Mémoire présenté le : 8 juillet 2015

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

Par : Thomas GUILLON VERNE

Titre : Construction de tables de mortalité d'expérience sur de petits échantillons pour l'estimation de la sinistralité décès.

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'IA		*Signature*
Mme	Brigitte ECARY	
Mr	Christian FETTIG	
Membres présents du jury de l'ISFA		
Mr	Christian ROBERT	
Mr	Stéphane LOISEL	

Entreprise
Nom : Axéria Prévoyance
Signature :
Directeur de mémoire en entreprise
Nom : Jérôme Vuarier
Signature :
Invité
Nom :
Signature :

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

Signature du responsable entreprise

Signature du candidat

Secrétariat :
Mme Christine DRIGUZZI
Bibliothèque :
Mme Patricia BARTOLO

Résumé

Mots clés : mortalité, tables de mortalité d'expérience, estimateur des moments de Hoem, estimateur de Kaplan-Meier, méthode de lissage de Whittaker-Henderson, ajustement logistique, modèle de Cox, modèle à durée de vie accélérée.

L'estimation des risques par les assurances est depuis toujours un enjeu majeur. Avec la mise en place de Solvabilité 2 et les possibilités apparues grâce aux données de plus en plus précises, les assureurs doivent aujourd'hui quantifier le plus précisément ces risques. Cette meilleure estimation possible est nommée *best estimate* par la directive européenne. Dans ce cadre, chaque risque assurantiel s'estime d'une façon différente faisant appel à des méthodes mathématiques adaptées. Axéria Prévoyance étant un assureur de personnes proposant des produits d'assurance de prêt et de prévoyance, il est particulièrement exposé au risque décès.

L'objet de cette étude va ainsi être de proposer une méthode et les résultats obtenus permettant d'améliorer l'estimation actuelle de ce risque. Pour cela, nous allons construire des tables de mortalité adaptées à la population du portefeuille de façon à pouvoir calculer tête par tête et le plus précisément possible les capitaux à provisionner dans le cadre du risque décès.

Dans la première partie de ce mémoire, nous allons commencer par étudier les données. Il sera tout d'abord question de leur fiabilité et de leur suffisance afin de réaliser notre étude. Ensuite, nous en effectuerons une étude descriptive de façon à mieux comprendre le cadre dans lequel nous nous plaçons et à déjà dessiner la tendance vers laquelle les résultats vont s'orienter.

La seconde partie permettra elle de construire les premières tables de mortalité d'expérience femmes/hommes du portefeuille. En premier lieu, les taux bruts de mortalité du portefeuille seront calculés à l'aide des estimateurs des moments de Hoem et de Kaplan-Meier. Ensuite, nous utiliserons des méthodes d'ajustement (Whittaker-Henderson et ajustements logistiques) ainsi que des méthodes de positionnement par rapport à une référence externe pour obtenir des courbes de mortalité d'expérience pour les hommes et les femmes.

La troisième partie consistera à affiner la segmentation du portefeuille de façon à prendre en compte son hétérogénéité. Tout d'abord, à l'aide du test du log-rank, nous définirons la segmentation la plus adaptée pour le portefeuille. Ensuite, à l'aide des modèles de Cox (simple et stratifié) et des modèles à durée de vie accélérée, nous construirons des tables de mortalité avec une segmentation plus fine.

La dernière partie consistera à calculer les estimations de charges décès et les comparer aussi bien aux résultats obtenus à l'aide des hypothèses actuellement utilisées qu'à ceux réels.

Abstract

Key words : mortality, experience-based mortality tables, method of moments of Hoem, Kaplan-Meier's estimator, Whittaker-Henderson's smooth, logistic adjustment, Cox's model, accelerated failure time model.

The estimation of risks has always been a major issue for insurers. And today, with Solvency 2 and big data, they can and must quantify precisely all these risks. This is named best estimate in the European directive. Insurers have to use fitted mathematical methods in order to estimate the different risks. Axéria Prévoyance is a life and health insurer which sells loan insurance and personal life products. So, the death risk is really significant for the company.

In this study, we will set a method to improve the estimation of the death risk and present the results. We need to develop mortality tables adapted to our insurance portfolio to this end. Therefore, we will be able to calculate provisions for the solvency capital requirement (SCR).

In the first part we will study the data of our portfolio. We will check that data are reliable and sufficient to continue the study. We will do some adjustments if there are errors. Then, we will describe the portfolio of the company. We will be enable to understand which type of insured persons the company has in its portfolio and the main results we can expected.

In the second part we build experience-based mortality tables for the segmentation of women/men. First, we will work out crude mortality rates with the method of moments of Hoem and Kaplan-Meier's estimator. Then, with two different types of methods (adjustment methods -Whittaker-Henderson and logistic adjustment- and external reference position methods) we will obtain mortality tables for men and women.

The third part will consist in modeling the heterogeneity of the portfolio. Log-rank test will enable to choose the right segmentation of data. And after, we can use basic and stratified Cox models and accelerated failure time models to build mortality tables for the new segmentation defined.

In the last part we will calculate an estimation of the cost of deaths for one year and then we will compare the findings of the study with current figures and real cost of deaths in the portfolio.

Remerciements

Je souhaite remercier les différentes personnes qui m'ont aidé à réaliser ce mémoire.

Tout d'abord je remercie particulièrement mon tuteur en entreprise Jérôme Vuarier. Il m'a proposé de travailler sur ce sujet et m'a guidé dans les orientations à prendre pour réaliser cette étude. De plus, son esprit critique m'a été utile dans la rédaction de ce mémoire et notamment dans l'interprétation des résultats.

Je tiens aussi à remercier ma tutrice pédagogique Anne Eyraud-Loisel dont les conseils sur la rédaction et la relecture ont été utiles.

Enfin, je remercie les collaborateurs d'Axéria Prévoyance qui ont répondu à mes questions durant ma formation et m'ont ainsi permis d'avancer dans la réalisation de mon mémoire. Je pense notamment à Stanislas Falaise, Wilson Chane-Kee, Mathieu Robaut, Gaspard Jallat et Michael Casalinuovo.

Sommaire

Introduction

Axéria Prévoyance, filiale du groupe April est un assureur de personnes fondé en 1989 et proposant aujourd'hui des produits dans ce domaine tels que l'assurance santé, la prévoyance, l'assurance dépendance ou encore l'assurance emprunteur. Parmi ces produits, la prévoyance et l'assurance emprunteur incluent une garantie décès. Ces produits représentant aujourd'hui une part importante du portefeuille global d'Axéria Prévoyance, le décès est un des principaux risques de la compagnie. Il est donc intéressant et nécessaire de l'estimer au plus juste. Nous nous intéresserons donc dans cette étude aux périmètres de la prévoyance et de l'assurance emprunteur pour lesquels le décès est une des deux causes de versement d'indemnités (avec l'arrêt de travail).

Les provisions concernant le risque décès doivent être estimées de la manière la plus précise possible et l'entrée en vigueur de Solvabilité 2 oblige de plus les assureurs à justifier leur calcul de *best estimate*. Or, il a été observé depuis de nombreuses années que la mortalité du portefeuille d'Axéria Prévoyance est bien inférieure à celle donnée par les tables réglementaires TF 00-02 et TH 00-02. La compagnie ne disposant pas de ses propres tables d'expérience, il a donc été décidé d'en mettre une en place de façon à pouvoir mieux appréhender les risques au niveau individuel.

En effet, aujourd'hui, afin de calculer la sinistralité estimée dans le futur, on utilise des coefficients d'abattement à appliquer sur les tables réglementaires. Ces coefficients sont définis à partir de la maille du sexe, des produits et du caractère fumeur. Ceci permet d'avoir donc des échantillons plus fins qu'en utilisant la seule variable homme/femme. Cependant, ces coefficients ne sont pas calculés directement par la société, il est donc intéressant de faire une étude en interne afin de pouvoir analyser les résultats. De plus, appliquer un coefficient d'abattement sur les tables réglementaires suppose que la distribution des décès du portefeuille est similaire (avec abattement) à celle de la population française. Or, ceci n'a pas été étudié et le fait que la mortalité observée soit largement inférieure à celle réglementaire nous amène à penser que cette hypothèse n'est pas vérifiée.

Nous allons donc dans cette étude construire des tables de mortalité d'expérience afin de mieux connaître notre portefeuille. Les tables de mortalité ne sont pas une fin en soi mais un outil permettant le calcul d'indemnisations estimées futures pour les contrats. Elles peuvent donc servir pour la tarification mais Axéria Prévoyance assure principalement, pour ce type de contrats, des produits commercialisés par d'autres entreprises (des courtiers grossistes notamment comme April l'est historiquement). Ainsi, pour la société, ces tables servent à calculer des estimations de sinistralité future et en particulier à un an dans le cadre de Solvabilité 2. Cette directive imposera à partir de 2016 de calculer un besoin en fonds propres (SCR) qui devra permettre d'honorer ses engagements à horizon un an dans 199 cas sur 200. Le SCR découlera notamment de calculs de *best estimate* (meilleure estimation des flux futurs). Or, la directive solvabilité impose que le *best estimate* soit justifié et corresponde aux spécificités des portefeuilles. La construction de tables de mortalité d'expérience afin de pouvoir calculer un *best estimate* pour le risque décès correspond ainsi parfaitement aux exigences de l'Union européenne.

Ce mémoire sera divisé en 4 parties afin d'arriver à des résultats concrets finaux. Tout d'abord, nous ferons une présentation du portefeuille et des données. Cela permettra notamment de confirmer la

nécessité pour Axéria Prévoyance d'utiliser des tables d'expérience. Ensuite, nous mettrons en place une première construction de tables de mortalité pour les hommes et les femmes ce qui sera justifié par la quantité de données dont nous disposons. Pour y parvenir, il y aura premièrement le calcul des taux bruts du portefeuille, deuxièmement l'ajustement de ces taux à l'aide de différentes méthodes et finalement la comparaison de ces dernières pour l'obtention des tables finales. Comme nous l'avons précédemment dit dans cette introduction, nous disposons actuellement de coefficients d'abattement pour des segmentations incluant les produits et le caractère fumeur en plus du sexe. L'observation du portefeuille nous ayant permis de mettre en évidence de fortes disparités de mortalité entre ces segmentations, il a paru intéressant ensuite de modéliser ces différences. Pour cela, comme les échantillons sont petits, nous avons utilisé des méthodes répandues dans le milieu actuariel. Leur présentation et leur mise en application sera ainsi la troisième partie du mémoire. Enfin, la dernière partie sera consacrée aux conséquences chiffrées sur l'estimation de la sinistralité de l'utilisation des courbes nouvellement construites.

Chapitre 1

Les tables de mortalité

Afin de pouvoir proposer des produits dont l'incertitude repose sur le décès de l'assuré, les assureurs ont besoin de disposer de tables de mortalité qui s'obtiennent grâce à l'observation de populations importantes.

1.1 Tables de mortalité réglementaires

Les tables de mortalité réglementaires sont produites par l'Insee grâce aux statistiques sur la population française. Elles se présentent sous la forme d'un tableau où diverses informations peuvent être présentes pour chaque âge x :
- un coefficient de mortalité à horizon un an : q_x
- le nombre d'individus en vie : $L(x)$
- l'espérance de vie résiduelle : $\mathbb{E}(x)$.

