Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Thomas Vilain

Titre : Segmentation en assurance emprunteur et effets de la sélection médicale

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

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

Membres présents du jury de l'Institut des Actuaires

Entreprise :
Nom : Hannover Re

hannover re
Hannover Rück SE
Succursale Française
33 avenue de Wagram, 75017 Paris
SIRET 451 428 684 APE 6520 Z

Signature :

Membres présents du jury de l'ISFA

Directeur de mémoire en entreprise :
Nom : Mathieu Darnis
Signature : Mathieu Darnis
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

Mathieu Darnis

Signature du candidat

Vilain

Résumé

Le secteur de l'assurance emprunteur individuel a connu de profondes mutations depuis l'introduction de la loi Lagarde il y a un peu plus de 10 ans. Ce marché est devenu très concurrentiel et les acteurs du milieu proposent des tarifs toujours plus compétitifs et plus personnalisés.

Hannover Re accompagne ses partenaires depuis longtemps sur ce risque. Cela a permis de constituer une masse assez importante de données en lien avec l'assurance des emprunteurs. Cette étude propose de rassembler les données à disposition et de produire une analyse d'expérience sur la mortalité des emprunteurs afin de repenser ou revalider la grille tarifaire d'Hannover Re et de conseiller au mieux ses partenaires face aux enjeux de segmentation du tarif.

Nous avons utilisé le modèle de Cox afin de déterminer les variables pouvant servir à une segmentation. Ce modèle permet également de quantifier l'impact de chacune de ces variables explicatives. Nous en avons déduit des taux de mortalité lissés pour chaque sous-population et ces tables ont été validées par différentes méthodes actuarielles.

Nous avons pu constater l'impact du genre, de la consommation de tabac et de la catégorie socio-professionnelle sur le risque de décès. L'impact de la sélection médicale sur la mortalité observée les premières années a également pu être estimé. Cette nouvelle approche tarifaire a été utilisée dans la dernière partie du mémoire pour tarifer le rachat d'un portefeuille emprunteur existant via un traité de réassurance en *quota share*.

Mots clés : Mortalité, réassurance, assurance des emprunteurs individuels, modèle de Cox

Abstract

The individual mortgage insurance sector has experienced some huge changes since Lagarde's law was passed a little more than 10 years ago. This market has become very competitive and the stakeholders offer prices that are always more attractive as well as more personalised.

The Hannover Re has been helping its partners for a long time on this risk. This has allowed the company to collect a substantial amount of data linked to mortgage insurance. This study proposes that we gather the data available to us and produce an experience study on the death rate of borrowers in order to readjust Hannover Re's price framework and to better advise its partners who face issues of pricing segmentation.

We used the Cox model in order to identify the variables needed for a segmentation. This model has also allowed us to quantify the impact of each one of the selected covariates. We have then estimated the smoothed death rates of every subpopulation and those fits have been validated through different actuarial methods.

We have been able to record the impact of gender, smoking and the socio-economic category on the risk of death. The impact of the medical selection on the mortality observed in the first few years of contract has been estimated. In the last part of this thesis, we apply this new pricing approach in the context of a mortgage insurance portfolio redemption via a quota share reinsurance treaty.

Key words: Mortality, reinsurance, individual mortgage insurance, Cox model

Remerciements

Je tiens à remercier tout particulièrement mon maitre de stage, Mathieu Darnis pour son accueil, sa disponibilité et pour le travail que nous avons accompli. Son suivi et son accompagnement ont été déterminants dans la réalisation de ce mémoire. Je lui suis très reconnaissant pour la confiance qu'il m'a accordée et pour les misions proposées.

J'adresse mes sincères remerciements à Stéphane Loisel, mon tuteur académique, pour son accompagnement pendant mon alternance et à Frédéric Planchet pour ses relectures et ses conseils.

Je remercie également Sophie Beng pour sa relecture et ses conseils.

Je tiens à remercier vivement Alain Turco pour son accueil au sein de l'équipe et pour le partage de son expérience du terrain.

Introduction

Depuis 2010 et la loi Lagarde, le législateur a entamé un processus de libéralisation du marché de l'assurance emprunteur qui s'est poursuivi par la loi Hamon (2014) et l'amendement Bourquin (2018). Ce marché était jusqu'alors tenu par les bancassureurs qui imposaient à leurs clients leur propre produit d'assurance. L'ouverture du marché aux assureurs traditionnels a accru la concurrence et a permis aux emprunteurs de bénéficier d'une chute des prix[1]. Cette concurrence ainsi que l'émergence de comparateurs en ligne impose aux assureurs d'afficher des taux toujours plus attractifs pour capter de nouveaux clients. La tarification d'un contrat d'assurance emprunteur est donc devenue une tâche qui requiert beaucoup de précision.

En effet, les assureurs sont tenus d'avoir une certaine stabilité financière afin de pouvoir garantir la couverture promise aux assurés en cas de sinistre. Dans cette optique, on comprend bien que les assureurs doivent avoir une idée la plus fine possible du risque afin de pouvoir faire baisser les prix sans mettre en péril leur solvabilité et leur rentabilité. Cette connaissance du risque doit aussi permettre d'identifier avec précision le risque individuel de chaque assuré. Comme les assureurs cherchent à cibler certaines populations pour pouvoir leur proposer les meilleurs taux possibles et ainsi augmenter leurs parts de marché, la segmentation est un enjeu important sur ce type de contrat.

La connaissance du risque emprunteur intéresse aussi les réassureurs comme Hannover Re. Ceux-ci accompagnent de nombreux partenaires sur ce marché car il s'agit d'un risque engageant le preneur de risque sur une longue période (un prêt immobilier peut durer plus de 30 ans) et le risque reste méconnu pour certains acteurs. La cédante bénéficie alors de l'expertise du réassureur et de sa surface financière. Ces partenariats prennent souvent la forme de traités en *quota share*. Le réassureur joue alors un rôle important dans la conception du produit et sa tarification. C'est pourquoi il est important pour Hannover Re d'avoir une connaissance fiable et à jour du risque emprunteur.

En assurance emprunteur, l'assuré se protège souvent contre le décès et l'arrêt de travail. Les travaux présentés dans ce mémoire ont pour objet l'analyse de la mortalité. En effet, l'arrêt de travail n'a pas pu être étudié avec les données reçues mais l'étude de la mortalité seule revêt déjà un enjeu important dans la tarification des contrats emprunteurs. L'objectif de ce mémoire est de déterminer les caractéristiques des assurés qui ont un impact sur leur risque décès et de quantifier cet impact afin de pouvoir proposer une tarification segmentée. Ces résultats permettront de challenger l'approche tarifaire actuelle d'Hannover Re et de proposer une nouvelle segmentation. Dans la dernière partie de ce mémoire, les nouveaux taux de mortalité seront utilisés pour la tarification d'un portefeuille cédé en *quota share* à Hannover Re.

L'approche retenue pour mener cette étude est le modèle de Cox. Ce modèle de durée semi-paramétrique permet d'intégrer des variables explicatives et semble donc être le plus adapté à notre problématique.

[1] https://reassurez-moi.fr/guide/assurance-pret-immobilier-loi-bourquin-banques

Ce mémoire s'articule en quatre parties. Nous commencerons par une présentation du contexte et des enjeux dans la partie « I. Cadre de l'étude », puis la partie « II. Présentation technique » nous permettra d'introduire les notions théoriques de modèle de durée que nous appliquerons et donnera également une vision macro des différentes données disponibles. La partie « III. Modélisation » présentera la mise en pratique des différentes approches permettant de répondre à la problématique. On y trouvera notamment l'application du modèle de Cox. Les résultats obtenus sont appliqués à la tarification d'un portefeuille emprunteur dans la partie « IV. Tarification d'un portefeuille emprunteur ».

Par souci de confidentialité, certaines données chiffrées ont été modifiées ou censurées et les partenaires ayant collaboré à cette étude ont été anonymisés.

I Cadre de l'étude

L'objet de cette partie est de familiariser le lecteur avec le contexte de l'étude. Le « risque emprunteur » et les activités de réassurance y sont présentés. On y trouvera également les éléments qui ont motivé ces travaux.

I.1 L'assurance emprunteur

Lorsqu'un particulier contracte un crédit immobilier auprès d'un organisme prêteur, celui-ci exige aujourd'hui quasi systématiquement des garanties en cas de décès de l'emprunteur. C'est l'assurance emprunteur qui fournit cette garantie en offrant aux emprunteurs un remboursement lorsque ceux-ci sont en incapacité de le faire.

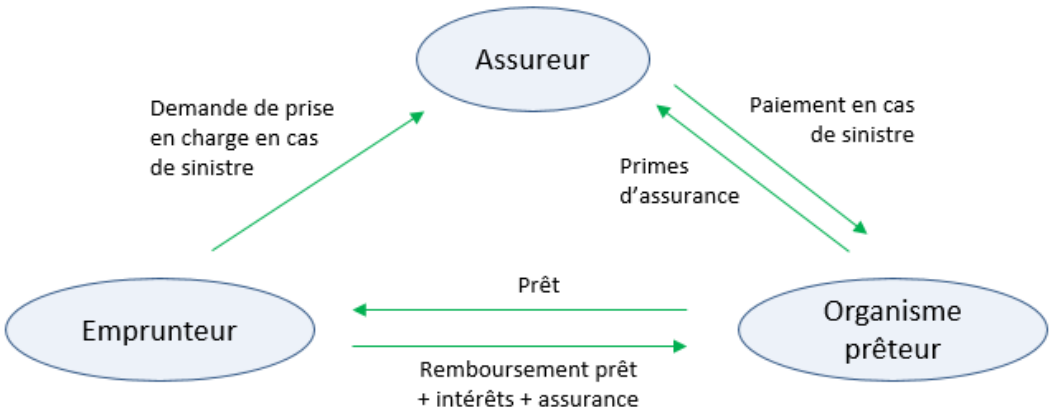
I.1.1 Principe

L'assurance emprunteur est une assurance temporaire qui couvre un emprunteur pendant la durée du prêt. Elle assure à l'établissement de crédit un paiement du capital restant dû en cas de décès de l'assuré. Des couvertures supplémentaires peuvent être exigées ou simplement optionnelles afin d'assurer le remboursement du prêt en cas d'arrêt de travail ou de perte d'emploi.

Le prêteur et l'emprunteur bénéficient tous les deux de cette assurance :

- Elle permet à l'assuré de protéger ses biens en cas de perte de revenus et ses héritiers qui en cas de décès de l'emprunteur se verraient hériter de ses dettes.
- Elle permet au prêteur de se protéger contre une partie du risque de non-remboursement et évite les démarches parfois difficiles pour récupérer l'argent prêté.

Le schéma suivant récapitule la relation tripartie qui s'établie entre l'emprunteur, l'organisme prêteur et l'assureur :



I.1.2 Principales garanties

Les principaux évènements qui peuvent déclencher l'intervention de l'assureur dans le remboursement du prêt sont :

- Le décès
- La perte totale et irréversible de l'autonomie (PTIA)
- L'invalidité
- L'incapacité
- La perte d'emploi

Il existe une grande variété d'options qui couvrent tout ou partie de ces évènements. Le remboursement en cas d'invalidité peut être conditionné à la cause de l'arrêt. Par exemple certaines garanties excluent les maladies dites « non-objectivables », c'est-à-dire les maladies psychiques ne pouvant être mesurées objectivement et les maladies physiques dont l'origine est subjective. La garantie perte d'emploi peut elle aussi prendre des formes variées. En général l'assureur intervient à la suite d'un licenciement économique. Elle s'adresse aux employés en CDI et peut consister à une prise en charge partielle ou totale du remboursement pendant une période d'inactivité ou bien consister en un report des échéances (l'assureur avance les sommes dues au prêteur et l'assuré rembourse lorsqu'il retrouve du travail ou à la fin du contrat d'assurance). La grande variété des garanties s'explique par la concurrence accrue entre les acteurs du secteur qui cherchent à créer des offres intéressantes tout en maitrisant le risque.

Dans la suite nous allons nous concentrer sur les risques décès et PTIA (qui est souvent assimilé à du décès).

I.1.3 La sélection médicale

Si l'âge est déterminant dans la détermination du risque de décès, l'état de santé de l'emprunteur intéresse aussi l'assureur. Celui-ci peut soumettre l'emprunteur à un questionnaire médical avant de lui accorder la couverture et avant même de lui donner un tarif définitif. C'est la pratique courante en assurance emprunteur individuelle.

L'assuré est alors tenu de répondre de bonne foi aux questions. En cas de fausse déclaration, l'assureur peut réclamer la nullité du contrat ou l'abaissement de la prise en charge s'il arrive à prouver que l'assuré avait délibérément menti au questionnaire médical. La réglementation encadre le spectre des questions que peuvent poser les assureurs afin de protéger le client contre des questions trop intrusives. La loi prévoit également un dispositif d'accès à l'assurance pour les personnes ayant vu leur dossier refusé (voir dispositif AERAS). D'après la fédération française de l'assurance, en 2010, 12% des demandes d'assurance de prêt ont été refusées suite à la détection d'un risque aggravé lors de la sélection médicale. La sélection médicale a donc un effet important sur la population assurée.

En pratique la sélection peut prendre différentes formes :

- Une déclaration d'état de santé : l'emprunteur complète et signe un formulaire pré-rempli où il déclare être en bonne santé.

- Un questionnaire médical simplifié : il comporte généralement une dizaine de questions qui permettent à l'assureur de vérifier rapidement les antécédents de l'emprunteur (traitements en cours, maladies en cours, opérations chirurgicales, etc.). Les réponses sont souvent présentées sous la forme d'un « oui/non ». Lorsque l'assuré répond non à une question, il lui est imposé de répondre au questionnaire médical détaillé, décrit plus bas, pour préciser ses antécédents.
- Un questionnaire médical détaillé : version plus détaillée du questionnaire médical simplifié. Les questions sont plus nombreuses et plus précises avec des questions ouvertes. Le médecin peut être sollicité pour remplir ce questionnaire.
- Un examen médical : le souscripteur subit une batterie de tests (sanguins, pulmonaires, cardiaques, …)

Le niveau de détail du questionnaire peut dépendre du produit emprunteur, de l'âge de l'assuré ou du montant de capitaux. Des questions nombreuses pourront permettre à l'assureur d'avoir une idée plus précise du risque et de sélectionner une population peu risquée mais risque aussi de décourager les éventuels souscripteurs (même bons).

Après étude d'un dossier médical, l'assureur peut choisir d'accepter ou de refuser d'assurer la personne. Elle peut aussi décider de majorer sa prime si le risque lui semble plus élevé.

I.1.4 L'anti-sélection

L'anti-sélection, ou sélection adverse, est un phénomène micro-économique qui joue un rôle important dans la conception d'offres d'assurance. Il résulte d'une asymétrie d'information entre l'assuré et l'assureur sur le niveau de risque porté par l'assuré et entraine une forte souscription de « mauvais risques » lorsqu'un produit d'assurance n'est pas assez segmenté. C'est ce qui arrive aux contrats « groupe » en assurance emprunteur.

Les banques négocient des contrats groupe avec les assureurs (ou leur propre filiale assurance) qu'elles proposent ensuite à leurs emprunteurs. Ces contrats permettent une mutualisation entre les adhérents de la banque et sont souvent tarifés en fonction de l'âge du souscripteur uniquement. Les contrats individuels font l'objet d'une tarification beaucoup plus personnalisée. C'est là qu'intervient le phénomène d'anti-sélection. Les assurés non-fumeurs portent moins de risque que les assurés fumeurs. Dans un contrat groupe, ils partagent à parts égales le risque avec les assurés fumeurs. Ils sont donc incités à souscrire un contrat individuel où la prime correspondra à leur faible niveau de risque. Si on fait l'hypothèse que les assurés peuvent résilier leurs contrats avant l'échéance pour souscrire un contrat plus intéressant, on comprend que le contrat groupe va se vider de ses bons assurés et attirer les personnes à risques car celles-ci paient des primes très importantes dans les contrats individuels. Le produit groupe se retrouvera donc rapidement déficitaire et devra réajuster son offre tarifaire pour les nouveaux entrants (incitant encore moins les bons risques à rejoindre le groupe).

Les offres individuelles ne sont pas à l'abri de l'anti-sélection. Une segmentation mal ajustée peut entrainer les mêmes effets que pour les contrats groupe. C'est pourquoi les questionnaires médicaux doivent être assez précis et la connaissance du risque assez fine.

I.1.5 Un marché récemment ouvert aux assureurs

Les banques exigent depuis longtemps une garantie en cas de décès ou d'arrêt de travail pour accorder leurs prêts. Pourtant ce marché est resté très longtemps fermé à la concurrence. Il a fallu attendre 2010 et la mise en place de la loi Lagarde pour que le législateur ouvre ce marché de 6 milliards d'euros. Avant cette loi les banques imposaient leur propre produit d'assurance aux emprunteurs et en profitaient pour réaliser des marges importantes sur ces produits. En effet, les banques se faisaient plutôt la guerre sur les taux d'intérêts pour attirer des clients et l'assurance permettait de faire plus de profit. Nous présenterons dans cette partie les trois lois qui ont permis d'ouvrir à la concurrence le marché de l'assurance emprunteur en France et nous discuterons succinctement de leur efficacité.

Les trois lois qui ont permis aux compagnies d'assurance d'accéder au marché de l'assurance emprunteur sont :

- La loi Lagarde (2010) : elle garantit à tout nouvel emprunteur le droit de choisir librement son assurance emprunteur. L'emprunteur peut présenter à son banquier, au moment de la signature, une assurance qu'il aura choisie lui-même. C'est ce que l'on appelle la délégation d'assurance. Le banquier se doit d'accepter dès lors que l'alternative propose des garanties au moins équivalentes à l'assurance proposée par la banque.
- La loi Hamon (2014) : elle permet à tout emprunteur de résilier son assurance dans les douze mois suivant la souscription à l'emprunt. Là encore, l'emprunteur doit présenter une autre offre d'assurance au moins équivalente. Ce dispositif vient pallier à une limite de la loi Lagarde. En effet, les banquiers pouvaient jouer sur le taux d'intérêt si leur client leur proposait une autre solution d'assurance que la leur avant la signature du prêt.
- L'amendement Bourquin (2018) : Il permet aux emprunteurs de changer d'assurance chaque année à la date d'anniversaire du contrat. Il ouvre donc au marché tous les contrats déjà souscrits en plus des nouveaux et de ceux qui sont dans leur première année. Là encore il faut pouvoir proposer des garanties équivalentes.

Cette ouverture devait permettre aux assureurs de pénétrer le marché et aux clients de profiter d'une baisse de prix dans un contexte de taux bas, où le prix de l'assurance représente un coût de plus en plus proche de celui des intérêts. En 2020, un couple de 34 ans empruntant 170 000 € sur 20 ans à un taux nominal de 1,30% paiera 23 148 € d'intérêt. Une couverture décès et arrêt de travail avec une franchise de 90 jours leur coûtera 19 040 € en prenant le tarif moyen des banques pour ce type de garantie[2]. On comprend donc bien que l'assurance emprunteur est devenue une variable importante dans le budget emprunt des ménages.

Une baisse de tarif a bel et bien été observée depuis 2010. Les assureurs proposent aujourd'hui des tarifs entre 30% et 40% moins chers[3], notamment grâce à des offres démutualisées et l'utilisation de tarifs en pourcentage du capital restant dû.

[2] Securimut (2020) Libre choix de l'assurance emprunteur immobilier : 3 lois pour quelle réalité ? Bilan des lois Lagarde, Hamon et Bourquin
[3] Actuélior (2020) Matinale emprunteur, La loi Lagarde, 10 ans après : analyses et perspectives

Pourtant la promesse d'ouverture du marché est à nuancer. S'il y a aujourd'hui beaucoup plus d'acteurs sur le marché (assureurs, mutuelles, courtiers, comparateurs en lignes etc.), 85% des emprunteurs sont encore assurés par leur banque[4]. Les banques ont réagi, se sont adaptées et ont fait des efforts pour conserver l'assurance des prêts.

Les nouveaux acteurs sont pourtant très agressifs tant d'un point de vu tarifaire que marketing. Certains proposent même à leurs futurs clients de réaliser les démarches auprès de la banque. Le client donne un mandat de mobilité à son nouvel assureur qui va gérer l'ensemble des démarches administratives. Cette pratique s'est développée car les assureurs ont compris que les banques jouaient sur la lourdeur administrative pour conserver leurs contrats. Mais malgré tout, les banques arrivent à rendre la tâche difficile au mandataire de mobilité.

Le gestionnaire/distributeur Securimut propose ce service à ses clients et a mené une étude[5] dans laquelle sont constatés sept points sur lesquels les banques jouent régulièrement pour rendre les démarches de mobilité quasiment impossibles aux particuliers isolés :

- Le faible respect des délais : La banque dispose légalement de 10 jours pour répondre à une demande de délégation. Seule une réponse sur deux est envoyée dans ce délai. Un quart des réponses arrive après un délai d'au moins un mois. Securimut note qu'un tiers des demandes ont dû être relancées après 20 jours.
- Le non-respect du mandat : Une partie des retards a été constaté parce que la banque ne reconnaissait pas le mandat de mobilité et répondait directement à l'emprunteur.
- Des objections erronées : Securimut considère que lorsqu'elles préfèrent dialoguer directement avec l'emprunteur, les banques utilisent souvent des techniques d'intimidation en avançant à tort que les garanties du nouvel assureur ne sont pas équivalentes.
- Des réponses partielles : Seul 40 % des demandes obtiennent une réponse unique et complète de la banque. Les autres reçoivent plusieurs réponses avec des informations arrivant au compte-goutte. Cela alourdit et ralentit les démarches. En gagnant du temps, la banque peut décaler d'un an les délégations Bourquin.
- Décalage de la date de substitution : En mettant en place une date de substitution différente de la date de demande, les banques font payer deux assurances à leur client sur des périodes de plusieurs mois.

Il est donc difficile pour un emprunteur de prendre une assurance qui ne soit pas celle de la banque qui accorde le prêt.

I.1.6 Les particularités du risque emprunteur

Le risque emprunteur n'est pas facile à appréhender pour les assureurs :

- C'est un risque long car l'assureur s'engage sur toute la durée du prêt. Les prêts immobiliers peuvent durer jusqu'à 30 ans aujourd'hui ce qui, d'un point de vu assuranciel, est peu fréquent pour ce type de garantie.
- La prime est définie à la souscription et ne peut plus être majorée par la suite. L'actuaire doit donc avoir une certaine confiance dans sa tarification.

[4] Les Echos, 12/12/2019, Assurance emprunteur : l'irrésistible assor des contrats individuels
[5] Securimut (2020) Libre choix de l'assurance emprunteur immobilier : 3 lois pour quelle réalité ? Bilan des lois Lagarde, Hamon et Bourquin

- Comme discuté plus haut, la connaissance du risque doit être suffisamment précise pour segmenter le tarif en fonction des caractéristiques de l'assuré.
- Le marché étant relativement nouveau, les assureurs qui sont arrivés récemment sur le marché n'ont pas toujours un historique suffisant pour faire leurs modèles.

Les preneurs de risques doivent donc avoir à la fois une bonne expertise technique pour tarifer correctement le risque et une surface financière suffisamment ample pour leur permettre de prendre un engagement de longue durée. Il est fréquent que les assureurs cherchent des partenaires pour se lancer sur ce nouveau marché, c'est là qu'intervient la réassurance. En effets les réassureurs ont souvent une expertise reconnue et une masse financière qui leur permet de souscrire des risques importants.

I.2 L'activité de réassurance

Cette partie a pour objectif de présenter la nature des relations que peut avoir Hannover Re avec ses partenaires sur le risque emprunteur.

I.2.1 Définition

La réassurance est communément appelée l'assurance des assurances. Picard et Besson en ont proposé une définition plus précise : « Une opération de réassurance est un contrat sur lequel un réassureur (dit cessionnaire) vis-à-vis d'un assureur professionnel (dit cédant) qui répond seul et intégralement vis-à-vis des assurés, prend en charge moyennant rémunération tout ou partie des sommes dues ou versées aux assurés à titre de sinistre. »[6]

I.2.2 Intérêt de la réassurance

Plusieurs raisons peuvent pousser un assureur à faire appel à la réassurance :

- Améliorer son ratio de solvabilité.
- Augmenter sa capacité de souscription
- Protéger son bilan contre des évènements extrêmes
- Protéger son bilan contre des dérives techniques
- Solliciter l'expertise technique du réassureur

Toutes ces raisons peuvent intervenir en assurance emprunteur.

I.2.3 Types de contrats

Les réassureurs proposent une large gamme de couvertures. Chacune correspond à un besoin bien précis de la cédante. Nous ne présenterons ici que rapidement les traités les plus classiques.

I.2.3.1 La distinction obligatoire/facultatif

On peut diviser la réassurance en deux catégories, la réassurance obligatoire et la réassurance facultative.

- La réassurance obligatoire : Le cessionnaire et la cédante s'accordent sur la définition d'une catégorie de risques et d'un ou plusieurs portefeuilles qui seront couverts. La

[6] Picard M. et Besson A. (1971) Les assurances terrestres en droit français

cédante est alors obligée de céder tous les risques correspondant à cette description et le cessionnaire s'engage à les prendre tous sans faire de sélection.

- La réassurance facultative : Les deux parties s'accordent sur une couverture d'un seul risque. Chaque affaire va faire l'objet d'une acceptation individuelle.

Pour l'assurance des emprunteurs, on trouve fréquemment de la réassurance obligatoire qui couvre des portefeuilles entiers mais aussi de la réassurance facultative pour couvrir le risque porté par un assuré empruntant des montants très importants.

Il est souvent avantageux pour une cédante de faire une assurance facultative, sur-mesure, pour les quelques très gros capitaux assurés et une réassurance obligatoire pour le reste du portefeuille.

I.2.3.2 La distinction proportionnel/non-proportionnel

On fait aussi la distinction entre réassurance proportionnelle et réassurance non-proportionnelle. Dans le premier cas, le réassureur s'engage à prendre une part proportionnelle des sinistres et des primes. Dans le second cas, il n'y a pas de relation de proportionnalité entre le montant des sinistres et la prise en charge du réassureur.

