Institut des ACTUAIRES

le cnam

Mémoire présenté devant le CNAM pour l'obtention du Master

Droit Economie Gestion, mention Actuariat et l'admission à l'Institut des Actuaires

le 22 novembre 2024

Par : Hicham BOUKHARSA

Titre: Construction des hypothèses Best Estimate biométriques pour un produit de Réassurance de dépendance (Long-Term Care) : Calcul de provision mathématique de rente et estimation des IBNR

Confidentialité : X NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

signatures

Présidente du Jury :
M. Stéphane LOISEL

Membres présents du jury de l'Institut des Actuaires :

M. Kamel ASSAM
M. Jean BRUNET
M. Guillaume GORGE

Membres présents du jury du Cnam :

M. Olivier DESMETTRE
M. Kristiano BEJKO

Secrétariat :

Bibliothèque :

Entreprise :
Nom : PwC

Directeur de mémoire en entreprise :

Nom : Vincent GIBRAIS
Signature :

Invité :
Nom :
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

Résumé

Les progrès médicaux et l'amélioration des conditions de vie se sont traduits par l'allongement de l'espérance de vie dans le monde. Cette augmentation de l'espérance de vie et l'arrivée à des âges avancés de la génération du baby-boom ont pour effet le vieillissement de la population mondiale. Il résulte du vieillissement de la population une augmentation du nombre de personnes âgées dépendantes. En outre, ces dernières décennies ont été marqué par une augmentation du nombre des enfants en état de dépendance à cause des maladies de trouble mentales, notamment l'autisme.

Dans ce contexte les assureurs ont un rôle important à jouer en proposant des couvertures adaptées pour couvrir le risque de dépendance, ils rencontrent néanmoins quelques difficultés à cause de la nature de ce risque récent et compliqué à définir. L'absence de maîtrise technique du risque de dépendance a conduit les organismes assureurs à recourir à la réassurance pour bénéficier de l'expertise du réassureur et atténuer les risques pris.

Le présent mémoire propose de répondre à un appel d'offre d'un réassureur international concernant le calcul des provisions technique d'un produit dépendance en run-off commercialisée dans un pays hors l'union européen. Il donne l'occasion de s'interroger sur la suffisance des montants des provisions et de s'assurer leur conformité par rapport à la réglementation en vigueur. Pour cela nous nous sommes employés à déterminer les lois de maintien qui permettront de calculer la provision mathématique pour rente ainsi que le calcul des IBNR pour déterminer le montant des sinistres déclarés tardivement.

Le premier chapitre du mémoire est un avant-propos permettant de poser quelques éléments de contexte sur la dépendance. On rappelle également quelques éléments concernant l'intérêt de la réassurance et le principe de provisionnement d'un produit dépendance.

La seconde partie sera dédiée à l'exploration et au retraitement des données qui ont servi à l'étude. Cette partie nous permet également de détailler le produit étudié et ses caractéristiques, notamment le produit présente la particularité de couvrir les adhérents dès la naissance et d'offrir une couverture bornée à 60 mois.

Le troisième chapitre sera consacré à la modélisation de la dépendance et à la détermination d'une loi de maintien à l'aide de l'estimateur de Kaplan Meier et de la régression de Cox. Cette section se terminera par la mise en place du modèle multi-états tenant en compte les différents états d'un assuré (Dépendant, valide ou décès).

Le quatrième chapitre présente quelques méthodes de calcul des IBNR, nous introduirons la méthode déterministe de Chain Ladder et les méthodes stochastiques de Mack et de Bootstrap ODP qui ont pour but de quantifier l'incertitude liée à l'estimation des IBNR.

Finalement, après avoir mis notre modèle en place, nous déterminerons le montant des provisions technique comme la somme de la Provision Mathématique de rente et des IBNR.

Abstract

Medical advancements and improvements in living conditions have resulted in an increase in life expectancy worldwide. This increase in life expectancy, along with the aging of the baby boomer generation, has led to the aging of the global population. As a result of population aging, there has been a rise in the number of elderly dependent individuals. Additionally, these past decades have been marked by an increase in the number of children with mental health disorders, particularly autism.

In this context, insurers have an important role to play by offering appropriate insurance policies to cover the risk of Long-Term Care. However, they encounter some difficulties due to the nature of this recent and complex risk. The lack of technical expertise of the risk of dependency has led insurance companies to resort to reinsurance to benefit from the expertise of the reinsurer and mitigate the risks taken.

This dissertation aims to respond to a tender from an international reinsurer regarding the calculation of technical provisions for a dependency product marketed in a country outside the European Union. It provides an opportunity to question the adequacy of the provision amounts and to ensure their compliance with current regulations. To achieve this, we have endeavored to determine the duration laws that will allow for the calculation of the annuity reserve as well as the calculation of IBNR (Incurred But Not Reported) to determine the amount of claims reported late.

The first part of this dissertation serves as an introduction to contextualize the subject of dependency. It also recalls the importance of reinsurance and the principles of provisioning for a dependency-related product.

The second part will be dedicated to the exploration and processing of the data used in this study. It will also We also describe the product specificities with notably the presence of people of every age and the cover duration which is 60 months.

The third part will focus on modeling dependency and determining a duration law using the Kaplan-Meier estimator and Cox regression. This section will conclude with the establishment of a multi-state model taking into account the different states of an insured (dependent, valid, or deceased).

The fourth chapter will present several methods for calculating IBNR. We will introduce the deterministic Chain Ladder method as well as the stochastic methods of Mack and Bootstrap ODP, which aim to quantify the uncertainty associated with IBNR estimation.

Finally, once our model is established, we will determine the amount of technical provisions as the sum of the mathematical provision and IBNR.

Remerciements

Je tiens à remercier Vincent Gibrais, pour son encadrement précieux et ses conseils avisés qu'il m'a dispensé ainsi que pour sa disponibilité afin de m'accompagner tout au long de ce mémoire.

J'adresse également mes remerciements à l'ensemble des professeurs et intervenants du Master Actuariat et notamment à mes tuteurs académiques, François Weiss et Stéphane Loisel, qui se sont montrés disponible pendant la réalisation de ce mémoire

Un grand merci également à mes collègues de PwC France pour leur sympathie et leurs conseils.

Et enfin, un énorme merci à mes parents, mon épouse et à ma fille, qui ont toujours su m'encourager, me motiver et m'accompagner durant ces années de cours du soir et ces week-ends à travailler. Merci.

Table des matières

Note de synthèse :

L'évolution de la science, l'amélioration des conditions de vie ou encore la mise en place de nouvelles règles sanitaires ont et continuent à contribuer à l'augmentation de l'espérance de vie ainsi qu'au vieillissement généralisé de la population mondiale, ceci provoque depuis une dizaine d'années une prise de conscience générale quant à l'émergence d'un nouveau risque dans le champ de l'assurance vie. La dépendance, qui se définit comme l'impossibilité pour une personne d'effectuer certains actes essentiels de la vie quotidienne, se traduit par un besoin de prise en charge visant à compenser la perte d'autonomie.

Dans ce contexte les assureurs ont un rôle important à jouer en proposant des couvertures de dépendance centrées sur des garanties couvrant en partie les coûts induits par les aides quotidiennes nécessaires et liées à la perte d'autonomie de l'assuré. Ils rencontrent néanmoins quelques difficultés. Le risque de dépendance reste un risque jeune et mal maitrisé. Le manque de données d'expérience notamment en termes de volume est un frein à l'assimilation de ce risque. Souvent, les assureurs font appel aux réassurances pour bénéficier d'un accompagnement technique spécialisé qui garantit une meilleure maîtrise des risques. Cet appui permet non seulement une évaluation plus précise et une gestion plus efficace des risques, mais aussi un pilotage optimisé du capital nécessaire à l'entreprise d'assurance. Ce dernier aspect est crucial pour maintenir la solvabilité de l'assureur et pour lui permettre de poursuivre ses activités tout en respectant les exigences réglementaires et en sécurisant les intérêts de ses assurés.

L'objectif de ce mémoire est la construction d'un modèle que nous jugeons complet pour le calcul des provisions techniques. Ce mémoire est organisé autour de quatre parties :

La première partie du mémoire permet de poser quelques éléments de contexte sur la dépendance, l'intérêt de la réassurance et le principe de provisionnement du risque de dépendance.

Dans une seconde partie, nous détaillerons l'approche théorique de l'analyse du risque de dépendance à l'aide de la théorie de durée. Après une présentation des modèles de durée nous nous intéresserons à estimer de manière fiable et à l'aide d'une démarche cohérente la loi de maintien nécessaires au calcul des provisions techniques, pour y parvenir, nous allons nous appuyer sur l'estimateur non paramétrique Kaplan Meier, ensuite nous allons déployer le modèle semi paramétrique de Cox afin de déterminer les variables les plus discriminantes qui affectent le risque de maintien. A la fin de cette section, nous allons essayer de mettre en place un modèle markovien que nous allons tester sa pertinence et ses hypothèses sous-jacentes.

La troisième partie s'attache à présenter les techniques actuarielles nécessaires pour la détermination de montants des sinistres tardifs (IBNR). Dans un premier temps, nous allons déterminer le montant des IBNR par la méthode déterministe de Chain Ladder, puis nous appliquerons des méthodes à fondement stochastique à savoir la méthode de Mack et la méthode de Bootstrap afin d'apporter une estimation de la variabilité des IBNR et de définir l'erreur d'estimation ainsi que des intervalles de confiance.

Le dernier chapitre vient mettre en application les différentes méthodes utilisées pour obtenir le montant de la provision technique pour le produit de dépendance.

Modélisation, estimation et principaux résultats :

Le produit de dépendance que nous allons explorer et puis modéliser présente la particularité de couvrir les adhérents dès la naissance et d'offrir une couverture qui s'étale sur 5 ans au maximum.

Le montant de la provision pour le risque de dépendance sera la somme de la provision mathématique de rente pour les sinistre connus et encours d'indemnisation et du montant IBNR pour les sinistres survenus mais non encore déclarés à l'assureur et ensuite au réassureur.

Notre approche consiste à proposer une modélisation de la loi de maintien pour la détermination de la provision mathématique de rente, puis à faire appel à des méthodes non-vie de projection basées sur les triangles de liquidation pour déterminer le montant des IBNR.

- **La provision mathématique de rente :**

Notre approche consiste à proposer une modélisation des lois de maintien, pour ce besoin deux modèles permettent de les modéliser : le modèle classique pour estimer la loi de survie, qui est très intuitif et le modèle multi-états, plus technique que le précédent.

Le premier modèle repose sur l'estimation d'une loi de maintien via les probabilités à chaque pas de temps de faire encore partie d'une population donnée, ou encore par les probabilités de sorties de cette population. Deux approches seront introduites :

1) L'estimateur de Kaplan Meier fera l'objet de notre étude, divers travaux démontrent la robustesse de cet estimateur et présente plusieurs avantages notamment la prise en compte des données censurées et sa facilité opérationnelle d'implémentation. A l'issue de cette modélisation, les taux bruts de maintien seront lissés en utilisant deux méthodes : la régression LOESS et le lissage de Whittaker-Henderson. Ces techniques sont appliquées pour corriger les irrégularités potentielles présentes dans les taux bruts.
2) La deuxième approche consiste à prendre en compte l'hétérogénéité pour établir une loi d'expérience robuste. Notre choix se porte sur le modèle de Cox : semi-paramétrique, multiplicatif, à hasard proportionnel. Nous procéderons ainsi à la vérification de l'hypothèse des risques proportionnels.

Le deuxième modèle concerne la modélisation multi-états à trois états dans lesquels l'individu assuré peut se trouver : **l'état de dépendance**, **l'état valide** ; suite à l'amélioration de son état de santé ; et **le décès**. L'individu ne peut se trouver que dans un seul des trois états précédents. Trois paramètres sont à estimer dans ce modèle : la probabilité de devenir valide, la probabilité de revenir à l'état dépendant et la probabilité de décéder sachant que l'individu est dépendant. A cette fin, nous proposerons un modèle markovien homogène par morceaux ainsi qu'un modèle semi-markovien localement homogène.

- IBNR : Provision pour sinistres inconnus

En addition de la provision mathématique de rente, le réassureur est obligé de constituer une provision dite IBNR pour se couvrir des sinistres survenus non encore connus.

Il existe un grand nombre de méthodes envisageables pour le calcul des provisions pour sinistres inconnus. Nous allons se référer aux méthodes de triangulation très connu en assurance non-vie, certaines de ces méthodes sont courantes, d'autres ne sont que rarement utilisées. Dans ce mémoire nous nous concentrerons sur les approches les plus utilisées sur le marché dans le cadre du provisionnement, à savoir

1. Chain-Ladder
2. Mack Chain ladder
3. Bootstrap ODP

Résultats de l'étude :

<u>Le calcul de la provision mathématique de rente :</u>

Pour le calcul de la provision mathématique de rente, nous avons utilisé les lois de maintien estimées à l'aide de l'estimateur de Kaplan Meier qui se présentent ci-dessous. Le choix de cet estimateur est motivé par sa simplicité, sa robustesse et son adaptabilité aux différents problèmes qu'on pourra affronter lors d'une analyse de survie (gestion des données censurées, facilité d'implémentation, explicabilité, adaptabilité aux études de petite taille, visualisation graphique…).

Nos lois de maintien sont estimées sur sept classes d'âge et selon le sexe.

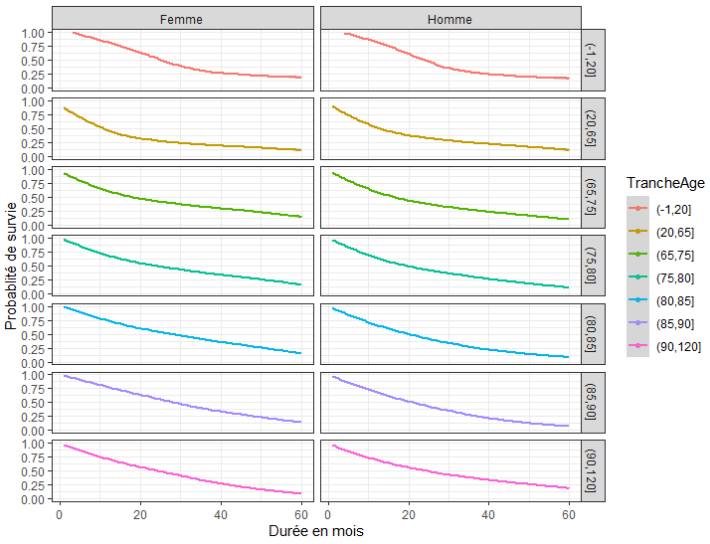


Figure 1 : La courbe de survie par sexe et par tranche d'âge : Lissage par Whittaker-Henderson

Les résultats de l'estimateur de Kaplan-Meier indiquent que les taux de maintiens sont en général un peu plus élevés pour les femmes que pour les hommes. L'analyse des lois de survie en dépendance montre également une sensibilité à l'âge d'entrée en dépendance, plus cet âge est élevé, plus le maintien en dépendance est élevé avec deux exceptions : 1) pour la tranche d'âge ''enfant'' qui montrait une courbe de maintien longue pour les durées moins de 24 mois et 2) au niveau de la tranche d'âge supérieur à 90 ans qui représentait une sortie rapide que nous avons jugé logique et en lien avec l'espérance de vie qui diminue avec l'âge.
Afin de résoudre les taux bruts erratiques obtenus via l'estimateur de Kaplan Meier, nous avons fait recours à des méthodes de lissage, notamment la méthode de LOESS et de Whittaker-Henderson, sur la base de critère de fidélité nous avons retenu le lissage de Whittaker-Henderson pour les deux variables de construction de la loi de maintien.

Le calcul des IBNR

Différentes méthodes d'estimation de montant des IBNR sont proposées dans la littérature : il y a les méthodes dites déterministes et celles dites stochastiques. Le montant des provisions IBNR est modélisé en se basant sur les données historiques du nombre des sinistres qu'on projette jusqu'à l'ultime. Le tableau ci-dessous montre le résultat du nombre ultime de sinistre selon les différentes méthodes :

	Méthode de Chain Ladder	Mack Chain ladder	Bootstrap
Moyenne de la distribution	125.04	125.04	121.1
Ecart type		15.2	18.4
VaR de niveau 99.5%		166	173
Coefficient de volatilité		12.2%	15.2%
Coefficient de comparaisons $\boldsymbol{VaR/BE}$		32.76%	42.98%

Tableau 1 : Le tableau de comparaison entre les différentes méthodes d'estimation des IBNR

A ce nombre de sinistres tardifs, nous rajoutons le nombre de sinistre en attente de 114 pondéré par la probabilité d'acceptation de 85% (acceptance rate).

La répartition finale de nombre IBNR de sinistres et des sinistres avec le statut ‘’en attente’’ se présente comme suit :

	IBNR + sinistre en attente			
Sexe	ageBand	% effectif	Age moyen	Nombre IBNR
Femme	(-1,20]	2,95%	8,6	3
	(20,65]	15,33%	54,2	18
	(65,75]	17,51%	71,0	21
	(75,80]	15,62%	78,3	18
	(80,85]	21,51%	83,1	25
	(85,90]	17,87%	87,8	21
	(90,120]	9,22%	93,2	11
Total Femme		100,00%	75,4	118
Homme	(-1,20]	8,77%	8,0	9
	(20,65]	15,83%	54,7	16
	(65,75]	20,61%	70,7	21
	(75,80]	15,71%	78,2	16
	(80,85]	17,39%	82,9	18
	(85,90]	13,34%	87,9	14
	(90,120]	8,35%	93,4	9
Total Homme		100%	70,2	104

Tableau 2 : Répartition du nombre IBNR et nombre de sinistre en attente par sexe et par tranche d’âge

Résultat final :

Les méthodes utilisées nous ont permis de calculer les lois de maintien en dépendance et de vérifier que les critères discriminants sont bien l’âge et le sexe. Sur le portefeuille étudié, la loi de maintien est assez bien estimée par l’estimateur de Kaplan-Meier. Ensuite le calcul de montant IBNR est basé sur la méthode de Mack Chain Ladder appliquée au triangle de nombre de sinistre déclaré, nous avons proposé une approche qui consiste à répartir le nombre ultime ainsi que les sinistres en attente, en fonction de la structure de portefeuille et puis appliqué les lois d’entrée en maintien produites auparavant.
Finalement le montant de la provision technique n’est simplement que la somme de la Provision mathématique de rente et le montant des IBNR.

Amount in m USD	Cedant account	Reinsurance's Evaluation
In-payment	30.78	37.5
IBNR + OS pnding	2.25	2.14
Total reserve	**33.03**	**39.64**
Paid claims	5.26	5.26
Total BE	**38.29**	**44.9**
Total BE + PAD(*)	**40.20**	**47.15**
Deficit(-)/Surplus(+)		**-6.94**

Tableau 3 : Résultat total de montant de provision technique

Chapitre 1 : Présentation Générale de la réassurance dépendance

I. Définition de l'assurance 'Long Term Care'

L'assurance ''long term care'' est traduite par l'assurance de soins de longue durée. Dans le contexte de notre étude, les deux termes : soins de longue durée et dépendance seront interchangeablement utilisés.

Une assurance de soins de longue durée est un programme d'assurance, dans le cadre duquel l'assuré paie une prime d'assurance mensuelle ou annuelle, en fonction de son âge. Ce paiement lui garantit que si par mégarde il devenait par la suite une personne assistée dont le fonctionnement quotidien n'est pas autonome, il aura droit à des indemnités d'assurance mensuelles de la compagnie d'assurance. Ces indemnités lui permettront d'obtenir une compensation, un remboursement de frais ou des prestations de soins de longue durée selon les termes de la police.

Cette assurance regroupe un ensemble de garanties qui ont pour but à couvrir les frais inhérents à la survenance d'une perte d'autonomie empêchant la personne dépendante à effectuer seul tout ou partie des actes de la vie quotidienne. Cet état peut être le résultat de causes diverses, par exemple :

– Le cancer ou les maladies cardio-vasculaires

– Les maladies neurologique : maladies d'Alzheimer, de Parkinson, état de démence sénile ou autisme.

Être dépendant, en tant que telle, n'est pas une notion relative à l'âge. On peut être dépendant à tout âge en raison d'un handicap, d'un accident, temporairement ou définitivement. Pour autant, la population la plus concernée par la perte d'autonomie est bien évidemment celle des personnes âgées. A cet effet les assureurs ont su adapter leurs produits aux besoins de leurs clients, les garanties d'une assurance des soins de longue durée sont principalement sous forme d'une rente accompagnées souvent des services d'assistance au domicile.

Évaluation de la dépendance

Il existe différents instruments qui permettent d'évaluer le niveau de dépendance d'un individu en perte d'autonomie. Nous allons se limiter à présenter le système d'évaluation proposé par notre assureur (cédante).

- Activité de la Vie Quotidienne (AVQ) :

Il s'agit d'un système simple et très largement utilisé dans le monde. Le principe est de compter le nombre d'activités de la vie quotidienne que l'individu est capable d'exercer avec ou sans équipements adaptés ou assistance d'une tierce personne pour évaluer le niveau de dépendance. Les AVQ sont au nombre de 6 :

1. Se lever et s'asseoir – la capacité autonome de l'assuré de passer de l'état couché à l'état assis et de se lever d'un siège y compris d'un fauteuil roulant ou d'un lit.
2. S'habiller et se déshabiller - la capacité autonome de l'assuré de mettre des articles d'habillement de toute sorte et de les enlever y compris le raccord ou l'assemblage d'une ceinture médicale ou d'un membre artificiel.
3. Toilette - la capacité autonome de l'assuré de se laver dans une baignoire, de se doucher dans une douche ou de toute autre manière usuelle, y compris entrer et sortir de la baignoire ou de la douche.
4. Manger et boire - la capacité autonome de l'assuré de s'alimenter par tout moyen sauf manger à l'aide d'une paille, mais y compris boire avec une paille, les aliments aillant été préparés et lui ayant été servis.
5. Incontinence - la capacité autonome de l'assuré de maîtriser sa défécation ou son urination ; l'incapacité de maîtriser une de ces opérations nécessitant par exemple l'usage permanent d'une stomie, d'un cathéter pour la vessie, de couches ou d'articles absorbants – sera considérée être une incontinence.
6. Mobilité - la capacité autonome de l'assuré de se déplacer, sans l'aide d'autrui ; l'assistance de béquilles, d'une canne, d'un déambulateur ou de toute autre accessoire y compris un accessoire mécanique, motorisé ou électronique, permettant à l'Assuré de se déplacer de manière autonome, ne seront pas considérés comme une atteinte à la capacité autonome de l'Assuré de se déplacer

La plupart des contrats d'assurance choisissent quatre AVQ pour définir le degré de dépendance des individus. Si trois AVQ sur quatre sont jugées invalides, l'individu sera atteint d'une dépendance dite lourde par comparaison à la dépendance légère où deux AVQ sur 4 sont validées.

Dans notre cas d'étude l'assureur a choisi que le déclenchement de la couverture se fait si l'assurer ne peut plus faire au moins 3 des AVQ sur 6.

Fonctionnement d'un contrat en Assurance des soins de longue durée (LTC).

La souscription d'une police indique qu'en contrepartie du paiement d'une prime d'assurance, et sous réserve des conditions, instructions et exceptions précisées dans le contrat, l'assureur accordera au bénéficiaire éligible une indemnité de soins en cas de dépendance. L'indemnité de soins sera accordée dans un cas de sinistre ayant lieu pendant la période d'assurance, selon les termes prescrits par cette police, ses conditions et réserves.

Dans le cadre d'une souscription individuelle, l'assuré paye une cotisation mensuelle, fixe en général, jusqu'à ce qu'il passe en dépendance ou décède. Dans le cadre d'une assurance collective, une entreprise souscrite auprès d'un assureur une assurance dépendance collective pour l'ensemble de ses salariés : l'employeur paye une partie des primes et les salariés payent le complément.

Le cadre de ce mémoire porte sur la modalisation de risque de dépendance (contrat collectif) en cas de réassurance, ainsi nous jugeons important d'introduire dans un premier temps les généralités et les mécanismes de base de la réassurance

II. L'activité de réassurance

1. Le besoin d'une réassurance

La compagnie d'assurance se tourne vers un réassureur pour plusieurs raisons, nous citons ci-dessous quelques-unes :

- Limiter la probabilité de ruine de l'assureur

Les assureurs font bien souvent appel à la réassurance pour diminuer leur probabilité de ruine.

En effet, la solvabilité d'un assureur peut être remise en cause par un sinistre de montant unitaire important. C'est par exemple le cas en assurance vie lorsqu'un capital décès de plusieurs millions repose sur une seule tête, ou encore en assurance de dommages en cas d'incendie d'un immeuble industriel. Les assureurs sont également redoutés par une importante fréquence des sinistres, elle se manifeste lorsque le portefeuille présente de la dépendance entre les risques.

Les compagnies d'assurance peuvent subir des écarts de résultats dus à la volatilité de la sinistralité en montant ou en nombre. L'un des rôles de la réassurance est de répondre au besoin d'une rémunération stable de l'actionnariat et donc de lisser le résultat.

- Un support financier

En cas de fort développement de l'activité d'un assureur, les fonds propres peuvent ne plus être réglementairement suffisants. La réassurance permet une diminution de l'exigence minimale de marge de solvabilité pour répondre aux exigences réglementaires.

Aussi, la réassurance permet d'augmenter la capacité de souscription de l'assureur, en effet, la réassurance lui permet d'accepter en garantie des risques pour des montants plus élevés que ses propres capacités d'engagement.

- Un support technique

L'assureur peut manquer de maîtrise concernant une branche inconnue et ainsi collaborer avec un réassureur qui l'accompagnera dans la conception d'un produit (tarification, sélection des risques, garanties du contrat, etc.). C'est surtout le cas en assurance de personnes et notamment en dépendance. Le réassureur mutualisant les risques de plusieurs compagnies d'assurance, il bénéficie d'une quantité importante de données. Il est alors plus à même de réaliser certaines études actuarielles telles que la construction de lois d'entrée en dépendance. Dans certains cas, la compagnie cédante nécessite une mission d'audit. Le réassureur inspectera alors le travail effectué par la compagnie d'assurance et émettra des recommandations. Le but étant pour le réassureur, d'acquérir de nouveaux contrats.

2. Les modes de réassurance

On distingue trois grands types de contrats en réassurance :

Les traités obligatoires

Le traité de réassurance permet de réassurer la totalité d'un groupe de polices, voire d'un portefeuille sans distinction particulière.

Pour l'assureur

- Avantage : Simplicité administrative, clarté des engagements réciproques.
- Inconvénient : Obligation systématique de céder tous les risques entrant dans le traité.

Pour le réassureur

- Avantage : Pas d'antisélection des risques, l'assureur cède aussi bien les affaires fortement exposées que celles qui ne le sont que faiblement.
- Inconvénient : Impossibilité pour le réassureur de sélectionner individuellement les risques et donc moins de visibilité à priori sur les risques acceptés. C'est la connaissance du sérieux de la compagnie d'assurance et des statistiques qui le guidera dans son choix de participation à un tel traité.

Les cessions facultatives

La réassurance facultative permet de céder des polices d'assurance spécifiques sous des conditions particulières. Le réassureur n'a pas l'obligation de l'accepter comme la cédante ne doit pas obligatoirement les céder.

Pour l'assureur

- Avantage : Pas de déséquilibre du portefeuille dû à un risque important.
- Inconvénient : Lourdeur de gestion et de souscription du risque.

Pour le réassureur

- Avantage : Très bonne maîtrise et information sur le risque, liberté de souscription.
- Inconvénient : Mutualisation des risques moins évidente.

Les cessions facultatives/obligatoires

L'assureur est libre de soumettre le risque au réassureur. Par contre, le réassureur doit l'accepter. Ce mode est très peu pratiqué car le réassureur est lésé. Il fait face à un risque d'antisélection, ainsi, il doit avoir une totale confiance en l'assureur. Ce mode de réassurance est aussi appelé « facultatif/obligatoire », « facob » ou « open-cover ».

3. Structure de réassurance

Une autre possibilité de distinction des types de réassurance se fait selon sa structure :

La réassurance proportionnelle

La réassurance proportionnelle consiste à partager proportionnellement les engagements pris par l'assureur et le réassureur au niveau des primes et des sinistres. Les deux méthodes de réassurance proportionnelle sont :

- Le Quote-Part (QP)

C'est la forme la plus simple de réassurance. Il se caractérise par un pourcentage de cession pour chacun des risques entrant dans le cadre du traité. Ce pourcentage s'applique aussi bien pour les primes que pour les sinistres. Ceci implique le partage du sort. On note $QP\ x$ avec x le taux de cession du traité.

Pour un portefeuille réassuré et composé de n polices versant chacune une prime P_i, la prime cédeé en réassurance sera $x \times \sum_{i=1}^{n} P_i$. De même si on observe m sinistres de de montant unitaire S_i, le montant à charge du réassureur sera $x \times \sum_{i=1}^{m} S_i$

Exemple :
Prenons le portefeuille suivant :

N° de police	Prime	Somme assurée	Sinistre
1	7	10	7
2	10	15	12
3	12	12	8
4	5	13	7
5	16	30	15
Totaux	50	80	49

Tableau 4 : Exemple de portefeuille

L'assureur décide de se réassurer avec un traité QP 40%

N° de police	Prime		Sinistre	
	Assureur	Réassureur	Assureur	Réassureur
1	4,2	2,8	4,2	2,8
2	6	4	7,2	4,8
3	7,2	4,8	4,8	3,2
4	3	2	4,2	2,8
5	9,6	6,4	9	6
Totaux	30	20	29,4	19,6

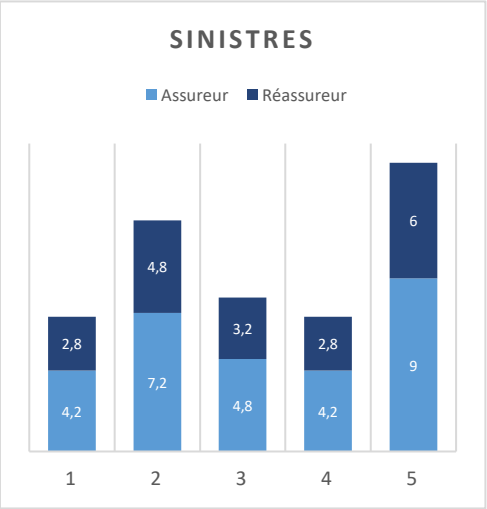


Figure 2 : Répartition des primes et sinistres pour un QP 40%

- L'Excédent de Plein (EP)

L'excédent de pleins (surplus share - SS) permet au réassureur de n'intervenir que dans le cas des sinistres qui dépassent un certain seuil prédéfini - appelé plein de rétention. Les primes et les sinistres sont partagés selon ce ratio appelé taux de cession.
Le taux de cession se calcule pour chacun des risques. Si on note la somme assurée du risque, alors le taux de cession de ce même risque est donné par la formule :

$$x_i = \frac{max(min(SA_i - plein; capacité); 0)}{SA_i}$$

En gardant les mêmes notations que précédemment, la prime cédée en réassurance sera $\sum_{i=1}^{n} x_i \times P_i$. De même si on observe m sinistres de de montant unitaire S_i, le montant à charge du réassureur sera $\sum_{i=1}^{m} x_i \times S_i$

Exemple :
Sur l'exemple précédent, nous procédons à une couverture de réassurance par un 6 EP 5. Le plein de rétention est de 5, la capacité 6 x 5 = 30. Les taux de rétention sont donc les suivants

N° de police	Somme assurée	Taux de cession
1	10	1/2
2	15	2/3
3	12	3/5
4	13	5/8
5	30	5/6

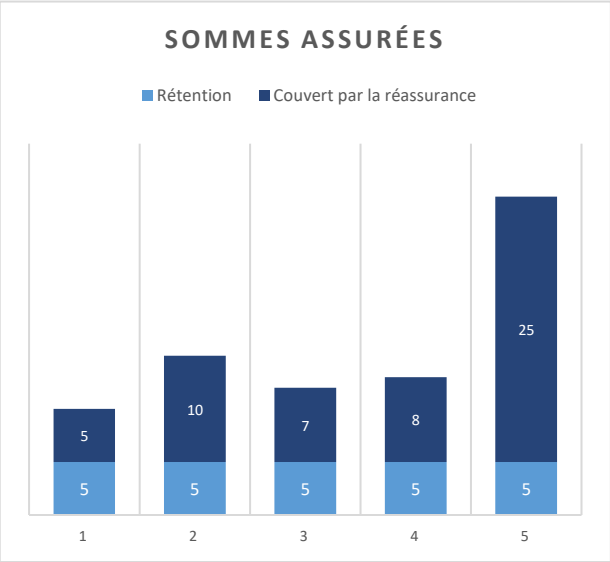


Figure 3 : Répartition des sommes assurées pour un 6 EP 5

Le taux de cession de chacun des risques est ensuite appliqué aux primes et aux sinistres :

N° de police	Prime		Sinistre	
	Assureur	Réassureur	Assureur	Réassureur
1	3,5	3,5	3,5	3,5
2	3,3	6,7	4,0	8,0
3	5,0	7,0	3,3	4,7
4	1,9	3,1	2,7	4,3
5	2,7	13,3	2,5	12,5
Totaux	16,4	33,6	16,0	33,0

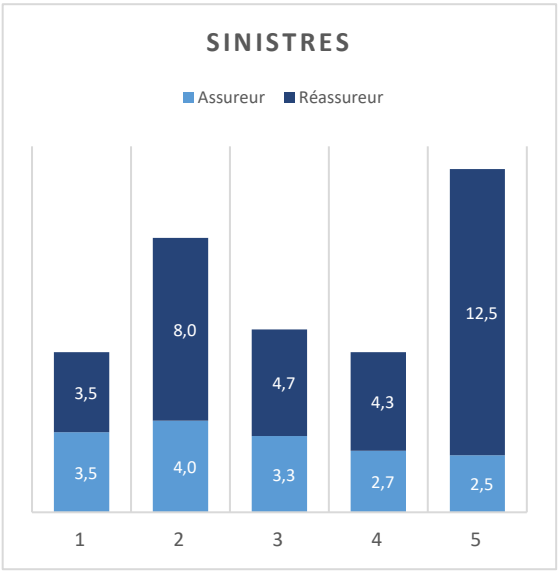


Figure 4 : Répartition des primes et sinistres pour un 6 EP 5

La réassurance non proportionnelle

Sous cette forme, le montant de sinistralité pris en charge par le réassureur est déterminé à priori avec un seuil d'intervention et une limite maximale, appelés respectivement priorité et portée. Il n'y a pas de lien direct entre la sinistralité à charge du réassureur et la prime originale ou la somme assurée du risque

- L'excédent de sinistre (XS)

Le contrat peut être établi par risque ou par évènement :

- Dans le premier cas, le réassureur intervient chaque fois qu'un sinistre supérieur à la priorité survient, pour une police donnée ;
- Dans le second cas, il intervient à chaque survenance d'évènement, bien souvent sur plusieurs polices.

Il existe aussi un cas où la couverture se fait d'abord par risque, puis par évènement sur la rétention. La prime du traité est généralement exprimée sous forme d'un taux s'appliquant aux primes acquises (ou émises) du portefeuille réassuré. On note ces traités $\boldsymbol{Portée\ XS\ Priorité}$. Si on note S_i le montant de la $i^{ème}$ sinistre (par risque ou par évènement selon le traité), la part à charge du réassureur, notée $S_i^{réa}$, sera donnée par la formule :

$$\boldsymbol{S_i^{réa} = max(min(S_i - Priorité; Portée); 0)}$$

Exemple

Appliquons une couverture de réassurance au portefeuille précédent avec un 10 XS 4. Le plafond de ce traité est de 14 (priorité + portée).

N° de police	Sinistre brut de réassurance	Cédé en réassurance	À charge de l'assureur
1	7	3	4
2	12	8	4
3	8	4	4
4	7	3	4
5	15	10	5

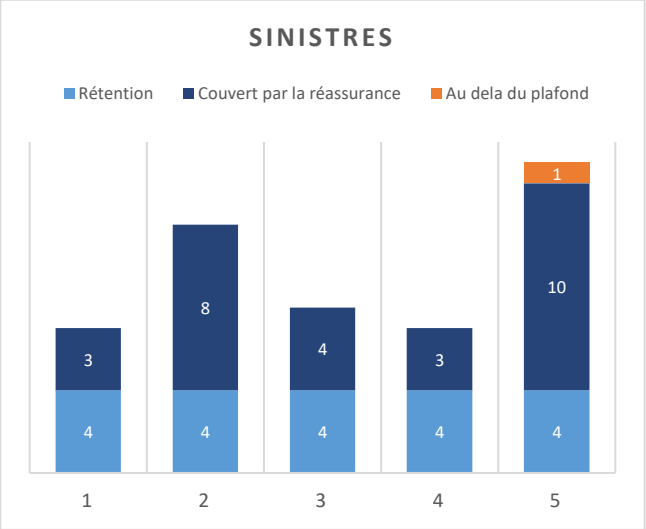


Figure 5 : Répartition des sinistres pour un 10 XS 4

- L'excédent de perte (Stop Loss)

L'excédent de perte intervient lorsque la cédante cherche à se prémunir contre les mauvais résultats sur une période donnée quelle qu'en soit la cause : fréquence ou montant des sinistres. Pour cela nous ne nous intéressons non plus aux montants des sinistres, mais aux résultats eux-mêmes. Le réassureur s'engage à protéger à concurrence d'un montant maximum, le montant dépassant le seuil financier au-delà duquel l'assureur est obligatoirement en perte. Généralement, la priorité et la portée de ce type de traité ne sont pas exprimées sous la forme d'un montant mais d'un pourcentage du S/P ou des sinistres sur capitaux.

Nous adoptons la notation suivante pour ce traité $Portée\ SL\ Priorité$ La sinistralité à charge du réassureur est établie à posteriori ; et dans le cas où le contrat de réassurance dépend du S/P constaté sur la période, elle est donnée par :

$$S^{réa} = max\left(min\left(\frac{S}{P} - Priorité; Portée\right); 0\right) \times P$$

Où P représente le total des primes perçues.

Exemple : Prenons un traité 40% SL 90%, et 5 cas de figure avec des résultats différents de S/P

Police	S/P	Cédé en réassurance (en % des primes)	À charge de l'assureur (en % des primes)
1	70%	0%	70%
2	80%	0%	80%
3	85%	0%	85%
4	110%	20%	90%
5	140%	40%	100%

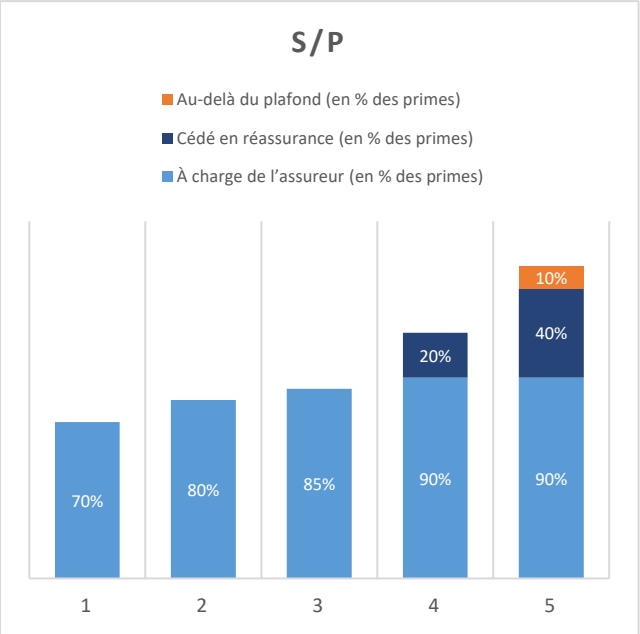


Figure 6 : Répartition du S/P pour un 40% SL 90%

III. Provision technique de L'assurance dépendance :

On rappelle que le sujet de ce mémoire est de proposer un modèle de provisionnement pour la garantie dépendance que les compagnies d'assurance et de réassurance doivent constituer pour faire face à leurs engagements.

Les provisions techniques d'une société d'assurance ou de réassurance représentent la réserve règlementaire et prudentielle qui doit être constituée pour faire face aux engagements pris envers les assurés par le biais des contrats de souscription. Ainsi, pour une société de réassurance, les provisions techniques correspondent à la réserve constituée pour faire face au règlement des engagements liés aux opérations de réassurance pris envers les compagnies d'assurance cédantes.

Les provisions techniques sont inscrites au passif du bilan des compagnies d'(ré)assurance ; elles représentent une part très importante de celui-ci environ 60 % à 80 %. Leur évaluation annuelle, voire trimestrielle a donc un impact important sur le résultat de la compagnie, ce qui en fait l'une des priorités majeures des compagnies d'(ré)assurance.

Les provisions techniques sont composées à la fois des provisions pour primes et des provisions pour sinistres, le portefeuille étant actuellement en arrêt de souscription depuis 2018 (run-off [1]). Il s'agit donc de couvrir, grâce aux provisions techniques pour sinistres, la liquidation complète du portefeuille dans le temps.

L'assurance dépendance prévaut le régime de survenance : chaque assureur paye aux assurés les sinistres qu'il couvre au moment de la survenance de sinistre. En revanche, il est fréquent qu'il s'écoule plusieurs années entre la date de survenance et la date du versement des indemnités, à cet effet, les (ré)assureurs constituent des provisions afin de procéder au règlement des sinistres. Selon la directive en vigueur, La provision technique peut se scinder en 3 types de provisions, énoncées dans les trois points suivants :

- Provision mathématique de rente (PM de rente) : est la valeur estimative ayant pour objectif d'assurer l'indemnisation complète de tous les sinistres survenus et connus mais encours d'indemnisation, autrement dit c'est le montant restant à payer pour chaque sinistre déclaré non clôturé.
- Les provisions constituées au titre des sinistres survenus mais non encore déclarés au réassureur, qui sont dus soit à un retard de déclaration de la part de la cédante, soit à un retard de déclaration de la

[1] Les portefeuilles en run-off font référence aux polices d'assurance ou aux contrats de réassurance résiliés mais pour lesquels l'assureur ou le réassureur reste engagé jusqu'à la liquidation de tous les sinistres.

part de l'assuré (la cédante n'en a alors elle-même pas encore connaissance). Les provisions constituées pour ces sinistres sont appelées IBNYR (Incurred But Not Yet Reported).

- Les provisions constituées au titre des sinistres survenus et déclarés au réassureur, mais dont le montant n'est pas encore bien connu, et peut sensiblement varier selon divers critères, à la hausse comme à la baisse. Une provision doit donc être dotée pour protéger l'entreprise en cas de sous-évaluation du montant ultime de ces sinistres. Les provisions constituées pour anticiper l'évolution de ces sinistres sont appelés IBNER (Incurred But Not Enough Reported).

La somme des IBNER et des IBNYR donne les IBNR (Incurred But Not Reported), estimation de la charge des sinistres tardifs qui doit être ajoutée à la provision mathématique pour donner le montant de la provision technique.

Remarque :
En général, les méthodes de calcul s'appliquent aux IBNR et ne distinguent pas les IBNER et les IBNYR.

Il est à noter qu'en réassurance, les cédantes estiment elles-mêmes leurs provision techniques, grâce aux informations qu'elles ont des sinistres et de leurs évolutions.
En générale, seules les PM de rente sont communiquées au réassureur, et donnent une idée, mais non définitive, du coût des sinistres restants à payer, en plus de ceux qui ont été payés. Le réassureur doit donc d'une part régler la partie des sinistres payés par la cédante correspondante à sa part suivant ses engagements, et d'autre part, au vu des informations fournies par la cédante, réestimer lui-même le montant des sinistres non encore réglés. Le réassureur doit donc ajouter aux PM estimées par les cédantes, des IBNR pour anticiper les retards de remontée d'information qui peuvent amener à des augmentations considérables des montants des sinistres.
Dans le cadre de ce projet de mission que nous étions mandatés de réaliser, nous allons procéder à un calcul complet de la provision technique (Provision Mathématique de rente et IBNR) qu'on comparera ensuite avec le montant des provisions techniques que la cédante nous a communiqué.

Conclusion :
L'augmentation du nombre de personnes âgées au sein de la population au cours des dernières décennies est liée à deux facteurs : la baisse de la fécondité et la baisse de la mortalité aux âges élevés, cette dernière est due aux progrès de la médecine et à l'amélioration des conditions de vie.
Dans ce contexte les assureurs ont un rôle important à jouer en proposant des couvertures adaptées. Ils rencontrent néanmoins quelques difficultés. Le risque de dépendance reste un phénomène récent et compliqué à définir, ainsi l'absence de maîtrise technique du risque dépendance a conduit les organismes assureurs à recourir à la réassurance pour bénéficier de l'expertise du réassureur et atténuer les risques pris.
Après avoir rappelé ces quelques éléments de contexte, on va présenter dans le prochain chapitre le produit de dépendance. En effet, l'exercice de provisionnement nécessite une compréhension claire des garanties du contrat que nous souhaitons provisionner. Nous présenterons également les données à notre disposition.

Chapitre 2 : Présentation et description de portefeuille

Cette partie a pour objectif d'établir le cadre de notre étude. Pour cela, nous introduisons les différentes caractéristiques du traité ainsi que le risque couvert. Nous évoquerons également le travail effectué sur les données que nous exploiterons pour nos études actuarielles d'élaboration d'une table d'expérience en dépendance.
Avant toute tentative de modélisation, Nous jugeons qu'une description claire de la garantie qu'on provisionne est essentielle, c'est pourquoi ce chapitre a pour objectif d'introduire les différentes caractéristiques du produit et du traité de réassurance.

