;
- *Standard Error* : Erreur standard de la probabilité de survie pour le calcul de l'intervalle de confiance ;
- *Cumulative Events* : le cumule des événements ;
- *Number Remaining* : l'effectif N_i des individus soumis au risque.

Étant donné la taille de la population que nous avions au départ ($N_0 = 3568$ individus), les tables de survie des chacune des variables reportent uniquement le début de cette table (Retourner à l'annexe).

6.3.1 Variable Age C

Ce premier tableau est suivi par un deuxième récapitulative du nombre d'individus présent au début de la période d'observation (Number of Cases), du nombre d'individus censurés (Censored) ainsi que du nombre d'individus ayant connu l'événement (Events) durant toute la période d'observation.

Age C	Total N	N of Events	Censored	
			N	Percent
1	3452	2400	1052	30,5%
2	116	97	19	16,4%
Overall	3568	2497	1071	30,0%

FIG. 6.4 – Tableau Age C

Ce deuxième tableau est suivi par un troisième dans lequel est reporté la durée moyenne du contrat ainsi que sa durée médiane,voir figure (6.5). Cette moyenne et cette médiane sont accompagnées de leur écart-type et de leur intervalle de confiance à 95%.

La moyenne est calculée sur l'ensemble de la période comprise entre l'instant initial t_0 et l'instant maximum au cours lequel est observée une échéance.

L'estimation de cette moyenne ne tient donc pas compte des sorties d'observation qui auront lieu au-delà de cet instant maximum.
Ainsi, la durée médiane d'un contrat pour les conducteurs avec ancienneté est estimée par 3145 jours et qui vaut 8 ans et demi, d'autres parts les conducteurs de type 2 la duré médiane est de 8 ans.

Plus que le conducteur est âgé plus la durée du contrat est longue.

Age C	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
1	3100,787	7,495	3086,098	3115,477	3145,000	11,208	3123,033	3166,967
2	2965,167	44,042	2878,844	3051,489	2916,000	38,533	2840,476	2991,524
Overall	3096,275	7,400	3081,772	3110,779	3140,000	10,927	3118,583	3161,417

FIG. 6.5 – Durée moyenne et médiane du temps de survie

Les graphiques de survie et de risque cumulé $S(t)$ et $H(t)$ sont édités à la suite de ces deux tables et constituent une aide précieuse pour l'interprétation.
Les croix sur ces graphiques représentent les moments auxquels des individus sortent d'observation (censures).

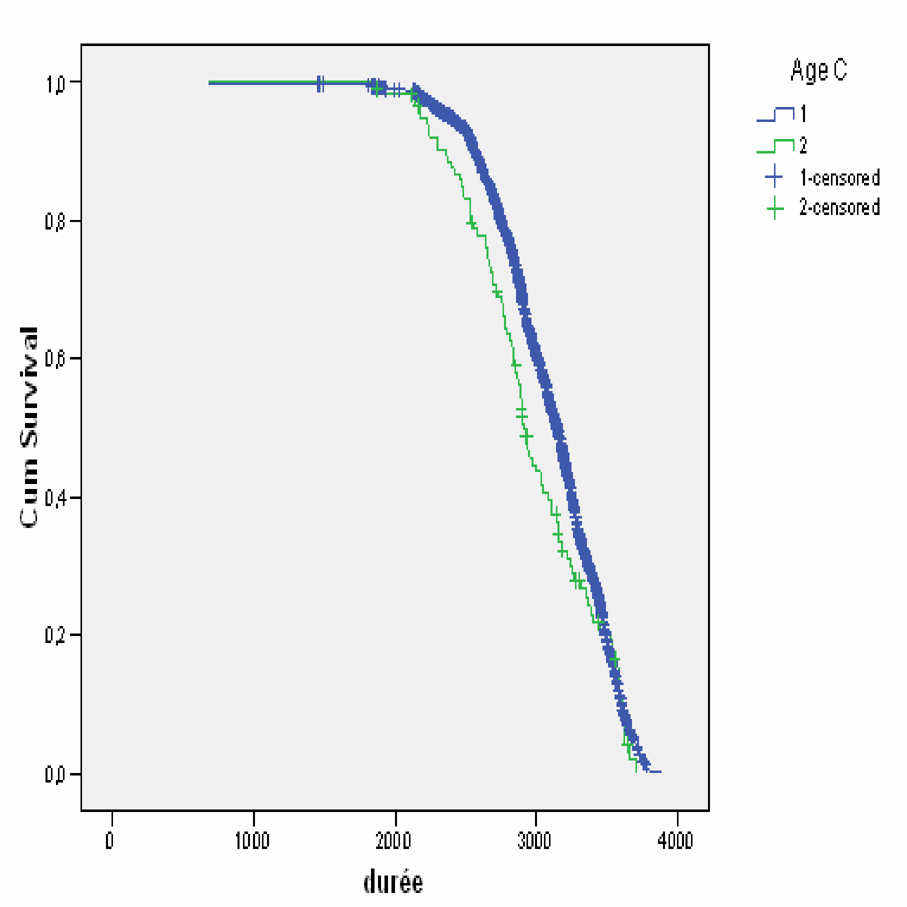


FIG. 6.6 – Fonction de survie des contrats

La plupart des sorties d'observations interviennent après le 2000eme jour, soit environ 5 ans et demi après l'entrée dans la compagnie d'assurance.

On conclusion, la durée de vie de contrat chez les conducteurs âgés est plus longue que les conducteurs jeunes.

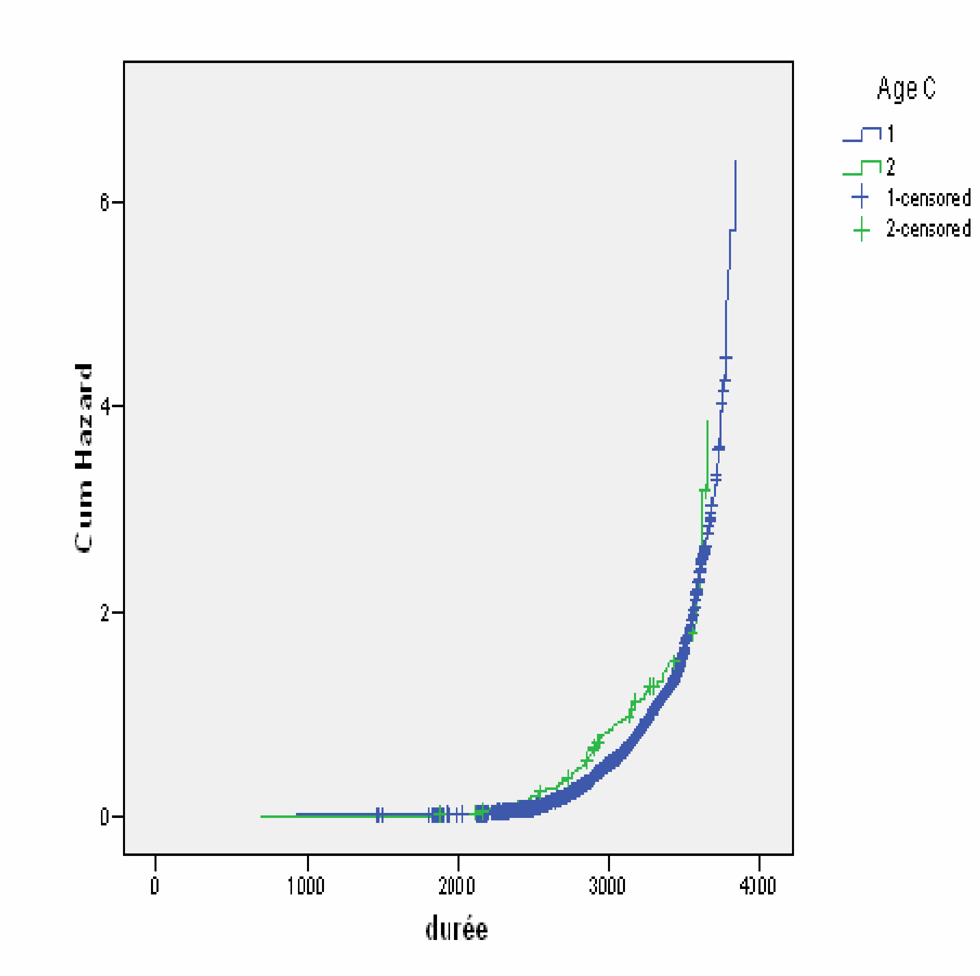


FIG. 6.7 – Fonction de risque cumulé des contrats

Un graphique $H(t)$ présente l'intérêt de donner une indication de l'évolution du risque et de connaître l'événement au cours du temps. Ainsi, une courbe d'allure convexe indique que le risque diminue au cours du temps. A l'opposé, une pente de forme concave indique une augmentation du risque au cours du temps, alors qu'une droite signifie un risque constant.