Nous n'allons pas ici présenter en détail le fonctionnement de chaque type de contrat. Le lecteur intéressé pourra se référer au cours ISFA de réassurance[7] ou au mémoire de Gabriel Pophillat[8].

Les deux principaux types de traités proportionnels sont :

- Le Quota-share (ou quote-part) : Assureur et réassureur conviennent d'un taux x de partage du risque. Le réassureur reçoit alors x% des primes et reverse en échange x% des sinistres à l'assureur. Il paie également à l'assureur des frais liés à la gestion des contrats.

  Ce type de garantie est très fréquent en assurance emprunteur lorsqu'assureur et réassureur construisent ensemble un produit.

- Le Surplus : Ce type de contrat fonctionne comme un quota-share qui ne commencerait qu'à partir d'un certain montant (appelé plein de rétention). En dessous de ce seuil il n'y a pas de partage du risque, au-dessus c'est un quota-share.

  L'assureur peut préférer ce type de garantie au quota-share afin de conserver une plus grande part de chiffre d'affaires.

Les deux principaux types de traités non-proportionnels sont :

- L'excédent de sinistre (XS) : L'assureur paie une proportion des primes qu'il reçoit au réassureur en échange desquelles celui-ci paie pour chaque sinistre la partie du montant qui excède un seuil appelé priorité et dans la limite d'un montant appelé portée. La priorité et la portée peuvent aussi dans certains cas s'appliquer à plusieurs risques ayant subis un même évènement. On parle alors d'XS CAT.

[7] Delacroix A. (2019) Réassurance Non-Vie. Cours, Institut de Science Financière et d'Assurances
[8] Pophillat G. (2019) Calcul de la meilleure estimation d'un traité emprunteur individuel français, Diplôme Universitaire d'Actuaire de Strasbourg

Ce type de garantie permet de limiter la part cédée à la réassurance tout en protégeant la cédante contre les évènements les plus importants. Elle peut souhaiter mettre en place cette couverture en assurance emprunteur pour céder au réassureur les invalidités et décès des assurés ayant d'importants capitaux sous risques.

- Le Stop-Loss (SL) : Dans le cas du Stop-Loss, on définit également une priorité et une portée mais celle-ci s'applique au ratio sinistre à prime du portefeuille considéré.

  Un assureur peut vouloir mettre en place ce type de couverture plutôt qu'un excédent de sinistre pour se protéger contre le risque d'une augmentation de la fréquence des sinistres. En effet, dans le cas de l'augmentation de la fréquence, les contrats XS ne sont pas nécessairement déclenchés car individuellement les sinistres ne dépassent pas forcément la priorité mais pourtant la charge totale des sinistres augmente pour l'assureur.

I.2.4 Hannover Re et le marché de l'emprunteur

Hannover Re est l'un des principaux réassureurs mondiaux, le troisième par son chiffre d'affaires. Bien placé également sur le marché français de l'assurance vie, son expertise en assurance des emprunteurs est reconnue.

Depuis l'ouverture progressive du marché de l'emprunteur, Hannover Re a créé de nombreux partenariats avec assureurs et distributeurs afin de proposer des alternatives aux couvertures bancaires. Les partenariats peuvent prendre la forme de couvertures non-proportionnelles avec la protection de portefeuilles emprunteurs contre de gros sinistres ou contre une dérive de la mortalité. Une partie importante du business se fait via des quota-shares. Hannover Re participe alors à la conception du produit, à sa tarification et discute de la distribution avec l'ensemble des partenaires participant au projet. Cette activité est non seulement très intéressante d'un point de vue actuariel mais elle est aussi source de chiffre d'affaires lorsque le produit se vend bien.

Dans le cadre de ces partenariats en quote-part, nous avons pu obtenir les données historiques des contrats en question afin de mener cette analyse sur le risque emprunteur.

I.3 Enjeux de l'étude

Cette étude a pour objet la garantie décès/PTIA des contrats emprunteurs individuels. Plusieurs enjeux l'ont motivée.

I.3.1 Un intérêt pour notre approche de provisionnement

Lorsqu'un assureur décide de couvrir un risque, il perçoit une prime et doit se tenir prêt à indemniser l'assuré en cas de survenance du sinistre (inversion du cycle de production). Pour être capable de faire face à cet engagement le moment venu, l'assureur doit constituer des provisions. Ce principe s'applique également aux réassureurs qui s'engagent envers leurs cédantes. Le niveau des provisions est suivi par le régulateur et obéit à certaines règles que nous ne détaillerons pas ici. Pour déterminer le montant des provisions en assurance vie, l'assureur peut utiliser soit une table réglementaire (en France les tables TH et TF 02 pour le risque de mortalité) soit une table d'expérience construite par l'assureur et certifiée par un actuaire agréé.

L'avantage d'une table d'expérience est qu'elle est plus proche du risque et permet donc d'avoir un montant de provisions techniques qui s'écoule bien au fil du temps. Comme la population emprunteur est sélectionnée sur des critères médicaux et que le fait d'avoir un projet immobilier indique une certaine stabilité financière, nous observons une sous-mortalité sur ce type de produits. Si bien que les provisions calculées à partir de ces tables sont souvent trop importantes par rapport à la sinistralité observée ce qui décale le résultat de l'assureur et lui fait immobiliser une partie de son capital inutilement. Ainsi, l'un des enjeux de l'étude menée sur le risque emprunteur est de faire certifier une table d'expérience.

I.3.2 Mesurer l'impact des caractéristiques des assurés

Il est aujourd'hui bien connu que le risque décès dépend de l'âge de l'assuré et de son sexe[9]. On le constate par exemple en observant les tables réglementaires TH et TF 02. Le deuxième enjeu de cette étude est d'identifier les autres caractéristiques qui ont un impact sur ce risque. Après avoir été identifiés ces impacts seront quantifiés et permettront :

- De challenger notre approche actuelle de tarification qui prend déjà en compte certaines variables explicatives.
- De concevoir de nouveaux produits, segmentés à partir des résultats obtenus.

I.3.3 Mesurer l'impact de la sélection médicale

Lors des premières années de contrat, il y a généralement peu de décès car les assurés viennent de passer la sélection médicale. Ils n'ont donc pas de pathologie grave ni de risque particulièrement aggravé ce qui fait que l'on n'observe presque que des décès accidentels les premières années. La mortalité est donc minorée chez les personnes venant de passer la sélection médicale. Une fois celle-ci passée, l'état de santé des assurés peut se dégrader. Après quelques années, des maladies graves peuvent se développer et on observe un regain de mortalité dans la population. L'un des objectifs de cette étude est aussi la mesure de l'impact (durée et effet) de la sélection médicale sur le risque de mortalité.

I.3.4 Les données

Pour atteindre les objectifs présentés ci-dessus, nous disposons des données venant de deux partenariats conclus avec des assureurs. Pour des raisons de confidentialité, ces assureurs seront appelés « assureur A » et « assureur B » dans la suite de ce rapport (ou « portefeuille A » et « portefeuille B » lorsque l'on parle de leurs données respectives). Une analyse plus détaillée des caractéristiques de ces portefeuilles est présentée en partie II.2.

Ces données ont l'avantage de correspondre exactement à la population qui souscrit des prêts. Elles sont assez récentes car toutes extraites courant 2020 et comme nous le verrons dans la partie II.2 les 749 décès observés permettront de qualifier la quantité de données de satisfaisante.

[9] L'arrêt « Test-Achats » de la Cour de Justice de l'Union Européenne interdit depuis 2011 l'utilisation du genre pour segmenter le tarif des produits d'assurance. Nous en discuterons plus en détail dans la partie III.2.3.

Nous nous intéresserons dans ce rapport aux points I.3.2 et I.3.3 de cette partie.

II Présentation technique

II.1 Éléments théoriques

II.1.1 Les modèles de durée

L'analyse de survie est une branche des statistiques qui cherche à modéliser le temps restant avant un évènement. Elle possède de nombreuses applications en fiabilité, biologie, médecine, économie, etc. Elle est aussi très utilisée en assurance vie pour modéliser la durée de vie des assurés ou leur loi de rachat. Nous allons présenter dans cette partie quelques éléments conceptuels et notations utiles à l'introduction des estimateurs d'Hoem et du modèle de Cox que nous utiliserons dans la suite. Le lecteur pourra également trouver une présentation de la technique de lissage retenue dans notre étude (Whittaker-Henderson).

II.1.1.1 Quelques notations

La durée de vie d'un individu peut être modélisée par une variable aléatoire T à valeurs dans $[0; +\infty[$. On s'intéresse souvent à la durée de survie d'un individu dont on sait qu'il est vivant à un certain âge. On considère donc souvent la grandeur $S_u(t) = P(T > u + t \mid T > u)$ qui est la fonction de survie conditionnelle.

On a que $S_u(t) = \frac{P(T>u+t)}{P(T>u)} = \frac{S(u+t)}{S(u)}$, avec $S(t) = P(T > t)$

Les assureurs utilisent beaucoup la notion de probabilité de décès à un âge donné. On introduit donc ${}_t\mathrm{q}_x = P(T < x + t \mid T > x)$ qui est la probabilité de mourir dans les t prochaines années pour un individu d'âge x. On note q_x la probabilité de mourir à l'âge x : $q_x = {}_1\mathrm{q}_x$. On passe facilement de la probabilité de décès à la probabilité de survie par : $q_x = 1 - S_x(1)$.

On introduit également $F(t) = P(T \leq t)$, la fonction de répartition de T à partir de laquelle on obtient la densité $f(t) = \frac{d}{dt}F(t) = \lim_{h\to 0} \frac{P(t\leq T\leq T+h)}{h}$ pour tout $t > 0$.

La densité et la fonction de survie permettent de définir la fonction de hasard :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\ln\left(S(t)\right)$$

Cette grandeur est très utilisée pour caractériser un modèle de durée. Elle peut être interprétée comme un taux de décès instantané.

II.1.1.2 Censure et troncature

Lorsqu'un assureur s'intéresse à la durée de vie de ses assurés, il ne peut les observer qu'entre la date de souscription et la date de fin de contrat. De plus, il y a souvent des contrats qui au moment de l'analyse sont encore en cours. Ainsi l'actuaire ne peut observer T que pendant une courte période appelée période d'observation. Pour extraire un maximum d'informations

des données non-complètes (celles où on n'observe pas de décès durant la période d'observation), on introduit les concepts de censure et de troncature.

La censure :

Le phénomène de censure intervient quand la réalisation de T intervient en dehors de la période d'observation. On parle de censure gauche lorsque l'observation de la censure C indique que $T \leq C$. On parle de censure droite lorsque l'observation de la censure C indique que $T \geq C$. C'est le cas, par exemple, pour les assurés encore en vie à la fin de la date d'observation. Nous savons qu'ils mourront un jour mais nous ne pouvons pas encore observer T.

Il existe trois types de censure :

- La censure de type I consiste à fixer le seuil C à une valeur commune pour toutes nos observations. C'est un cas de figure qui intervient beaucoup dans l'industrie : on teste des pièces pendant leurs x premières heures d'utilisation.
- La censure de type II consiste à fixer le seuil C lorsque l'évènement survient pour la $k^{ème}$ fois dans la population considérée.
- La censure de type III généralise la censure de type I au cas où C est une variable aléatoire. En pratique, on observe l'individu jusqu'à ce que cela ne soit plus possible. Cet âge maximal d'observation est une variable aléatoire.

En assurance on rencontre principalement des censures droites de type III.

Troncature :

Il s'agit cette fois du cas où la date de début du phénomène n'appartient pas à la période d'observation. On dit que les données sont tronquées à gauche quand T n'est pas observable lorsqu'elle est inférieure à un seuil C. On dit que les données sont tronquées à droite quand T n'est pas observable lorsqu'elle dépasse un seuil C.

La troncature est très différente de la censure car lorsque des données sont tronquées, on n'a aucune information sur celles-ci. Dans le cas de la censure, on sait qu'il existe une information sur T mais on ne connait pas sa valeur exacte.

En assurance, on rencontre principalement des censures gauches.

Nous allons maintenant présenter les principaux modèles utilisés lors de notre analyse. L'estimateur d'Hoem qui nous a servi à déterminer des taux de mortalité bruts pour certaines sous populations, le modèle de Cox et la méthode de lissage de Whittaker Henderson.

II.1.2 L'estimateur des moments de Hoem

L'estimateur des moments de Hoem permet d'estimer la valeur des différents q_x. C'est un estimateur paramétrique, sans biais et qui intègre les phénomènes de censure et de troncature. Il repose sur un principe assez intuitif qui veut que l'on puisse estimer la probabilité de décès à

un âge x en faisant le nombre de décès à l'âge x sur le nombre d'années d'exposition à l'âge x en comptant une année complète d'exposition pour les observations non-censurées.

La construction de l'estimateur des moments de Hoem est inspirée de Pophillat (2019)

Introduisons les notations suivantes :

- $\alpha_i \in [0 ; 1]$: fraction d'année entre le début de l'observation de l'assuré i, sur la classe d'âge [x ; x+1[et la date à laquelle l'assuré i a eu x ans ;
- $\beta_i \in [0 ; 1]$: fraction d'année entre la fin de l'observation de l'assuré i, sur la classe d'âge [x ; x+1] et la date à laquelle l'assuré i a eu x ans ;
- ${}_tq_x \in [0 ; 1]$: probabilité de décéder entre les âges x et x + t ;
- $q_x \in [0 ; 1]$: probabilité de décéder à l'âge x ;
- $n_x \in \mathbb{N}$: nombre d'assurés en vie à l'âge x ;
- X_i : variable aléatoire égale à 1 si l'assuré i décède dans la classe d'âge $[x+\alpha_i ; x+\beta_i[$, sinon à 0 ;
- $E_x = \sum_{i=1}^{n_x}(\beta_i - \alpha_i)$: exposition de la classe d'âge [x ; x+1[;
- $D_x = \sum_{i=1}^{n_x} X_i$: variable aléatoire représentant le nombre de décès observés dans la classe d'âge [x ; x+1[;
- d_x : réalisation de la variable aléatoire D_x.

Pour construire l'estimateur des moments de Hoem, nous sommes obligés de faire trois hypothèses :

- (H1) : les X_i sont indépendants ;
- (H2) : la probabilité de décès sur toute classe d'âge [x ; x+1[est linéaire,
$\forall t \in [0 ; 1]$ ${}_tq_x = t \cdot q_x$;
- (H3) : $\forall s \geq t$, ${}_{s-t}q_{x+t}$ est approximée par ${}_sq_x - {}_tq_x$.

La première hypothèse implique que la variable aléatoire D_x suit une loi binomiale de paramètres $(n_x, {}_{\beta_i-\alpha_i}q_{x+\alpha_i})$.
Les hypothèses 2 et 3 permettent d'écrire $\mathbb{E}(X_i)$ comme suit,

$$\mathbb{E}(X_i) = {}_{\beta_i-\alpha_i}q_{x+\alpha_i} = {}_{\beta_i}q_x - {}_{\alpha_i}q_x = (\beta_i - \alpha_i)q_x$$

En notant $Y_i = \frac{X_i}{\beta_i - \alpha_i}$, on obtient :

$$\mathbb{E}(Y_i) = \frac{\mathbb{E}(X_i)}{\beta_i - \alpha_i} \approx q_x$$

En appliquant la loi des grands nombres à Y_i, on trouve finalement l'expression de l'estimateur des moments de Hoem :

$$\hat{q}_x^{Hoem} = \frac{d_x}{E_x}$$

Il existe d'autres estimateurs des taux de mortalité bruts. Le plus utilisé est l'estimateur de Kaplan Meier qui prend lui aussi en compte les censures et troncatures et qui a l'avantage de ne faire aucune hypothèse sur la répartition des décès sur l'intervalle [x ; x+1[. En pratique les

estimateurs d'Hoem et de Kaplan Meier[10] donnent des résultats assez proches. Nous avons décidé d'utiliser l'estimateur d'Hoem pour sa simplicité.

II.1.3 Présentation du modèle de Cox

Le modèle de Cox permet d'introduire des variables explicatives à notre modèle de durée. Ce modèle est donc très utile pour analyser l'impact des caractéristiques d'un assuré sur sa mortalité. Il fait partie de la famille des modèles à hasard proportionnel.

II.1.3.1 Les modèles à hasard proportionnel

Les modèles à hasard proportionnel sont des modèles semi-paramétriques dans lesquels on introduit une fonction de survie de base $S_0(t)$ et on fait l'hypothèse que la fonction de survie du phénomène à analyser, $S_\theta(t)$, est reliée à la fonction de survie de base par la relation :

$$S_\theta(t) = S_0(t)^\theta$$

avec $\theta > 0$, un paramètre inconnu.

En dérivant, on a que $f_\theta(t) = \theta S_0(t)^{\theta-1} f(t)$, et en utilisant $h_\theta(t) = \frac{f_\theta(t)}{S_\theta(t)}$, on obtient :

$$h_\theta(t) = \theta . h_0(t)$$

La fonction de hasard du phénomène observé est donc proportionnelle à la fonction de hasard de base (ou fonction de hasard de référence ou « *baseline* »), d'où le nom de modèles à hasard proportionnel.

II.1.3.2 Le modèle de Cox

Le modèle de Cox appartient à la famille des modèles à hasard proportionnel. Dans ce modèle $\theta = e^{z'\beta}$, où $z = (z_1, z_2, \dots, z_p)$ est un vecteur de p variables explicatives et $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ est un vecteur de p coefficients. Ces coefficients sont obtenus à l'aide du modèle de régression linéaire suivant qui découle simplement de la définition du modèle de Cox :

$$\ln\left(h(t|Z = z) = \ln\left(h(t)\right) + \sum_{j=1}^{p} z_j \beta_j\right.$$

Il existe deux méthodes pour déterminer les paramètres de ce modèle. La première consiste à considérer la fonction de survie de base $S_0(t)$ connue. On peut, par exemple, ajuster un modèle de Weibull aux données de la population globale pour obtenir $h(t)$ puis estimer le vecteur β. Une autre méthode consiste à considérer $S_0(t)$ inconnue. Cette dernière approche est plus difficile à mettre en œuvre mais a l'avantage d'être plus souple. Lorsque l'on parle du modèle de Cox, on fait plutôt référence à cette méthode semi-paramétrique.

[10] Voir Pophillat G. (2019) pour une présentation de cet estimateur.

II.1.3.3 Estimation du modèle

Considérons que nous disposons de k décès. Ils ont lieu presque sûrement à des instants distincts $T_1 < T_2 < \cdots < T_k$. On regroupe les individus décédés entre les indices 1 et k (l'individu i décède en T_i) et le reste de la population entre k+1 et n le nombre total d'individus observés. Notons $R(T_i)$ l'ensemble des individus sous risque en T_i^-, c'est-à-dire juste avant le i^{ème} décès.

La probabilité qu'un individu i décède en T_i sachant qu'il est vivant en T_i^- est égale à :

$$h_0(T_i)\exp\left(z'_{(i)}\beta\right)\ dt$$

La probabilité que, de l'ensemble des individus vivant en T_i^-, le décès en T_i soit celui du i^{ème} individu s'écrit alors :

$$\frac{h_0(T_i)\exp\left(z'_{(i)}\beta\right)}{\sum_{j\in R(T_i)} h_0(T_i)\exp\left(z'_{(j)}\beta\right)}$$

On en déduit l'écriture de la vraisemblance partielle de Cox :

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\left(z'_{(i)}\beta\right)}{\sum_{j\in R(T_i)}\exp\left(z'_{(j)}\beta\right)}$$

L'intérêt de passer par la vraisemblance partielle du modèle est d'obtenir une vraisemblance $L(\beta)$ ne dépendant pas de la fonction de référence h_0. Cela permet de d'obtenir des estimations de β (en maximisant la vraisemblance partielle) sans avoir à spécifier h_0.

II.1.3.4 Obtention du maximum de vraisemblance

Le calcul du maximum de vraisemblance s'effectue par méthode numérique. La maximisation de ce type de modèle ne pose pas beaucoup de problèmes aux logiciels actuels. La difficulté réside dans la gestion des ex aequo.

En effet, nous avons fait l'hypothèse que les dates de décès étaient toutes distinctes. Cette hypothèse est tout à fait justifiée en temps continu mais en pratique nous utilisons des données discrétisées où des ex aequo peuvent apparaitre. Dans notre jeu de données, la précision est au jour près ce qui signifie que les ex aequo sont peu fréquents mais peuvent arriver.

Nous allons essayer d'expliciter le problème des ex aequo au travers d'un exemple. Prenons un jeu de données avec quatre individus. Les deux premiers décèdent le même jour. Pour simplifier les notations, notons $r_i = \exp\left(z'_{(i)}\beta\right)$ le « risk score » de l'individu i. On peut alors réécrire la vraisemblance partielle de la façon suivante :

$$L(\beta) = \prod_{i=1}^{k} \frac{r_i}{\sum_{j\in R(T_i)} r_j}$$

Seulement, dans notre exemple, on ne sait pas très bien s'il faut écrire :

$$L_{exemple} = \frac{r_1}{r_1 + r_2 + r_3 + r_4} \cdot \frac{r_2}{r_2 + r_3 + r_4}$$

ou bien :

$$L_{exemple} = \frac{r_2}{r_1 + r_2 + r_3 + r_4} . \frac{r_1}{r_1 + r_3 + r_4} .$$

Il existe plusieurs méthodes pour gérer les ex aequo. Le package *survival* du logiciel R en propose trois que nous allons présenter succinctement :

Méthode de Breslow :

De nombreux programmes utilisent cette méthode par défaut. Elle consisterait dans notre exemple à approximer la vraisemblance partielle par :

$$L_{exemple} \approx \frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)^2}$$

Cette méthode simplifie grandement le problème mais sous-estime la vraisemblance car le dénominateur est plus grand que dans les deux formules précédentes.

Méthode d'Efron :

C'est la méthode utilisée par R par défaut. Elle est plus précise et plus rapide en temps de calcul que la méthode de Breslow. Elle consisterait dans notre exemple à approximer la vraisemblance partielle par :

$$L_{exemple} \approx \frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)(\frac{r_1}{2} + \frac{r_2}{2} + r_3 + r_4)}$$

Dans cette approximation, on décide de prendre une valeur moyenne entre les deux valeurs correspondant aux évènements "$T_1 < T_2$" et "$T_2 < T_1$".

Méthode exacte :

Cette méthode calcule la vraisemblance exacte partielle. Elle peut conduire à des temps de calcul importants lorsque les ex aequo sont nombreux. Elle consisterait dans notre exemple à approximer la vraisemblance partielle par :

$$L_{exemple} \approx \frac{r_1 r_2}{\sum_{i \neq j} r_i r_j}$$

Nous avons décidé d'utiliser la méthode d'Efron pour sa simplicité. On note que, dans le cas d'un jeu de données sans ex aequo, ces méthodes sont toutes équivalentes.

II.1.3.5 Interprétation des β_i

Le programme de maximisation de la vraisemblance partielle nous donne un vecteur β qu'il convient d'interpréter avec précaution si ces coefficients passent le test de significativité. Trois cas de figure peuvent survenir :

- $\beta_i = 0 \Rightarrow h(t) = e^{\beta_i} h_0(t) = h_0(t)$ correspond au cas où la covariable n'a pas d'impact sur la mortalité.
- $\beta_i \leq 0 \Rightarrow h(t) = e^{\beta_i} h_0(t) \leq h_0(t)$ correspond au cas où la covariable diminue le risque de mortalité en augmentant.
- $\beta_i \geq 0 \Rightarrow h(t) = e^{\beta_i} h_0(t) \geq h_0(t)$ correspond au cas où la covariable augmente le risque de mortalité en augmentant.

II.1.3.6 Évaluation de la fonction de hasard de référence h_0

Une fois l'estimation des paramètres effectuée, nous pouvons estimer la fonction de hasard cumulée de base $H_0(t)$ définie pour tout $t \geq 0$ par : $H_0(t) = \int_0^t h_0(u)du$, à l'aide de l'estimateur de Breslow.

Supposons que nous disposons des données sous la forme suite : $(T_i, D_i, Z_{(i)})_{i=1,\dots,n}$ avec T_i la durée observée pour l'individu i, D_i l'indicatrice égale à 1 lorsque l'on observe un décès pour cet individu et $Z_{(i)} = (Z_{(i)}^1, \dots, Z_{(i)}^p)'$ les covariables de cette personne. L'estimateur de Breslow s'écrit alors :

$$\widehat{H}_0(t) = \sum_{i=1}^{n} \frac{I(T_i < T, D_i = 1)}{\sum_{j=1}^{n} I(T_j \geq T_i).\exp\left(Z'_{(j)}\beta\right)}$$

Si nous avons décidé dans un premier temps de ne pas spécifier la fonction de hasard de base pour évaluer au mieux $\hat{\beta}$, il est pratique une fois les paramètres estimés d'ajuster une fonction de hasard de référence à notre modèle. En effet, en tarification, nous nous intéressons non seulement à l'impact relatif des variables explicatives mais aussi au niveau observé des taux de mortalité. Sans h_0, les $(\hat{\beta}_i)_{i=1,\dots,p}$ ne permettent pas de donner le niveau de risque à affecter aux assurés. À partir de la fonction de hasard de référence, on obtient facilement les q_x :

On calcule d'abord la fonction de survie de base estimée : $\hat{S}_0(t) = \exp\left(-\widehat{H}_0(t)\right)$.
On en déduit ensuite les taux de mortalité de base :

$$\hat{q}_x^0 = 1 - \frac{\hat{S}_0(x+1)}{\hat{S}_0(x)}$$

Puis les taux de mortalité pour un individu de covariables $z = (z_1, \dots, z_p)$:

$$\hat{q}_x^{COX} = 1 - \left(\frac{\hat{S}_0(x+1)}{\hat{S}_0(x)}\right)^{\exp(z'\beta)} = 1 - (1 - \hat{q}_x^0)^{\exp(z'\beta)}$$

II.1.3.7 Résidus de Schoenfeld :

Les résidus de Schoenfeld servent à vérifier l'hypothèse des hasards proportionnels. En effet, une hypothèse forte du modèle de Cox est que l'effet de chaque covariable doit être indépendant du temps. L'idée de ces résidus est de calculer pour chaque date de mort la différence entre les caractéristiques des individus décédés et une moyenne pondérée des caractéristiques des individus à risque à ce moment. Les résidus calculés par le logiciel R se présentent sous la forme d'une matrice à p colonnes et avec autant de lignes que d'observations non-censurées. Ils sont définis comme suit :

$$\widehat{Scho}_i^j = Z_{(i)}^j - \bar{Z}^j(\hat{\beta}, T_i)$$

Avec

$$\bar{Z}^j\left(\hat{\beta}, T_i\right) = \sum_{k=1}^{n} \frac{I(T_k < T, D_k = 1)}{\sum_{l=1}^{n} I(T_l \geq T_k).\exp\left(Z'_{(l)}\beta\right)} Z_{(k)}^j$$

On peut voir $\bar{Z}(\beta, t) = (\bar{Z}^1(\beta, t), \dots, \bar{Z}^p(\beta, t))$ comme le vecteur moyenne pondérée à l'instant t des vecteurs de covariables chez les individus à risque.