L'Insee donne ainsi cette définition qui explique la méthode générale pour obtenir les tables chaque année :

« Une table de mortalité annuelle suit le cheminement d'une génération fictive de 100 000 nouveaunés à qui l'on fait subir aux divers âges les conditions de mortalité observées sur les diverses générations réelles, durant l'année étudiée. Pour éviter les aléas des tables annuelles et pour disposer d'une table détaillée par âge aussi précise que possible, on calcule également une table de mortalité couvrant une période de trois années ».

Cependant, l'espérance de vie de la population française augmentant chaque année, il existe 2 sortes de tables.

1.1.1 Les tables de mortalité du moment

Ces tables sont éditées chaque année par l'Insee et reflètent la mortalité de la population au moment de l'étude, un âge donné correspond ainsi à un taux de mortalité. En assurance vie non viagère, les tables réglementaires utilisées par les companies sont les TF 00-02 (femmes) et TH 00-02 (hommes) qui ont été établies à partir des données de l'Insee de 2000 à 2002. Les courbes de mortalité de ces tables se présentent sous la forme suivante (on utilise l'échelle logarithmique pour une meilleure lisibilité aux âges faibles).

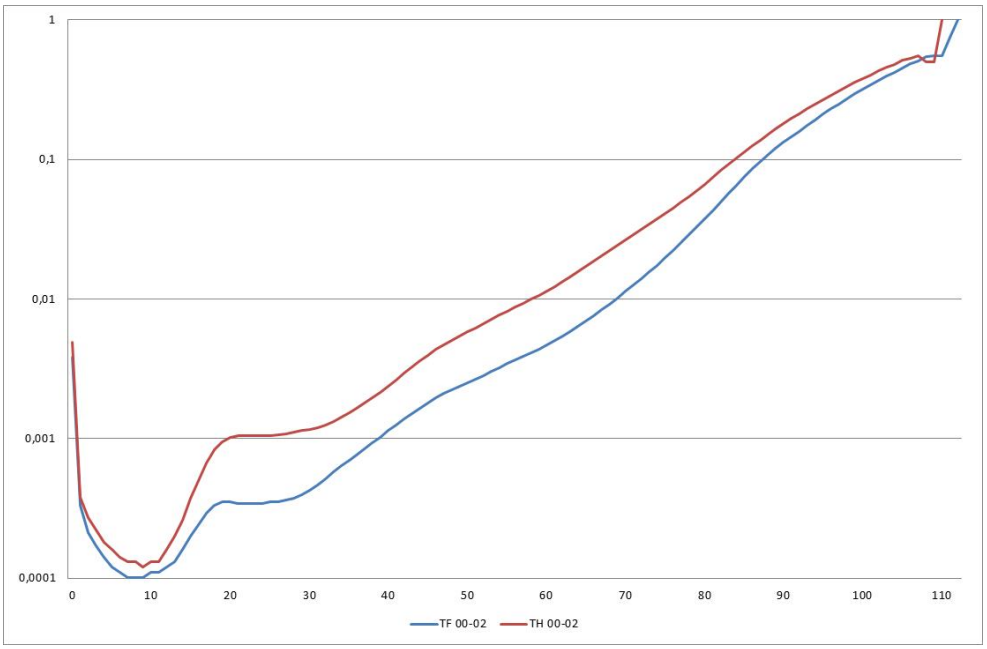


FIGURE 1.1 – Taux de mortalité des tables TF 00-02 et TH 00-02 par âge

On peut remarquer à travers ces courbes que les mortalités masculines et féminines suivent la même tendance. Le premier pic de mortalité correspond à la mortalité infantile et le second vers 20 ans aux mortalités accidentelles et par suicide représentant un pourcentage important des décès à cet âge. Ensuite la mortalité croît approximativement linéairement (car il s'agit d'une échelle logarithmique) jusqu'aux âges ultimes. On note enfin que la mortalité masculine est toujours plus élevée en étant jusqu'à presque trois fois plus importante vers l'âge de 20 ans.

1.1.2 Les tables de mortalité générationnelles

Ces tables référencent la mortalité de la population par génération. Il s'agit donc de produire une table de mortalité pour chaque année de naissance. Pour les produits d'assurance de rentes viagères, les tables réglementaires utilisées sont les TGF 05 (femmes) et TGH 05 (hommes) qui fournissent la mortalité des générations de 1900 à 2005. Ces tables tiennent ainsi compte de la tendance d'évolution de la mortalité.

1.2 Tables de mortalité d'expérience

Dans le cadre de Solvabilité 2, les assureurs sont autorisés à produire leurs propres tables de mortalité basées sur leur portefeuille d'assurés afin de mieux quantifier leur risque (les provisions mathématiques) aussi bien lors de la souscription du contrat que par la suite.

Depuis le 20 décembre 2012, dans le cadre de la lutte contre les discriminations, la commission européenne interdit aux assureurs d'utiliser le sexe comme critère de tarification. Ainsi, ils doivent soit utiliser une table femmes, soit utiliser une table unisexe pour calculer le tarif. Cependant, dans notre cas, comme nous voulons évaluer des provisions, la différence hommes-femmes sera faite.

1.3 Intérêt et utilisation des tables de mortalité

Les tables de mortalité sont utilisées par toutes les compagnies d'assurance (ainsi que les mutuelles et les institutions de prévoyance) afin de tarifer et de provisionner les contrats proposés aux clients. En effet, le risque de décéder dans les 10 prochaines années est très différent que l'on ait 30 ou 60 ans. Ces

tables permmettent donc de modéliser les espérances de durée de vie à chaque âge (une segmentation plus faible que l'âge entier ne présentant pas un grand intérêt du fait de la faible variation de la mortalité comparativement à l'âge) et donc de tarifer ou de provisionner des contrats viagers ou non.

Cependant, pour les assureurs qui ciblent volontairement des personnes présentant un risque moins ou plus élevé que la moyenne (en segmentant à l'aide de la sélection médicale ou des catégories socio-professionnelles), les tables réglementaires ne sont en général pas adaptées. Les tables de mortalité d'expérience peuvent alors permettre de mieux modéliser le risque spécifique auquel la compagnie est soumise. En effet, au-delà de taux de décès moins importants, la forme générale des courbes peut être très différente. Ainsi, les taux de décès d'un portefeuille peuvent être moitié moins élevés à 30 ans par rapport à la population globale mais être identiques à 80 ans. La courbe de mortalité d'expérience a alors une incidence plus importante que celle de référence qu'il est important de prendre en compte pour les calculs à court, moyen ou long terme.

Ainsi, Axéria Prévoyance propose plusieurs types de contrat qui justifient l'utilisation de tables de mortalité d'expérience. Tout d'abord, les contrats d'assurance de prêt standards sont généralement des bons risques au niveau médical. Ensuite, la compagnie a une assurance de prêt spécifique aux gros capitaux. Comme la sélection médicale est poussée pour ces contrats et que les personnes concernées font généralement partie d'une catégorie socio-professionnelle avec moins de décès, la mortalité d'expérience de ces contrats est beaucoup plus faible (nous verrons par la suite que l'abattement est de plus de 70%). Enfin, il est aussi proposé des contrats aux personnes considérées comme ayant un risque "aggravé" car elles ont déjà eu une maladie ou un accident précédemment, ceux-ci impliquant statistiquement un taux de décès plus élevé que la moyenne par la suite.

Chapitre 2

Présentation des données et de l'étude

Dans ce chapitre nous allons présenter le portefeuille d'Axéria Prévoyance, la façon dont sont obtenues puis retraitées les données ainsi que présenter des statistiques descriptives de ces données.

2.1 Présentation du portefeuille d'Axéria Prévoyance étudié

Axéria Prévoyance est un assureur de personnes qui propose des produits de santé et de prévoyance. La particularité de cet assureur est d'exercer son activité en délégation de gestion. Les contrats sont ainsi souscrits et gérés par des organismes qui sont principalement des courtiers grossistes. Nous allons donc maintenant présenter les produits de la compagnie.

Comme nous nous intéresserons au risque de mortalité dans cette étude, nous allons présenter uniquement les produits qui sont concernés par une garantie décès. Les produits d'Axéria Prévoyance concernés par cette garantie peuvent être regroupés dans trois grands types de contrats :

- Les produits d'assurance emprunteur
- Les produits de prévoyance individuelle
- Les produits de prévoyance professionnelle.

Les produits d'assurance emprunteur concernent le remboursement du capital restant dû. Les produits de prévoyance individuelle sont souscrits par des particuliers qui souhaitent s'assurer temporairement pour un capital en cas de décès. Enfin, les produits de prévoyance professionnelle sont similaires à ceux de prévoyance individuelle mais sont souscrits dans le cadre d'une activité professionnelle. Ces assurances peuvent donner lieu à des rentes conjoint ou d'éducation mais la proportion du nombre de contrats concernés et le peu de données à disposition nous amène à ne pas construire de tables de mortalité prospectives. Les tables réglementaires TGF-05 et TGH-05 seront donc utilisées pour le calcul des rentes. De plus, comme nous le verrons par la suite ce sont des produits destinés principalement à des personnes en activité, ce qui implique que les effectifs sous risque au-delà de 65 ans sont faibles contrairement à des garanties qui impliqueraient des rentes viagères.

L'assurance emprunteur représentant la majorité du portefeuille des garanties décès, il existe une seconde segmentation pour ces produits :

- Les produits d'assurance de prêt standards (ADP standard)
- Les produits d'assurance de prêt pour les gros capitaux (ADP gros capitaux)
- Les produits d'assurance de prêt pour les risques aggravés (ADP risques aggravés).

Ainsi, la majorité des ADP sont des standards. Les ADP gros capitaux concernent quant à eux les prêts de sommes importantes. Afin d'accepter d'assurer de gros capitaux qui font donc peser un risque de pointe sur la compagnie, la sélection médicale à la signature du contrat est plus importante (tests complémentaires au classique questionnaire de santé), ceci permettant a priori de ne garder que des personnes à faible risque de décès. Enfin, les ADP risques aggravés sont destinés aux personnes considérées comme ayant un risque plus élevé de mortalité. En effet, une personne ayant eu précédemment dans sa

vie une grave maladie, comme un cancer par exemple, ne pourra pas souscrire une assurance de prêt à un taux identique aux personnes ne présentant pas de risque particulier, des études précédentes ayant montré que le risque de décès est alors plus élevé. Ainsi, ce type d'assurance permet à ces personnes de s'assurer contre une majoration du taux de la prime demandée pour un ADP Standard.

Il faut enfin noter que le portefeuille ADP est en run-off depuis 2011 (qui représente la majorité des contrats étudiés), il n'y aura donc un risque limité de changement de mortalité du portefeuille pour cette segmentation.

2.2 Besoin de mise en place de tables de mortalité d'expérience

Nous avons présenté dans le premier chapitre les tables de mortalité et leur intérêt. Nous allons maintenant étudier statistiquement le portefeuille d'Axéria Prévoyance afin de pouvoir vérifier si les différentes segmentations présentent des disparités de taux de décès par rapport aux tables réglementaires mais aussi entre elles. Pour cela, nous avons comparé la mortalité théorique qui aurait dû être observée sur le portefeuille selon les tables réglementaires à celle réellement observée. Il faut noter que les résultats qui seront présentés ont été obtenus à partir de la base de donnée finale obtenue après retraitement, qui sera présentée par la suite.