L'examen de ce graphique indique que plus la durée n'augmente et plus la pente de croissance diminue, en ce qui concerne l'âge du conducteur on remarque que plus l'âge augmente plus sa courbe se situe à un niveau plus bas.
Ce qui nous permet de mentionner que les chances de rester assurée à la compagnie pour un conducteur jeune diminuent avec le temps après une période.

Le test statistique de Log Rank est édité sur la table si dessous qui représente les résultats du test, le degré de liberté (ddl) et la significative du test.

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6,130	1	,013

FIG. 6.8 – Résultats du test de comparaison

Le χ_2 calculé pour le test du Log-Rank est égal à 6.130 et il est à comparer au χ_2 théorique qui vaut $3,84$ pour un risque $\alpha = 5\%$.
Le χ_2 observé est supérieur au χ_2 théorique, ce qui conduit à rejeter l'hypothèse statistique d'homogénéité des distributions de durée.

En d'autres termes, la distribution de la durée des contrats diffèrent entre les deux âges.

Le χ_2 est accompagné d'une *p*-value inférieure à $\alpha = 5\%$ indique que l'hypothèse nulle doit être rejetée.

Notre hypothèse selon laquelle la durée du contrat est plus longue chez les personnes âgés que chez ceux des jeunes apparaît ainsi se vérifier.

6.3.2 Variable AMC

La variable Age de Mise en Circulation a été divisée en 4 groupes :

- $AMC1$: pour les voitures de durée de circulation inférieure ou égale à 6 ans ;
- $AMC2$: durée de circulation comprise entre 6 et 7 ans ;
- $AMC3$: entre 7 et 8 ans ;
- $AMC4$: strictement supérieur à 8 ans.

Le tableau si dessous est un tableau récapitulatif qui regroupe toute l'information initiale de notre variable étudié,la deuxième colonne contient le nombre d'individu présent au début de la période d'observation ,la troisième *number of events* représente les individus ayant connu l'événement durant toute la période d'étude et la dernière colonne représente le nombre et le pourcentage des voitures censurés.

AMC*	Total N	N of Events	Censored N	Censored Percent
1	481	309	172	35,8%
2	1108	764	344	31,0%
3	1901	1362	539	28,4%
4	78	62	16	20,5%
Overall	3568	2497	1071	30,0%

FIG. 6.9 – Tableau AMC

Puisque les données sur la survie de notre étude sont censurées, alors le temps de survie médian est une meilleure mesure de tendance centrale que le temps moyen de survie.
D'après l'estimateur de Kaplan-Meier, le temps médian de survie des contrats est de :

- 6 ans et demi pour le premier groupe,
- 7 ans et demi pour le deuxième groupe,
- 9 ans pour le troisième groupe ;
- 10 ans pour le dernier.

La durée médiane est estimée par 3140 jours (8 ans et demi) et un intervalle de confiance à 95% de borne inférieur et supérieur respectivement comme suit $3118,583$ et $3118,583$.

La moyenne de survie est $3096,275$ et $IC_{\alpha}(95\%) = [3081,772 \quad - \quad 3110,779]$

AMC*	Mean[a] Estimate	Mean[a] Std. Error	Mean[a] 95% Confidence Interval Lower Bound	Mean[a] 95% Confidence Interval Upper Bound	Median Estimate	Median Std. Error	Median 95% Confidence Interval Lower Bound	Median 95% Confidence Interval Upper Bound
1	2347,430	10,960	2325,948	2368,911	2424,000	13,737	2397,076	2450,924
2	2778,522	3,523	2771,617	2785,426	2810,000	6,529	2797,202	2822,798
3	3302,150	5,344	3291,676	3312,623	3293,000	8,935	3275,487	3310,513
4	3724,931	6,259	3712,663	3737,198	3719,000	8,361	3702,612	3735,388
Overall	3096,275	7,400	3081,772	3110,779	3140,000	10,927	3118,583	3161,417

FIG. 6.10 – Durée moyenne et médiane du temps de survie

Les résultats du test statistique dans le cas de la comparaison de l'ensemble des sous populations sont présentées dans le tableau figure (6.11) .

Le nombre de degré de liberté (df) correspond au nombre de sous-populations moins une, Ici quatre sous-populations sont prises en compte, le nombre de degrés de liberté est donc de 3. Le χ_2 calculé pour le test du Log-Rank est égal à $6390,759$ et par conséquent plus grand que le χ_2 théorique de $7,81$ pour un risque $\alpha = 5\%$.

On conclusion la différence entre les 4 groupes de l'AMC est statistiquement significative au risque $\alpha = 5\%$.

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6390,759	3	,000

FIG. 6.11 – Résultats du test de Log-Rank

Voici la représentation des courbes de survie pour chaque âge de voiture :

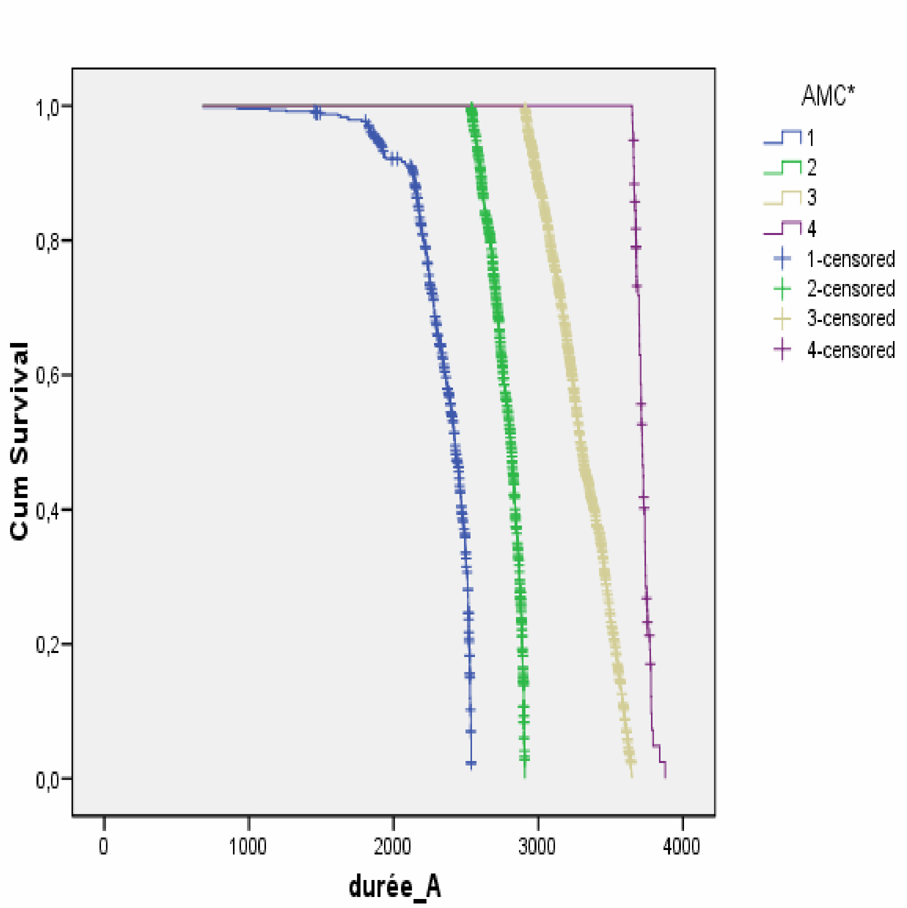


FIG. 6.12 – Fonction de survie des contrats

Les croix sur les courbes représentent les censures

Les contrats des voitures de type 4 survivent plus longtemps que celles de type 3 puis celles de type d'*AMC*2 et enfin les voitures d'*AMC*1 survivent le moins longtemps.

On pourra interpréter ce résultat comme suit :
les conducteurs ayant de nouvelles voitures sont exposés à un risque élevé cela est due à l'excès de vitesse et par suite des accidents de coût élevé qui obligera l'assureur a résilié le contrat.