II.1.4 La méthode Whittaker-Henderson

La méthode Whittaker-Henderson permet de lisser les taux bruts $\hat{q}_x$ obtenus avec les différentes méthodes décrites ci-dessus. En effet, les $\hat{q}_x$ présentent souvent des irrégularités. On peut légitimement penser que le phénomène analysé, ici la mortalité en fonction de l'âge, est un phénomène régulier. Il semble, par exemple, raisonnable de penser que la probabilité de décéder dans l'année augmente avec l'âge ce qui rend un peu suspects les pics observés sur les taux bruts. Ces irrégularités sont imputées au manque de données. Avec plus de données, les $\hat{q}_x$ seraient plus proches des q_x et seraient, de ce fait, plus réguliers. On comprend alors la nécessité de lisser les taux bruts afin de rapprocher nos modèles de la réalité.

La méthode de Whittaker-Henderson est une méthode non-paramétrique, très utilisée pour lisser des résultats. Elle cherche à construire un lissage qui soit un équilibre entre un critère de fidélité et un critère de régularité. Ce lissage prend deux paramètres en compte : h et z deux entiers positifs dont nous allons voir l'utilité.

Soient $U' = (u_1, \dots, u_n)$ et $V' = (v_1, \dots, v_n)$ deux vecteurs avec dans U, les valeurs à lisser et dans V, les valeurs lissées.

On définit un critère de fidélité F tel que :

$$F = \sum_{i=1}^{n} w_i (v_i - u_i)^2$$

et un critère de régularité S tel que :

$$S = \sum_{i=1}^{n} [\Delta_z(v_i)]^2$$

avec w_i un vecteur de pondération (dans notre cas on peut prendre l'exposition) et Δ_z la différence d'ordre z.

Pour rappel, la différence d'ordre z est définie comme suit :

$$\Delta_z(v_i) = \sum_{k=0}^{z} \binom{z}{k} \times (-1)^{z-k} \times v_{i+k}$$

Elle peut être vue comme une approximation de la dérivée d'ordre z. En effet si on considère que v est une fonction z fois dérivable alors on peut approcher $v'(x)$ par $v(x+1) - v(x)$ qui n'est autre que $\Delta_1(x)$. De même, $v''(x) \approx \Delta_2(x)$, etc.

Le but de cette méthode est de trouver le lissage V qui minimise $M = F + hS$, qui est une combinaison linéaire des critères de fidélité et de régularité. Plus h est grand, plus on met l'accès sur la régularité.

La résolution de ce problème peut être trouvée dans le cours de modèles de durée de Frédéric Planchet[11].

II.1.5 Validation de l'ajustement

Une fois le modèle établi et les résultats lissés, il est bon de vérifier que les probabilités de décès obtenues sont bien cohérentes avec les données d'origine. Ainsi, on met en place un certain nombre de tests actuariels pour vérifier que le nombre de décès attendus par le modèle correspond bien à ce que l'on observe dans le jeu de données. Dans la partie III, nous utiliserons les cinq tests présentés ici.

II.1.5.1 Le test du SMR

Le *Standardized Mortality Rate* (SMR) est défini comme étant le ratio du nombre de décès observés sur le nombre de décès attendus. Il est parfois appelé *actual to expected ratio* (A to E ratio) dans la littérature anglo-saxonne.

$$SMR = \frac{A}{E}$$

Avec A le nombre de décès observés (*actual*) et E le nombre de décès attendus (*expected*).

Le SMR est souvent donné sous forme de pourcentage. Lorsqu'il est inférieur à 100%, on surestime les taux de mortalité et à l'inverse, lorsqu'il est supérieur à 100%, on observe plus de décès que l'on en prédit donc on sous-estime le risque de mortalité. Naturellement, un modèle est d'autant plus satisfaisant qu'il est proche de 100%, en particulier dans notre contexte de calcul d'un *best-estimate*. Toutefois, il peut arriver dans certains contextes que, par souci de prudence, l'on privilégie les modèles avec un SMR inférieur à 100%.

En faisant l'hypothèse que le SMR peut s'écrire comme suit, on peut construire un test permettant d'évaluer si le SMR est suffisamment proche de 100% pour que l'on puisse considérer le modèle comme satisfaisant (voir Liddell 1984).

[11] Planchet F., Thérond P-E. (2006) Modèles de durée, Applications actuarielles. Economica

Soit

$$SMR = \frac{\sum_x D_x}{\sum_x N_x . \hat{q}_x}$$

Avec

- D_x le nombre de décès observés à l'âge x. On suppose que D_x suit une loi de poisson.
- N_x l'exposition au risque à l'âge x.

Afin de tester l'hypothèse $H_0 : SMR = 1$, on introduit la statistique ξ comme suit :

- Si $SMR > 1,\ \xi = 3 \times D^{1/2}\left(1 - \frac{1}{9D} - \left(\frac{D}{E}\right)^{1/3}\right)$
- Si $SMR < 1,\ \xi = 3 \times D^{1/2}\left(\frac{1}{9D^*} + \left(\frac{D^*}{E}\right)^{1/3} - 1\right)$
- Si $SMR = 1,\ \xi = 0$

Avec $D = \sum_x D_x$, $D^* = D + 1$ et $E = \sum_x -N_x . \hat{q}_x$

Sous l'hypothèse H_0, ξ suit une loi normale centrée réduite.

Ainsi nous rejetons l'hypothèse H_0 (qui suppose une modélisation satisfaisante) lorsque la statistique ξ excède le quantile à 95% de la loi normale.

Cette approche permet aussi de déterminer un intervalle de confiance pour les valeurs estimées. Les bornes de l'intervalle de confiance à 95% du SMR sont données par Liddell :

$$SMR_{inf} = \frac{D}{E}\left(1 - \frac{1}{9D} - 1{,}96 \times \frac{1}{3\sqrt{D}}\right)^3$$

$$SMR_{sup} = \frac{D^*}{E}\left(1 - \frac{1}{9D^*} - 1{,}96 \times \frac{1}{3\sqrt{D^*}}\right)^3$$

On peut s'intéresser au SMR par âge ou pour n'importe quelle sous-population. Lorsque la quantité de données est limitée, le SMR par âge a peu de sens. Dans ce cas, on s'intéresse plutôt au SMR de la population globale.

II.1.5.2 Le test du χ^2

Le test du χ^2 permet de vérifier l'homogénéité de l'estimation au sein d'une variable catégorielle. C'est donc un excellent complément au test du SMR lorsque celui-ci est uniquement effectué sur la population globale.

En effet, un modèle peut donner un SMR global de 100% mais très mal prédire la répartition des décès entre les hommes et les femmes par exemple. Imaginons un portefeuille composé d`hommes et de femmes pour lequel on observe 200 décès. Imaginons encore que l'on

surestime la mortalité des hommes de moitié, disons 60 décès prédits de trop. Si on sous-estime la mortalité des femmes de 60 décès, alors le SMR global reste de 100% et l'on passe le test présenté ci-dessus bien que l'estimation ne soit absolument satisfaisante. C'est pourquoi il est utile d'introduire le test du χ^2.

On se donne une variable catégorielle à p modalités. Comme présenté dans les travaux de Hymans Robertson, on peut créer une statistique qui suit une loi du χ^2 en agrégeant les erreurs d'estimation de chaque modalité :

$$\sum_{i=1}^{p} \frac{\left(A^{(i)} - E^{(i)}\right)^2}{E^{(i)}} \sim \chi_p^2$$

Avec $A^{(i)}$ le nombre de décès observés pour la i-ème modalité et $E^{(i)}$ le nombre de décès attendus pour cette même modalité.

Pour utiliser ce test, il convient de supprimer tous les doublons de la base d'origine car il est très sensible à la dépendance entre les observations.

Dans le cas où l'ajustement testé a été obtenu à l'aide d'un lissage non-paramétrique, le nombre de degrés de liberté est difficile à déterminer. Nous rencontrerons ce cas de figure dans la partie III.3.3 et nous utiliserons alors la méthode décrite par Giesecke (1981).

II.1.5.3 Le test des signes

Si le modèle considéré estime de manière satisfaisante la probabilité que chaque individu a de décéder à l'âge x, alors le nombres de décès observés à l'âge x a une probabilité de 50% d'être au-dessus du nombre de décès attendus et une probabilité de 50% d'être au-dessous du nombre de décès attendus. Autrement dit, les SMR par âge doivent se répartir équitablement autour de 100%.

On peut construire un test basé sur cette observation. Celui-ci compare le nombre de fois où le SMR par âge est supérieur à 100% au nombre de fois où il est inférieur à 100% et si cet écart n'est pas trop important, le modèle est validé.

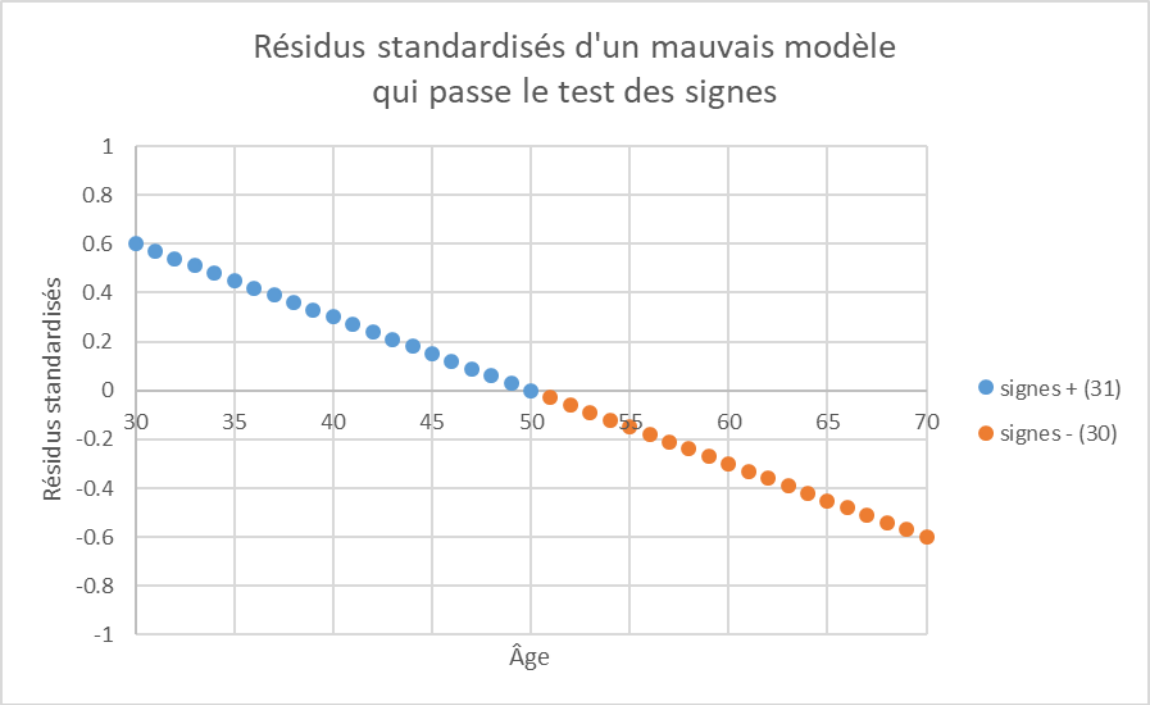
Prenons un modèle qui prédit la probabilité de décès à l'âge x pour n âges distincts. Notons r_x le résidu standardisé du modèle à l'âge x, $r_x = \frac{A_x - E_x}{E_x}$, avec A_x le nombre de décès observés à l'âge x et E_x le nombre de décès attendus à l'âge x. Notons, pour finir, $n_+ = \mathbb{I}_{\{r_x \geq 0\}}$ le nombre de résidus positifs et $n_- = n - n_+$ le nombre de résidus inférieurs à 0. Alors, on a :

$$\xi = \frac{n_+ - n_- - 1}{\sqrt{n}} \sim \mathcal{N}(0,1)$$

Si l`hypothèse selon laquelle les résidus sont équitablement répartis autour de zéro est exacte, alors ξ a 95% de chances d'appartenir à l'intervalle $[-1{,}96; 1{,}96]$. On rejette cette hypothèse si ce n'est pas le cas.

II.1.5.4 Le test des *runs*

Le test des signes est un bon indicateur mais il ne décrit qu'un aspect de la validité de la prédiction. Seul, il ne permet pas de trancher si un modèle est satisfaisant comme le montre l'exemple ci-dessous :



Le modèle donnant ces résidus passe le test des signes et on peut très bien imaginer que si les SMR sous-estimés compensent les SMR surestimés, alors il peut également passer les tests du χ^2 et du SMR global. C'est pourquoi il est important d'ajouter à notre liste de tests, le test des *runs*.

Un *run* est une séquence de résidus voisins de même signe. Par exemple dans la représentation des résidus proposés ci-dessus, il y aurait deux *runs* : un premier composé des résidus entre les âges 30 et 50 et un second composé des résidus entre 51 et 70 ans. L'idée de ce test est d'apprécier si le nombre de *runs* est suffisant pour que les observations soient considérées comme aléatoirement réparties autour des estimations. Sous l'hypothèse H_0, le nombre de *runs* d'une séquence de n âges est une variable aléatoire dont la distribution conditionnelle sachant le nombre n_+ de signes positifs et n_- le nombre de signes négatifs, avec $n = n_+ + n_-$ est approximativement normal, avec :

$$\mu = \frac{2n_+n_-}{n_+ + n_-} + 1$$

$$\sigma^2 = \frac{2n_+n_-(2n_+n_- - n)}{n^2(n-1)}$$

On a alors,

$$\xi = \frac{Nombre\ de\ runs - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

Comme pour le test des signes, on accepte l'hypothèse nulle, synonyme de bon modèle si $\xi \in [-1{,}96; 1{,}96]$.

II.1.5.5 Le test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov permet de vérifier que les probabilités de décès estimées $\hat{q}_x$ n'ont pas été lissées avec excès, au point de créer une différence significative entre les fonctions de répartition empiriques des échantillons observés et prédits.

Ce test a été décrit par Forfar et al (1988). Soit D l'écart maximal entre les deux fonctions de répartition et A et B définis comme plus haut. Alors la statistique ci-dessous suit une loi connue :

$$\xi = D\sqrt{\frac{AE}{A+E}}$$

On rejette l'hypothèse que l'ajustement est conforme aux données d'origine si cette statistique excède le quantile à 95% de cette loi.

Nous disposons donc d'un bel arsenal de tests pour valider l'adéquation de l'ajustement avec les données d'origine. Comme nous l'avons vu, ces tests sont complémentaires et une modélisation est réellement satisfaisante lorsqu'elle passe tous ces tests.

II.2 Présentation des données

Nous disposons donc de données fournies par deux partenaires. Ces données proviennent toutes de contrats d'assurance emprunteur individuelle et leurs principales caractéristiques sont présentées dans cette partie.

II.2.1 Format des données

Les données sont arrivées sous des formats divers selon le partenaire qui nous les a fournies. On dispose de différentes bases (assurés, sinistres, contrats, cotisations, …) que l'on peut lier entre elles via des identifiants communs à chaque base.

La première étape consiste à vérifier la cohérence des données entre les bases et à sélectionner le risque qui nous intéresse (ici l'assurance emprunteur individuelle).

II.2.2 Période d'observation

Après avoir vérifié la fiabilité des données, nous avons déterminé une période d'observation pour cette étude. Ce choix est important car le risque décès nécessite beaucoup de données pour être bien estimé. Le besoin de données nombreuses est particulièrement vrai pour l'étude d'une

population où l'on observe peu de décès ce qui est le cas des emprunteurs car il s'agit d'une population plutôt jeune, avec un emploi et ayant passé des formalités médicales.

On est donc tenté de prendre une période d'observation la plus large possible en prenant la date de début la plus ancienne possible et la date de fin la plus récente possible. Trois choses nous empêchent de procéder ainsi :

- Nous cherchons à estimer le risque décès qui pèse aujourd'hui sur la population assurée. En prenant des contrats trop anciens, on risque de modéliser le risque de mortalité actuel à partir d'une population aux caractéristiques différentes de celles d'aujourd'hui. La date de début d'observation doit donc être relativement proche de la date actuelle.
- Il y a toujours un délai entre le décès et l'acceptation définitive du sinistre par l'assureur. Les démarches prennent toujours du temps et parfois les dossiers nécessitent une analyse approfondie avant d'être acceptés. La date de fin d'observation doit donc être choisie de telle façon que les sinistres survenus avant cette date soient presque tous acceptés à la date d'extraction des données.
- Il est préférable que la période d'observation en mois soit un multiple de 12 car on observe une saisonnalité au cours de l'année[12] et une saison surreprésentée pourrait biaiser les résultats.

La date d'extraction la plus récente dont nous disposons est au 01/05/2020. L'analyse de la cadence d'acceptation des sinistres a été menée séparément entre les données de l'assureur A et les données de l'assureur B car ces délais peuvent varier d'un système de gestion à l'autre. En prenant tous les sinistres survenus avant 2017 et en nous intéressant à la durée entre la survenance et la clôture des sinistres acceptés, nous avons créé une fonction de survie, à retrouver en annexe, qui nous donne le pourcentage de sinistres enregistrés à la date d'extraction en fonction de la date de fin d'observation. On ne prend que les sinistres survenus entre 2010 et 2016 pour construire cette fonction car les sinistres recevables des années 2017, 2018, 2019 et 2020 ne sont peut-être pas encore tous acceptés. Grâce à cet outil nous constatons que si la cadence de règlement est restée constant depuis 2010, les sinistres survenus en juin 2019 ont plus de 95% de chance d'avoir été clôturés à la date d'extraction. On en déduit que la quasi-totalité des sinistres survenus entre 2010 et juin 2019 sont présents dans notre base.

En prenant en compte les considérations discutées ci-dessus, nous avons fixé la période d'observation entre le 01/07/2012 et le 30/06/2019.

II.2.3 Les variables disponibles

Les données reçues comportent de nombreuses variables. Seules les variables apportant des informations utiles à notre étude ont été conservées.

Certaines variables sont indispensables à la construction d'une table :

- L'identifiant assuré
- La date de naissance

[12] L'INSEE fait mention de la saisonnalité du taux de décès : https://www.insee.fr/fr/statistiques/4804802#consulter

- Le sexe
- La date de souscription
- L'identifiant sinistre
- La date de sortie
- Le motif de sortie

Des variables optionnelles apportent une information intéressante pour notre analyse :

- La catégorie socio-professionnelle (CSP)
- Le tabagisme éventuel de l'assuré
- Le code postal

Malheureusement nous ne disposons pas d'informations fiables sur le montant du prêt. Les données sont parfois incohérentes, voire manquantes. Nous disposons seulement d'une variable servant à marquer les prêts dont le montant est supérieur à 500 000 €.

Les deux assureurs ont la même définition de la variable « fumeur » et les catégories socio-professionnelles sont très proches. Il est aussi à noter que les deux cédantes ont des processus de sélection médicale équivalents et que ceux-ci peuvent être qualifiés de standards.

À partir de la date de souscription est créée la variable *Ancienneté*. L'ancienneté est le temps que l'assuré a passé dans le contrat et c'est donc le temps qui sépare l'assuré de la sélection médicale. L'ancienneté est comptée en années et commence à 1 pour les 365 premiers jours du contrat puis passe à 2 pour l'année suivante et ainsi de suite. La base de données comportait une ligne par assuré. L'introduction de cette nouvelle variable a nécessité la création de nouvelles lignes car chaque assuré doit posséder une ligne par année d'ancienneté. Les variables dépendantes du temps (notamment le début et la fin de l'observation) ont été adaptées en conséquence.

II.2.4 Statistiques descriptives

Dans cette partie nous présenterons quelques données chiffrées sur les différents portefeuilles.

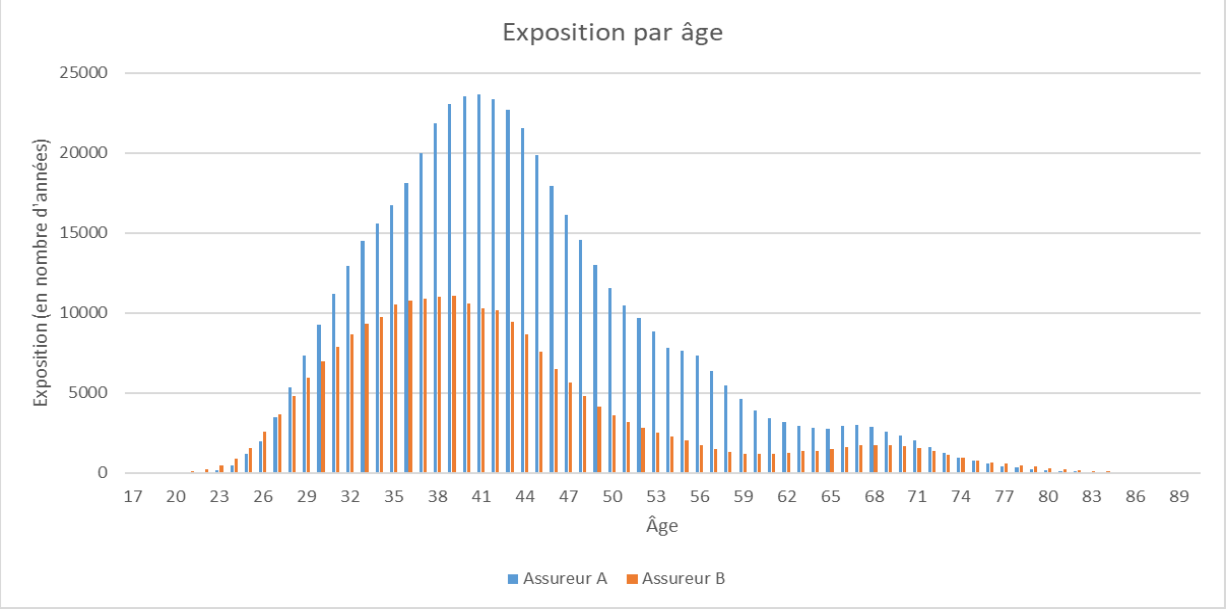
Le tableau ci-dessous présente les principales caractéristiques des deux jeux de données :

	Assureur A	Assureur B
Période d'observation :	01/07/2012 - 30/06/2019	01/07/2012 - 30/06/2019
Âge moyen :	42,8 ans	41,3 ans
Exposition totale :	585 172 années	319 011 années
Part Outre-mer :	2,75%	14,30%
Nombre de décès :	430	319
Ancienneté moyenne :	3,01 années	5,07 années

Le poids de l'assureur A est plus important car son exposition totale est substantiellement supérieure à celle de l'assureur B. Les deux portefeuilles diffèrent par le pourcentage d'assurés hors-métropole et par l'ancienneté moyenne. Nous testerons l'impact de ces variables dans la suite à l'aide du modèle de Cox. L'âge moyen est jugé assez proche.

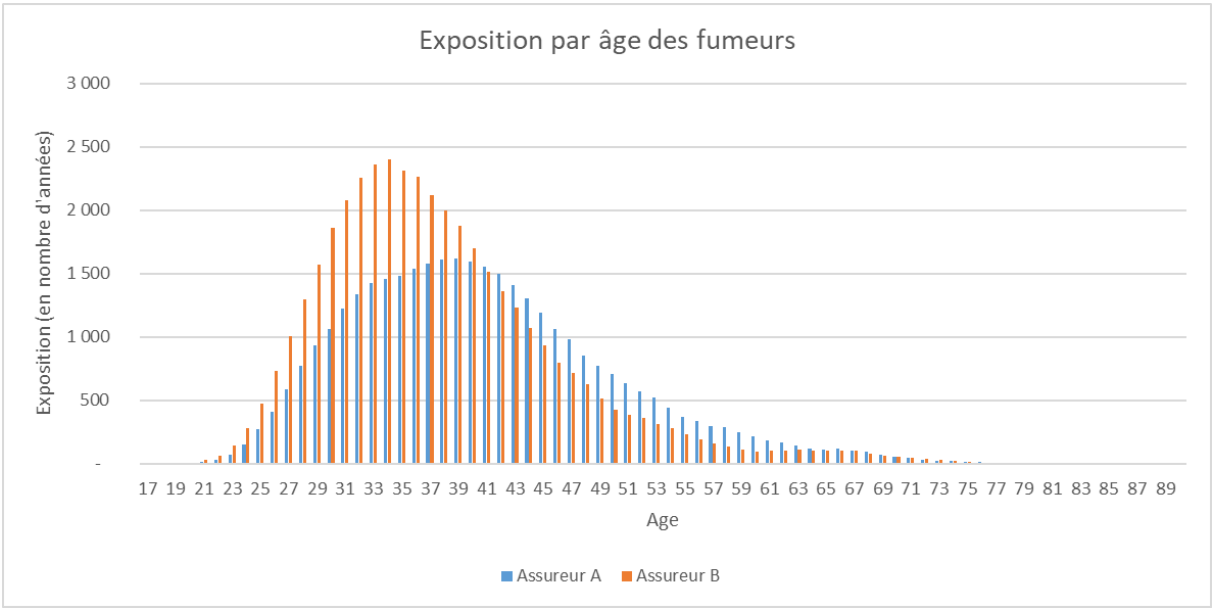
L’ancienneté est une variable à surveiller en assurance emprunteur. Elle est définie comme le nombre d’années depuis la souscription. En effet, comme discuté au paragraphe I.3.3, la sélection médicale a un impact sur la mortalité des assurés dans leurs premières années après souscription. Elle est assez nettement supérieure chez l’assureur B.