Afin de calculer la probabilité de décès théorique de chaque personne par rapport à sa durée d'exposition au risque de décès(selon les tables réglementaires TH 00-02 et TF 00-02), les âges exacts d'entrée dans la période d'étude et de sortie (à 3 décimales) ont été calculés. De plus, afin de calculer les coefficients de mortalité nous avons fait deux hypothèses :

- Les taux de décès instantanés sont constants sur une année.
- Les probabilités de décès théoriques des personnes ayant souscrit un produit ont été majorées suivant un taux correspondant à leur risque médical (cela se traduisant par une surprime dépendant des antécédents médicaux). Ainsi, celui-ci est calculé lors de la tarification du contrat de ces personnes.

L'hypothèse de constance des taux de décès instantanés (hypothèse exponentielle) permet de calculer les taux de mortalité théoriques aux âges non entiers et se traduit mathématiquement par la formule :

$$\forall 0 \leq t \leq 1, \ {}_tq_x = 1 - (1 - q_x)^t$$

où ${}_tq_x$ représente la probabilité de décéder avant t an sachant que la personne est vivante à l'âge x, q_x étant une notation simplifiée de ${}_1q_x$.

Ainsi, grâce aux taux de mortalités théoriques calculés, nous avons pu déterminer le nombre de décès théoriques du portefeuille que nous avons comparé au nombre de décès réels pour la période allant de début 2010 à fin 2013. Les résultats pour le portefeuille total ainsi que pour les différents produits sont les suivants (le nombre de décès théoriques ainsi que le taux d'abattement ont été arrondis).

Portefeuille	Décès observés	Décès théoriques	Abattement global
ADP standard	859	2964	71%
ADP gros capitaux	32	163	80%
ADP risques aggravés	176	198	11%
Prévoyance individuelle	346	610	43%
Prévoyance professionnelle	57	187	69%
Portefeuille global	1451	4049	64%

Remarque
Une personne décédée pouvant avoir souscrit plusieurs contrats, il est normal que le nombre de décès de tous les portefeuilles assemblés soit légèrement supérieur au nombre de décès comptabilisés en ne faisant pas de segmentation.

On peut donc observer que la mortalité du portefeuille est bien plus faible que la mortalité théorique des tables réglementaires, excepté pour l'ADP risques aggravés qui concerne un risque particulier. De plus, une différence si importante de mortalité implique sûrement des variations de ces coefficients selon les âges. Ainsi, le besoin d'Axéria de mise en place de tables de mortalité d'expérience est mis en évidence et va donc être étudié dans la suite. De plus, on peut déjà voir que les taux de mortalité des assurés

semblent être différents suivant les portefeuilles. Il serait donc intéressant de modéliser l'hétérogénéité des données.

Avant de se lancer dans les calculs pour la mise en place des tables d'expérience, nous allons maintenant présenter les bases de données ainsi que les traitements qui ont été effectués.

2.3 Récupération des données et traitement

2.3.1 Données de l'infocentre

Afin d'optimiser tous les travaux effectués sur les données, le groupe April (dont Axéria fait partie) a mis en place un infocentre sous SAS regroupant notamment la majorité des portefeuilles gérés par Axéria (et notamment tous ceux qui seront étudiés ici). Les données disponibles sur l'infocentre sont tous les contrats ayant été actifs au cours des cinq dernières années. Ainsi, initiallement nous avons décidé d'utiliser les données de début 2009 à fin 2013 pour effectuer notre étude de façon à ce que les taux de mortalité du moment ne soient pas biaisés à cause de l'augmentation de l'espérance de vie au cours des années.

Les différentes données que nous avons utilisées étant réparties sur plusieurs tables de l'infocentre, nous en avons utilisées plusieurs en les liant grâce à différentes clés. Les contrats sont en effet identifiés par trois clés principales :

- Le numéro d'adhérent qui est attribué lorsqu'une nouvelle personne souscrit un contrat qui ne pré-existait pas. Celà signifie qu'un numéro d'adhérent peut regrouper plusieurs personnes (en général des conjoints pour les contrats d'assurance emprunteur) mais aussi plusieurs contrats souscrits par la suite par les personnes ayant ce numéro d'adhérent.
- Le numéro d'affaire qui regroupe le numéro d'adhérent et un code produit ce qui permet d'identifier les contrats associés à ce numéro d'adhérent.
- Le numéro de personne qui identifie chaque assuré individuellement.

Dans le cas des contrats avec bénéficiaires désignés ou pour les personnes ayant souscrit également un contrat santé famille par exemple, les enfants peuvent être identifiés également sous le numéro d'adhérent de leurs parents. Comme ces personnes ne sont pas incluses dans la garantie décès (pas de versement de capital en cas de décès des enfants) nous les avons enlevées de la base de données.

Ainsi, les données collectées à l'aide de l'infocentre ont permis d'obtenir de nombreuses informations dont nous listons les principales ici (d'autres informations ont été ajoutées et ont notamment permis de faire des vérifications) :

- le numéro de personne, le numéro d'adhérent et le code produit
- la date de naissance, le nom et la civilité de la personne
- les dates d'effet et de fin de contrat
- la cause de résiliation du contrat (le décès est indiqué comme cause de fin)
- le caractère fumeur, non fumeur ou non renseigné de la personne

Une fois que les personnes ayant des contrats avec des garanties décès faisant partie des produits que nous modélisons ont été identifiés, nous avons pu les traiter de façon à obtenir une base de données où chaque ligne correspond à une personne.

2.3.2 Traitement des données

Les traitements effectués sur les données peuvent être catégorisés. Nous allons les lister afin que cela soit plus aisé à comprendre. Les traitements effectués peuvent ainsi être dus aux causes suivantes :

- Une hypothèse générale sur les données
- Un regroupement de plusieurs lignes d'une même personne
- Une modification des données suite à des erreurs dans la base.

Ainsi nous allons présenter ces traitements un à un.

Hypothèse

Dans les cas où une personne est décédée en ayant un contrat avec garantie décès mais que l'indemnisation a été refusée (généralement pour une maladie non déclarée dans le questionnaire de santé ou pour une cause non couverte par le contrat), le contrat est conservé dans la base de données en considérant la personne sous risque jusqu'au décès, mais en ne retenant pas le décès comme cause de sortie. En effet, comme il n'y a pas eu le versement d'indemnités, on ne considère pas que le risque s'est réalisé pour la compagnie mais qu'il s'agit d'une censure.

Modification des données

Malgré le fait que l'infocentre ait été créé pour améliorer la fiabilité des données, des erreurs peuvent toujours être présentes, notamment car les données sont toujours renseignées manuellement par les courtiers ou gestionnaires à la base. Ainsi, de nombreuses vérifications ont été effectuées de façon à corriger les éventuelles erreurs. Nous allons présenter ici ce qui a été fait.

Tous les retraitements qui vont être présentés ne seront pas forcément chiffrés mais il faut savoir que l'ensemble des modifications effectuées représente moins de 1% des données utilisées.

En analysant les données nous avons pu effectuer des retraitements automatiques qui sont les suivants :
- Si les dates de contrats sont incohérentes (dates d'effet après date de fin), les lignes de ces contrats ont été supprimées. 40 contrats étaient concernés soit moins de 0,01% des lignes.
- Si pour un même numéro d'adhérent on identifie deux personnes ayant les mêmes dates de naissance et le même code civilité, on modifie l'un des deux numéros de façon à n'en garder qu'un seul.

Cependant, d'autres retraitements ne peuvent être réalisés automatiquement et demandent d'aller vérifier les informations dans les fichiers de gestion. Nous avons donc listé ces retraitements dans le tableau suivant en indiquant à chaque fois le nombre de personnes concernées et les actions effectuées.

Type d'erreur	Nombre observé	Action effectuée
Décès faussement comptabilisé (décès réel du co-assuré)	12	Modification de la cause de fin de contrat
Dates de décès inexacte (le co-assuré étant toujours assuré)	4	Modification de la date
Deux numéros de personnes dont un non référencé en gestion	194	Suppression de la ligne pour le numéro non référencé
Deux numéros de personne référencés avec une date de naissance différente	17	Changement du numéro et de la date de naissance
Date de naissance incohérente (trop récente ou ancienne)	42	Modification de la date de naissance
Décès non comptabilisé (obtenu par recoupement cause de fin)	6	Modification cause et date de fin
IBNR (à fin juin)	21	Modification cause et date de fin
Dates non possibles ($mois > 12$ par exemple)	0	Modification de la date

On peut remarquer que les retraitements effectués sur les décès sont relativement importants mais le fait que l'infocentre permette d'identifier les décès de plusieurs manières permet d'obtenir au final des données fiables.

Regroupement des lignes d'une même personne

De nombreuses personnes assurées dans le portfeuille ont souscrit plusieurs contrats qu'ils aient été actifs en même temps (assurance emprunteur et garantie temporaire décès par exemple) ou non (assurance de deux emprunts consécutifs). Ainsi, pour ces personnes nous retrouvons plusieurs lignes dans la base de donnée. Deux cas généraux peuvent se présenter que nous présentons dans les graphiques en dessous.

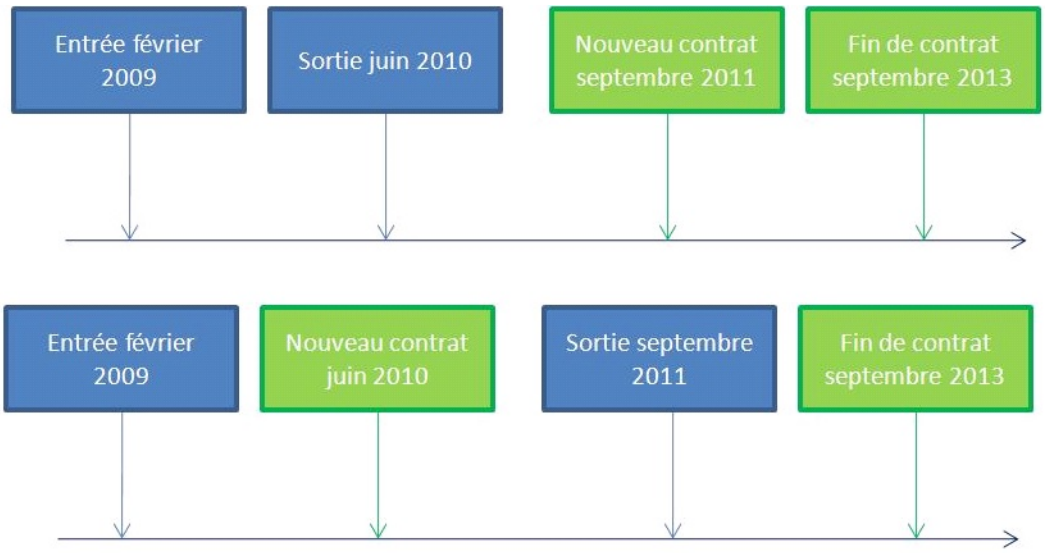


FIGURE 2.1 – Présentations des cas possibles pour des personnes ayant plusieurs contrats

Ainsi, dans le cas ou une personne a eu deux contrats ne se chevauchant pas, on garde deux lignes car sinon cela reviendrait à considérer une personne sous risque alors qu'elle n'est pas assurée et donc à diminuer artificiellement le taux de mortalité.

Remarque 1
Dans le cas ou la période entre la résiliation d'un contrat et la souscription d'un nouveau est inférieure à un mois, nous avons considéré que les contrats se chevauchaient comme dans le second cas présenté ici.

Ainsi, dans le cas où les contrats se chevauchent on les regroupe en une seule ligne en prenant en compte comme dates d'effet et de fin de contrat les dates extrêmes.