I. Description du portefeuille étudié et du traité de réassurance :

1. Description du portefeuille

Le produit d'assurance dépendance commercialisé par la cédante se présente comme suit :

Thème	Condition
Type d'assurance	Soins de longue durée (LTC)
Description de l'assurance	Indemnisation mensuelle au titre des frais de séjour de l'assuré dans une institution pour dépendants ou dédommagement mensuel fixe pour les assurés dépendants vivant chez eux.
Nombre de mois durant lesquels l'indemnisation sera versée	60 mois au maximum.
Délais de carence	30 jours : pendant cette période l'assuré ne pourra prétendre à aucune indemnisation de sinistres. De même pendant cette période d'attente l'assuré devra acquitter sa cotisation.
Montant de la franchise	Aucune franchise
Définition du cas de sinistre	Le mauvais état de santé et le fonctionnement diminué de l'Assuré du fait d'une maladie, d'un accident ou d'un problème de santé en raison desquels il ne peut effectuer par lui-même une partie essentielle (au moins 50% de l'opération) d'au moins 3 sur 6 des opérations quotidiennes précisées dans la définition du cas de sinistre, ou le mauvais état de santé et le fonctionnement diminué de l'Assuré du fait d'un « épuisement psychique » (selon la définition du cas de sinistre) prescrit par un médecin spécialisé.
Type d'indemnités d'assurance	Pour un bénéficiaire en hospitalisation de longue durée – compensation. Pour un bénéficiaire recevant des soins de longue durée à domicile – prestation de services de longue durée à domicile par le biais d'entreprises de soins de longue durée, ou compensation par indemnité de soins de longue durée mensuelle fixe pour l'emploi d'un travailleur étranger, ou compensation par indemnité mensuelle fixe de soins de longue durée.
Type de garantie	Soin Silver et soins Gold : la différence réside dans les critères d'indemnisation ainsi que le degré de dépendance.

<table>
<tr><td>Montant d’assurance</td><td>
<table>
<tr><td>Lieu de résidence de l’assuré</td><td>Silver</td><td>Gold</td></tr>
<tr><td>Indemnisation mensuelle pour les assurés résidant en institution</td><td>40% du cout réel de la prestation plafonné à 1500 $/mois</td><td>80% du cout réel de la prestation plafonné à 3000 $/mois</td></tr>
<tr><td>Indemnisation mensuelle pour les assurés résidant chez eux (dédommagement)</td><td>1000 $/mois</td><td>1200 $/mois</td></tr>
</table>
</td></tr>
<tr><td>Dépendance entre le montant d’assurance et l’âge de l’Assuré</td><td>Il n’y a aucun rapport entre le montant d’assurance et l’âge de l’Assuré au moment du cas de sinistre, ou l’âge de l’Assuré à la date d’adhésion à la police</td></tr>
<tr><td>Assuré</td><td>Adhérent, conjoint/e et les enfants de mois de 18 ans.</td></tr>
<tr><td>Exceptions à la responsabilité de l’Assureur</td><td>Cas de sinistre survenu à un enfant avant qu’il n’atteigne l’âge de 12 mois.
Cas de sinistre survenu avant la date de début d’assurance ou après la fin de la période d’assurance.
Participation à une activité illégale.
Ivresse chronique ou usage de drogues, sans prescription médicale.
Accident de la route. Le terme « accident de la route » sera interprété conformément à la Loi sur les indemnités aux victimes des accidents de la route, 1975, ou conformément à toute autre loi la remplaçant
Fission ou fusion nucléaire, pollution radioactive.
Toute maladie congénitale, pour des raisons d’hérédité ou autres, y compris une tare ou des lésions causées du fait de la grossesse ou de l’accouchement au cours duquel l’Assuré est né, à condition que la chose ait été prescrite par un diagnostic médical documenté au cours des 12 mois ayant suivi sa naissance.</td></tr>
</table>

2. Les caractéristiques du traité :

Il s'agit d'un traité en quote-part avec un taux de cession de 25%, c'est-à-dire le réassureur s'engage à prendre en charge 25% de tous les risques du portefeuille considéré moyennant le pourcentage 25% de la prime perçue pour ce portefeuille.
Comme mentionnée précédemment, ce type de couverture proportionnel est intéressant pour les cédantes disposant d'un produit complexe à caractère évolutif avec un risque de dérive important.

Résumé :

- Un traité est en Quote-Part (Part = 25%), il couvre le risque de dépendance
- Il a été souscrit en 2008 puis en run-off depuis le 31.12.2018.
- Le délai de carence est de 30 jours
- Le délai maximal de prestation est 60 mois.
- Deux types de garantie (Silver et Gold).

II. Présentation des données

La constitution de la base de données constitue en soi un travail délicat et décisif, car il conditionne la qualité et la robustesse des estimations qui seront effectuées ensuite.
Nous avons veillé durant l'élaboration de notre table d'expérience sur la pertinence et l'exactitude des données, pour y parvenir des échanges réguliers avec la cédante ont été effectué afin de comprendre les bases de sinistres qu'elle nous communique et les différents changements de statuts qui ont été opérés.
Les informations à exploiter proviennent de la base de donnes sinistres envoyées par la cédante - appelée également bordereaux-. Ces données sont nettoyées mensuellement par ''l'Inforce Management'' sur une plate-forme dédiée à ce but, nommée *Life Administration Platform (LAP)* et revue par les équipes de la comptabilité technique pour s'assurer de l'adéquation des informations comptable avec les caractéristiques du traité.
Les actuaires ont l'accès à la plateforme LAP pour extraire les différentes données envoyées par la cédante, que ça soit des données de primes ou de sinistres. Nous rappelons le lecteur que le traité de réassurance qui fait l'objet de notre étude est fermé, il s'agit d'un portefeuille en run off, par conséquent, les données qui nous intéressent seront les données sinistres.

Présentation de l'export de sinistres brut :

L'export de la base de données sinistre comprend la liste de tous les adhérents à la garantie étant ou ayant été bénéficiaire de la garantie LTC depuis l'origine du contrat jusqu'au 29/06/2021. L'extraction a été réalisée à partir de la base de données le 30/06/2021.
Chaque ligne correspond à un sinistre (et non pas un identifiant). Parfois un identifiant peut avoir plusieurs numéros de sinistres s'il y'a un changement dans le statut de sinistre (par exemple, un changement dans le type d'hospitalisation).
Au 30/06/2021, le fichier compte 31 671 sinistres/dossiers.
La liste des attributs du fichier est la suivante :

- Caractéristiques individuelles des assurés
  - Le sexe (M pour homme et F pour femme)
  - La date de naissance,
  - La date de décès si l'assuré décède
- Données concernant le contrat
  - Le numéro du contrat,
  - Le numéro de sinistre
  - Le type de garantie
  - La date du fait générateur
  - La date de situation du contrat (date de dernière mise à jour),
  - La date de souscription ou d'effet,

- La nombre de jour indemnisé au titre de dépendance.
- Le délai de carence
- Le montant déjà réglé au titre du sinistre
- L'état du sinistre (ouvert, clos, en attente, annulé),
- Le type d'hospitalisation (à domicile ou dans un centre d'hospitalisation)

1. Retraitement de la base de données

- Le traitement des données manquantes

Les variables concernant la date de naissance, les dates d'effet d'entrée et de sortie du contrat, les dates de début et de fin de garantie ne sont pas toujours renseignées dans la base de données. Les retraitements effectués sont les suivants :

- La date de naissance étant indispensable à la construction des lois, les contrats pour lesquels il manque cette information sont supprimés de la base.
- Si la date d'effet de l'entrée du contrat est manquante alors la date de souscription est retenue.

- Retraitement des données aberrantes

Les retraitements des données aberrantes sont les suivants :

- Si la date de décès précède la date d'effet du contrat alors le contrat est supprimé.
- Si la date de décès survient avant la date de sortie alors la date de sortie prend la valeur de la date de décès.

Des échanges réguliers avec la cédante ont abouti à construire une base fiable dont plus de 95% des données sont correctement renseignées.

- Contrats à ôter du portefeuille

Dans un premier temps, nous retirons de la base de données les contrats qui ne sont jamais réellement entrés dans le portefeuille, c'est-à-dire qu'ils ont été enregistrés mais qui n'ont pas pris effet.
La table ci-dessous représente les différents statuts pour un contrat donnée vu à la date du 30/06/2021.

Statut du contrat	Nombre de contrats
Benefit End	2 594
Canceled	1 653
Closed	1 176
Deceased	10 909
In Payment	5 179
OS	303
Refused	9 568
Settelment	289
Total	**31 671**

Tableau 5 : Répartition du nombre de contrat par statut du contrat ''Silver''

Les statuts ''refusé'' et ''annulé'' seront écartés de notre base de données, ainsi le nombre de contrats restant dans le portefeuille est de 20 450.

Contrats en attente ''OS (Outstanding)'':

Les contrats avec le statut ''en attente'', ce sont les contrats pour lesquels la cédante est en attente de complément d'information, ces derniers seront retenus dans notre analyse avec une probabilité de passage de statut en attente vers le statut en payement ''In payment''. Le présent mémoire ne fera pas l'objet d'une étude approfondie pour estimer la probabilité de conversion de ces contrats, nous allons nous contenter d'une estimation empirique pour déterminer cette probabilité.

2. Statistique descriptive du produit

Type de garantie.

Notre portefeuille de dépendance est constitué de 67% d'assurés ayant souscrit à un contrat avec la garantie Silver (dépendance totale seule) et de 33% d'assuré ayant souscrit la garantie Gold (dépendance totale et partielle), cette garantie Gold séduit une partie moins importante du portefeuille. Cela peut s'expliquer par la volonté des assurés de se couvrir uniquement pour les risques les plus graves.

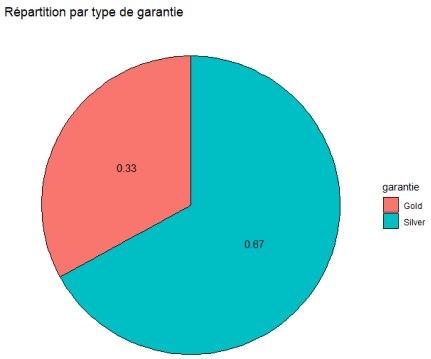


Figure 7 : Ventilation des assurés par type de produit souscrit

Étant donné que le nombre de contrats souscrits pour une couverture en dépendance totale et partielle ''Gold'' n'est pas très important par rapport à une couverture en dépendance total, le présent mémoire s'intéressera uniquement aux entrées en dépendance pour le produit ''Silver'', par conséquent les analyses qui vont suivre seront focalisées sur ce dernier produit.

Répartition par âge et par sexe

Afin d'analyser le risque de dépendance, des études statistiques descriptives sont préalablement effectuées sur le produit étudié (Silver).

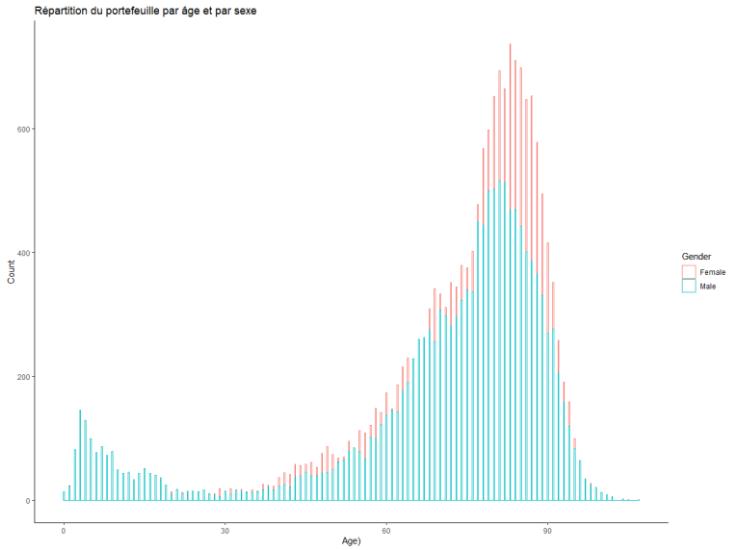


Figure 8 : Pyramide des âges des assurés du portefeuille étudié par sexe

Le portefeuille de dépendants total ''Silver'' est principalement constitué d'assurés entre 61 et 90 ans, avec un âge moyen de 73 ans, 70 ans pour les hommes et 75 ans pour les femmes. Le plus jeune a 0 an (les enfants des assurés sont automatiquement couverts par cette assurance d'où les âges proches de 0) et le plus âgé a 106 ans. On observe, entre autres, une augmentation nette de nombre de dépendants à partir de 75 ans et puis une

décroissance après l'âge de 85 ans. Nous constatons également que le poids des femmes (53%) est supérieur à celui des hommes (47%) notamment pour les âges au-dessus de 28 ans, cette tendance est inversée pour les âges inférieur à 28 ans.
La table ci-dessous affiche la répartition et âge moyen des assurés par sexe et degré de dépendance

	Produit Silver		
	Femme en dépendance	Homme en dépendance	Total
Age moyen	75,37	70,17	72,93
Effectif	10 857	9 593	20 450
Proportion	53,1%	46,9%	100%

Tableau 6 : Répartition et âge moyen des assurés par sexe pour le produit ''Silver''

Type d'indemnités :

La prestation peut être de deux types : une compensation financière pour une personne hospitalisée et qui ne vit plus à domicile, tandis qu'une personne dépendante qui vit chez elle peut recevoir la prestation d'un infirmier qui vient dispenser les soins chez elle ou une allocation correspondant au coût d'un infirmier pour un certain nombre d'heures de prestation. En s'intéressant à la répartition selon le lieu de soin, nous observons une proportion des assurés recevant des soins à domicile plus élevé que la proportion de des assuré recevant des soins dans une institution de soin (70,8% contre 29,2%).

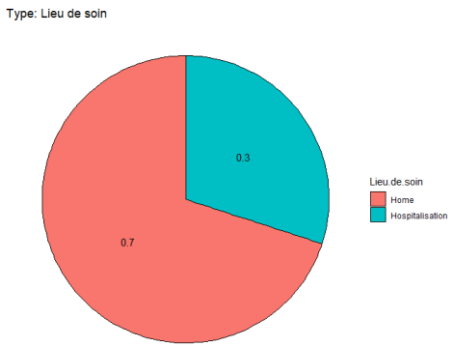


Figure 9 : Ventilation des assurés par lieu de soin

Survenance des sinistres

Un autre regard doit également être porté sur le portefeuille dépendance, celui de la survenance des sinistres ; c'est-à-dire, l'entrée des assurés dans l'état de dépendance.
Nous allons baser notre étude sur 20 450 sinistres avec un historique des changements d'état depuis 2008 et un arrêt de l'observation en 2018.

Sinistres survenus par année de survenance :

En suivant l'évolution de l'entrée en dépendance des assurés, une accélération du nombre de sinistres est observée.

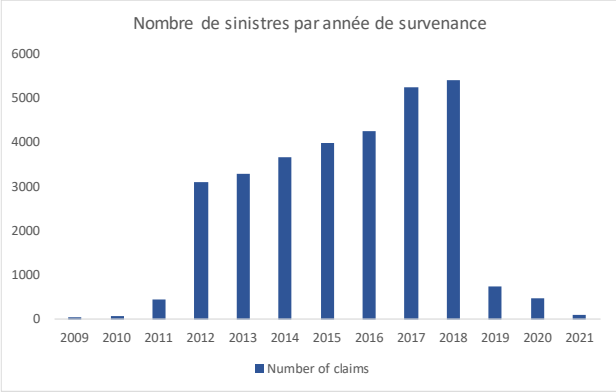


Figure 10 : Répartition du nombre de sinistres par année de survenance

Le nombre de sinistres déclarés a augmenté régulièrement depuis 2012, atteignant un pic en 2018, cette augmentation concerne les deux sexes, comme le montre le graphique ci-dessous.

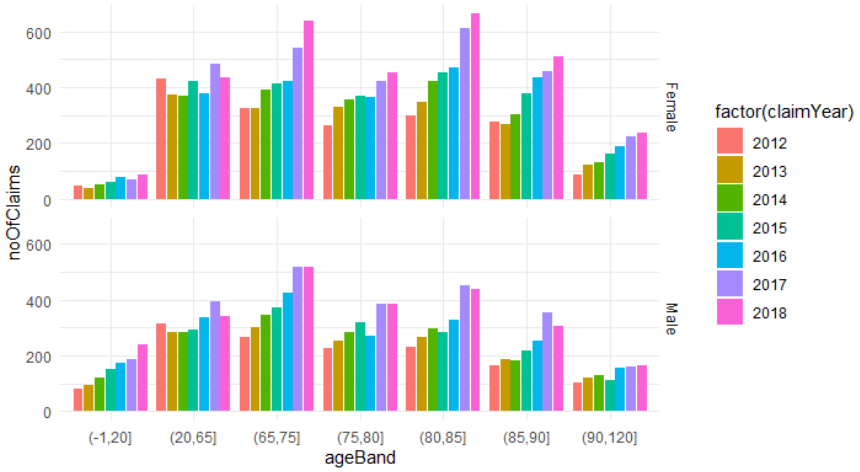


Figure 11 : Répartition du nombre de sinistres par année de survenance, par sexe et par tranche d'âge

On remarque que le nombre de sinistres survenus en 2018 est légèrement supérieure à celui enregistré en 2017, nous jugeons que cette augmentation sera plus accentuée dans le futur dû au retard de réception des données de la part de la cédant (*reporting lag*). En effet, une période de 6 mois minimum s'écoule entre la survenance du sinistre et la remontée de l'information à la compagnie de réassurance.

Nous remarquons que le nombre de sinistres est en hausse notamment à partir de 2016, cette augmentation s'explique par trois raisons :

- Vieillissement du portefeuille des dépendants.
- Hausse de taux d'acceptation (acceptance rate) des sinistres suite à un contrôle effectué par le régulateur local.
- Augmentation du nombre des sinistres pour les enfants des assurés (notamment les troubles du neurodéveloppement tel que l'autisme)

Nous avons volontairement affiché des sinistres survenus postérieurement à 2018, année de rentrer en run-off, ces sinistres sont pour le moment exclus de notre analyse, ainsi des échanges avec la cédante sont en cours pour comprendre leur nature.

Le graphique ci-après représente le taux de refus des sinistres par la cédante, ainsi nous constatons une forte diminution des dossiers refusés pour les sinistres postérieurs à 2015.

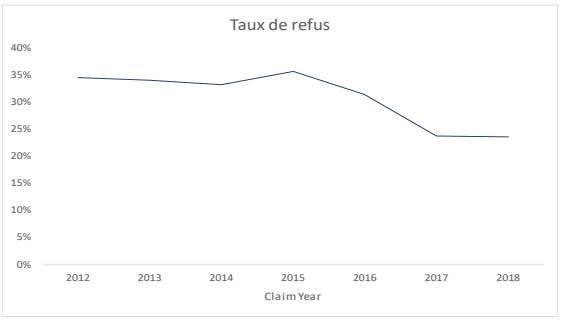


Figure 12 : Graphique d'évolution du taux de refus par année de survenance.

On considère un enfant une personne âgée de moins 14 ans. Le graphique ci-dessous montre l'évolution du nombre de sinistre pour cette catégorie d'assurés, ceci peut être expliqué par la baisse du taux de refus des sinistres. On note que cette observation concerne les deux types de garantie, ''Silver'' et ''Gold'', avec plus d'intensité pour la garantie silver à cause de l'effet volume.

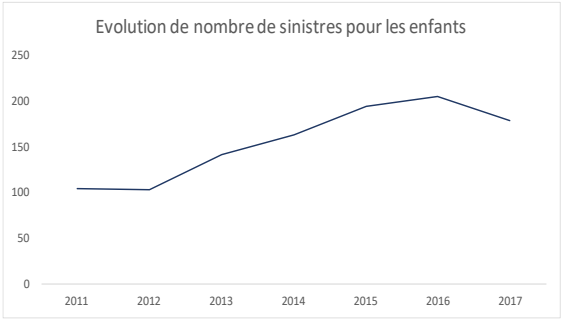


Figure 13 : Graphique d'évolution du nombre de sinistres pour les enfants

Nous allons se concentrer sur les statuts des sinistres ouverts qui font l'objet d'une provision, soit un sinistre en statut de paiement ''In payment'' ou en statut suspens "'OS'':

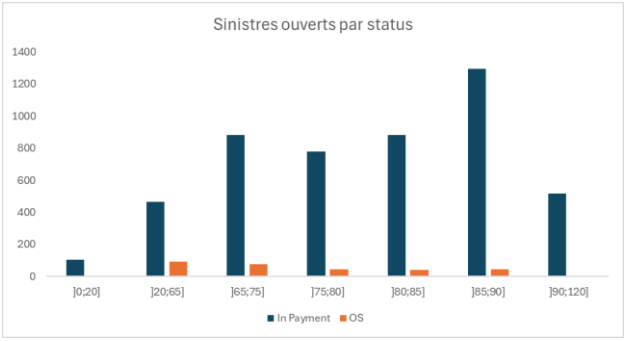


Figure 14 : Nombre de sinistre ouvert par statut et par tranche d'âge

Nous constatons que les sinistres ouverts en statut de paiement sont concentrés sur la tranche d'âge]85 ;90]. En ce qui concerne les sinistres en statut 'OS[2]', leur nombre reste relativement faible par rapport aux sinistres en cours de paiement dans chaque intervalle, ce qui montre que la plupart des sinistres sont activement traités ou déjà réglés. Par ailleurs, la distribution du nombre de sinistres est globalement uniforme

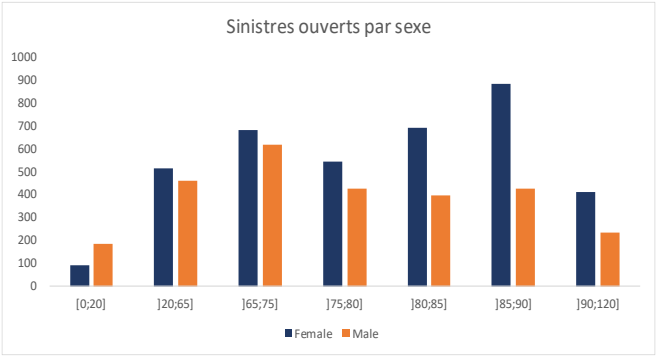


Figure 15 : Nombre de sinistre ouvert par sexe et par tranche d'âge

Sur le portefeuille d'assurés avec le statut sinistre ouvert, le poids des femmes est plus élevé que celui des hommes pour toutes les tranches d'âge à l'exception de la tranche]0 ;20 ans]. Cela est cohérent avec les observations relevées plus tôt.

[2] Statut 'OS' : Le sinistre a été déclaré à l'entreprise et enregistré dans son système. L'entreprise est en train de collecter des informations et n'a pas encore confirmé ou rejeté le sinistre

Conclusion :

Ce travail est effectué dans le cadre de partenariat auprès d'un grand réassureur international qui, dans le cadre des clôtures trimestrielles, réalise des analyses approfondies ''deep dive analysis'' sur des portefeuilles caractérisés par des provisions matérielles.

Dans un premier temps, il est nécessaire de bien définir le périmètre de notre étude.

Bien que le réassureur partenaire ait mis à notre disposition un produit d'assurance couvrant la dépendance totale et partielle ''Gold'' et uniquement totale ''Silver'', nous nous sommes restreints à l'analyse du produit d'assurance couvrant uniquement la dépendance totale.

Le travail préliminaire d'analyse de portefeuille a permis de décrire la répartition des assurés et des sinistres selon le sexe, l'âge des assurés et les statuts des sinistres.

L'augmentation des sinistres pour la tranche d'âge enfant pourrait relever un besoin intéressant de disposer des données détaillées pour comprendre ce phénomène, parce qu'on ne dispose pas de données nous avons jugé que cette augmentation est due à la baisse du taux de refus des sinistres à la suite d'un audit réalisé par le régulateur local en 2015.

Finalement, nous avons remarqué que le risque de dépendance est plus élevé chez les femmes que chez les hommes avec une proportion de souscription plus élevée chez les femmes, qui laisse entrevoir un effet d'antisélection étant que la prime est unisexe est donc moins chère pour les femmes que pour les hommes, toutefois on peut également considérer que cette différence sous-tend une aversion au risque différente.

Chapitre 3 : Modélisation du risque de dépendance :

Il nous parait important dans un premier temps d'introduire les outils mathématiques nécessaire pour la modélisation de durée de vie et plus particulièrement leur application dans l'élaboration des tables d'expérience en dépendance.

I. La théorie des modèles de durée

1. Description de la survie

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenance d'un évènement précis. L'analyse des données de survie est l'étude du délai de la survenance de l'événement étudié.
Les modèles de durée présentent certaines particularités :

- Les données de durée sont engendrées par des variables aléatoires positives ;
- La fonction de survie et le taux de hasard ont une interprétation physique naturelle dans le cadre des modèles de durée que nous détaillerons plus bas ;
- Les données de durée sont souvent incomplètes (observation tronquée et/ou censurée), l'expérience ayant souvent une durée limitée, les assurés ne sont que partiellement observables sur une durée donnée.

Supposons que la durée de survie T soit une variable positive ou nulle, et absolument continue, alors sa loi de probabilité peut être définie par l'une des six fonctions équivalentes suivantes (chacune des fonctions ci-dessous peut être obtenue à partir de l'une des autres fonctions) :

a) La fonction de survie

Soit T une variable aléatoire positive d'une durée de maintien en état de dépendance, sa fonction de survie est, pour t fixé :

$$S(t) = P(T > t), t \geq 0$$

Elle est la probabilité que la sortie de l'état de dépendant de l'individu se produise après l'instant t. Avant l'instant t l'individu est en état de dépendance.

b) Fonction de répartition

La fonction de répartition représente, pour t fixé, la probabilité que la sortie de l'état dépendant de l'individu se produise entre 0 et t c'est-à-dire :

$$F(t) = P(T \leq t) = 1 - S(t), t \geq 0$$

Remarque :
Il est arbitraire de décider que $S(t) = P(T > t)$ ou $S(t) = P(T \geq t)$. Cela n'a aucune importance quand la loi T est continue. Dans le cas où F a des sauts (quand le temps est discret, compté en mois ou en semaine) nous retrouvons les notations suivantes : $F^-(t) = P(T < t)$ et $F^+(t) = P(T \leq t)$
Où F^- est la limite à gauche et F^+ la limite à droite (définition et notations sont identiques pour la fonction S), avec $F^- \leq F^+$ et $S^- \geq S^+$.
Afin de ne pas alourdir les notations, nous omettons cette distinction pour le cas discret.

c) La fonction de densité

C'est la fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$

$$F(t) = \int_0^t f(u)du$$

Si la fonction de répartition F admet une dérivée au point t alors :

$$f(t) = \lim_{h \to 0^+} \frac{P(t \leq T < t + h)}{h} = F'(t) = -S'(t)$$

Pour t fixé, la densité de probabilité représente la probabilité d'être dépendant dans un petit intervalle de temps après l'instant t, on peut également la voir comme la probabilité de sortir de l'état de dépendance sur un intervalle infinitésimal autour de t.

d) La fonction de survie conditionnelle :

La probabilité que l'individu quitte l'état entre les instants t et $t + u$, sachant son maintien dans l'état à l'instant t, est donnée par le Théorème de Bayes, tel que :

$$P(t < T \leq t + u/T > t) = \frac{P(t \leq T < t + u)}{P(T > t)} = 1 - \frac{S(u + t)}{S(t)}$$

La fonction de survie conditionnelle représente la probabilité que le temps passé dans l'état dépendant d'un individu dépasse l'instant $t + u$, sachant qu'il était dépendant à l'instant t.

$$S(t/u) = P(T > u + t/T > t) = \frac{P(T > u + t)}{P(T > t)} = \frac{S(u + t)}{S(t)}$$

e) Le taux de hasard :

Le taux d'hasard λ d'une variable aléatoire T est souvent utilisé pour spécifier un modèle de durée car il a une interprétation physique naturelle, on peut l'assimiler à la probabilité de quitter l'état de dépendance dans un petit intervalle de temps après l'instant t, conditionnellement au fait d'être resté dépendant auparavant. De grandes valeurs de cette fonction indiquent les assurés pour lesquels la probabilité de sortie de l'état de dépendance est importante.

On peut définir cette relation dans le cas continu par :

$$\forall t \geq 0, \lambda(t) = \lim_{h \to 0^+} \frac{P(t \leq T < t + h/T \geq t)}{h}$$

Dans le cas discret, on suppose que T prend ses valeurs dans un ensemble dénombrable $\tau = \{t_1, t_2, \dots, t_n, \dots\}$. On définit le taux de hasard λ par :

$$\forall t, \lambda(t) = P(T = t/T \geq t) = \frac{P(T = t)}{P(T \geq t)}$$

On peut réécrire le taux de hasard dans le cas continu comme :

$$\forall t \in \mathbb{R}+, \lambda(t) = \lim_{h \to 0^+} \frac{1}{h} \times \frac{P(t \leq T < t + h \cap T \geq t)}{P(T \geq t)} = \lim_{h \to 0^+} \frac{1}{h} \times \frac{P(t \leq T < t + h)}{P(T \geq t)} = \frac{f(t)}{S(t)}$$

$$\lambda(t) = -ln\big(S(t)\big)'$$

f) Le taux de hasard cumulé

Le taux de hasard cumulé Λ d'une variable aléatoire T est défini dans le cas continu comme l'intégrale du taux de hasard λ

$$\Lambda(t) = \int_0^t \lambda(u)du = -ln\big(S(t)\big)$$

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé (ou du risque instantané) :

$$S(t) = exp\big(-\Lambda(t)\big) = exp\left(-\int_0^t \lambda(u)du\right)$$

Dans le cas discret, le taux de hasard cumulé Λ vaut :

$$\Lambda = \sum_{i | t_i \leq t} \lambda(t)$$

L'intérêt de présenter les modèles de durée c'est de pouvoir modéliser la durée de maintien d'un assuré dans un état définit, ce qui permet de garantir le paiement probable des rentes mensuelles et de constituer de provision mathématique de dépendance.

Afin de modéliser la variable de durée T, nous pouvons avoir recours à différents types de modèles, à savoir :

- Les modèles paramétriques : on émet l'hypothèse que les temps de survie de l'échantillon étudié sont distribués selon une loi connue (loi exponentielle, Weibull, Pareto, etc.)
- Les modèles non-paramétriques : aucune hypothèse a priori n'est faite sur l'allure générale de la fonction de survie (estimateur de Kaplan Meier, Nelson-Aalen)
- Les modèles semi-paramétriques : ils sont basés sur un modèle de base adapté à la cohorte puis différenciés pour chaque sous-groupe. Ils s'avèrent plus adaptés en cas d'échantillons hétérogènes (modèle de Cox).

Le choix du modèle est orienté par la qualité et la quantité des données à disposition.

Remarque [3]:

Dans le cadre des modèles bidimensionnels, divers travaux démontrent que la méthode la plus robuste pour produire les taux bruts est l'estimateur de Kaplan-Meier pour chaque âge à l'entrée, sous réserve de disposer de données en quantité suffisante. En effet, la perte d'information que représente la non prise en compte de la loi conjointe selon les deux dimensions du problème est faible et peu pénalisante en pratique. Par ailleurs, des travaux ont généralisé l'estimateur de Kaplan-Meier (Adjusted Kaplan-Meier Estimator) en dimension deux en présence d'hétérogénéité (cf. XIE et LUI [2000]).

Une adaptation des modèles Lee-Carter ou de ses variantes de type log-Poisson, développés dans le cadre de la construction de tables de mortalité prospectives, aurait pu être envisagée mais les travaux de LELIEUR V. et PLANCHET F. ont démontré que lorsque l'on sort du contexte de très grands échantillons dans lesquels ces modèles ont été développés, une instabilité des coefficients liée aux fluctuations d'échantillonnage apparaît et rend ces modèles peu efficients. Si cette difficulté a pu être contournée dans le cadre de l'étude du risque de mortalité, c'est ici plus complexe. En effet, le modèle de Lee-Carter repose sur l'hypothèse d'homoscédasticité des taux ce qui constitue une contrainte du modèle forte et peu réaliste étant donné que la variance des taux croît avec l'âge du fait de la diminution des effectifs sous risque. Ainsi nous avons choisi de retenir un modèle non paramétrique à savoir l'estimateur de Kaplan-Meier ce qui semble être le plus naturel et le plus efficace.

→ Le but est de construire des tables de maintien par âge, sexe (Homme/Femme) ; on parle de l'estimateur de Kaplan-Meier avec une stratification par cohorte.

→ La méthodologie : Nous calculons l'estimateur non paramétrique de Kaplan Meier. Comme notre base de données est volumineuse, la convergence de cet estimateur vers la fonction de survie théorique est bonne. Nous rappelons que :

$Sup_{t<\tau_L}\left|\widehat{S_n}(t) - S(t)\right| \xrightarrow[n\to+\infty]{} 0$ Où τ_L d´esigne par exemple la date limite des observations de X

Dans la deuxième section nous allons modéliser le taux de maintien à l'aide du modèle semi-paramétrique de Cox.

2. Spécificité de l'analyse de la survie

La période d'observation nous conduit à définir des notions primordiales des modèles de durées : la censure et la troncature

- **<u>Censure et troncature :</u>**

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet les données sont souvent partiellement recueillies notamment à cause des processus de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer des réalisations indépendantes et identiquement distribuées de durées T, on observe la réalisation de cette variable T soumise à diverses perturbations.

Dans le cadre de l'établissement d'une table d'expérience pour estimer les taux bruts, on fixe une fenêtre d'observation $[c, C]$ qui couvre plusieurs périodes (mois, années). Cependant les contrats d'assurance du portefeuille observé ne sont pas tous entièrement couverts par l'intervalle choisi. Ainsi la non-prise en considération doit être tenue compte dans l'élaboration de la table d'expérience. On distingue deux phénomènes qui peuvent affecter nos observations :

- On dit qu'il y a une censure si la variable aléatoire T_x désignant la durée de survie à l'âge x n'est pas observable au-dessus d'une certaine période C. Alors on observe seulement l'incidence s'il a eu lieu avant C. La variable aléatoire Z_x désigne la durée de vie future tenant compte de la fenêtre d'observations, c'est-à-dire $Z_x = T_x \wedge C$.
- On parle de troncature si la variable aléatoire T_x n'est pas observable en-dessous d'une certaine période c. Contrairement à la censure, dans ce cas on est conscient de l'existence de l'information par contre on n'arrive pas à mesurer sa valeur.

[3] https://www.ressources-actuarielles.net/EXT%5CISFA%5Cfp-isfa.nsf/0/1430AD6748CE3AFFC1256F130067B88E/$FILE/Seance7.pdf?OpenElement

Les deux phénomènes sont présentés sur la figure ci-dessous. Il est sous-entendu que les variables qui ne sont pas observables en dehors de l'intervalle $[c, C]$, elles ne sont pas prises en compte dans l'étude.

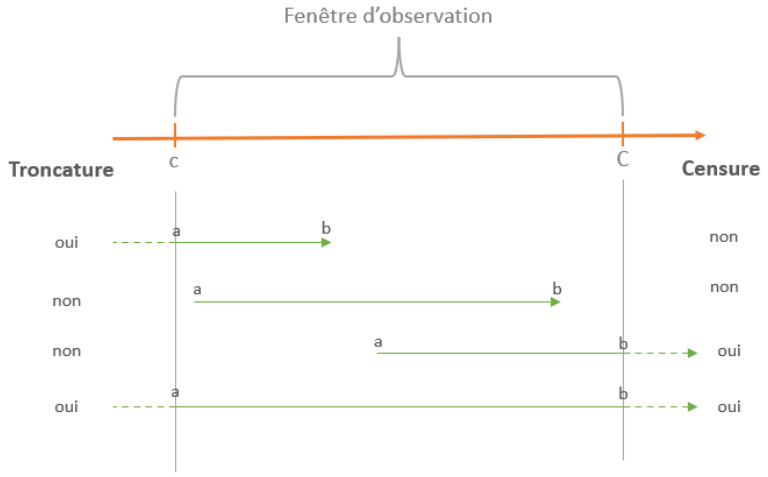


Figure 16 : Présentation des troncatures et censures

- **<u>Application à notre jeu de donnée :</u>**

Dans notre étude, les observations pour lesquelles l'occurrence de l'événement (entrée en dépendance ou sortie de la dépendance) se produit avant le début de l'observation ne sont pas retenues, c'est de la troncature gauche. De plus, on retient comme fin d'observation l'occurrence de l'événement s'il se produit dans la fenêtre d'observation, sinon, c'est de la censure droite.

Nos données étant arrêtées au 29/06/2021, donc on considère comme des censures les sinistres dont la date de sortie est postérieure au 29/06/2021.

En conclusion notre base de données doit inclure trois informations principales : la date d'entrée en dépendance, la date de sortie de dépendance et l'information sur la censure. La différence entre la date d'entrée et de sortie de dépendance permet d'obtenir la durée de maintien en dépendance.

II. Estimateur de Kaplan-Meier

1. Construction de l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier découle de l'idée suivante : être encore en état de dépendance après un instant t, c'est être en état de dépendance juste avant cet instant t, c'est-à-dire, si $t'' < t' < t$

$$P(T_x > t) = P(T_x > t', T_x > t) = P(T_x > t/T_x > t') \times P(T_x > t')$$
$$= P(T_x > t/T_x > t') \times P(T_x > t'/T_x > t'') \times P(T_x > t'')$$

En considérant les temps d'événements distincts $T_{(i)}$ $(i = 1, \dots, n)$ rangés par ordre croissant, on obtient :

$$P\left(T_x > T_{(j)}\right) = \prod_{k=1}^{j} P\left(T_x > T_{(k)} \,/\, T_x > T_{(k-1)}\right)$$

Avec $T_{(0)} = 0$. Considérons les notations suivantes :

- $d_{x,i}$: le nombre de sorties ayant lieu à l'instant $T_{x,i}$ pour l'âge à la survenance x ;
- $n_{x,i}$: l'effectif sous risque à l'âge à la survenance x avant $T_{x,i}$.

On note la probabilité p_i de sortir de l'état de dépendance dans l'intervalle $]T_{(i-1)}\ ;\ T_{(i)}]$ sachant que l'on était dépendant en $T_{(i-1)}$, i.e. $p_{x,i} = P\left(T_x \leq T_{(i)}/T_x > T_{(i-1)}\right)$ peut être estimé par :

$$\widehat{p_{x,i}} = \frac{d_{x,i}}{n_{x,i}}$$

En chaque instant $T_{x,i}$, on observe, en l'absence d'ex-æquo, si l'événement réalisé est une sortie ou une censure. La nature de l'événement est décrite par la variable $D_{x,i}$

$$D_{x,i} = \begin{cases} 1\ si\ T_{x,i} \leq C \\ 0\ si\ T_{x,i} > C \end{cases}$$

On obtient alors l'estimateur de Kaplan-Meier :

$$\widehat{S(t)} = \prod_{\substack{i=1,,,n \\ T_i \leq t}} \left(1 - \frac{D_{x,i}}{n_{x,i}}\right) = \prod_{\substack{i=1,,,n \\ T_i \leq t}} \left(1 - \frac{1}{n_{x,i} - i + 1}\right)^{D_{x,i}}$$

Avec $n_{x,i} = n_{x,i-1} - d_{x,i-1} - c_{x,i-1} + t_{x,i-1}$ où $c_{x,i-1}$ est le nombre de personnes censurées entre $]T_{(i-1)}\ ;\ T_{(i)}]$ $t_{x,i-1}$ est l'effectif des personnes tronquées de même période

En pratique cependant nous sommes confrontés à la présence d'ex-æquo :

- si ce sont des événements de nature différente, on considère que les observations non censurées ont lieu avant les censurées.
- si il y a plusieurs sorties au même temps $T_{x,i}$, alors $d_{x,i} > 1$ et on a

$$\widehat{S(t)} = \prod_{\substack{i=1,,,n \\ T_i \leq t}} \left(1 - \frac{d_{x,i}}{n_{x,i}}\right)$$

A partir de l'estimateur de la fonction de survie $\widehat{S(t)}$, on détermine la probabilité de maintien en état de dépendance pour chaque âge à la survenance avec un pas temporel :

$$\widehat{q_{x,t}} = \frac{\widehat{S(t+1)}}{\widehat{S(t)}}$$

Étant donné qu'estimer la probabilité de sortie du risque est équivalent à estimer la probabilité de maintien, on estime la probabilité de sortie de l'état de dépendance par

$$\widehat{p_{x,t}} = 1 - \frac{\widehat{S(t+1)}}{\widehat{S(t)}} = 1 - \widehat{q_{x,t}}$$

On rappel toujours que le terme de durée désigne le temps écoulé jusqu'à la sortie de l'état de dépendance (décès, guérison, fin de prestation).

Intervalles de confiance

Pour des raisons de simplification des notations mathématiques nous omettons le paramètre âge x.

Afin de construire l'intervalle de confiance de cet estimateur, on rapproche sa variance à l'aide de l'estimateur de Greenwood. On pose les hypothèses suivantes :

- Les variables $ln(1-qi)$ sont indépendantes deux à deux ;
- $n_i \times (1-q_i)$ suit une loi binomiale $B(n_i; 1-q_i)$

L'expression de l'estimateur de la fonction de survie nous permet d'écrire :

$$ln\left(\hat{S}(t)\right) = \sum_{\substack{i=1,,,n \\ T_i \leq t}} ln\left(\left(1-\frac{d_i}{n_i}\right)\right)$$

La variance de $ln\left(\hat{S}(t)\right)$peut être rapprochée à l'aide de la méthode delta

$$Var\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{\substack{i=1,,,n \\ T_i \leq t}} \frac{\frac{d_i}{n_i}}{n_i\left(1-\frac{d_i}{n_i}\right)} = \hat{S}(t)^2 \sum_{\substack{i=1,,,n \\ T_i \leq t}} \frac{d_i}{n_i(n_i-d_i)}$$

Cet estimateur s'appelle l'estimateur de Greenwood. En s'appuyant sur la propriété de la normalité asymptotique de l'estimateur de Kaplan-Meier, on peut déterminer les intervalles de confiance de la fonction de survie de niveau de confiance de $1-\alpha$

$$IC_\alpha = \left[\hat{S}(t)\left\{1+\Phi^{-1}\left({}^{\alpha}/_{2}\right)\sqrt{\sum_{\substack{i=1,,,n \\ T_i \leq t}} \frac{d_i}{n_i(n_i-d_i)}}\right\};\ \hat{S}(t)\left\{1+\Phi^{-1}\left(1-{}^{\alpha}/_{2}\right)\sqrt{\sum_{\substack{i=1,,,n \\ T_i \leq t}} \frac{d_i}{n_i(n_i-d_i)}}\right\}\right]$$

Où Φ^{-1} représente la fonction de répartition inverse d'une loi normale centrée réduite. Saporta (2006) indique que la normalité asymptotique de l'estimateur de Kaplan-Meier est vérifiée dans le cas où $n_i > 30$.

Finalement l'estimateur de taux de sortie ''termination rate'' peut être déduit à partir de l'intervalle de confiance de la fonction de survie via la relation suivante : $\widehat{p_t} = 1 - \frac{\widehat{S(t+1)}}{\widehat{S(t)}}$. Ainsi l'intervalle de confiance se présente comme suit :

$$IC_\alpha = \left[\widehat{p_t} + (1-\widehat{p_t})\Phi^{-1}\left({}^{\alpha}/_{2}\right)\sqrt{\sum_{\substack{i=1,,,n \\ T_i \leq t}} \frac{d_i}{n_i(n_i-d_i)}}; \widehat{p_t} + (1-\widehat{p_t})\Phi^{-1}\left(1-{}^{\alpha}/_{2}\right)\sqrt{\sum_{\substack{i=1,,,n \\ T_i \leq t}} \frac{d_i}{n_i(n_i-d_i)}}\right]$$

2. Estimation des taux bruts de maintien.

L'objectif de cette section est l'application de la théorie présentée dans La théorie des modèles de durée à l'aide du logiciel R.

Pour établir la loi de maintien à l'aide de l'estimateur de Kaplan-Meier, nous devons recenser pour chaque individu ayant connu un état de dépendance, sa date d'entrée ainsi que son ancienneté dans cet état. Si l'individu est toujours en dépendance à la date de l'étude, nous noterons l'ancienneté actuelle.

Une fois que le recensement terminé, les taux bruts de maintien seront estimés sur la base d'une segmentation qui permettra d'expliquer la durée de maintien en dépendance ainsi que de pérenniser notre étude de construction de la table d'expérience.

- L'âge est le premier facteur à prendre en compte dans la construction de lois biométriques. Les causes de dépendance chez les personnes âgées et chez les personnes plus jeunes sont souvent différentes et vont donc avoir un impact sur la gravité de la situation et la durée de survie en état de dépendance.
- La segmentation par sexe : semble également pertinente à prendre en compte

Pour répondre à la question : La segmentation utilisée est-elle pertinente ? nous présentons le test de log-rank

Log Rank :

Il sert à déterminer si deux distributions (dans notre cas des fonctions de survie) ou plus suivent significativement la même loi. Il permet en fait de déterminer si les fluctuations entre les courbes sont dues au hasard ou non. Ce test est particulièrement bien adapté en présence de données censurées (il est dans ce cas plus indiqué que le

test de Wilcoxon par exemple), ce qui est notre cas. Il s'applique lorsque les deux courbes de survie sont calculées par la méthode de Kaplan-Meier. De façon plus rigoureuse, le test consiste à comparer les hypothèses :

- Hypothèse $H_0 : S_A = S_B$ les fonctions de survie des deux échantillons sont égales
- Hypothèse $H_1 : S_A \neq S_B$ les fonctions de survie des deux échantillons sont différentes.

Soit $d_{i,j}$ le nombre de sorties dans le groupe j $(par\ exemple\ j \in \{1,2\})$. Le test de Log-Rank revient à construire des statistiques fondées sur les sommes des différences entre les taux de sortie théoriques et les taux de sorties observées $(d_{i,j}^{th} - d_{i,j}^{obs})$), qui sont asymptotiquement gaussiennes. Sous l'hypothèse H_0, nous utilisons les statistiques suivantes :

$$\varphi_j = \frac{\left[\sum_i \left(d_{i,j}^{th} - d_{i,j}^{obs}\right)\right]^2}{\sum_i Var\left(d_{i,j}^{obs}\right)}$$

Les φ_j suivent asymptotiquement une loi khi2 de degré de liberté $j - 1$.

Nous nous intéressons maintenant à l'analyse de la durée de maintien en l'état de dépendance des individus sinistrés.

1. Segmentation par Sexe :

Nous allons diviser notre échantillon en deux sous populations, par une segmentation relative au sexe des assurés. En effet, nous avons remarqué dans le premier chapitre que les femmes étaient plus vulnérables que les hommes vis-à-vis de la dépendance. Nous pouvons nous interroger sur l'existence d'une différence significative entre les taux bruts d'entrée en dépendance masculins et féminins et si la création de deux lois de maintien apporterait plus de précision.