Nous nous intéressons maintenant à la courbe des âges pour mieux appréhender le profil de notre population et pour déceler éventuellement une différence entre les assureurs :

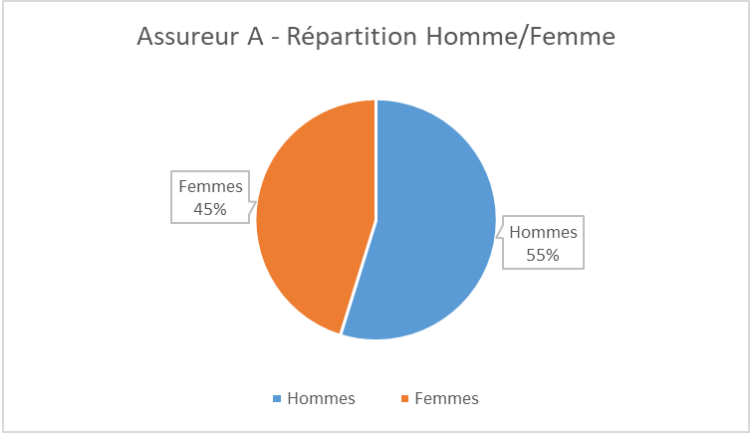


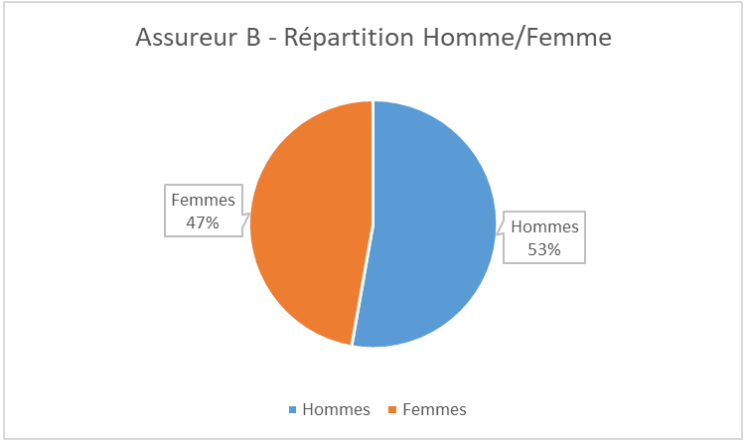
Les deux assureurs ont des profils très proches.

En vérifiant que les deux portefeuilles avaient la même courbe des âges pour chaque sous-catégorie (par csp, par sexe, …) nous avons constaté un écart pour la sous-population « Fumeurs » :

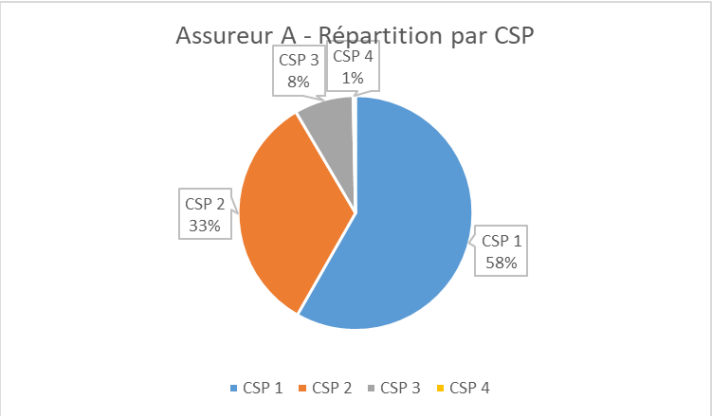


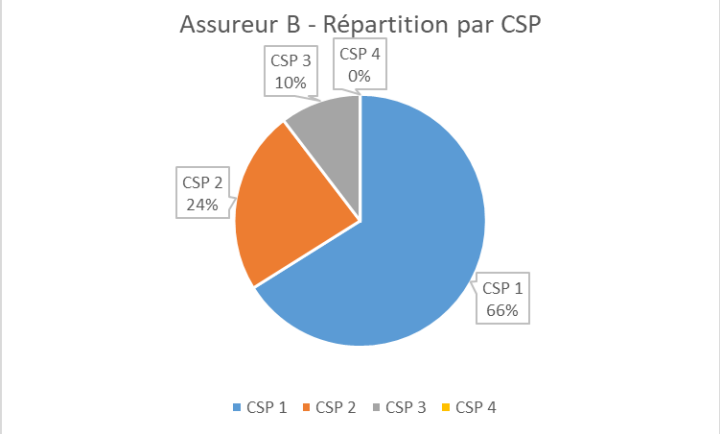
La représentation des différentes sous catégories est présentée ci-dessous :





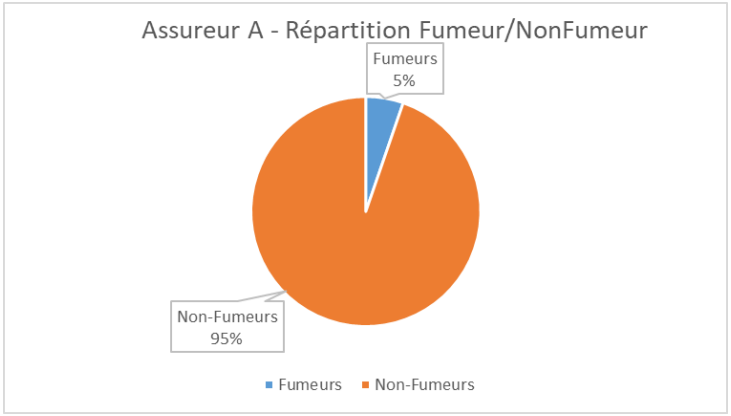
La répartition homme/femme est sensiblement la même dans les deux portefeuilles. Il y a très légèrement plus d'homme que de femmes dans nos données.

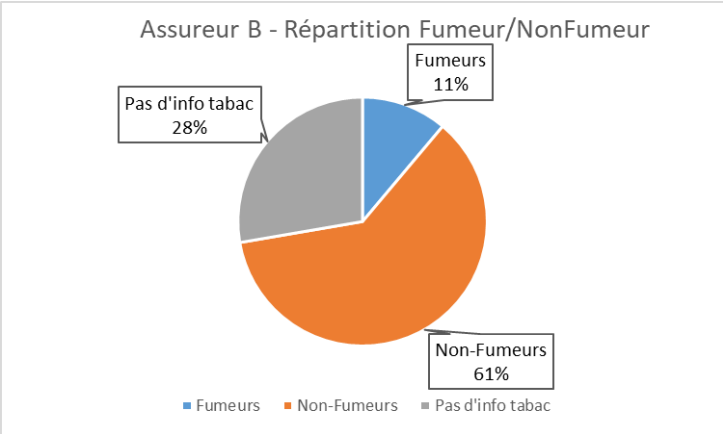




Il y a plus de Csp 1 chez l'assureur B que chez l'assureur A. La part de Csp 2 est plus importante dans le portefeuille A que dans le portefeuille B. On note également qu'il n'y a pas de Csp 4 dans le portefeuille B.

Ces deux graphiques montrent donc une hétérogénéité entre les portefeuilles. Il apparait également que la masse de données sera sans doute trop faible en Csp 4 pour effectuer une analyse précise de cette sous-population.

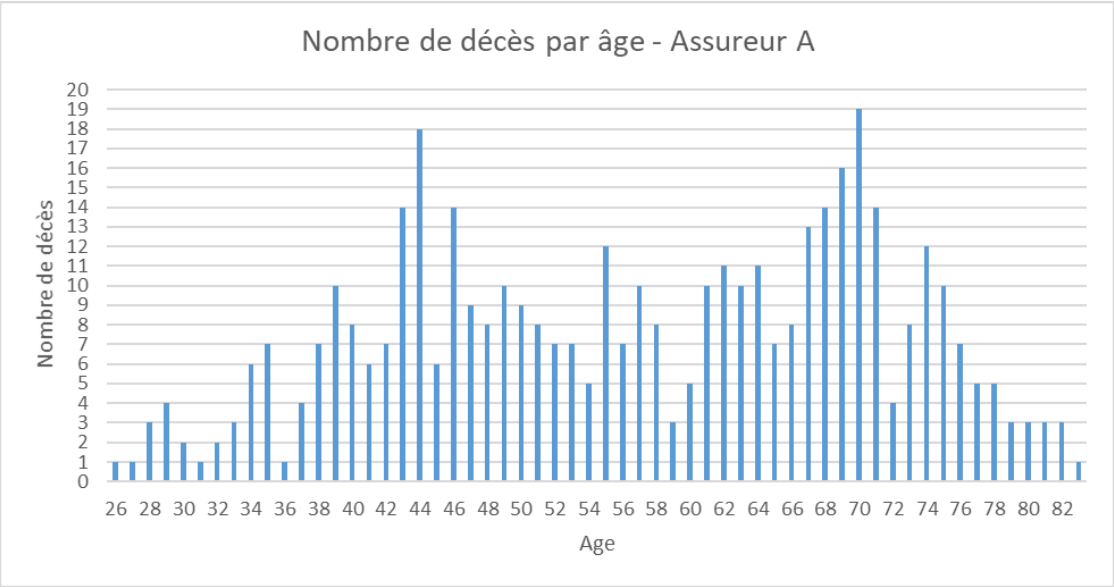


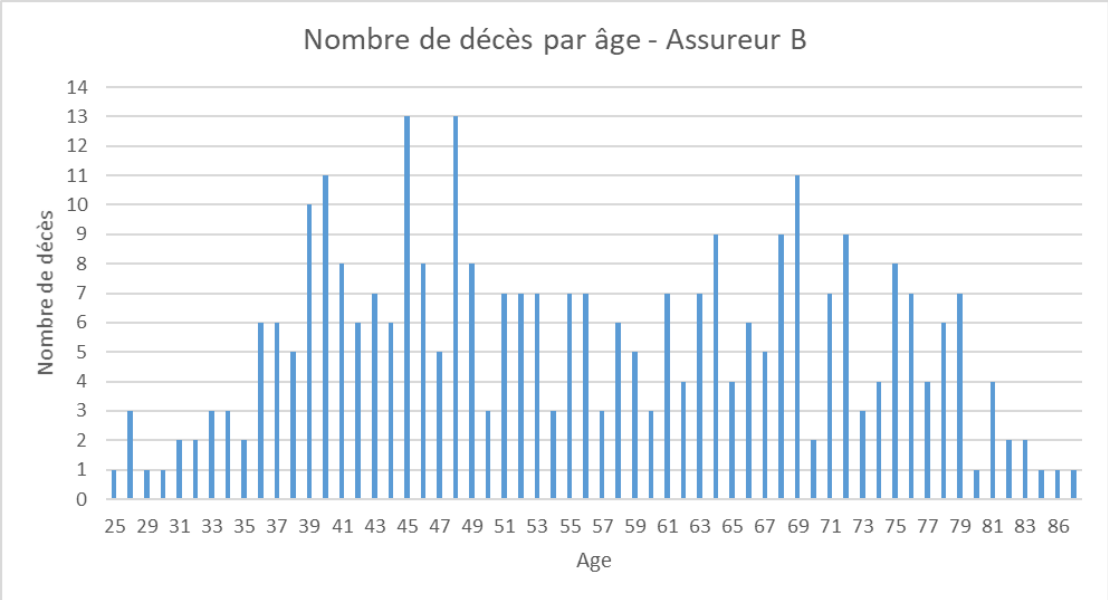


Il y a une part plus importante de fumeurs dans le portefeuille B. On remarque aussi que de nombreuses données sont manquantes dans le portefeuille B. Cela correspond aux contrats où il n'y a pas eu de segmentation Fumeur/Non-fumeur. Lorsque nous avons appliqué le modèle de Cox, nous n'avons pas pris en compte ces lignes.

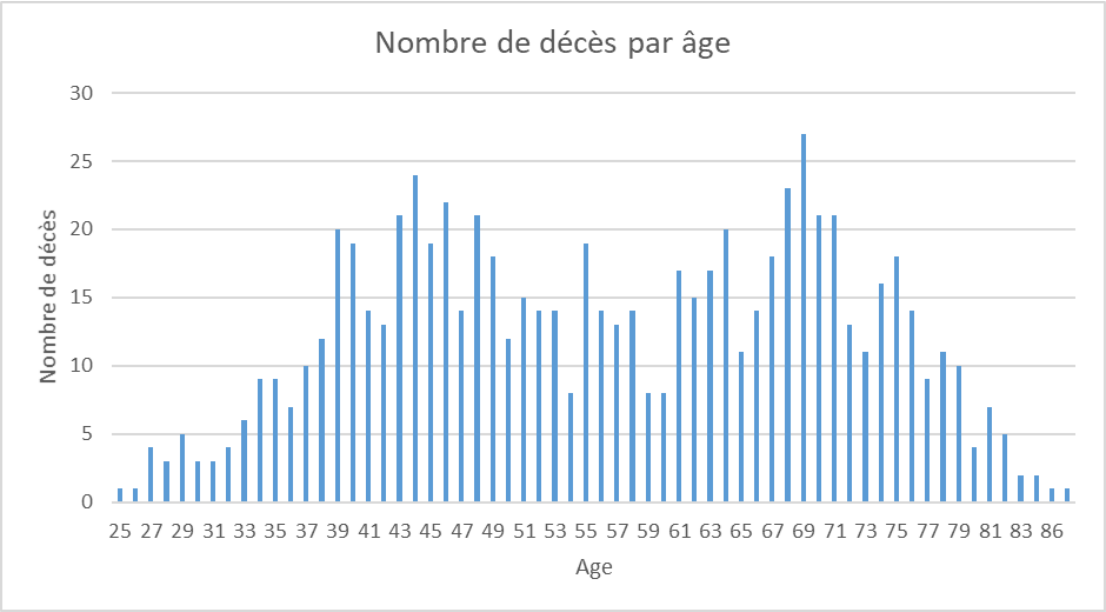
Lorsque l'on sélectionne une sous-population, les proportions se conservent plutôt bien. Par exemple, la proportion Homme/Femme des Non-fumeurs du portefeuille A est à peu près la même que la proportion Homme/Femme des Non-fumeurs du portefeuille B. Nous ne pouvons pas présenter toutes les combinaisons possibles ici.

Un enjeu important est le nombre de décès. Ces évènements sont assez rares, particulièrement aux âges jeunes, ce qui peut compliquer l'estimation des taux bruts. L'utilisation d'un modèle non-modèle paramétrique comme l'estimateur des moments de Hoem peut être délicate car il faut pouvoir proposer une estimation de la probabilité de décès à chaque âge, or si l'on n'observe pas ou peu de décès à certains âges, l'estimation sera peu fiable. Les modèles paramétriques peuvent alors être une alternative intéressante car, du fait que l'on n'estime pas les paramètres âge par âge, ils sont moins sensibles à ce type de problématiques. Ces modèles nécessitent toutefois l'établissement d'une hypothèse sur la forme générale de la mortalité. Ne souhaitant pas formuler une telle hypothèse car n'ayant pas une idée précise de la forme de la fonction de hasard de la population des emprunteurs, nous décidons de vérifier que la quantité de données est suffisante pour appliquer un modèle non-paramétrique. Le critère sélectionné est le critère de Cochran qui permet de juger de la suffisance des données pour faire l'approximation gaussienne que nous utilisons par la suite. Selon ce critère, il faut avoir au moins cinq décès par âge et au moins cinq survivants. Ces conditions ne sont vérifiées ni pour le portefeuille de l'assureur A ni pour le portefeuille de l'assureur B :





Cette constatation nous motive à regrouper les deux populations, après avoir vérifié l'homogénéité des données, afin d'obtenir un nombre suffisant de décès à chaque âge. Le nombre de décès par âge pour la population globale est présenté ci-dessous :



Tous les âges compris entre 33 et 79 ans ont plus de cinq décès.

III Modélisation

Nous allons présenter dans cette partie la mise en place de modèles de durée sur nos données. L'objectif est de déterminer quelles variables influent sur la durée de vie et nous allons également essayer de quantifier cet impact.

III.1 Approche naïve

Une première approche consiste à créer des sous-populations et à estimer sur chacune d'entre elles les taux de mortalité. Cette approche est très simpliste mais elle permet d'obtenir une première visualisation des taux de mortalité. Nous discuterons des limites de cette technique à la fin de cette partie.

La base construite en regroupant les données de l'assureur A et de l'assureur B doit servir de base pour nos tarifications futures. Dans ce contexte, nous aimerions tester si le fait d'appartenir à l'assureur A ou à l'assureur B a un impact sur la mortalité. En effet si l'origine des données joue un rôle important, nous pourrons conclure que la construction de notre base ne pourra pas servir de référence pour l'ensemble de la population des emprunteurs car il faudra ajuster la mortalité pour chaque cédante. Nous pourrons simplement utiliser les coefficients relatifs à appliquer suivant les caractéristiques de l'assuré. Par exemple le coefficient correspondant à la surmortalité des assurés fumeurs par rapport aux non-fumeurs. Dans le cas où les données de l'assureur A et de l'assureur B sont homogènes, nous pourrons considérer que l'origine des données ne joue pas un rôle très important et les taux de mortalité obtenus pourrons servir de référence pour la population des emprunteurs[13]. Nous allons donc appliquer l'approche naïve à la variable *Origine*.

III.1.1 Obtention des courbes

Dans la suite, les taux bruts sont obtenus grâce à l'estimateur des moments de Hoem. On effectue une légère transformation des données avant d'appliquer l'estimateur. En effet, comme on l'a montré plus haut, il est souhaitable d'avoir au moins cinq décès par âge pour construire une table de mortalité. Or le fait de considérer des sous-populations diminue drastiquement le nombre de décès observés. Pour tenter de remédier à ce problème, nous avons utilisé une technique dite de glissement par âge pour augmenter artificiellement le nombre de décès. Cette technique consiste à changer l'exposition et le nombre de sinistres de la façon suivante pour chaque tranche d'âge (voir le mémoire de Gabriel Prophillat pour plus de précisions sur cette méthode) :

- $$E_x^{agrégée} = \begin{cases} \sum_{i=x}^{x+1} E_x^{initiale}, \text{si } x = x_{min} \\ \sum_{i=x-1}^{x+1} E_x^{initiale}, \text{si } x \in [\![x_{min}+1\,;x_{max}-1]\!] \\ \sum_{i=x-1}^{x} E_x^{initiale}, \text{si } x = x_{max} \end{cases}$$

[13] Cette hypothèse reste très forte et lorsque cela est possible, il est important de vérifier cette hypothèse au cours de la tarification.

- $d_x^{agrégé} = \begin{cases} \sum_{i=x}^{x+1} d_x^{intial}, \text{si } x = x_{min} \\ \sum_{i=x-1}^{x+1} d_x^{initial}, \text{si } x \in [\![x_{min}+1 ; x_{max}-1]\!] \\ \sum_{i=x-1}^{x} d_x^{initial}, \text{si } x = x_{max} \end{cases}$

avec,

- $E_x^{initiale}$: exposition à l'âge x calculée à partir des données initiales ;
- d_x^{intial} : nombre de décès observés à l'âge x à partir des données initiales ;
- $E_x^{agrégée}$: exposition à l'âge x calculée en utilisant la méthode de glissements par âge ;
- $d_x^{agrégé}$: nombre de décès à l'âge x déterminé en utilisant la méthode de glissements par âge.

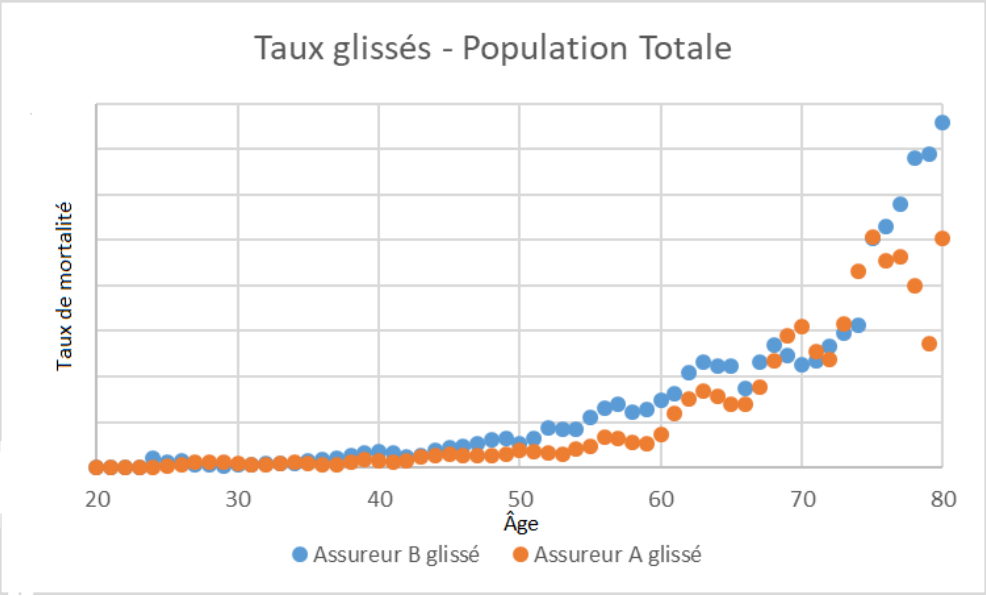
Les taux bruts obtenus avec l'estimateur de Hoem après l'application de cette méthode reviennent à prendre les taux bruts de la population originale (sans cette modification) et à transformer pour chaque âge x, q_x en la moyenne entre q_{x-1}, q_x et q_{x+1} pondérée par les expositions.
Le nombre de décès observés par tranche d'âge x passant artificiellement à la somme des décès observés en âge x-1, x et x+1, on vérifie plus facilement le critère de Cochran.

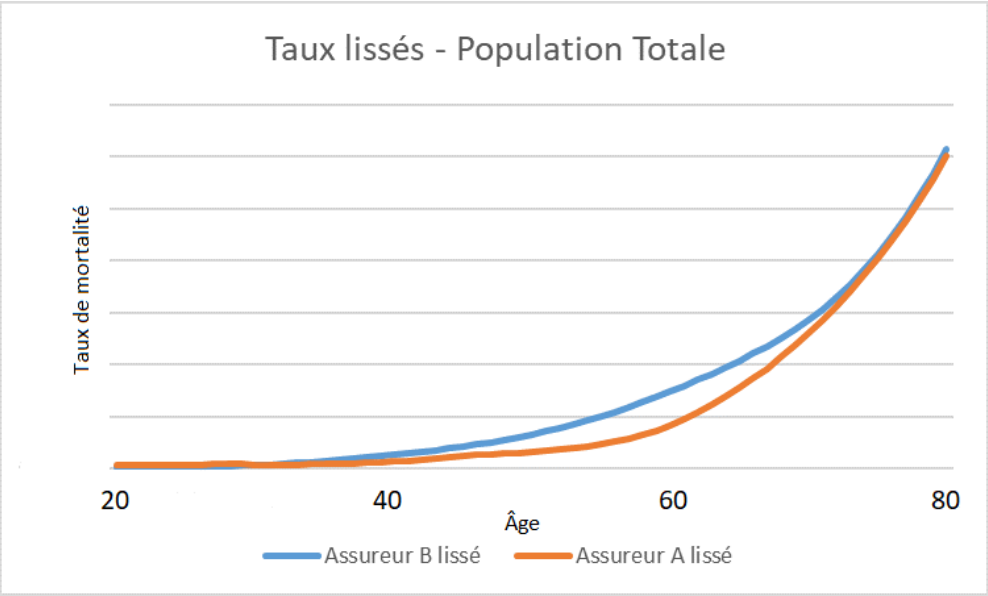
Ces taux sont ensuite lissés avec la méthode Whittaker Henderson (avec z = 2 et h = 5). On compare alors l'écart sur les taux glissés et sur les taux lissés.

III.1.2 Population globale

Dans un premier temps nous avons comparé les populations totales des deux bases. Les taux obtenus sont représentés ci-dessous.

Notons que dans toute cette partie, nous avons censuré l'échelle de l'axe des ordonnés afin de ne pas révéler de données sensibles. Ce retrait n'est en rien dérangeant dans notre démarche car nous nous intéressons ici à la différence de forme entre les taux de l'assureur A et de l'assureur B.





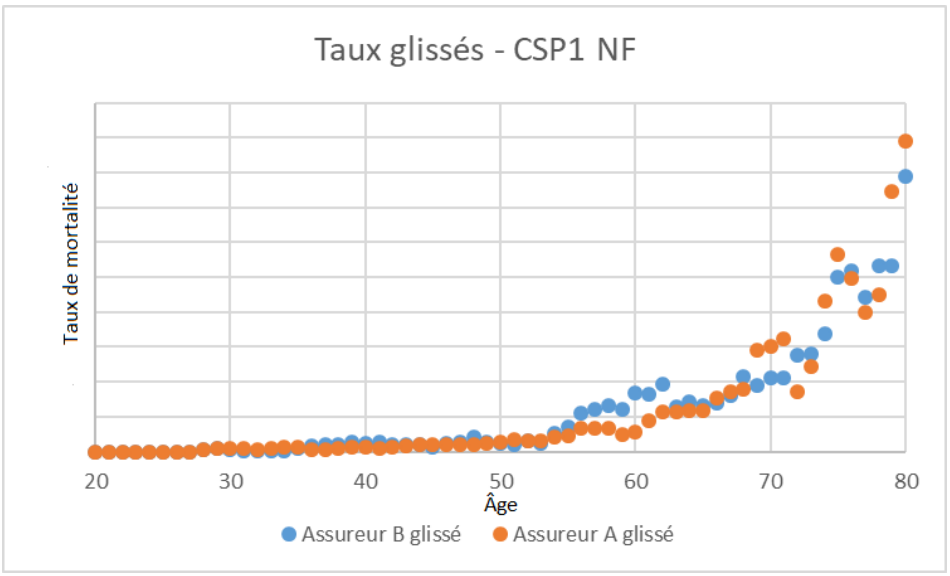
On observe une réelle différence entre les estimations des deux sous-populations. En faisant une moyenne des écarts pondérée par l'exposition, on trouve que les assurés de l'assureur B ont une probabilité de décès plus de deux fois supérieure aux assurés de l'assureur A.