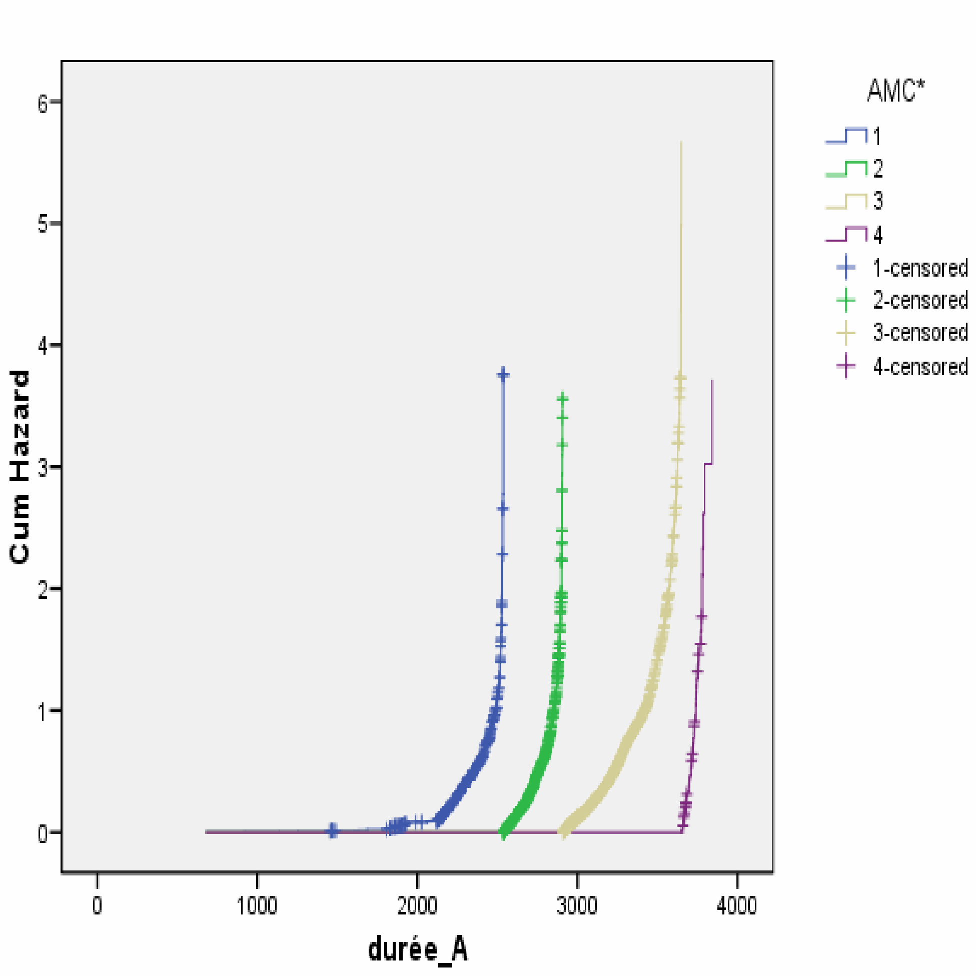


FIG. 6.13 – Fonction de risque cumulé des contrats

6.3.3 Variable Permis de conduire

Le tableau de la figure (6.14) est le tableau descriptif de la statistique de notre variable Permis de conduire.

Permis	Total N	N of Events	Censored	
			N	Percent
1	3527	2456	1071	30,4%
2	41	41	0	,0%
Overall	3568	2497	1071	30,0%

FIG. 6.14 – Tableau Permis

Sur le graphique si dessous, nous constatons une différence entre les deux catégories de permis.

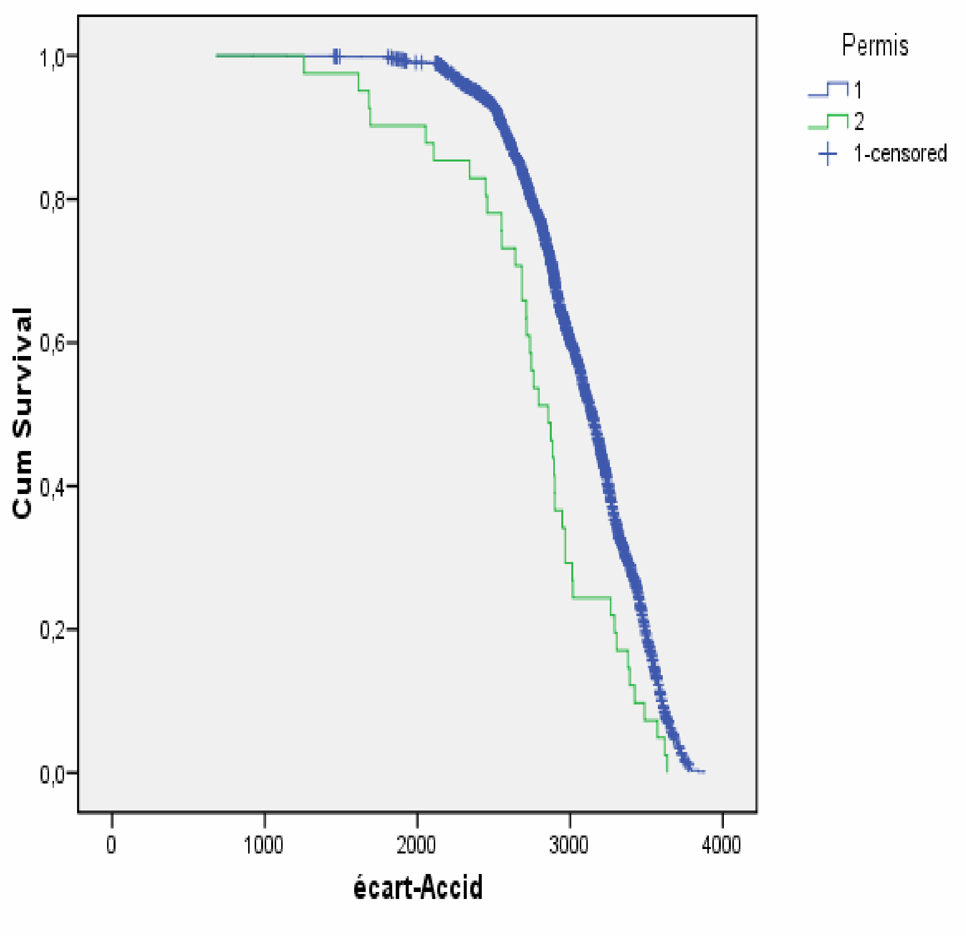


FIG. 6.15 – Fonction de survie des contrats

La courbe correspondante au permis 2 c'est à dire : personnes ayant un nouveau permis laisse à penser que ce groupe a plus de chance que sa durée de vie à la compagnie soit plus faible que pour l'autre groupe.

Là encore, la statistique du Log-Rank prenant la valeur $15,395$ confirmant une différence significative entre les deux groupes avec une p-value de 0.0.

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	15,395	1	,000

FIG. 6.16 – Résultats du test de Log-Rank

Puisque les courbes sont significativement différentes, cela signifie que les variables exogènes jouent un rôle dans la durée de vie des contrats.

6.3.4 Variable Bonus-Malus

Le Bonus-malus (noté BM), cette variable aléatoire a été regroupé en trois classes :
- $BM1$ correspond à moins de 30% de bonus ;
- $BM2$ correspond à un bonus compris entre 30% et 50%;
- $BM3$ correspond à un code bonus strictement supérieur à 50%..

Et voici si dessous le tableau descriptif de la statistique de notre variable bonus-malus :

			Censored	
BM	Total N	N of Events	N	Percent
1	3522	2467	1055	30,0%
2	16	13	3	18,8%
3	30	17	13	43,3%
Overall	3568	2497	1071	30,0%

FIG. 6.17 – Tableau Bonus-Malus

Le graphique si dessous présente une estimation de la fonction de survie correspondant aux trois codes Bonus-Malus toujours par la méthode de Kaplan-Meier, mais ici en tenant compte du découpage de la variable BM (bonus-malus) en 3 classes comme indiqué au par avant.