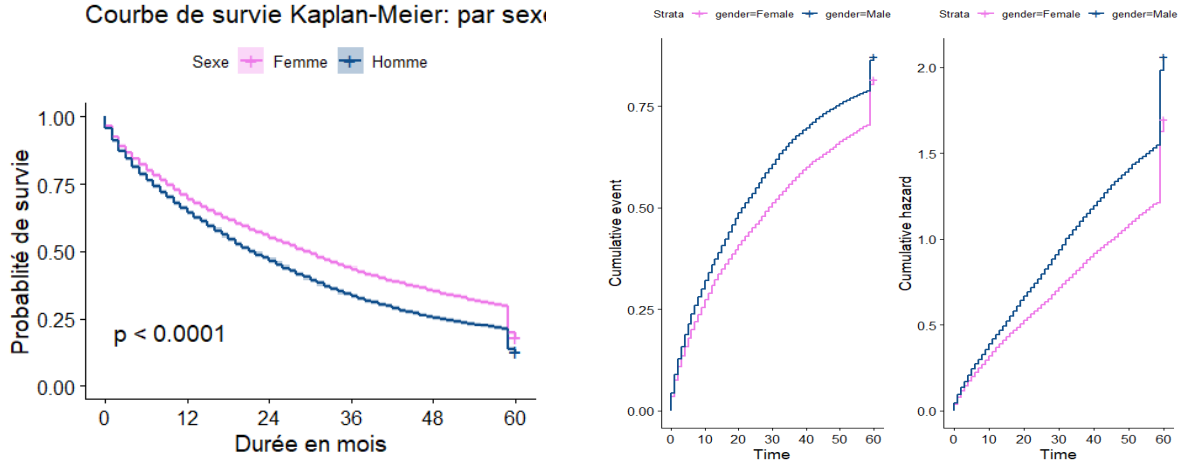


Figure 17 : Taux de servie : KM selon le sexe (Courbe à droite) Courbes de risque cumulé (à gauche)

Une première observation que l'on peut faire est que les courbes de survie homme et femme évoluent de façon quasi-proportionnelle (quelques croisements des deux courbes). On observe graphiquement que pour les durées inférieures à 12 mois le taux de maintien est presque indépendant du sexe, après cette durée les deux courbes commencent à se distinguer montrant que le taux de maintien est légèrement plus faible chez les hommes que chez les femmes.

Les deux courbes à droite, courbe de risque cumulé et courbe Le taux de hasard cumulé, montrent une augmentation du risque au cours du temps, les hommes ont tendance de se maintenir en dépendance moins que les femmes.

Justification statistique.

Afin de valider statistiquement cette hypothèse de différenciation des lois de maintien homme et femme, nous procéderons au test de Log Rank.

Le test de Log-Rank est effectué à partir de deux échantillons (Homme/Femme).

```
> surv_diff <- survdiff(Surv(claimDurationAdjusted, status) ~ gender, data = data_etude)
> surv_diff
Call:
survdiff(formula = Surv(claimDurationAdjusted, status) ~ gender,
    data = data_etude)

                N Observed Expected (O-E)^2/E (O-E)^2/V
gender=Female 10946     8946     9893      90.7       223
gender=Male    9773     8534     7587     118.3       223

 Chisq= 223  on 1 degrees of freedom, p= <2e-16
>
```

Figure 18 : Sortie R du Test de Log-Rank

Dans les sous populations, la p-value est inférieure au seuil de 5 % qui est le risque maximal toléré pour cette étude ; nous pouvons donc rejeter H_0 ce qui implique que les probabilités issues de la segmentation Hommes-Femmes sont significativement différentes. Ces résultats permettent de conclure qu'une telle segmentation apporte une information supplémentaire.

On peut alors conclure que la distinction suivant le sexe concernant la survie des personnes dépendantes est justifiée statistiquement.

Le taux après lissage :

La courbe des taux bruts présentée précédemment est peu régulière là où il y a peu d'observation. C'est pourquoi nous allons procéder au lissage de ces taux pour pouvoir construire notre table de maintien. Il existe deux types de lissage : Les lissages non paramétriques (e.g. Moyennes mobile, Witttaker- Henderson et Loess) et les lissages paramétriques (e.g. Gompertz-Makeham et Thatcher).

L'avantage des méthodes non-paramétriques est qu'elles ne nécessitent pas de faire des hypothèses à priori sur l'allure de la courbe, elles ne prennent en compte que les données disponibles tandis que les méthodes paramétriques reposent sur des hypothèses de distribution de la loi qu'il est alors nécessaire de tester sur les données.

En général, le choix du lissage doit donc dépendre d'un arbitrage entre 2 critères :

- La fidélité des taux lissés aux taux bruts ;
- La régularité des taux.

Les données que nous avons ne montrent pas de variations erratiques importantes. Ainsi, la cohérence entre les taux ajustés et les taux observés sera le critère principal pour sélectionner la méthode de lissage appropriée.

Pour vérifier la qualité de lissage, nous nous baserons sur deux critères.

1) Distance entre les taux bruts et les taux lissés

Une méthode pour tester la fidélité des taux bruts aux données consiste à calculer la distance en valeur absolue entre les taux bruts et les taux lissés.

Nous vérifions alors la qualité des ajustements par application de ce critère qui repose sur le fait que plus la somme des distances des taux bruts aux taux ajustés est proche de 0 plus la fidélité est importante.

$$\sum |\hat{t}_x^* - \hat{t}_x| \to 0$$

Avec $\hat{t}_x^*$ représente les taux bruts

2) Intervalle de confiance

Une façon classique d'effectuer la validation d'une table d'expérience une fois la courbe lissée, est de comparer les sorties modélisées des sorties brutes, par sexe et par ancienneté. Et nous construisons un intervalle de confiance autour des sorties en effectuant l'approximation usuelle par une loi normale. L'objectif étant de vérifier que les sorties effectivement observées appartiennent à l'intervalle de confiance du modèle étudié

Dans la suite de ce mémoire, nous allons effectuer 2 types de lissages des taux bruts afin d'en faciliter leur exploitation et dans le but de prendre en compte le phénomène de régularité et de rendre ainsi plus crédibles les taux trouvés

Pour la partie théorique de ces deux méthodes nous renvoyons le lecteur à ANNEXES 1

- **Lissage de Whittaker – Henderson**

Le graphique ci-dessous présente ce lissage obtenu pour femmes et les hommes. Les courbes apparaissent bien lissées et en adéquation avec la courbe de base.

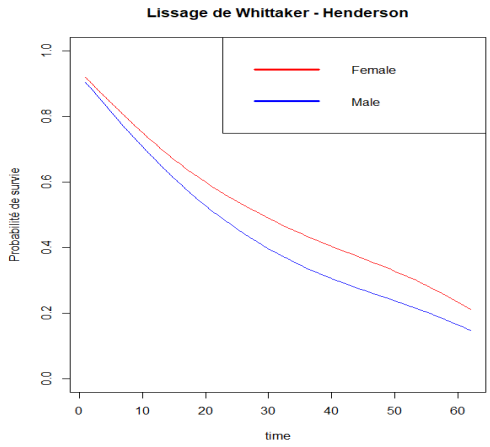


Figure 19 : La courbe de survie : Lissage par Whittaker-Henderson

Fidélité des taux lissés aux taux bruts

Distance entre les taux bruts et les taux lissés : la distance absolue entre les taux bruts et lissé est de l'ordre de 0.035

Intervalle de confiance : A la lecture du graphique on peut voir que les taux lissés sont bien présents dans l'intervalle de confiance asymptotique des taux bruts.

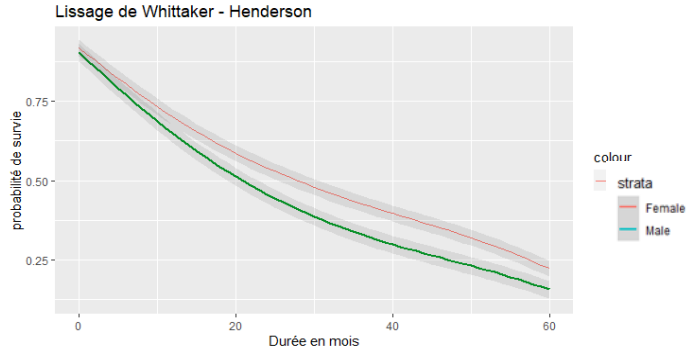


Figure 20 : La courbe de survie : Lissage par Whittaker-Henderson avec intervalle de confiance

- **Lissage LOESS**

Le graphique ci-dessous représente la courbe de survie lissé avec la régression non paramétrique ''Loess''. Comme pour le lissage effectué auparavant, nous allons procéder à tester la fidélité de cet ajustement à l'aide de la distance entre les taux bruts et les taux lissés ainsi que sur la base de l'intervalle de confiance.

Après calculer la distance absolue entre les taux lissé et les taux brute, nous avons trouvé la valeur de 0.056. En ce qui concerne l'intervalle de confiance, nous constatons que les taux lissés sont quasi- inclus dans cet intervalle.

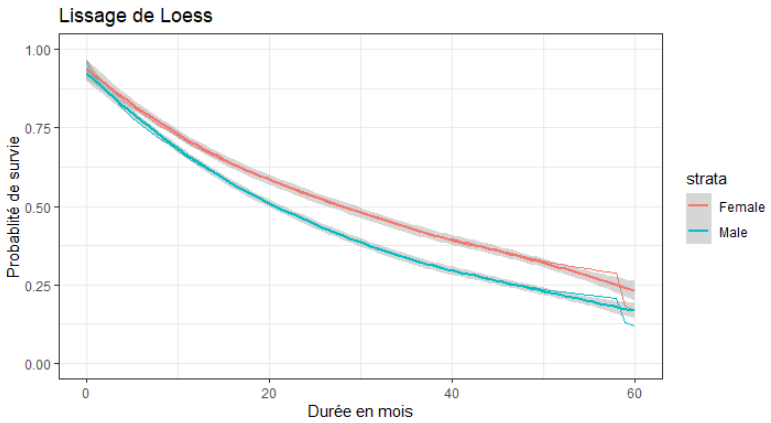


Figure 21 : La courbe de survie : Lissage par régression de LOESS

Comparaison entre les deux méthodes de lissage :

Les deux méthodes fournissent des taux lissés proches qui appartiennent à l'intervalle de confiance à 95%, néanmoins le lissage par Whittaker-Henderson permet d'obtenir des taux lissés les plus proche des taux bruts selon le critère de la distance absolue.

Nous utiliserons donc par la suite les taux de maintien lissés à l'aide de la méthode de Whittaker-Henderson.

2. Segmentation selon l'âge.

Nous avons retenu les différentes classes d'âges d'entrée en dépendance que celles utilisés lors de la tarification par la cédante, cette segmentation du portefeuille par classes d'âge est nécessaire pour estimer l'impact de l'âge d'entrée en dépendance sur le maintien de l'individu en dépendance.

Le tableau ci-dessous représente la répartition des rentiers dépendants ainsi que la moyenne d'âge pour chaque classe :

Age de l'entrée en dépendance	Nombre de rentier	Poids de la classe d'âge	Moyenne classe d'âge
(-1,20]	960	4,69%	8,25
(20,65]	3 029	14,81%	59,64
(65,75]	3 808	18,62%	70,78
(75,80]	3 349	16,38%	78,2
(80,85]	4 260	20,83%	82,98
(85,90]	3 222	15,76%	87,79
(90,120]	1 822	8,91%	93,18
Total	20 450	100%	72,93

Tableau 7 : Répartition et âge moyen des assurés par classe d'âge

Pour segmenter les lois de maintien par âge, on doit faire attention à ne pas avoir des écarts trop importants entre les tailles des différents groupes de segmentation, ainsi nous constatons que la répartition est quasi-homogène, exception faite pour la tranche d'âge minimale (-1, 20] et maximale (90,120].

Face à cette répartition, ci-dessous les lois de maintien en dépendance par tranche d'âge d'entrée en dépendance :

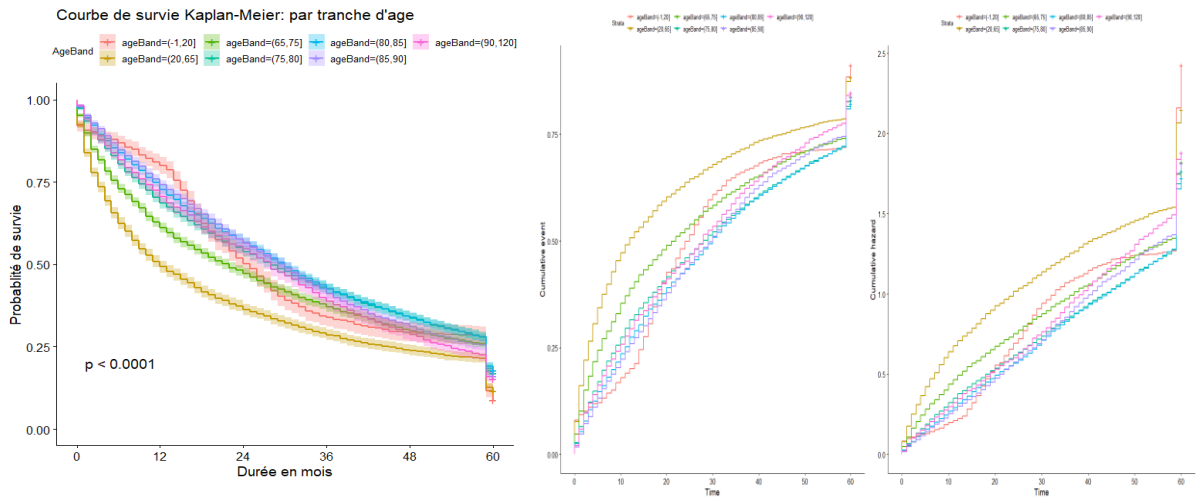


Figure 22 : Taux de servie : KM selon les classes d'âge (Courbe à droite) Courbes de risque cumulé (à gauche)

L'âge semble nous apporter une information complémentaire quant au maintien en dépendance.

On observe que la vitesse de sortie de l'état de dépendance décroit avec l'âge, à l'exception des âges très jeune (-1,20] où le taux de maintien est fort durant les premiers 11 mois (courbe concave), ensuite la courbe s'inverse entre les durée 11 mois et 24 mois, après 24 mois la courbe décroit rapidement, ce mouvement dans la courbe de survie de cette tranche peut être expliqué par le manque de donnée, et/ou les personnes entrées dans cet état à de jeunes âges ont un important risque de dépendance, causé par le motif d'entrée dans cet état (en général c'est des maladies enfantines à risque grave).

On remarque ainsi que la classe d'âge (20-65] sortent plus rapidement que le reste des classes d'âge, cette observation est assez cohérente si on considère que les assurés de cette tranche jouissent d'un état de santé meilleurs que le reste (classe des assurés jeune et senior), suivi de la tranche d'âge (65-75].
Pour un âge d'entrée en dépendance très avancés de 75 ans, nous observons clairement une survie plus faible qu'aux âges moins avancés, ceci est due à une résistance de la population des dépendants face à cet état, et donc d'un maintien plus important.
Nous remarquons que les assurés les plus vieux (>90 ans) ont une courbe de survie qui décroit rapidement après la durée de 30 mois, ceci est expliqué par l'espérance de vie en dépendance qui est faible (sortie pour cause de décès).
Ces remarques confirment que l'établissement d'une loi de maintien en dépendance selon la tranche d'âge peut s'avérer très pertinente pour la modélisation de la dépendance.

Justification statistique :
Le test de Log Rank confirme que les courbes de maintien selon la tranche d'âge sont significativement différentes à savoir la p value calculé est très petite p-value ≪ 0.05

```
> surv_diff <- survdiff(Surv(claimDurationAdjusted, status) ~ ageBand, data = data_etude)
> surv_diff
Call:
survdiff(formula = Surv(claimDurationAdjusted, status) ~ ageBand,
    data = data_etude)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
ageBand=(-1,20]   960      875      832    2.2391    2.5084
ageBand=(20,65]  3029     2675     2037  199.7939  241.5175
ageBand=(65,75]  3808     3191     3072    4.6101    5.9650
ageBand=(75,80]  3349     2776     2975   13.3530   17.1462
ageBand=(80,85]  4260     3497     3881   37.9821   51.9756
ageBand=(85,90]  3491     2922     3131   13.9186   18.0180
ageBand=(90,120] 1822     1544     1552    0.0424    0.0493

 Chisq= 290  on 6 degrees of freedom, p= <2e-16
>
```

Figure 23 : Sortie R du test de Log-Rank

On peut alors conclure que la distinction suivant l'âge concernant la survie des personnes dépendantes est justifiée statistiquement.

Le taux après lissage par tranche d'âge :
Comme pour la variable ''sexe'' nous avons retenu le lissage de Whittaker – Henderson.

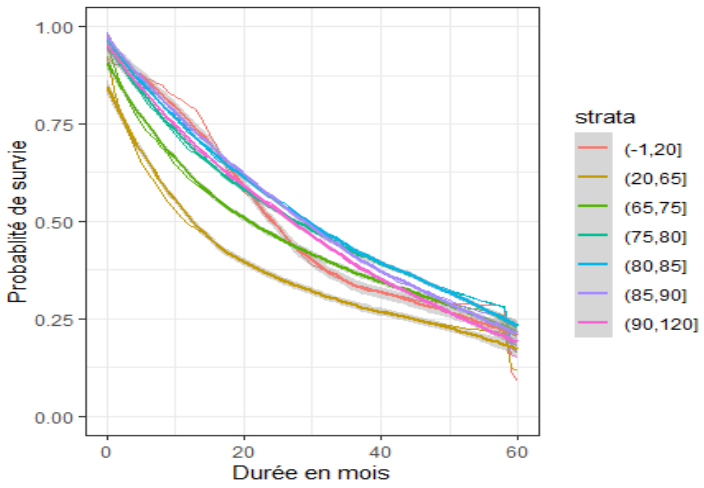


Figure 24 : La courbe de survie par tranche d'âge : Lissage par Whittaker-Henderson avec intervalle de confiance

Point d'attention :
Nous avons pensé à établir une courbe de dépendance par âge, mais on a remarqué que les taux bruts sont très erratiques et irréguliers, notamment pour les petits et les grands âges, ceci est dû au manque d'informations sur

les données de ces âges. En outre, comme nous allons voir ultérieurement dans la section du modèle semi paramétrique de Cox, il est utile de procéder à la discrétisation des variable quantitatives pour remédier au problème de la non-linéarité avec la variable à expliquer (taux d'hasard dans le cas du modèle semi paramétrique de Cox).

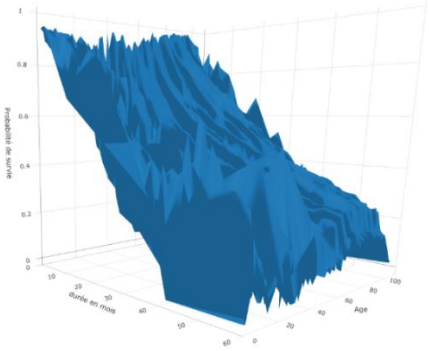


Figure 25 : La courbe de survie en fonction de la variable âge ''continue''

3. Segmentation par tranche d'âge d'entrée en dépendance et selon le Sexe de l'assuré

Nous avons vu précédemment que les deux variables : "Age d'entrée en dépendance" et "Sexe" ont un impact significatif sur le maintien de l'individu en dépendance.

Nous avons donc décidé de construire deux lois de maintien en dépendance : l'une issue de la population masculine et l'autre de la population féminine. Chacune prenant en paramètre l'ancienneté en dépendance et la tranche d'âge d'entrée en dépendance.

En général, les courbes des hommes et des femmes montrent une tendance similaire de diminution de la probabilité de survie au fil du temps. Cependant, dans certaines tranches d'âge, on peut remarquer des différences dans la vitesse de diminution de cette probabilité. Par exemple, dans les tranches d'âge avancées ((80-85] ans, (90-120] ans), les femmes semblent avoir une probabilité de survie légèrement plus élevée que les hommes à durée égale. En revanche, pour la tranche d'âge des enfants (-1,20], la courbe de survie des hommes décroît plus lentement que celle des femmes, indiquant une probabilité de survie supérieure pour les garçons dans cette catégorie.

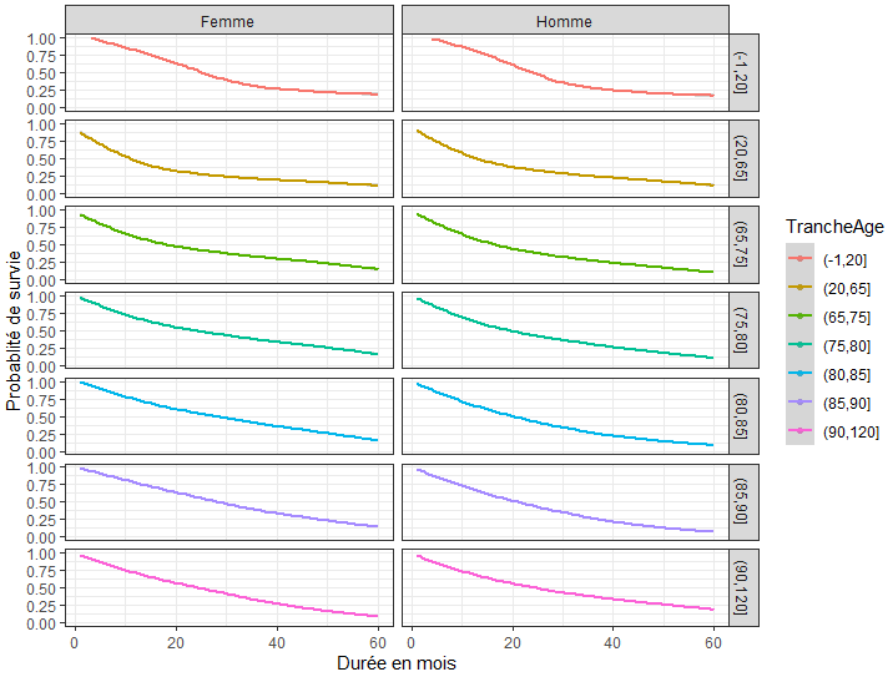


Figure 26 : La courbe de survie par sexe et par tranche d'âge : Lissage par Whittaker-Henderson

Nous allons conserver cette courbe de survie pour le calcul de la provision technique en dépendance.

Test de Log Rank

La p-value est inférieure du 5% induisant le rejet de l'hypothèse H_0. On peut conclure que la distinction suivant le sexe, l'âge concernant la survie des personnes dépendantes est justifié statistiquement.

```
> surv_diff <- survdiff(Surv(claimDurationAdjusted, status) ~ ageBand+gender, data = data_etude)
> surv_diff
Call:
survdiff(formula = Surv(claimDurationAdjusted, status) ~ ageBand +
    gender, data = data_etude)

                                N Observed Expected (O-E)^2/E (O-E)^2/V
ageBand=(-1,20], gender=Female    271      246      242    0.0671    0.0727
ageBand=(-1,20], gender=Male      689      629      590    2.5954    2.8645
ageBand=(20,65], gender=Female   1574     1383     1039  113.7630  129.2000
ageBand=(20,65], gender=Male     1455     1292      998   86.6949   98.0592
ageBand=(65,75], gender=Female   1868     1520     1600    4.0051    4.7171
ageBand=(65,75], gender=Male     1940     1671     1472   26.9193   31.2659
ageBand=(75,80], gender=Female   1741     1391     1669   46.4292   54.9017
ageBand=(75,80], gender=Male     1608     1385     1306    4.7890    5.4928
ageBand=(80,85], gender=Female   2428     1930     2396   90.7305  112.2858
ageBand=(80,85], gender=Male     1832     1567     1485    4.5670    5.2928
ageBand=(85,90], gender=Female   2078     1678     2034   62.3306   75.1657
ageBand=(85,90], gender=Male     1413     1244     1097   19.7901   22.3227
ageBand=(90,120], gender=Female   986      798      913   14.3798   16.1279
ageBand=(90,120], gender=Male     836      746      640   17.7141   19.4039

 Chisq= 529  on 13 degrees of freedom, p= <2e-16
>
```

Figure 27 : Sortie R du test de Log-Rank par tranche d'âge et par sexe

Conclusion :

L'estimateur de Kaplan Meier de par sa simplicité et sa robustesse produit des résultats à la fois satisfaisants et exploitables pour la suite de l'étude. Nous avons démontré que la segmentation par sexe semble pertinente à prendre en compte, chose que nous avons prouvé ''visuellement'' par l'apparition d'une différence entre les courbes de survie en fonction du sexe. Nous avons conclu à l'aide de test Log Rank que l'âge pourrait être un facteur à prendre en compte dans la construction de lois de maintien. Les causes de dépendance chez les personnes âgées et chez les personnes plus jeunes sont souvent différentes et vont donc avoir un impact sur le montant de la provision.

Les premiers résultats indiquent que les taux de maintiens sont en général un peu plus élevés pour les femmes que pour les hommes. L'analyse des lois de survie en dépendance montre également une sensibilité à l'âge d'entrée en dépendance, plus cet âge est élevé, plus le maintien en dépendance est élevé avec deux exceptions : 1) pour la tranche d'âge ''enfant'' qui montrait une courbe de maintien longue pour les durées moins de 11 mois et 2) au niveau de la tranche d'âge supérieur à 90 ans qui représentait une sortie rapide que nous avons jugé logique et en lien avec l'espérance de vie qui diminue avec l'âge.

Afin de résoudre les taux bruts erratiques obtenus via l'estimateur de Kaplan Meier, nous avons fait recours à des méthodes de lissage, notamment la méthode de Loess et de Whittaker-Henderson, sur la base de critère de fidélité nous avons retenu le lissage de Whittaker-Henderson pour les deux variables de construction de la loi de maintien.

La méthode de Kaplan Meier nous a permis de prouver que la population étudiée est hétérogène, donc il est nécessaire de prendre en compte les spécificités de chaque sous-groupe. En supposant que l'hétérogénéité est la conséquence d'un mélange de sous-populations caractérisées chacune par des variables observables, on s'intéresse donc à des modèles dits d'hétérogénéité intégrant l'effet des variables explicatives observables. Pour étudier l'influence de certaines caractéristiques sur la durée en dépendance, nous allons utiliser 3 variables explicatives : le sexe, l'âge à la survenance et le type de service offert aux assurés (soins dans un centre d'hospitalisation ou les soins à domicile). Cette question sera abordée dans un contexte semi-paramétrique avec le modèle de Cox.

III. Modèle de Cox

Le modèle à hasard proportionnel de Cox est un modèle semi-paramétrique qui introduit un effet multiplicatif des variables explicatives. Il cherche donc à évaluer l'effet de variables explicatives sur la fonction de hasard de base.

1. Principe du modèle

Le modèle de Cox permet d'exprimer le risque instantané de survenue de l'événement en fonction de l'instant t et des variables explicatives X_j

Pour rappel, le risque instantané de survenue de l'événement représente la probabilité d'apparition de l'événement à l'instant t sachant qu'il ne s'est pas encore réalisé juste avant l'instant t.

Le risque instantané (appelé également fonction de hasard) est de la forme :

$$\mu(t, X) = \mu_0(t) \times exp(\beta' X)$$

Avec :

- $X = (X_1, \dots, X_p)$ un vecteur de variable aléatoires explicatives
- μ_0 la fonction de hasard de base commune à tous les individus et qui ne dépend que du temps
- $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ paramètre ne dépendant pas du temps : il représente l'effet des covariables sur le risque instantané

Le modèle de Cox peut être vu comme le produit d'une fonction de hasard de référence qui représente le risque instantané de sortie de dépendance et un risque relatif représenté par le facteur exponentiel $exp\left(\sum_{i=1}^{p} \beta_i X_i\right)$.

C'est un modèle semi-paramétrique puisqu'aucune supposition n'est faite sur $\mu_0(t)$. La fonction de risque de base ne dépend pas de l'individu considéré et traduit donc l'hypothèse que la dépendance au temps du risque de connaître l'évènement est identique pour tous les individus.

Hypothèses du modèle de Cox

Ce modèle suppose plusieurs hypothèses :

- Il existe une relation log-linéaire entre fonction de risque instantané et les covariables :

$$ln\left[\frac{\mu(t, X)}{\mu_0(t)}\right] = \beta' X$$

- Le rapport des fonctions de risque instantané pour deux individus $i_1$et i_2 n'ayant qu'une covariable qui diffère entre eux, est proportionnel, ne dépend que de X_{i1} et X_{i2} et ne dépend pas du temps.

$$\frac{\mu(t, X_{i1})}{\mu(t, X_{i2})} = \frac{\mu_0(t) \times exp(\beta' X_{i1})}{\mu_0(t) \times exp(\beta' X_{i2})} = exp\left(\beta'(X_{i1} - X_{i2})\right)$$

  Cette hypothèse signifie que le rapport en logarithme entre les courbes de hasard de 2 groupes d'individus est non seulement proportionnel à la différence entre les valeurs de la variable, mais aussi et surtout indépendant du temps.

1.1 Estimation du modèle

Soient :

- T_i la durée observée jusqu'à la sortie de l'état de dépendance de l'assuré i ou la durée jusqu'à la date de fin d'observation si l'individu est toujours en dépendance
- X_i la valeur de la covariable de l'individu i
- $R(T_i)$: le risque à la période T_i, l'ensemble des individus à risque juste avant l'instant T_i

La probabilité conditionnelle que cet $i^{ème}$ individu n'est plus en dépendance en T_i s'exprime comme suit :

$$\frac{\mu_0(T_i) \times exp\left(\beta' X_{(i)}\right)}{\sum_{j \in R(T_i)} \mu_0(T_i) \times exp\left(\beta' X_j\right)} = \frac{exp\left(\beta' X_{(i)}\right)}{\sum_{j \in R(T_i)} exp\left(\beta' X_j\right)}$$

Cette probabilité dépend uniquement du paramètre β

La fonction de hasard de base est vue comme une contrainte au modèle justifiant l'utilisation d'une vraisemblance dite partielle :

$$L_p = \prod_{i=1}^{D} \frac{exp\left(\beta' X_{(i)}\right)}{\sum_{j \in R(T_i)} exp\left(\beta' X_j\right)}$$

Avec D le nombre de cas de sortie de dépendance observés parmi les n assurés.
La vraisemblance partielle ne dépend pas de la fonction de hasard de base $\mu_0(t)$. La maximisation de la vraisemblance partielle permet d'estimer les coefficients β sans connaître la fonction de hasard de base.
On obtient alors une estimation du vecteur de paramètre β en égalisant à zéro les dérivées premières de la log-vraisemblance.

Evénements simultanés

Le raisonnement précédent suppose des temps d'événements distincts. Dans le cas des données réelles, l'hypothèse de continuité, n'est pas toujours vérifiée. En effet, il pourrait arriver que deux ou plusieurs individus sortent de la dépendance au même âge. Pour assurer la continuité, il faudrait alors disposer d'informations plus fines sur l'instant de sortie comme le jour ou l'heure.
On rappelle que la probabilité que l'individu j sort de l'état de dépendance en T_i est donnée par la formule suivante :

$$p_j = \frac{exp(\beta' X_j)}{\sum_{k \in R(T_i)} exp(\beta' X_k)}$$

En présence de plusieurs événements, la méthode ''exacte'' consiste à admettre que les événements se produisent les uns à la suite des autres. Cependant, on ne connaît pas l'ordre des événements, il faut donc considérer toutes les possibilités. Dans le cas de deux sujets s_1 et s_2 de caractéristiques X_1 et X_2 qui sortent de la dépendance en T_i la contribution exacte à la vraisemblance est :

$$\frac{exp(\beta' X_1) \times exp(\beta' X_2)}{\sum_{j \in R(T_i)} exp(\beta' X_j) \times \sum_{j \in R(T_i) \setminus s_1} exp(\beta' X_j)} + \frac{exp(\beta' X_1) \times exp(\beta' X_2)}{\sum_{j \in R(T_i)} exp(\beta' X_j) \times \sum_{j \in R(T_i) \setminus s_2} exp(\beta' X_j)}$$

Le problème de cette méthode est que le temps de calcul devient très long quand il y a beaucoup d'événements simultanés. Ainsi, on utilise le plus souvent l'approximation de Breslow qui consiste à supposer que la contribution des d_i événements en T_i est le produit des probabilités p_j pour les unités sorties en T_i (i.e. $\sum_{j \in R(T_i)} exp(\beta' X_j) \approx \sum_{j \in R(T_i) \setminus k} exp(\beta' X_j)$

$$L_B(T_i) = \prod_{\substack{j:unités \\ sorties\ en\ T_i}} p_j = \frac{exp\left(\beta' \left(\sum_{\substack{j:unités \\ sorties\ en\ T_i}} X_j\right)\right)}{\left(\sum_{k \in R(T_i)} exp(\beta' X_k)\right)^{d_i}}$$

L'approximation de Breslow de la vraisemblance totale est :

$$\prod_{i=1}^{D} L_B(T_i)$$

La maximisation de cette vraisemblance est rapide. De plus, si le nombre d'événements simultanés n'est pas trop grand alors la méthode est assez précise.
On peut estimer le risque cumulé de base par l'estimateur de Breslow :

$$\widehat{\Lambda_0(t)} = \sum_{i:T_i \leq t} \frac{d_i}{\sum_{j \in R(T_i)} exp(\beta' X_j)}$$

Avec $\Lambda_0(t) = \int_0^t \mu_0(s) ds$ et d_i est le nombre de sortie en T_i
On peut alors en déduire un estimateur de la fonction de survie pour un vecteur de covariable X.

$$S(t, X) = exp\left(-\int_0^t \mu(u, X) du\right)$$

Soit :

$$\widehat{S(t, X)} = exp\left(-\widehat{\Lambda_0(t)} \times exp(\hat{\beta}' X)\right)$$

Une fois un modèle et les coefficients de celui-ci estimés, il faut le valider à travers différents tests.

1.2 Test de significativité des coefficients

Le test de significativité des coefficients permet de savoir si les variables explicatives sont significatives et donc s'il est cohérent de les garder dans notre modèle. Pour chaque coefficient on teste l'hypothèse suivante :

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

Pour vérifier la validité de l'hypothèse H_0, on utilise la statistique de test :

$$\sqrt{n}\frac{\hat{\beta}}{\hat{\sigma}} \xrightarrow{n\to+\infty} N(0,1)$$

On peut par conséquent obtenir un intervalle de confiance de niveau 1-α (α risque d'erreur du premier espèce), $\hat{\beta} \pm z_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n}}$

Effet d'une co-variable (exemple de la variable sexe) :

Soit X_i une variable qualitative tel que : $X_i = 1$ si l'individu est un homme sinon 0

$$\frac{\mu_0(t) \times exp\big(\beta_1 \times 1 + \cdots + \beta_p X_p\big)}{\mu_0(t) \times exp\big(\beta_1 \times 0 + \cdots + \beta_p X_p\big)} = exp(\beta_1)$$

Ainsi le coefficient $exp(\beta_1)$ est le rapport de risque entre un homme et une femme toutes choses étant égales par ailleurs, c'est à dire si les deux individus ont des caractéristiques communes, excepté pour la variable "Sexe".

- Si $\beta_1 > 0$, $exp(\beta_1) > 1$: le risque que l'événement d'intérêt se produise est plus élevé chez les hommes que chez les femmes.
- Si $\beta_1 < 0$, $exp(\beta_1) < 1$: le risque que l'événement d'intérêt se produise est plus faible chez les hommes que chez les femmes.
- Si $\beta_1 = 0$, $exp(\beta_1) = 1$: le risque que l'événement d'intérêt se produise est le même chez les hommes que chez les femmes

1.3 Vérification des hypothèses du modèle :

Test de l'hypothèse de proportionnalité

Le modèle de Cox postule que les risques sont proportionnels entre individus. Il est nécessaire de vérifier cette hypothèse, afin de s'assurer de la fiabilité des résultats.

Plusieurs approches sont possibles :

- Comparaison graphique des courbes de survie (Log minus Log curve)
- Analyse des résidus de Schoenfeld ;
- Test de corrélation des résidus de Schoenfeld avec le temps (test de Therneau et Grambsch)

- **<u>Courbes LML (Log minus Log)</u>**

Nous considérons la transformation suivante des courbes de survie : $\ln\big(-ln\big(S(t,x)\big)\big)$

Cette transformation, dite LML pour "Log Minus Log", a la propriété suivante : Si l'hypothèse de risques proportionnels est valide, alors

$$\begin{aligned} S(t,x) = \exp\big(-\Lambda(t)\big) &= \exp\left(-\int_0^t \mu(u,X)du\right) \\ &= \exp\left(-\int_0^t \mu_0(u)\exp(\beta' X)\,du\right) = \exp\left(-\int_0^t \mu_0(u)du\right)^{\exp(\beta' X)} = \exp\left(-\Lambda_0(t)\right)^{\exp(\beta' X)} \\ &= {S_0}^{\exp(\beta' X)} \end{aligned}$$

$$\ln\big(-ln\big(S(t,x)\big)\big) = \ln\big(-\ln\big(S_0(t)\big)\big) + \beta' X$$

Pour deux profils différents X_1 et X_2, la différence entre les courbes LML vaut $(X_2 - X_1)\beta$

Cette quantité est indépendante du temps.

- Les courbes de survie après transformation LML sont donc parallèles pour différentes valeurs de x.
- Il suffit alors de tracer les courbes LML correspondant aux différents niveaux d'une covariable, les autres co-variables restant constantes, et de les comparer.

- S'il est possible de superposer les différentes courbes par simple translation, alors l'hypothèse de proportionnalité est vérifiée

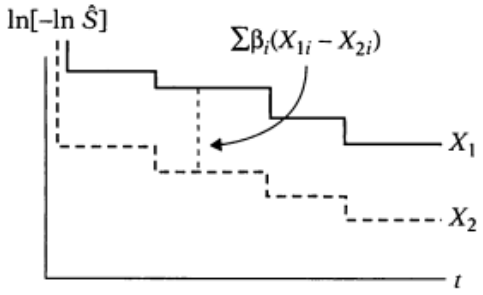


Figure 28 : Illustration de la courbe LML

- **Résidus de Schoenfeld**

Dans un modèle de régression linéaire traditionnel, les résidus mesurent la différence entre les valeurs observées de la variable dépendante et les valeurs prédites par le modèle.
Dans le cas d'un modèle de Cox, c'est le risque instantané qui est expliqué et la notion de résidu n'a alors pas de sens, car il n'y a pas moyen de calculer une différence entre valeurs observées et prédites.
Plusieurs autres notions de résidus ont ainsi dû être définies (Schoenfeld, déviance, martingale, score, ...). Ce sont ceux de Schoenfeld qu'on développe ici.
Les résidus de Schoenfeld sont calculés par observation par co-variable. Ils sont définis uniquement aux moments observés de l'événement étudié (donnée non censurée). Pour le $i^{ème}$ assuré et la $k^{ème}$ co-variable, l'estimation du résidu de Schoenfeld, $r_{i,k}$, est donné par (notation de Hosmer et Lemeshow) : $\widehat{r_{i,k}} = x_{ik} - \widehat{\overline{x_{w_i k}}}$ où :

- x_{ik} est la $k^{ème}$ covariable de l'assuré k
- $\widehat{\overline{x_{w_i k}}}$ est une moyenne pondérée des valeurs des covariables pour ceux qui sont soumis au risque fixé au moment de l'événement donné.
- Une valeur positive de $r_{i,k}$ indique une valeur X qui est plus élevée que prévu à l'heure de la sortie (e.g. décès).

Les résidus de Schoenfeld ne sont pas une mesure de l'ajustement du modèle aux données. Ils s'interprètent comme une mesure de la différence de profil (par rapport aux covariables du modèle) entre un individu subissant l'événement étudié et l'ensemble des individus exposés au risque.
Les résidus de Schoenfeld peuvent être analysés sur la base de graphiques, afin de détecter un éventuel non-respect de l'hypothèse de proportionnalité. L'idée consiste à représenter les résidus en fonction d'une transformation du temps. On peut aussi rajouter sur le même graphique une courbe représentant l'évolution moyenne des résidus en fonction du temps. Cette courbe donne la tendance générale. Toute différence par rapport à une droite horizontale représente une déviation par rapport à l'hypothèse de proportionnalité.

- **Test de corrélation des résidus de Schoenfeld avec le temps**

Une méthode de test des résidus consiste à calculer la corrélation entre les résidus et les durées de survie. Si l'hypothèse nulle d'absence de corrélation est acceptée, alors l'hypothèse de proportionnalité est vérifiée. Sinon, elle est rejetée. Alternativement, il est aussi possible de calculer une régression linéaire expliquant les résidus à l'aide du logarithme du temps, puis à tester si la pente de la droite de régression est bien nulle.

Le rejet de l'hypothèse de proportionnalité des risques implique que le rôle des variables explicatives du modèle évolue au fil du temps. Une autre méthode de test consiste alors à introduire des coefficients de régression évoluant en fonction du temps dans le modèle de Cox et à tester leur significativité. S'ils sont significatifs, alors l'hypothèse de proportionnalité des risques est remise en question.

Point d'attention : Quelques solutions en cas de rejet de l'hypothèse de proportionnalité :

En cas de rejet de l'hypothèse de proportionnalité, deux solutions relativement simples peuvent être retenues. La première consiste à modéliser l'évolution des coefficients β dont les tests nous amènent à rejeter la constance. Pratiquement, il s'agit simplement de retenir le modèle avec les interactions entre les co-variables concernées et une fonction adéquate de la durée t.

La seconde possibilité est le recours à un modèle stratifié. En effet, la stratification consiste à calculer un modèle de Cox en attribuant une valeur différente du risque de base $\mu_0(t)$ à chaque catégorie de la variable de stratification. En revanche, l'influence des variables explicatives, et donc les valeurs estimées des paramètres β, est commune à toutes les catégories. Cette méthode permet d'inclure dans un modèle de Cox une variable ne satisfaisant pas à l'hypothèse de proportionnalité des risques.

2. Application du modèle de Cox.

Pour tester la significativité de l'impact des covariables sur la loi de maintien, nous utilisons le modèle de Cox décrit dans le paragraphe précédent à l'aide de la fonction *coxph* de la *library (survival)* pour obtenir les résultats sous le logiciel R.

Cette section est structurée comme suit : nous commençons par estimer les paramètres du modèle, ensuite, nous procédons à la vérification de l'hypothèse fondamentale du modèle de Cox, à savoir l'hypothèse des risques proportionnels, en cas de non-respect de cette hypothèse, nous explorons des modèles alternatifs, tels que le modèle avec interaction temporelle et le modèle de Cox stratifié.

2.1. Segmentation selon le sexe ''gender''

2.1.1. Estimation d'un modèle de Cox avec la covariable ''gender''

La sortie R ci-dessous présente le résultat du modèle Cox pour tester et estimer l'impact de la variable ''gender'' sur la loi de maintien.

```
> summary(cox1)
Call:
coxph(formula = Surv(duration_adj, status) ~ gender, data = data_etude)

  n= 20450, number of events= 17480

              coef exp(coef) se(coef)     z Pr(>|z|)
genderMale 0.21121   1.23517  0.01516 13.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
genderMale     1.235     0.8096     1.199     1.272

Concordance= 0.529  (se = 0.002 )
Likelihood ratio test= 193.5  on 1 df,   p=<2e-16
Wald test            = 194.2  on 1 df,   p=<2e-16
Score (logrank) test = 194.9  on 1 df,   p=<2e-16

> print(paste0("Le critère AIC = ",AIC(cox1)))
[1] "Le critère AIC = 315209.096323234"
```

Figure 29 : Sortie R de la régression de Cox pour la variable sexe

La sortie R affiche un coefficient Bêta pour la variable ''gender'' de 0.21 avec une p-value significativement faible et inférieure à $2e^{-16}$, ceci indique qua la variable est significativement non nulle et par conséquent, une variable discriminante pour la loi de maintien.

En outre, le coefficient β étant positif, $exp(\beta)$ est supérieur à 1, indique que le risque de dépendance est plus fort chez les femmes que chez les hommes (le risque de sortie est plus faible chez les femmes que chez les hommes), on peut dire qu'en moyenne, les hommes ont 23% de moins de chance de rester en dépendance que les femmes. Ce qui conforte les premiers résultats émis lors de l'utilisation de l'estimateur de Kaplan-Meier.

Le graphique ci-après présente l'évolution de la valeur de Bêta avec la durée en dépendance. Nous constatons que la valeur de Beta ne varie pas significativement en fonction du temps.

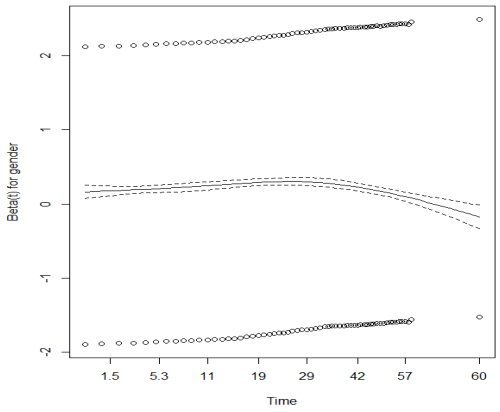


Figure 30 : Le coefficient Beta de la variable sexe en fonction du temps

Dans le cas de la régression sur la variable sexe nous allons montrer que l'hypothèse de proportionnalité est vérifiée justifiant le non nécessité de procéder à une modélisation d'un effet dépendant du temps.

2.1.2. Test de de l'hypothèse de proportionnalité

La courbe LML (Log Minus Log)

Pour rappel, le test graphique consiste à tracer les courbes LML pour chaque valeur d'une covariables, les autres étant maintenues constantes, et de vérifier si cette translation se vérifie. Il y a proportionnalité si on peut obtenir la courbe LML du haut par simple glissement de celle du bas vers le haut. Dans la figue ci-dessous, on remarque que c'est quasiment le cas pour la variable ''gender'' où le parallélisme semble mieux marqué pour des durées d'au-delà de 22 mois.