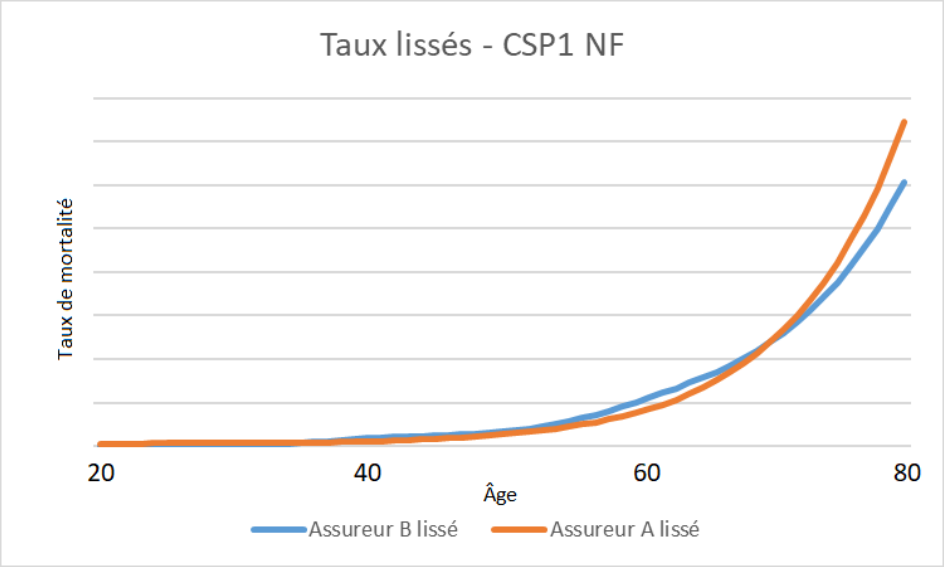
Pourtant nous n'en concluons pas qu'il y a une différence due à la variable *Origine*. En effet, les deux portefeuilles n'ont pas la même proportion de Csp, de fumeurs etc. La différence que nous observons vient peut-être de ces variables. Nous conservons ces graphiques à l'esprit car ils donnent une première vision des taux de mortalités de ces deux populations.

III.1.3 Population Csp 1 non-fumeur

Pour davantage cibler l'impact de la variable *Origine* nous avons appliqué la même méthode à la population Csp 1, non-fumeur. Aucune segmentation au niveau du sexe n'a été effectuée car les deux sous-populations sélectionnées ont des repartions homme/femme très proches.

On obtient alors les taux suivants :





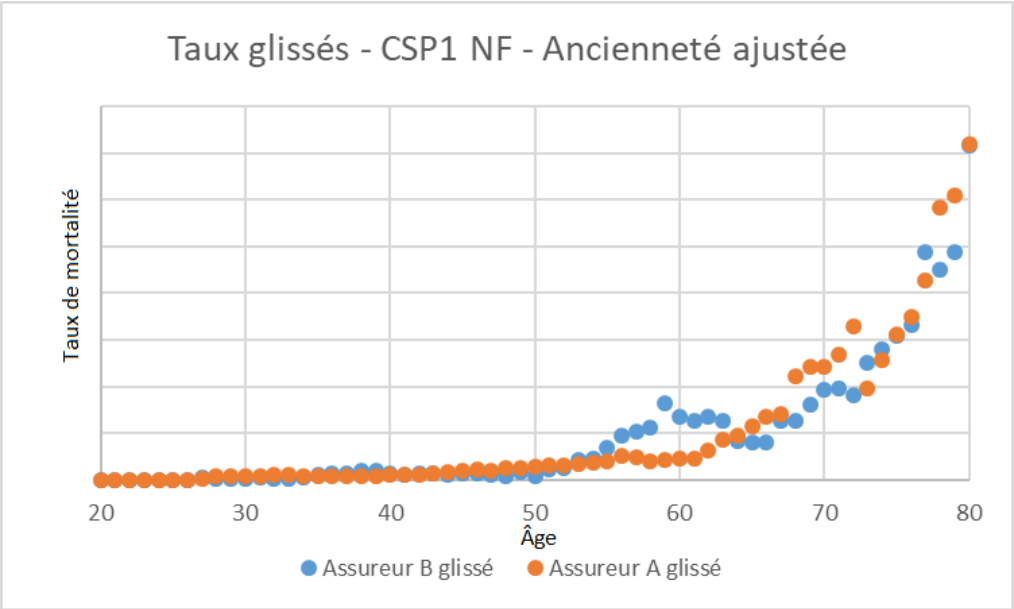
On observe un rapprochement de la mortalité entre les assurés A et les assurés B. L'écart moyen passe à 40% sur les taux glissés et à 15% sur les taux lissés. Cet écart entre taux lissés et taux glissés sera discuté un peu plus loin dans la partie III.1.5 Limites de l'approche naïve.

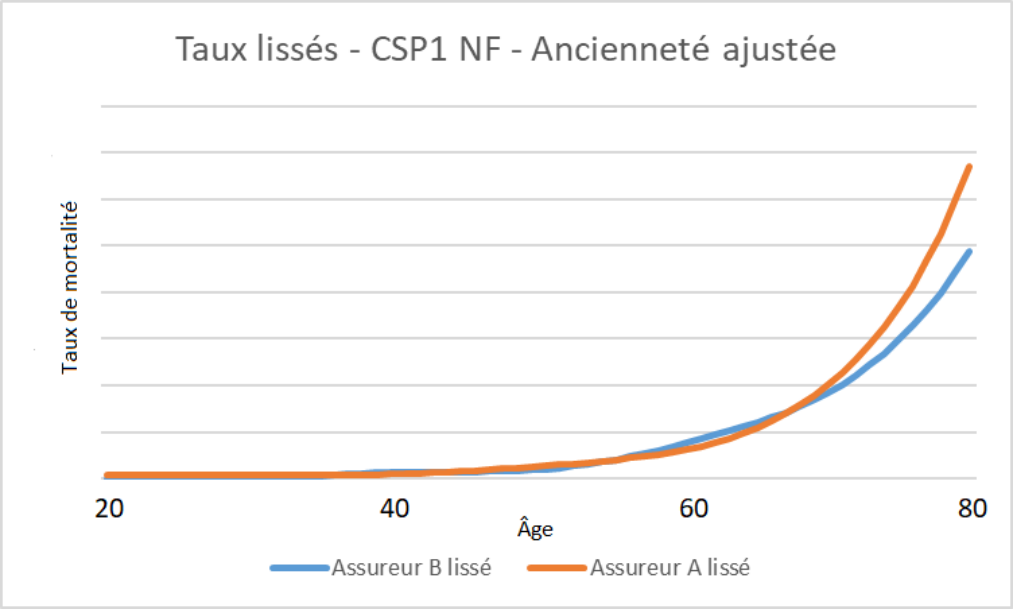
Nous avons essayé de rapprocher les deux courbes en ne considérant que la population métropolitaine mais cette sélection n'a pas apporté de résultats très probants. Une variable qui a nécessairement un impact et qui n'est pas égalisée dans les sous-populations Csp 1 non-fumeurs est l'ancienneté.

III.1.4 Population Csp 1 non-fumeur avec ancienneté ajustée

Pour que les deux sous-populations soient comparables, il faut qu'elles aient la même ancienneté moyenne. Pour cela nous avons conservé des deux sous-populations que l'exposition correspondant à une ancienneté inférieure ou égale à cinq ans. Une fois cette nouvelle sous-population sélectionnée, nous observons des ancienneté moyennes très proches pour les portefeuilles A et B (0,3 année d'ancienneté de différence).

On obtient les résultats suivants :





Les écarts moyens pondérés par l'exposition sont plus faibles que précédemment : 8% pour les taux glissés et 4% pour les taux lissés. On observe tout de même une "bosse" chez l'assureur B autour de 60 ans. Celle-ci casse un peu la forme habituellement observée en mortalité (croissance exponentielle). Nous l'attribuons à un manque de données plutôt qu'à un phénomène propre à cette population.

III.1.5 Limites de l'approche naïve

Le premier reproche que l'on peut faire à cette méthode est le fait qu'il peut y avoir un écart important entre les écarts des taux glissés et lissés. Cela pose naturellement des problèmes d'interprétation mais est aussi révélateur d'une certaine faiblesse de nos estimations quand bien même les sous-populations testées dans cette partie étaient probablement parmi les plus représentées du portefeuille.

Un deuxième problème apparait lorsque nous cherchons à conclure sur les résultats obtenus. Quel est l'impact de la variable *Origine* ? Difficile à dire même après les différents ajustements effectués car on peut difficilement savoir si les 8% d'écart sont révélateurs d'un véritable phénomène sous-jacent ou si on a simplement affaire à deux réalisations différentes d'une même variable aléatoire.

On note également que nous avons testé les variables que nous pensions intéressantes pour expliquer la mortalité mais peut-être qu'il en existe d'autres et cette méthode ne nous permet pas de les déterminer.

Cette approche est quasiment impossible à mettre en œuvre pour segmenter notre portefeuille. En effet, il faudrait faire une table de mortalité pour chaque sous-population. En plus d'être fastidieuse, cette tâche n'est pas réalisable pour les catégories très peu représentée dans la population totale car les estimateurs n'auraient que trop peu de valeur statistique. Par exemple pour les Csp 4 – Fumeurs – Femmes, il n'y a pas de décès. Même avec une approche paramétrique, il est impossible de construire une courbe de mortalité sans décès.
Il faut donc choisir une approche différente pour segmenter notre portefeuille.

III.2 Test d'homogénéité et détermination des variables intéressantes

III.2.1 Le log-rank test

Le log-rank test permet de tester l'hypothèse H_0 : « Les distributions de survie sont identiques » pour deux jeux de données censurés.

III.2.2 Homogénéité des données

L'application de ce test à la variable *Origine* doit nous permettre de déterminer si les données sont homogènes. Cette validation est très importante car c'est elle qui nous permet de justifier le fait que l'on peut fusionner ces deux jeux de données.

Le test donne une p-value de 20% ce qui signifie que l'on conserve l'hypothèse H_0. En d'autres termes il n'y a pas de différence significative entre les données de l'Assureur A et celles de l'Assureur B.

III.2.3 Les variables intéressantes

L'application de ce test permet de déterminer les variables qui pourraient être intégrées au modèle de Cox. Les résultats du test sont présentés ci-dessous :

Variable	p-value
Civilité	$2{,}71.10^{-9}$
Tabagisme	$< 2.10^{-16}$
DOMTOM	0,26
Csp1	0,038
Ancienneté 2	$1{,}67.10^{-13}$
Origine	0,20

On voit que la variable *Outre-mer* n'est pas significative. La variable *Csp* n'est significative que si l'on teste la Csp 1 contre toutes les autres et l'ancienneté est significative de 1 à 15 ans (nous n'avons pas repris tous les résultats dans le tableau ci-dessus).

La variable *Civilité* ne peut pas être utilisée pour tarifer un contrat d'assurance. Son interdiction a été établie par la Cour de Justice de l'Union Européenne en 2011 avec l'arrêt « Test-Achats » afin d'empêcher toute discrimination entre les hommes et les femmes. Pourtant nous la conserverons dans notre modèle car cette variable a un fort pouvoir explicatif. En pratique lorsqu'un nouveau produit est lancé, on utilise une estimation de la répartition homme-femme pour obtenir un tarif indifférencié.

III.3 Application du modèle de Cox

Pour appliquer le modèle de Cox, nous utilisons le package *survival* du logiciel R. La démarche est présentée ci-dessous et a été appliquée aux données présentées en II.2. moins les Csp4 et les personnes dont on n'a pas d'information sur le tabagisme.

Comme nous l'avons vu dans la partie théorique, l'estimation des taux de mortalité $\hat{q}_x$ se fait en deux étapes. Nous commencerons par estimer le vecteur des coefficients $\hat{\beta}$ pour un certain modèle qu'il faudra sélectionner. Une fois cette première étape effectuée, nous pourrons estimer les taux de mortalité par âge grâce à l'application de l'estimateur de Breslow. Ces taux de mortalité de référence seront alors lissés afin d'obtenir les $\hat{q}_x$ finaux. Les résultats seront observés entre les âges de 18 et de 84 ans. Dans cette partie nous procéderons également à la validation du modèle. Pour des raisons de confidentialité, les résultats ont été modifiés ou censurés avant d'être partagés dans ce mémoire.

III.3.1 Sélection du modèle

Comme nous l'avons vu dans la partie théorique, le modèle de Cox consiste dans un premier temps à trouver le vecteur β par maximisation de la vraisemblance partielle. L'objectif de cette partie est de déterminer le modèle qui permet d'estimer la probabilité de décès avec le plus de précision et la plus grande simplicité. Nous qualifierons de meilleur modèle celui qui a la plus grande concordance, la plus petite AIC et dont toutes les variables sont significatives au seuil de 5%.

La concordance est une mesure usuelle pour mesurer la validité de l'ajustement d'un modèle de durée. C'est la probabilité que les prédictions aillent dans le même sens que les données observées (voir Therneau et Atkinson 2020). C'est une valeur qu'il faut essayer de maximiser.

L'AIC (*Akaike information criterion*) est une mesure de la qualité d'un modèle statistique. Il part du constat que l'on peut toujours améliorer la vraisemblance d'un modèle en ajoutant des variables explicatives. L'idée est donc de pénaliser l'ajout de variables explicatives afin d'éviter que celui-ci ne soit abusif. L'AIC s'écrit comme suit :

$$AIC = 2p - 2\log\left(\mathcal{L}(\hat{\beta})\right)$$

Avec p le nombre de variables explicatives et $\mathcal{L}(\hat{\beta})$ la vraisemblance du modèle avec les paramètres $\hat{\beta}$.

L'AIC est une mesure que l'on cherche à minimiser.

Pour déterminer le modèle optimal, nous partons du modèle qui intègre toutes les variables explicatives puis nous retirons progressivement les variables qui ne sont pas significatives.

Le modèle complet :

Variable	Exp($\widehat{\beta}$)	p-value	Significativité
Civilité M	$\beta_M^{Civilité}$	$4{,}29.10^{-8}$	***
Fumeur OUI	β_{oui}^{Fumeur}	$< 2.10^{-16}$	***
Csp 2	β_2^{Csp}	0,14	
Csp 3	β_3^{Csp}	0,11	
Csp 4	β_4^{Csp}	0,98	
Origine Ass-2	$\beta_2^{Origine}$	0,16	
DOMTOM Oui	β_{oui}^{DomTom}	0,67	
Ancienneté 2	β_2^{Anc}	$1{,}96.10^{-5}$	***
Ancienneté 3	β_3^{Anc}	$2{,}53.10^{-10}$	***
Ancienneté 4	β_4^{Anc}	$8{,}00.10^{-12}$	***
Ancienneté 5	β_5^{Anc}	$6{,}84.10^{-9}$	***
Ancienneté 6	β_6^{Anc}	$7{,}96.10^{-9}$	***
Ancienneté 7	β_7^{Anc}	$2{,}16.10^{-14}$	***
Ancienneté 8	β_8^{Anc}	$1{,}37.10^{-8}$	***
…			

*** désigne une significativité à 0,1%, ** à 1%, * à 5%, . à 10%.

AIC :	11 612,86
Concordance :	0,663

Les anciennetés supérieures à 8 n'ont pas été reprises dans ce tableau pour des raisons de clarté. En effet, on peut observer des assurés avec 20 ans d'ancienneté. Sur les 12 anciennetés supérieures à 8, seules 4 modalités sont significatives.

Ces résultats nous indiquent que la variable *DOMTOM* n'est pas significative. Nous la retirons donc du modèle.

La sélection des variables à conserver :

Après avoir retiré la variable *DOMTOM*, nous obtenons des résultats assez semblables au modèle complet. Nous décidons alors de retirer une nouvelle variable non-significative. La variable la moins significative est alors la variable *Origine*, c'est donc elle que nous retirons du modèle.

Nous obtenons alors un modèle avec quatre variables explicatives : la civilité, la catégorie socio-professionnelle, le tabagisme et l'ancienneté. Un extrait des résultats obtenus avec ce modèle est donné ci-dessous :

Variable	Exp($\widehat{\beta}$)	p-value	Significativité
Civilité M	$\beta_M^{Civilité}$	$5{,}25.10^{-8}$	***
Fumeur OUI	β_{oui}^{Fumeur}	$< 2.10^{-16}$	***
Csp 2	β_2^{Csp}	0,16	
Csp 3	β_3^{Csp}	0,11	
Csp 4	β_4^{Csp}	0,98	
Ancienneté 2	β_2^{Anc}	$2{,}31.10^{-5}$	***
Ancienneté 3	β_3^{Anc}	$3{,}48.10^{-10}$	***
…			

*** désigne une significativité à 0,1%, ** à 1%, * à 5%, . à 10%.

AIC :	11 611,26
Concordance :	0,663

L'AIC s'améliore légèrement car nous retirons des variables. Nous constatons que la variable *Csp* n'est pas significative. Malgré tout, des avis d'experts nous portent à croire que cette variable joue un rôle dans la mortalité des assurés. De plus, l'utilisation de cette variable est extrêmement répandue sur le marché. Si nous n'utilisons pas cette variable, alors nous serons exposés au risque d'anti-sélection. Nous attirerons beaucoup d'assurés en *csp* 3 et 4 dans nos portefeuilles et si l'avis des experts, qui est aussi un avis de bon sens, se révèle correct, nous pouvons perdre beaucoup d'argent. Nous décidons donc de conserver cette variable.

Discutons rapidement des variables retenues :

- *Civilité* : Nous ne sommes pas surpris de la présence de cette variable. Les tables réglementaires TH 00-02 et TF 00-02, respectivement pour les hommes et pour les femmes, ont déjà montré la différence de mortalité entre ces deux populations. D'après ces tables, à 45 ans les hommes ont deux fois plus de chance que les femmes de décéder dans l'année.
- *Fumeur* : Cette variable fait naturellement partie des variables sélectionnées. Fumer augmente la probabilité de développer des maladies pulmonaires ou cérébrales[14] et augmente donc le risque de mortalité.

[14] https://www.santepubliquefrance.fr/determinants-de-sante/tabac/articles/quelles-sont-les-consequences-du-tabagisme-sur-la-sante

- *Ancienneté* : Comme discuté dans la partie I.3.3, l'un des enjeux de cette analyse est de déterminer l'influence de la sélection médicale. Non sommes donc satisfaits de voir cette variable parmi les variables significatives. Nous regrouperons certaines modalités afin que toutes les modalités soient significatives.

- *Csp* : Comme discuté plus haut, nous conservons cette variable malgré le peu de significativité détectée. Il semble naturel de considérer que les conditions de travail aient une influence sur la durée de vie.

Regroupement de modalités :

Maintenant que nous avons sélectionné les variables du modèle, il nous faut réarranger certaines d'entre elles afin d'obtenir une significativité conséquente pour chaque modalité. Ce réarrangement doit aussi donner plus de sens à notre analyse car l'impact de la sélection médicale ne peut conduire à 20 coefficients différents pour les vingt premières années.

Nous commençons par la variable *Csp*. Comme nous l'avons vu dans la partie II.2.4, la base de données ne comporte qu'une très petite quantité d'assurés en *Csp 4*. La p-value obtenue pour cette modalité est de 0,98, ce que nous ne pouvons pas accepter dans notre modèle, nous avons donc décidé de retirer ces quelques lignes du jeu de données.

La variable *Ancienneté* est plus délicate à manier. L'objectif est de faire un regroupement de modalités qui ait du sens d'un point de vu statistique mais également d'un point de vu souscription. On imagine que l'ancienneté joue un rôle les premières années car les personnes malades ont été écartées suite à la sélection médicale. Après quelques années d'ancienneté, l'assuré a pu développer une maladie grave et sa mortalité revient à un niveau standard.

Notre première approche a été de créer 3 classes d'ancienneté de 3 ans et une classe d'ancienneté de 10 ans et plus. Les résultats sont présentés ci-dessous :

Variable	Exp($\widehat{\beta}$)	p-value	Significativité
Civilité M	$\beta_M^{Civilité}$	$4{,}14.10^{-8}$	***
Fumeur OUI	β_{oui}^{Fumeur}	$< 2.10^{-16}$	***
Csp 2	β_2^{Csp}	0,14	
Csp 3	β_3^{Csp}	0,13	
Ancienneté 4-6	β_{4-6}^{Anc}	$3{,}25.10^{-7}$	***
Ancienneté 7-9	β_{7-9}^{Anc}	$5{,}12.10^{-10}$	***
Ancienneté 10&+	$\beta_{10\&+}^{Anc}$	$6{,}20.10^{-5}$	***

*** désigne une significativité à 0,1%, ** à 1%, * à 5%, . à 10%.

AIC :	11 618,07
Concordance :	0,641

Ces résultats confirment l'augmentation de la mortalité avec l'ancienneté. La variable *Csp* reste non-significative.

Cette première approche nous donne un aperçu de modèle avec un regroupement de modalités pour la variable *Ancienneté*. Comparons ces résultats avec un modèle où l'on regroupe les anciennetés par groupes de 2 :

Variable	Exp($\widehat{\beta}$)	p-value	Significativité
Civilité M	$\beta_M^{Civilité}$	$2{,}70.10^{-8}$	***
Fumeur OUI	β_{oui}^{Fumeur}	$< 2.10^{-16}$	***
Csp 2	β_2^{Csp}	0,17	
Csp 3	β_3^{Csp}	0,14	
Ancienneté 3-4	β_{3-4}^{Anc}	$1{,}36.10^{-10}$	***
Ancienneté 5-6	β_{5-6}^{Anc}	$4{,}82.10^{-8}$	***
Ancienneté 7-8	β_{7-8}^{Anc}	$1{,}84.10^{-11}$	***
Ancienneté 9-10	β_{9-10}^{Anc}	$3{,}75.10^{-6}$	***
Ancienneté 11&+	$\beta_{11\&+}^{Anc}$	$4{,}63.10^{-4}$	***

*** désigne une significativité à 0,1%, ** à 1%, * à 5%, . à 10%.

AIC :	11 597,45
Concordance :	0,653

Dans ce nouveau modèle, l'estimation des variables *Civilité*, *Fumeur* et *Csp* reste inchangée. L'AIC et la concordance se sont améliorées, on peut considérer que ce modèle est plus intéressant que le précédent. Malgré tout, ce groupement de la variable *Ancienneté* n'est pas très précis. L'augmentation de la mortalité avec l'ancienneté semble se faire par paliers : un premier entre 1 et 2 ans d'ancienneté, un deuxième entre 3 et 6 ans et un dernier pour 7 ans et plus.

Les deux dernières approches sont un peu simplistes car on attend une forte différence entre les premières années d'ancienneté[15]. Dès lors, il semble que les premières années devraient être considérées seules. C'est pourquoi nous avons introduit le modèle dans lequel les anciennetés 1, 2 et 3 sont seules. On crée également un groupe 4-6 et un groupe 7 et plus car d'après le dernier modèle, les estimations semblent assez homogènes au sein de ces deux groupes. Les résultats sont les suivants :

Variable	Exp($\widehat{\beta}$)	p-value	Significativité
Civilité M	$\beta_M^{Civilité}$	$6{,}76.10^{-8}$	***
Fumeur OUI	β_{oui}^{Fumeur}	$< 2.10^{-16}$	***
Csp 2	β_2^{Csp}	0,17	
Csp 3	β_3^{Csp}	0,13	
Ancienneté 2	β_2^{Anc}	$2{,}32.10^{-5}$	***

[15] Voir par exemple S. Sanchez d'Hondt (2012)

Ancienneté 3	β_3^{Anc}	$3{,}47.10^{-10}$	***
Ancienneté 4-6	β_{4-6}^{Anc}	$6{,}11.10^{-13}$	***
Ancienneté 7&+	$\beta_{7\&+}^{Anc}$	$< 2.10^{-16}$	***

*** désigne une significativité à 0,1%, ** à 1%, * à 5%, . à 10%.

AIC :	11 576,03
Concordance :	0,663

L'AIC et la concordance sont nettement meilleures. On voit que l'on a bien fait de distinguer la première année d'ancienneté de la deuxième car leurs coefficients sont très différents. Il semble ici que l'ancienneté 3 peut intégrer le groupe « Ancienneté 4-6 » car les coefficients β_3^{Anc} et β_{4-6}^{Anc} sont très proches. Le nouveau modèle est présenté ci-dessous :

Variable	Exp($\widehat{\beta}$)	p-value	Significativité
Civilité M	$\beta_M^{Civilité}$	$6{,}76.10^{-8}$	***
Fumeur OUI	β_{oui}^{Fumeur}	$< 2.10^{-16}$	***
Csp 2	β_2^{Csp}	0,17	
Csp 3	β_3^{Csp}	0,13	
Ancienneté 2	β_2^{Anc}	$2{,}32.10^{-5}$	***
Ancienneté 3-6	β_{3-6}^{Anc}	$3{,}84.10^{-13}$	***
Ancienneté 7&+	$\beta_{7\&+}^{Anc}$	$< 2.10^{-16}$	***

*** désigne une significativité à 0,1%, ** à 1%, * à 5%, . à 10%.

AIC :	11 574,47
Concordance :	0,664

La concordance et l'AIC s'améliorent très sensiblement suite à ce dernier regroupement. Le travail sur la variable *Ancienneté* est considéré comme terminé.

Ici on a tenté de faire un croisement entre les variables *Fumeur* et *Csp* afin d'augmenter la significativité de la variable *Csp* mais cet essai est resté infructueux.

Nous introduisons maintenant une nouvelle modalité à la variable *Csp* : la *Csp 0*. La différence de niveau de vie entre les catégories socio-professionnelles dont nous disposons n'est pas suffisante pour avoir un impact significatif sur la mortalité. Il est possible que la définition des catégories socio-professionnelles ne soit pas assez efficace pour représenter le niveau de vie des assurés. Afin d'affiner le concept de catégorie socio-professionnelle, les personnes ayant souscrit à un prêt de plus de 500 000 € sont classés dans une nouvelle catégorie appelée *Csp 0*. Le modèle construit avec cette nouvelle variable donne les résultats suivants.