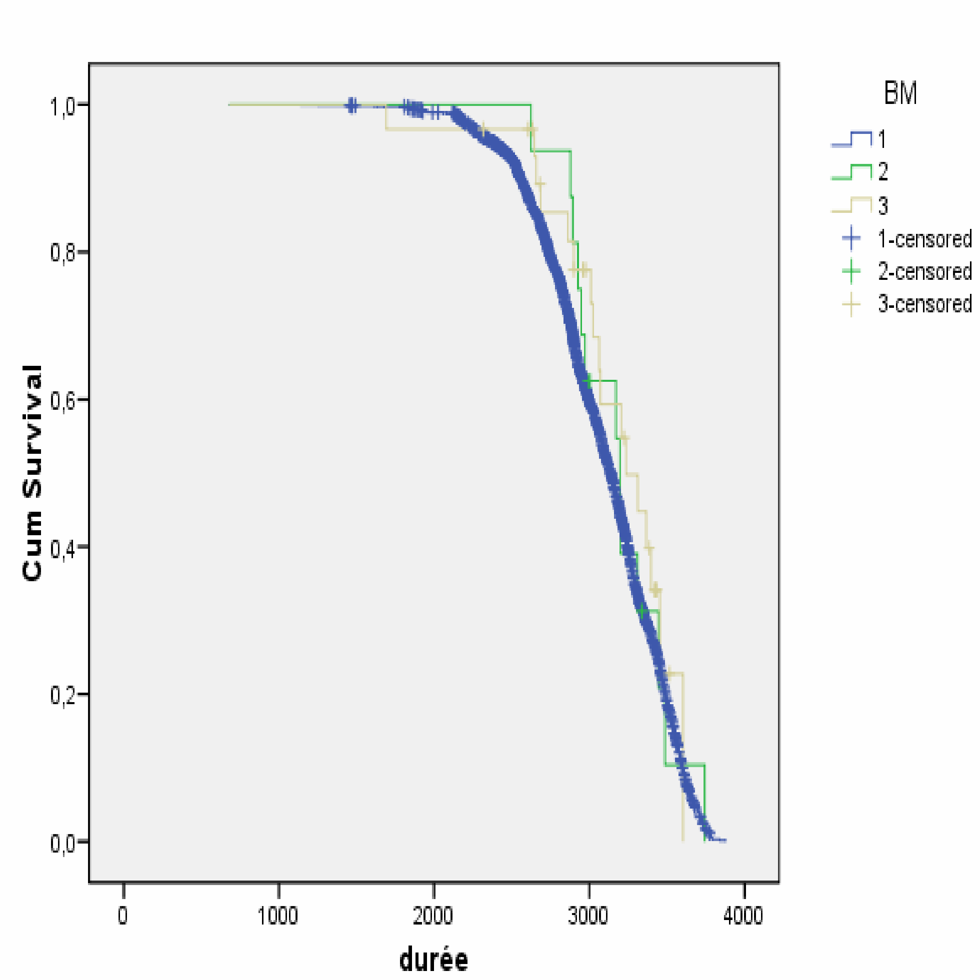


FIG. 6.18 – Fonction de survie des contrats

Remarquons que la statistique du Log-Rank prenant la valeur 1.452 confirme l'inexistence de la différence significative entre les trois courbe avec une p-value de 0.484 environ.

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1,452	2	,484
Breslow (Generalized Wilcoxon)	2,742	2	,254
Tarone-Ware	2,252	2	,324

FIG. 6.19 – Résultats de test de Log-Rank

6.4 Estimation par le modèle de COX

Le modèle de COX (1972) en temps continu, connu aussi sous le nom de *modèle semi-paramétrique* à risques proportionnels est un modèle d'analyse de survie du type régression.

Il exprime le risque instantané *(the hazard rate)*, il permet de quantifier et de tester les effets propres de caractéristiques individuelles telles que : *Age C, BM, Permis* sur le risque de transition.

Dans cette partie on va essayer d'expliquer étape par étape l'ajustement du modèle semi-paramétrique à risques proportionnels avec la procédure de régression de COX.

6.4.1 Estimation d'un premier modèle simple

Nous commençons par illustrer l'estimation d'un premier modèle dans lequel nous cherchons à expliquer le risque de terminer un contrat en fonction de l'âge du conducteur (Age C), bonus-malus (BM) et enfin en fonction du Permis.

Interprétation des coefficients estimés

L'estimation proprement dite du modèle est donnée dans le tableau "Variables in the Equation"

	B	SE	Wald	df	Sig.	Exp(B)
Permis	-5,815	1,008	33,291	1	,000	,003
AgeC	,018	,104	,030	1	,863	1,018
BM	-,103	,110	,875	1	,349	,902

FIG. 6.20 – Résultat des variables

Sous *B* on y lit le coefficient estimé de chaque facteur explicatif, celui-ci mesure l'effet du facteur sur le logarithme du risque.

Par exemple, le coefficient Permis est égale à -5.815 c'est à dire qu'à chaque nouvelle catégorie de permis le logarithme du risque diminue de -5.815,mais il est souvent plus aisé d'interpréter l'exponentiel du coefficient donné sous *Exp(B)* qui correspond à hazard ratio.

Pour le permis, l'exponentiel du coefficient est environ 0.003,ce qui signifie que pour un même BM et pour la même catégorie d'âge le risque de voir un nouveau permis terminé son contrat est 0.003 fois celui d'un ancien permis.

La colonne *"SE"* (Standard Error) donne l'erreur standard du coefficient qui mesure la variabilité de l'estimateur utilisé.
Les trois colonnes *"Wald", "df"* ou *ddl* et *"Sig"* concernent la statistique de Wald utilisée pour tester la significative individuelle de chaque coefficient, elles donnent respectivement la valeur de la statistique, ses degrés de liberté et son degré de signification.

On remarque ici que, si l'effet permis est très significatif au seuil $\alpha = 5\%$, l' âge du conducteur (Age C) et Bonus-Malus (BM) n'affectent pas significativement sur le risque de terminer un contrat ou la durée d'assurance.

Autres résultats

Le tableau *"Case Processing Summary"* nous indique que nos données sont constituées de 3568 épisodes dont 1071 (30%) sont censurés.

		N	Percent
Cases available in analysis	Event[a]	2497	70,0%
	Censored	0	,0%
	Total	2497	70,0%
Cases dropped	Cases with missing values	0	,0%
	Cases with negative time	0	,0%
	Censored cases before the earliest event in a stratum	1071	30,0%
	Total	1071	30,0%
Total		3568	100,0%

FIG. 6.21 – Tableau descriptif des données

Les informations sous "Block 0 : Beginning Block" de la figure (6.22) concernent le modèle sans les facteurs explicatifs : permis, BM, Age C.

En fait, on a ici simplement la valeur du χ_2 log vraisemblance que l'on peut interpréter comme la distance entre les prédictions fournit par ce modèle et les observations. Cette statistique est utile en tant que valeur de référence pour juger les facteurs explicatifs introduits ultérieurement.

-2 Log Likelihood
34093,086

FIG. 6.22 – Block 0

Les résultats globaux du modèle spécifié figurent sous “Block 1” de la figure (6.23) dans le tableau intitulé “Omnibus Tests of Model Coefficients” qui contient des indicateurs globaux d’ajustement soit :

1. Le $-2\log$ vraisemblance mesure la qualité de reproduction des données ;

2. La statistique du khi-deux du Score Test ou “ Overall Score ”,

3. Statistique du khi-deux du rapport de vraisemblance donnée sous “ Change From Previous Step ” et “Change From Previous Block ”.

Il s’agit ici des statistiques du khi-deux qui mesurent l’amélioration par rapport au modèle de référence

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
33690,382	219,416	3	,000	402,705	3	,000	402,705	3	,000

FIG. 6.23 – Block 1

En règle générale, on considère le gain comme significatif lorsque le degré de signification est inférieur à 5%, ce qui est clairement le cas dans notre étude.

Nos trois facteurs permettent d’améliorer significativement l’ajustement par rapport au modèle sans facteurs explicatifs car $-2\log(m_1) = 33690.382$ est inférieur à

$-2log(m_0) = 34093.086.$

Enfin, l'output si dessous donne la valeur moyenne des variables explicatives.

	Mean
Permis	1,016
AgeC	1,039
BM	1,019

FIG. 6.24 – Moyenne des variables

6.4.2 Le traitements des attributs catégorielles

Les variables explicatives catégorielles sont admises mais doivent être recodés sous forme d'un ensemble de variables auxiliaires. L'avantage de traiter une variable comme catégorielle est de capter des non linéarités. C'est par exemple le cas de la variable âge en mise en circulation (AMC) construite pour notre en jeux de donnée.

De façon générale, une variable avec p catégories doit être recodée avec $p-1$ variables auxiliaires,le codage le plus fréquent est sous forme de variables indicatrices prenant les valeurs 0 ou 1.

La variable AMC compte quatre catégories, ces quatre variables sont linéairement dépendantes : il suffit de connaître la valeur de deux d'entre elles pour en déduire la valeur de la troisième.

L'avantage du recodage est d'obtenir ainsi une statistique de Wald pour la variable catégorielle en plus celles des variables indicatrices.
Sous l'hypothèse d'effet nul, la statistique de Wald est distribuée comme une khi deux pour un degré de liberté (le nombre de variables servant au codage).

		Frequency	(1)	(2)	(3)
AMC_A[a]	1	481	1	0	0
	2	1108	0	1	0
	3	1901	0	0	1
	4	78	0	0	0

FIG. 6.25 – Codage des variables catégorielles

A' partir du tableau donné si dessus la première variable auxiliaire $AMC(1)$ est associé à la première catégorie de voiture ayant circulé une durée inférieur ou égale à 6 ans ; d'autre part la deuxième variable auxiliaire $AMC(2)$ est associé au véhicule d'une durée de circulation inférieur ou égale à 7 ans et la dernière associé aux voitures de 8 ans.

Les effets mesurés sont les écarts par rapport à $AMC(4)$.

Du tableau de la figure (6.26), le premier coefficient mesure l'écart entre $AMC(2)$ par rapport à $AMC(4)$, le deuxième coefficient l'écart entre $AMC(3)$ et $AMC(4)$. On remarque que le premier coefficient est non significatif mais par contre le coefficient de l'$AMC(3)$ est très significatif.

On dira alors que le risque de terminer un contrat est sur une voiture ayant circulé au moins 8 ans.
On note d'autre part que la statistique de Wald pour l'effet *AMC* reste très significatif, cette significativité est due à la forte différence entre le 3^{eme} AMC et les autres ages(le nombre grand).