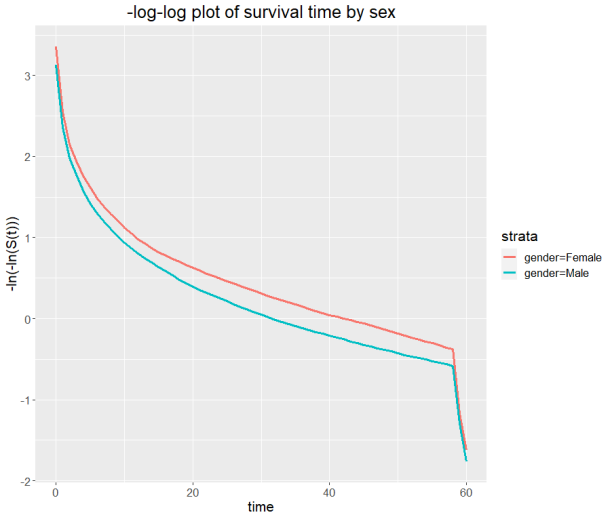


Figure 31 : la courbe LML de la variable sexe

Vérification de l'hypothèse de proportionnalité à l'aide du test des résidus Schoenfeld

Utilisant R, on peut extraire les résidus partiels de Schoenfeld du modèle de Cox qui exclut des interactions. Comme il a été précédemment mentionné, ces résidus partiels de Schoenfeld sont spécifiques à chaque covariable.

```
> cox.model <- cox.zph(cox1)
> cox.model
        chisq df    p
gender   2.18  1 0.14
GLOBAL   2.18  1 0.14
> ggcoxzph(cox.model)
```

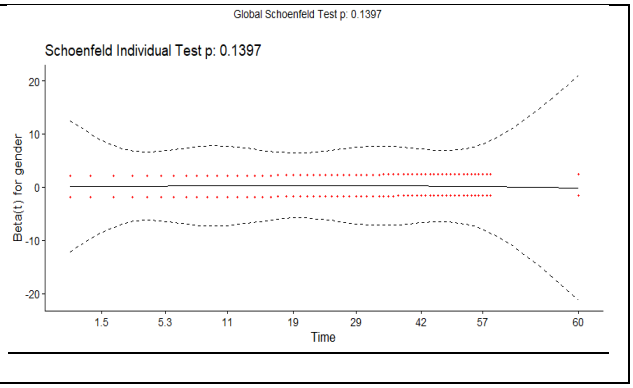


Figure 32 : Test de proportionnalité de la variable sexe à l'aide du test des résidus Schoenfeld

Dans la sortie R, nous voyons deux lignes : la première nommée ''**gender''** représente le test de Schoenfeld pour la covariable sexe. Le chi-square (chisq) est de 2.18 avec 1 degré de liberté (df). La valeur p correspondante est de 0.14. Puisque la p-value est supérieure à 0.05, cela indique qu'il n'y a pas de preuve statistiquement significative que les risques proportionnels sont violés pour la covariable sexe. En d'autres termes, l'effet du sexe sur le risque semble être constant dans le temps. Tandis que la deuxième « **GLOBAL »** représente un test global pour toutes les covariables dans le modèle. Le chi-square est également de 2.18 avec 1 degré de liberté, et la p-value est de 0.14. Cela signifie qu'il n'y a pas de preuve globale de violation des risques proportionnels dans le modèle. Étant donné que notre modèle ne comprend qu'une seule variable, le test global correspond au test individuel.

Ce constat est ainsi confirmé par le graphe des résidus qui suggère que l'hypothèse de risques proportionnels est raisonnable pour la covariable sexe, car les résidus de Schoenfeld ne montrent pas de tendances systématiques dans le temps et la ligne de régression des résidus est proche de l'horizontale.

2.1.3. La représentation graphique de la courbe de survie

Grâce aux paramètres estimés, on peut avoir un aperçu de la fonction de survie pour les deux modalités de la variable ''gender''.

Sur le graphe ci-dessous, on remarque bien que les fonctions de survie sont belles et bien différentes par sexe. On constate que la courbe de survie des femmes est toujours bien au-dessus de celle des hommes : les femmes ont tendance à plus se maintenir en dépendance que les hommes.

On remarque que cette tendance s'intensifie (l'écart entre les deux courbes de survie s'élargie davantage) à partir du 20ème mois d'ancienneté en dépendance.

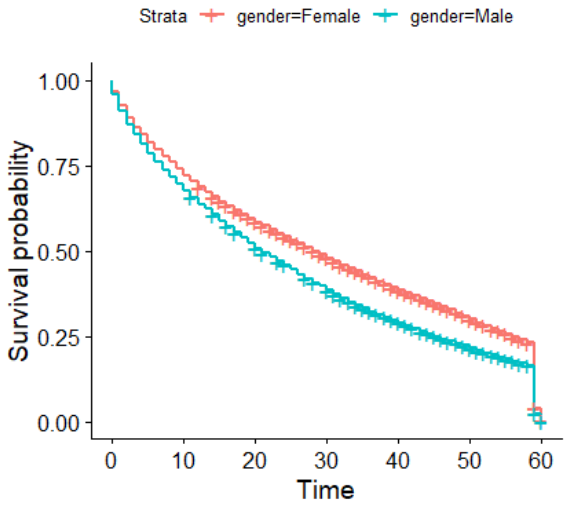


Figure 33 : la courbe de survie pour la variable sexe

Les résultats du modèle de Cox a permis d'affirmer que le genre de l'individu a une place importante dans l'estimation du maintien en dépendance.

2.2. Segmentation selon le type de service

2.2.1. Estimation d'un modèle de Cox avec la covariable type de service

Nous procédons comme pour la variable ''gender'' et nous allons utiliser le modèle de Cox pour vérifier la significativité de la variable ''type de service''.
Pour rappel, la variable ''type de service'' contient deux modalités : prestations ayant lieu à domicile ou en établissement spécialisé.

```
> cox3 <- coxph(Surv(duration_adj, status) ~  Last.Type.of.Services , data = data_etude)
> summary(cox3)
Call:
coxph(formula = Surv(duration_adj, status) ~ Last.Type.of.Services,
    data = data_etude)

  n= 15116, number of events= 12220
   (5334 observations effacées parce que manquantes)

                                      coef exp(coef) se(coef)     z Pr(>|z|)
Last.Type.of.ServicesHospitalisation 0.06618   1.06842  0.02494 2.653  0.00797 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                     exp(coef) exp(-coef) lower .95 upper .95
Last.Type.of.ServicesHospitalisation     1.068      0.936     1.017     1.122

Concordance= 0.498  (se = 0.002 )
Likelihood ratio test= 6.93  on 1 df,   p=0.008
Wald test            = 7.04  on 1 df,   p=0.008
Score (logrank) test = 7.04  on 1 df,   p=0.008

> print(paste0("Le critère AIC = ",AIC(cox3)))
[1] "Le critère AIC = 213014.147766979"
```

Figure 34 : Sortie R de la régression de Cox pour la variable type d'hospitalisation

Le coefficient Bêta de la variable ''type de service'' est 0.07 avec une p-value significativement faible et inférieure à 5%. La variable ''type de service'' est une variable faiblement discriminante pour la loi de maintien. De plus β étant positif, $exp(\beta)$ est légèrement supérieur à 1 : le risque de dépendance est légèrement inférieur chez les assurés qui domicilient dans des centres d'hospitalisation que chez les assurés dépendants domiciliés chez eux, ceci pourrait s'expliquer par la détérioration de l'état de santé des individus recherchant des soins intensifs dans les établissements de santé, où la majorité des sorties résultent en décès.
Le graphique ci-après présente l'évolution de la valeur de Bêta en fonction de la durée en dépendance. Elle est négative au début pour les durées inférieures à 12 mois et ensuite positive pour les durées entre 12 mois et 55 mois et puis négatif, ce qui se traduit par une dépendance légèrement supérieure chez les assurés hospitalisés dans un établissement de santé pour les durées courtes et longue et ensuite une dépendance accrue chez les assurés vivant à domicile pour les durées entre 12 mois et 55 mois.

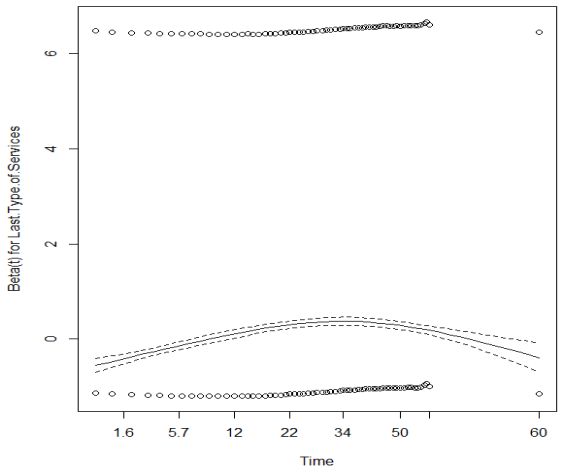


Figure 35 : le coefficient Beta de la variable ''type de service'' en fonction du temps

2.2.2. Test de l'hypothèse de proportionnalité

La courbe LML de la variable ''type de service'' :

D'après le graphique ci-dessous, nous constatons que l'hypothèse de proportionnalité n'est pas respectée (les deux courbes LML ne sont pas parallèles) pour la variable ''type de service''. Dans la pratique, cela signifie que la fonction de hasard de base ne sera pas la même selon que l'individu reçoit une aide dans un établissement spécialisé ou à domicile.

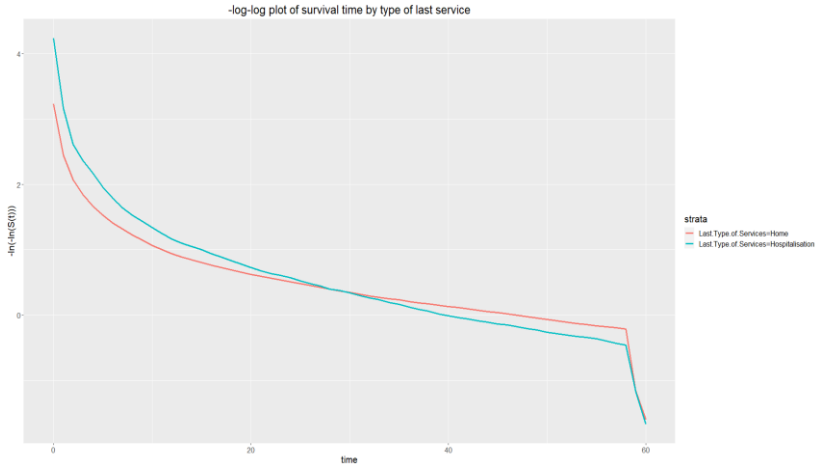


Figure 36 : la courbe LML de la variable ''type de service''

Pour remédier au problème de non-proportionnalité, nous pouvons considérer un modèle de Cox stratifié ou bien introduire une variable d'interaction avec le temps, pour rappel, une violation de l'hypothèse de proportionnalité implique que les coefficients des variables ne sont pas constants au cours du temps.

2.2.3. Covariable ''type de service'' dépendant du temps

Le graphique sur l'évolution de la valeur de Bêta avec la durée en dépendance (Figure 35 : le coefficient Beta de la variable ''type de service'' en fonction du temps), nous a illustré la nécessité d'introduire une variable d'interaction définie comme le produit de la variable ''type de service'' et la durée de survie.
L'approche consiste à tester une spécification de coefficient de la variable '' type de service'' variant avec t. On le réalise en complétant le modèle de Cox avec les interactions entre notre variable et t et en testant la significativité de ces interactions. Si l'on considère la variable ''type de service'' son effet est ainsi mesuré par :

$$\beta_{type\ de\ service=Hospitalisation} * \mathbb{1}_{type\ de\ service=Hospitalisation} + \beta_{type\ de\ service=Hospitalisation\_t} * \mathbb{1}_{type\ de\ service=Hospitalisation}$$
$$= \left(\beta_{type\ de\ service=Hospitalisation} + \beta_{type\ de\ service=Hospitalisation\_t}\right) * \mathbb{1}_{type\ de\ service=Hospitalisation}$$

Il apparait donc que l'introduction de l'interaction avec t n'est qu'une façon de spécifier un coefficient $\beta(t)$ qui varie avec t. Sous R, la déclaration d'une telle variable $X(t)$ peut être réalisée par l'option *time-transform tt* de *coxph*. Sa prise en compte dans le modèle s'effectue lors de sa déclaration en ajoutant $tt(X)$ comme covariable.

```
> cox3_interaction=coxph(Surv(duration_adj, status) ~ Last.Type.of.Services + tt(Last.Type.of.Services),data = data_etude,
+                        tt = list(
+                          function(Last.Type.of.Services, duration_adj, ...){
+                            # As Last.Type.of.Services is a categorical variable
+                            # we must transform it into a design matrix that
+                            # we can use for multiplication with time
+                            # Note: the [,-1] is for dropping the intercept
+                            mtrx <- model.matrix(~Last.Type.of.Services)[,-1]
+                            mtrx * duration_adj
+                          }
+                        ))
> summary(cox3_interaction)
Call:
coxph(formula = Surv(duration_adj, status) ~ Last.Type.of.Services +
    tt(Last.Type.of.Services), data = data_etude, tt = list(function(Last.Type.of.Services,
    duration_adj, ...) {
    mtrx <- model.matrix(~Last.Type.of.Services)[, -1]
    mtrx * duration_adj
}))

  n= 15116, number of events= 12220
   (5334 observations effacées parce que manquantes)

                                      coef exp(coef)  se(coef)      z Pr(>|z|)
Last.Type.of.ServicesHospitalisation -0.128996  0.878978  0.040242 -3.206  0.00135 **
tt(Last.Type.of.Services)             0.007616  1.007645  0.001179  6.459 1.05e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                     exp(coef) exp(-coef) lower .95 upper .95
Last.Type.of.ServicesHospitalisation     0.879     1.1377    0.8123    0.9511
tt(Last.Type.of.Services)                1.008     0.9924    1.0053    1.0100

Concordance= 0.518  (se = 0.002 )
Likelihood ratio test= 48.1  on 2 df,   p=4e-11
Wald test            = 50.75  on 2 df,   p=1e-11
Score (logrank) test = 51.01  on 2 df,   p=8e-12
```

Figure 37 : Sortie R de la régression de Cox avec la variable temps

Le résultat ci-dessus du test de significativité rejette l'hypothèse nulle au seuil de 5% pour la variable d'interaction.

On remarque d'après la sortie R que le la statistique du rapport du vraisemblance est supérieur à 0, avec une p-value <5%, ceci montre que le gain est clairement significatif par rapport au modèle naïf où tous les coefficients β sont nuls. Pour rappel, l'idée de la statistique du rapport de vraisemblance permet d´évaluer si globalement l'ensemble des facteurs explicatifs considérés améliore significativement l'ajustement du modèle naïf qui ne tient compte d'aucun facteur.

Dans notre exemple, le paramètre β(t) est estimé par -0.129 + 0.007616 × t

Point d'attention :

La création de la variable d'interaction améliore le modèle mais au détriment de l'interprétabilité, ceci conduit à ce que l'on appelle le compromis interprétabilité /performance, c'est à dire, plus un modèle est performant, moins il est explicable, et vice versa.

Vérification de l'hypothèse de proportionnalité à l'aide du test des résidus Schoenfeld

R permet d'obtenir les résidus partiels de Schoenfeld du modèle de Cox estimé (modèle avec interaction). Comme mentionné auparavant, les résidus partiels de Schoenfeld sont associés à chaque covariable, donc il y aura 3 résidus comme illustré ci-dessous.

```
> test.residu <- cox.zph(cox3_interaction)
> test.residu
                                        chisq df      p
Last.Type.of.Services                1.219380  1 0.2695
duration_adj                         0.000522  1 0.9818
Last.Type.of.Services:duration_adj 2.06e-02  1   0.89
GLOBAL                               7.518423  3 0.0571
```

Figure 38 : Sortie R du test de vérification de l'hypothèse de proportionnalité

Une valeur de p-value inférieure à 5% indique que l'hypothèse n'est pas vérifiée. Notre objectif pour mettre en évidence les situations de non-proportionnalité des risques, est donc d'avoir pour chacune des variables explicatives ainsi que pour l'ensemble du modèle une p-value > 5%.

Il apparaît que p-value est supérieur à 5% globalement et pour chaque variable prise individuellement. Toutes les p-values n'étant pas significatives nous rejetons l'hypothèse de non-indépendance entre le temps et les résidus de Schoenfeld. De plus le test global nous donne une p-value de 6% >5%. L'hypothèse des risques proportionnels est ainsi validée. On peut opérer une vérification supplémentaire en regardant les résidus de Schoenfeld en fonction du temps comme sur la figure ci-dessous.

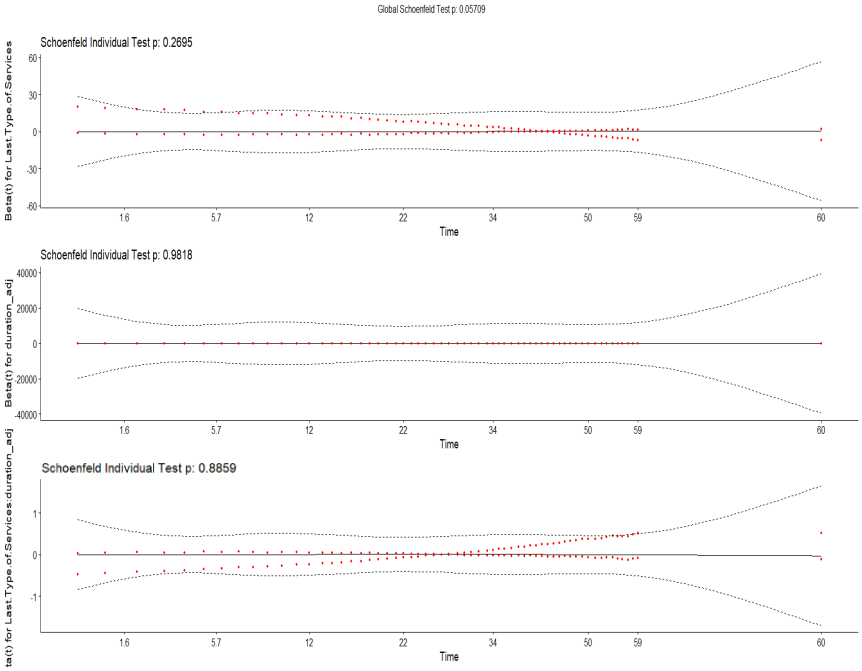


Figure 39 : Représentation graphique des résidus de Schoenfeld

Graphiquement les résidus de Schoenfeld sont alignés autour de 0, sans tendance temporelle : ceci confirme a posteriori l'hypothèse de proportionnalité.

2.2.4. La représentation graphique de la courbe de survie

Grâce aux paramètres estimés du modèle avec interaction temporelle, nous pouvons obtenir un aperçu de la fonction de survie par modalités :

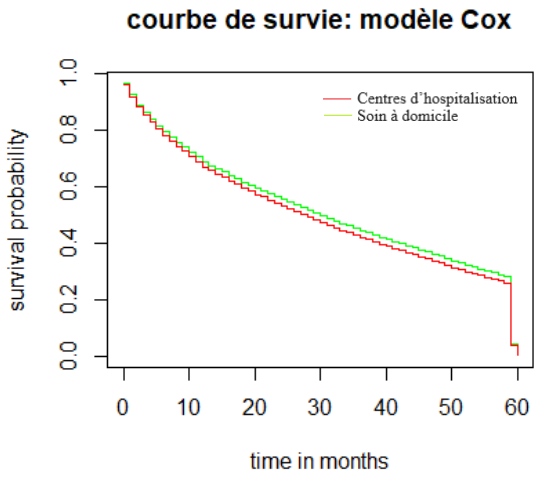


Figure 40 : la courbe de survie pour la variable type de service

Sur le graphe on remarque que les fonctions de survie ne sont pas très différentes par type de service fourni, ceci est due à l'estimation du coefficient β qui s'écarte légèrement du 0 ($exp(\beta) \cong 1$) entre les deux modalités. La lecture du graphe montre que les le risque de dépendance est légèrement inférieur chez les assurés qui domicilient dans des centres d'hospitalisation (courbe rouge) que chez les assurés dépendants domiciliés chez eux (courbe vert).

2.3. Segmentation selon tranche d'age ''AgeBand''

2.3.1. Estimation d'un modèle de Cox avec la covariable ''ageBand''

Pour tester la significativité de l'impact de la variable ''ageBand'' sur la loi de maintien, nous utilisons à nouveau le modèle de Cox. Pour remédier au problème de la non-linéarité de la variable ''âge'' nous avons retenu la discrétisation utilisée lors de l'élaboration de l'estimateur de Kaplan-Meier.

Le tableau ci-dessous présente le résultat de ce modèle. La première tranche d'âge :"(-1,20]" est retenue comme la modalité de référence du modèle, un bêta est alors estimé pour chacune des six autres classes, pour comparer l'impact de chacune par rapport à la première tranche ("(-1,20]").

```
> summary(cox2)
Call:
coxph(formula = Surv(duration_adj) ~ ageBand, data = data_etude)

  n= 20450, number of events= 20450

                   coef exp(coef) se(coef)     z Pr(>|z|)
ageBand(20,65]  0.34958   1.41847  0.03725 9.384  < 2e-16 ***
ageBand(65,75]  0.19984   1.22121  0.03634 5.499 3.82e-08 ***
ageBand(75,80]  0.08314   1.08670  0.03684 2.257  0.02400 *
ageBand(80,85]  0.07844   1.08160  0.03596 2.181  0.02916 *
ageBand(85,90]  0.09742   1.10232  0.03669 2.655  0.00793 **
ageBand(90,120] 0.17140   1.18697  0.04016 4.269 1.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
ageBand(20,65]      1.418     0.7050     1.319     1.526
ageBand(65,75]      1.221     0.8189     1.137     1.311
ageBand(75,80]      1.087     0.9202     1.011     1.168
ageBand(80,85]      1.082     0.9246     1.008     1.161
ageBand(85,90]      1.102     0.9072     1.026     1.185
ageBand(90,120]     1.187     0.8425     1.097     1.284

Concordance= 0.545  (se = 0.002 )
Likelihood ratio test= 190  on 6 df,   p=<2e-16
Wald test            = 196.2  on 6 df,   p=<2e-16
Score (logrank) test = 197.1  on 6 df,   p=<2e-16

> print(paste0("Le critère AIC = ",AIC(cox2)))
[1] "Le critère AIC = 315190.856978947"
>
```

Figure 41 : Sortie R de la régression de Cox pour la variable tranche d'âge''ageBand''

Les coefficients Bêtas des modalités de la variable âge sont strictement positifs cela signifie que la loi de maintien est plus courte par rapport à la variable de référence (-1,20]. Le taux de hasard est plus important chez la tranche d'âge (20,65] que chez la tranche d'âge (-1,20], avec un risque relatif de 1.41847, ceci veut dire que les assurés de cette tranche d'âge sont 41.847% plus susceptibles de sortir de l'état de dépendance par rapport aux jeunes de 0 à 20 ans. On constate également que le modèle prédit que les assurés de la tranche d'âge (90,120] auront un taux de sortie plus important par rapport aux assurés d'âge 75 ans à 90 ans, ceci s'explique par l'espérance de vie qui décroit avec l'âge.

Les estimations retenues de ce modèle illustrent quelques contradictions avec le modèle de Kaplan Meier (KM), notamment sur la tranche (80,85] car nous attendions avoir un coefficient négatif qui reflète une courbe de maintien plus longue par rapport à la modalité de référence (-1,20], ainsi nous jugeons important de vérifier l'hypothèse de proportionnalité.

2.3.2. Test de l'hypothèse de proportionnalité

La courbe LML de la variable ''ageBand'' :

Le graphique ci-dessous indique une violation de l'hypothèse de proportionnalité, puisque les courbes LML ne sont pas parallèles pour la variable ''ageband''. En pratique, cela implique que la fonction de risque de base variera en fonction de la tranche d'âge de l'individu.

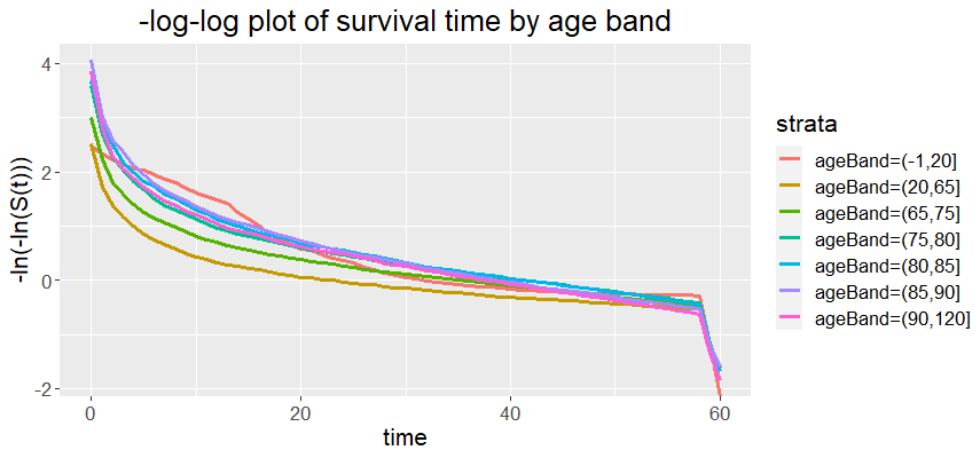


Figure 42 : la courbe LML de la variable tranche d'âge ''AgeBand''

Vérification de l'hypothèse de proportionnalité à l'aide du test des résidus Schoenfeld

Comme prévu, la p-valeur du test est en effet inférieure à 5% justifiant que la variable tranche d'âge ne vérifie pas l'hypothèse de hasards proportionnels.

```
> cox.zph(cox2, transform="identity")
         chisq df       p
ageBand    474  6 <2e-16
GLOBAL     474  6 <2e-16
```

Figure 43 : Sortie R du test de vérification de l'hypothèse de proportionnalité

2.3.3. Covariable ''ageband'' dépendante du temps

Pour tenir compte de la non-proportionnalité, La démarche consiste donc à intégrer au modèle des covariables dépendantes du temps comme déjà fait dans la modélisation de la variable ''type de service''.

```
> cox2_interaction=coxph(Surv(duration_adj, status) ~ ageBand + tt(ageBand),data = data_etude,
+       tt = list(
+         function(ageBand, duration_adj, ...){
+           # As ageBand is a categorical variable
+           # we must transform it into a design matrix that
+           # we can use for multiplication with time
+           # Note: the [,-1] is for dropping the intercept
+           mtrx <- model.matrix(~ageBand)[,-1]
+           mtrx * duration_adj
+         }
+       ))
> summary(cox2_interaction)
Call:
coxph(formula = Surv(duration_adj, status) ~ ageBand + tt(ageBand),
    data = data_etude, tt = list(function(ageBand, duration_adj,
        ...) {
        mtrx <- model.matrix(~ageBand)[, -1]
        mtrx * duration_adj
    }))

  n= 20450, number of events= 17480

                                 coef  exp(coef)   se(coef)      z Pr(>|z|)
ageBand(20,65]               0.7091752  2.0323143  0.0613095 11.567  < 2e-16 ***
ageBand(65,75]               0.2852525  1.3300978  0.0610563  4.672 2.98e-06 ***
ageBand(75,80]              -0.0215528  0.9786778  0.0627915 -0.343 0.731415
ageBand(80,85]              -0.1432584  0.8665301  0.0615052 -2.329 0.019848 *
ageBand(85,90]              -0.1546753  0.8566933  0.0627668 -2.464 0.013729 *
ageBand(90,120]             -0.0789969  0.9240428  0.0688386 -1.148 0.251148
tt(ageBand)ageBand(20,65]   -0.0185335  0.9816372  0.0019124 -9.691  < 2e-16 ***
tt(ageBand)ageBand(65,75]   -0.0072936  0.9927329  0.0018338 -3.977 6.97e-05 ***
tt(ageBand)ageBand(75,80]    0.0004933  1.0004934  0.0018448  0.267 0.789160
tt(ageBand)ageBand(80,85]    0.0046356  1.0046463  0.0017952  2.582 0.009815 **
tt(ageBand)ageBand(85,90]    0.0063279  1.0063480  0.0018329  3.452 0.000555 ***
tt(ageBand)ageBand(90,120]   0.0069859  1.0070103  0.0020653  3.382 0.000718 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                            exp(coef) exp(-coef) lower .95 upper .95
ageBand(20,65]                 2.0323     0.4920    1.8022    2.2918
ageBand(65,75]                 1.3301     0.7518    1.1801    1.4992
ageBand(75,80]                 0.9787     1.0218    0.8653    1.1068
ageBand(80,85]                 0.8665     1.1540    0.7681    0.9775
ageBand(85,90]                 0.8567     1.1673    0.7575    0.9688
ageBand(90,120]                0.9240     1.0822    0.8074    1.0575
tt(ageBand)ageBand(20,65]      0.9816     1.0187    0.9780    0.9853
tt(ageBand)ageBand(65,75]      0.9927     1.0073    0.9892    0.9963
tt(ageBand)ageBand(75,80]      1.0005     0.9995    0.9969    1.0041
tt(ageBand)ageBand(80,85]      1.0046     0.9954    1.0011    1.0082
tt(ageBand)ageBand(85,90]      1.0063     0.9937    1.0027    1.0100
tt(ageBand)ageBand(90,120]     1.0070     0.9930    1.0029    1.0111

Concordance= 0.559  (se = 0.003 )
Likelihood ratio test= 711.4  on 12 df,   p=<2e-16
Wald test            = 754.5  on 12 df,   p=<2e-16
Score (logrank) test = 774.4  on 12 df,   p=<2e-16
```

Figure 44 : Effet dépendant du temps dans le modèle de Cox

Les trois tests (Rapport de vraisemblance, Wald et Score) indiquent l'hypothèse de non-nullité des coefficients. La valeur de la p-value est inférieure à 5% pour les trois tests.

Concernant l'examen des coefficients du modèle, nous constatons que le test de significativité ne rejette pas l'hypothèse nulle au seuil de 5% des modalités : (75,80] et (90,120].

En ce qui concerne les modalités avec interaction temporelle, le test montre qu'elles sont significatives à l'exception de la modalité (75-80].

Nous constatons qu'une personne appartenant à la classe d'âge (80,85] a un risque de rester en dépendance plus élevé que celui d'une personne qui appartient à la tranche d'âge de référence, soit la tranche (-1,20] (toutes choses égales par ailleurs). Tenant en compte de l'interaction temporelle, ce coefficient est ajusté de 0.004*t, autrement le paramètre $\beta(t)$ est estimé par $\beta(t) = -0.14 + 0.004 \times t$, justifiant un impact négatif sur le risque de dépendance (maintien en dépendance plus long) pour les durées inférieures à 35 mois. De même, pour la tranche d'âge (85-90], le coefficient $\beta(t) = -0.15 + 0.006 \times t$ justifie un impact négatif sur le risque de dépendance pour les durées inférieures à 25 mois, cette observation est assez cohérente si on considère que plus un assuré est âgé moins sa probabilité de maintien est importante.

En revanche les coefficients obtenus pour les modalités (20,65] et (65,75] illustrent un risque de dépendance moins élevé par rapport à la modalité de référence (autrement une sortie plus rapide que la modalité de référence en gardant toute chose égale par ailleurs).

La pertinence de cette estimation peut être mise en évidence sur le graphique réalisé à partir des résidus de Schœnfeld

Vérification de l'hypothèse de proportionnalité à l'aide du test des résidus Schoenfeld

Les résultats ci-dessus ne rejettent pas l'hypothèse de risque proportionnel au seuil de 5%. C'est-à-dire, le modèle est au risque proportionnel.

```
> test.residu <- cox.zph(cox2_interaction)
> test.residu
                       chisq df    p
ageBand              1.94542  6 0.92
duration_adj         0.00057  1 0.98
ageBand:duration_adj 5.49887  6 0.48
GLOBAL               9.73351 13 0.72
> ggcoxzph(test.residu)
```

Figure 45 : Sortie R du test de vérification de l'hypothèse de proportionnalité

On peut opérer une vérification supplémentaire en regardant les résidus de Schoenfeld en fonction du temps comme sur la figure ci-dessous. Nous apercevons que la représentation des résidus est horizontale sur toute la période considérée pour la covariable ''ageBand''. L'hypothèse de proportionnalité est donc graphiquement vérifiée, et cela est confirmé par les p-values qui sont en effet supérieures à 5%,

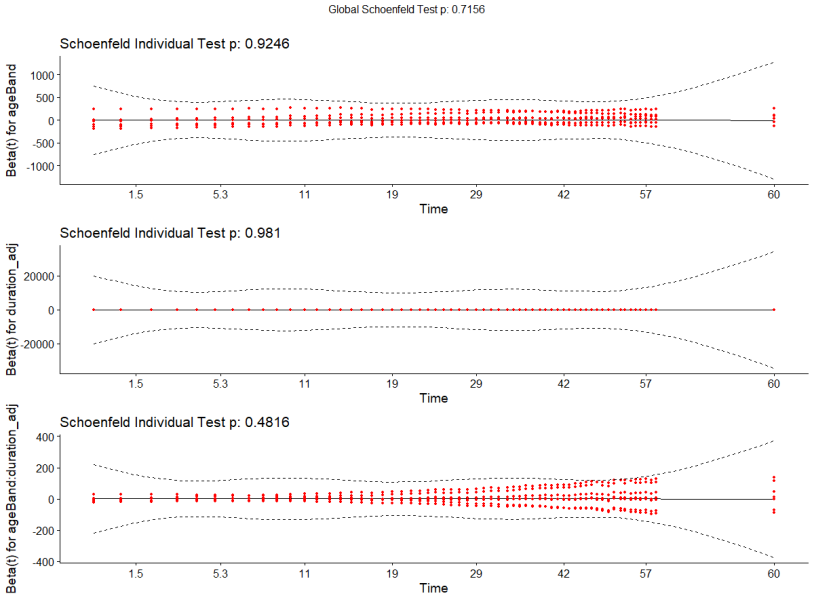


Figure 46 : Représentation graphique des résidus de Schoenfeld

Point d'attention :

Le résultat de la modélisation de l'effet temporel de la variable 'ageBand' semble raisonnable en comparaison avec un modèle sans interaction temporelle. Cependant, l'absence de significativité des modalités (75,80] et (90,120], pour lesquelles nous nous attendions à des conclusions statistiquement significatives, suggère que la forme fonctionnelle de β(t) n'est pas linéaire. Elle pourrait être mieux représentée par une fonction en escalier, c'est-à-dire avec des coefficients distincts pour chaque intervalle de temps. Il s'agit donc cette fois d'intégrer au modèle des interactions entre la variable et un indicateur propre à chaque intervalle, de manière à prendre en compte des variables évoluant dans le temps.

L'application de cette méthode est apparue complexe à mettre en œuvre et doit faire appel à des technique de simulation qui permettrons de trouver les intervalles de partitions de temps les plus adéquates.

Dans la littérature, il est jugé pertinent pour les variables qualitatives de procéder à un modèle stratifié pour corriger la non-proportionnalité que d'introduire une fonction d'interaction temporelle. Nous allons détailler ce modèle dans la section Modèle stratifié :

Remarque : Justification de la discrétisation de la variable âge.

Nous sommes conscients de la nécessité de justifier la discrétisation de la variable âge, la plupart des mémoires retiennent la variable quantitative ''âge''.

Le modèle de Cox suppose que chaque variable apporte une contribution linéaire au modèle, mais la relation peut parfois être plus complexe. Nous pouvons diagnostiquer ce problème graphiquement en utilisant des tracés résiduels.

La figure ci-dessous représente la variable âge continue par rapport aux résidus de martingale du modèle de Cox basé sur la variable âge brut (sans discrétisation en modalités). Cela permet de choisir la forme fonctionnelle de la covariable continue dans le modèle de Cox. Les lignes ajustées doivent être linéaires pour satisfaire aux hypothèses du modèle des risques proportionnels de Cox.

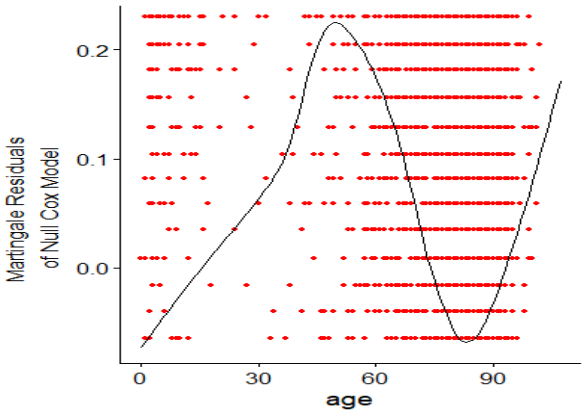


Figure 47 : Résidus de martingale en fonction de l'âge

D'après ce graphe nous constatons clairement une linéarité par morceaux, et ceci confirme le choix également des tranches que la cédante avait déterminée pour son modèle de tarification. On peut également penser à introduire une fonction polynomiale en âge pour décrire la relation entre l'âge et la fonction de hasard, mais par souci d'interprétabilité, nous avons retenu la discrétisation de la variable âge.

2.4. Le modèle avec 3 variables (''ageBand'', ''gender'' et ''type de service'').

2.4.1. Estimation d'un modèle de Cox avec les 3 covariables.

Nous avons vu précédemment que les trois types de variables : "ageBand", "gender" et '' 'type de service'' ont un impact significatif sur le maintien de l'individu en dépendance notamment les deux covariables sexe et tranche d'âge. Dans cette section nous avons donc décidé de construire un modèle avec les 3 covariables discriminantes.

La sortie R ci-dessous affiche l'estimation du modèle de Cox :

```
> cox <- coxph(Surv(duration_adj, status) ~ gender +ageBand+Last.Type.of.Services , data = data_etude)
> summary(cox)
Call:
coxph(formula = Surv(duration_adj, status) ~ gender + ageBand +
    Last.Type.of.Services, data = data_etude)

  n= 15116, number of events= 12220
   (5334 observations effacées parce que manquantes)

                                    coef exp(coef) se(coef)      z Pr(>|z|)
genderMale                      0.21813   1.24375  0.01836 11.881  < 2e-16 ***
ageBand(20,65]                  0.82279   2.27684  0.06266 13.131  < 2e-16 ***
ageBand(65,75]                  0.62408   1.86653  0.06214 10.043  < 2e-16 ***
ageBand(75,80]                  0.53604   1.70923  0.06278  8.538  < 2e-16 ***
ageBand(80,85]                  0.51996   1.68195  0.06205  8.380  < 2e-16 ***
ageBand(85,90]                  0.56309   1.75609  0.06272  8.978  < 2e-16 ***
ageBand(90,120]                 0.69551   2.00472  0.06533 10.646  < 2e-16 ***
Last.Type.of.ServicesHospitalisation 0.07824   1.08138  0.02510  3.117  0.00183 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                exp(coef) exp(-coef) lower .95 upper .95
genderMale                          1.244     0.8040     1.200     1.289
ageBand(20,65]                      2.277     0.4392     2.014     2.574
ageBand(65,75]                      1.867     0.5358     1.653     2.108
ageBand(75,80]                      1.709     0.5851     1.511     1.933
ageBand(80,85]                      1.682     0.5945     1.489     1.899
ageBand(85,90]                      1.756     0.5694     1.553     1.986
ageBand(90,120]                     2.005     0.4988     1.764     2.279
Last.Type.of.ServicesHospitalisation     1.081     0.9247     1.029     1.136

Concordance= 0.574  (se = 0.003 )
Likelihood ratio test= 400  on 8 df,   p=<2e-16
Wald test            = 390.9  on 8 df,   p=<2e-16
Score (logrank) test = 395.6  on 8 df,   p=<2e-16

> print(paste0("Le critère AIC = ",AIC(cox)))
[1] "Le critère AIC = 212635.10105571"
```

Figure 48 : Estimation du modèle de Cox pour les variables :''ageBand'', ''gender'' et ''type of service''

Toutes les variables sont significatives, ainsi elles sont discriminantes pour la loi de maintien. Nous constatons que les coefficients sont tous positifs indiquant que le risque de dépendance est plus fort par rapport à la variable de référence. Par exemple pour la variable ''gender'', la dépendance est plus forte chez les femmes que chez les hommes, toutes choses égales par ailleurs

Nous pouvons constater que la modalité ''ageBand'' (90,120] possède l'effet le plus élevé sur le taux de hasard. Ainsi, nous concluons les assurés les plus âgés qui appartiennent à la tranche d'âge (90,120] ont le moins risque de se maintenir en dépendance que ceux assurés des autres tranches d'âge (la courbe de survie décroît plus rapidement).

La variable ''type de service'' a un coefficient proche de zéro ($\boldsymbol{exp(\beta)} \cong \mathbf{1}$), ce qui montre que cette variable apporte une explication légère à la fonction de hasard. Ce constat a déjà été repéré.

2.4.2. Test de l'hypothèse de proportionnalité

Comme vu auparavant, l'hypothèse de hasards proportionnels suppose qu'avec des covariables constantes dans le temps, le ratio de risques entre deux personnes est aussi constant dans le temps. Dans le cas multivarié, cette hypothèse doit tenir pour toutes les covariables.

Pour tester cette hypothèse nous avons fait appel au test des résidus de Schoenfeld.

Vérification de l'hypothèse de proportionnalité à l'aide du test des résidus Schoenfeld

```
> cox.model
                       chisq df       p
gender                  5.65  1   0.017
ageBand               542.06  6 < 2e-16
Last.Type.of.Services  62.79  1 2.3e-15
GLOBAL                584.00  8 < 2e-16
>
```

Figure 49 : Sortie R du test de vérification de l'hypothèse de proportionnalité

Il apparaît que la p-value est inférieure à 5% globalement et pour chaque variable prise individuellement.

Toutes les p-values étant significatives nous acceptons l'hypothèse de non-indépendance entre le temps et les résidus de Schoenfeld.

On peut opérer une vérification supplémentaire en regardant les résidus de Schoenfeld en fonction du temps comme sur la figure ci-dessous.

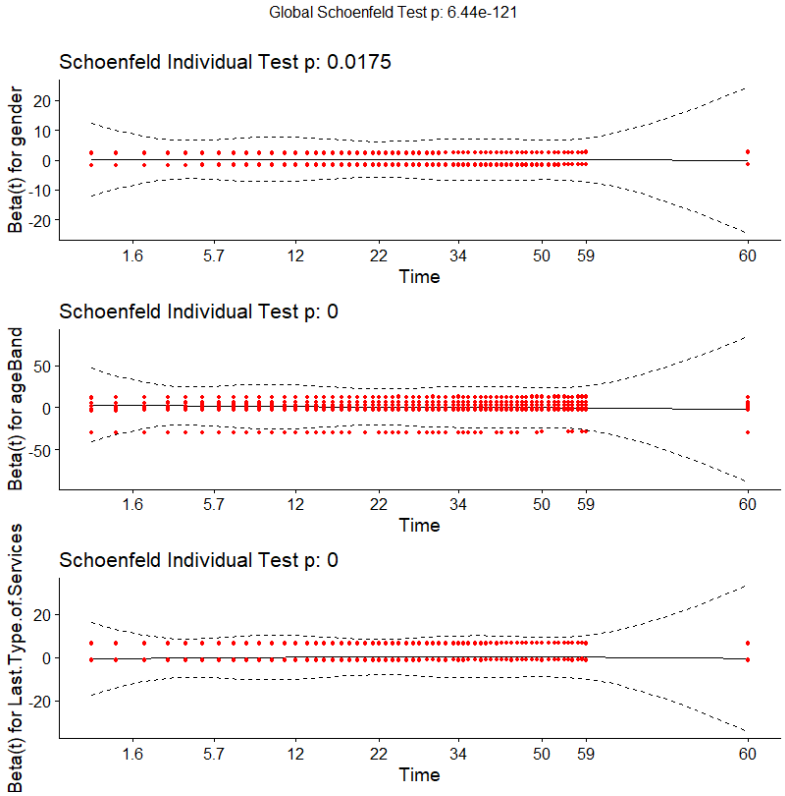


Figure 50 : Représentation graphique des résidus de Schoenfeld

Pour tenir en compte de la non-proportionnalité, nous avons envisagé de faire une starification selon la variable âge pour les raisons suivantes :

- L'âge est une variable très discriminante dans le modèle de cox non stratifié
- L'âge présente un caractère non linéaire
- La section précédente montrait qu'une stratification de la variable âge pourra être mieux adapté que l'introduction d'une variable temporelle.

2.4.3. Modèle stratifié :

Le principe est d'ajouter des strates au modèle de Cox afin que l'hypothèse de proportionnalité n'ait plus besoin d'être respectée pour la variable ageBand. Dans la pratique, cela signifie que la fonction de hasard de base ne sera pas la même selon que l'individu soit enfant, jeune ou vieux, mais dans ce cas l'effet de la variable tranche d'âge sur la probabilité de maintien en dépendance ne peut plus être estimé. Formellement, si l'on considère K groupes distincts, alors le modèle stratifié classique correspondant pose que :

$$h(t/\beta, X) = h_{0,k}(t) exp\left(\sum_{i=1}^{n} \beta_i x_i\right), k = 1, \dots, K$$

On notera que dans cette modélisation, on suppose que les autres covariables dans le modèle ont un effet multiplicatif constant sur le risque relatif à travers toutes les strates, contrairement à une analyse de sous-groupes complètement séparée.

Application du modèle de Cox stratifié

Sous R, l'estimation de ce modèle s'effectue en déclarant simplement *strata(var)* comme une variable exogène dans l'option *''formula''.*

Cas 1 : Les variables ''gender'' et ''type d'hospitalisation'' avec stratification par tranche d'âge :

Le résultat de l'application du modèle de Cox stratifié est obtenu ci-dessous :

```
> cox_strata1 <- coxph(Surv(duration_adj, status) ~ gender +Last.Type.of.Services + strata(ageBand), data = data_etude)
> summary(cox_strata1)
Call:
coxph(formula = Surv(duration_adj, status) ~ gender + Last.Type.of.Services +
    strata(ageBand), data = data_etude)

  n= 15116, number of events= 12220
   (5334 observations effacées parce que manquantes)

                                   coef exp(coef) se(coef)      z Pr(>|z|)
genderMale                      0.22976   1.25830  0.01839 12.493  < 2e-16 ***
Last.Type.of.ServicesHospitalisation 0.09510   1.09977  0.02512  3.785 0.000154 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                exp(coef) exp(-coef) lower .95 upper .95
genderMale                          1.258     0.7947     1.214     1.304
Last.Type.of.ServicesHospitalisation     1.100     0.9093     1.047     1.155

Concordance= 0.535  (se = 0.003 )
Likelihood ratio test= 166.8  on 2 df,   p=<2e-16
Wald test            = 167.8  on 2 df,   p=<2e-16
Score (logrank) test = 168.5  on 2 df,   p=<2e-16
```

Figure 51 : Sortie R pour le modèle de Cox stratifié

Usuellement, dans les modèles stratifiés, l'effet des covariables n'est pas différencié selon le groupe. Il est donc essentiel d'inclure dans le modèle une interaction entre chaque covariable et l'appartenance au groupe. Si les interactions entre toutes les covariables et la variable de stratification sont intégrées, les résultats correspondront à des estimations effectuées séparément pour chaque sous-population.