Variable	Exp($\widehat{\beta}$)	p-value	Significativité
Civilité M	$\beta_M^{Civilité}$	$6{,}76.10^{-8}$	***
Fumeur OUI	β_{oui}^{Fumeur}	$< 2.10^{-16}$	***

Csp 1	β_1^{Csp}	0,0055	**
Csp 2	β_2^{Csp}	0,0015	**
Csp 3	β_3^{Csp}	0,0015	**
Ancienneté 2	β_2^{Anc}	$2{,}32.10^{-5}$	***
Ancienneté 3-6	β_{3-6}^{Anc}	$3{,}84.10^{-13}$	***
Ancienneté 7&+	$\beta_{7\&+}^{Anc}$	$< 2.10^{-16}$	***

*** désigne une significativité à 0,1%, ** à 1%, * à 5%, . à 10%.

AIC :	11 565,62
Concordance :	0,67

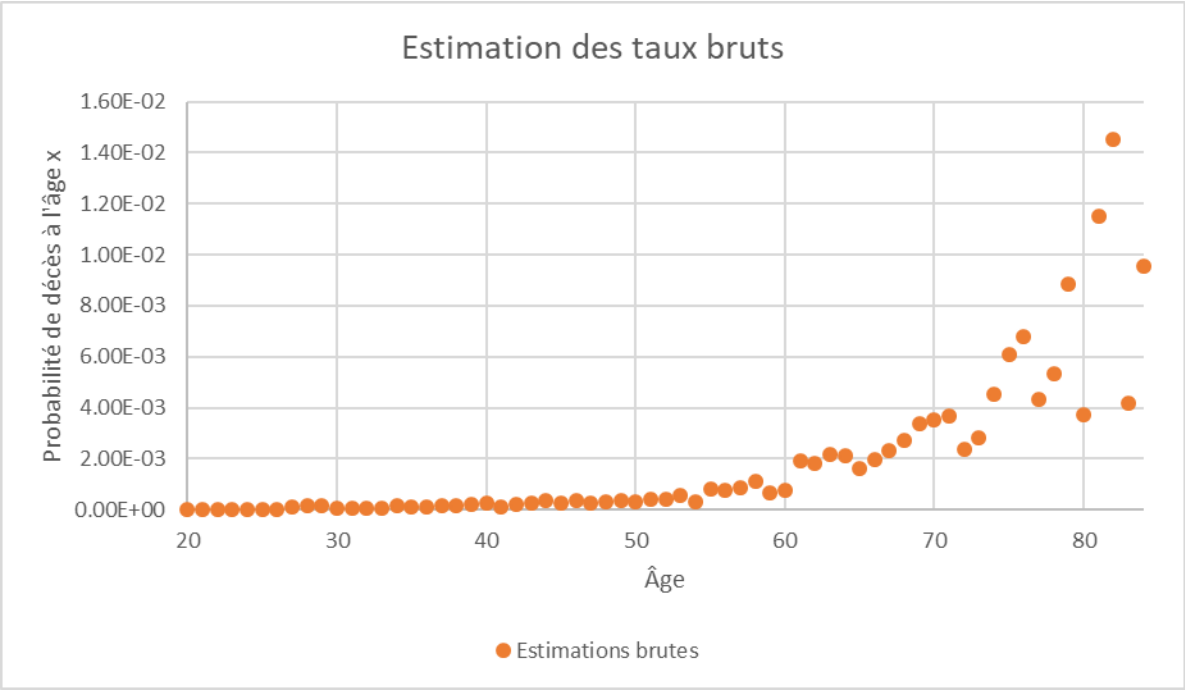
La catégorie des assurés ayant emprunté une somme importante a donc une mortalité significativement plus faible que le reste des assurés. On peut très bien utiliser cette distinction à des fins de tarification. La différence entre les Csp 1, 2 et 3 reste assez faible car les coefficients de ces modalités ont un intervalle de confiance à 95% de plus ou moins 20%. L'AIC et la concordance sont très bons.

Nous avons ensuite tenté de mettre dans cette catégorie *Csp 0* que les assurés de *Csp 1* avec un capital initial important. Les résultats présentés ci-dessous sont légèrement moins bons. Nous conserverons donc le modèle précédent.

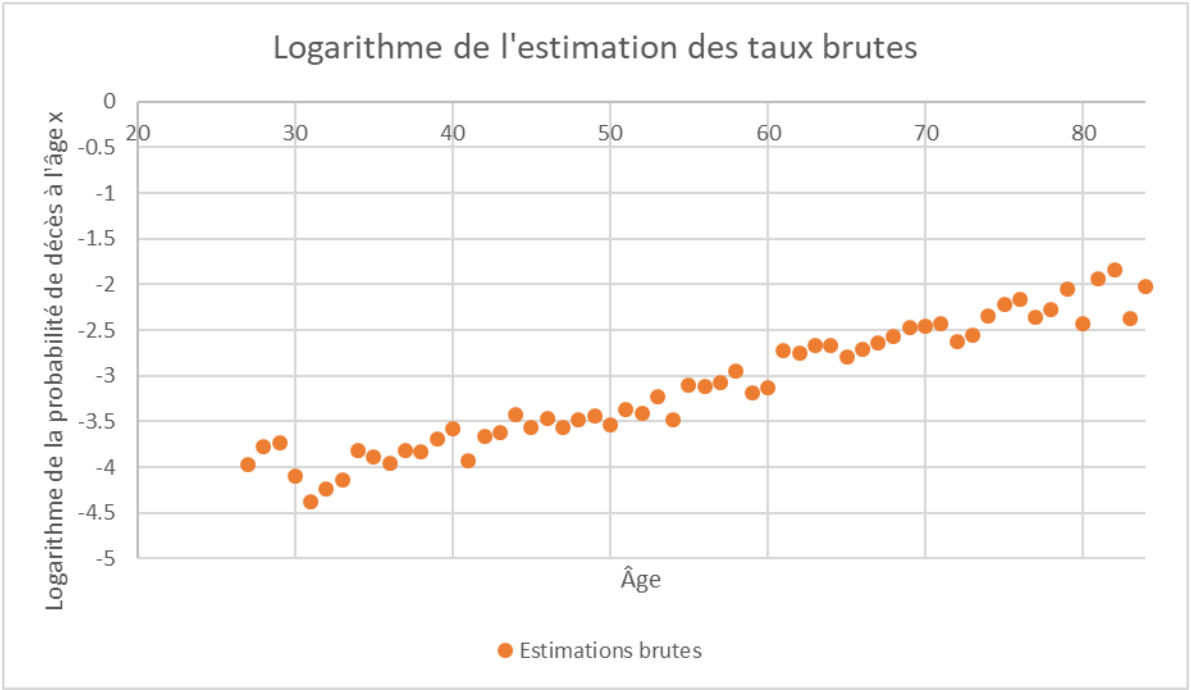
AIC :	11 571,1
Concordance :	0,67

III.3.2 Estimation des $\hat{q}_x$

Le modèle a été sélectionné et le vecteur des coefficients $\hat{\beta}$ a été estimé. Nous appliquons donc maintenant l'estimateur de Breslow pour obtenir une estimation brute $\tilde{q}_x$ des taux de mortalité de référence par âge, qui correspondent ici à la population de référence, c'est-à-dire les femmes non-fumeurs de csp 0 et d'ancienneté 1. Les résultats sont présentés ci-dessous :



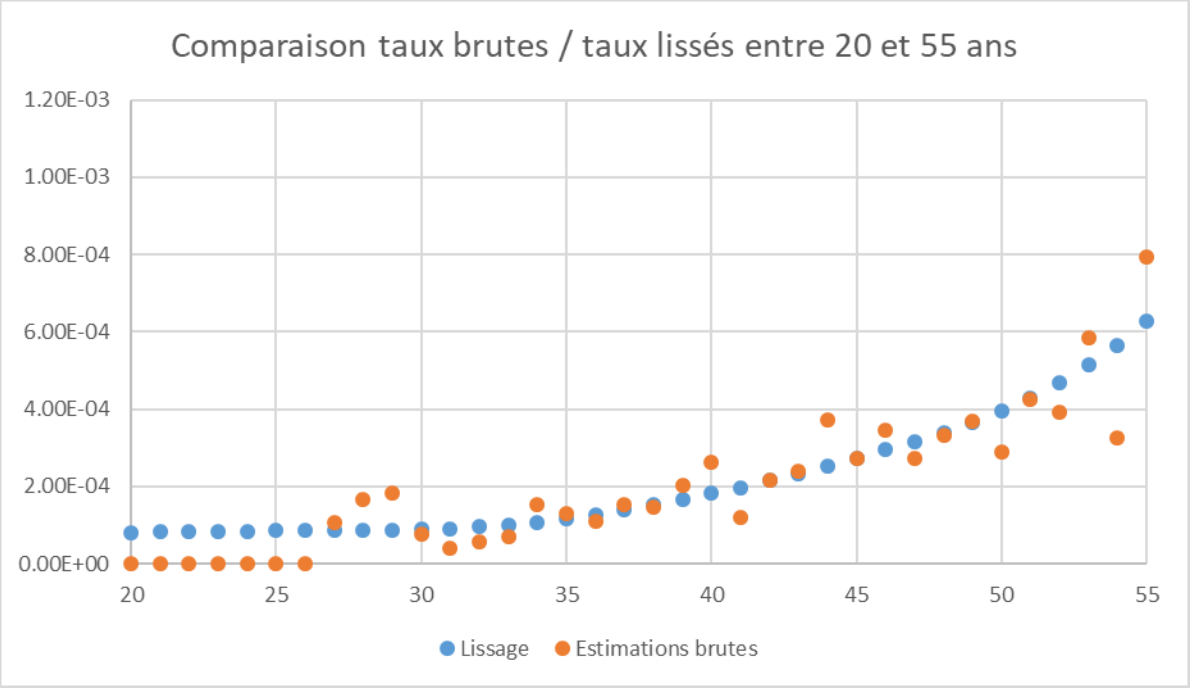
Ces taux de mortalité bruts semblent satisfaisants car on devine une forme exponentielle, en particulier à partir de 50 ans. Cette forme est plus difficile à apprécier avec cette représentation pour les âges inférieurs à 50 ans. En représentant les taux bruts sur une échelle logarithmique, nous constatons que les $\tilde{q}_x$ ont bien une forme exponentielle.

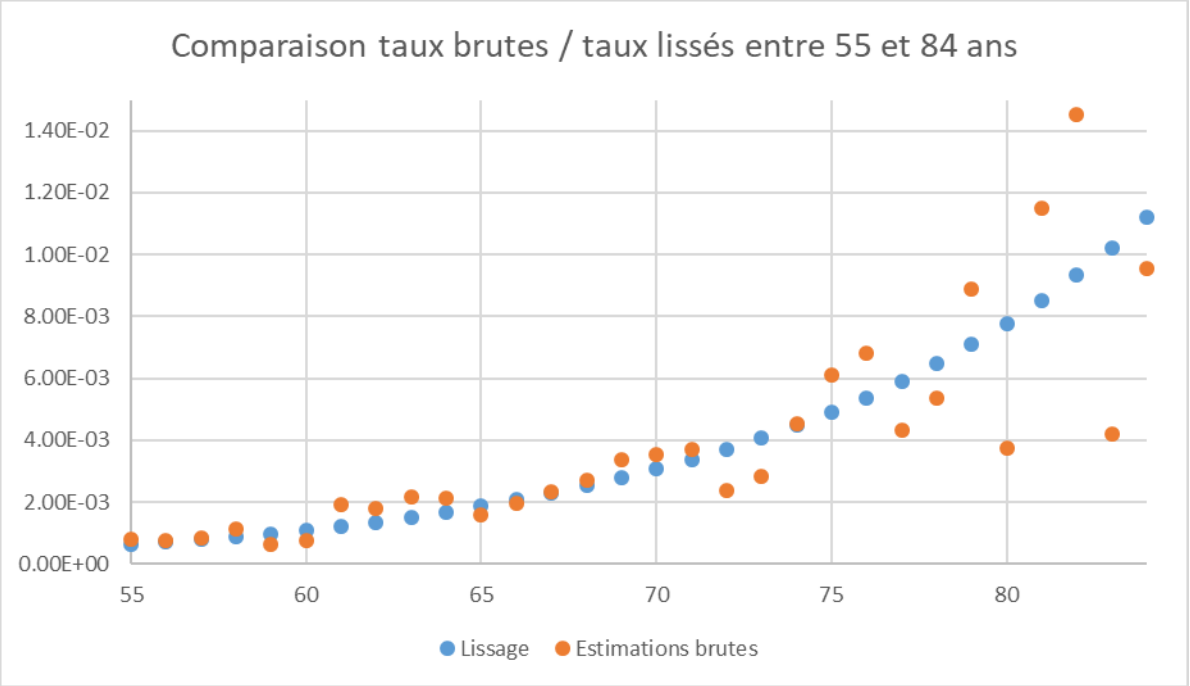


Si la forme générale de ces estimations est très satisfaisante, il n'en reste pas moins qu'avec ces taux bruts, la probabilité de décès à l'âge x n'est pas croissante avec l'âge. Or il semble raisonnable de penser que les taux de mortalité q_x sont croissants et réguliers. Il est normal d'observer des irrégularités dans la répartition des taux bruts mais ces taux ne peuvent pas servir tels quels à la tarification de contrats d'assurance pour des raisons de cohérence vis à vis du

client. Nous décidons donc d'appliquer la méthode de lissage de Whittaker et Henderson. Les paramètres h et z, introduits dans la partie théorique, ont été testés pour plusieurs valeurs mais les résultats étaient graphiquement assez proches. Nous avons finalement choisi le couple $(h, z) = (5, 2)$ car il permet d'obtenir le *SMR* le plus proche de 100%.

Les taux lissés $\hat{q}_x$ sont représentés ci-dessous et comparés aux taux bruts. Afin de pouvoir bien visualiser l'ajustement des taux lissés aux âges les plus petits, nous avons découpé la représentation graphique en deux. Le premier graphique correspond aux âges inférieurs à 55 ans et le second aux âges supérieurs à 55 ans.





Visuellement, ce lissage semble très satisfaisant. Ce sont ces taux lissés $\hat{q}_x$ qui seront validés dans la partie suivante.

III.3.3 Validation du modèle

Une fois le modèle sélectionné, il convient de vérifier la pertinence des coefficients et de la fonction de hasard de référence estimés. La validation du modèle se fait en deux étapes :

- La validation des hypothèses du modèle de Cox
- L'adéquation des $\hat{q}_x$ estimés avec les valeurs d'origines

Un modèle qui se distingue des autres lors de la phase de sélection peut très bien échouer lors de cette phase de validation et être rejeté.

III.3.3.1 Validation des hypothèses du modèle de Cox

La première hypothèse à vérifier est la non-nullité du vecteur β. Pour cela, il suffit d'effectuer le test du score qui considère l'hypothèse nulle H_0 : « $\beta_1 = 0\ et\ \beta_2 = 0\ et\ \dots et\ \beta_p = 0$ » pour un modèle à p coefficients.

Modèle	p-value
Modèle Cox	$< 2.10^{-16}$

La p-value étant extrêmement faible, on rejette donc l'hypothèse nulle.

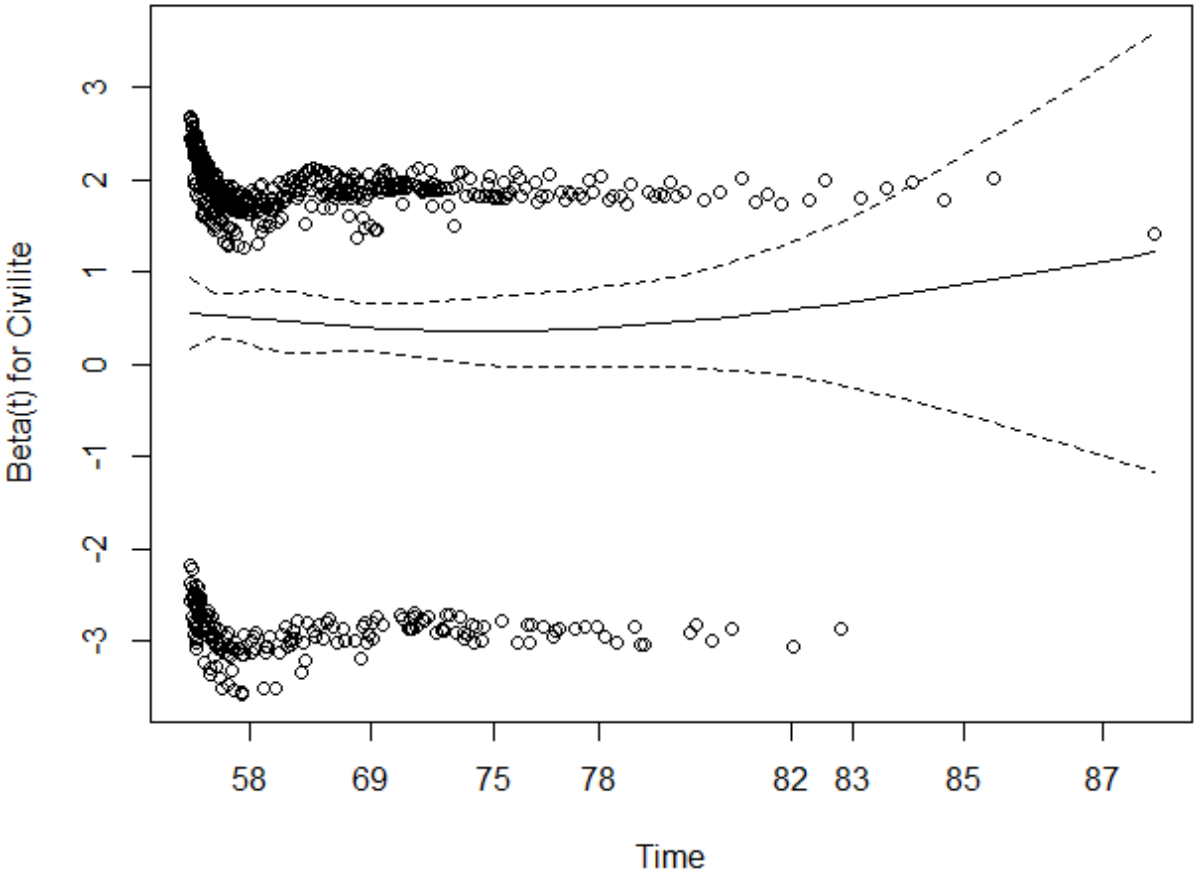
La deuxième hypothèse à vérifier est l'hypothèse des hasards proportionnels. En effet, le modèle de Cox est fondé sur la constance des coefficients avec le temps. Autrement dit, on fait l'hypothèse $H_0: \beta_i(t) = \beta_i$ pour tout $i \in [\![1, p]\!]$. Pour la vérifier, nous utilisons un test développé par Grambsch et Therneau (1994) basé sur les résidus de Schoenfeld. Les résultats de ce test sont repris ci-dessous :

Variable	Khi 2	Degrés de liberté	p-value
Civilité	0,0941	1	0,76
Fumeur	2,3197	1	0,13
CSP	0,2951	1	0,59
Ancienneté	4,4668	3	0,22
Global	7,1717	6	0,31

On utilise un niveau de confiance à 95% et l'on constate que toutes les variables, ainsi que le modèle dans sa globalité, passent le test de Grambsch et Therneau.

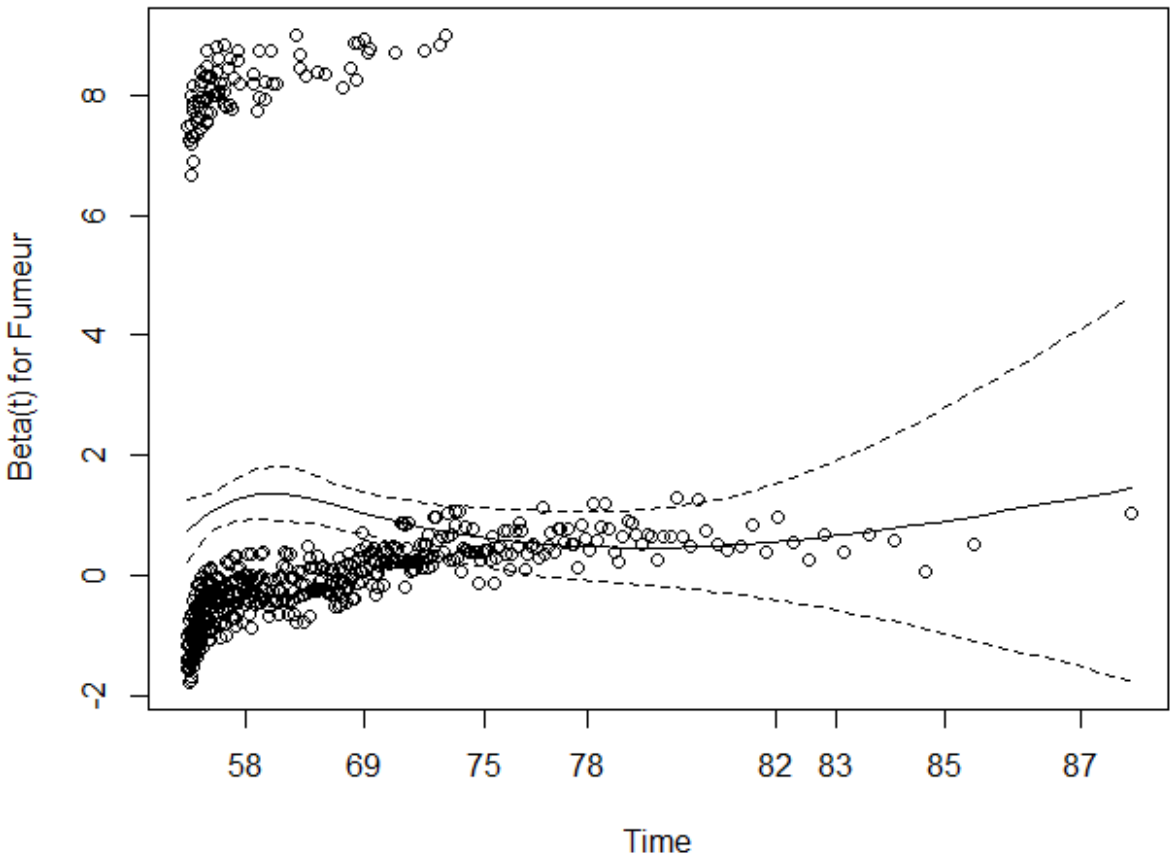
Il est toutefois conseillé de poursuivre l'investigation en faisant apparaitre les résidus de Schoenfeld afin de vérifier l'hypothèse visuellement. Il faut alors vérifier qu'il n'y a pas de tendance dans la représentation des résidus au cours du temps.

Commençons par la variable *Civilité* :



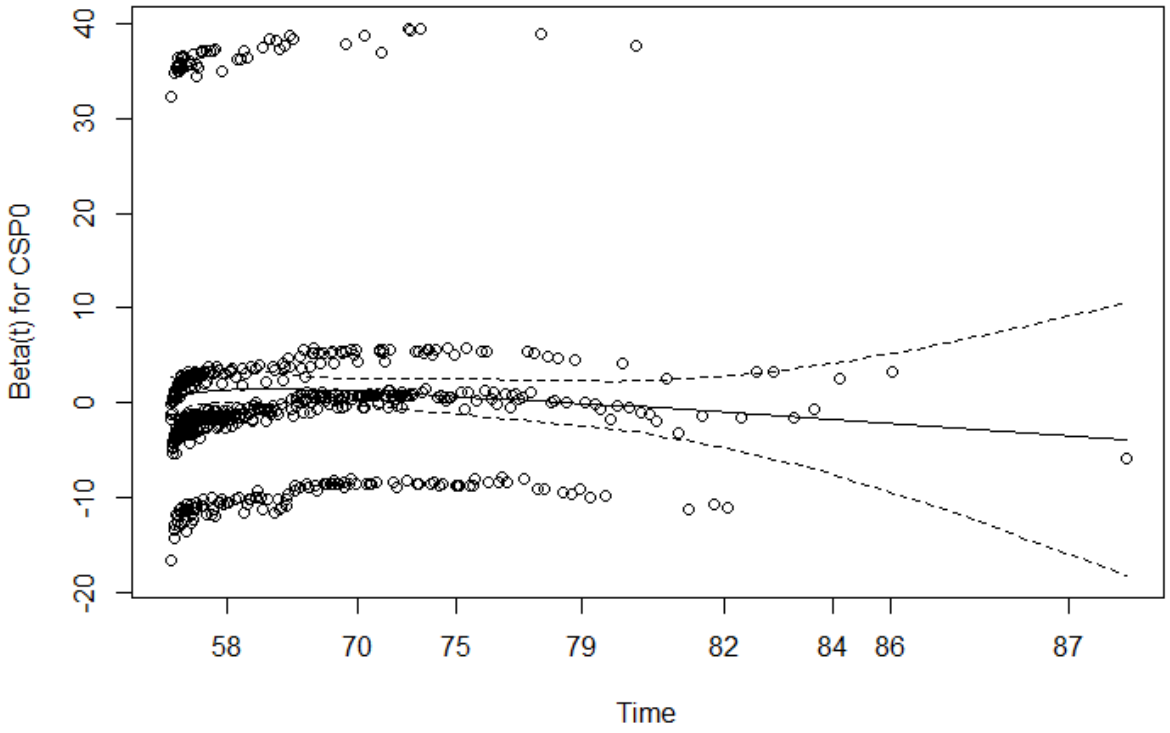
On observe une légère vaguelette pour les âges les plus jeunes mais dans l'ensemble les résidus ne semblent pas avoir de réelle tendance dans le temps. Cela confirme la grande p-value observée dans le test ci-dessus.