	B	SE	Wald	df	Sig.	Exp(B)
BM	-,024	,111	,047	1	,828	,976
AgeC	,103	,104	,986	1	,321	1,109
Permis	-4,882	,946	26,613	1	,000	,008
AMC_A			41,092	2[a]	,000	
AMC_A(2)	19,572	11,602	2,846	1	,092	3E+008
AMC_A(3)	5,435	,864	39,599	1	,000	229,392

FIG. 6.26 – Résultats avec les variables indicatrices

EVALUATION GLOBALE

Les indications sur l'ajustement global du modèle sont données si dessous, on y trouve :

1. $-2\log$ de vraisemblance ;

2. Statistique du khi 2 du score test ;

3. Statistique du khi 2 de la vraisemblance.

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
30842,604	4552,269	5	,000	3250,482	5	,000	3250,482	5	,000

FIG. 6.27 – Résultats des tests statistiques

Le moins log de vraisemblance

L'ajustement d'un modèle est meilleur lorsque son $-2\log Lik$ est petit.
On observe ici pour notre exemple que l'ajustement du modèle $-2\log(m_2) = 30842.604$ meilleur que celle du modèle sans facteurs explicatifs $-2Log(m_0) = 34093.086$

Les statistiques du khi-deux

Ces statistiques permettent d'évaluer si globalement l'ensemble des facteurs explicatifs considérés améliore significativement l'ajustement du modèle initiale qui ne tient compte d'aucun facteur.

Pour un modèle avec p coefficients, on teste l'hypothèse

$$H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_p = 0$$

contre l'hypothèse

$$H_1 : \beta_1 \neq 0, \beta_2 \neq 0, ..., \beta_p \neq 0$$

Sous l'hypothèse H_0, les trois statistiques sont distribuées asymptotiquement selon une loi de khi-deux à 5 degrés de liberté. On remarque une amélioration très significative par rapport au modèle initiale car le degré de signification est égal à 0.

Conclusion

Le modèle initiale doit être rejeté au profit du modèle ajusté ; cela ne signifie pas que ce dernier est satisfaisant mais simplement qu'il est meilleur.

6.4.3 Estimation de la fonction de survie

Pour un modèle à risque proportionnel, la fonction de survie $S(t,x)$ d'un individu avec profil x se déduit de la fonction de survie de $S_0(t)$, correspondant au profil virtuel de référence $x = 0$, par la relation

$$S(t,x) = S_0(t)^{exp(x^{'}\beta)}$$

Le modèle de Cox donne l'estimation $\beta^{'}$ des coefficients β. Pour obtenir une estimation de la probabilité de survie $S(t,x)$, il nous faut encore une estimation de la fonction $S_0(t)$ de référence.

Sous la procédure regression de cox du logiciel SPSS, on obtient par défaut la courbe de la survie pour un profit moyen $\bar{x}$ et pour l'obtenir pour un autre profit il faut préciser et fixer les valeurs des covariables pour obtenir la fonction de survie de référence $S_0(t)$. On note que pour la variable catégorielle, nous avons donné la valeur utilisée comme catégorie de référence donc la variable $AMC = 4$, $BM = 0$, $AgeC = 0$, $Permis = 0$ et voici en dessous la courbe représentative

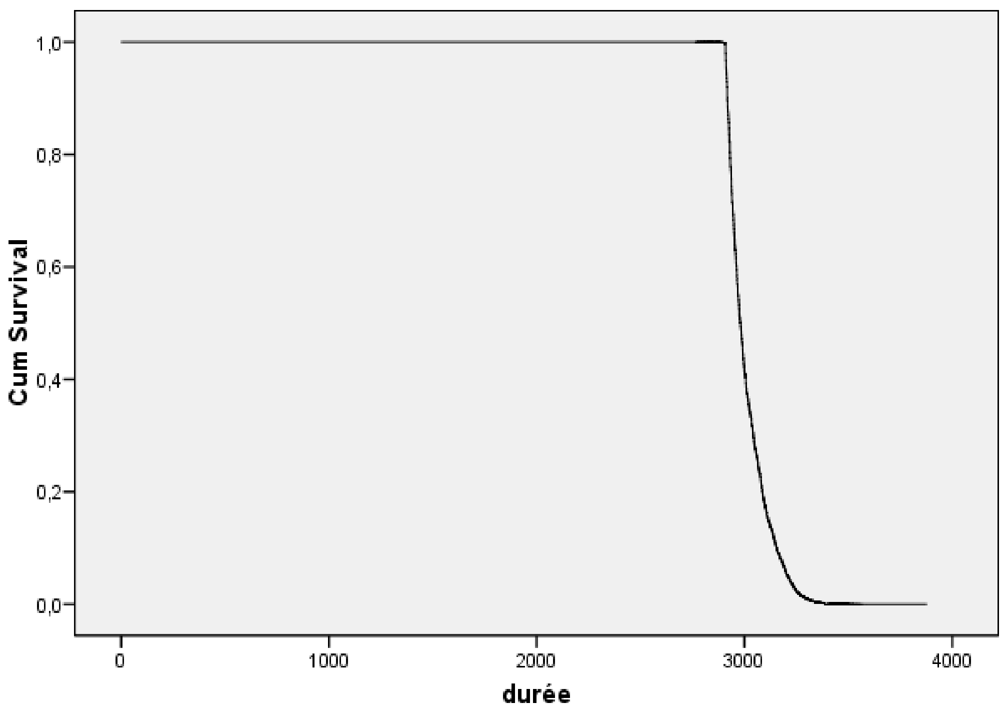


FIG. 6.28 – Fonction de survie éstimé

6.5 Modèle de COX stratifié

Un modèle stratifié est un modèle où le risque de référence est propre à chaque strate. La stratification se fait selon une variable que l'on spécifie.

6.5.1 Stratification selon le type du permis

A un premier titre d'illustration, considérons l'estimation de notre modèle avec une stratification selon le Permis, on doit évidemment enlever la variable Permis de la liste des covariables.

La figure si dessous montre les résultats on stratifiant selon le permis :

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
30843,209	2188,325	3	,000	2840,537	3	,000	2840,537	3	,000

FIG. 6.29 – Estimation du modèle avec stratification selon le Permis

Remarque la statistique du "Score Test" ("Overall Score") et celle du rapport de vraisemblance donn´ee sous 'Change From Previous Step' et 'Change From Previous Block'.

On remarque que par rapport au modèle sans stratification (Figure 6.30)on gagne un degré de liberté puisque l'on estime un paramètre de moins celui de Permis.
Il est naturel alors que les statistiques du khi-deux soient plus faibles et elles restent toujours aussi significatives.

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
30849,367	2350,624	4	,000	3243,719	4	,000	3243,719	4	,000

FIG. 6.30 – Estimation du modèle sans stratification

Notons que la statistique du modèle initial stratifié est de $m_s = 33683.746$ contre $m_0 = 34093.086$ sans stratification.

-2 Log Likelihood
33683,746

FIG. 6.31 – Block 0

Les coefficients des covariables autres que Permis restent très proches de ceux du modèle sans stratification la où le Permis est traité comme une covariable.

	B	SE	Wald	df	Sig.	Exp(B)
BM	-,024	,111	,048	1	,827	,976
AgeC	,103	,104	,979	1	,322	1,108
AMC_A	-9,257	1,006	84,662	1	,000	,000

FIG. 6.32 – Résultat des variables avec stratification selon le Permis

Le coefficient $\beta_1$associé à la variable BM est estimé par $\hat{\beta}_1 = -0.024$ avec une p-value associée à 0.827, Le coefficient β_2 associé à la variable Age C est estimé par $\hat{\beta}_2 = 0.103$ à une p-value de 0.322 et β_3 le coefficient associé à la variable AMC donnée par $\hat{\beta}_3 = -9.257$ à une P-value de 0.00.

Des deux courbes de la figure si dessous , on peut remarquer que les personnes d'un ancien permis ont tendance à rester plus longtemps dans une même compagnie que ceux d'un nouveau permis.

Cette conclusion va dans le même sens que l'interprétation que l'on peut tirer de la valeur de 0.003 de l'exponentiel du coefficient Permis lorsqu'on l'introduit comme une covariable.