Remarque 2
Le traitement des contrats se chevauchant est effectué indépendemment selon que l'on considère la totalité des contrats ou l'ADP seul par exemple. Ainsi, si une personne a un contrat ADP standard et un autre de prévoyance individuelle, la durée du contrat sera uniquement celle de l'ADP standard si l'on étudie la segmentation ADP standard et la durée totale des deux contrats si l'on se place dans le cas général.

2.4 Suffisance des données

Après avoir obtenu une base de données avec une ligne correspondant à une personne, nous allons vérifier que les données sont en quantité suffisante pour pouvoir établir une table de mortalité d'expérience. Pour cela nous allons utiliser le critère de Cochran qui énonce qu'il faut pour cela que les données vérifient :

$$N_x \, \hat{q}_x \geq 5 \qquad \text{et} \qquad N_x \, (1 - \hat{q}_x) \geq 5 \, .$$

N_x représente ici l'effectif sous risque à l'age x et $\hat{q}_x$ l'estimateur empirique du taux de mortalité tel que $\hat{q}_x = \frac{d_x}{N_x}$, avec d_x le nombre de décès à l'âge x. Les données doivent donc finalement respecter les conditions suivantes :

$$d_x \geq 5 \qquad \text{et} \qquad N_x - d_x \geq 5 \, .$$

Voici les valeurs obtenues pour les hommes, les femmes, les hommes du contrat ADP dont le caractère fumeur n'est pas renseigné et les femmes avec les mêmes caractéristiques. On notera par la suite ces deux dernières segmentations la segmentation des hommes de référence et la segmentation des femmes de référence. Ce sont les quatre segmentations retenues pour le calcul des taux bruts et l'ajustement qui sera effectué par la suite. Le choix de ces deux dernières segmentations servira à partir du chapitre 5 afin de modéliser l'hétérogénéité des données.

	Hommes		Femmes		ADP Hommes non renseigné		ADP Femmes non renseigné	
Age	N_x*q_x	$N_x*(1-q_x)$	N_x*q_x	$N_x*(1-q_x)$	N_x*q_x	$N_x*(1-q_x)$	N_x*q_x	$N_x*(1-q_x)$
18	0	45	0	26	0	0	0	0
19	0	89	0	73	0	0	0	1
20	0	169	0	131	0	1	0	0
21	1	272	0	234	0	1	0	1
22	0	424	0	436	0	1	0	2
23	0	757	0	902	0	2	0	8
24	0	1309	0	1772	0	4	0	18
25	1	2285	0	3189	0	23	0	34
26	4	3860	1	5232	0	55	0	106
27	0	6191	2	8293	0	142	0	276
28	2	9554	2	12509	0	360	0	704
29	1	13893	4	17847	0	813	0	1618
30	**8**	**19072**	2	23804	0	1699	0	3116
31	**8**	**24517**	3	29264	0	3059	0	5210
32	**9**	**30138**	**8**	**34209**	1	5008	4	7823
33	**15**	**35281**	**5**	**38244**	6	7329	1	10595
34	**21**	**39544**	**10**	**40672**	**4**	**9985**	3	13181
35	**22**	**44013**	**8**	**43127**	**6**	**12952**	3	15763
36	**33**	**47910**	**15**	**44703**	**8**	**16033**	5	17956
37	**27**	**51561**	**16**	**45670**	**12**	**19059**	8	19823
38	**35**	**54046**	**10**	**45512**	**15**	**21351**	2	20773
39	**22**	**54203**	**13**	**43311**	**7**	**22358**	7	20329
40	**32**	**51950**	**13**	**39432**	**14**	**21860**	8	18715
41	**24**	**47976**	**11**	**34380**	**9**	**20275**	3	16419
42	**35**	**43355**	**13**	**29539**	**19**	**18214**	9	14060
43	**27**	**38594**	**10**	**25174**	**13**	**16140**	4	11892
44	**16**	**34065**	**14**	**21467**	**5**	**14100**	8	9962
45	**28**	**29624**	**8**	**18118**	**10**	**11971**	4	8237
46	**31**	**25471**	**14**	**15028**	**12**	**10026**	5	6580
47	**21**	**21705**	**12**	**12393**	**6**	**8177**	5	5183
48	**27**	**18346**	**8**	**10217**	**8**	**6394**	6	4022
49	**31**	**15352**	**11**	**8449**	**15**	**4870**	1	3063
50	**30**	**12887**	**12**	**7073**	**10**	**3658**	5	2325
51	**24**	**11009**	**13**	**6005**	**4**	**2714**	2	1743
52	**19**	**9530**	**8**	**5274**	**4**	**2065**	3	1327
53	**19**	**8366**	**6**	**4666**	**5**	**1604**	0	1057
54	**22**	**7467**	**10**	**4111**	**4**	**1267**	2	866
55	**34**	**6564**	**9**	**3669**	**7**	**1016**	1	726
56	**25**	**5897**	5	3265	**5**	**848**	1	601
57	**26**	**5327**	8	2900	7	718	0	513
58	**26**	**4737**	2	2599	6	643	0	463
59	**27**	**4270**	5	2341	6	609	0	445
60	**29**	**3766**	6	2066	4	574	1	404
61	25	3257	5	1793	2	511	0	341
62	25	2881	9	1599	4	416	0	284
63	27	2573	4	1427	2	342	1	227
64	20	2203	5	1289	1	281	1	202
65	16	1835	5	1140	2	239	0	185

FIGURE 2.2 – Tableau présentant les valeurs du critères de Cochran par âge

On a indiqué ici en bleu les âges où les données retenues pour chaque segmentation sont suffisantes (ce n'est pas tout à fait le cas pour les derniers âges du portefeuille ADP Hommes non renseigné mais la tendance observée étant satisfaisante nous les avons gardés). Ainsi, sur ces plages seront utilisées des méthodes d'ajustement ou de lissage des taux bruts, alors que sur le restant des âges des méthodes de positionnement par référence externe seront utilisées. A noter que pour les hommes, les critères de Cochran sont bons jusqu'à 80 ans mais la volatilité importante des taux de décès (due au faible effectif sous risque) et le fait que ces âges ne représentent pas un pourcentage important de l'effectif sous risque, nous préférerons utiliser une autre méthode pour l'estimation. Enfin, pour la dernière segmentation des femmes, on peut voir que le critère de Cochran n'est pas respecté. Par conséquent la courbe de mortalité de cette population sera obtenue par ajustement à la courbe de mortalité des femmes avec les méthodes présentées dans le chapitre 4.

Remarque
Les plages d'âges pour lesquelles une modélisation sans référence externe est possible sont relativement faibles, il est donc intéressant de se demander si une telle modélisation plus complexe à effectuer que le positionnement par rapport à une référence externe est utile. Ainsi, la réponse est oui car le coeur de cible et des effectifs sous risques du portefeuille se trouve entre 25 et 55 ans (avec un pic d'effectifs sous risque à 40 ans), il est donc utile de modéliser le plus finement possible la mortalité des personnes présentes dans cette tranche d'âges.

2.5 Statistiques descriptives du portefeuille

Maintenant que les données ont été traitées et que la faisibilité de l'étude a été mise en évidence, nous allons décrire les données de façon à ce que les différentes hypothèses effectuées tout au long de l'étude puissent être justifiées et mises en rapport avec la particularité des données.

Comme nous l'avons vu précédemment, la plage d'étude des contrats que l'on peut utiliser est de 5 années (de 2009 à 2013). Les tables de mortalité du moment impliquent une stabilité des taux de mortalité du portefeuille sur toute la durée sélectionnée. Nous avons donc calculé l'abattement de mortalité par année du portefeuille global par rapport aux tables de référence TF 00-02 et TH 00-02 de façon à voir si cette hypothèse était vérifiée. Voici les résultats obtenus.

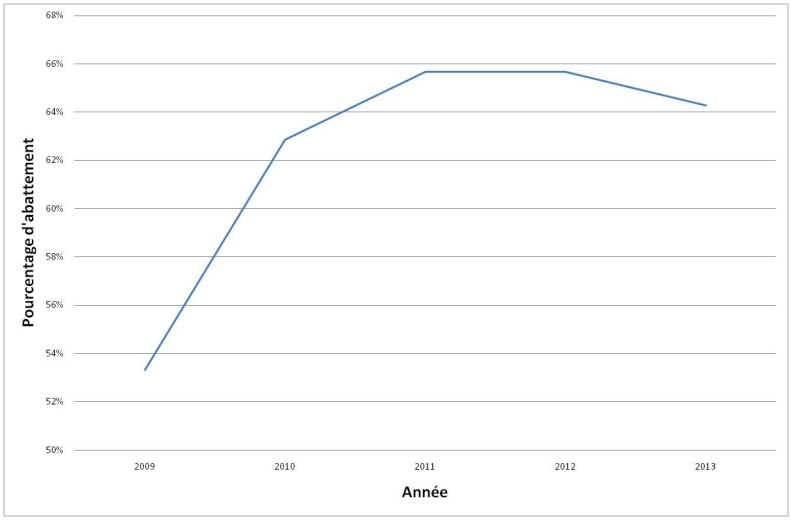


FIGURE 2.3 – Taux d'abattement de mortalité du portefeuille global par année par rapport aux tables TH 00-02 et TF 00-02

On remarque que le taux d'abattement de l'année 2009 est entre 15% et 20% plus faible que les 4 années suivantes. Cette différence étant très importante et ne respectant pas le critère de stabilité de la mortalité sur la période observée, nous avons décidé d'exclure l'année 2009 de l'étude et donc d'étudier les données sur la période 2010-2013.

Remarque
Cette différence de mortalité par rapport aux années suivantes n'a pas pu être expliquée par des raisons de surmortalité en France ou encore par un nombre important de résiliations ou d'adhésions par rapport aux autres années. De plus, en regardant le nombre d'années d'exposition et le nombre de décès, on voit que seul ce second critère varie fortement.

La période 2010-2013 pour l'observation des données ayant été retenue nous pouvons maintenant présenter des données chiffrées sur le portefeuille modélisé. Ainsi, voici la présentation des effectifs ayant eu au moins 1 contrat actif durant cette période. Nous présentons ici les effectifs globaux ainsi qu'avec différentes segementations qui seront modélisées par la suite.

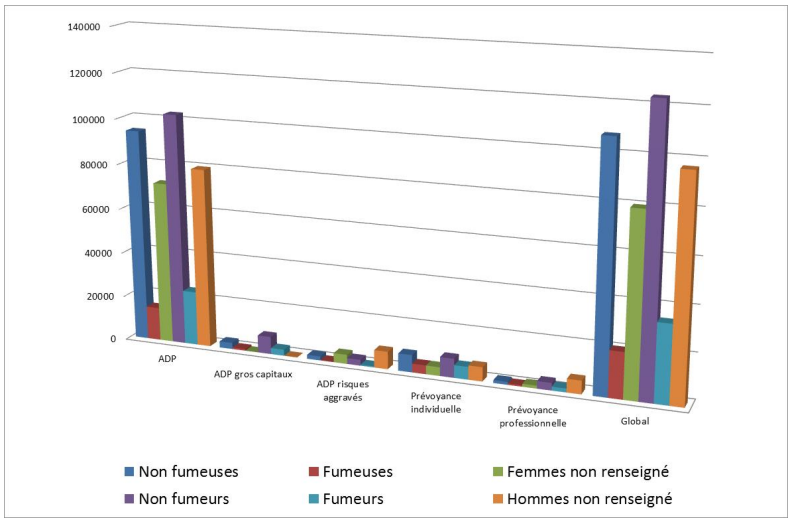


FIGURE 2.4 – Effectifs des différents portefeuilles sur la période 2010-2013

Grâce à ce graphique on s'aperçoit bien que l'ADP standard représente la majorité du portefeuille modélisé. De plus, on peut observer le fait que le caractère fumeur/non fumeur n'a été pris en compte qu'à partir des années 1990 dans la tarification car seules les personnes ayant souscrit un contrat ADP gros capitaux ont toutes ce critère renseigné (les contrats ADP Master étant vendus depuis le début des années 2000 alors que les autres le sont depuis le début des années 80 ou 90).