Considérons maintenant la variable *Fumeur* :



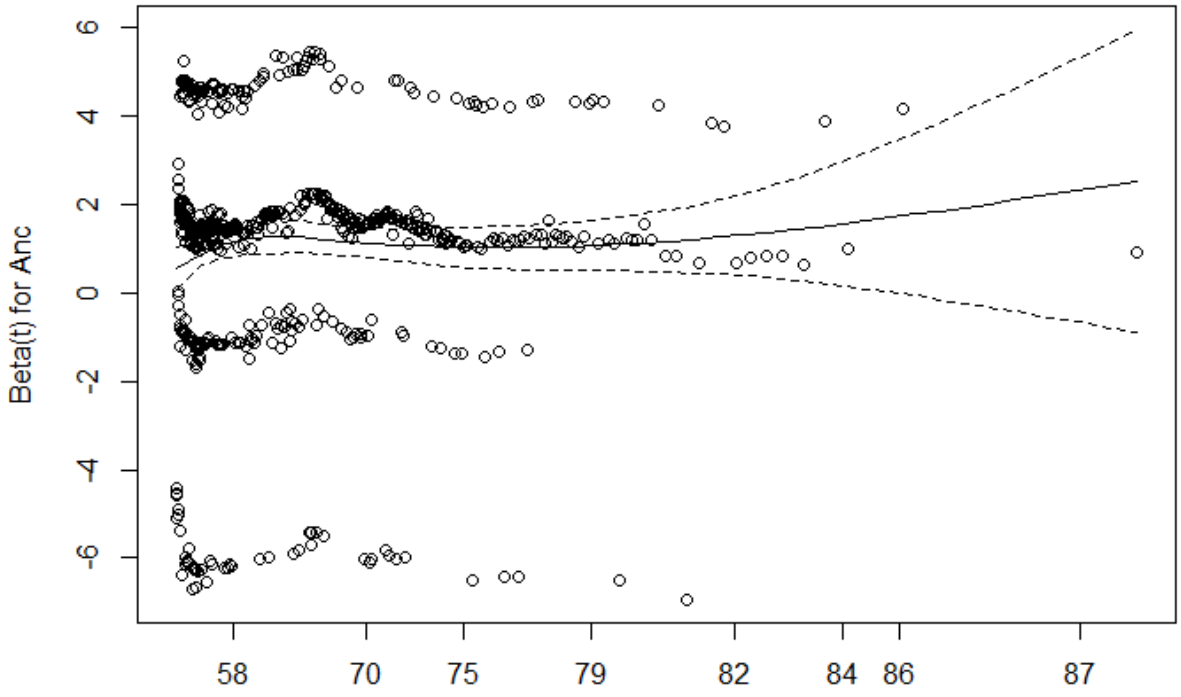
Ici l'on pourrait peut-être détecter une très légère tendance à la hausse jusqu'à l'âge de 75 ans, après quoi, les résidus semblent se stabiliser. En effet, la p-value de la variable *Fumeur* était la plus faible et des avis d'experts médicaux en interne nous indique que le risque de tomber malade des suites du tabagisme est plus important pour une personne qui fume depuis longtemps que pour une personne jeune qui adopte cette pratique depuis moins de temps. Toutefois cet effet n'est pas confirmé par le test de Grambsch et Therneau et la tendance observée sur les résidus n'est pas flagrante. On peut donc tout à fait maintenir l'hypothèse des hasards proportionnels pour cette variable.

Passons à la variable *CSP* :



Les résidus sont relativement stables dans le temps.

Pour la variable *Ancienneté*, les résidus de Schoenfeld sont les suivants :



Là encore, les résidus semblent assez stables avec le temps donc ces observations confirment le test de Grambsch et Therneau.

III.3.3.2 Validation de l'ajustement des $\hat{q}_x$

Une autre approche pour valider nos estimations est de comparer les résultats obtenus avec le modèle avec les données d'origine. Il est assez naturel de vérifier que les $\hat{q}_x$ obtenus décrivent bien les données d'origine. Nous disposons, à cet effet, de plusieurs tests qui ont été décrits dans la partie II.1.5. Les résultats sont présentés dans le tableau ci-dessous :

Test	Résultat
Test du *SMR*	Validé
Test du χ^2 – Civilité	Validé
Test du χ^2 – Fumeur	Validé
Test du χ^2 – CSP	Validé
Test du χ^2 – Ancienneté	Validé
Test des signes	Validé
Test des *runs*	Validé
Test de Kolmogorov	Validé

Tous les tests semblent confirmer la validité de l'ajustement effectué. Regardons chacun d'entre eux plus en détail.

Le test du *SMR* :

Le *SMR* global donne une première idée de la précision de notre ajustement.

SMR	p-value
103%	0,22

Nous observons un *SMR* de 103% donc sous-estimons légèrement les taux de décès. Cependant, la *p-value* du test est de plus de 20% donc on peut considérer ce *SMR* comme satisfaisant, c'est-à-dire suffisamment proche de 100%.

Pour avoir une idée plus détaillée du ratio *observés sur attendus*, il est possible d'utiliser un *SMR* par âge. Nous avons tenté cette approche mais la faible quantité de données a été un peu gênante. En effet, les *SMR* par âge sont parfois très éloignés de 100% et les intervalles de confiance peuvent être très larges à cause du manque de données. Nous avons donc jugé cette approche peu satisfaisante.

En regroupant les individus en trois groupes d'âge (18-39 ans, 40-59 ans et 60-84 ans), on construit des *SMR* par groupe d'âge. Les données disponibles sont alors plus importantes et on obtient les résultats suivants :

Classe d'âges	*SMR*	p-value
18-39 ans	103%	0,37
40-59 ans	101%	0,40
60-84 ans	105%	0,20

Les tests des signes et des *runs* donneront une appréciation globale des *SMR* par âge.

Le test du χ^2 :

Comme expliqué dans la partie II.1.5.2, le test du *SMR* doit être complété par le test du χ^2 afin de vérifier l'homogénéité des estimations au sein de chaque variable explicative. L'application du test du χ^2 valide nos résultats pour chaque variable explicative. Pour donner au lecteur un aperçu de cette homogénéité, nous présentons les *SMR* pour chaque modalité.

Pour la variable *Civilité* :

Modalité	*SMR*
Femme	103%
Homme	103%

Pour la variable *Fumeur* :

Modalité	*SMR*
Fumeur	104%
Non-Fumeur	103%

Pour la variable *Csp* :

Modalité	*SMR*
CSP 0	104%
CSP 1	103%
CSP 2	103%
CSP 3	103%

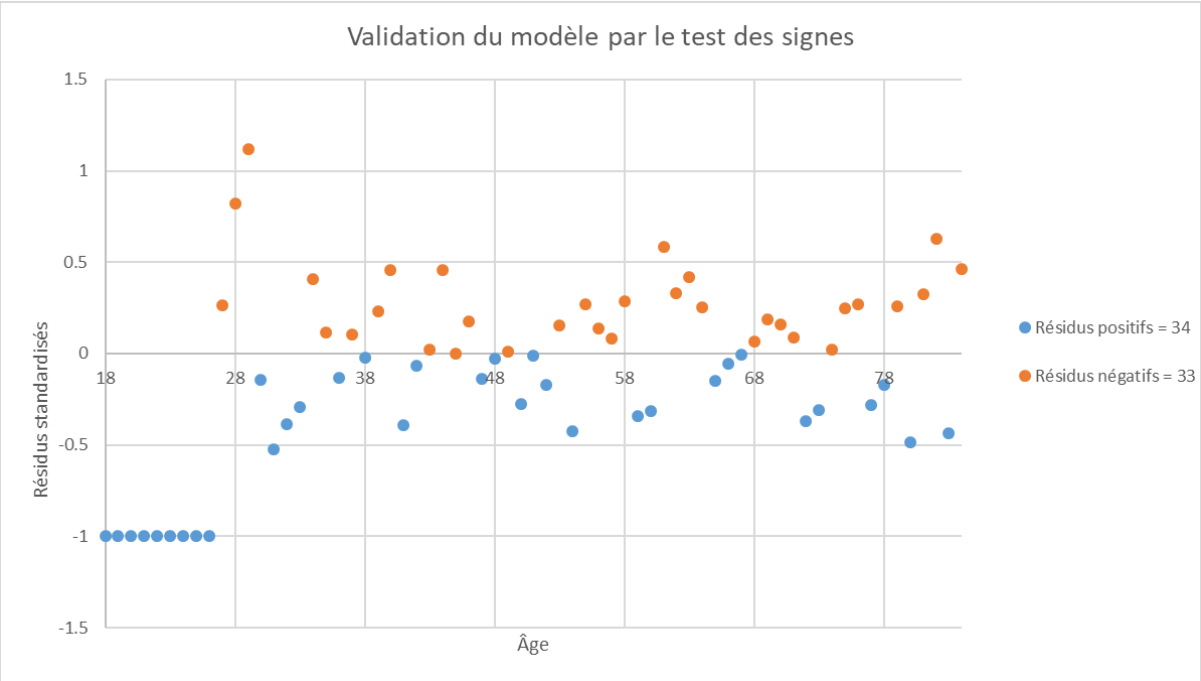
Pour la variable *Ancienneté* :

Modalité	*SMR*
Ancienneté 1	102%
Ancienneté 2	103%
Ancienneté 3-6	104%
Ancienneté 7&+	103%

Pour chaque modalité, le *SMR* est très proche de 103%. On comprend donc pourquoi le test du χ^2 est vérifié. On conclut de cette analyse que la modélisation de chaque sous-modalité est très satisfaisante.

Le test des signes :

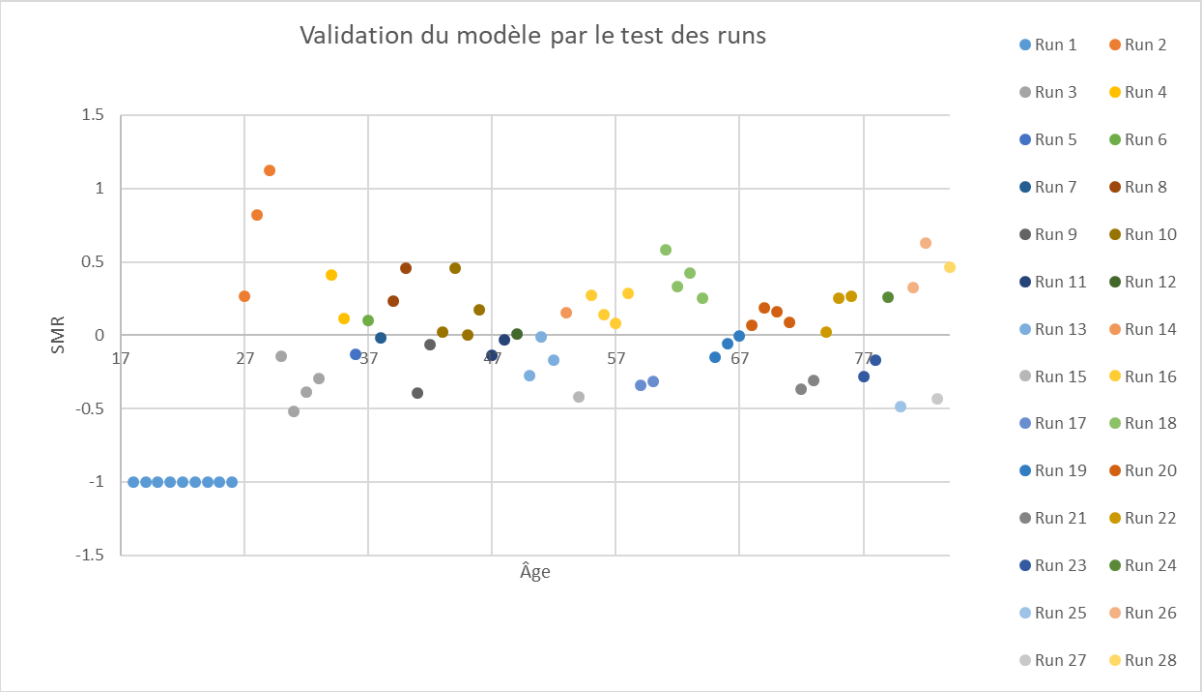
Pour vérifier que les observations par âge sont bien réparties aléatoirement par rapport aux estimations, on met en place le test des signes.



Comme on le voit sur ce nuage de points, les résidus du modèle se répartissent assez équitablement autour de 0. Avec une différence entre le nombre de résidus positifs et le nombre de résidus négatifs de 1 pour 67 résidus, le test des signes est validé. On peut toutefois se demander s'il est bien judicieux de considérer dans ce calcul les premiers résidus car on n'observe aucun décès pour les premiers âges. Même dans le cas où l'on ne considère que les années avec au moins un décès observé, le test des signes est validé. On peut donc considérer que le modèle passe ce test.

Le test des runs :

Comme nous l'avons vu dans la partie théorique, le test des signes doit être complété du test des *runs*. Une visualisation des différents *runs* est proposée ci-dessous :

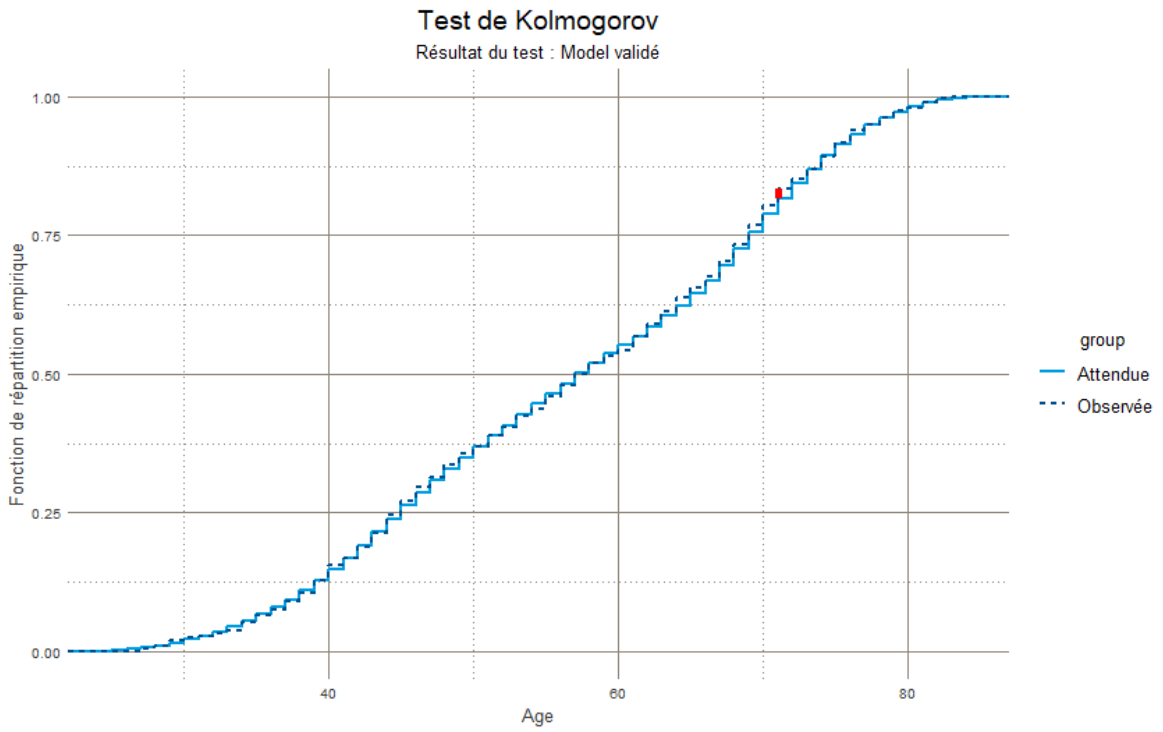


Visuellement, les résidus semblent être répartis harmonieusement autour de 0. On observe 28 *runs* sur l'ensemble des résidus entre 18 et 84 ans. La valeur de la statistique de ce test, décrite dans la partie théorique, est égale à -1,59. Elle est donc supérieure à -1,96 et ainsi on peut valider le test des *runs*.

Là encore, il est légitime de se demander si le test est toujours validé si l'on ne prend que les âges où l'on observe des décès. Dans ce cas la statistique est de -0,66 donc elle permet également de valider le test des *runs*.

Le test de Kolmogorov-Smirnov:

Ce test est, lui aussi, validé par le modèle. Afin d'observer visuellement son fonctionnement, nous avons représenté les fonctions de répartition empiriques *Observée* et *Estimée*, avec en rouge l'écart maximal entre ces deux courbes :



III.3.3.3 Conclusion sur la validation du modèle

Comme nous l'avons vu, les hypothèses du modèle de Cox sont validées, en particulier la contraignante hypothèse des hasards proportionnels. La vérification de l'ajustement qui dépend à la fois de la précision de l'estimation des paramètres, de l'estimation de la fonction de hasard de référence et du lissage, a montré au travers des cinq tests d'ajustements que les résultats du modèle étaient très satisfaisants. On retient donc ce modèle qui donne une probabilité de décès par âge dépendant de la civilité, du tabagisme, de la catégorie socio-professionnelle et de l'ancienneté.

III.3.4 Résultats

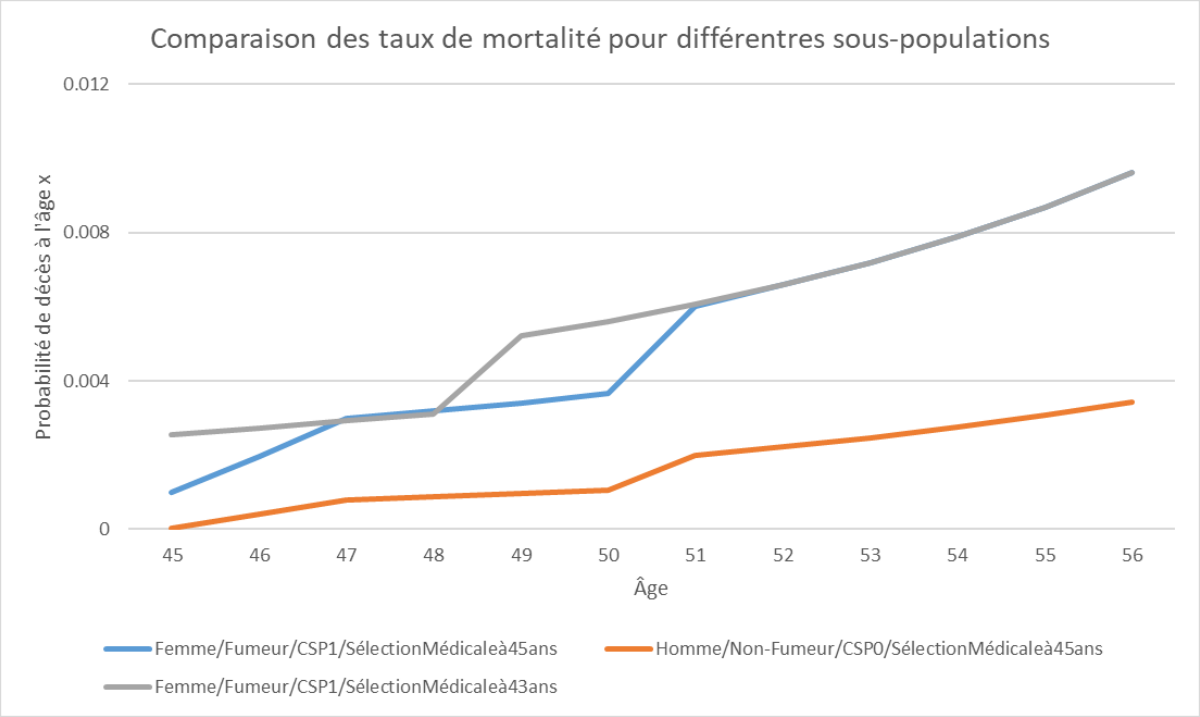
L'application du modèle de Cox a donc permis de déterminer les variables qui ont un rôle sur la mortalité des emprunteurs. L'âge, le genre, le niveau de vie et le tabagisme ont un impact sur la mortalité. Dans notre portefeuille, la catégorie socio-professionnelle n'a pas un impact très significatif sur la mortalité mais son utilisation est usuelle sur le marché et les avis d'expert indiquent qu'au contraire, cette variable a un effet sur la mortalité. Nous avons donc conservé cette variable importante. L'estimation des coefficients à appliquer aux variables explicatives semble confirmer les tendances déjà observées par le passé. Elle permet cette fois de les quantifier précisément.

L'analyse a aussi permis de démontrer et de quantifier l'impact de la sélection médicale. Cet impact est observé sur les six premières années qui suivent le test médical. La première année on observe une mortalité 4 fois moins importante que pour des assurés de plus de 7 ans d'ancienneté. Cet effet diminue de moitié dès la deuxième année suivant la sélection médicale et de 3 à 6 ans d'ancienneté, la sous-mortalité n'est plus que de 25%.

Le SMR global étant de 103%, nous sous-estimons la mortalité des emprunteurs de la base de données. Cet écart n'est pas significatif au sens du test du SMR car il est assez faible et nous ne nous attendons pas à obtenir exactement un SMR de 100%. Cependant, par application du principe de prudence, nous opérerons une majoration de 3% aux taux discutés ci-dessus dans le cadre des futures tarifications.

Pour illustrer les résultats obtenus, les taux lissés de plusieurs sous-populations ont été représentés ci-dessous. Nous suivons l'évolution sur 12 années des $\hat{q}_x$ de trois assurés :

- Une femme fumeur en *csp* 1 ayant passé la sélection médicale à l'âge de 45 ans
- Une homme non-fumeur en *csp* 0 ayant passé la sélection médicale à l'âge de 45 ans
- Une femme fumeur en *csp* 1 ayant passé la sélection médicale à l'âge de 43 ans



Cette représentation est assez instructive. Nous observons d'abord des sauts dus aux changements d'ancienneté. Les $\hat{q}_x$ modélisés ne sont pas tout à fait lisses. On voit apparaitre sur la courbe bleue les quatre niveaux d'ancienneté : deux coefficients distincts sont appliqués aux deux premières années, puis un coefficient intermédiaire entre les âges 47 et 50 ans et, enfin, une dernière majoration vient s'appliquer à partir de 51 et pour le reste de la vie de l'assurée. C'est la différence d'ancienneté qui explique l'écart entre la courbe bleue et la courbe grise car il s'agit de deux assurées de mêmes caractéristiques. A partir d'un certain temps, la sélection médicale n'a plus d'effet et les deux courbes se rejoignent. En comparant la courbe bleue et la courbe orange, nous pouvons constater l'importance du tabagisme et de la catégorie socio-économique sur la mortalité. En effet, on s'attend à ce qu'une femme ait des $\hat{q}_x$ inférieurs à ceux d'un homme mais ici, le fait que cette femme fume et qu'elle soit en *csp* 1 et non 0 comme l'homme, augmente sa probabilité de décès jusqu'à la faire dépasser largement celle de l'homme.

Les résultats donnés par ce modèle pourront servir à la tarification de nouveaux contrats emprunteur. Nous allons voir dans la partie suivante, comment passer de la prime pure qui découle directement de l'application de ces taux, à la prime commerciale.

IVTarification d'un portefeuille emprunteur

Dans cette partie, nous allons décrire le processus qui permet d'obtenir le tarif commercial d'un portefeuille à partir des taux calculés dans la partie III. Cela nous permettra de quantifier l'écart entre les prix actuellement proposés aux clients et ceux qui pourraient être proposés en utilisant la segmentation obtenue à l'aide du modèle de Cox. Ces résultats sont donc très importants en interne mais ils seront fortement modifiés dans ce mémoire car strictement confidentiels. Le lecteur pourra toutefois suivre la démarche permettant de tarifer la réassurance d'un portefeuille emprunteur.

Nous nous intéresserons au cas d'un portefeuille d'assurance emprunteur qu'une cédante voudrait céder au réassureur.

IV.1 Le portefeuille à tarifer

Un assureur français ayant lancé un produit emprunteur en 2013 souhaite céder son portefeuille afin de sécuriser son profit. Le portefeuille est en *run-off*, c'est-à-dire qu'il ne prend plus de nouveaux assurés. L'assureur a constaté de bons résultats lors des premières années mais ne sait pas exactement quels gains attendre dans le futur. La cession du portefeuille pourra lui permettre de s'assurer un profit avec ce produit et de libérer du capital pour lancer un nouveau produit par exemple.

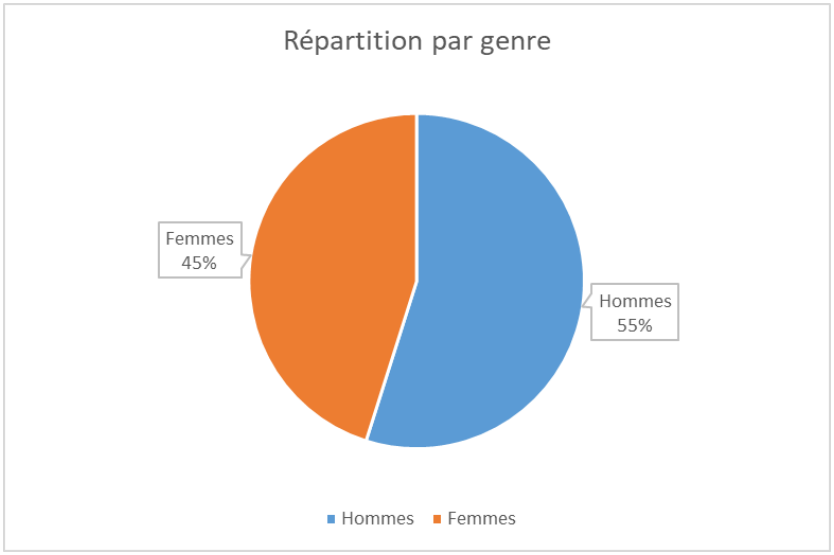
Il est donc envisagé de céder le portefeuille via un contrat en *quota-share* avec un taux de cession de 100%. Dans ce type de contrat, le réassureur prend en charge 100% des sinistres et reçoit 100% de la prime perçue par l'assureur, aussi appelé cédante. Toutefois, une partie de la prime cédée est finalement conservée par la cédante car si le risque est totalement transféré au réassureur, la gestion du portefeuille reste à la charge de l'assureur et celle-ci représente un coût certain. Une fois le contrat mis en place, l'assureur s'occupe des relations avec le client et perçoit la part de la prime correspondant aux frais de gestion. La variabilité due au risque a disparu. L'assureur peut alors libérer les provisions constituées pour assurer ce risque. On peut noter qu'il constitue tout de même une petite provision pour faire face au risque de contrepartie en cas de défaut du réassureur.