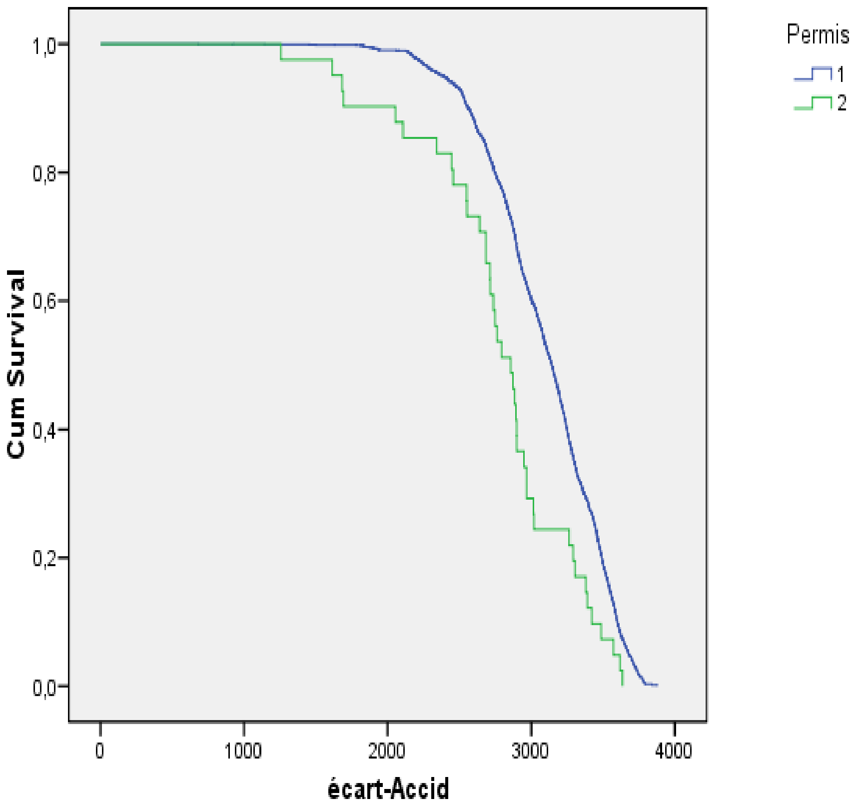


FIG. 6.33 – Courbe de survie stratifié selon le Permis

La figure si contre correspond à l'estimation de la fonction de survie pour BM= 3 ; Age C= 2 ; AMC= 2

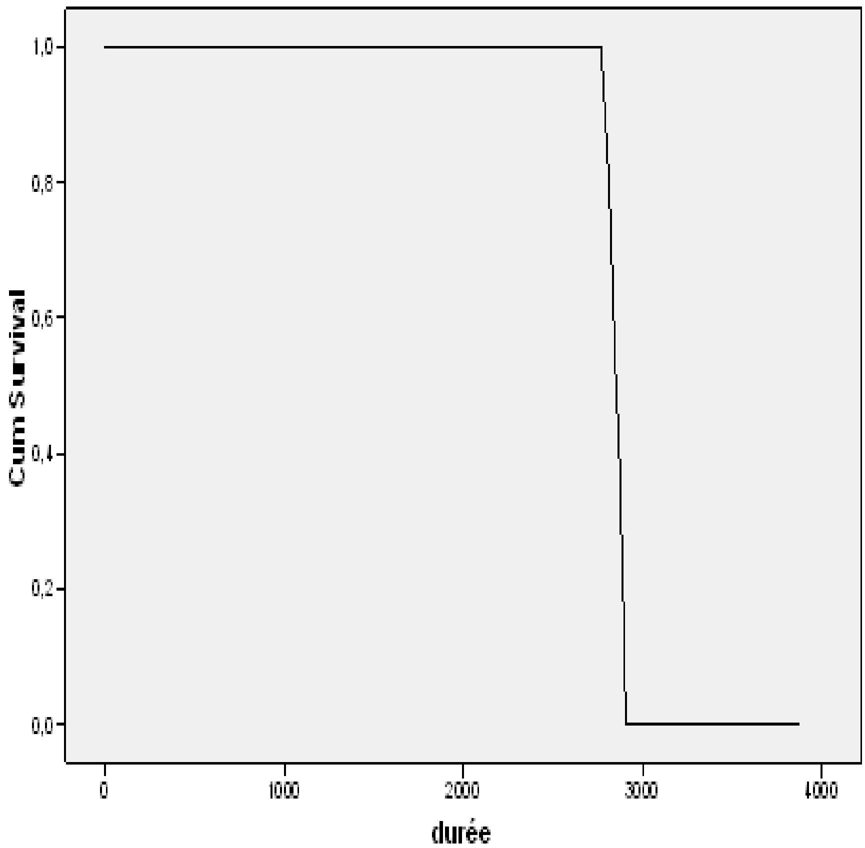


FIG. 6.34 – Fonction de survie estimé

6.5.2 Stratification selon l'AMC

Nous avons utilisé le modèle de COX en stratifiant selon la variable exogène AMC.

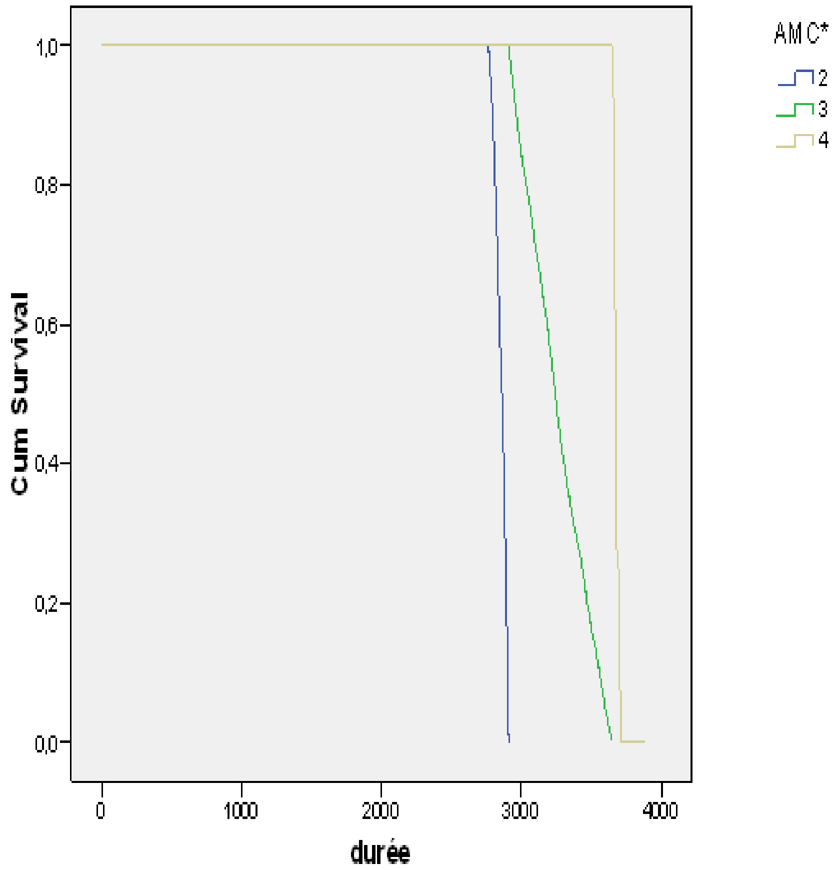


FIG. 6.35 – Fonction de survie stratifié selon l'AMC

Le coefficient β_1 associé à la variable BM est estimé par $\hat{\beta}_1 = -0.024$ (la p-value associée est de 0.827), Le coefficient β_2 associé à la variable Age C est estimé par $\hat{\beta}_2 = 0.103$ à une p-value de 0.322 et $\beta_3$le coefficient associé à la variable Permis donnée par $\hat{\beta}_3 = -5.076$ à une p-value de 0.00.

	B	SE	Wald	df	Sig.	Exp(B)
BM	-,024	,111	,048	1	,827	,976
AgeC	,103	,104	,980	1	,322	1,108
Permis	-5,076	1,031	24,213	1	,000	,006

FIG. 6.36 – Résultat des variables avec stratification selon le AMC

Le graphique suivant contre correspond à l'estimation de la fonction de survie pour

BM= 3 ; Age C= 2 ; Permis= 2

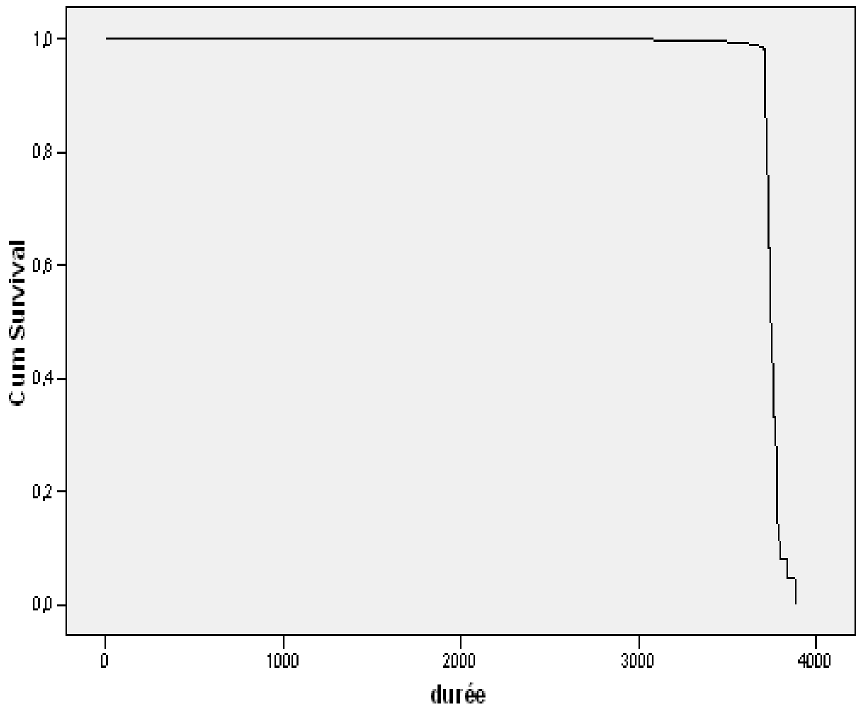


FIG. 6.37 – Fonction de survie estimé

6.5.3 Stratification selon l'AgeC

On obtient le graphique si dessous représentant la courbe de survie stratifié selon l'age du conducteur