Une autre donnée intéressante du portefeuille est la répartition par âges selon le sexe, dont voici le graphique.

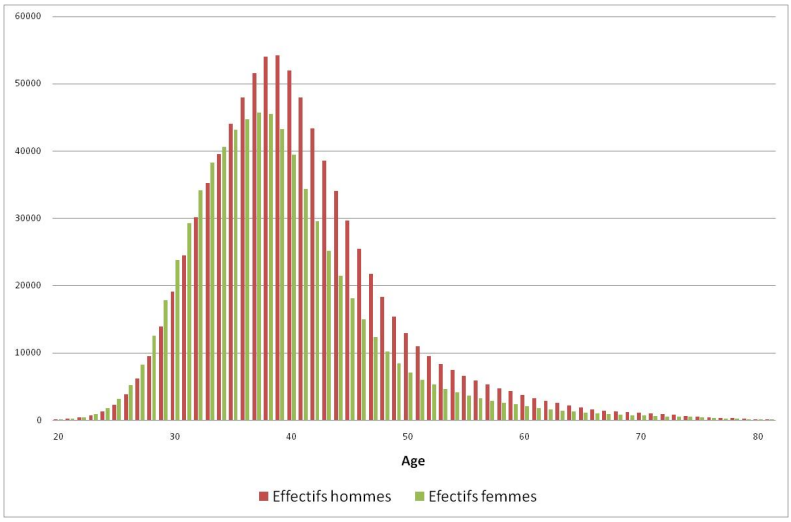


FIGURE 2.5 – Effectifs sous risque du portefeuille global par âge et par sexe sur la période 2010-2013

Avec ce graphique, on peut tout d'abord remarquer que la majorité des effectifs sous risque se situent entre 30 et 50 ans ce qui confirme le choix de modéliser la mortalité avec des modèles permettant de lisser ou d'ajuster les taux bruts sur ces âges.

La seconde observation est le fait qu'il y ait globalement plus d'hommes que de femmes dans le portefeuille. Ceci s'explique à priori par le fait que les produits proposés sont majoritairement destinés aux personnes exerçant une activité professionnelle (ou ayant un revenu stable), sachant que le taux d'actifs chez les hommes est plus élevé que chez les femmes (entre 15 et 64 ans, trois quart des hommes et deux tiers des femmes sont en activité selon l'Insee).

On peut aussi observer la proportion identique d'hommes et de femmes sous risque à partir de 50 ans. Cela peut notamment s'expliquer par 2 facteurs :

- Les prêts sont souvent souscrits pour des périodes entre 10 et 20 ans ce qui favorise la stabilité du portefeuille.
- Les personnes souscrivant un prêt au-delà de 50 ans sont moins ciblées par les produits proposés donc les nouveaux clients sont moins nombreux (ceci est compensé par les garanties temporaires décès qui peuvent être souscrites à tout âge)

Enfin, on remarque que jusqu'à plus de 30 ans, le nombre de femmes est plus important que celui des hommes. Il ne semble pas y avoir d'explication statistique de la population française expliquant cela. Ainsi, on peut supposer qu'à cet âge les offres proposées aux femmes sont plus concurrentielles par rapport à celles proposées aux hommes.

Après avoir analysé les effectifs, nous allons maintenant nous intéresser aux décès qui constituent le coeur de cette étude. Une des données intéressante à regarder est la répartition des décès au cours de l'année. Pour cela, nous avons tracé le graphique du nombre de décès par mois au cours de la période 2010-2013.

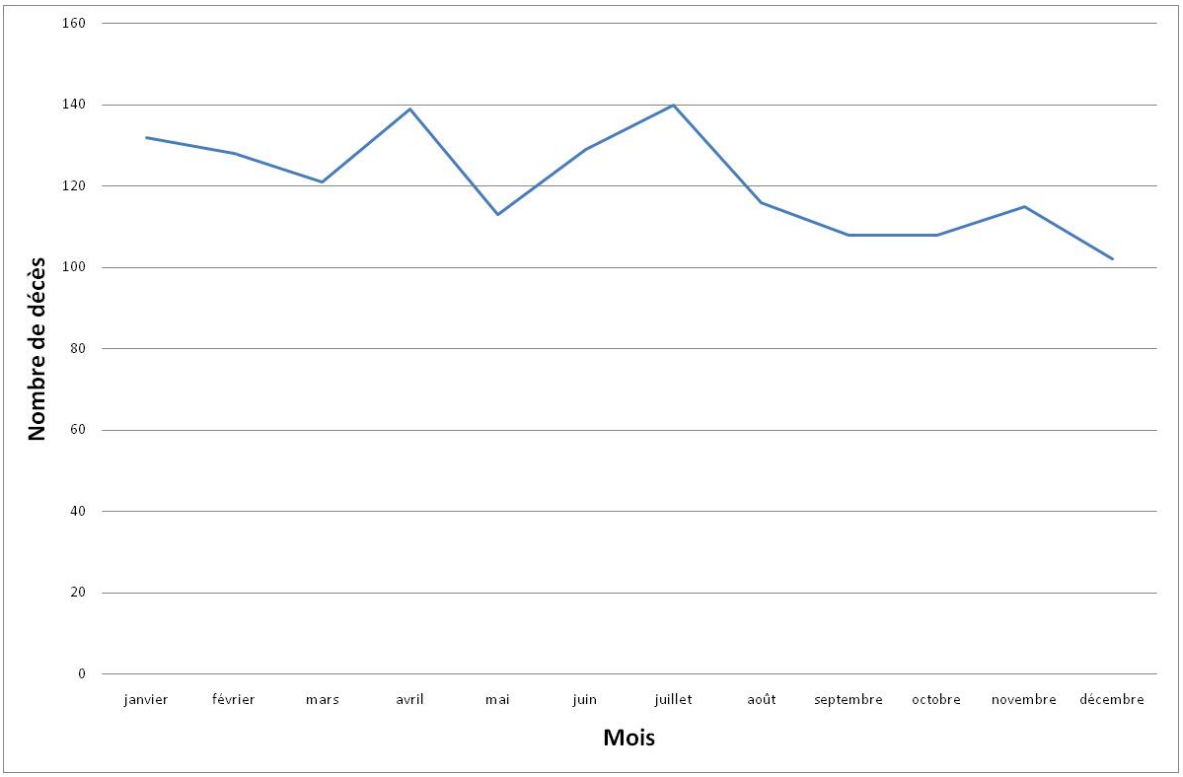


FIGURE 2.6 – Nombre de décès par mois au cours de la période 2010-2013

Ce graphique montre qu'il y a un nombre moins élevé de décès dans les derniers mois de l'année qui peut être dû aux décès non encore déclarés (IBNR). Ceci permet de valider le choix d'utiliser des données sur des années civiles complètes car ne pas le faire aurait pu entraîner un biais supplémentaire dans les calculs. Il faut noter que la baisse des décès durant l'été se vérifie sur la population française mais elle devrait remonter ensuite sur les derniers mois de l'année. Les IBNR ayant été pris en compte, la particularité du portefeuille n'a pas à priori d'explication naturelle sinon que les clients ne représentent pas fidèlement la population française (ce qui est le cas et qui explique notamment la mise en place des tables de mortalité d'expérience).

Voici maintenant le graphique s'intéressant au nombre de décès par tranche d'âges pour les hommes et les femmes.

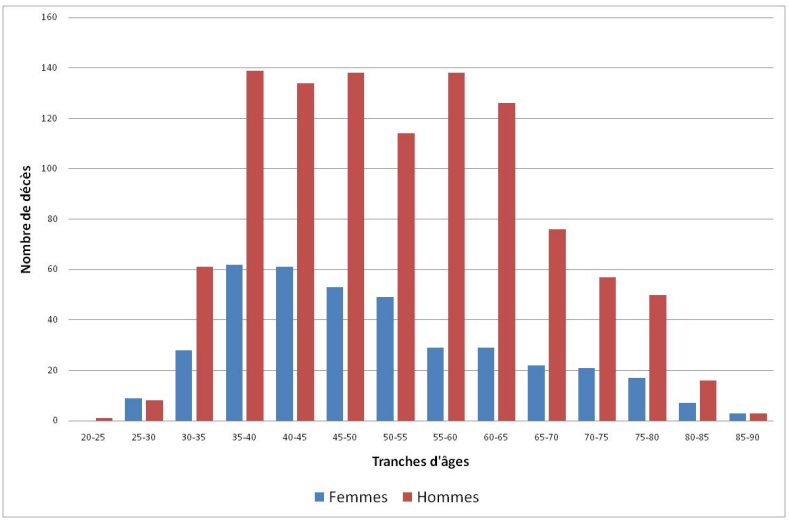


FIGURE 2.7 – Nombre de décès par tranche d'âges pour les hommes et les femmes

Ce graphique illustre bien l'importante différence de mortalité entre les hommes et les femmes car le nombre de décès est 2 à 3 fois moins important à partir de 30 ans ce qui ne se retrouve pas dans les effectifs précédemment présentés. De plus, on peut voir que malgré la baisse significative des effectifs à partir de 40 ans, le nombre de décès ne baisse que légérement jusqu'à 55 ans pour les femmes et 65 ans pour les hommes. Enfin, on peut voir grâce à ce graphique que les plages retenues pour la modélisation des taux de mortalité sans utiliser de référence externe (30-60 ans pour les hommes et 35-55 ans pour les femmes) correspondent bien aux âges où le nombre de décès est le plus élevé.

La partie descriptive des données étant terminée, nous allons présenter les segmentations de portefeuille retenues pour la modélisation

2.6 Segmentation des données pour l'estimation et l'ajustement des taux bruts

Comme nous l'avons précédemment présenté, le but de cette étude est de construire des tables de mortalité pour le portefeuille de garantie décès avec capitaux d'Axéria Prévoyance. Pour cela, nous avons tout d'abord décidé de construire des tables d'expérience hommes et femmes pour la totalité du portefeuille. En effet, on sait que la mortalité féminine est de façon générale moitié moindre que celle des hommes, la segmentation selon le sexe est donc assez naturelle.

Cependant, comme nous avons pu le voir dans la partie "Besoin de mise en place de tables de mortalité d'expérience", les taux de mortalité suivant les différents produits proposés sont très éloignés. Cela peut aussi être mis en évidence dans le cas des personnes fumeuses, non fumeuses, ou dont le caractère n'est pas renseigné (les taux d'abattement par rapport aux tables de référence sont de l'ordre de 45%, 75% et 56% respectivement). Ainsi, nous mettrons en place, après avoir construit les tables de mortalité pour les hommes et les femmes, des modèles permettant de prendre en compte l'hétérogénéité des données et donc essayer d'obtenir des courbes de mortalité pour chaque sous-segmentation. Ainsi, cela demnandera de créer des tables des mortalité d'expérience pour des populations de référence. Afin de ne pas surcharger le mémoire de trop nombreux graphiques, les 2 chapitres suivants présentant l'estimation des taux bruts ainsi que l'ajustement des taux ne seront appliqués que sur la segmentation hommes/femmes. Ainsi, les courbes obtenues avec des segmentations différentes seront présentées à la fin du chapitre 4 en expliquant brièvement les modèles retenus.