Sous R, il suffit de déclarer var1*strata(var2). Dans l'exemple ci-dessous, l'effet de la variable sexe et l'effet de la variable type de service sont supposés différer selon la tranche d'âge.

```
> cox_strata <- coxph(Surv(duration_adj, status) ~ gender+Last.Type.of.Services +Last.Type.of.Services*strata(ageBand) +gender*strata(ageBand), data = data_etude)
> summary(cox_strata)
Call:
coxph(formula = Surv(duration_adj, status) ~ gender + Last.Type.of.Services +
    Last.Type.of.Services * strata(ageBand) + gender * strata(ageBand),
    data = data_etude)

  n= 15116, number of events= 12220
   (5334 observations effacées parce que manquantes)

                                                        coef exp(coef) se(coef)      z Pr(>|z|)
genderMale                                           -0.17423   0.84010  0.12236 -1.424 0.154461
Last.Type.of.ServicesHospitalisation                 -0.21401   0.80734  1.00521 -0.213 0.831401
Last.Type.of.ServicesHospitalisation:strata(ageBand)(20,65] -0.16094   0.85134  1.01021 -0.159 0.873420
Last.Type.of.ServicesHospitalisation:strata(ageBand)(65,75]  0.05855   1.06030  1.00709  0.058 0.953636
Last.Type.of.ServicesHospitalisation:strata(ageBand)(75,80]  0.25095   1.28524  1.00694  0.249 0.803192
Last.Type.of.ServicesHospitalisation:strata(ageBand)(80,85]  0.39233   1.48043  1.00650  0.390 0.696687
Last.Type.of.ServicesHospitalisation:strata(ageBand)(85,90]  0.53007   1.69905  1.00668  0.527 0.598503
Last.Type.of.ServicesHospitalisation:strata(ageBand)(90,120] 0.61295   1.84588  1.00778  0.608 0.543039
genderMale:strata(ageBand)(20,65]                    0.13770   1.14763  0.13069  1.054 0.292048
genderMale:strata(ageBand)(65,75]                    0.37887   1.46063  0.12945  2.927 0.003425 **
genderMale:strata(ageBand)(75,80]                    0.45909   1.58264  0.13049  3.518 0.000434 ***
genderMale:strata(ageBand)(80,85]                    0.47487   1.60780  0.12896  3.682 0.000231 ***
genderMale:strata(ageBand)(85,90]                    0.57566   1.77831  0.13027  4.419 9.92e-06 ***
genderMale:strata(ageBand)(90,120]                   0.48236   1.61989  0.13525  3.566 0.000362 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                                     exp(coef) exp(-coef) lower .95 upper .95
genderMale                                              0.8401     1.1903    0.6610     1.068
Last.Type.of.ServicesHospitalisation                    0.8073     1.2386    0.1126     5.790
Last.Type.of.ServicesHospitalisation:strata(ageBand)(20,65]    0.8513     1.1746    0.1175     6.166
Last.Type.of.ServicesHospitalisation:strata(ageBand)(65,75]    1.0603     0.9431    0.1473     7.633
Last.Type.of.ServicesHospitalisation:strata(ageBand)(75,80]    1.2852     0.7781    0.1786     9.249
Last.Type.of.ServicesHospitalisation:strata(ageBand)(80,85]    1.4804     0.6755    0.2059    10.644
Last.Type.of.ServicesHospitalisation:strata(ageBand)(85,90]    1.6991     0.5886    0.2362    12.221
Last.Type.of.ServicesHospitalisation:strata(ageBand)(90,120]   1.8459     0.5417    0.2561    13.305
genderMale:strata(ageBand)(20,65]                       1.1476     0.8714    0.8883     1.483
genderMale:strata(ageBand)(65,75]                       1.4606     0.6846    1.1333     1.882
genderMale:strata(ageBand)(75,80]                       1.5826     0.6319    1.2255     2.044
genderMale:strata(ageBand)(80,85]                       1.6078     0.6220    1.2487     2.070
genderMale:strata(ageBand)(85,90]                       1.7783     0.5623    1.3776     2.296
genderMale:strata(ageBand)(90,120]                      1.6199     0.6173    1.2427     2.112

Concordance= 0.546  (se = 0.003 )
Likelihood ratio test= 309.2  on 14 df,   p=<2e-16
Wald test            = 314.1  on 14 df,   p=<2e-16
Score (logrank) test = 316.7  on 14 df,   p=<2e-16

>
```

Figure 52 : Estimation du modèle de Cox pour les variables, sexe, type d'hospitalisation stratifié selon la variable ageband

Les hypothèses de nullité globale des coefficients ainsi que de leur non-significativité sont donc rejetées. En ce qui concerne l'examen des coefficients du modèle, nous constatons que la p-value est supérieur à 5% pour la variable ''type de service'' justifiant le manque de significativité statistique pour cette variable.

Quant à la variable ''gender'', les assurés hommes ont un risque de dépendance moins élevé par rapport aux femmes, le risque de sortie de l'état de dépendance est plus élevé de 22.7% ($exp(37.9\% - 17.4\%) - 1$) pour la tranche d'âge (65,75], de 33.0% pour la tranche d'âge (75,80], de 35,1% pour la tranche d'âge (80,85], de 49.4% pour la tranche d'âge (85,90] et de 36.1% pour la tranche d'âge (90,120].

La dernière validation à effectuer est l'hypothèse de proportionnalité (la plus contraignante en général).

Comme pour les tests d'avant, nous avons obtenu résidus de Schoenfeld numériques et graphiques. Pour le test analytique, les résultats obtenus sont les suivants.

```
> cox.model <- cox.zph(cox_strata)
> cox.model
                                      chisq df       p
gender                                 2.19  1  0.1385
Last.Type.of.Services                 32.77  1 1.0e-08
Last.Type.of.Services:strata(ageBand) 72.26  6 1.4e-13
gender:strata(ageBand)                18.60  6  0.0049
GLOBAL                                94.90 14 4.5e-14
```

Figure 53 : Sortie R du test de vérification de l'hypothèse de proportionnalité

D'après le test analytique on constate que l'hypothèse des risques proportionnels n'est toujours pas vérifiée en raison de l'existence d'une corrélation entre les résidus et la durée de survie.

Donc nous avons décidé de ne pas tenir ce modèle en compte et de procéder au modèle stratifié sans la variable ''type de service'', à noter que la cette variable n'est pas statistiquement significative, en plus d'être une variable qui ne satisfait pas l'hypothèse de proportionnalité.

Cas 2 : de la variable ‘gender’ avec stratification par tranche d’âge :

La sortie du modèle R se présente comme suit :

```
> summary(cox_strata)
Call:
coxph(formula = Surv(duration_adj, status) ~ gender + strata(ageBand),
    data = data_etude)

  n= 20450, number of events= 17480

             coef exp(coef) se(coef)     z Pr(>|z|)
genderMale 0.21417   1.23884  0.01539 13.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
genderMale     1.239     0.8072     1.202     1.277

Concordance= 0.53  (se = 0.002 )
Likelihood ratio test= 193.1  on 1 df,   p=<2e-16
Wald test            = 193.8  on 1 df,   p=<2e-16
Score (logrank) test = 194.4  on 1 df,   p=<2e-16
```

Figure 54 : Sortie R du modèle Cox stratifié pour la variable ‘gender’

Les fonctions de survie correspondant à chaque groupe peuvent être représentées graphiquement.

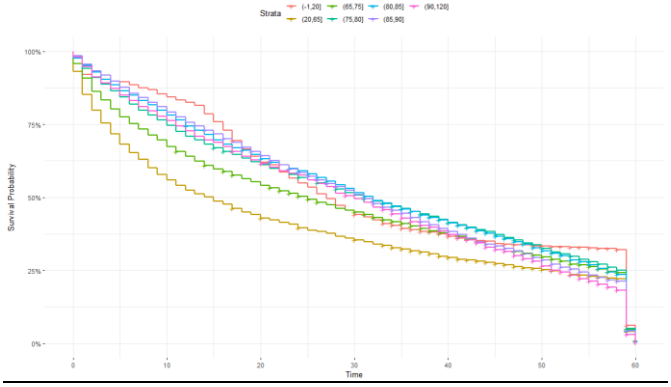


Figure 55 : La courbe de survie du modèle Cox-stratifié

Comme mentionné précédemment, les modèles stratifiés ne permettent pas de différencier l'effet des covariables entre les groupes. Dans de tels cas, il est nécessaire d'inclure dans le modèle l'interaction entre une covariable et l'appartenance à chaque groupe.

Dans notre cas, l’effet de la variable ‘gender’ est supposé différer selon l’appartenance à chaque tranche d’âge.

La sortie R ci-dessous nous permet de valider globalement le modèle (l’hypothèse de nullité globale des coefficients est rejetée au seuil de risque de 5%).

```
> cox_strata <- coxph(Surv(duration_adj, status) ~ gender+gender*strata(ageBand), data = data_etude)
> summary(cox_strata)
Call:
coxph(formula = Surv(duration_adj, status) ~ gender + gender *
    strata(ageBand), data = data_etude)

  n= 20450, number of events= 17480

                                    coef exp(coef) se(coef)      z Pr(>|z|)
genderMale                       0.06515   1.06732  0.07526 11.881  < 2e-16 ***
genderMale:strata(ageBand)(20,65] -0.11103   0.89491  0.08463 13.131  < 2e-16 ***
genderMale:strata(ageBand)(65,75]  0.10078   1.10603  0.08322  8.978  < 2e-16 ***
genderMale:strata(ageBand)(75,80]  0.19611   1.21666  0.08436  2.325 0.020083 *
genderMale:strata(ageBand)(80,85]  0.23588   1.26602  0.08264  2.854 0.004314 **
genderMale:strata(ageBand)(85,90]  0.30364   1.35478  0.08415  3.608 0.000308 ***
genderMale:strata(ageBand)(90,120] 0.20841   1.23172  0.09091  2.292 0.021881 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                   exp(coef) exp(-coef) lower .95 upper .95
genderMale                            1.0673     0.9369     0.9209     1.237
genderMale:strata(ageBand)(20,65]     0.8949     1.1174     0.7581     1.056
genderMale:strata(ageBand)(65,75]     1.1060     0.9041     0.9396     1.302
genderMale:strata(ageBand)(75,80]     1.2167     0.8219     1.0313     1.435
genderMale:strata(ageBand)(80,85]     1.2660     0.7899     1.0767     1.489
genderMale:strata(ageBand)(85,90]     1.3548     0.7381     1.1488     1.598
genderMale:strata(ageBand)(90,120]    1.2317     0.8119     1.0307     1.472

Concordance= 0.533  (se = 0.002 )
Likelihood ratio test= 270  on 7 df,   p=<2e-16
Wald test            = 273.5  on 7 df,   p=<2e-16
Score (logrank) test = 275.6  on 7 df,   p=<2e-16
```

Figure 56 : Estimation du modèle de Cox stratifié selon la variable ‘’ageband’’

Toutes les variables sont significativement différentes de 0 au seuil de risque de 5%, ce qui revient à dire que les variables choisies sont discriminantes pour la loi de maintien.

L'estimation des coefficients permet de conclure que les assurés hommes ont un risque de dépendance moins élevé par rapport aux femmes, le risque de dépendance est moins élevé de 18.0% ($exp(6.5\% + 10.1\%) - 1$) pour la tranche d'âge (65,75], de 29.9% pour la tranche d'âge (75,80], de 35,1% pour la tranche d'âge (80,85], de 44.6% pour la tranche d'âge (85,90] et de 31.5% pour la tranche d'âge (90,120].
On remarque ainsi que les hommes de la tranche d'âge (20,65] sortent moins rapidement de la dépendance que les femmes, toutes choses égales par ailleurs.
L'étape de validation à effectuer est l'hypothèse de proportionnalité, nous avons ci-dessous les résultats du test de Grambsch et Therneau et la représentation graphique des résidus :

```
> print(paste0("Le critère AIC = ",AIC(cox_strata)))
[1] "Le critère AIC = 249568.795891712"
> cox.model <- cox.zph(cox_strata)
> cox.model
                        chisq df     p
gender                   1.28  1 0.259
gender:strata(ageBand) 15.42  6 0.017
GLOBAL                 15.43  7 0.1385
```

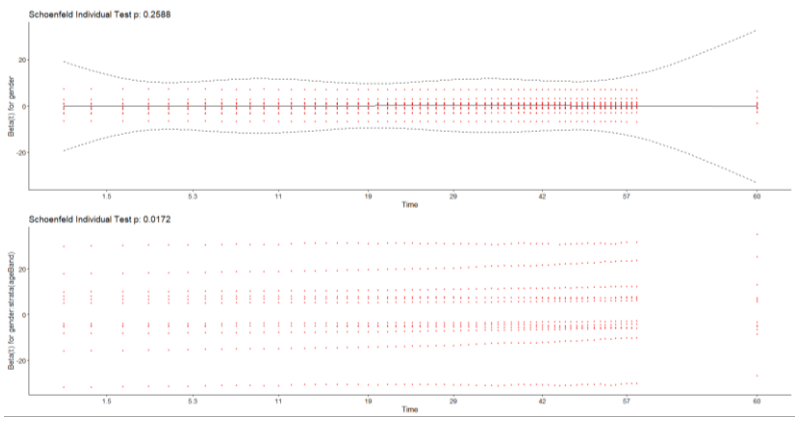


Figure 57 : Sortie R du test de vérification de l'hypothèse de proportionnalité et représentation graphique des résidus Schoenfeld

L'hypothèse des risques proportionnels est désormais vérifiée, et de plus nous jugeons une amélioration du modèle.

Remarques :

Pour réconforter notre postulat nous avons essayé de stratifier avec les 2 autres variables, sexe et type de service, or l'hypothèses de proportionnalité n'était pas respectée.

3. Conclusion et choix du modèle.

Après avoir testsé 6 modèles, nous avons décidé le modèle à retenir sur la base de critère AIC, le meuilleur modèle est celui qui minimise AIC, ainsi le modèle retenue est le modèle startifié par la variable age avec l'effet sexe.

Modèles :	AIC
Unqiuement la variable ''Gender''	315 210
Uniquement la varaiable ''Age Band''	315 191
Uniquement la varaiable ''type de service''	289 579
Les trois varaiable combinées	265 366
Modele stratifié par age, uniquement la varaible gender + type de service	252 134
Modele stratifié par age, uniquement la varaible gender	249 569

Tableau 8 : Comparaison de différents modèles sur la base du critère IC

Conclusion :

L'objectif de cette partie, était de mettre en place le modèle semi-paramétrique de Cox pour verifier et puis confrmer l'hétérogénéité existante au sein de notre portefeuille. En effet, plusieurs modèles ont été élaborés, dans un premier temps nous avons procédé au test de significativité des coefficients de regression et ensuite nous avons veillé sur la validité des hypothèse du modèle de Cox à l'aide des résidus de Schoenfeld (hypothèse de proportionalité des risques) qui est peu réalisé en pratique.
Pour pallier à cette violation d'hypothèse de proportionnalité nous avons proposé deux extensions du modèle de Cox, un modèle dont les régresseurs dépendant du temps et un modèle de Cox stratifiés.
Ainsi, sur la base de critère AIC, le modèle retenu à la fin de notre analyse se base sur le modele de Cox stratifié par age avec une seule varaible descriminante sexe.

IV. Modèle multi états

1. Contexte général de l'étude.

Les modèles multi-états sont une généralisation des modèles du survie, caractérisés par un processus stochastique à espace fini d'états utilisé pour décrire l'évolution des individus à travers différents états dans le temps.
Cette section aura pour objectif principal de présenter les modèles multi-états markoviens et semi markoviens en développant la théorie et en l'appliquant à notre base de données.
Dans un premier temps, nous présenterons la méthode relative au modèle de Markov homogène, ce modèle est le plus simple, il suppose que les intensités de transition entre les états sont constantes dans le temps. Nous allons présenter les aspects théoriques indispensable pour présenter ensuite le modèle semi-markovien, dans ce modèle les intensités de transition dépendent de temps de séjour dans l'état présent (la durée passée dans l'état).
La modélisation multi-états consiste à définir trois états dans lesquels l'individu assuré peut se trouver : l'état actif/sain, l'état de dépendance et un état qui conduit à la sortie de la dépendance autre que l'état sain. L'individu ne peut se trouver que dans un seul des trois états précédents.

Remarque :
Dans l'objectif de réduire la volumétrie du modèle, l'étude des données nous conduit à n'envisager que certaines transitions au cours de la survie en dépendance, correspondant à celles qui sont les plus observées et significativement plus élevées que les autres. À ce titre, seules les transitions suivantes sont considérées

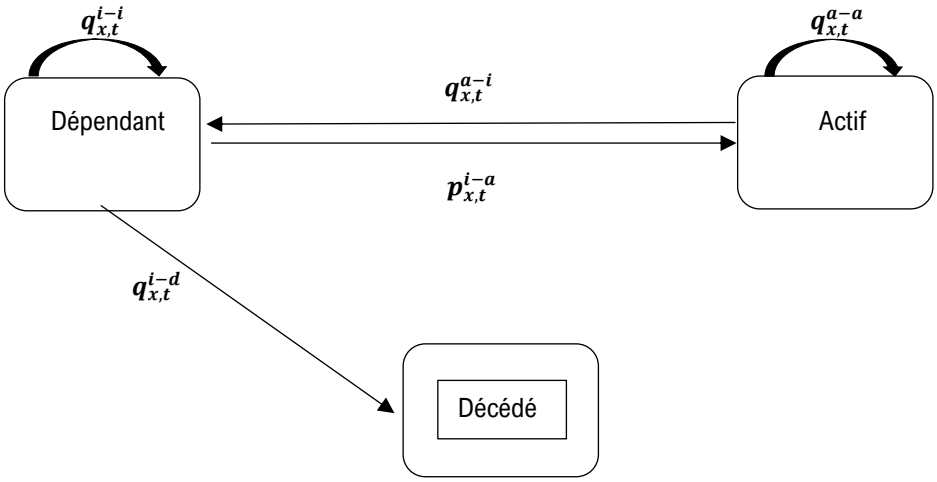


Figure 58 : Etats relatifs au risque dépendance du produit ''Silver''

Cinq paramètres sont à estimer dans ce modèle : la probabilité qu'un individu dépendant devient actif $\boldsymbol{p_{x,t}^{i-a}}$, la probabilité de décéder sachant que l'individu est dépendant $\boldsymbol{q_{x,t}^{i-d}}$, la probabilité qu'un individu soit reconnu comme actif revient à l'état dépendant $\boldsymbol{q_{x,t}^{a-i}}$, la probabilité qu'un individu reste dépendant $\boldsymbol{q_{x,t}^{i-i}}$ et la probabilité qu'un individu actif reste dans le même état $\boldsymbol{q_{x,t}^{a-a}}$.
Ce schéma montre que les probabilités de transition peuvent être abordées à l'aide d'un processus semi-Markovien non homogène délicat à modéliser car la probabilité de transition entre les états évolue dans le temps et dépend de plusieurs paramètres, tel que l'âge de l'assuré, le sexe, le type de soin etc. Néanmoins nous trouvons des études qui supposent une homogénéité locale, l'hypothèse sous-jacente est de considérer que les taux de transition sont constants par morceaux.
Dans le présent mémoire, nous allons nous placer dans un cadre multi-états markovien homogène avec introduction des co-variables : âge et sexe (modèle markovien homogène par morceaux). En d'autres termes, nous allons starifier l'ensemble de donnée pour observer l'effet du sexe et de l'âge sur l'état futur du processus qui ne dépend que de l'état actuel et non du temps passé dans cet état.

Nous allons dans un deuxième temps présenter le modèle semi-markovien homogène par morceaux dont la loi de durée de transition entre états sera estimée par la loi de Weibull, qui est une généralisation de la loi exponentielle classique de la modélisation des lois de survie.

Remarque :
Le volume limité de données concernant la transition de l'état de dépendance à l'état actif aura un impact significatif sur l'estimation du paramètre de taux d'intensité pour cet état.

Afin de comprendre la nature des probabilités de transition d'un modèle multi-états, il est primordial d'avoir une compréhension complète de l'ensemble des éléments de la théorie markovienne

2. Le Cadre théorique du modèle markovien :

Les processus markoviens sont très utiles dans la construction de modèles de durée en assurance. A l'aide de ces modèles nous souhaitons étudier les processus d'évolution de l'état de dépendance.

2.1. Généralité

L'état occupé par l'assuré est représenté à partir d'un processus stochastique $(X_t)_{t\geq 0}$ càdlàg (continue à droite et avec limite à gauche) à valeur dans un espace d'états fini $S = \{e_1, e_2, \dots, e_m\}$ que l'on appelle processus multi-états. Ce processus est défini sur un espace probabilisé $(\Omega; \mathrm{A}; \mathbb{P})$ on note sa filtration engendrée $(\mathcal{F}_t)_{t\geq 0}$
Le processus $(X_t, \mathcal{F}_t)_{t\geq 0}$ est markovien s'il satisfait à la propriété de Markov, c'est-à-dire s'il vérifie l'égalité suivante pour $0 \leq s \leq t$ et pour tout $e \in S$

$$P(X_t = e/\mathcal{F}_t) = P(X_t = e/X_s)$$

Cette définition signifie que le processus dispose d'une propriété « d'oubli » qui peut s'interpréter comme le fait qu'un événement futur ne dépend pas du passé si le présent est connu.
Le processus de Markov se caractérise par les probabilités de transition entre tous les états possibles h et j dans l'espace des états :

$$S \times S \to [0,1];\ (h,j) \longmapsto P(X_t = j/X_s = h)$$

En pratique, nous utiliserons la notation matricielle $p(s,t) = \left(p_{h,j}(s,t)\right)_{h,j\in S}$ telle que :
$p_{h,j}(s,t) = P(X_t = j/X_s = h)$ avec $p(s,s) = \mathbb{1}_d$, correspondant à la matrice identité $\sum_{j\in S} p_{hj}(s,t) = 1$
Le calcul des probabilités de transition entre deux états est susceptible d'être effectué à partir de la propriété de Chapman-Kolmogorov, qui permet d'introduire un état intermédiaire dans l'expression de ces probabilités. Cette relation s'écrit sous forme matricielle pour tout $0 \leq s \leq u \leq t$

$$p(s,t) = p(s,u) \times p(u,t)$$

Cette propriété est extrêmement utile dans la mesure où elle permet de calculer toute matrice des transitions du triplet $(P(s,u), P(s,t), P(t,u))$ connaissant les deux autres. C'est l'une des raisons importantes de la facilité d'utilisation prédictive des modèles de Markov, et qui font tout leur intérêt.

2.2. Le lien avec les modèles de survie :

Il est nécessaire dans le cadre de la construction de processus markovien de définir des intensités de transition (ou taux instantanés de transition) entre les états $\forall t \geq 0\ et\ pour\ tout\ h\ et\ j\ \in S$

$$\mu_{hj}(t) = \lim_{\Delta t \to 0} \frac{p_{hj}(t, t+\Delta t)}{\Delta t}\ pour\ h \neq j\ et\ \mu_{hh}(t) = -\sum_{j\neq h} \mu_{hj}(t)$$

On définit la matrice d'intensité de transition markovienne au temps $t \geq 0$ par $\mathbb{Q}(t)$.
Les éléments de cette matrice sont définis par :

$$q_{hj}(t) = \delta_{h,j}\, \mu_{hh}(t) + \left(1 - \delta_{h,j}\right)\mu_{hj}(t)$$

Avec $\delta_{h,j} = 1\ si\ h = j\ sinon\ \delta_{h,j} = 0\ si\ h \neq j$
$q_{hj}(t)$ peut être interprété comme le risque instantané de transiter de l'état h à l'état j à l'instant t :

- Les termes diagonaux sont négatifs, ils décrivent l'envie de sortir de l'état h
- Les termes non diagonaux sont positifs ; ils décrivent l'envie d'arriver en j depuis h
- La somme des lignes est égale à 0.

On peut les voir comme la « vitesse » de l'évènement qui fait passer le processus de l'état h à j.

Les fonctions d'intensité cumulée de transition correspondantes sont alors obtenues pour tout $t \geq 0$ et $h, j \in S$

$$A_{hj}(t) = \int_0^t \mu_{hj}(u)du \ \ pour\ h \neq j\ et\ A_{hh} = \sum_{j \neq h} A_{hj}(t)$$

Les grandeurs introduites fournissent une généralisation de la fonction de hasard et de la fonction de hasard cumulée utilisées dans les modèles de durée classiques.

Intéressons-nous maintenant au lien entre intensités et probabilités de transitions. Les équations de Chapman-Kolmogorov et la définition des intensités de transition permettent à présent d'obtenir les équations intégro-différentielles forward de Kolmogorov, exprimées sous forme matricielle, pour tout $0 \leq s \leq t$

$$p(s,s) = \mathbb{1}_d\ et\ \frac{\partial p(s,t)}{\partial t} = p(s,t)\mathbb{Q}(t)$$

Où $\mathbb{Q}(t)$ comme défini précédemment correspond à la matrice des fonctions d'intensité de transition en t. De manière symétrique mais en s'intéressant cette fois à l'instant initial, il est possible d'établir les équations backward de Kolmogorov, associées aux processus X_t pour tout $0 \leq s \leq t$

$$p(s,s) = \mathbb{1}_d\ et\ \frac{\partial p(s,t)}{\partial s} = -p(s,t)\mathbb{Q}(s)$$

Les équations forward de Kolmogorov admettent, dans le cas d'un processus markovien, une unique solution, de plus, si nous supposons que les fonctions d'intensité cumulée de transition $A(t) = \left(A_{hj}(t)\right)_{h,j \in S}$ sont constantes par morceaux, l'expression des probabilités de transition prend la forme d'un simple produit fini sur les temps de sauts $s = t_0 \leq t_1 \leq \ldots \leq t_K = T$ de A, la solution s'exprime sous la forme suivante :

$$p(s,t) = \prod_{k=1}^{K} \left(\mathbb{1}_d + \Delta A(t_k)\right)$$

Avec : $\Delta A(t_k) = \mathrm{A}(t_k) - \mathrm{A}(t_{k-1})$

Lorsque les intensités de transition, supposées constantes, sont connues pour chaque segment $\mu_1, \ldots, \mu_K$, chaque terme du produit s'obtient sous la forme d'une exponentielle de matrice :

$$\left(\mathbb{1}_d + \Delta A(t_k)\right) = exp\left(\mu_k(t_k - t_{k-1})\right)$$

2.3. Le processus Markovien homogène :

Pour tout $0 \leq s \leq t, P(s,t) = P(0, t-s)$ on notera alors simplement $P(t-s) = P(0, t-s)$, on dit que les probablités de transition sont stationnaire, ainsi : $P(x) = exp(Q \times x)$

Dans le cas particulier où les probabilités de transition ne dépendraient pas de t et de s mais uniquement de $t - s$, le processus markovien associé est qualifié de processus homogène, en d'autres termes le processus est indépendant du temps local (temps passé dans l'état) et global (l'âge de l'individu).

Dans le cas homogène on peut calculer explicitement les probabilités de transition.

En effet, pour une matrice diagonalisable dans $\mathbb{R}$, l'exponentielle de matrice s'exprime de manière analytique à l'aide d'un produit d'une matrice diagonale dont les termes correspondent aux exponentielles des valeurs propres avec des matrices de passage.

Pour les matrices non diagonalisables, on peut utiliser la décomposition en matrices de Jordan pour se ramener au cas diagonalisable.

Pour tout h, le temps de séjour dans l'état h avant d'effectuer une transition vers un autre état est une variable aléatoire de loi exponentielle.

2.4. Le processus Markovien non homogène :

Bien que dans ce contexte, l'étude des processus markoviens soit simplifiée, cette hypothèse que les probabilités de transition ne dépendraient pas de t et de s mais uniquement de $t - s$ n'est pas valide dans la plupart des applications actuarielles, les probabilités de transition variant par exemple avec l'âge. Par conséquent, le recours aux processus inhomogènes semble nécessaire. Dans ce processus les probabilités de transition dépendent uniquement de temps et non de la durée passée dans l'état.

Ainsi, la solution des équations de Kolmogorov s'écrit :

Pour tout $0 \leq s \leq t$: $P(x) = exp\left(-\int_s^t Q(u)\, du\right)$.

3. Cadre théorique du modèle semi-markovien :

3.1. Généralités :

Comme nous l'avons mentionné, le modèle de Markov homogène possède un caractère trop restrictif pour décrire les risques biométriques couvert par les contrats d'assurance de dépendance. Lorsqu'un ou plusieurs états correspondent à une situation de fragilité (p. ex. maladie, accident) de l'assuré, il est généralement nécessaire de faire intervenir le temps de séjour dans ces états. Cette spécificité est alors prise en compte à partir de modèles qualifiés de semi-markoviens.

On définit le modèle semi-markovien homogène comme suit :

S_k correspond, pour tout $k \in \mathbb{N}^*$, à la date du k ^{ème}saut du processus $(X_t)_{t\geq 0}$

$$S_k = inf\{t > S_{k-1}/X_t \neq X_{S_{k-1}}\}$$

J_k correspond, pour tout $k \in \mathbb{N}^*$, à l'état du processus $(X_t)_{t\geq 0}$ entre les dates S_k et S_{k+1}

Le processus $(S_k; J_k)_{k\geq 0}$ permet de notifier les temps de saut ainsi que l'état occupé entre chaque saut, et fournit une représentation complète de la trajectoire suivie par l'état de l'individu.

Le nombre de transitions $h \to j$ en date $t \geq 0$ est enregistré à partir du processus de comptage $N_{hj}(t)$, N_t correspond au nombre total de transitions effectuées à cette date. L'introduction de ce dernier processus de comptage permet de relier les processus $(S_k; J_k)_{k\geq 0}$ et $(X_t)_{t\geq 0}$ en notant $X_t = J_{N(t)}$. La durée de séjour dans l'état courant à une date $t \geq 0$ est par ailleurs donnée par $U_t = t - S_{N(t)}$.

Le processus $(X_t)_{t\geq 0}$est semi_markovien s'il vérifie l'égalité suivante, pour $k \in \mathbb{N}^*$ et pour tout $C \in [0, +\infty[\times S$

$$P((S_{k+1}; J_{k+1}) \in C/(S_l; J_l);\ l = 1,,,,k) = P((S_{k+1}; J_{k+1}) \in C/(S_k; J_k))$$

Cette condition signifie que l'évolution future d'un processus semi-Markovien (temps de séjour dans l'état présent et état suivant) ne dépend que de l'état présent de celui-ci, et non des états antérieurs ou temps de séjour en ces derniers (Figure ci-dessous) :

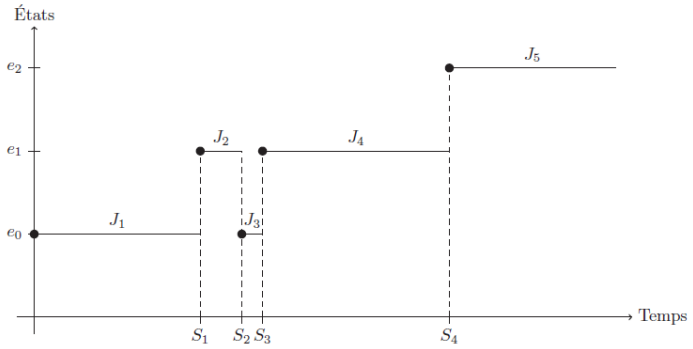


Figure 59 : Exemple de chemin parcouru au cours de la durée de vie d'un individu pour un modèle multi-etats quelconque a 3 états {e0, e1, e2}.

- **<u>Noyau semi-markovien</u>**

On définit le <u>noyau semi-markovien</u> par :

$$Q_{hj}(s,u) = P\left(\Delta S_{N(s)+1} \leq u,\ J_{N(s)+1} = j/\left(S_{N(s)}; J_{N(s)}\right) = (s,h)\right)$$

Avec $\Delta S_{k+1} = S_{k+1} - S_k$, Dans cette définition, le noyau semi-markovien (NSM) dépend explicitement de deux échelles temporelles : la date s et une durée u. S'il ne dépend pas de la date s, le processus $(X_t)_{t\geq 0}$ sera qualifié de semi-markovien homogène et $(S_k; J_k)_{k\geq 0}$ de processus de renouvellement markovien homogène.

La matrice Q porte le nom de noyau semi-markovien car sa seule définition permet de caractériser complètement le processus et, en particulier, de retrouver toutes les lois marginales qui lui sont associées.

Le noyau semi-markovien correspond à la probabilité sachant qu'on est entré dans l'état h que la prochaine transition ait lieu vers l'état j, avant qu'une durée u ne se soit écoulé

- **Quantité d'intérêt :**

Les processus $(S_k; J_k)_{k\geq 0}$ et $(X_t; U_t)_{t\geq 0}$ peuvent être exprimés en fonction de grandeurs d'intérêt qui permettent de procéder à l'estimation du modèle.
Lorsque l'on s'intéresse au processus $(X_t; U_t)_{t\geq 0}$, on se réfère aux probabilités de transition, définies pour tout $0 \leq u \leq s \leq t, v \geq 0$ et pour tout $h, j \in S$ par :

$$p_{hj}(s,t,u,v) = P(X_t = j, U_t \leq v / (X_s; U_s) = (h,u))$$

Ces probabilités permettent de caractériser le processus semi-markovien $(X_t; U_t)_{t\geq 0}$, et constituent les quantités que l'on cherchera à exprimer. Pour des applications actuarielles, on introduit également les probabilités associées au prochain changement d'état.

$$\overline{p_{hj}}(s,t,u) = P\left(X_{S_{N(s)+1}} = j, S_{N(s)+1} \leq t / (X_s; U_s) = (h,u)\right); h \neq j \; et$$

$$\overline{p_{hh}}(s,t,u) = P\left(X_{S_{N(s)+1}} \leq t / (X_s; U_s) = (h,u)\right)$$

Le processus $(S_k; J_k)_{k\geq 0}$ quant à lui, permet d'introduire la probabilité de saut $h \to j$ en date s : la loi de survie marginale est égale à la probabilité de passer à l'état j sachant le processus est entré à l'état h à l'instant s :

$$r_{hj}(s) = P\big(J_{N(s)+1} = j / \big(S_{N(s)}; J_{N(s)}\big) = (s,h)\big)$$

La matrice $R = \left(r_{hj}(s)\right)_{(h,j)\in S^2}$ est une matrice de transitions markovienne non homogène au sens de la définition d'un processus de Markov. Elle ne dépend que du temps d'entrée dans l'état courant. Cette dernière équation fait apparaître le processus de Markov comme cas limite d'un processus semi-markovien et identifie, en ce sens, l'approche semi-markovienne comme une généralisation de la modélisation markovienne.
On définit les fonctions de répartition du temps de séjour, connaissant la date de dernier saut et la prochaine transition :

$$K_{hj}(s,u) = P\big(\Delta S_{N(s)+1} \leq u / J_{N(s)+1} = j, J_{N(s)} = h, S_{N(s)} = s\big)$$

La probabilité de rester dans un état h pendant une durée au plus égale à u en date s est ensuite définie par

$$H_h(s,u) = P\big(\Delta S_{N(s)+1} \leq u / J_{N(s)} = h, S_{N(s)} = s\big) = \sum_{j\neq h} Q_{hj}(s,u)$$

Sous l'hypothèse d'indépendance de durées de transition.

$$Q_{hj}(s,u) = K_{hj}(s,u) \times p_{hj}$$

La fonction de survie associée est donnée par :

$$S_{hj}(s,u) = 1 - K_{hj}(s,u)$$

Sous la condition d'hypothèse d'indépendance on peut écrire :

$$S_h(s,u) = \sum_{h\neq j} S_{hj}(s,u) \times p_{hj}(s)$$

En pratique nous utiliserons la densité de probabilité associée à la fonction de survie S_{hj} donnée par

$$f_{h,j} = F'_{h,j} = -S'_{h,j}$$

Ces expressions permettent de jeter les bases de la modélisation semi-markovien que nous souhaitons mettre en œuvre.

- **Lois de durées des séjours dans un état donné**

Les temps de séjour dans l'état h avant le transfert vers l'état j peuvent être distribués selon une loi continue quelconque. En effet, nous avons vu pour le cas d'un processus markovien que la classe des lois exponentielles sont naturellement liées à ce processus. Dans ce mémoire nous allons procéder par le choix d'une loi plus générale : Weibull. Nous avons discrétisé la loi exponentielle car elle ne comporte pas suffisamment de paramètres pour prendre en compte la complexité du processus de survie.

3.2. Processus semi-markovien non homogène :

Nous nous plaçons dans un cadre multi-états semi-markovien inhomogène. Cela signifie que les probabilités de transition dépendent du temps (e.g. âge de l'assuré), et non seulement de la durée passée dans l'état (durée de transition vers un autre état). Dans le présent mémoire, nous ne possédons pas assez d'éléments pour opter une modélisation semi-markovienne non homogène, mais nous pourrions par exemple supposer que l'influence de

l'âge sur les probabilités de transition est linéaire ou de représenter les probabilités de passage par la fonction Logit pour borner ses valeurs entre 0 et 1, tel que : $p_{h,j} = \frac{e^{a_{h,j} \times age + b_{h,j}}}{1 + e^{a_{h,j} \times age + b_{h,j}}}$

3.3. Calibrage du modèle semi markovien.

Le modèle décrivait dans la section précédente permet de générer l'évolution probable d'une personne en dépendance à travers les états jusqu'au le décès. La méthode la plus simple sera de faire appel aux techniques de simulation de Monte Carlo. Celle-ci consiste à simuler un très grand nombre de trajectoires avec les mêmes conditions initiales puis à moyenner les résultats obtenus pour évaluer un comportement central.
Plus les trajectoires simulées sont nombreuses et plus l'image du processus correspondant à notre modèle sera complète et fidèle.
L'algorithme utilisé est très simple.

1. Nous prenons notre portefeuille initial constitué des personnes en état de dépendance, soit l'état h et on génère l'état suivant, j à l'aide des probabilités estimées des passages $\left(p_{h,j}\right)_{h \neq j}$
2. A partir de l'état initial h et l'état suivant j, nous générons le temps de passage correspondant par une réalisation d'une loi de Weibull simple.
3. Si le dernier état visité ne s'agit pas d'un état de sortie (décès), nous répétons les étapes 1 puis 2 ;

La complexité de cet algorithme réside dans son implémentation dans le logiciel de modélisation, R dans notre cas de figure.

Conclusion :

La construction de lois d'expérience en best estimate se devrait de tenir compte de l'interaction entre les différents états, sous peine de générer un biais dans l'estimation des engagements. Chaque changement d'états donne lieu à des flux financiers différents pour le réassureur. D'où l'objectif de cette section qui nous permettra de tenir en compte ces interactions en s'appuyant sur l'utilisation de modèles multi-états qui permettent d'avoir une représentation détaillée de chaque loi de transition. 3 états ont été identifiés :

- Actif/Valide/Sain
- Dépendance/perte d'autonomie
- Décès

Toutefois, on distingue plusieurs types de modèles multi-états :

- Markovien homogène (MH) : les intensités de transition sont constantes ;
- Markovien non-homogène (MNH) : les intensités de transition dépendent du temps ;
- Semi-markovien homogène (SMH) : les intensités de transition dépendent de la durée passée dans l'état
- Semi-markovien non-homogène (SMNH) : les intensités de transition dépendent de la durée passée dans l'état et du temps ;

Dans le cadre de notre mémoire, les modèles : markovien homogène par morceau et Semi-markovien homogène par morceau constituent les classes de modèles retenus que nous croyons à la fois réaliste et fructueux. Nous souhaitons mentionner que ces approches relèvent d'une complexité en termes de développement mathématique ainsi qu'en termes d'implémentation.

4. Cadre pratique : construction du modèle

Ce chapitre décrit l'estimation des paramètres de modèle exposé dans la précédente section. Nous rappelons le lecteur qu'on est dans logique de run-off, ainsi le portefeuille contient que des assurés en cas de dépendance, autrement notre point de départ est constitué des assurés dépendants dont on suivra l'évolution des états par une approche Markovienne.
La modélisation est effectuée à l'aide des fonctions du package ''*msm*'' (Multi-state Models).
Pour l'application de l'analyse markovienne, il est nécessaire de formater les données de la façon suivante :

ID	Mois	Etat	Sexe	Date de sinistre	Type de service
1	1	1	Homme	12/03/2016	Maison
1	2	1	Homme	12/03/2016	Maison
1	3	2	Homme	12/03/2016	Maison
2	1	1	Femme	31/06/2017	Centre d'hospitalisation
2	2	1	Femme	31/06/2017	Centre d'hospitalisation
2	3	1	Femme	31/06/2017	Centre d'hospitalisation
2	4	3	Femme	31/06/2017	Centre d'hospitalisation
…	…	…	…	…	…
3000	1	1	Femme	15/08/2015	Maison
3000	2	2	Femme	15/08/2015	Maison
3000	3	2	Femme	15/08/2015	Maison
3000	4	1	Femme	15/08/2015	Maison

Tableau 9 : Exemple d'une base de données.

En construisant notre jeu de données pour cette application, nous faisons hypothèse forte sur les covariables, celles-ci n'évoluent pas au cours des mois.
La méthodologie suivie pour la détermination du nombre de transition est simple, elle se base sur les bordereaux reçus de la cédante, si l'assuré n'a pas changé d'état entre deux bordereaux consécutifs donc l'état est maintenu ainsi le nombre de mois de séjour dans cet état sera la différence entre la date d'entrée dans l'état et la date de réception du bordereau, à noter que l'évènement de décès, s'il survient, est bien reporté à chaque date de clôture. Dans de nombreux cas, on n'observe pas de changement d'état de dépendance (transition i → i).
Notons qu'entre deux bordereaux consécutifs, il peut y avoir des changements d'état qui ne sont pas observés par la cédante et qui ne sont pas renseignés dans la base de données.
La sortie R ci-dessous décrit la représentativité des transitions observées et renseignées dans la base de données

```
> View(data_msm3)
> statetable.msm(Status._inverse, Numerator, data=data_msm3)
    to
from      1      2      3
   1 337579     78   6884
   2    355   9389      0
>
```

1 : état de dépendance, 2 : état valide et état 3 : décès.

4.1. Modèle Markovien homogène :

L'estimation de la matrice d'intensité de transition Q est effectuée à l'aide de l'estimateur du maximum de vraisemblance dont les résultats sont obtenus par la fonction « msm » :

```
> cav.msm

Call:
msm(formula = Status._inverse ~ Time, subject = Numerator, data = data_msm1,     qmatrix = Q, deathexact = 3, control = list(fnscale = 4000))

Maximum likelihood estimates

Transition intensities
                  Baseline
State 1 - State 1 -0.02071 (-0.02120,-0.02022)
State 1 - State 2  0.02071 ( 0.02022, 0.02120)
State 2 - State 1  0.03664 ( 0.03302, 0.04066)
State 2 - State 2 -0.57768 (-0.59132,-0.56436)
State 2 - State 3  0.54104 ( 0.52826, 0.55414)

-2 * log-likelihood:  93770
>
```

Figure 60 : Sortie R de la matrice d'intensité de transition estimée par la fonction « msm »

Le temps moyen de séjour dans un état r est déterminé par la fonction « sojourn.msm » : le temps moyen est estimé par $1/q_{rr}$, où q_{rr} est le terme diagonal de la matrice d'intensité de transitions estimée. Par exemple une

personne dépendante reste en moyenne dans cet état pendant 48.3 mois ($= |\frac{1}{-0.02071}|$) avant qu'il change d'état.

```
> sojourn.msm(cav.msm)
         estimates     SE      L      U
State 1    48.296 0.5797 47.173 49.446
State 2     1.731 0.0206  1.691  1.772
>
```

Figure 61 : Sortie R, temps moyen de séjour dans un état

« se » correspond à l'écart type, « L » et « U » sont les bornes de l'intervalle de confiance.

4.1.1. Construction de la courbe de survie et des probabilité de transition :

Courbe de survie

L'estimation de la matrice de transition permet de calculer les probabilités de transition et par conséquent tracer une courbe de survie pour chaque état :

La courbe rouge reflète l'évolution de la courbe de survie chez la classe des assurés dépendants, tandis que la courbe bleue est relative au statut ''valide''. D'après le graphe présenté ci-dessus, nous constatons que les courbes de survie décroisent exponentiellement, plus vite chez les assurés en état ''valide'', ceci s'explique par le retour rapide à l'état dépendant (donc une sortie de l'état valide). On note également que l'allure de la courbe est fortement sensible au nombre de transition de cet état, comme mentionné, nous avons un nombre faible de données pour les cas ''valides'' ce qui rend sa lecture non représentative et affaiblie significativement sa pertinence.

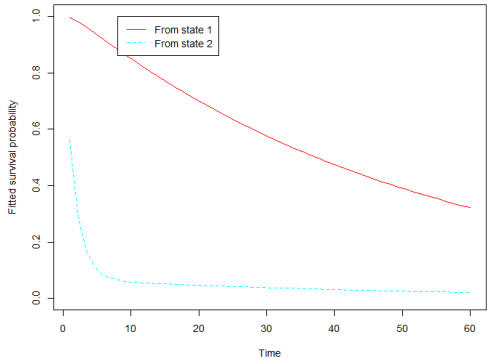


Figure 62 : La fonction de survie pour les deux états '' dépendant'' et ''valide''.

Probabilité de transition de l'état dépendant :

La fonction *« pmatrix »* du package *msm* extrait les probabilités de transition estimées à un temps donné par la fonction *msm*. De fait, il est alors possible de présenter les courbes des probabilités de transition et de passage au cours du temps.