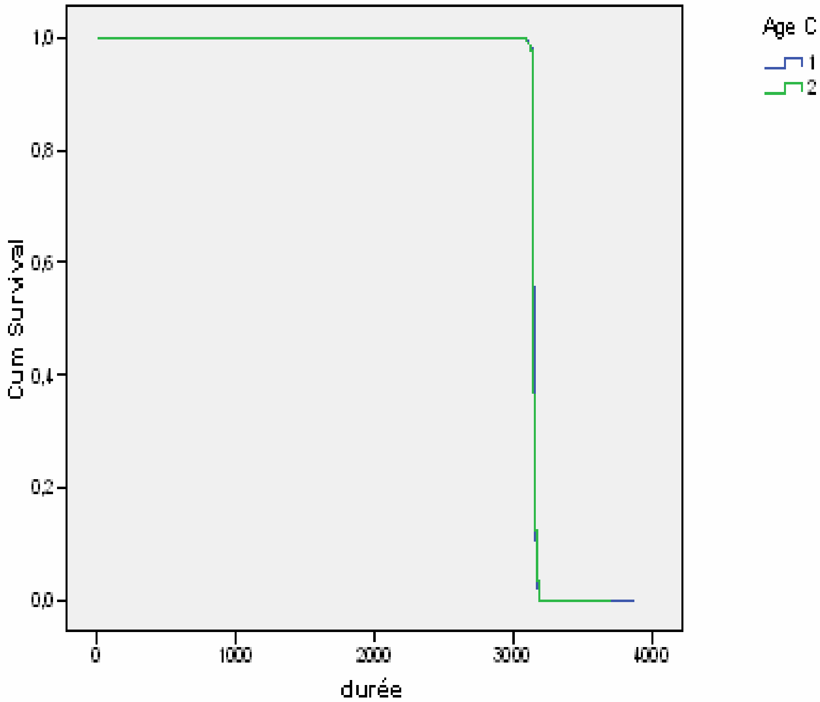


FIG. 6.38 – Fonction de survie stratifié selon Age C

par suite voici le tableau pour lesquels les variables sont estimées

	B	SE	Wald	df	Sig.	Exp(B)
BM	,041	,111	,134	1	,715	1,041
Permis	-3,891	1,104	12,429	1	,000	,020
AMC	-6,392	,229	782,491	1	,000	,002

FIG. 6.39 – Résultat des variables avec stratification selon l'AgeC

Le coefficient β_1 associé à la variable BM est estimé par $\hat{\beta}_1 = 0.041$ (la p-value associée est de 0.715), Le coefficient β_2 associé à la variable Permis est estimé par $\hat{\beta}_2 = -3.891$ à une p-value de 0.00 et β_3 le coefficient associé à la variable AMC donnée par $\hat{\beta}_3 = -6.392$ à une P-value de 0.00.

Le graphique de la figure (6.3) correspond à l'estimation de la fonction de survie pour BM= 3 ; AMC= 2 ; Permis= 2

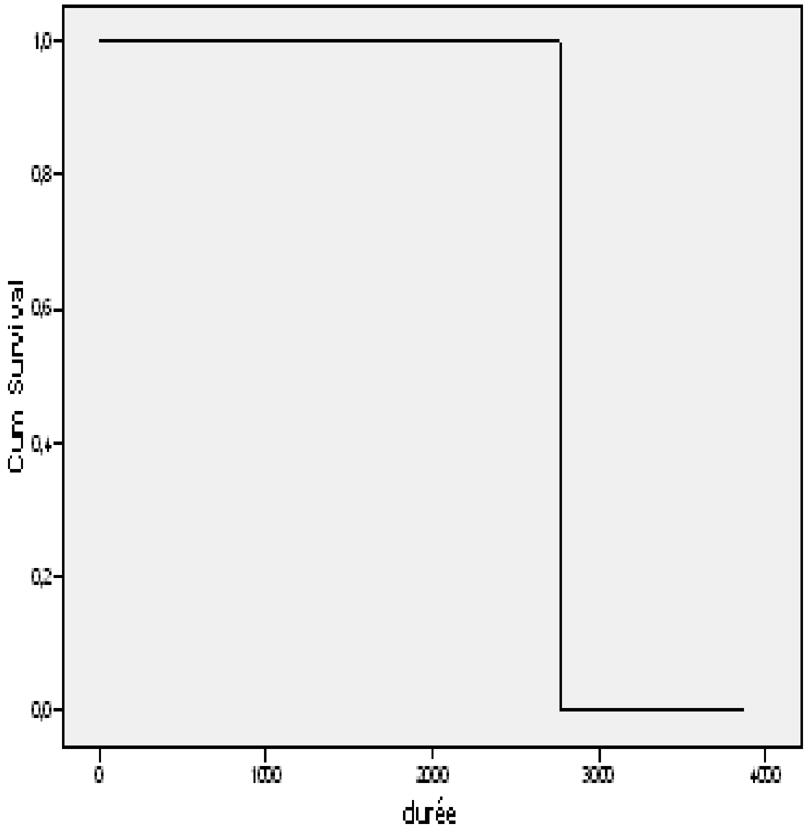


FIG. 6.40 – Fonction de survie estimé

Conclusion générale

Tout au long de notre travail et , dans un premier temps nous avons détaillé le comportement des modèles de survie issus de contrats assurances automobiles, afin d'examiner les effets des facteurs de risques sur les modèles statistiques.
Notre principal objectif était d'étudier la durée de vie d'un contrat d'assurance automobile sur le marché Algérien, et de determiner les facteurs qui induisent les accidents routiers.

Pour notre étude sur le phénomène de résiliation de contrats, nous avions en premier lieu fixé un seuil de censure ,par suite on fait appel à la méthode non-paramètrique de Kaplan-Meier qui doit être considérée comme un préalable à la mobilisation d'autres méthodes d'analyses semi-paramétriques.
En effet,Kaplan-Meier est une méthode qui permet de mieux comprendre la distribution du parcours d'un événement (une variable) au cours du temps et permet aussi de comparer la distribution des covariables en sous population.
Par suite un modèle semi paramètrique de COX , pour quantifier les caractéristiques des variables sur le risque proportionnel.
Ces résultats restent valables pour notre seuil choisi à 30% du coût de la sinistralité. En conclusion, cette étude nous a permis d'illustrer les méthodes d'estimations des durées de vie en assurance non vie, d'obtenir et de determiner des éléments intéressants à l'assureur pour bien choisir ces assurés.

Annexe

Table de survie d'Age.C

Table de survie de l'AMC

Table de survie pour Permis

Table de survie B.M

Survival Table

Age C		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
1	1	679,000	2	1,000	,000	1	3451
	2	922,000	2	,999	,000	2	3450
	3	1142,000	2	,999	,001	3	3449
2	1	1256,000	2	,999	,001	4	3448
	2	1453,000	1	.	.	4	3447
	3	1458,000	2	,999	,001	5	3446
	4	1463,000	1	.	.	5	3445
	5	1466,000	1	.	.	5	3444
	6	1471,000	1	.	.	5	3443
	7	1472,000	1	.	.	5	3442
	8	1492,000	1	.	.	5	3441
	9	1497,000	2	,998	,001	6	3440
	10	1613,000	2	,998	,001	7	3439
	11	1632,000	2	,998	,001	8	3438
	12	1682,000	2	,997	,001	9	3437
	13	1691,000	2	,997	,001	10	3436
	14	1780,000	2	,997	,001	11	3435
	15	1806,000	1	.	.	11	3434
	16	1808,000	2	,997	,001	12	3433
	17	1814,000	2	,996	,001	13	3432
	18	1822,000	2	,996	,001	14	3431
	19	1828,000	1	.	.	14	3430
	20	1829,000	2	,996	,001	15	3429
	21	1831,000	2	,995	,001	16	3428
	22	1832,000	1	.	.	16	3427
	23	1833,000	2	,995	,001	17	3426
	24	1841,000	2	.	.	18	3425