Chapitre 3

Estimation des taux bruts

3.1 Données censurées et tronquées

Avant de se lancer dans l'estimation des taux bruts, il est important de présenter les censures et les troncatures qui sont un paramètre important à prendre en compte dans les modèles.

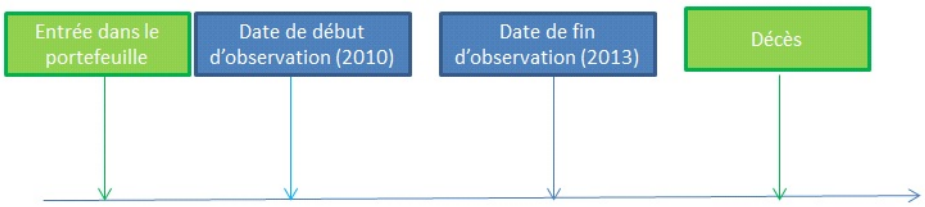
3.1.1 Censure

Voici la définition des censures dans le cas de nos données.

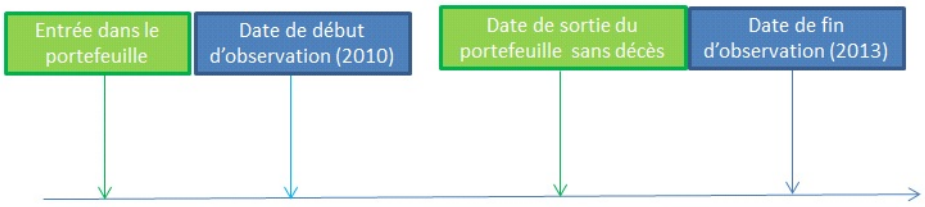
Censure à droite : l'individu est dit censuré à droite s'il n'est pas décédé lors de sa dernière observation.

Censure à gauche : l'individu est dit censuré à gauche s'il était déjà présent dans le portefeuille au moment du début d'observation des données.

De plus, comme les données étudiées se situent entre 2010 et 2013, il peut y avoir des censures fixes et des censures aléatoires. Voici des schémas présentant les différents types de censures pour un individu.



Cas d'une censure fixe à gauche et à droite



Cas d'une censure fixe à gauche et aléatoire à droite

De façon mathématique, on note (X_i) les variables aléatoires de durée de survie des individus et (C_i) les variables aléatoires positives représentant la censure (dans le cas de la censure fixe, les C_i sont fixés, égaux et positifs. On associe à chaque X_i l'observation (T_i, D_i) avec T_i la durée d'observation de l'individu

i et D_i la variable décès. Ainsi on peut écrire l'expression suivante caractérisant la censure :

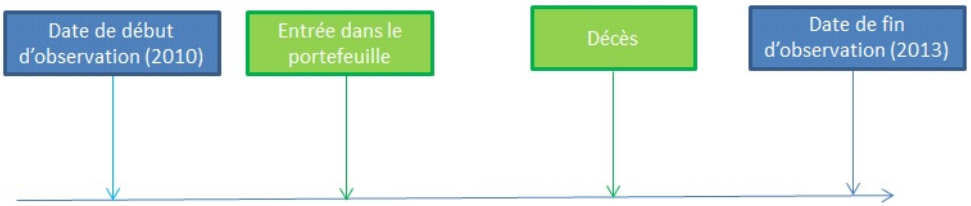
$$T_i = min(X_i, C_i) \quad , \quad D_i = \begin{cases} 1 \, si \, X_i \leq C_i \\ 0 \, si \, X_i > C_i \end{cases} \tag{3.1}$$

3.1.2 Troncature

Dans le cadre de nos données la troncature est définie de la façon suivante.

Troncature à gauche : l'individu est dit troncaturé à gauche s'il est entré dans le portefeuille après le début de la période d'observation (2010).

Ainsi voici un schéma de troncature.



Cas d'une troncature à gauche

Dans un cadre mathématique, en notant c la troncature on obtient la fonction de survie suivante :

$$S(t|c < T < C) = \begin{cases} 1 \;\; si \;\; t < c \\ \frac{S(t)-S(C)}{S(c)-S(C)} \;\; si \;\; c \leq t \leq C \\ 0 \;\; si \;\; t > C \end{cases} \tag{3.2}$$

3.2 Choix des estimateurs

Les bases de données ayant été traitées, on va donc maintenant pouvoir calculer les taux bruts de mortalité qui seront la base des tables finales d'expérience.

Pour calculer les taux bruts à partir des données d'expérience, plusieurs estimateurs existent, il faut donc choisir à priori lesquels sont les plus adaptés à nos données. Il existe deux grands types d'estimateurs : les estimateurs paramétriques et les non paramétriques.

Dans le cas de tables d'expérience de mortalité, les estimateurs paramétriques supposent une distribution à priori de la table de mortalité. Ainsi, l'estimateur couramment utilisé pour estimer la mortalité est l'estimateur binomial qui fait l'hypothèse d'une distribution binomiale de la mortalité à chaque âge x. Du fait du caractère binaire du décès (l'individu est vivant ou est décédé à un âge donné), cet estimateur permet en général d'obtenir de bons résultats. Cependant, il oblige à étudier une population où toutes les personnes commencent à être observées en même temps et toutes les sorties doivent être des décès. Comme on a vu précédemment qu'on était en présence de censure et de troncature avec les données utilisées, cet estimateur ne semble pas adéquat. Ainsi, Hoem a proposé un modèle généralisant cet estimateur et permettant de prendre en compte les censures et les troncatures. Afin de pouvoir l'utiliser, il faut disposer des dates d'entrée et de sortie afin de calculer une exposition au risque. Comme les données ont été préparées afin d'obtenir des dates précises de début et de fin de contrat, ce sera une des deux méthodes retenues pour estimer les taux bruts de mortalité.

Les estimateurs non paramétriques ont eux l'avantage de ne pas présupposer une distribution des lois de mortalité. Un premier estimateur était utilisé auparavant dans l'actuariat, l'estimateur actuariel introduit en 1912 par Böhmer. Il a l'avantage de ne pas s'appuyer sur des données individuelles ce

qui était plus facile à mettre en oeuvre lorsque l'informatique n'existait pas. Cependant, dans le cas d'échantillons de taille réduite comme c'est le cas ici, l'adéquation aux données du portefeuille étudié est souvent mauvaise et par conséquent nous ne l'utiliserons pas. Le second estimateur pouvant être utilisé est celui de Nelson-Aalen. Cependant, il a tendance à sous-estimer la fonction de hasard cumulé, ce qui est dérangeant pour une étude sur la mortalité (voir le cours de Frédéric Planchet *Statistique des modèles non paramétriques* pour plus d'informations). Ainsi, l'estimateur couramment utilisé aujourd'hui en actuariat et en biologie médicale est celui de Kaplan-Meier. Introduit en 1958, il permet de calculer à chaque instant de sortie due à un décès, la probabilité de survie des individus présents à ce moment-là. Cet estimateur demande ainsi de connaître pour chaque personne les dates exactes d'entrée et de sortie, ce qui le rend difficile à mettre en oeuvre sur des milliers de données. Cependant, comme nous disposons de ces informations, cette méthode sera aussi utilisée pour le calcul des taux bruts.

Avant de mettre en application ces méthodes nous allons maintenant présenter théoriquement ces estimateurs. Ainsi, la présentation de ceux-ci se base largement sur les cours de Monsieur Planchet ainsi que sur le livre *Modèles de durée, Applications actuarielles* de Messieurs Planchet et Thérond.

3.3 Estimateur des moments de Hoem

3.3.1 Principe

L'estimateur des moments de Hoem consiste à découper la plage des données étudiées en n intervalles qui seront ici les âges entiers $[x, x+1]$, puis, pour chaque individu à calculer l'exposition $[a_i, b_i]$ dans chaque intervalle $[x, x+1]$. Chaque individu pourra donc être présent dans plusieurs intervalles consécutifs, mais n'aura pas toujours le même poids sur l'ensemble des individus d'un intervalle en fonction de sa durée d'exposition.

3.3.2 Hypothèses et calcul

Les hypothèses pour cet estimateur sont les suivantes :

- Les décès sont indépendants les uns des autres : cette hypothèse est considérée comme vérifiée par le portfeuille car il est très rare que deux personnes assurées décèdent de la même cause (cependant, cela peut se produire dans un accident de voiture par exemple).

- La probabilité de décéder entre x et $x+t$ pour $t \in [0,1]$ est une fonction linéaire du temps de telle sorte que ${}_tq_x = tq_x$. De plus, cette hypothèse permet de faire l'approximation

$$ {}_{s-t}q_{x+t} \approx {}_tp_x - {}_sp_x \, . $$

  Ceci entraîne que la probabilité de décéder dans l'intervalle $[a_i, b_i]$ est $(b_i - a_i)q_x$.

- On pose n_x le nombre de personnes vivantes à l'âge x, D_x et d_x la variable aléatoire du nombre de décès à l'âge x ainsi que sa réalisation et $Y_1, ..., Y_{n_x}$ les variables de Bernoulli indépendantes de paramètre ${}_{b_i-a_i}q_{x+a_i}$ représentant la vie ou le décès de chaque individu ayant été exposé au risque à l'âge x. $D_x = \sum_{k=1}^{n_x} Y_k$ est donc la somme de variables de Bernoulli indépendantes et suit donc une loi binomiale $B(n_x, {}_{b_i-a_i}q_{x+a_i})$.

En posant $Z_i = \frac{Y_i}{b_i - a_i}$, on a alors $\mathbb{E}[Z_i] = \frac{\mathbb{E}[Y_i]}{b_i - a_i}$. Or, on a vu dans les hypothèses faites précédemment que $\mathbb{E}[Y_i] = {}_{b_i-a_i}q_{x+a_i} \approx (b_i - a_i)q_x$. Ceci entraîne pour x fixé

$$ \forall i \in [\![1, n_x]\!] \, , \, \mathbb{E}[Z_i] \approx q_x. $$

En appliquant la loi des grands nombres aux Z_i (l'étape précédente ayant permis d'obtenir des variables aléatoires indépendantes de même espérance) on obtient

$$ \hat{q} = \frac{d_x}{\sum_{i=1}^{n_x} b_i - a_i} \, . $$

Remarque
Lors du calcul de cet estimateur, on obtient ainsi des effectifs par âge tenant compte de l'exposition réelle au risque. Ces effectifs seront utilisés dans la suite de l'étude.

3.3.3 Intervalles de confiance de l'estimateur de Hoem

Après avoir estimé les taux bruts, il est nécessaire de calculer des intervalles de confiance, un chiffre brut ne présumant pas la qualité de l'estimateur si on ne connaît pas l'ordre de grandeur de l'erreur possible.

Intervalles de confiance ponctuels

On va commencer par estimer un intervalle de confiance pour chaque âge sur les taux bruts, ce qui explique le terme ponctuel.