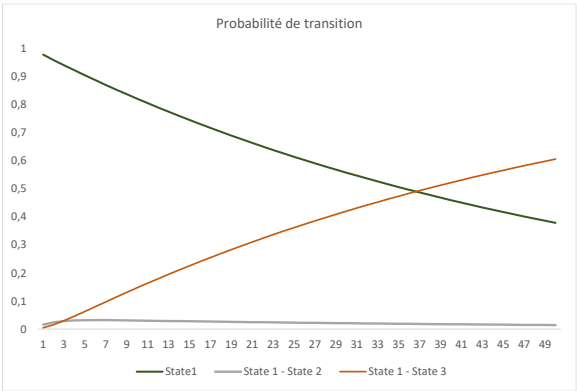


Figure 63 : Probabilité de transition de l'état dépendant

D'après le graphique ci-dessous, nous remarquons que les probabilités de maintien dans un état de dépendance décroissent tandis que les probabilités de transition vers le décès croissent jusqu'à 100%.

4.1.2. Validation du modèle de Markov homogène

Il est difficile de valider un modèle en se basant sur la vraisemblance. La fonction « msm » fournit $-2 * \log -$ vraisemblance = 93 770, valeur difficilement interprétable en l'état. Pour valider le modèle, les valeurs prédites peuvent être comparées aux valeurs observées. La fonction « prevalence » construit le stock des personnes dépendantes observées par niveau au cours du temps en nombre et en pourcentage ainsi que les prédictions à partir du modèle. Les graphiques ci-dessous présentent la prévalence relative à nos données ainsi que celle estimée par le modèle en fonction du niveau de dépendance.

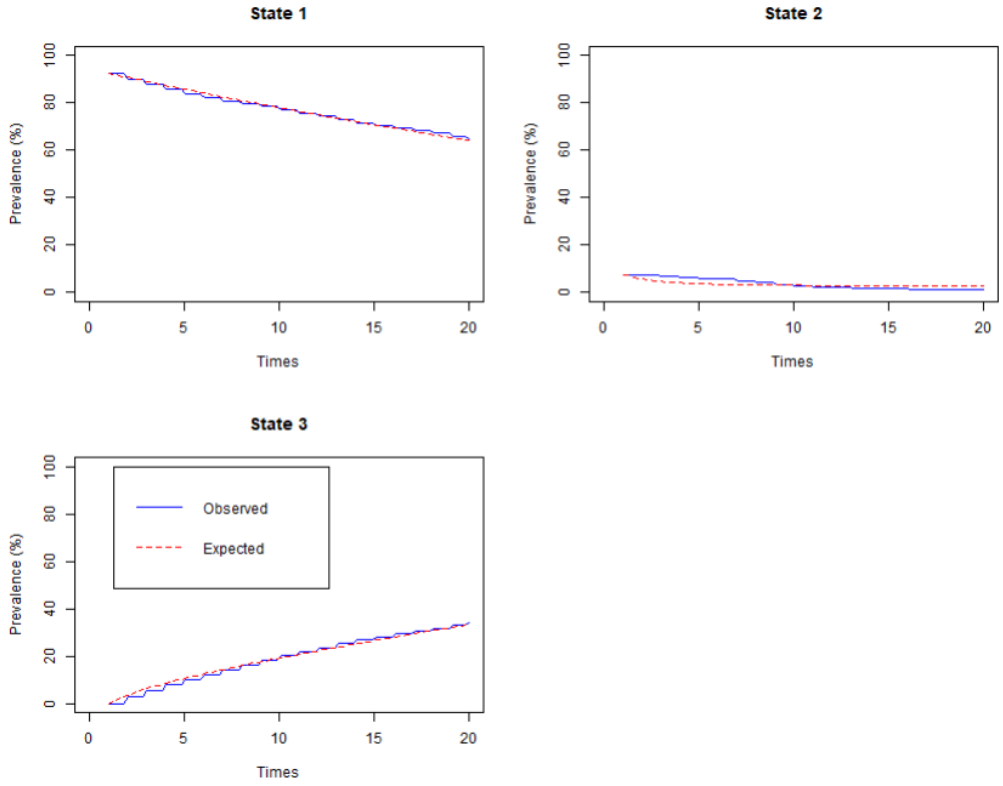


Figure 64 : Sortie R de la fonction de ''prévalence'' pour les différents statuts

Pour l'état 1 et 3 (dépendant et décès respectivement), le modèle colle aux données. Pour l'état 2 (sain), la prédiction sous-estime la prévalence pour les durées inférieures à 10 mois et ensuite la courbe s'inverse (la prévalence est sur-estimée), ce problème est probablement dû au nombre faible des observations sur cet état.
En général, Il apparaît que l'hypothèse d'un modèle markovien homogène est valable.

4.1.3. Trajectoire du modèle markovien homogène :

Le modèle markovien estimé nous permet de générer l'évolution probable d'une personne dépendante jusqu'à sa sortie.
Pour le faire nous avons suivi l'algorithme défini ultérieurement.
L'illustration ci-dessous montre 25 trajectoires. Le point de départ est toujours l'état de dépendance.

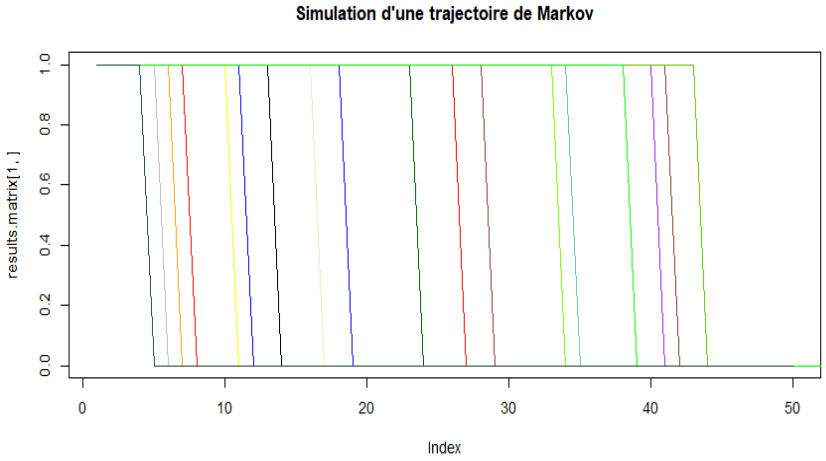


Figure 65 : Simulation de 25 trajectoires de l'état ''dépendant''

Même si le modèle de Markov homogène semble adéquat pour la modélisation de notre jeu de données, il est jugé nécessaire, sur la base des conclusions précédentes, d'explorer également un modèle localement homogène.

4.2. Modèle markovien localement homogène :

Dans cette section, nous allons présenter une modélisation des chaines de Markov localement homogène, autrement l'hypothèse d'homogénéité dans le temps n'est vérifiable que par des morceaux (modèle avec des intensités constantes par morceaux).

4.2.1. Processus markovien homogène par sexe :

Nous allons déterminer un modèle markovien homogène par morceaux en différenciant les matrices d'intensités et de transition par sexe.

Pour ce faire, nous rajoutons à notre modèle la covariable ''sexe'' qui, pour un individu, n'évolue dans le temps.

Nous estimons donc une matrice d'intensité "baseline" à laquelle s'appliqueront des correctifs pour obtenir la matrice d'intensité en fonction du sexe.

La matrice d'intensité de transition se présente comme suit :

```
> qmatrix.msm(cavgender.msm, covariates=list(sexe="1") ) #Female
        State 1                        State 2                        State 3
State 1 -0.01807 (-0.01868,-0.01748)  0.01807 ( 0.01748, 0.01868) 0
State 2  0.04127 ( 0.03601, 0.04729) -0.57040 (-0.58928,-0.55213)  0.52914 ( 0.51170, 0.54717)
State 3 0                              0                           0

> qmatrix.msm(cavgender.msm, covariates=list(sexe="0") ) #Male
        State 1                        State 2                        State 3
State 1 -0.01807 (-0.01868,-0.01748)  0.01807 ( 0.01748, 0.01868) 0
State 2  0.04127 ( 0.03601, 0.04729) -0.57040 (-0.58928,-0.55213)  0.52914 ( 0.51170, 0.54717)
State 3 0                              0                           0
```

Figure 66 : Matrice d'intensité de transition pour différents sexes.

On en déduit le séjour par sexe, comme attendu, les femmes restent plus longtemps en état de dépendance que les hommes, le temps passé en état de dépendance est de 55 mois contre 40 mois pour les hommes. Cette observation conforte les conclusions établies dans les précédentes sections.

```
> sojourn.msm(cavgender.msm,covariates=list(sexe="1") ) #Female
        estimates      SE      L      U
State 1    55.343 0.93552 53.540 57.208
State 2     1.753 0.02912  1.697  1.811
> sojourn.msm(cavgender.msm,covariates=list(sexe="0") ) #Male
        estimates      SE      L      U
State 1     40.57 0.69251 39.237 41.952
State 2      1.41 0.02392  1.364  1.458
>
```

Figure 67 : Le temps moyen de séjour dans un état en fonction du sexe.

La courbe de survie issue du modèle de markovien homogène par sexe :

L'interprétation graphique des courbes de survies relatives à la variable sexe est conforme aux résultats précédents : la dépendance chez les femmes (durée de maintien) est supérieure à celles des hommes.

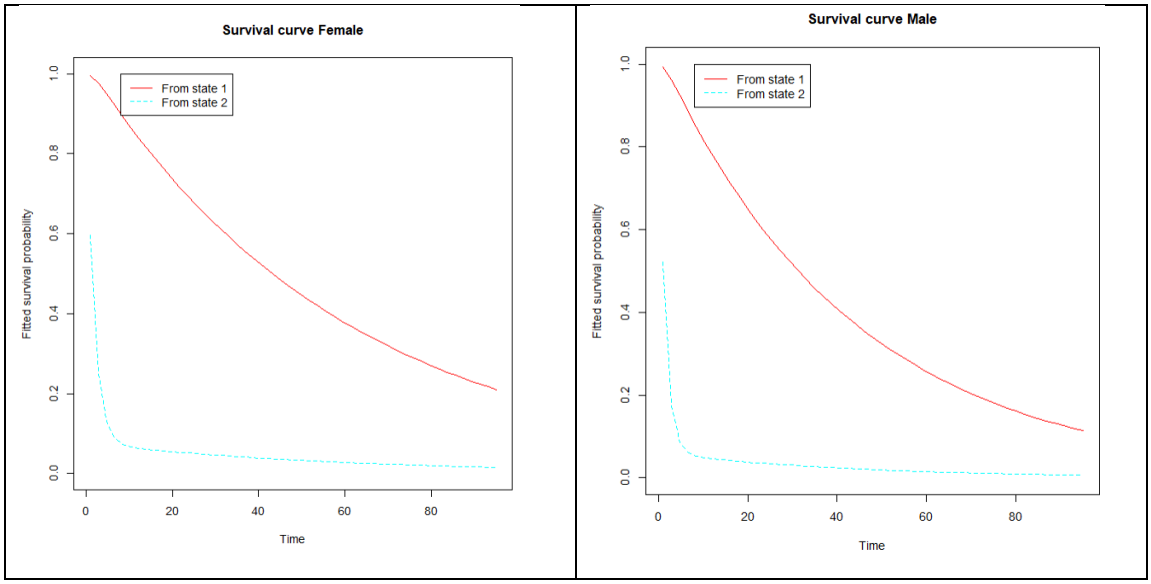


Figure 68 : Fonction de survie par sexe en fonction de l'état

Validation du modèle via la fonction Prevalence

Comme pour le modèle de Markov homogène, on pourra s'appuyer sur la fonction prévalence, celle-ci permet de construire le stock de personnes dépendantes observées ainsi que les prédictions du modèle et donc de les comparer graphiquement :

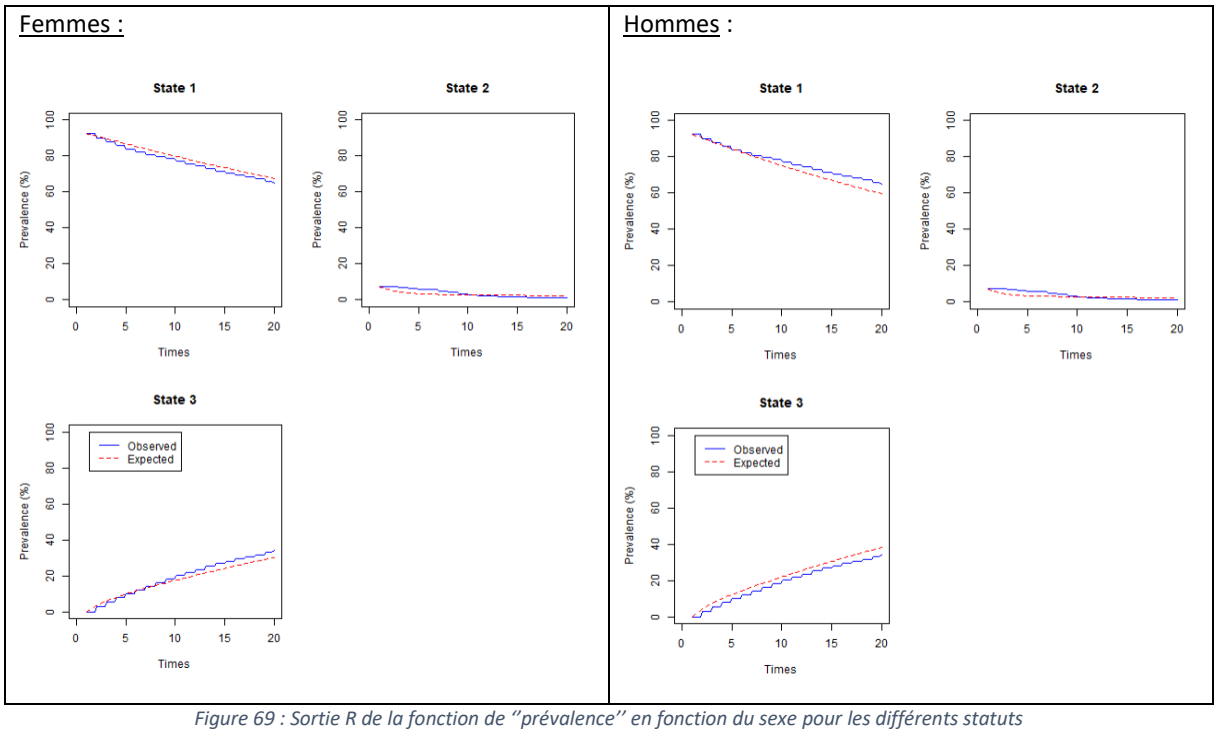


Figure 69 : Sortie R de la fonction de ''prévalence'' en fonction du sexe pour les différents statuts

On remarque que le modèle prédit globalement bien la prévalence par sexe, il sur-estime légèrement la prévalence pour l'état dépendant chez les femmes et l'état décès chez les hommes, et il sous-estime la prévalence pour l'état dépendant chez les hommes et l'état décès chez les femmes.

4.2.2. Processus markovien homogène par tranche d'âge ''AgeBand'' :

Nous allons déterminer un modèle de Markov localement homogène par tranche d'âge.
La matrice d'intensité de transition se présente comme suit :

```
> qmatrix.msm(cavageband.msm,covariates=list(ageBand="(75,80]") ) #
        State 1                      State 2                      State 3
State 1 -0.01976 (-0.02096,-0.01863)  0.01976 ( 0.01863, 0.02096) 0
State 2  0.05309 ( 0.04292, 0.06567) -0.57610 (-0.61021,-0.54389)  0.52301 ( 0.49269, 0.55520)
State 3 0                            0                            0
> qmatrix.msm(cavageband.msm,covariates=list(ageBand="(65,75]") ) #
        State 1                      State 2                      State 3
State 1 -0.01986 (-0.02105,-0.01874)  0.01986 ( 0.01874, 0.02105) 0
State 2  0.03836 ( 0.03099, 0.04749) -0.44644 (-0.47259,-0.42173)  0.40808 ( 0.38468, 0.43290)
State 3 0                            0                            0
> qmatrix.msm(cavageband.msm,covariates=list(ageBand="(90,120]") ) #
        State 1                      State 2                      State 3
State 1 -0.02638 (-0.02818,-0.02469)  0.02638 ( 0.02469, 0.02818) 0
State 2  0.07390 ( 0.04194, 0.13020) -2.23669 (-2.40347,-2.08149)  2.16279 ( 2.01167, 2.32527)
State 3 0                            0                            0
> qmatrix.msm(cavageband.msm,covariates=list(ageBand="(-1,20]") ) #
        State 1                       State 2                       State 3
State 1 -0.002131 (-0.003236,-0.001404)  0.002131 ( 0.001404, 0.003236) 0
State 2  0.007776 ( 0.002595, 0.023302) -0.060695 (-0.089487,-0.041167)  0.052919 ( 0.034942, 0.080144)
State 3 0                             0                             0
> qmatrix.msm(cavageband.msm,covariates=list(ageBand="(20,65]") ) #
        State 1                      State 2                      State 3
State 1 -0.02042 (-0.02196,-0.01899)  0.02042 ( 0.01899, 0.02196) 0
State 2  0.01681 ( 0.01233, 0.02293) -0.27847 (-0.29907,-0.25929)  0.26166 ( 0.24316, 0.28156)
State 3 0                            0                            0
> qmatrix.msm(cavageband.msm,covariates=list(ageBand="(85,90]") )
        State 1                      State 2                      State 3
State 1 -0.02337 (-0.02461,-0.02220)  0.02337 ( 0.02220, 0.02461) 0
State 2  0.07976 ( 0.06060, 0.10496) -1.34673 (-1.41912,-1.27804)  1.26698 ( 1.20123, 1.33632)
State 3 0                            0                            0
>
```

Figure 70 : Matrice d'intensité de transition pour différentes tranches d'âge.

Le nombre faible des transitions sur la tranche (-1,20] crée du bruit dans l'estimation des paramètres du modèle, nous avons décidé de merger cette tranche avec la tranche (20,65] pour former une nouvelle tranche (-1,65].
On en déduit le séjour par tranche d'âge, nous remarquons un séjour long dans l'état dépendant chez les assurés dans la tranche d'âge (75,80] suivi de la tranche (65-75], les âges au-delà de 80 ans ont un séjour inférieur au reste des tranches, ceci nous parait logique car l'espérance de vie en dépendance décroit avec l'âge.

```
> sojourn.msm(cavageband.msm,covariates=list(ageBand="(75,80]") ) #
        estimates         SE         L         U
State 1 50.632288 1.52684354 47.726449 53.715051
State 2  1.739802 0.05113865  1.642404  1.842975
> sojourn.msm(cavageband.msm,covariates=list(ageBand="(65,75]") ) #
        estimates         SE         L         U
State 1 50.278044 1.49513887 47.431389 53.295544
State 2  2.239306 0.06510437  2.115271  2.370614
> sojourn.msm(cavageband.msm,covariates=list(ageBand="(90,120]") ) #
         estimates         SE          L          U
State 1 37.9416207 1.2803411 35.5133840 40.5358886
State 2  0.4477568 0.0164441  0.4166596  0.4811749
> sojourn.msm(cavageband.msm,covariates=list(ageBand="(20,65]") ) #
        estimates        SE         L         U
State 1 48.958422 1.8119444 45.532821 52.641743
State 2  3.600128 0.1312597  3.351841  3.866807
> sojourn.msm(cavageband.msm,covariates=list(ageBand="(-1,65]") ) #
        estimates         SE         L         U
State 1 50.160807 0.6736548 48.857694 51.498677
State 2  1.443798 0.0192215  1.406612  1.481968
> sojourn.msm(cavageband.msm,covariates=list(ageBand="(85,90]") )
         estimates         SE          L          U
State 1 42.7644758 1.12464619 40.6160549 45.0265393
State 2  0.7436574 0.01988207  0.7056926  0.7836645
```

Le séjour chez la tranche (20-65] avant la merger avec la tranche (-1,20] : A titre indicatif

```
> sojourn.msm(cavageband.msm,covariates=list(ageBand="(80,85]") )
        estimates       SE      L     U
State 1      48.3 1.22064 45.963 50.75
State 2       1.2 0.03007  1.143  1.26
```

Figure 71 : Le temps moyen de séjour dans un état en fonction de la variable tranche d'âge.

La courbe de survie issue du modèle markovien homogène par tranche d'âge ''AgeBand'' :

Les courbes de survie par tranche d'âge permettent de déduire que les assurés de la tranche d'age (75-80] se maintiennent en dépendance plus longuement que les autres tranches. En revanche les assurés les plus vieux possèdent la courbe qui décroit le plus vite.

En raison du nombre faible des données de l'état ''valide'', on constate que la sortie est rapide avec une allure similaire pour toutes les tranches d'âges.

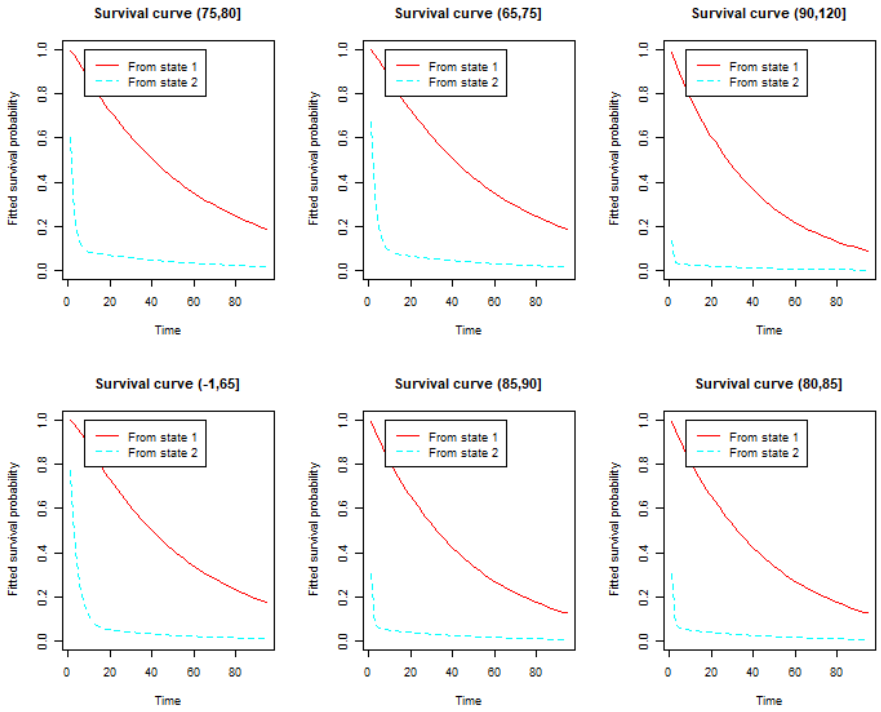
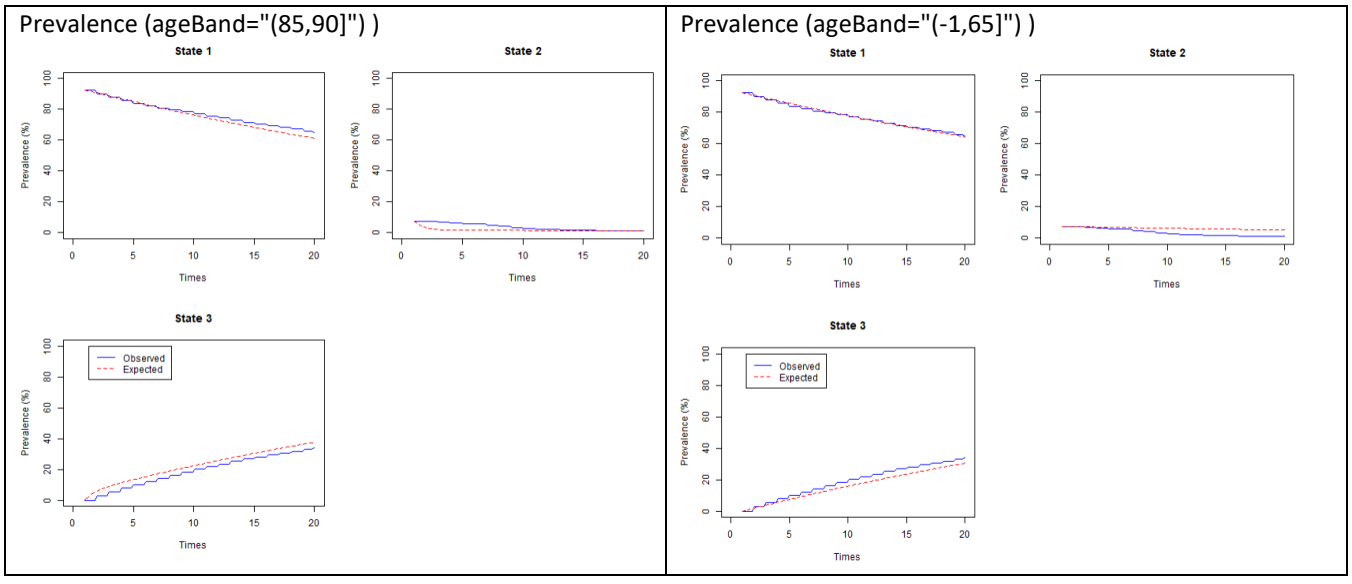


Figure 72 : Fonction de survie par sexe en fonction de la tranche d'âge.

Validation du modèle via la fonction Prevalence

Pour valider le modèle nous procédons de la même manière que pour la variable sexe, à savoir la fonction de prévalence.



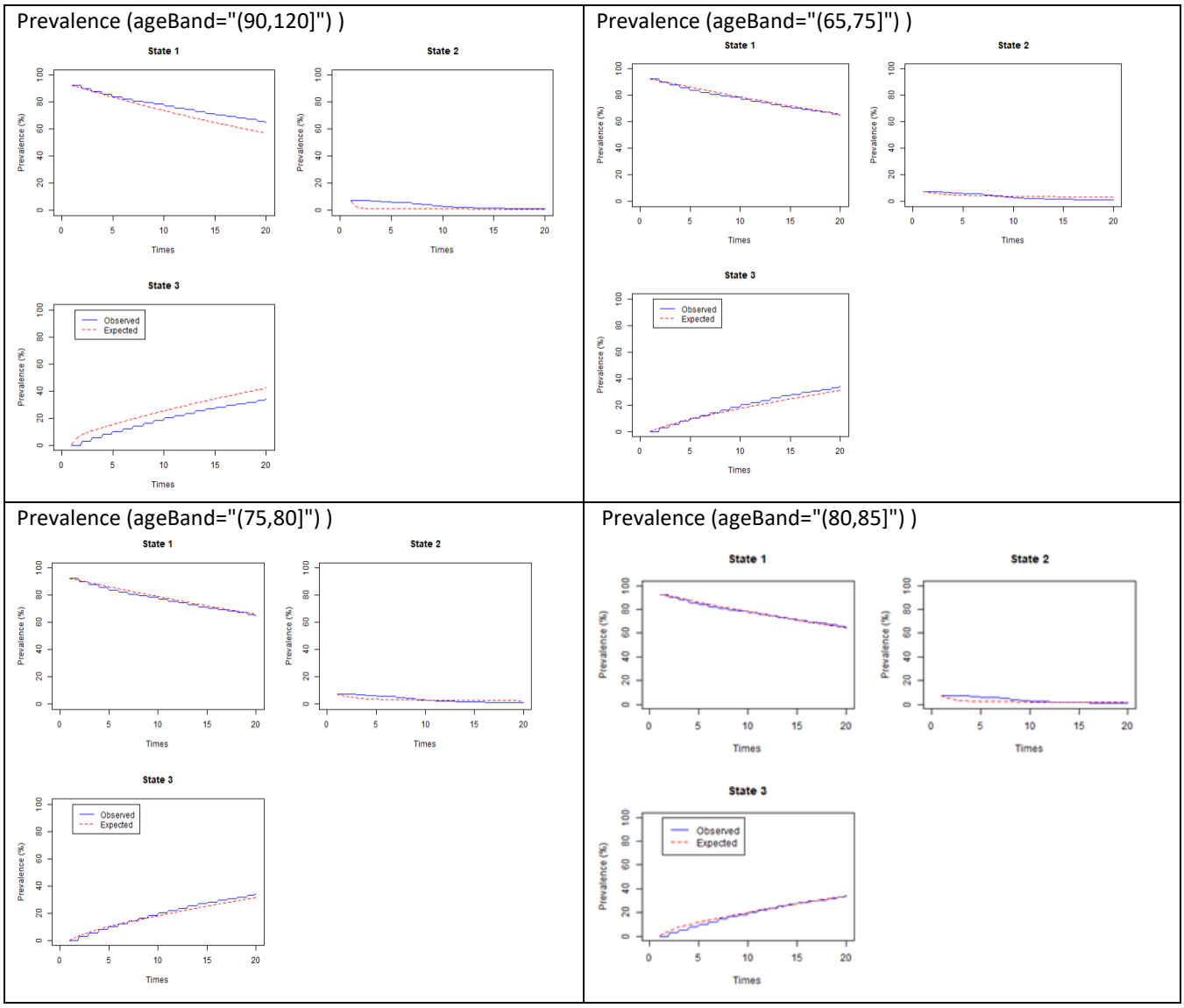


Figure 73 : Sortie R de la fonction de ''prévalence'' en fonction de la tranche d'âge pour les différents statuts

On remarque que le modèle prédit globalement moins bien la prévalence par tranche d'âge par rapport à la variable sexe, Plusieurs explications sont possibles (homogénéité n'est pas respecté au sein de chaque sous-groupe, ou bien la propriété de Markov (l'état futur ne dépend que de l'état présent) n'est pas remplie), mais nous jugeons que le manque de données par tranche d'âge notamment pour les tranches d'âge les moins représentatifs impacte fortement les fonctions de prévalence.

4.3. Modèle semi-markovien.

Dans le processus semi-Markovien homogène les intensités de transition dépendent de la durée passée dans l'état et de l'état occupé.

4.3.1. Le principe du modèle semi-Markovien :

Ce modèle nous permet de générer l'évolution probable d'une personne dépendante à travers différents états jusqu'à son décès. L'approche qu'on retient consiste à modéliser le processus de saut et une loi de durée de séjour dans l'état ceci pour chaque sexe de l'assuré. En ce qui concerne le processus de saut ; nous allons nous référer au modèle markovien développer dans la section précédente, quant à la loi de durée nous allons faire appel à la loi de Weibull, plus de détail sur cette loi sera présenté ultérieurement.

Pour obtenir une vue complète sur les trajectoires possibles en dépendance, nous construisons des tables d'expérience des personnes dépendantes à l'aide de la méthode de Monte-Carlo. Celle-ci consiste à simuler un

très grand nombre de trajectoires avec les mêmes conditions initiales puis à moyenner les résultats obtenus pour évaluer un comportement central.
Plus les trajectoires simulées sont nombreuses et plus l'image du processus correspondant à notre modèle sera complète et fidèle. Étant donné un jeu de conditions initiales, toutes les trajectoires générées sont équiprobables. La loi des grands nombres permet d'approcher par moyenne simple la valeur centrale du processus à un instant donné, ce pour tous les instants possibles.

L'algorithme de simulation :

Comme mentionné auparavant, nous proposons de simuler les états successifs d'un assuré d'âge x au cours de la vie de son contrat selon un algorithme de rejet se basant sur des tirages d'une loi uniforme en fonction des probabilités de passage définies dans la matrice M définissant la chaîne de Markov initiée précédemment.
L'évolution de l'état de l'assuré d'un mois à un autre s'effectue de la manière suivante ; soit u une réalisation de variable aléatoire uniforme :

1) A partir de l'état initial de dépendance i_1 nous générons l'état suivant i_2, si $u < p^{dep,dep}$, l'assuré reste en dépendance
2) À partir de l'état initial i_1 et l'état suivant i_2, nous générons le temps de passage correspondant par une réalisation d'une loi de Weibull simples si i_2. Les paramètres de la loi de Weibull considérée sont ceux estimés dans la section suivante.
3) Si le dernier état visité n'est pas la mort, nous répétons les étapes 2 puis 3 jusqu'au décès de l'individu.

Le manque de données entre le passage de l'état dépendant à l'état actif limite de tirer des conclusions statistiquement robustes, alors nous avons décidé d'étudier que l'état de transition entre la dépendance et le décès.

4.3.2. Estimation du temps de séjour à l'aide de la loi de Weibull

Tout d'abord, nous allons décrire les caractéristiques de la loi de Weibull que nous utiliserons dans le contexte du modèle AFT (Accelerated Failure Time) pour estimer la durée de séjour dans un état.

La loi de Weibull est une loi de probabilité à deux facteurs k et λ. k représente le paramètre de forme et λ le paramètre d'échelle.
La fonction de survie pour un modèle de Weibull s'exprime sous la forme suivante :

$$\forall t \geq 0, S(t) = exp\left(-\left(\frac{t}{\lambda}\right)^{\mathrm{k}}\right), \text{avec } \mathrm{k} > 0, \lambda > 0,$$

Lorsque $k = 1$, on retrouve la densité d'une loi Exponentielle de paramètre λ.
Pour tenir en compte la segmentation par sexe nous allons procéder à l'application du modèle AFT (Accelerated failure Time model).

Présentation de la méthode de régression par un modèle de vie accélérée (AFT)

Les modèles à temps de vie accélérée sont des modèles de régression utilisés en analyse de survie pour la modélisation et l'analyse des distributions de survie. Ils se distinguent des modèles à hasard proportionnel dans leur formulation mathématique par l'introduction d'une loi de probabilité. Ces modèles peuvent être des alternatives utiles au modèle à risque proportionnel de Cox. Tout comme le modèle de Cox, ils permettent d'intégrer l'effet des différentes covariables sur la fonction de survie.
Soit T une variable aléatoire continue, positive ou nulle mesurant la durée de survie d'un individu. Considérons toujours un groupe de n individus ayant p covariables $X = (X1, \ldots, Xp)$; $\beta = (\beta 1, \ldots, \beta p)$ les coefficients de régression. Les modèles à temps de vie accélérée s'écrivent :
$log(T) = \beta_1 X_1 + \cdots + \beta_P X_P + \varepsilon$ (Equ.1)
Contrairement au modèle de Cox qui explique le logarithme de la fonction de hasard par une régression linéaire, les modèles à temps de vie accélérée expliquent par une régression linéaire sur les covariables le logarithme des temps de survie. La particularité est qu'ils supposent de plus que les temps de survie suivent une distribution dictée par la variable aléatoire ε. Les lois de probabilité les plus utilisées dans la modélisation sont la loi

Exponentielle, la loi de Weibull, la loi Log-normal, la loi Gamma, la loi gamma généralisé, la loi log-logistique, la loi de Fisher, etc.

L'équation (Equ.1) peut être réécrite de la façon suivante :

$$T = T_0 exp(\beta' X)$$

Avec $T_0 = exp(\varepsilon)$

Le calcul de la fonction de survie au temps t donne :

$$\begin{aligned} S(t/X) &= P(T > t/X) \\ &= P(T_0 exp(\beta' X) > t) \\ &= P\big(T_0 > t * exp(-\beta' X)\big) \\ &= S_0\big(t * exp(-\beta' X)\big) \end{aligned}$$

La fonction de survie est donc de la forme :

$$S(t/X) = S_0\big(t * exp(-\beta' X)\big)$$

Application du modèle AFT

On effectue une régression par maximum de vraisemblance dans un univers de données censurées à l'aide du package "flexsurvreg". On s'intéresse à la loi de Weibull.

Un des intérêts de notre modélisation AFT est qu'elle nous donne la possibilité de résumer nos courbes de survie homme et femme par une relation de proportionnalité. Pour ce faire, on teste la significativité des coefficients. La statistique de Wald z dont la construction et la p-value associée nous assurent que le coefficient estimé est significativement différent de 0. On présente la sortie R ci-dessous.

```
> summary(weibull1)

Call:
survreg(formula = Surv(duration_adj1, status) ~ gender, data = data_weibull,
    dist = "w")
               Value Std. Error     z      p
(Intercept)  3.58031    0.00967 370.2 <2e-16
genderMale  -0.18229    0.01382 -13.2 <2e-16
Log(scale)  -0.09149    0.00644 -14.2 <2e-16

Scale= 0.913

Weibull distribution
Loglik(model)= -78186   Loglik(intercept only)= -78272
        Chisq= 173.4 on 1 degrees of freedom, p= 1.3e-39
Number of Newton-Raphson Iterations: 5
n= 20450
```

Figure 74 : Sortie R de la régression AFT

Nos coefficients sont bien significatifs et on peut proposer la relation de proportionnalité suivante :

$$S_{femme}(t) = S_{homme}(t * \exp(-0.18)) \Leftrightarrow S_{femme}(t) = S_{homme}(t * 0.83)$$

Nous constatons donc que le temps de survie est 0,83 fois plus courte par rapport à la survie de base ''femme''

La fonction de survie pour les deux sexes est présentée ci-dessous.

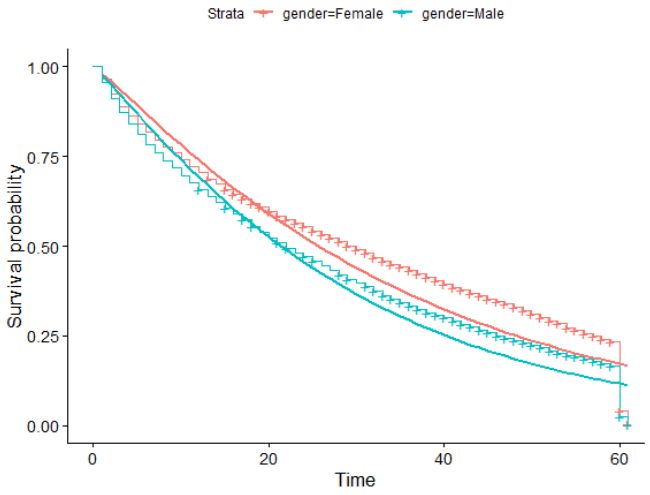


Figure 75 : la fonction de survie estimé par la régression d'AFT en fonction du sexe.

On constate que les fonctions de survie sont discernables. Les femmes ont une probabilité de survie plus élevée que celle des hommes, ceci est en adéquation avec les modèles précédents.
En comparant le modèle de Weibull avec l'estimateur de Kaplan-Meier, nous constatons que pour les durées inférieures à 20 mois, la loi de Weibull surestime la fonction de survie par rapport à l'estimateur de Kaplan-Meier, après cette date, le constat s'inverse et la courbe estimée par la loi de Weibull se trouve en dessous de la courbe de survie de Kaplan Meier.

Estimation des paramètres du Weibull :

Les paramètres de la loi de Weibull (forme et échelle) sont obtenus par estimation de maximum de vraisemblance dont le résultat est fourni par le logiciel R.

```
> beweibull1<- flexsurvreg(Surv(duration_adj1,status)~gender, data=data_weibull, dist="weibull")
> beweibull1
Call:
flexsurvreg(formula = Surv(duration_adj1, status) ~ gender, data = data_weibull,
    dist = "weibull")

Estimates:
           data mean  est       L95%      U95%      se        exp(est)  L95%      U95%
shape             NA    1.09581   1.08206   1.10972   0.00706        NA        NA        NA
scale             NA   35.88461  35.21078  36.57134   0.34707        NA        NA        NA
genderMale   0.47330   -0.18229  -0.20938  -0.15520   0.01382   0.83336   0.81109   0.85625

N = 20450,  Events: 17480,  Censored: 2970
Total time at risk: 569087
Log-likelihood = -78186, df = 3
AIC = 156377

>
```

Figure 76 : Sortie R de l'estimation des paramètres de la loi de Weibull.

La sortie R permet de conclure que :
La variable de forme (shape) : Le coefficient estimé est d'environ 1.09. La loi de Weibull se réduit à une loi exponentielle lorsque ce paramètre est égal à 1, donc dans ce cas, la distribution du temps de survie est légèrement différente d'une distribution exponentielle, suggérant une légère variation dans le taux de risque sur le temps.
La variable d'échelle (scale) : Elle a une valeur estimée d'environ 35.88 (=$e^{3,58}$). Ce paramètre détermine la dispersion des durées de vie. Plus la valeur d'échelle est élevée, plus les durées de vie sont étalées.

4.3.3. Simulation de trajectoire semi-Markovien.

L'estimation des paramètres de la loi de Weibull ainsi que l'utilisation de la matrice de passage estimée dans la section de modèle de Markov homogène par morceau, nous permettront facilement à l'aide de l'algorithme de simulation de générer des trajectoires. La figure ci-dessous représente 5 simulations de trajectoire pour un assuré de sexe feminin.

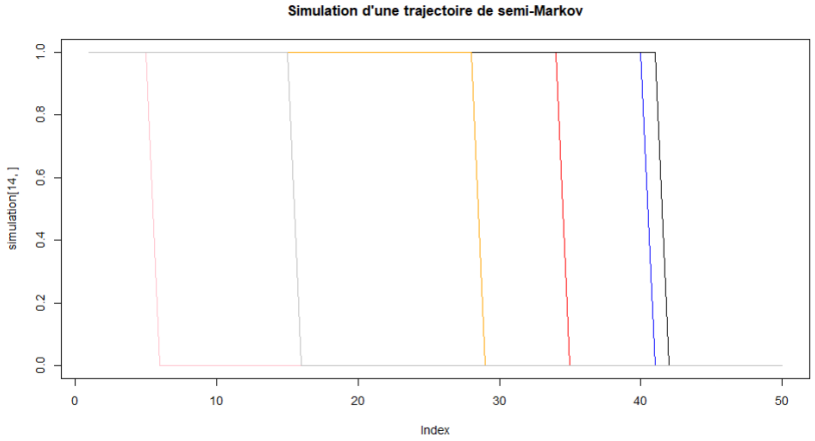


Figure 77 : Simulation semi-markovienne de 5 trajectoires de l'état ''dépendant''

Exemple numérique (fictif) pour le calcul de la provision mathématique :
Les 5 trajectoires simulées permettent de calculer une provision mathématique théorique comme suit.
Posons les hypothèses suivantes :

- Assuré femme
- Taux d'intérêt = 0%
- La rente = 100 €

Scénarios	Temps de séjour (mois)	Montant de provision
1	7	700 €
2	17	1700 €
3	29	2900 €
4	35	3500 €
5	42	4200 €
Moyenne	**26**	**2600 €**

Tableau 10 : Exemple fictif de calcul de la provision mathématique selon le modèle semi – markovien

Ainsi la provision mathématique pour ce cas fictif s'établie à 2600€

Conclusion :

Nous avons vu que l'approche markovienne se présente comme une alternative efficace pour la modélisation de la dépendance une fois que le volume de données le permette.

Dans un premier temps nous avons proposé le cas d'un processus markovien homogène, clairement le plus simple à manipuler, mais qui suppose des hypothèses que nous avons jugé inadaptées à notre jeu de données notamment l'hypothèse d'homogénéité des intensités qui ne dépendent ni du temps ni de la durée passé dans l'état, en effet le modèle homogène est aussi très bien adapté aux jeux de données observés sur une période trop courte pour pouvoir présumer d'un effet de dérive temporelle. Pour pallier le problème de l'homogénéité nous avons proposé un modèle markovien homogène par morceau en introduisant les covariables sexe et tranche d'âge, par souci de significativité de données nous avons conclu que le modèle homogène par morceau selon le sexe est le plus adéquat. Et enfin nous avons développé un modèle semi-markovien homogène par morceau, qui est une généralisation du précédent dont l'intérêt principal pour nous, est qu'il gère de manière indépendante les transitions et les durées entre transitions, nous avons retenu la loi de Weibull comme loi décrivant la durée passée dans un état et les lois de transitions définies dans le modèle markovien homogène par morceau pour estimer les lois de passage. Ce modèle nous a permis de générer l'évolution probable d'une personne dépendante jusqu'à sa sortie. Pour obtenir une vue complète sur les trajectoires possibles nous avons fait appel à la méthode de Monte-Carlo qui consiste à simuler un très grand nombre de trajectoire conditions initiales puis à moyenner les résultats obtenus pour évaluer un comportement central.

Chapitre 4 : Méthode d'estimation des tardifs et leur application

En addition des provision mathématiques de rente (PM) présentées dans la section précédente, le réassureur est imposé de déterminer les provisions pour sinistres déclarés tardivement (IBNR) des prestations futures au titre des sinistres en cours, mais encore inconnus à la date de clôture.

I. Construction du triangle de développement

La méthodologie que nous allons retenir pour la détermination des IBNR se basera sur :

1. Une estimation de nombre de sinistres ultime auquel on déduit le nombre de sinistres connus
2. Calcul de montant IBNR similairement aux sinistres connus tout en respectant la structure du portefeuille.

Généralement, les triangles de nombre de sinistres sont construits avec un pas annuel. Dans le cadre de cette étude, les triangles sont réalisés au pas mensuel, pour plus de précision.

Les notations utilisées dans cette partie sont les suivantes :

- i : Désigne le mois de survenance du sinistre $i = 0, \dots, n$
- j: Désigne le mois de développement $j = 0,1.., n$
- X_{ij} : le nombre de sinistres survenus en mois i connu après j mois de développement
- $C_{i,j}$: le nombre cumulé de nombre de sinistres le mois i, au$j^{ème}$mois de développement, on a donc $C_{i,j} = \sum_{k=}^{j} X_{kj}$

	1	*2*	…	*j*	…	*n-1*	*n*
1	$C_{1,1}$	$C_{1,2}$	…	…	…	$C_{1,n-1}$	$C_{1,n}$
2	$C_{2,1}$	$C_{2,2}$	…	…	…	$C_{2,n-1}$	
…	…	…	…	…	…		
i	…	…	…	$C_{i,j}$			
…	…	…	…				
n-1	$C_{n-1,1}$	$C_{n-1,2}$					
n	$C_{n,1}$						

Figure 78 : Triangle de nombre de sinistres cumulé

Le triangle peut se lire de différentes façons :
- En colonne : pour un mois de développement donné, on observe le nombre de sinistres indépendamment du mois de survenance ;
- En ligne : pour un mois de survenance donné, on observe la cadence de nombre de sinistres.
- En diagonale : celle-ci correspond à un mois comptable

L'information connue correspond au triangle supérieur, soit :

$$I = \{(C_{i,j}) | 0 \leq i + j \leq n\}$$

L'objectif est d'estimer la partie inférieure du triangle par les différentes méthodes présentées dans les parties suivantes, $\{(\widehat{C_{i,j}}) | i + j > n, i \leq n, j \leq n\}$

	1	*2*	…	*j*	…	*n-1*	*n*
1	$C_{1,1}$	$C_{1,2}$	…	…	…	$C_{1,n-1}$	$C_{1,n}$
2	$C_{2,1}$	$C_{2,2}$	…	…	…	$C_{2,n-1}$	$\hat{C}_{2,n}$
…	…	…	…	…	…	…	…
i	…	…	…	$C_{i,j}$	…	…	…
…	…	…	…	…	…	…	…
n-1	$C_{n-1,1}$	$C_{n-1,2}$	…	…	…	$\hat{C}_{n-1,n-1}$	$\hat{C}_{n-1,n}$
n	$C_{n,1}$	$\hat{C}_{n,2}$	…	…	…	$\hat{C}_{n,n-1}$	$\hat{C}_{n,n}$

Figure 79 : Triangle de nombre de sinistres complété

II. Méthode déterministique de Chain Ladder :

La méthode Chain Ladder est une méthode déterministe fréquemment utilisée car facile à mettre en œuvre. Elle s'applique à des triangles de paiements cumulés, de nombre de sinistre ou des triangles de charges.
L'idée de cette méthode est de supposer que la liquidation future est similaire à la liquidation passée. Les hypothèses du modèle sont les suivantes :
– le nombre de sinistres incrémentaux sont indépendants par mois de survenance,
– les mois de développement sont les variables explicatives du comportement du nombre de sinistres futurs.
Le déroulement des sinistres est régi par des facteurs de développement strictement positifs $f_j; j \in \{1, \dots, n-1\}$ qui ne dépendent que du mois de développement j.