FIG. 6.41 – Table de la fonction de survie Age.C

Survival Table

AMC*		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
1	1	679,000	2	,998	,002	1	480
	2	922,000	2	,996	,003	2	479
	3	1142,000	2	,994	,004	3	478
	4	1256,000	2	,992	,004	4	477
	5	1453,000	1	.	.	4	476
	6	1458,000	2	,990	,005	5	475
	7	1463,000	1	.	.	5	474
	8	1466,000	1	.	.	5	473
	9	1471,000	1	.	.	5	472
	10	1472,000	1	.	.	5	471
	11	1492,000	1	.	.	5	470
	12	1497,000	2	,987	,005	6	469
	13	1613,000	2	,985	,005	7	468
	14	1632,000	2	,983	,006	8	467
	15	1682,000	2	,981	,006	9	466
	16	1691,000	2	,979	,007	10	465
	17	1780,000	2	,977	,007	11	464
	18	1806,000	1	.	.	11	463
	19	1808,000	2	,975	,007	12	462
	20	1814,000	2	,973	,007	13	461
	21	1822,000	2	,971	,008	14	460
	22	1828,000	1	.	.	14	459
	23	1829,000	2	,969	,008	15	458
	24	1831,000	2	,966	,008	16	457
	25	1832,000	1	.	.	16	456

FIG. 6.42 – Table de la fonction de survie AMC

Survival Table

Permis		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
1	1	679,000	2	1,000	,000	1	3526
	2	922,000	2	,999	,000	2	3525
	3	1142,000	2	,999	,000	3	3524
	4	1453,000	1	.	.	3	3523
	5	1458,000	2	,999	,001	4	3522
	6	1463,000	1	.	.	4	3521
	7	1466,000	1	.	.	4	3520
	8	1471,000	1	.	.	4	3519
	9	1472,000	1	.	.	4	3518
	10	1492,000	1	.	.	4	3517
	11	1497,000	2	,999	,001	5	3516
	12	1632,000	2	,998	,001	6	3515
	13	1780,000	2	,998	,001	7	3514
	14	1806,000	1	.	.	7	3513
	15	1808,000	2	,998	,001	8	3512
	16	1814,000	2	,997	,001	9	3511
	17	1822,000	2	,997	,001	10	3510
	18	1828,000	1	.	.	10	3509
	19	1829,000	2	,997	,001	11	3508
	20	1831,000	2	,997	,001	12	3507
	21	1832,000	1	.	.	12	3506
	22	1833,000	2	,996	,001	13	3505
	23	1841,000	2	.	.	14	3504
	24	1841,000	2	,996	,001	15	3503
	25	1844,000	1	.	.	15	3502
	26	1846,000	2	,995	,001	16	3501
	27	1859,000	2	,995	,001	17	3500

FIG. 6.43 – Table de la fonction de survie Permis

BM		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
1	1	679,000	2	1,000	,000	1	3521
	2	922,000	2	,999	,000	2	3520
	3	1142,000	2	,999	,000	3	3519
	4	1256,000	2	,999	,001	4	3518
	5	1453,000	1	.	.	4	3517
	6	1458,000	2	,999	,001	5	3516
	7	1463,000	1	.	.	5	3515
	8	1466,000	1	.	.	5	3514
	9	1471,000	1	.	.	5	3513
	10	1472,000	1	.	.	5	3512
	11	1492,000	1	.	.	5	3511
	12	1497,000	2	,998	,001	6	3510
	13	1613,000	2	,998	,001	7	3509
	14	1632,000	2	,998	,001	8	3508
	15	1682,000	2	,997	,001	9	3507
	16	1780,000	2	,997	,001	10	3506
	17	1806,000	1	.	.	10	3505
	18	1808,000	2	,997	,001	11	3504
	19	1814,000	2	,997	,001	12	3503
	20	1822,000	2	,996	,001	13	3502
	21	1828,000	1	.	.	13	3501
	22	1829,000	2	,996	,001	14	3500
	23	1831,000	2	,996	,001	15	3499
	24	1832,000	1	.	.	15	3498
	25	1833,000	2	,995	,001	16	3497
	26	1841,000	2	.	.	17	3496
	27	1841,000	2	,995	,001	18	3495
	28	1844,000	1	.	.	18	3494
	29	1846,000	2	,995	,001	19	3493

FIG. 6.44 – Table de la fonction de survie B.M

Resumé

Dans cet mémoire nous appliquons les méthodes non paramétriques et semi paramétriques pour étudier la durée de vie de contrats assurance non-vie sur le marché Algérien et les facteurs qui influent sur cette durée.

Bibliographie

[1] Alberti.C, Timsit.J.F, and Chevret.S. Analyse de survie : le test du logrank. *Rev Mal Respir*, 2005.

[2] Berlanger.I. *Excel 97 Petit manuel d'utilisation*, Aout 2001.

[3] Boukhari.M. Sur les courbes de survie et leurs applications en assurance de personnes. Master's thesis, USTHB, 2008.

[4] Boukhetala.K, Marion.J.M, and Oulidi.A. Apport des méthodes de durée de vie au domaine de l'assurance, application aux contrats d'assurances automobiles. *INRIA*, 2009.

[5] Brouhns.N. *Ingéniérie actuarielle. Les modèles de régression non linéaires comme solutions à divers problèmes actuariels*. PhD thesis, Université Catholique de Louvain, Décembre 2005.

[6] Charpentier.J.C. Stage latex niveau débutant, 2006.

[7] Charreaux.G. *L'Analyse Financère*.

[8] Chimka.J.R. *An introduction to the observation of graduation as survival data*.

[9] Collett.D. Modelling survival data in medical research, 2003.

[10] Courilleau.E and Marion.J.M. Comparaison de modèles d'estimation de la fonction de survie,appliquée à des données routières. *Rev.Statistique Appliquée*, 47(1) :81–97, 1999.

[11] Dessovic.A and Eigner.M.F. Cox regression analysis, November 2008.

[12] Dietz.K, Gail.M, Krickeberg.K, Samet.J, and Tsiatis.J. *Statistics for Biology and Health, Survival Analysis*. Edition Springer, 2003.

[13] FFSA. *Le bonus-malus*, Mai 2009.

[14] Fox.J. *Cox Proportional-Hazards Regression for Survival Data*, February 2002.

[15] Le Goff.J.M and Forney.Y. *Estimations non-paramétriques avec SPSS*, Décembre 2003.

[16] Huber.C. *Cours de Biostatistique et Modélisation*. Université Paris V, René Descartes.

[17] Huber.C. *Modèles pour des durées de survie*.

[18] Jalby.V, editor. *Traitement des données avec EXEL*, Novembre 2007.

[19] Kalbf'Leisch.J.D and Ross.L.P. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, 2002.

[20] Kelly.J and Simmons.C. *The unofficial guide to Microsoft Office Excel 2007*. Wiley, 2007.

[21] Letué.F. A semiparametric shock model for a pair of event-related dependent censored failure times. *Elsiver*, pages 3869–3884, March 2008.

[22] Lorino.M.T. *Modèles statistiques pour des données de survie corrélées*. PhD thesis, I.N.A Paris, mai 2002.

[23] Lozano.V. *Tout ce que vous avez toujours voulu savoir sur LATEX*. In Libro Veritas, 2008.

[24] Mariane.M and Mzioud.O. *Etude de l'équilibre financier du portefeuille d'assurance maladie*. PhD thesis, INSEA, juin 2006.

[25] Marion.J.M, Loizeau.J.M, and Oulidi.A. Méthode d'éstimation de durées de vie de contrats d'assurances automobiles. *INRIA*, Mai 2009.

[26] Olivier.B. *Utilisation de modèles à direction révélatrice unique pour les modèles de durée*. PhD thesis, Université Paris VI, novembre 2009.

[27] Perrigot.R, Cliquet.G, and Mesbah.M. Possible applications of survival analysis in franchising research. *Int. Rev. of Retail, Distribution and Consumer Research*, page 129–143, January 2004.

[28] Pilo.G. Le chômage et les conséquences de la réduction des indemnités de l'assurance chômage. Master's thesis, Université de Genève, 2006.

[29] Pintilie.M. *Competing Risks .A Practical Perspective*. Wiley, 2006.

[30] Planchet.F. *Statistique des modèles paramétrique et semi-paramétriques*, Octobre 2008.

[31] Posière.J.P. *Mathématique appliquées à la gestion*. Les zoom's, août 2005.

[32] Shih.J.H and Shou.E.L. Semiparametric estimation of a nasted random effects model for the analysis of multi-level clustered failure time data. *Science Direct*, 2009.

[33] Tsiatis.A.A. *Semiparametric Theory and Missing Data*. Springer, 2006.

[34] ZHENG.Z.K. A class of estimators of the mean survival time from interval censored data with application to linear regression. *Applied matematics*, 23(4) :377–390, 2008.