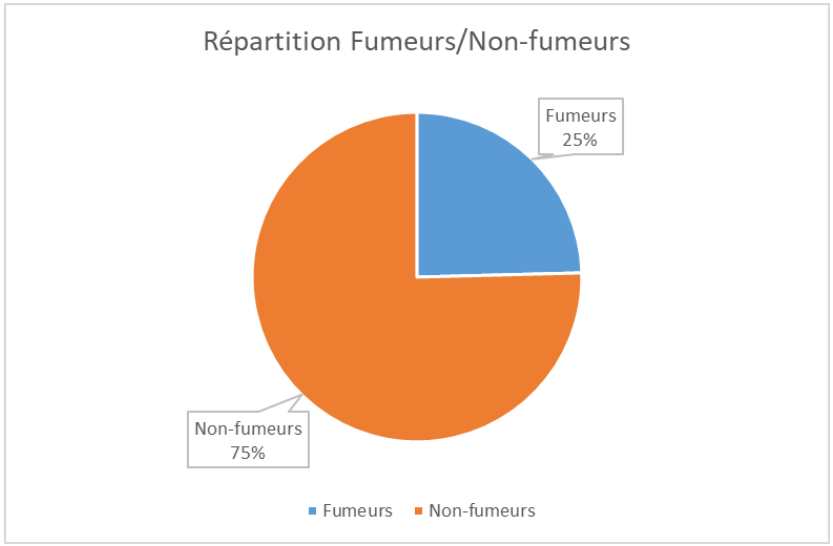
En tant que réassureur, nous allons donc recevoir les primes et devoir indemniser les sinistres. L'objectif de la tarification va donc être de déterminer la valeur des flux échangés avec la cédante. Il faudra aussi s'assurer que le prix permette bien de couvrir tous les frais du réassureur.

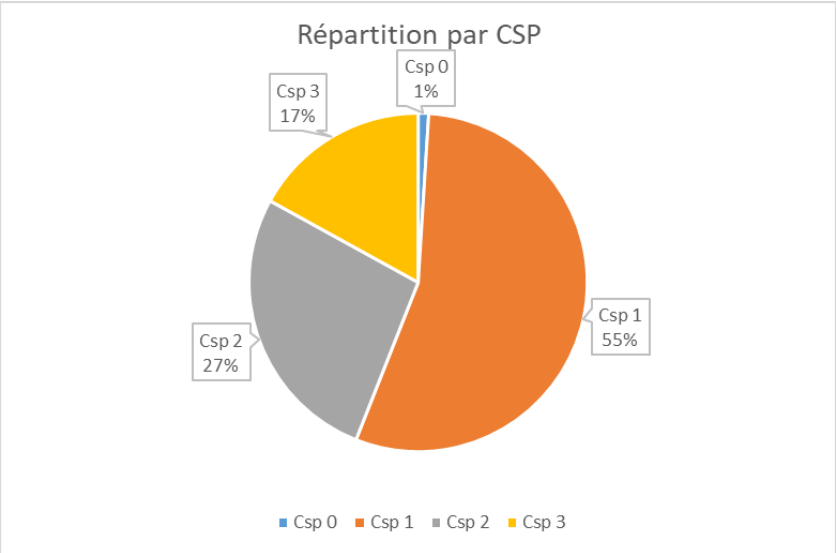
Nous disposons d'un bordereau de 70 311 polices comportant les principales informations nécessaires à la tarification. Les variables *civilité*, *fumeur* et *csp* y apparaissent et nous permettrons d'appliquer la segmentation proposée par le modèle que nous avons construit dans la partie III. Les autres informations importantes sont la date de naissance et la date de souscription afin de déterminer l'âge et l'ancienneté. Il faut aussi pouvoir connaitre à chaque instant le capital restant dû. C'est pourquoi les informations sur le capital emprunté, une

éventuelle quotité, la périodicité des remboursements, le taux d'intérêts et la durée du prêt sont également très importantes. Ici nous disposons de toutes ces informations.

Comme nous l'avons déjà écrit plus haut, le portefeuille comporte 70 311 polices. L'âge moyen est de 43 ans et le capital restant dû moyen de 115 000 €. Les différentes sous-populations sont réparties comme suit :







Le portefeuille comporte une importante proportion de fumeurs et un bon nombre de personnes de catégorie socio-professionnelle 2 ou 3. Nous avions remarqué que ces populations étaient particulièrement à risque.

IV.2 Méthode de tarification

HannoverRe est une entreprise privée qui a des coûts et qui se doit de respecter la réglementation européenne afin de pouvoir effectuer ses opérations de réassurance en toute légalité. La tarification doit donc répondre à des enjeux réglementaires et de rentabilité.

La réglementation Solvabilité II impose aux assureurs comme aux réassureurs de pouvoir être capable de répondre à leurs engagements vis-à-vis des assurés ou des cédantes et ce, même lors de la survenance d'un risque majeur. Ainsi il est demandé aux assureurs et réassureurs de mesurer leurs risques et de posséder un niveau de fonds propres suffisant pour faire face à 99,5% des cas de figures possibles. Ce montant de capital cible est appelé *SCR* (*Solvency Capital Requirement*). Le montant de capital possédé par une entité est souvent rapporté au SCR pour définir le ratio de couverture de la compagnie. Il doit alors être supérieur à 100% selon la réglementation, mais les compagnies mènent souvent une politique de gestion des risques plus sélective pour assurer un ratio de couverture plus élevé. Or dans la pratique, le montant des primes reçues par l'assureur, hors taxe et hors frais de gestion, ne permet pas à lui seul d'apporter le capital nécessaire pour assurer une solvabilité dans 99,5% des cas de figure. L'immobilisation du capital d'HannoverRe est donc nécessaire. Or dans une société par actions comme HannoverRe, le capital appartient aux actionnaires. Engager des fonds représente un risque et immobilise du capital, c'est pourquoi il est nécessaire qu'une partie de la prime serve à rémunérer les investisseurs.

Il faut également que la tarification couvre les frais d'HannoverRe.

Pour prendre en compte tous ces facteurs, on utilise la méthode suivante. On commence par calculer la valeur actuelle des profits futurs ou *PVFP (Present Value of Futur Profits)*, puis on calcule le montant de provisions à allouer à ce risque. La différence entre ces deux valeurs est

le montant de capital à apporter. C'est sur cette base que l'on calcule le coût du capital qui viendra s'ajouter aux frais de gestion.

IV.2.1 Calcul de *PVFP*

Pour calculer la *PVFP* du portefeuille considéré, on projette les flux futurs à l'aide des taux de mortalité obtenus avec le modèle de Cox. Il est également nécessaire de considérer le taux de résiliations car lorsqu'un assuré sort du contrat, l'individu n'est plus sous risque et l'assureur ne perçoit plus de primes. Nous avons utilisé une hypothèse interne pour modéliser les résiliations. Les taux utilisés dépendent de l'ancienneté du contrat.

Soit Π_i la prime annuelle reçue par l'assureur pour l'assuré i. Même lorsque le contrat en *quota-share* est à 100%, le réassureur ne conservera pas 100% de la prime commerciale perçue par l'assureur car malgré la cession du risque, celui-ci conserve des frais liés à la gestion du contrat. Ainsi le réassureur redonne une partie de la prime de réassurance à l'assureur pour couvrir ces frais de gestion via le mécanisme de la commission de réassurance. Notons $\pi_i = (1-\alpha)\Pi_i$ la part de la prime qui revient au réassureur, avec α le taux correspondant aux frais de la cédante.

On commence par projeter le capital restant dû de chaque assuré, noté CRD_i^t pour l'assuré i en année t. On détermine ensuite pour chaque année la probabilité que cet assuré soit encore couvert. Les sorties peuvent être dues à un décès ou à un rachat, ainsi en notant p_i^t la probabilité de survie de l'assuré i en t et avec $p_i^0 = 1$, on peut calculer p_i^t pour tout t avec la formule suivante :

$$p_i^{t+1} = p_i^t . (1 - q_i^t)(1 - r_i^t)$$

Avec q_i^t la probabilité de décès de l'assuré i entre t et $t+1$ et r_i^t la probabilité de résiliation de l'assuré i entre t et $t+1$.

Chaque année, on encaisse donc

$$\pi^t = \sum_{i=1}^{n} p_i^t . \pi_i^t$$

Le capital restant dû est plus faible en fin d'année qu'en début d'année. Pour modéliser les décès, on considère qu'ils ont lieu en milieu d'année et avant les résiliations. Ainsi avec des notations analogues :

$$Sinistres_t = \sum_{i=1}^{n} p_i^t . q_i^t . CRD_i^{t-1/2}$$

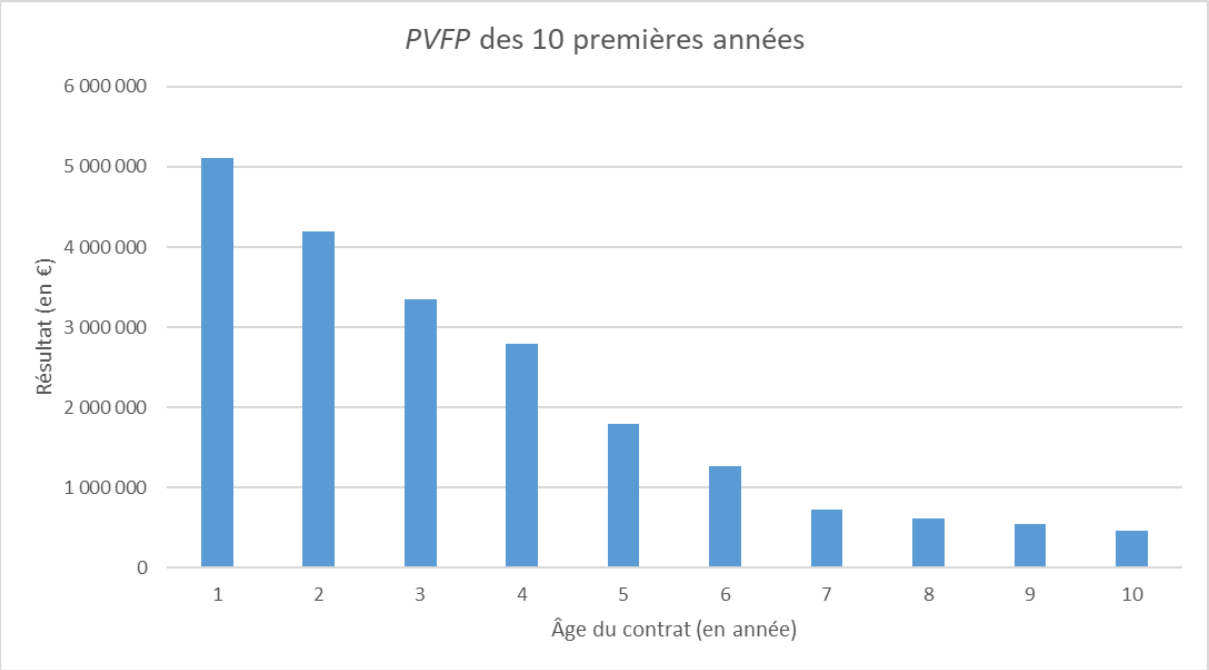
Le résultat de l'année t est égal à la différence entre les primes perçues et les sinistres. Ce résultat est imposé à hauteur de I_t et chaque année, la gestion du contrat coûtera ρ au réassureur. On obtient donc un résultat net d'impôts comme suit.

$$RésultatNet_t = (\pi^t - Sinistres_t) . (1 - I_t) - \rho$$

On actualise ces résultats avec le facteur d'actualisation d_t pour obtenir la *PVFP* :

$$PVFP_t = d_t . RésultatNet_t$$

Les *PVFP* sont représentées ci-dessous. Le plus long prêt dure 32 ans mais dans un souci de clarté nous avons sélectionné les 10 premières années seulement. Les dernières valeurs sont proches de 0.



Les résultats sont très largement positifs. On s'attend donc à ce que ce portefeuille génère des revenus importants dans les années futures et jusqu'à expiration de toutes les polices.

PVFP (Best estimate)	23 492 667 €

IV.2.2 Coût du capital

Si le réassureur récupère le portefeuille pour 23 492 667 €, il doit immobiliser du capital afin d'honorer ses engagements même si les résultats futurs subissent un choc à la baisse. Ce capital immobilisé est soumis à un risque et doit donc être rémunéré. Ce coût sera ensuite intégré au prix.

On commence par calculer le montant qui devra être immobilisé en face de ce risque. Ce montant dépend du modèle interne de l'entreprise et son calcul ne pourra pas être détaillé ici. Le souscripteur utilise une version simplifiée spécialement pour la tarification. Le capital ainsi calculé dépend de chocs à effectuer sur les hypothèses de calcul ainsi que du type de *business* signé. En effet la diversification est un enjeu clé en réassurance et la politique de souscription en dépend beaucoup.

On estime le capital immobilisé chaque année puis l'on calcule la valeur actuelle de ces montants. Dans notre cas le capital utilisé est de 21 428 025 €. Nous choisirons dans cet exemple de le rémunérer à hauteur de 10%.

Capital immobilisé	21 428 025 €
Coût du capital	2 142 803 €

IV.2.3 Obtention du tarif commercial

On peut maintenant intégrer au prix le coût du capital ainsi que les frais de gestion. Dans notre exemple nous tarifons 30 000 € de frais à la souscription. Nous souhaitons également ajouter une marge de sécurité. Celle-ci permettra d'assurer un profit malgré des résultats moins bons qu'attendus. Nous choisissons dans cet exemple une marge de 15%.

Le tarif suivant pourra être proposé au client :

***PVFP* (*Best estimate*)**	23 492 667 €
Coût du capital	2 142 803 €
Frais de gestion	30 000 €
Tarif sans marge	21 319 864 €
Tarif final	18 121 884 €

La grille tarifaire segmentée obtenue dans la partie III donne une vision assez précise du risque décès de ce portefeuille proposé pour cession ce qui nous a permis d'évaluer sa valeur actuelle. Nous nous attendons à ce que les primes cédées soient supérieures au coût des sinistres et estimons la valeur actuelle du portefeuille à 23 492 667 €. A partir de cette valeur nous avons construit le tarif commercial en prenant en compte les coûts de fonctionnement du réassureur, sa marge et la rémunération du capital investi. Nous proposerons finalement 18 121 884 € à la cédante pour entrer dans ce contrat en *quota share*.

Conclusion

L'objectif de ces travaux était de mener une analyse de mortalité à partir de l'expérience accumulée par Hannover Re en assurance des emprunteurs individuels, afin de déterminer les variables ayant un effet sur la mortalité. Après avoir vérifié que cela était cohérent, les bases *assurés* et *sinistres* de deux partenaires ont été agrégées dans le but d'obtenir un volume de données suffisant pour mener l'étude.

Une première approche, dite naïve, consistait à isoler chaque sous-population et à effectuer une analyse de mortalité sur chacune d'entre elles. Nous avons rapidement montré les nombreux problèmes qu'implique une telle démarche. L'utilisation d'un modèle intégrant des covariables s'est donc révélé nécessaire et nous avons choisi le modèle de Cox dont les variables ont été sélectionnées par une approche descendante. Les variables retenues sont la *civilité*, la *csp*, le *tabagisme* et l'*ancienneté*. Après avoir sélectionné ces variables, certaines d'entre elles ont été modifiées. C'est notamment le cas de la variable *Ancienneté* dont les modalités ont été réparties en 4 groupes afin de modéliser au mieux l'effet de la sélection médicale. La variable *Csp* a elle aussi été modifiée. La *Csp 4* a été supprimée et une *Csp 0* correspondant aux personnes ayant emprunté une somme supérieure à 500 000 € a été introduite. Cette reclassification du niveau de vie des assurés permet de donner plus de sens à la variable *Csp*. Malgré tout, la significativité de la variable *Csp* reste faible. Un autre résultat intéressant déterminé dans cette partie est le fait que la mortalité en métropole ne semble pas être significativement différente de la mortalité dans les territoires d'outre-mer.

Une fois le modèle sélectionné et les variables réarrangées, le modèle de Cox donne une estimation des coefficients à appliquer à chaque sous-population. Ces coefficients ont indiqué une surmortalité des hommes par rapport aux femmes, des fumeurs par rapport aux non-fumeurs et des personnes moins favorisées professionnellement par rapport aux plus aisées. Les valeurs des coefficients n'ont pas été partagées dans ce mémoire mais elles ont confirmé les impressions des souscripteurs d'Hannover Re et ont permis de les quantifier. L'impact de la sélection médicale a également pu être apprécié au travers de l'analyse de la variable *Ancienneté*. Elle indique une minoration de la mortalité pendant les six années suivant la sélection médicale avant une normalisation à partir de la 7ème année. La première année la mortalité est quatre fois inférieure à celle observée à partir de la 7ème année d'ancienneté. Des offres commerciales sur les premières années de contrat pourront peut-être être proposées aux clients. Cependant, l'effet de la sélection médicale nous invite aussi à rester prudent lorsque l'on observe de bons résultats techniques sur un portefeuille assez récent.

Après avoir obtenu ces coefficients, nous avons estimé la fonction de hasard de base h_0 en appliquant l'estimateur de Breslow. Nous en avons déduit les taux bruts de mortalité par âge de la population de référence. Les taux obtenus avaient une forme générale très satisfaisante mais qui comportait de nombreuses irrégularités. Nous avons donc lissé ces taux bruts pour obtenir les taux de mortalité de référence $\hat{q}_x$. À partir de ces taux de référence, nous pouvons déterminer les courbes de mortalité de chaque individu en appliquant les coefficients correspondants à chaque profil de risque.

Pour utiliser ces taux à des fins de tarification, nous avons dû vérifier la pertinence du modèle. Sa validation a été menée en deux étapes : la validation des hypothèses du modèle de Cox et la validation de l'adéquation des $\hat{q}_x$ estimés avec les valeurs d'origine. L'exigeante hypothèse des hasards proportionnels a été validée par le test de Grambsch et Therneau et une batterie de tests a permis de vérifier que l'estimation des $\hat{q}_x$ était bien fidèle aux données d'origine. La modélisation est donc jugée satisfaisante.

Une première limite de cette méthode est l'hypothèse des hasards proportionnels. Le test de Grambsch et Therneau indique que cette hypothèse peut être faite sur ce jeu de données, toutefois nous savons que ce n'est pas le cas dans la réalité[16]. Par exemple, l'impact sur la mortalité du genre ou de la consommation de tabac n'est pas constant avec le temps. Pour pallier cette limite, nous pourrions mettre en place un modèle Accelerated Failure Time (AFT) comme dans Dal Pont (2020) qui a l'avantage de ne pas faire l'hypothèse des hasards proportionnels. L'inconvénient de cette approche est qu'elle nécessite de faire une hypothèse sur la forme de la fonction de survie, là où le modèle de Cox est plus flexible.

Une seconde limite de la modélisation actuelle est le fait qu'elle ne prend pas en compte de variables croisées. On peut tout à fait imaginer que l'impact du tabagisme n'est pas le même sur la population masculine que féminine, à cause de différences d'habitudes de consommations ou de morphologies. Nous pourrions envisager la mise en place d'un modèle linéaire généralisé qui expliquerait la variable *décès*. Cette approche permettrait de faire apparaitre des variables croisées et permettrait également de faire varier les coefficients du modèle en fonction de l'âge. Toutefois, lorsque le modèle est complexe, le faible nombre de décès observés peut être un frein à la modélisation.

Pour améliorer la procédure de sélection du modèle, une approche par pénalisation pourrait être envisagée (voir Verweij P. et Van Houwelingen H. (1974)).

La mise en place du modèle décrit dans ce mémoire implique donc plusieurs simplifications. Il pourrait être intéressant de mettre en place les modèles mentionnés plus haut afin de comparer les résultats avec ceux de ce mémoire.

L'exemple du rachat d'un portefeuille emprunteur a été choisi pour montrer une application des taux de mortalité obtenus avec le modèle de Cox. Nous avons présenté la démarche permettant de passer des taux de mortalité à la prime commerciale de réassurance. Les notions de *PVFP* et de rémunération du capital ont particulièrement été développées.

Les résultats de cette étude ont permis à Hannover Re d'affiner sa connaissance du risque emprunteur et seront utilisés à des fins de tarification. Ces travaux pourraient être complétés dans le futur par une analyse du risque d'arrêt de travail et du risque de rachat.

[16] Planchet F., Thérond P-E. (2006) Modèles de durée, Applications actuarielles. Economica

Bibliographie

Planchet F., Thérond P-E. (2006) Modèles de durée, Applications actuarielles. Economica

Delacroix A. (2019) Réassurance Non-Vie. Cours, Institut de Science Financière et d'Assurances

Berchtold A. (2013) Données longitudinales et modèles de survie. Cours, Université de Genève

Geffray S. Analyse des durées de vie avec le logiciel R, Université de Marseille

Chen N. (2016) Evolution de la variance d'un modèle de Cox et prise en compte dans le cadre du modèle interne sur un portefeuille de crédit caution. Impact sur le processus ORSA, Institut de Statistique de l'Université de Paris

Koye G.-K. (2019) Le machine learning dans la construction de tables de mortalité d'expérience, ENSAE ParisTech

Babin S. (2016) Tarification en Assurance Emprunteur : Création de tables de mortalité d'expérience après segmentation du portefeuille par scoring, Université Paris Dauphine

Pophillat G. (2019) Calcul de la meilleure estimation d'un traité emprunteur individuel français, Diplôme Universitaire d'Actuaire de Strasbourg

Mietton M. (2013) Construction de tables d'incidence et de maintien en Arrêt de Travail dans le cadre de l'assurance emprunteur, Institut de Science Financière et d'Assurances

S. Sanchez d'Hondt (2012) Analyse de l'effet de la sélection des risques sur la garantie décès/P.T.I.A. en assurance emprunteur

Dal Pont M. (2020) Construction d'une table de mortalité d'expérience en assurance emprunteur

Ahmadi S. et Brown R. (2018) Key factors for explaining differences in Canadian pensioner baseline mortality

Hymans Robertson (2009) What longevity predictors should be allowed for when valuing pension scheme liabilities?

Actuélior (2020) Matinale emprunteur, La loi Lagarde, 10 ans après : analyses et perspectives

Securimut (2020) Libre choix de l'assurance emprunteur immobilier : 3 lois pour quelle réalité ? Bilan des lois Lagarde, Hamon et Bourquin

Tomas J. et Planchet F., Critères de validation : Aspects méthodologiques, Note de travail II1291-14 v1.4 Institut des Actuaires

Forfar et al. (1988) On graduation by mathematical formula

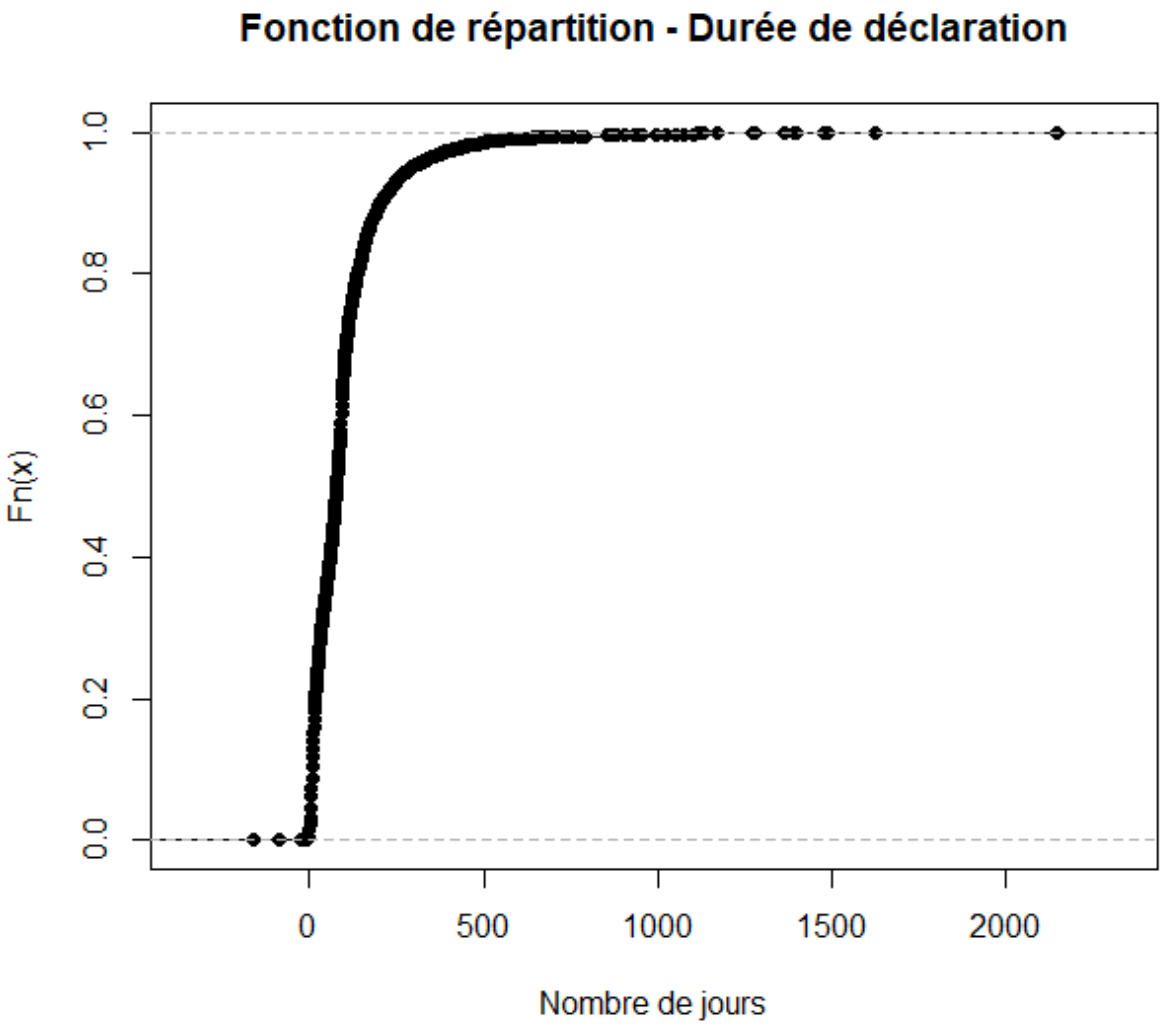
Therneau T. et Aktinson E. (2020) Concordance

Sanchez d'Hondt S. (2012) Analyse de l'effet de la sélection des risques sur la garantie décès/P.T.I.A. en assurance emprunteur

Verweij P. et Van Houwelingen H. (1974) Penalized likelihood in Cox regression

Giesecke L. (1981) Use of the chi-square statistic to set Whittaker-Henderson smoothing coefficient

ANNEXE


Fonction de répartition - Durée de déclaration
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Nombre de jours