Remarque :
Pour appliquer cette méthode, il est nécessaire de disposer d'un portefeuille homogène et grand, sans événements extrêmes. Ces conditions sont indispensables pour obtenir des résultats satisfaisants.

On définit le facteur de développement individuel :

$$f_{i,j} = \frac{C_{i,j+1}}{Ci,j} \; pour \; i,j = 1, \dots, n-1$$

Nous allons alors considérer des coefficients de passage, d'un mois à l'autre, commun pour les mois de survenance, et dont l'estimation est donnée par :

$$\widehat{f_j} = \frac{\sum_{i=0}^{n-j+1} C_{i,j+1}}{\sum_{i=0}^{n-j+1} C_{i,j}}$$

Grâce à ces facteurs, nous pouvons estimer :

- Le nombre de sinistres ultimes par mois de survenance $\widehat{C_{\iota,n}} = C_{i,n-i} \times \prod_{j=n-i}^{n-1} \widehat{f_j}$
- Le nombre de sinistres tardifs par mois de survenance $I\widehat{BNR}_\iota = \widehat{C_{\iota,n}} - C_{i,n-i}$

Nous rappelons qu'il existe une correspondance entre les coefficients de passage et les cadences de nombre de sinistres $Cadence_j = \frac{1}{f_j \times \dots \times f_{n-1}}$
Malgré sa simplicité d'utilisation, il est important d'effectuer certaines vérifications pour valider cette méthode. Nous allons retenir deux tests, un est graphique et l'autre est numérique :

- L'alignement des couples $(C_{i,j}, C_{i,j+1})$
- Le coefficient de détermination R^2

Pour j fixé, nous avons supposé l'existence d'un coefficient $\widehat{f_j}$ tel que $C_{i,j+1} = \widehat{f_{i,j}} \times C_{i,j}$
Les couples $(C_{i,j}, C_{i,j+1})_{i=0,\dots,n-j-1}$ doivent donc être sensiblement alignés par une droite passant par l'origine.

1. Application de la méthode de Chain Ladder

1.1. Vérification des hypothèses sous-jacentes au Chain -Ladder

L'alignement des couples $(C_{i,j}, C_{i,j+1})$

Afin de vérifier cette hypothèse nous allons examiner l'existence d'une relation linéaire entre le nombre de sinistres cumulés d'un mois de déroulement à l'autre.

Nous concentrerons notre analyse sur les données du triangle de nombre à compter de l'année 2016. Cette décision est motivée principalement par deux facteurs : premièrement, l'harmonisation des données en raison d'un changement observé dans le comportement des sinistres à partir de 2015, deuxièmement, la spécificité de la durée de couverture qui s'étend sur 5 ans.

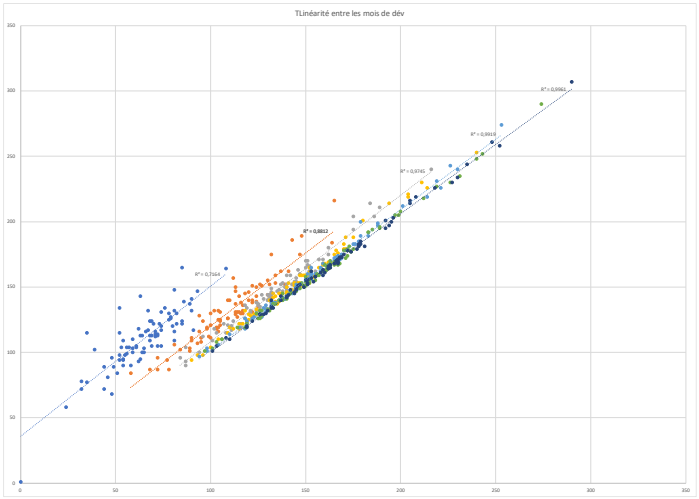


Figure 80 : C-C plot de nombre de sinistres cumulés

Ces graphiques représentent les éléments du mois $i+1$ de développement du triangle selon le mois i.
On peut noter que l'ensemble des points est relativement bien aligné sur une droite passant par l'origine. Ces éléments valident bien l'utilisation de la méthode de Chain Ladder avec les données présentes.
Pour réconforter notre hypothèse de linéarité, nous pouvons faire recours au coefficient de détermination R²
Coefficient de détermination R²

Le coefficient de détermination (R²) détermine à quel point l'équation de régression linéaire est adaptée pour décrire la distribution linéaire des points.

Mois de dev	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12
R²	71,63892%	88,117643%	97,44672%	98,85562%	99,02748%	99,60824%	99,54946%	99,70672%	99,72414%	99,71551%	99,56371%	99,57521%

Mois de dev	12-13	13-14	14-15	15-16	16-17	17-18	18-19	19-20	20-21	21-22	22-23	23-24
R²	99,67576%	99,88029%	99,94169%	99,90450%	99,79755%	99,89858%	99,91754%	99,94760%	99,92689%	99,95348%	99,95583%	99,96520%

Mois de dev	24-25	25-26	26-27	27-28	28-29	29-30	30-31	31-32	32-33	33-34	34-35	35-36
R²	99,98245%	99,98595%	99,98312%	99,97045%	99,93953%	99,96009%	99,94140%	99,94977%	99,96929%	99,85333%	99,84984%	99,94980%

Tableau 11 : Coefficient de détermination

Les tableaux ci-dessous affichent des coefficients proches de 1, ce qui indique que l'hypothèse de linéarité est respectée.

1.2. Facteur de développement de Chain Ladder

Le calcul des coefficients de passage présente les résultats suivants, la courbe ''rose'' représente la cadence de liquidation.

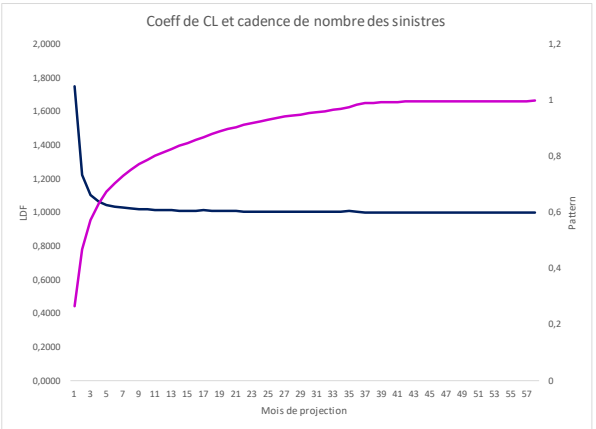


Figure 81 : Coefficient de Chain Ladder et cadence de nombre de sinistres

Nous constatons une convergence des facteurs de développement vers 1 à partir du 11 -ème mois ce qui signifie que le nombre de sinistres tardives évoluent très peu à partir de la vision de l'année comptable 2022.

2. Résultat de la méthode de Chain-Ladder:

Les résultats des estimations issues de la méthode de Chain Ladder standard sont résumés dans le tableau suivant.

Le nombre de sinistre tardif que le réassureur estime recevoir s'élève à 125 sinistres

Mois	Nombre de sinistre observé	Nbre de sinistre Ultime	IBNR nombre
2016_1	193,00	193,00	-
2016_2	156,00	156,00	-
2016_3	203,00	203,00	-
2016_4	173,00	173,00	-
2016_5	177,00	177,00	-
2016_6	214,00	214,00	-
2016_7	274,00	274,25	0,25
2016_8	226,00	226,20	0,20
2016_9	167,00	167,15	0,15
2016_10	139,00	139,12	0,12
2016_11	241,00	241,34	0,34
2016_12	219,00	219,31	0,31
2017_1	281,00	281,40	0,40
2017_2	204,00	204,29	0,29
2017_3	205,00	205,29	0,29
2017_4	195,00	195,28	0,28
2017_5	247,00	246,35	0,25
2017_6	244,00	244,42	0,42
2017_7	248,00	248,49	0,49
2017_8	253,00	253,56	0,56
2017_9	210,00	210,61	0,61
2017_10	219,00	219,74	0,74
2017_11	273,00	274,10	1,10
2017_12	281,00	282,42	1,42
2018_1	264,00	265,69	1,69
2018_2	226,00	227,90	1,90
2018_3	307,00	309,96	2,96
2018_4	274,00	277,80	3,80
2018_5	297,00	303,70	6,70
2018_6	243,00	250,06	7,06
2018_7	323,00	333,98	10,98
2018_8	285,00	295,88	10,88
2018_9	179,00	186,63	7,63
2018_10	298,00	311,76	12,94
2018_11	331,00	348,01	16,50
2018_12	610,00	643,78	33,78
Total	8 879	9 004,48	125,04

Tableau 12 : Estimation de nombre IBNR par mois de survenance

<u>Méthode de détermination de queues de développement :</u>

Pour une vision plus prudente sur le facteur de Chain ladder, nous proposons d'inclure un facteur de queue pour tester la profondeur de l'historique et aussi pour compenser l'effet du facteur de développement inférieur à un qui peut sous-estimer les provisions :

La définition de nouveaux facteurs de développement est en général effectuée à l'aide d'une régression par courbe de référence (exponentielle, Weibull, quadratique…) sur les facteurs connus. Le critère des moindres carrés est utilisé pour estimer les paramètres de la courbe ($a \in \mathbb{R}$, $b \in \mathbb{R}+$) reflétant au mieux les coefficients de développement connus. C'est l'extrapolation exponentielle qui a été retenue dans ce mémoire pour la projection

des coefficients de développement. Celle-ci est justifiée par la bonne adéquation entre la modélisation des coefficients historiques et modélisés détaillée par la suite.
La fonction exponentielle inverse est définie comme :

$$f: \mathbb{N} \rightarrow \mathbb{R}+$$
$$j \rightarrow exp\{a - b \times j\} + 1$$

Le calcul du tail factor repose sur l'estimation des paramètres de la fonction exponentielle inverse par la méthode des moindres carrés ordinaires.
Nous considérons que le triangle est entièrement développé lorsque l'estimateur de son dernier facteur de développement vérifie la condition :

$$\hat{f}_j - 1 < 10^{-5}$$

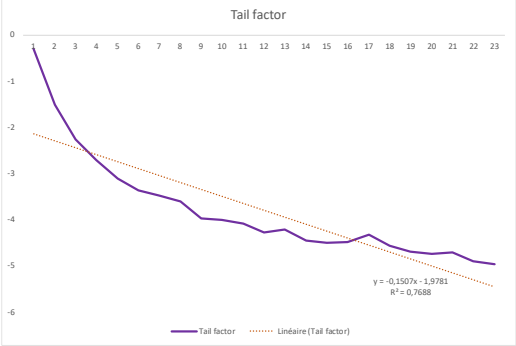


Figure 82 : Graphique de Tail factor

L'application du modèle de régression aux facteurs de développement du triangle de nombre de sinistre nous permet d'obtenir $n_{ult} = 62\ et\ f_{ult} = 1{,}00001041567 \cong 1$, donc nous nous contentons sur la profondeur de développement initiale.

Remarque :
Il peut arriver que les données brutes que l'on souhaite utiliser ne vérifient pas de manière satisfaisante les hypothèses nécessaires à l'application de la méthode. La présence de points aberrants est souvent courante. Un changement dans les systèmes informatiques, dans la gestion des sinistres, dans la réglementation… peut impacter grandement les cadences de paiement des sinistres, leurs réserves dossier/dossier le nombre de sinistres. Afin de pallier ces problèmes, il est possible de modifier les calculs des coefficients de développement en essayant de ne pas prendre en compte certains coefficients de passage individuel :
Les ajustements suivants peuvent être simplement mis en place pour le calcul des coefficients de développement :

- Remplacer les points aberrants au niveau des coefficients de passage par une moyenne des autres coefficients de passage pour une année de développement donnée.
- Supprimer une partie des données du triangle, cela revient à « nettoyer » certaines années pour lesquelles on sait qu'elles sont non significatives.
- Considérer les historiques sur une période plus courte pour calculer les coefficients de développement, sur les 3 dernières années par exemple, si on sait que le business a considérablement évolué.
- Choisir de pondérer ou non les coefficients en fonction du volume des charges de sinistres constatés afin de donner plus de poids aux années où l'activité est plus révélatrice de la situation réelle du marché.

Bien sûr, ce genre d'analyse doit s'accompagner d'une recherche des causes plus approfondie auprès des différents services de gestion des sinistres.
Dans notre exemple, le triangle de nombre ne présente pas de point aberrant, ceci est dû à la prise en compte que des sinistres postérieurs à 2016 permettant d'avoir des données homogènes, en outre les coefficients de passage converge rapidement vers 1 sans présenter d'irrégularité prouvant une stabilité des données.

Avantage et critiques :
Nous avons vu que cette méthode a l'avantage d'être aisée à appliquer, ce qui explique son succès, ainsi c'est une méthode qui prend en compte les variations stables qu'un triangle de liquidation peut présenter (exemple d'une l'inflation constante dans le temps). Cependant, elle repose sur des hypothèses qui peuvent ne pas s'avérer réalistes dans la pratique : les facteurs de développement sont supposés constants pour toutes les années

d'origine, c'est-à-dire que le nombre cumulés $C_{i,j}$ et $C_{i,j+1}$ doivent être proportionnels. De plus, le nombre de sinistres tardifs dépend fortement du dernier nombre connu ; si le triangle est peu stable et irrégulier, ce modèle n'est pas adapté.

III. Méthodes stochastiques :

L'idée dans cette partie est de pouvoir quantifier la variabilité du nombre des sinistres tardifs estimées, notamment par la construction d'intervalles de confiance, et pour obtenir une marge d'erreur sur le nombre ultime. Pour cela, les méthodes stochastiques sont adaptées, puisqu'elles considèrent les nombres ultimes sous un angle probabiliste en considérant leur distribution.

1. Méthode de Mack :

1.1. Cadre théorique du modèle de Mack

La méthode de Mack est la première méthode faisant intervenir la notion d'incertitude dans la méthode déterministe Chain-Ladder. En effet, elle permet de mesurer l'incertitude associée à la prédiction du montant des provisions que doit faire l'assureur.
Cette méthode s'appuie sur trois hypothèses :

(H1) : Les années de survenance des sinistres sont indépendantes les unes des autres, c'est-à-dire, $C_{i,j}$et $C_{k,j}$ sont indépendants si $i \neq k$.

(H2) :

$$\forall i = 1, \dots, n, \forall j = 1, \dots, n-1, E\left(C_{i,j+1}/C_{i,1}, C_{i,2}, \dots, C_{i,j}\right) = C_{i,j} \times f_j$$

Cette seconde hypothèse suppose alors que le passage d'une année de développement à l'autre est décrit en termes d'espérance
Avec

$$\widehat{f_j} = \frac{\sum_{i=0}^{n-j+1} C_{i,j+1}}{\sum_{i=0}^{n-j+1} C_{i,j}}$$

H(3)

$$\forall i = 1, \dots, n, \forall j = 1, \dots, n-1, il\ existe\ un\ paramètre\ \sigma_j$$
$$Var\left(C_{i,j+1}/C_{i,1}, C_{i,2}, \dots, C_{i,j}\right) = C_{i,j} \times {\sigma_j}^2$$

Dans le raisonnement de Mack propose également un estimateur sans biais de σ_j^2, noté S_j^2,

$$\begin{cases} \forall 1 \leq j \leq n-2, S_j^2 = \dfrac{1}{n-j-1} \displaystyle\sum_{i=1}^{n-j} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - f_j\right)^2 \\ j = n-1, \qquad {S^2}_{n-1} = \min\left(\dfrac{{S^4}_{n-2}}{{S^2}_{n-3}}, \min\left({S^2}_{n-2}, {S^2}_{n-3}\right)\right) \end{cases}$$

Le modèle de Mack considère implicitement que les nombres cumulés suivent une distribution normale avec une moyenne et une variance qui sont décrites dans les hypothèses (H2) et (H3).

1.2. <u>La mesure de l'incertitude à l'ultime :</u>

Il s'agit désormais de quantifier l'incertitude sur le nombre IBNR ainsi estimées, c'est à- dire l'erreur de prédiction du modèle de provisionnement utilisé :
Soit D_nl'information disponible au moment de l'estimation des provisions, l'erreur de prédiction $MSEP\left(C_{i,n}\right)$ et l'estimation du nombre ultime est définie comme suit :

$$MSEP\left(\hat{C}_{i,n}\right) = E\left(\left(\hat{C}_{i,n} - C_{i,n}\right)^2 / C_{i,j} : i + j \leq n+1\right)$$

$$= Var\left(C_{i,n}/C_{i,j}: i+j \leq n+1\right) + \left(E\left(C_{i,n}/C_{i,j}: i+j \leq n+1\right) - \hat{C}^2{}_{i,n}\right)^2$$

En posant : $\widehat{IBNR}_i = \hat{C}_{i,n} - C_{i,n+i-1}$, la provision étudiée, : $\widehat{IBNR}_i - IBNR_i = \hat{C}_{i,n} - C_{i,n}$ conduit à $MSEP\left(\widehat{IBNR}_i\right) = MSEP\left(\hat{C}_{i,n}\right)$.

Puis avec $\hat{R} = \sum_{i=2}^{n} \hat{R}_i$ des estimateurs de $MSEP\left(\hat{R}_i\right)$ et $MSEP\left(\hat{R}\right)$ sont respectivement :

$$\begin{cases} \widehat{MSEP\left(\widehat{IBNR}_i\right)} = \widehat{C^2{}_{i,n}} \sum_{j=n-i+1}^{n-1} \frac{S^2{}_j}{\widehat{f^2}_j} \times \left(\frac{1}{\widehat{C_{i,j}}} + \frac{1}{\sum_{k=1}^{n-j} C_{k,j}}\right), pour\ i = 2, \dots, n \\ \widehat{MSEP\left(\widehat{IBNR}\right)} = \sum_{i=2}^{n} \widehat{MSEP\left(\widehat{IBNR}_i\right)} + \widehat{C_{i,n}} \times \left(\sum_{k=i+1}^{n} \widehat{C_{k,n}}\right)\left(\sum_{j=n-i+1}^{n-1} \frac{2 \times S^2{}_j}{\widehat{f^2}_j \sum_{i=1}^{n-j} C_{i,j}}\right) \end{cases}$$

1.3. Application du modèle du Mack sur le triangle de nombre.

Vérification des hypothèses de Mack

H1) cette hypothèse sera vérifiée, quant à elle, en effectuant un test reposant sur l'étude des coefficients individuels de développement et permettant de détecter l'existence de tendances sur ces facteurs pour chaque diagonale du triangle.

Afin de vérifier l'indépendance des mois de survenance, nous étudions l'influence de l'effet calendaire sur les facteurs de développement individuels.
Le test consiste à classer les coefficients individuels de développement de chaque colonne j (mois développement n° j) selon 3 groupes : les coefficients inférieurs à la médiane de la colonne notés " P" et les coefficients supérieurs à la médiane notés " G ". Pour valider l'absence d'effet calendaire, les nombres de coefficients classés P et G sur chaque diagonale doivent être proches, chaque facteur de développement ayant la même probabilité d'être classé G ou P. Il est à noter que, lorsque le nombre d'éléments de la colonne est impair, un des facteurs est égal à la médiane et sera dans ce cas classé dans un troisième groupe noté " *". Pour chaque diagonale n°k , la variable $n_k = \min(G_k, P_k)$ est définie où Gket Pk sont le nombre d'éléments de la diagonale classés G et P respectivement. Si l'hypothèse d'absence d'effet calendaire est justifié pour un inventaire, alors $\boldsymbol{Z_j}$ suit une loi binomiale de paramètres $n_k = G_k + P_k$ et p = 0.5 telle que :

$$\begin{cases} \mathrm{E}(\mathrm{Z_k}) = \frac{\mathrm{n_k}}{2} - \binom{\mathrm{n_k}-1}{\mathrm{m_k}} \times \frac{\mathrm{n_k}}{2^{\mathrm{n_k}}} \text{avec } \mathrm{m_k} = \left|\frac{\mathrm{n_k}-1}{2}\right| \\ \mathrm{V}(\mathrm{Z_k}) = \frac{\mathrm{n_k} \times (\mathrm{n_k}-1)}{4} - \binom{\mathrm{n_k}-1}{\mathrm{m_k}} \times \frac{\mathrm{n_k} \times (\mathrm{n_k}-1)}{2^{\mathrm{n_k}}} + \mathrm{E}(\mathrm{Z_k}) - \mathrm{E}(\mathrm{Z_k})^2 \end{cases}$$

L'approximation d'une loi normale pour Z est alors possible. L'absence d'effets calendaires est validée et donc l'indépendance des années de réclamation si Z appartient à l'intervalle de confiance défini par la formule suivante (sous l'hypothèse d'un intervalle de confiance à 95%) : $\lfloor \mathbb{E}(Z) \mp 1.96 * \sqrt{\mathbb{V}(Z)} \rfloor$.

diagonale	Gj	Pj	nj	Zj	mj	E(Zj)	Var(Zj)
2016_1	15	9	24	9	11,5	10,07	2,26
2016_2	7	13	20	7	9,5	8,24	1,90
2016_3	9	12	21	9	10	8,65	1,83
2016_4	10	15	25	10	12	10,49	2,19
2016_5	6	16	22	6	10,5	9,15	2,08
2016_6	8	15	23	8	11	9,57	2,01
2016_7	11	7	18	7	8,5	7,33	1,71
2016_8	13	7	20	7	9,5	8,24	1,90
2016_9	19	8	27	8	13	11,41	2,37
2016_10	25	8	33	8	16	14,19	2,92
2016_11	14	14	28	14	13,5	11,91	2,62
2016_12	20	8	28	8	13,5	11,91	2,62
2017_1	13	15	28	13	13,5	11,91	2,62
2017_2	16	8	24	8	11,5	10,07	2,26
2017_3	21	7	28	7	13,5	11,91	2,62
2017_4	20	8	28	8	13,5	11,91	2,62
2017_5	15	14	29	14	14	12,33	2,55
2017_6	20	9	29	9	14	12,33	2,55
2017_7	21	7	28	7	13,5	11,91	2,62
2017_8	16	15	31	15	15	13,26	2,74
2017_9	14	14	28	14	13,5	11,91	2,62
2017_10	27	14	41	14	20	17,93	3,64
2017_11	21	18	39	18	19	16,99	3,46
2017_12	19	11	30	11	14,5	12,83	2,80
2018_1	20	17	37	17	18	16,06	3,28
2018_2	28	11	39	11	19	16,99	3,46
2018_3	17	19	36	17	17,5	15,62	3,35
2018_4	27	11	38	11	18,5	16,56	3,53
2018_5	25	12	37	12	18	16,06	3,28
2018_6	20	10	30	10	14,5	12,83	2,80
2018_7	18	18	36	18	17,5	15,62	3,35
2018_8	22	16	38	16	18,5	16,56	3,53
2018_9	22	15	37	15	18	16,06	3,28
2018_10	22	14	36	14	17,5	15,62	3,35
2018_11	20	15	35	15	17	15,12	3,10
2018_12	28	7	35	7	17	15,12	3,10

Z	437	
E(Z)	402,00	
Var(Z)	525,00	
alpha	5,0%	
intervalle 0%	**357,09**	**446,91**
	H1 est acceptée	

Tableau 13 : Résultat de test d'indépendance

H2) L'existence d'une relation linéaire est mise en évidence entre les nombres cumulés d'un mois de déroulement à l'autre (voir : partie de la méthode de Chain Ladder)
H3) Cette hypothèse peut se faire de manière graphique à l'aide des résidus standardisés des nombres observés.

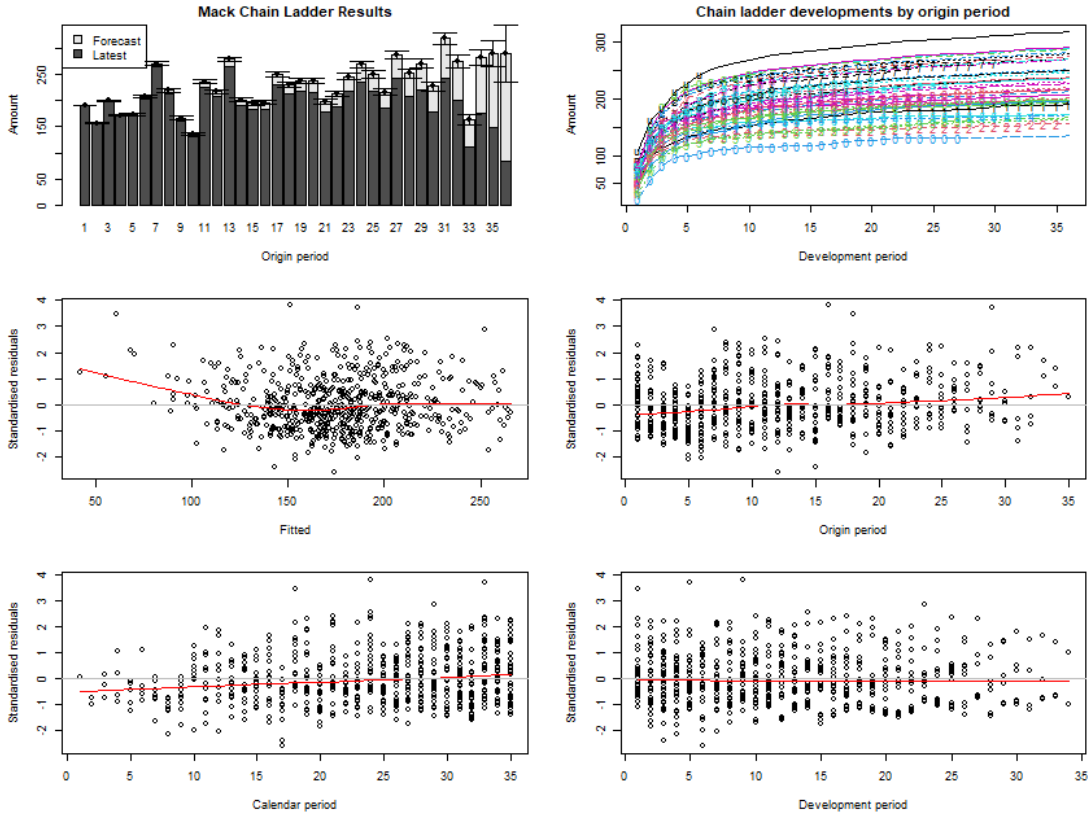


Figure 83 : Comparaison entre le nombre de sinistre estimée et le nombre déjà survenu, à gauche, et les cadences de sinistre prédites par la méthode Chain Ladder, en bas le critère aléatoire des résidus par mois de développement et par mois d'origine

Que l'on regarde les résidus par mois d'origine, de développement ou par mois calendaire, on observe une distribution de ces derniers autour de zéro. De plus, ils semblent aléatoirement distribués, aucune tendance particulière ne se dégage. Enfin, si on suppose que les résidus sont indépendants et identiquement distribués selon une loi normale centrée réduite, ils doivent être compris entre -2.5 et 2.5 dans 99% des cas. D'après les graphiques, ceci semble bien être le cas. D'après toutes ces remarques, on peut donc supposer que le modèle de Mack est adapté aux données sur ce segment.

Estimation de σ^2 pour chaque j :

A l'aide de R nous pouvons obtenir l'estimation de l'écart type pour les mois de développement :

```
> M$f.se
 [1] 0.055260145 0.014499502 0.005702002 0.004392480 0.003900171 0.002463493 0.002604786 0.002312802 0.002423367 0.002196548
[11] 0.002513422 0.001102724 0.001917617 0.001560697 0.001200676 0.001536905 0.001738934 0.001601025 0.001157668 0.001487520
[21] 0.001559785 0.001209078 0.001722288 0.001271254 0.001078850 0.001317592 0.001184404 0.001736282 0.001740935 0.001812021
[31] 0.001996193 0.004446088 0.002062603 0.002648042 0.003489661
>
```

A l'aide du vecteur des estimations de l'écart type, et sous l'hypothèse de normalité du nombre estimé des IBNR, l'intervalle de confiance (à 95%) du nombre d'IBNR se définit comme suit :

$$\left[\widehat{IBNR} - 1{,}96 \times se\left(\widehat{IBNR}\right); \widehat{IBNR} + 1{,}96 \times se\left(\widehat{IBNR}\right)\right]$$

L'intervalle de confiance de nombre des sinistres tardifs est $[95.24\,;154.83]$ avec un nombre moyen IBNR= 125.04, cet intervalle de confiance ne s'écarte pas beaucoup de la moyenne IBNR, ce qui montre un risque moins important dans l'estimation de nos nombres des tardifs.

2. Approche avec Bootstrap non paramétrique.

2.1. Rappels sur les GLM utilisés dans les modèles stochastiques :

Les modèles linéaires généralisés ont été introduits en 1972 par J. Nelder et R. Wedderburn. Ils sont une généralisation du modèle linéaire normal et sont formés de trois composantes : la composante aléatoire, la composante systématique et la fonction de lien.

- La composante aléatoire

Nous cherchons toujours à expliquer les variables réponses ($X_{i,j}$). Nous supposons maintenant qu'elles suivent une loi de probabilité de type exponentielle. Leur densité est définie par la formule suivante :

$$f(x_{i,j};\theta_{i,j};\phi) = exp\left\{\frac{[\theta_{i,j}x_{i,j} - b(\theta_{i,j})]w_{i,j}}{\phi} + c(x_{i,j};\phi)\right\}$$

Où : $\theta_{i,j}$est un paramètre réel, nommé paramètre naturel
Φ : un paramètre de dispersion strictement positif
$w_{i,j}$: est une pondération (=1 par la suite)
b et c sont des fonctions caractéristiques du modèle, b étant deux fois dérivables à valeurs dans R et c à valeurs dans R^2.

- La composante systématique :

Soit M la matrice de régression et ζ le vecteur des paramètres. La composante systématique η est notée et est définie par $\eta = M\zeta$.

- La fonction de lien :

Notée g, c'est la fonction qui fait le lien entre la composante aléatoire et la composante systématique. Une fois ces trois composantes définies, nous avons alors :

$$\begin{cases} \mu_{i,j} = g^{-1}(\eta_{i,j}) \\ E(X_{i,j}) = \mu_{i,j} \\ V(X_{i,j}) = \Phi V(\mu_{i,j}) \end{cases}$$

Nous nous sommes particulièrement intéressés aux familles Poisson sur-dispersé et Gamma à lien log ce qui implique $\boldsymbol{\mu_{i,j} = e^{\mu+\alpha_i+\beta_j}}$:

Famille	Fonction de répartition/Densité	Variance
Poisson $\mathbf{P}(\boldsymbol{\lambda})$	$P(X = x) = e^{xln(\lambda)-\lambda+c(x)}$	$V(\mu) = \mu$
Gamma $\boldsymbol{\Gamma\left(\nu;\frac{\nu}{\mu}\right)}$	$f(x) = exp\left(\left(-\frac{x}{\mu} - ln(\mu)\right)\nu + c(x;\nu)\right)$	$V(\mu) = \mu^2$

L'estimation des incréments est alors donnée par $\boldsymbol{\widehat{\mu_{i,j}} = e^{\hat{\mu}+\widehat{\alpha_i}+\widehat{\beta_j}}}$.

2.2. Bootstrap non paramétrique avec le modèle ODP

En pratique deux lois sont proposées pour réaliser un Bootstrap GLM, nous trouvons la loi Gamma et la loi de Poisson. Le choix de la loi de poisson sur-dispersée est basé sur la nature de triangle de nombre qui ne présente pas des valeurs extrêmes ou atypiques, et par conséquent, une courbe de distribution homogène non dispersée est jugée adéquat. Or la courbe de distribution selon une Gamma est plus étendue avec une queue plus épaisse

La procédure Bootstrap

Le Bootstrap est une méthode de simulation basée sur le rééchantillonnage des données avec un tirage aléatoire avec remise.

L'utilisation du Bootstrap suppose que les éléments de l'échantillon de départ soient indépendants et identiquement distribués (iid). Les variables ($\boldsymbol{X_{i,j}}$) ne sont en général pas identiquement distribuées. Il est donc préférable d'avoir recours aux résidus du modèle, en particulier les résidus de Pearson car plus simples à calculer. Le calcul des résidus se réalise avec un modèle ODP (nous renvoyons le lecteur à l'annexe 2 pour plus de détail), en effet le modèle Log Poisson donnant les mêmes résultats que la méthode Chain-Ladder, l'utilisation de cette dernière est une solution pour diminuer le temps de calcul.

Remarques :

- Pour corriger le biais dans la comparaison des estimations analytiques et Bootstrap d'erreur de prédiction, il est conseillé d'ajuster les résidus en intégrant le nombre de paramètres de régression dans l'erreur Bootstrap de prédiction. On définit les résidus ajustés par $\boldsymbol{r_{i,j}^{adj} = \sqrt{\frac{n}{\frac{1}{2}n(n+1)-2n+1}} \times r_{i,j}}$ où n est le nombre d'éléments de l'échantillon

La procédure Bootstrap dans le cadre du provisionnement est la suivante :

- $\boldsymbol{X_{i,j}} \sim ODP(\boldsymbol{m_{i,j}},\phi)$ par hypothèse
- Les coefficients $\boldsymbol{\widehat{m}_{i,j}}$ et $\boldsymbol{\widehat{\varphi_j}}$ sont estimés

- Calcul des résidus de Pearson de la loi de Poisson (dans ce cas c'est $\frac{X_{i,j}}{\varphi_j}$ qui suit une loi de Poisson et non $X_{i,j}$) $r_{i,j} = \frac{X_{i,j} - \widehat{m}_{i,j}}{\sqrt{\widehat{\varphi_j} \cdot \widehat{m_{ij}}}}$
- Ajustement des résidus $r_{i,j}^{adj} = \sqrt{\frac{n}{\frac{1}{2}n(n+1)-2n+1}} \times r_{i,j}$
- Les résidus sont rééchantillonnés de façon aléatoire et avec remise
- Pour chaque nouveau triangle de résidus qui est renommé $r_{i,j}^{b}$
  - Calcul de la table des pseudo-montants incrémentés en inversant la formule des résidus
  - Par la méthode de Chain Ladder nous calculons à nouveau la partie inférieure de triangle
  - Estimation des montants finaux et déduction des réserves à constituer

Schéma de la procédure à effectuer B fois :

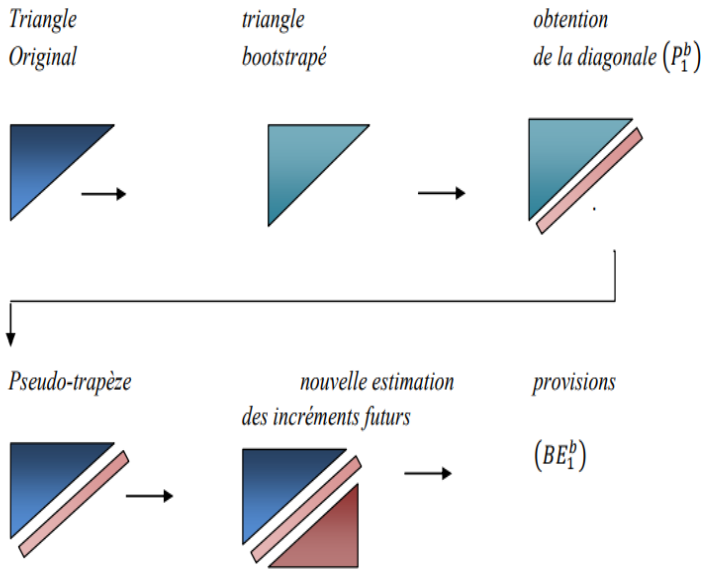


Figure 84 : Schéma de la procédure Bootstrap

2.3. Application de la méthode de Bootstrap :

Le nombre de simulation choisi est de 10 000. Généralement dans des simulations type Monte Carlo, le nombre minimum de simulations nécessaires pour que les résultats soient stables est de 5000. Puisque le temps de simulation du programme n'est pas exorbitant, il semble plus prudent de prendre 10 000 simulations.
Après avoir générer un échantillon de 10 000 il est possible de déduire facilement toutes les statistiques liées à la variabilité de la distribution, notamment les quantiles et l'écart type du nombre IBNR.
Pour éviter toute ambiguïté, le nom IBNR qui figure dans les sorties de R représente la différence entre le nombre ultime des sinistres et la diagonale cumulée du triangle de nombre.
A l'aide de la sortie R, ci-dessous quelques statistiques intéressantes de la procédure de Bootstrap ODP :

```
> Bot
BootChainLadder(Triangle = Tri_Nbre, R = nbsim, process.distr = "od.pois"
```

	Latest	Mean Ultimate	Mean IBNR	IBNR.S.E	IBNR 75%	IBNR 95%
1	193	193	0.000	0.000	0	0
2	156	156	0.000	0.000	0	0
3	203	203	0.000	0.000	0	0
4	173	173	0.000	0.000	0	0
5	177	177	0.000	0.000	0	0
6	214	214	0.000	0.000	0	0
7	274	274	0.237	0.643	0	1
8	226	226	0.187	0.571	0	1
9	167	167	0.138	0.481	0	1
10	139	139	0.120	0.439	0	1
11	241	241	0.323	0.754	1	2
12	219	219	0.291	0.698	0	2
13	281	281	0.368	0.806	1	2
14	204	204	0.268	0.677	0	2
15	205	205	0.271	0.676	0	2
16	195	195	0.264	0.665	0	2
17	246	246	0.318	0.759	1	2
18	244	244	0.389	0.828	1	2
19	248	248	0.454	0.883	1	2
20	253	254	0.505	0.937	1	2
21	210	211	0.574	0.955	1	2
22	219	220	0.673	1.018	1	3
23	273	274	0.990	1.268	2	3
24	281	282	1.283	1.381	2	4
25	264	266	1.554	1.509	2	4
26	226	228	1.780	1.559	3	5
27	307	310	2.742	1.939	4	6
28	274	278	3.573	2.202	5	8
29	297	303	6.369	2.882	8	11
30	243	250	6.759	2.955	9	12
31	323	334	10.596	3.712	13	17
32	285	295	10.483	3.724	13	17
33	179	186	7.352	3.056	9	13
34	298	311	13.217	4.176	16	20
35	331	347	16.407	4.675	19	24
36	610	643	32.595	6.778	37	44

```
                  Totals
Latest:         8,878.0
Mean Ultimate:  8,999.1
Mean IBNR:        121.1
IBNR.S.E           18.4
Total IBNR 75%:   133.0
Total IBNR 95%:   152.0
>
```

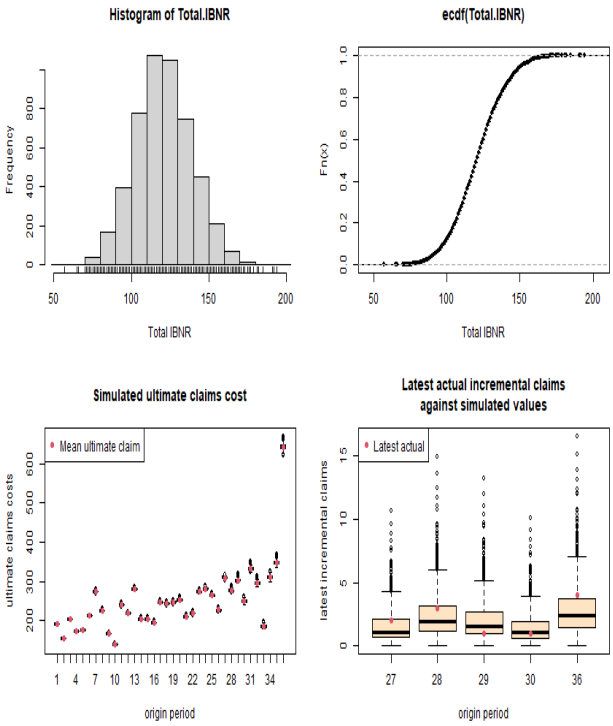


Figure 85 : Sortie R du résultat de l'estimation de nombre IBNR avec la méthode de Bootstrap.

Après avoir générer un échantillon de 10 000 il est possible de déduire facilement toutes les statistiques liées à la variabilité de la distribution.

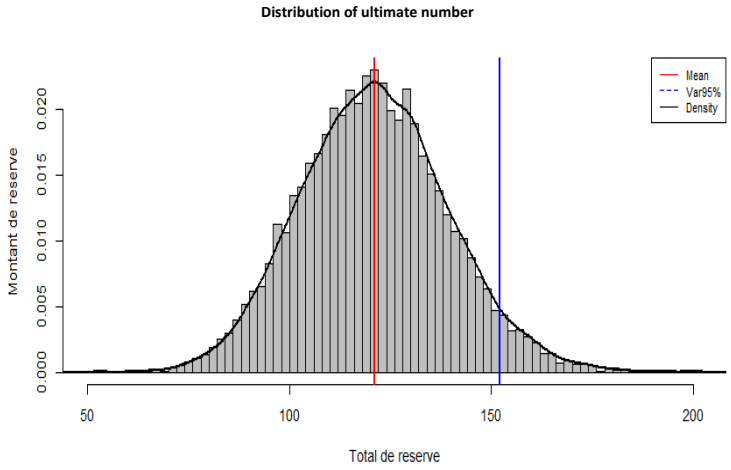


Figure 86 : Graphique de distribution de nombre de sinistres

La moyenne de nombre IBNR, obtenue par le Bootstrap des résidus de Pearson de la modélisation ODP, s'établie à 121.1 sinistres avec un quantile à 95% de 152 sinistres.
Nous remarquons que la courbe de distribution des IBNR est symétrique par rapport à la moyenne.
A noter que la distribution des IBNR de nombre est fortement impactée par le choix du nombre de simulations et que la convergence peut ne pas être pratiquement atteignable.

Equi-distribution des résidus

L'hypothèse généralement retenue est que les résidus sont distribués selon une loi normale centrée-réduite. Nous confrontons donc nos données à cette hypothèse de distribution.

Tests graphiques

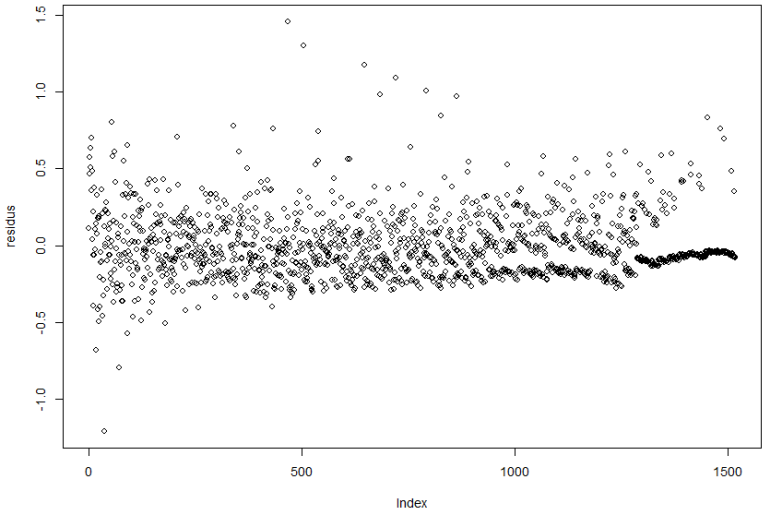


Figure 87 : Graphe des résidus du modèle de Bootstrap.

Le Q-Q Plot ou diagramme quantile-quantile est un test d'adéquation graphique qui repose sur la comparaison des quantiles empiriques et théoriques. Les quantiles théoriques sont ici ceux d'une loi normale centrée réduite tandis que les quantiles empiriques sont évalués à partir des résidus. En cas d'adéquation des données à la loi normale centrée réduite, nous devrions observer un alignement des points le long de la bissectrice

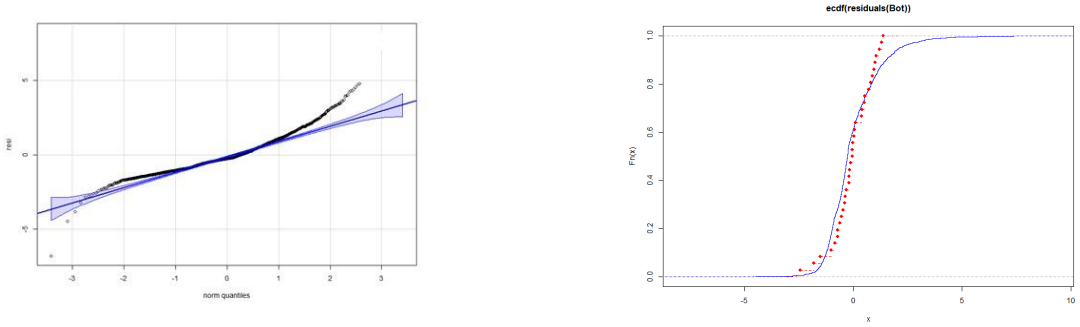


Figure 88 : Graphe à droite : Comparaison des fonctions de répartitions théoriques et empiriques des résidus, Graphe à gauche : QQ-plot des résidus du Bootstrap

Les QQ-plots montrent un alignement des couples (quantiles théoriques, quantiles empiriques) le long de la bissectrice pour les valeurs centrales. Bien que cet alignement ne soit pas observé pour les valeurs les plus extrêmes, nous conclurons dans un premier temps que ce test confirme l'hypothèse de normalité des résidus. Compte tenu du faible nombre de données, les tests graphiques de comparaisons des fonctions de répartition montrent une bonne adéquation des résidus à la loi normale centrée réduite, avec toujours cependant une réserve quant aux valeurs extrêmes.