On va ainsi calculer cet intervalle pour les q_x réels à partir des $\hat{q}_x$ observés. Grâce aux hypothèses qui ont été faites sur la loi de décès de chaque individu au cours de l'âge (modélisation à l'aide d'une loi de Bernoulli) et l'approximation de la probabilité de décès d'un individu entre a_i et b_i ($_{s-t}q_{x+t} = (s-t)q_x$), on peut dans le cas où l'on dispose de suffisamment d'informations (pour cela on utilise le critère de Cochran présenté dans le chapitre 2) utiliser l'approximation gaussienne pour la loi Q_x (dont la réaisation est q_x). On a donc

$$Q_x \left(q_x, \sqrt{\frac{q_x(1-q_x)}{N_x}} \right)$$

N_x étant l'effectif sous risque calculé à l'âge x. Ceci permet d'obtenir un intervalle de confiance asymptotique exact en fonction des q_x. Cependant, comme on ne connaît pas la variance réelle obtenue à partir de ces coefficients on approxime l'intervalle de confiance de niveau α à l'aide des $\hat{q}_x$

$$I_\alpha = \left[\hat{q}_x + u_{\alpha/2}\sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{N_x}} \ , \ \hat{q}_x - u_{\alpha/2}\sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{N_x}} \right] .$$

Avec $u_{\alpha/2}$ représentant le quantile d'ordre $\alpha/2$ de la loi normale centrée réduite.

Cependant, si l'on ne dispose pas d'informations suffisantes (ce qui est le cas ici pour les âges faibles et élevés), il faut utilser l'intervalle de confiance à distance finie qui est présenté dans le livre *Modèles de durée, Applications actuarielles* de Messieurs Planchet et Thérond. .

Bandes de confiance

Il existe aussi des méthodes pour obtenir des bandes de confiance qui encadrent les taux bruts de décès sur toute une plage d'âges $[x_0, x_1]$ simultanément (x_0 et x_1 étant des âges entiers).

La méthode d'estimation de ces bandes de confiance utilisée sera celle de Sidak car elle permet d'estimer directement les bandes sur les taux de décès et non les taux de survie (dans la littérature, les estimations les plus répandues se basent sur l'estimateur de Kaplan-Meier).

On cherche ainsi la fonction c telle que $\mathbb{P}\{q_x \in \hat{q}_x \pm c(\hat{q}_x), x \in [\![x_0, x_0+n]\!]\} = 1-\alpha$. Avec n la plage sur laquelle on souhaite obtenir la bande de confiance et $1-\alpha$ le niveau de confiance souhaité.

Afin d'obtenir la bande de confiance on utilise l'hypothèse d'encadrement indépendant entre deux âges. Ainsi, si on considère un encadrement de probabilité $1-\beta$ identique pour x_0 et x_0+1, l'encadrement simultané des deux âges sera le produit des probabilités d'encadrement à chaque âge. Ainsi, mathématiquement si $\mathbb{P}\{q_x \in \hat{q}_x \pm c(\hat{q}_x), x = x_0\} = 1-\beta$ et $\mathbb{P}\{q_x \in \hat{q}_x \pm c(\hat{q}_x), x = x_0+1\} = 1-\beta$, alors, grâce à l'hypothèse d'indépendance des encadrements

$$\mathbb{P}\{q_x \in \hat{q}_x \pm c(\hat{q}_x), x \in [\![x_0, x_0+1]\!]\} = (1-\beta)^2 .$$

En continuant ainsi jusqu'à $x_0 + n$ on obtient donc l'égalité $(1-\beta)^{n+1} = 1-\alpha$ (on rapelle que $1-\alpha$ est le niveau de confiance souhaité sur toute la plage $[x_0, x_0+n]$). On a donc le résultat suivant

$$\beta = 1 - (1-\alpha)^{\frac{1}{n+1}} .$$

En reprenant l'estimation précédente des intervalles de confiance ponctuels, on obtient donc la bande confiance de niveau $1-\alpha$ sur la plage d'âges $[x_0, x_0+1]$

$$\mathbb{P}\left\{ q_x \in \hat{q}_x \pm u_{\beta/2}\sqrt{\frac{\hat{q}_x(1-\hat{q}_x)}{N_x}}, x \in [\![x_0, x_0+n]\!] \right\} = (1-\alpha) .$$

3.4 Estimateur de Kaplan-Meier

3.4.1 Principe

L'estimateur de Kaplan-Meier consiste à calculer la probabilité de survie des individus à chaque instant de décès en comptabilisant à chacun de ces moments le nombre exact de personnes présentes dans le portefeuille. Ainsi, ce modèle permet de ne faire aucune hypothèse sur la loi sous-jacente des décès, il est non paramétrique. De plus, nous verrons que les censures et les troncatures peuvent facilement être prises en compte avec cet estimateur.

3.4.2 Présentation théorique de l'estimateur

Le but étant de créer un estimateur discret (par âges), nous allons présenter le calcul effectué pour un âge x donné (la plage d'étude étant donc $[x, x+1[$). Les notations utilisées seront les suivantes :

- S_x loi discrète de la forme (t_i, s_i) , les s_i état les valeurs prises par S_x en t_i
- T_x la durée de vie résiduelle d'un individu vivant en x
- q_i la probabilité de décéder en t_i
- n_i le nombre de personnes vivantes à la date t_i
- d_i le nombre de personnes décédées en t_i
- c_{i-1} le nombre de personnes censurées sur $[t_{i-1}, t_i[$
- tr_{i-1} le nombre de personnes censurées à gauche (troncature due à l'entrée dans le portefeuille d'un individu) sur $[t_{i-1}, t_i[$.

L'estimateur de Kaplan-Meier sur l'intervalle $[x, x+1[$ consiste à calculer le produit des probabilités de décès à chaque instant. Pour justifier cela voici l'écriture mathématique

$$S(t) = \mathbb{P}(T > t | T > s)\, S(s) .$$

En développant de nouveau $S(s)$ dans cette formule, on obtient un produit de termes de la forme $\mathbb{P}(T > t|T > s)$. Comme on se trouve dans le cas d'une étude sur la mortalité, on choisit les instants t et s comme les âges de décès des individus, soit les t_i. On doit donc estimer les termes $\mathbb{P}(T > t_i | T > t_{i-1}) = p_i$. Comme $p_i = 1 - q_i$, on choisit un estimateur naturel de q_i

$$q_i = \frac{d_i}{n_i}$$

avec $n_i = n_{i-1} - d_{i-1} - c_{i-1} + tr_{i-1}$.

Ainsi, on obtient un estimateur de la fonction de survie qui se note dans notre cas (segmentation par âge)

$$\hat{S}_x(1) = \prod_{i/t_i<1} \left(1 - \frac{d_i}{n_i}\right) .$$

Ainsi, on obtient le taux brut $\hat{q}_x$ suivant

$$\hat{q}_x = 1 - \prod_{i/t_i<1} \left(1 - \frac{d_i}{n_i}\right) .$$

Remarque

Dans le cas où un décès survient au même âge qu'une censure ou une troncature, par convention on considère que le décès précède celle-ci.

3.4.3 Intervalles de confiance de l'estimateur de Kaplan-Meier

Nous allons au même titre que pour l'estimateur de Hoem calculer des intervalles de confiance ponctuels. Afin d'estimer la variance de l'estimateur de Kaplan Meier, nous allons utiliser l'estimateur de la variance de Greenwood présenté en annexes

$$\hat{V}(\hat{S}(x)) = \hat{S}(x)^2\gamma(x)^2$$

avec $\gamma(x) = \sqrt{\sum_{i,x\leq t_i\leq x+1} \frac{d_i}{N_i*(N_i-d_i)}}$. Ceci permet ainsi d'obtenir l'intervalle de confiance de l'estimateur qui est

$$I_\alpha = \left[1-(1-\hat{q}_x)\left(1+u_{1-\alpha/2}\sqrt{\sum_{i,x<t_i<x+1}\frac{d_i}{n_i(n_i-d_i)}}\right)\ ,\ 1-(1-\hat{q}_x)\left(1-u_{1-\alpha/2}\sqrt{\sum_{i,x<t_i<x+1}\frac{d_i}{n_i(n_i-d_i)}}\right)\right].$$

3.5 Mise en place du calcul des estimateurs

Les données de l'infocentre d'Axéria se trouvent sous SAS. Comme l'objectif de cette étude est qu'elle puisse être actualisée dans les prochaines années, le but est de réaliser la majorité des calculs sous SAS de façon à ce qu'une grande partie de l'étude soit automatisée. Ainsi, nous allons présenter ici comment les calculs des taux bruts ont été réalisés à partir des bases de données retraitées du chapitre 2.

3.5.1 Estimateur de Hoem

On dispose pour cet estimateur d'une base de données des personnes présentes de début 2010 à fin 2013. Pour chaque personne identifiée par une ligne nous avons aussi calculé l'âge exact d'entrée dans le portefeuille (à partir de 2010) et de sortie du portefeuille (au plus tard à fin 2013). Ainsi, il est facile d'obtenir la durée totale d'exposition au risque de chaque individu en faisant la soustraction des deux âges (ce calcul est effectué dans le but de vérifier que les âges décimaux calculés sont corrects, la durée d'exposition devant être comprise entre 0 et 4 ans).

Ensuite, le programme calcule pour chaque personne la durée d'exposition à tous les âges dans une plage définie arbitrairement : on a choisi de 18 à 92 ans. L'âge minimal a été choisi car les personnes ayant un capital sous risque avant 18 ans sont très peu nombreuses (quelques dizaines au maximum) et la probabilté de décès étant très faible, les données d'expérience n'ont pas de signification. L'âge de 92 ans correspond à l'âge maximum observé sur le portefeuille (on garde les grands âges malgré la population sous risque très faible car la probabilité de décès est beaucoup plus élevée qu'à 18 ans). Ainsi, une personne pourra être présente à cinq âges consécutifs maximum avec une durée d'exposition à chaque fois comprise entre 0 et 1. Comme il faut calculer pour chaque personne 75 nouvelles lignes, cette étape est longue et n'est pas optimisée.

Ensuite, on regroupe toutes les personnes par âges afin d'obtenir pour chacun la durée totale d'exposition du portefeuille ainsi que le nombre de décès survenus. On peut alors calculer les taux de mortalité grâce à l'estimateur présenté précédemment, ainsi que les intervalles de confiance ponctuels et les bandes de confiance. Dans la pratique les intervalles de confiance ponctuels ne sont calculés que lorsque le critère de Cochran (calculé au même moment) est supérieur ou égal à 5 (comme cela est présenté au-dessus). Les intervalles de confiance à distance finie se calculent au cas par cas mais ne peuvent pas toujours l'être car ils dépendent du nombre de décès.

3.5.2 Estimateur de Kaplan-Meier

Pour cet estimateur, on utilise la même base de données que pour l'estimateur de Hoem, il n'y a donc pas de risque que les âges calculés ainsi que les dates soient différents.

Comme cet estimateur demande de connaître à chaque instant le nombre de personnes présentes pour pouvoir calculer un taux de survie, la première étape consiste à dupliquer toutes les lignes. Ainsi chaque personne sera représentée par deux lignes dont une indiquera l'âge d'entrée dans le protefeuille et l'autre l'âge de sortie (avec l'indiquation de la cause de sortie : décès ou censure). Ensuite, on trie la table par rapport à la colonne indiquant soit un âge d'entrée, soit un âge de sortie. On obtient donc une table triée par âge (le pas des instants d'entrée ou de sortie étant 1/365 soit le jour) où chaque personne entre puis sort du portefeuille ce qui est indispensable pour réaliser le calcul de l'estimateur de Kaplan-Meier.