<u>Test analytique : Kolmogorov-Smirnov :</u>

Nous utilisons à présent le test d'adéquation de Kolmogorov-Smirnov (fonction *ks.test()* de R) pour tester l'adéquation des données à la loi normale centrée réduite. L'hypothèse de base (H_0) du test étant que les données sont distribuées selon cette loi, nous rejetterons donc cette hypothèse pour un seuil $\alpha = 5\%$ si les p-values obtenues sont inférieures à 5%. Les résultats obtenus ne remettent donc pas en question notre hypothèse de distribution :

```
> ks.test(Residuals_boot,"pnorm")

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Residuals_boot
D = 0.16967, p-value = 0.3168
alternative hypothesis: two-sided
```

Figure 89 : Sortie R du test de Kolmogorov-Smirnov

L'étude des résidus a également globalement permis de valider l'hypothèse, généralement retenue, de distribution selon une loi normale centrée réduite avec une p-value correspondante supérieure à 5%.

Conclusion et limite :

La méthode du Bootstrap est très pratique car elle permet de s'affranchir d'hypothèses parfois contraignantes sur une famille de lois de probabilité. En particulier, cette méthode est très utilisée pour obtenir des meilleures estimations des réserves à partir de peu de donnée.
Avec la loi des grands nombres on pourra rapidement converger vers l'espérance des réserves (n=10 000).
Limite :
L'utilisation du Bootstrap pour des variables aléatoires à queues épaisses peut en pratique poser un certain nombre de problèmes. En effet, les très grandes valeurs ont alors tendance à apparaître beaucoup trop souvent. Le Bootstrap ne peut alors être utilisé pour estimer une moyenne ou une probabilité d'excès pour une valeur très éloignée de la moyenne.
Dernièrement, la technique de Bootstrap non paramétrique n'est pas très pertinente quand on ne dispose que d'un échantillon de faible taille. Prenons l'exemple d'un réassureur qui vient de s'engager dans une nouvelle branche d'activité, l'historique de la sinistralité de cette branche n'est donc pas assez grand pour utiliser cette technique. Dans ce cas, on peut employer un Bootstrap paramétrique. Au lieu de rééchantillonner les résidus, on les simule à partir d'une loi (Normale, Log-normale, Gamma… selon les résultats des tests d'adéquation). Le principe est ensuite exactement le même que pour le Bootstrap non paramétrique. Il faut noter aussi que les queues de distribution peuvent être sous-estimées quand on utilise un Bootstrap non paramétrique sur un petit triangle de résidus.

IV. Comparaison entre les méthodes et mise en place de calcul des IBNR :

1. Comparaison entre les méthodes

Les principales caractéristiques de la distribution qui feront office d'outils de comparaison entre les méthodes de provisionnement sont les suivantes :

- Moyenne de la distribution, ou Best Estimate des réserves
- L'écart-type de la distribution
- La Value-At-Risk de niveau 99,5% de la distribution
- Le coefficient de volatilité, qui est défini comme $\frac{ecart\ type}{espérance}$
- Le coefficient de comparaison entre Value-At-Risk et la meilleure estimation des IBNR $\frac{VaR_{99,5\%}-IBNR}{IBNR}$

Le tableau suivant résume les résultats :

```
$Totals
           Totals
IBNR 60%:     125
IBNR 95%:     152
IBNR 99%:     167
IBNR 99.5%:   173
```

	Méthode de Chain Ladder	Mack Chain ladder	Bootstrap
Moyenne de la distribution	125.04	125.04	121.1
Ecart type		15.2	18.4
VaR de niveau 99.5%		166	173
Coefficient de volatilité		12.2%	15.2%
Coefficient de comparaisons VaR/BE		32.76%	42.98%

Tableau 14 : Le tableau de comparaison entre les différentes méthodes d'estimation des IBNR.

Les résultats obtenus avec les méthodes de Mack Chain Ladder et Bootstrap sont proches, avec un nombre de sinistre tardif faible selon la méthode de Bootstrap, dans une approche prudentielle, il est préférable de retenir une méthode moins optimiste quant à la possibilité de récupération.
En revanche la méthode de Bootstrap affiche une variabilité élevée par rapport à la méthode de Mack. Nous constatons que la méthode de Mack est légèrement moins volatile autour de la moyenne par rapport à la méthode de Bootstrap, comme on l'a vu précédemment, le graphe des résidus du modèle de Mack semble présenter des structures aléatoires contrairement au graphe des résidus de l'approche bootstrap.
Le modèle de Mack semblerait être réellement adapté à notre portefeuille d'assurance.

2. Mise en place de calcul des IBNR.

A ce stade, nous avons estimé le nombre IBNR nécessaire qui s'établie à 125 nombres, il nous reste à effectuer une répartition par tranche d'âge et sexe, pour réaliser cette répartition, nous nous sommes basés sur la répartition du portefeuille.

Sexe	ageBand	% effectif	Age moyen	Nombre IBNR
Femme	(-1,20]	2,95%	8,6	2
	(20,65]	15,33%	54,2	10
	(65,75]	17,51%	71,0	12
	(75,80]	15,62%	78,3	10
	(80,85]	21,51%	83,1	14
	(85,90]	17,87%	87,8	12
	(90,120]	9,22%	93,2	6
Total Femme		100,00%	75,4	66
Homme	(-1,20]	8,77%	8,0	5
	(20,65]	15,83%	54,7	9
	(65,75]	20,61%	70,7	12
	(75,80]	15,71%	78,2	9
	(80,85]	17,39%	82,9	10
	(85,90]	13,34%	87,9	8
	(90,120]	8,35%	93,4	5
Total Homme		100,00%	70,2	59

Tableau 15 : Répartition du nombre des IBNR par sexe et par tranche d'âge

Point d'attention :

Pour les sinistres avec le statut en attente, nous avons retenu l'hypothèse de tenir en compte le nombre de ces sinistres avec une probabilité d'acceptation de sinistres. Le nombre de sinistre faible ne permet pas de développer une méthode statistique robuste. Ces sinistres en statut ''attente'' sont majoritairement des sinistres en cours d'investigation de la part de la cédante.
Le nombre de sinistre en attente se présentent comme suit :

Mois	Année 2018	Année 2017
Janvier	10	1
Février	15	2
Mars	20	2
Avril	6	1
Mai	5	1
Juin	5	1
Juillet	5	1
Août	7	1
Septembre	4	0
Octobre	9	1
Novembre	10	1
Décembre	7	1
Total	103	11

Tableau 16 : Le nombre de sinistres en attente

Nous comptons 114 sinistres en attentes qu'on ajoute au nombre ultime avec une probabilité d'acceptation établie à 85%.

Conclusion :

Cette section nous a permis de calculer le nombre des sinistres tardifs, le recours à des méthodes stochastiques permettent d'avoir une distribution des IBNR et d'estimer le risque au tour du scénario central. En effet, l'approche simulatoire ODP a permis d'obtenir un nombre IBNR légèrement faible par rapport à la méthode de Mack Chain Ladder, cependant, la volatilité des IBNR est un peu élevée dans le modèle de Mack.

Nous sommes conscients de la difficulté de conclure de manière précise sur ces résultats. En effet, par prudence nous avons décidé de retenir la méthode de Mack est par conséquent estimé un nombre tardif de 125 sinistres à qui on ajoute le nombre de sinistre en attente de 114 dont le nombre doit être multiplié par le taux d'acceptation à 85%. Ainsi le nombre total de sinistre tardif est $\mathbf{125} + \mathbf{114} * \mathbf{85}\% = \mathbf{222}\ \boldsymbol{sinistres}$.

La répartition finale de nombre IBNR de sinistres et des sinistres avec le statut ''en attente'' se présente comme suit :

	IBNR + sinistre en attente			
Sexe	ageBand	% effectif	Age moyen	Nombre IBNR
Femme	(-1,20]	2,95%	8,6	3
	(20,65]	15,33%	54,2	18
	(65,75]	17,51%	71,0	21
	(75,80]	15,62%	78,3	18
	(80,85]	21,51%	83,1	25
	(85,90]	17,87%	87,8	21
	(90,120]	9,22%	93,2	11
Total Femme		100,00%	75,4	118
Homme	(-1,20]	8,77%	8,0	9
	(20,65]	15,83%	54,7	16
	(65,75]	20,61%	70,7	21
	(75,80]	15,71%	78,2	16
	(80,85]	17,39%	82,9	18
	(85,90]	13,34%	87,9	14
	(90,120]	8,35%	93,4	9
Total Homme		100%	70,2	104

Tableau 17 : Répartition du nombre IBNR et nombre de sinistre en attente par sexe et par tranche d'âge.

Chapitre 5 : Calcul du provisionnement technique en dépendance

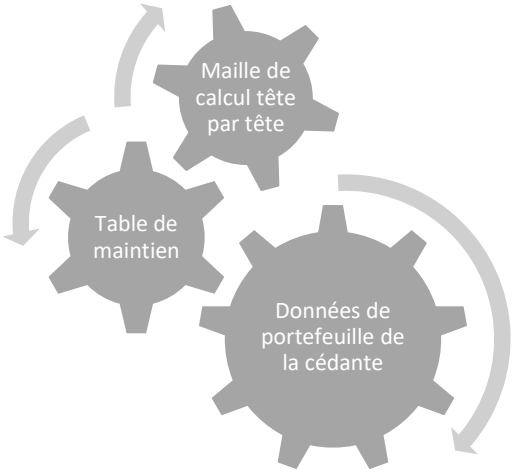
L'objectif de Cette section est d'exploiter les travaux de modélisation présentés dans les sections précédentes pour apporter une estimation des provisions technique (PM+IBNR) pour notre produit de dépendance.

I. La mise en place du modèle

Comme mentionné auparavant, le calcul de la provision technique revient au calcul de la PM de rente en dépendance à laquelle on ajoute le montant des IBNR.

1. Le calcul de la PM de rente en dépendance :

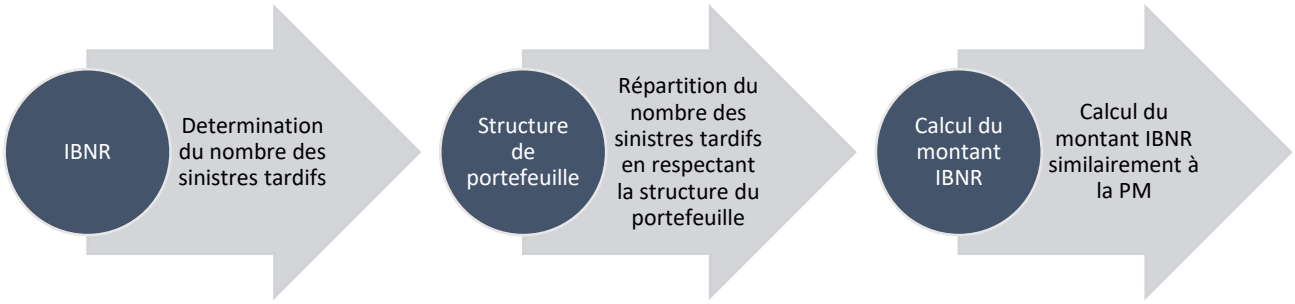
La Provision mathématiques de rente correspond à la provision pour rente en cours de service. A fortiori, les sinistres sont déjà survenus et connus auprès de l'assureur et par la suite auprès du réassureur.
Les lois biométriques d'expérience réalisées dans les sections du Chapitre 3 : Modélisation du risque de dépendance **:** permettent d'effectuer des projections sur la population observée à une date d'arrêté. Ces dernières permettent d'estimer l'évolution de cette population et d'en déduire les montants de provisions à constituer pour faire face à ses engagements contractuels vis-à-vis de la cédante.
Ci-dessous une représentation schématique du processus actuel de calcul de la PM :



Grâce aux inputs de la cédante, la PSAP est calculée comme suit :
$PM = a^{t,s} \times R$ Où $a^{t,s} = \sum_{k=1}^{T} {v_k}^k \prod_{j=0}^{k-1} {}_{j|1}p_x^{d,S}$ correspond au capital constitutif d'une rente dépendance payable à terme échu pour un assuré d'âge x en dépendance.
${}_{k|1}p_x^{d,S}$: la probabilité qu'un individu de sexe S, dépendant totalement d'âge x reste dans le même état pour les âges compris entre $x + k \; et \; x + k + 1$.
Avec : $v = 1/\, i + 1$ où i est le taux d'intérêt et R montant de l'annuité de la garantie.
La réglementation locale du réassureur oblige d'utiliser la courbe des taux ''locked-in'' c'est-à-dire le taux estimé lors de la première comptabilisation forwardisé pour le compte de résultat ''Income statement'' et utiliser la courbe actuelle ''current interest rate'' pour le calcul des provisions techniques qui figurent dans le bilan ''Balance sheet''.
Remarque :
Comme la table de maintien était construite sur la base des données mensuelles, nous concluons qu'aucune transformation de rente n'est à prévoir.

Le calcul des IBNR.

Le Chapitre 4 : Méthode d'estimation des tardifs et leur application nous a permis de déterminer le nombre de sinistres tardifs qui doivent être provisionnés. La démarche que nous avons retenue et puis de projeter ce nombre de sinistre tardif au rythme des sinistres survenus et connus, autrement, le nombre de sinistre tardif sera bifurqué en respectant la structure du portefeuille pour ensuite utiliser les tables de maintien retenus pour le calcul de la PM afin de projeter dans l'avenir les dates possibles de paiement.



2. Mise en place du modèle

L'estimateur non paramétrique de Kaplan Meier est l'estimateur le plus simple à mettre en place et le plus intuitif, ce dernier permet d'estimer la fonction de survie sans faire d'hypothèses sur la distribution des temps de survie. En outre cette estimateur présente l'avantage d'être le moins couteux en termes de gain en temps de calcul.

Pour aider le lecteur à comprendre la méthode de calcul nous allons prendre un exemple concret d'un adhérent en dépendance avec les caractéristiques suivantes :

- Age : 73 ans
- Sexe : Homme
- 26 mois qu'il est en état de dépendance

Les caractéristiques suivantes nous permettent de déterminer le taux de maintien à utiliser, le calcul de la provision aura pour objectif de s'assurer le versement des rentes après cette date, de ce fait la probabilité avant les 26 mois est égale à 1.

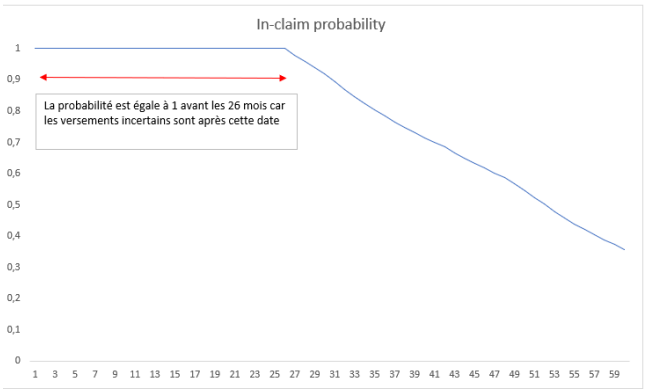


Figure 90 : La probabilité conditionnelle de survie pour un adhérent homme, d'âge 73 ans et d'ancienneté de 26 mois.

Après avoir déterminé la courbe de survie, nous pouvons à l'aide de la courbe du taux à terme -Forward rate- (Cf Annexe 5) de déterminer le nombre d'annuités à servir dans ce cas, il sera de l'ordre de 23.

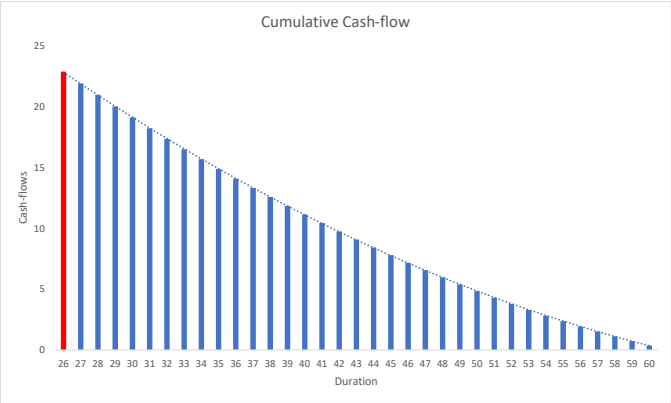


Figure 91 : Exemple de montant de la provision mathématique pour une rente d'1$

On remarque que le montant de la provision mathématique pour une rente d'1$ diminue au fil des années. Plus la personne dépendante vieillit plus elle risque de mourir, cela explique cette diminution.
Le calcul de la provision mathématique pour cet adhérent se détermine comme suit :

$$Reserve = \text{PV}\ Annuity\ number \times \min\left(Observed\ Average\ Paid\ claims,\ Average\ expected\ claim\right) \times reinsurance\ share$$

Application numérique :

$$Reserve = 22.9 \times 1\,001\ \times 25\% = 5730.72\$$$

II. Résultat final du montant de provision technique.

Le calcul tête par tête de la PM plus les IBNR nous a permis de déterminer le montant BEL de la garantie dépendance :

Amount in m USD	Cedant account	Reinsurance's Evaluation
In-payment	30.78	37.5
IBNR + OS pnding	2.25	2.14
Total reserve	**33.03**	**39.64**
Paid claims	5.26	5.26
Total BE	**38.29**	**44.9**
Total BE + PAD(*)	**40.20**	**47.15**
Deficit(-)/Surplus(+)		**-6.94**

Tableau 18 : Résultat total de montant de provision technique

Le montant USD 45 million est la somme dont l'entreprise devra disposer aujourd'hui pour permettre raisonnablement d'assurer ces engagements jusqu'à extinction du groupe fermé, le montant USD 2. 25 million qui correspond à la provision pour risque adverse qui a pour but de compenser l'incertitude existant sur l'estimation des flux futurs de trésorerie relative aux risques non financiers. Le montant de la provision pour déviation défavorable est déterminé de manière à ce que l'évaluation des engagements garanties soit suffisante à 75%.

CONCLUSION

Ce mémoire a notamment été un prétexte pour construire une loi de maintien en dépendance. Les caractéristiques de ce produit et notamment son système d'adhésion à la naissance et d'une couverture maximum de 60 mois nous ont amené à nous intéresser au caractère pertinent d'un tel produit dans le cadre de la modélisation de la dépendance.
Tout d'abord, un lourd travail préliminaire de gestion et de traitement des bases est nécessaire, entraînant toujours un biais d'estimation. Une fois cette étape décisive franchie, l'estimation a été possible.
Pour obtenir notre modèle de provisionnement nous avons déterminé les lois de maintien pour le calcul de la provision mathématique de rente et le montant des IBNR. Les lois de maintiens ont été estimés selon trois approches :

- L'utilisation de l'estimateur de Kaplan Meier en fonction du sexe et de la tranche d'âge de l'assuré afin d'obtenir des taux bruts sur lesquels nous avons appliqué un lissage de Whittaker-Henderson, les résultats nous ont apparu satisfaisants et en adéquation avec les données de notre portefeuille.
- La mise en place du modèle semi-paramétrique de Cox auquel nous avons testé l'hypothèse de proportionnalité et avons conclu sa violation. Pour contourner ce problème nous avons proposé un modèle de Cox staratifié par tranche d'age et incluant uniquement le sexe comme variable descrimanante.
- L'élaboration d'un modèle multi-états à l'aide de la théorie markovienne, nous avons proposé un modèle markovien homogène par morceau ainsi qu'un modèle semi-markovien dont la loi de probabilité des temps de station dans un état donné est estimée par une loi de Weibull.

Pour l'estimation des IBNR, nous avons fait appel à des méthode de triangulation notamment la méthode détérministe de Chain-Ladder et les deux méthodes stochastiques les plus utulisées, Mack et le Bootstrap sur les résidus ODP, nous avons ensuite mis en évidence les pricnipales limites de chaque méthodes. Enfin, le modèle de Mack a été retenu grace à la faible variabilité du montant IBNR estimé.

Bien qu'essayant d'être aussi complet que possible dans la démarche scientifique visant à proposer une solution fiable du problème de construction d'une rente adaptée au niveau de dépendance d'un assuré, nos travaux ne se sont pas étendus à certains champs qui mériteraient une attention certaine, notamment l'hypothèse d'indépendance des lois de durée par rapport aux variables descriptives, et le manque d'une évaluation approfondie de l'écart type pour les lois de durée.

L'absence d'une table de référence pour tester notre modèle restreint l'évalation de la robustesse de notre modèle, en revanche le rapprochement de nos provisions techniques à celles de la cédante nous parait prudent.

Finalement après avoir mis en place notre modèle de provisionnement, le montant total des provisions techniques s'est établi à USD 47.51 million incluant la provision pour deviation adverse. Toutefois, un suivi régulier des évolutions du portefeuille doit être effectué pour réaliser des projections plus pertinentes en vue du pilotage.

Le risque dépendance est récent et il est susceptible d'évoluer dans les années à venir. Ainsi, cette étude n'est pas définitive et elle pourrait être révisée selon les évolutions probables en matière de réglementation et de jurisprudence.

BIBLIOGRAPHIE

[1] ABECERA N. Actuariat branche dommages. Cours CNAM 2021
[2] ALEGRE A., POCIELLO E., Pons M.A, SArrasi J., Varea J., VICENTE A. [2002], Actuarial valuation of long-term care annuities
[3] AMERICAN ACADEMY OF ACTUARIES. [2021], Developed by the Long-Term Care Practice Note Work Group of the Health Practice Council of the American Academy of Actuaries, Long-Term Care Insurance Practice Note Work Group. « State of the U.S. Long-term Care Insurance Industry »
[4] ARMBRUSTER A. [2010], Estimation du Best Estimate sur le risque dépendance, mémoire d'actuaire, Diplôme Universitaire d'actuaire de Strasbourg
[5] ATCHAMA C. [2009], Suivi et optimisation d'un contrat dépendance collectif à adhésion facultative, mémoire d'actuariat, l'Université Paris DAUPHINE.
[6] Axco. Israel : Life ans benefits. Insurance Market Report, 2019
[7] BESSIOUD R. [2015], Modélisation de la partie attritionnelle de la provision RC médicale, mémoire d'actuaire, ISFA.
[8] BIESSY G. [2016], Modélisation semi-markovienne de la perte d'autonomie chez les personnes âgées : application à l'assurance dépendance, Thèse de doctorat, Université Paris-Saclay
[9] BIESSY G. [2013], Construction d'un modèle multi-états semi-markovien dans le contexte de l'assurance dépendance, mémoire d'actuaire.
[10] BONNET C, CAMBOIS E, CASES C, and GAYMU J. La dépendance : quelles différences entre les hommes et les femmes ? Gérontologie et société, 36(2) :55–66, 2013.
[11] CHARPENTIER A. et al. [2010] « Mesurer le risque lors du calcul des provisions pour sinistres à payer », Revue Risques, no 83.
[12] David Roxbee Cox. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2) :187–220, 1972
[13] DE LA MORINERIE A. [2016] Conceptualisation d'un modèle multi-états en arrêt de travail et application à une loi d'incidence en incapacité, mémoire d'actuaire, ENSAE
[14] DELFAU C. Assurance Dépendance : « Comment rendre l'assurance individuelle plus attractive pour les clients et les assureurs », thèse professionnelle 2009-2011 MBA, ENASS
[15] Denuit M. et Robert C. [2007], Actuariat des Assurances de Personnes. Economica
[16] ENGLAND V. Analytic and bootstrap estimates of prediction errors in claims reserving.1999. Insurance : Mathematics and Economics.
[17] DUPIN-NIVEL, L. [2014], Etude du risque dépendance : comportement des lois d'incidence et de longévité dans le temps, mémoire d'actuaire.
[18] FEUVRIER A ; [2011] Analyse de risque de dépendance et construction de tables d'éxperience, mémoire d'actuaire, EURIA
[19] FURET A [2006] : Impacts de la volatilité des IBNR dans une cotation en réassurance, mémoire d'actuaire
[20] GAUTHIER G. Simulation des processus de diffusion. 2009. 6-601-09 Simulation Monte Carlo
[21] GAUTHIER A. [2019], Comparaison de méthodes de constructiol} de table de maintien en invalidité, méoire d'actuaire, ISFA
[22] Guibert Q, [2015], Sur l'utilisation des mod`eles multi-´etats pour la mesure et la gestion des risques d'un contrat d'assurance, Thèse de doctorat, l'Universite Claude Bernard Lyon I
[23] GRAMBSCH P., THERNEAU T., [2000] « Modeling Survival Data: Extending the Cox Model», Springer.
[24] HAYNE Roger M. (1985): "An Estimate of Statistical Variation in Development Factor Methods", Casualty Actuarial Society.
[25] INSTITUT DES ACTUAIRES, Guide de provisionnement des sinistres en assurance non-vie, 2023, Rédigé par le Groupe de Travail sur le Provisionnement des Sinistres Non-Vie
[26] JACKSON, C. [2023], Multi-state modelling with R: the msm package, MRC Biostatistics Unit Cambridge, U.K
[27] JEQUIER A, [2016] Construction d'une table de maintien en incapacité et mesure du risque d'estimation, mémoire d'actuaire Dauphine

[28] Kaplan E.L, Meier P. Non parametric estimation from incomplete observation. Journal of the American Statistical Association ; 1958 ; 53, 457-481
[29] KIPKOECH CP. [2012] An actuarial multi-state modelling of long term care insurance products – a case study of the kenyan insurance industry. UNIVERSITY OF NAIROBI
[30] LEMAIRE J. [2014] « Impact du provisionnement en norme actuelle et en norme Solvabilité II », Mémoire d'actuariat.
[32] LEPEZ V. [2006]. Trajectoires en Dépendance des personnes âgées : modélisation, estimation et application en assurance vie. Mémoire d'actuariat, Centre d'Etudes Actuarielles
[33] LIAGRE A. Modélisation et réassurance de la dépendance dans le contexte de solvabilité 2. Mémoire d'actuariat, 2017.
[34] LIN. N [2012] Etude des lois de maintien et de passage en dépendance, mémoire d'actuaire, ISFA
[35] Lugand A. [2010], Evaluation des barèmes de provisions du contrat Dépendance du BCAC, mémoire d'actuaire, CNAM
[36] LIU Huijuan et VERRALL Richard (2010): "Bootstrap Estimation of the Predictive Distributions of Reserves Using Paid and Incurred Claim", VARIANCE, Casualty Actuarial Society.
[37] MACK T. [1993] « Distribution-free calculation of the standard error of chain ladder reserve estimates », Astin bulletin, vol. 23, no 02, p. 213-225.
[38] MATHIEU, E. [2006]. Modélisations multi-états markoviennes et semi-markoviennes. Applications à l'état de santé des patients atteints par le virus du SIDA. Thèse, Université de Montpellier.
[39] Ouhbi B., Limnios N., Nonparametric Estimation for Semi-Markov Processes
Based on its Hazard Rate Functions, Statistical Inference for Stochastic Processes, vol. 2, 1999.
[40] PLANCHET F. Statistiques des modèles non paramétriques. Cours ISFA 2021
[41] PLANCHET F., THEROND P. [2011] « Modélisation statistique des phénomènes de durée », Economica
[42] PINHEIRO PJR. [2003], Bootstrap methodology in claim reserving. Journal of Risk and Insurance.
[43] PUTTER H, FIOCCO M, GESKUS R. B. [2007], Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine, 26:2389–2430.
[44] Renshaw, Verall. A stochastic model underlying the Chain-Ladder technique.1998. British Actuarial Journal.
[45] Shu, Y., J. Klein, and M.-J. Zhang (2007). Asymptotic theory for the Cox semi-Markov illness-death model. Lifetime Data Anal 13, 91–117.
[46] THATCHER A.R. [1999], The Long-term Pattern of Adult Mortality and the Highest Attained Age, Journal of the Royal Statistical Society.
[47] TSAPBAZE E. [2013]. MODELISATION DYNAMIQUE DES ETATS DE DEPENDANCE, mémoire d'acatuaire ENSAE ParisTech
[48] WANNEVEICH M, [2016] Estimations et projections d'indicateurs de santé pour maladies chroniques et prise en compte de l'impact d'interventions, Thèse de doctorat de l'Université de Bordeaux.
[49] WINTER P. [2005] « Méthodes bidimensionnelles pour l'ajustement de lois de maintien d'expérience en arrêt de travail ». ISFA, Mémoire d'actuariat.
[50] XIE J., LIU C. [2000] « Adjusted Kaplan-Meier Estimator and Log-rank Test with Inverse Probability of Treatment Weighting for Survival Data », Statist. Med., vol. 00:1-6

ANNEXES

Annexe 1 : Lissage Whittaker-Henderson et Loess

On suppose disposer de N observations $(X_1, Y_1), \dots, (X_N, Y_N)$ où la variable à expliquer Y et le régresseur X sont continus. On cherche à ajuster un modèle de la forme :

$$Y_i = f(X_i) + \varepsilon_i$$

où les résidus $\varepsilon_i$sont supposés indépendants de loi $N(0, \sigma^2)$. L'estimation de $f(.)$ peut se faire à l'aide de techniques de lissage diverses. En particulier, $\hat{f}(X_i)$peut s'exprimer comme une combinaison linéaire des $Y_1, \dots, Y_N$,i.e :

$$\hat{f}(X_i) = \sum_{k=1}^{N} q_{ik}\, Y_k$$

où les poids $q_{ij} = q(X_i, X_k)$ dépendent du point $X_i$où la réponse doit être estimée. En définissant le vecteur $f = \left(f(X_1), \dots, f(X_N)\right)^T$, (2.16) se réécrit $\hat{f} = Q\, y$.

- **Lissage Whittaker-Henderson**

Cette méthode a été présentée en 1923 par E.T. Whittaker et complétée par R. Henderson en 1924. Son objectif est d'obtenir le meilleur compromis entre la fidélité et la régularité, c'est-à-dire entre l'adéquation aux données brutes et la régularité des estimations. Il est présenté tout d'abord la méthode développée en dimension un qui peut être appliquée directement au risque de dépendance.

Principe du lissage

Elle fait appel à deux critères : une contrainte de fidélité et une contrainte de régularité. Le premier consiste à prendre en compte le fait que les taux estimés doivent être proches des taux bruts. Le second inclut la contrainte selon laquelle la courbe des taux estimés doit être aussi régulière que possible. Le but est de chercher les taux qui minimisent ces deux critères. Nous définissions pour cela un opérateur de différentiation Δ pour toute fonction u

$$\forall x, \Delta u(x) = u(x+1) - u(x)$$

On peut montrer par récurrence la proposition suivante :

$\Delta^n u(x) = \sum_{i=1}^{N} \binom{n}{i} (-1)^{n-i} u(x+i)$ eq 1.1

Nous définissons ensuite mathématiquement les critères de fidélité et de régularité utilisés par la méthode de Whittaker-Henderson :

Le critère de fidélité F est défini par la relation suivante :

$$F = \sum_{i=1}^{p} w_i (q_i - \hat{q}_i)^2$$

Le critère de régularité R est vu comme :

$$R = \sum_{i=1}^{p-z} (\Delta^n q_i)^2$$

Où z est un le paramètre contrôlant le lissage des taux. Plus z est élevé, et plus la courbe lissée va être grossière. Le lissage de Whittaker-Hendersen consiste tout simplement à résoudre le programme de minimisation suivant :

$\underbrace{min}_{q_x} (M = F + h \times R)$ (Equ : 1.2)

h est un paramètre inconnu, que l'on cherche à estimer. Il peut se comprendre comme un critère de pénalisation reliant linéairement le critère de fidélité et de régularité. Résoudre 1.1 revient à satisfaire la condition du premier ordre (CPO) suivante :

$$\forall\, 1 \le i \le p, \frac{\partial M}{\partial q_i} = 0$$

Notons $Q = (q_i)_{1\le i\le p}$ le vecteur des taux d'entrée en incapacité, $\hat{Q} = (\hat{q}_i)_{1\le i\le p}$ le vecteur des taux bruts, $W = (diag(w_i))_{1\le i\le p}$ la matrice des poids affectés à chaque individu et $\Delta^z Q = (\Delta^z q_i)_{1\le i\le p}$la matrice de l'opérateur de différentiation. On obtient alors une expression matricielle de F et R :

$$F = \left(Q - \hat{Q}\right)^T W \left(Q - \hat{Q}\right)$$
$$R = (\Delta^z Q)^T (\Delta^z Q)$$

En introduisant la matrice $K_z \in \mathcal{M}_{p-z,p}(\mathbb{R})$ qui regroupe les coefficients de l'équation eq 1.1 on obtient la relation suivante :

$$\Delta^z Q = K_z Q \text{ (Equ 5.8)}$$

Le critère M s'écrit donc comme :

$$M = \left(Q - \hat{Q}\right)^T W \left(Q - \hat{Q}\right) + h(\Delta^z Q)^T (\Delta^z Q)$$
$$M = \hat{Q}^T W Q - 2\hat{Q}^T W \hat{Q} + \hat{Q}^T W \hat{Q} + h \times \hat{Q}^T {K_z}^T K_z Q$$

La condition du premier ordre (CPO) 5.8 se développe de la façon suivante :

$$\frac{\partial M}{\partial Q} = 2WQ - 2W\hat{Q}^T + 2 \times h \times {K_z}^T K_z$$
$$\frac{\partial M}{\partial Q} = 0$$

L'expression précédente permet de donner la solution du programme de minimisation Equ : 1.2, lorsque la matrice $W + h \times {K_z}^T K_z Q$ est inversible :

$$Q^l = (W + h \times {K_z}^T K_z)^{-1} W\hat{Q}$$

Remarque : La méthode de Whittaker-Henderson peut être vue comme un lissage bayésien, dans la mesure ou le critère de régularité R définit une loi a priori pour le vecteur Q.

- **Méthode Loess (LOcally Weighed Scatterplot Smoothing)**

Cette méthode, proposée par Cleveland [13] et [16], fait partie de la famille des régressions polynômiales locales. Elle consiste à approximer localement f(.) par une droite. Pour estimer f en un point x, les observations reçoivent des poids afin que celles proches du point d'intérêt jouent un rôle prépondérant dans l'estimation. La méthode Loess se décompose précisément comme suit :

- (i) Les ϑ plus proches voisins de x (le voisinage s'apprécie en termes de proximité entre variables explicatives), où observations qui seront contenues dans la fenêtre, sont identifiés et leur distance au point x est calculée. Ces ϑ plus proches voisins correspondent en général à un pourcentage v des observations que le praticien souhaite inclure dans chaque fenêtre.
- (ii)Le paramètre de lissage h(x), qui définit la fenêtre $V(x) = [x - h(x), x + h(x)]$, est déterminé de façon à ce que la fenêtre contienne ces ϑ observations. On retient alors la distance de x à son voisin le plus éloigné contenu dans la fenêtre. On la définit par
  $h(x) = \underbrace{max}_{i\in \mathrm{V(x)}} |x - X_i|$
- (iii) Les poids $w_i(x)$ attribués aux éléments X_i du voisinage ou de la fenêtre V(x) s'obtiennent par la formule : $w_i(x) = K\left(\frac{|x-X_i|}{h(x)}\right)$ où $K(u) = (1 - |u|^3)^3 \mathbb{I}[|u| < 1]$
  la fonction K(.) est une fonction de pondération continue, symétrique, décroissante sur [0, 1[, unimodale en 0, nulle hors de [−1, 1]. D'autres fonctions de pondération conviennent bien entendu mais Cleveland suggère d'utiliser cette fonction tricube.
- $\hat{f}(x)$ est calculée en régressant les Y_i, $i \in$ V(x), , sur les X_i correspondants à l'aide d'un ajustement par moindres carrés pondérés en se servant de la droite de régression pour prédire la réponse correspondant à x. La réponse est de la forme (2.16).

Formellement, pour une approximation linéaire locale de la forme $\beta_0(x) + \beta_1(x)x$ dans la fenêtre (x–h(x), x+ h(x)), les estimateurs $\widehat{\beta_0}(x)$et $\widehat{\beta_1}(x)$ des paramètres de la régression sont déterminés en minimisant

$$\sum_{k=1}^{N} w_k(x)\left[Y_k - \widehat{\beta_0}(x) - \widehat{\beta_1}(x)X_k\right] = \sum_{i\in \mathrm{V(x)}} w_k(x)\left[Y_k - \widehat{\beta_0}(x) - \widehat{\beta_1}(x)X_k\right]$$

Qui donne finalement :

$$\hat{f}(x) = \widehat{\beta_0}(x) + \widehat{\beta_1}(x)x = \frac{\sum_{k=1}^{N} w_k(x)Y_k}{\sum_{k=1}^{N} w_k(x)} + (x - \bar{x}_w)\ \frac{\sum_{k=1}^{N} w_k(x)(X_k - \bar{x}_w)Y_k}{\sum_{k=1}^{N} w_k(x)\ (X_k - \bar{x}_w)}$$

où $\bar{x}_w$ est une moyenne pondérée (par la fonction de pondération tricube) des observations dans la fenêtre de x :

$$\bar{x}_w = \frac{\sum_{k=1}^{N} w_k(x)X_k}{\sum_{k=1}^{N} w_k(x)}$$

En prenant $x = X_i$, on retrouve une expression de la forme (2.16) :

$$\widehat{Y_i} = \hat{f}(\ X_i) = \sum_{k=1}^{N} q_k(X_i)Y_k$$

où les poids $q_k(X_i)$ ne dépendent que des régresseurs.

Annexe 2 : Modèle ODP

Soient μ, φ deux réels strictement positifs. Une variable aléatoire X suit la loi de Poisson surdispersée de paramètre (μ, ϕ) si et seulement si X/ϕ suit une loi de Poisson de paramètre $m\ /\ \phi$. Dans ce cas, nous notons $X \sim ODP(m, \phi)$

Avec le paramètre ϕ en plus, une loi de Poisson surdispersée généralise une loi de Poisson habituelle. Elle permet ainsi une relation plus flexible entre la variance et l'espérance de la variable. Concrètement, $Var[X] = \phi E[X]$. Par ailleurs, la famille de loi de Poisson surdispersée possède une propriété intéressante : elle est invariante par l'additivité montrée dans le théorème connu suivant :

Soient X1, X2 deux variables aléatoires telles que $X_1 \sim ODP(m_1, \phi)$ et $X_2 \sim ODP(m_2, \phi)$. Alors, nous avons $X_1 + X_2 \sim ODP(m_1 + m_2, \phi)$

Ces propriétés un rôle important dans le modèle linéaire généalisé basé sur l'hypothèse suivante :

Les observations sont les règlements cumulés $C_{i,j}$.

- Hypothèse sur la loi des observations : $C_{i,j} \sim ODP(m_{i,j}, \phi)$
- Une fonction de lien logarithme entre l'espérance des observations et les variables explicatives
- Un choix de variables explicatives : $ln(m_{i,j}) = c + \alpha_i + \beta_j$ où les coefficients sont estimés par maximum de vraisemblance

Le modèle ODP est un modèle dit non-récursif : les montants futurs sont complètement spécifiés par le modèle.

- Nous travaillons ici avec un modèle additif : $ln(m_{i,j}) = c + \alpha_i + \beta_j$ avec $\alpha_1 = \beta_1 = 0$
- Dans ce cadre, l'estimation des paramètres se fait par maximum de vraisemblance avec des contraintes dites de « nullité aux coins ».

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{j=1}^{n-i+1} \phi^{-1} \times \frac{1}{2} \times \left(C_{i,j} \times \left(c + \alpha_i + \beta_j\right) - e^{c + \alpha_i + \beta_j}\right) + constante$$

L'estimation de l'espérance des montants cumulés $\widehat{m}_{i,j}$ est alors possible grâce à la fonction de lien. Nous noterons cependant que l'estimation du paramètre d'échelle ϕ est faite de façon séparée avec la formule suivante :

$$\hat{\phi} = \sum_{i,j} \left(\frac{C_{i,j} - \widehat{m}_{i,j}}{\sqrt{\widehat{m}_{i,j}}}\right)^2 \times \frac{1}{N - p}$$

- Avec N le nombre d'observations et $p = 2n - 1$ le nombre de paramètres estimés.

L'estimation du coefficient par année de développement est aussi souvent mentionnée :

$$\hat{\phi} = \sum_{i,j} \left(\frac{C_{i,j} - \widehat{m}_{i,j}}{\sqrt{\widehat{m}_{i,j}}}\right)^2 \times \frac{1}{n_j(N - p)} \ avec\ n_j = n - j$$

Annexe 3 : Le taux à terme (forward rate)

Avant de présenter la notion du taux forward, il est nécessaire d'introduire quelques notion notions liées aux outils financiers :

Taux

Définition 1 (compte du marché monétaire). Un compte de marché monétaire est un investissement sans risque où les gains sont cumulés de manière continue au taux sans risque du marché. On note sa valeur B(t) avec B(0) = 1 et on admet qu'elle suit l'équation différentielle suivante : $dB(t) = r_t B(t)dt$

La résolution de cette équation différentielle donne :

$B(t) = exp\left(\int_0^t r_s\, ds\right)$ (Equ 3.1)

où r_s est le taux spot instantané auquel le compte bancaire croît continûment, appelé aussi taux court.

Si nous considérons que le taux court suit une dynamique stochastique, nous pouvons appliquer la formule d'Itô à $ln(B(t))$ et nous retrouvons la même formule (Equ 3.1).

La question suivante peut être posée : quelle est la valeur à l'instant t d'une unité monétaire versée à l'instant T (t < T) :

Pour y répondre nous introduisons la notion de déflateur.

Définition 2 (Facteur d'actualisation stochastique) Un déflateur ou facteur d'actualisation stochastique entre t et T, est le montant en date t qui est équivalent à une unité monétaire payable à la date T. On le note D(t, T) avec : $D(t,T) = \frac{B(t)}{B(T)} = exp\left(-\int_t^T r_s\, ds\right)$

Définition 3 (Obligation zéro-coupon (ZC)) Une obligation zéro-coupon de maturité T est un actif qui garantit à son détenteur le paiement d'une unité monétaire à l'instant T sans aucun paiement intermédiaire. On note P(t, T) sa valeur à l'instant t < T avec P(T, T) = 1.

En univers risque neutre, les prix actualisés au taux sans risque forment une martingale sous la probabilité risque neutre Q. Ainsi en supposant que H est le prix d'un actif qui suit un processus $(\mathcal{F}_t)$ adapté et intégrable de payoff H_T à maturité T, sa valeur en t est donnée par :

$$H_t = E^Q[(D(t,T)H_T|\mathcal{F}_t)]$$

En remplaçant H par P avec P(T, T) = 1, on a :

$$P(t,T) = E^Q[(D(t,T)|\mathcal{F}_t)] = E^Q\left[\left(exp\left(-\int_t^T r_s\, ds\right)\Big|\mathcal{F}_t\right)\right]$$

Le prix d'une obligation zéro coupon est donc l'espérance sous la mesure risque-neutre du déflateur. Nous pouvons également remarquer que lorsque r est déterministe D(t,T) = P(t, T)

- *Taux comptant (spot)*

Définition 4 (Taux d'intérêt instantané en composition continue)Le taux instantané composé continûment est le taux constant auquel un investissement de P(t, T) à l'instant t croît continûment pour donner une unité de monnaie à T, on le note R(t, T). La définition décrit formellement l'équation suivante :

$$P(t,T)\exp\,(\,R(t,T)(T\,-t))\;=\;1$$

On a ainsi la formule :

$$R(t,T)\;=\;-\frac{lnP(t,T)}{\tau(t,T)}$$

Où : τ (t, T) : La mesure du temps restant jusqu'à maturité entre t et T > t

- *Taux à terme (forward)*

Le taux forward en t peut être perçu comme une généralisation du taux spot. Il établit un lien entre deux maturités T_1 et T_2 avec $T_2 \geq T_1 \geq$ t et non plus entre t et T avec T ≥ t

Définition 2.1.7. Le taux forward Le taux forward F(t, T_1, T_2) est le taux d'intérêt en t d'un emprunt entre T_1 et T_2. Il est possible de le représenter schématiquement à travers le schéma ci-dessous :

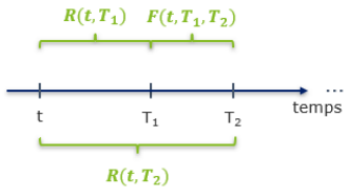


FIGURE 2.4 – Schéma du taux *forward*

Ainsi, en absence d'opportunité d'arbitrage (AOA), il est possible de déduire de ce schéma l'équation définissant un tel taux :

en intérêts composés annuellement :

$$\left(1+R(t,T_2)\right)^{T_2-t}=\left(1+R(t,T_1)\right)^{T_1-t}\left(1+\mathrm{F}(t,T_1,T_2)\right)^{T_2-T_1}$$

$$\Rightarrow \mathrm{F}(t,T_1,T_2)=\left(\frac{\left(1+R(t,T_2)\right)^{T_2-t}}{\left(1+R(t,T_1)\right)^{T_1-t}}\right)^{\frac{1}{T_2-T_1}}-1$$

en intérêts continus :

$$exp\left(-(T_2-t)R(t,T_2)\right)=exp\left(-(T_1-t)R(t,T_1)\right)exp\left(-(T_2-T_1)\mathrm{F}(t,T_1,T_2)\right)$$

$$\Rightarrow \mathrm{F}(t,T_1,T_2)=\frac{1}{T_2-T_1}\left(R(t,T_2)(T_2-t)-R(t,T_1)(T_1-t)\right)$$

$$\Rightarrow \mathrm{F}(t,T_1,T_2)=\frac{1}{T_2-T_1}ln\left(\frac{P(t,T_1)}{P(t,T_2)}\right)$$

La courbe taux spot et taux forward

Period [descriptive]	Period [months]	Spot rate	Forward rate
3 Month	3	1,600%	0,132%
6 Month	6	1,600%	0,132%
1 Year	12	1,650%	0,144%
2 Year	24	1,600%	0,124%
3 Year	36	1,550%	0,116%
4 Year	48	1,600%	0,148%
5 Year	60	1,700%	0,181%
7 Year	84	1,800%	0,177%
8 Year	96	1,850%	0,185%
9 Year	108	1,900%	0,193%
10 Year	120	1,900%	0,157%
15 Year	180	2,150%	0,237%
20 Year	240	2,250%	0,218%
25 Year	300	2,350%	0,234%
30 Year	360	2,300%	0,166%