L'étape suivante consiste à chaque instant d'âge (36,612 par exemple) à identifier s'il y a un décès et si oui à placer la ligne en question avant toutes les autres afin de se conformer à la convention selon laquelle si un décès et une censure ou une troncature se passent au même instant, le décès précéde ces dernières. De plus, si plusieurs décès surviennent au même âge, on regroupe ceux-ci sur une seule ligne en indiquant qu'il y a eu n décès au même instant (une colonne indique 1 ou 0 pour le décès, il suffit donc de noter n à la place). Ensuite, on implémente un compteur qui a chaque ligne va être :

- majoré de 1 si la personne entre dans le portefeuille
- minoré de 1 si la personne en sort sans décéder
- minoré du nombre de décès si la cause de sortie est le décès (du fait que l'on a regroupé les décès au même instant dans une même ligne).

Afin de vérifier l'implémentation il faut vérifier que le compteur ne soit jamais négatif et qu'il finisse à 0 sur la dernière ligne. La table SAS va donc se présenter sous la forme suivante.

Personne	Age entrée	Age sortie	Age estimateur	Décès	Compteur
152	32,33	32,41	32,33	0	512
6954	32,34	34,82	32,34	0	513
7416	29,02	32,41	32,41	2	511
152	32,33	32,41	32,41	0	510

Tableau 1 : Exemple de table servant à calculer l'estimateur de Kaplan-Meier

Enfin, il suffit de calculer pour chaque ligne la probabilité de survie en indiquant 1 s'il n'y a pas de décès et $1 - \frac{\text{nombre de décès}}{\text{compteur}}$ en cas de décès. Pour calculer la probabilité de survie entre deux âges, on multiplie alors ligne à ligne la probabilité de survie qui vient d'être calculée en la réindexant à 1 à chaque changement d'âge entier. Il n'y a alors plus qu'à garder la dernière ligne de chaque âge entier, ce qui donne la probabilté de survie sur 1 an et finalement le taux de mortalité brut ($q_x = 1 - S(x)$).

3.6 Résultats des estimateurs

Dans cette partie nous allons maintenant présenter les résultats obtenus à l'aide de ces deux estimateurs. Nous allons développer les résultats pour l'obtention des taux bruts pour les hommes et les femmes, et nous présenterons succintement les résultats pour les segmentations des hommes ayant souscrit un contrat ADP standard, gros capitaux ou prévoyance professionnelle et dont le caractère non fumeur n'est pas renseigné ainsi que pour les femmes ayant les mêmes caractéristiques. Ces deux dernières segmentations seront utilisées dans la suite de l'étude à partir du chapitre 5.

3.6.1 Estimateur des taux bruts de Hoem

Les résultats obtenus avec l'estimateur de Hoem pour les hommes sur la totalité du portefeuille sont les suivants. On commence par montrer l'allure générale de la courbe obtenue.

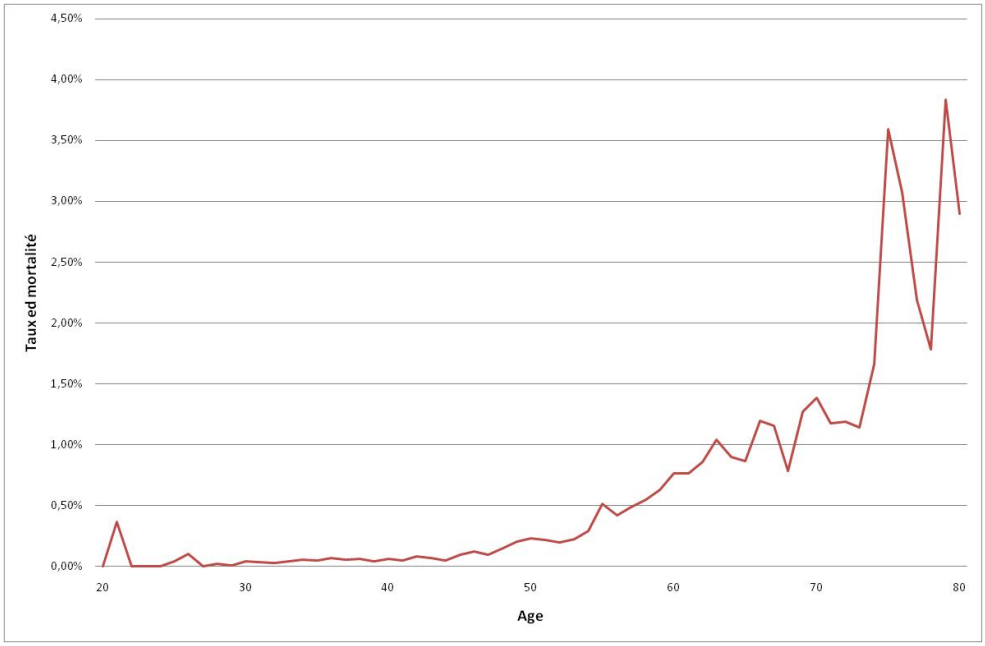


FIGURE 3.1 – Estimateur de Hoem des taux bruts de mortalité des hommes entre 20 et 80 ans

On peut voir que les taux bruts fluctuent de façon importante notamment aux âges élevés. Cela confirme donc bien que la plage d'âges avec des données suffisantes est restreinte (30 à 60 ans pour les hommes). L'allure de la courbe pour les femmes n'est pas représentée mais les fluctuations sont plus prononcées du fait d'une mortalité moins élevée pour un nombre de données similaire.

FIGURE 3.2 – Estimateur de Hoem des taux bruts de mortalité des hommes

On a représenté ici les taux bruts calculés pour les hommes de 20 à 92 ans avec les intervalles de confiance associés ainsi que les bandes de confiance. On a choisi une échelle logarithmique de façon à mieux visualiser les résultats. On peut déjà remarquer sur ce graphique que les bandes de confiance sont toujours supérieures aux intervalles, elles semblent donc plus adaptées pour cette étude.

Pour les femmes, on a représenté les résultats uniquement avec les bandes de confiance car du fait du nombre de données faible, les intervalles ponctuels ne couvraient pas une plage d'âges suffisante.

FIGURE 3.3 – Estimateur de Hoem des taux bruts de mortalité des femmes

Là-aussi nous avons représenté les résultats sur une échelle logarithmique. On peut remarquer sur ce graphique que la bande de confiance inférieure est en majorité égale à 0 ce qui confirme que le nombre de données est faible. Cependant, par hypothèse de prudence nous aurons plutôt tendance à choisir les taux qui ne sous-estiment pas la mortalité observée, la bande de confiance égale à 0 n'est donc pas trop problématique.

3.6.2 Estimateur des taux bruts de Kaplan-Meier

Nous allons présenter les résultats avec une échelle logarithmique pour les hommes et les femmes (car en échelle standard les intervalles de confiance sont difficilement lisibles pour les femmes). Les intervalles ont aussi été représentés pour les deux graphiques. Comme les résultats graphiques sont similaires aux précédents, on pourra faire les mêmes remarques.

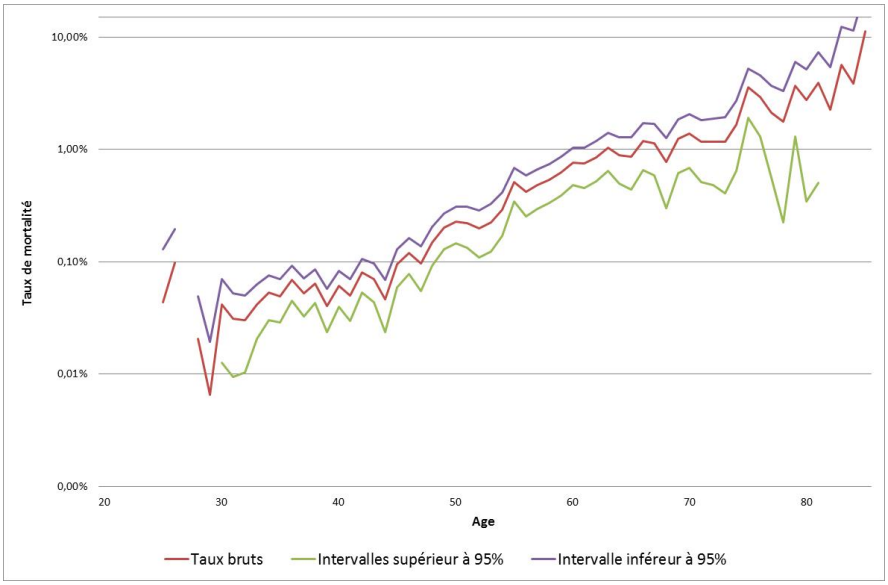


FIGURE 3.4 – Estimateur de Kaplan-Meier des taux bruts de mortalité des hommes

FIGURE 3.5 – Estimateur de Kaplan-Meier des taux bruts de mortalité des femmes

3.6.3 Comparaison des taux bruts

Comparaison des estimateurs

Afin de comparer les estimateurs, nous avons tracé le graphe des taux bruts suivant en échelle logarithmique.

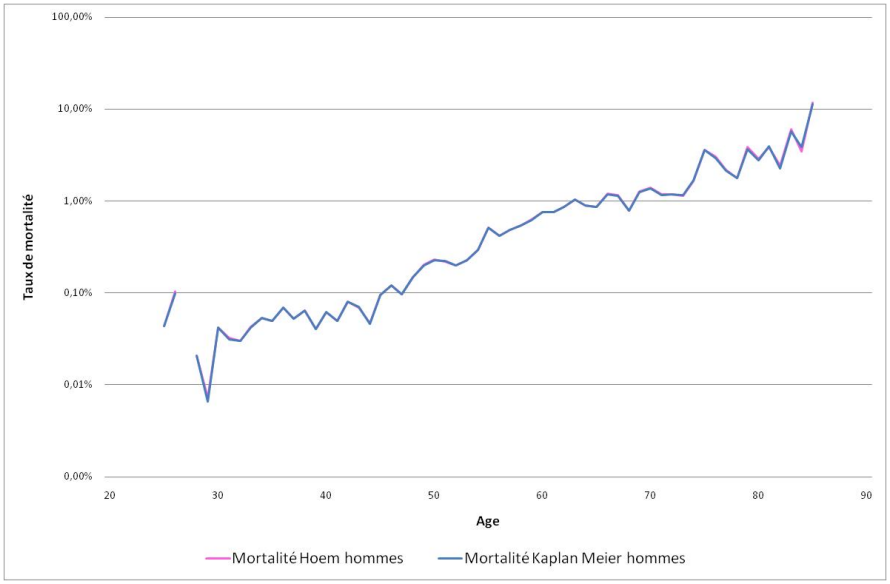


FIGURE 3.6 – Estimateurs de Kaplan-Meier et de Hoem pour les hommes

On peut tout de suite s'apercevoir que les estimateurs de Kaplan-Meier et de Hoem donnent des résultats presque identiques. Nous allons donc pouvoir choisir celui que l'on souhaite pour la suite de l'étude. Pour cela, nous avons tracé le graphique représentant les intervalles de confiance de l'estimateur de Kaplan-Meier et les bandes de confiance de l'estimateur de Hoem.

FIGURE 3.7 – Intervalles et bandes de confiance de Kaplan-Meier et de Hoem pour les hommes

On peut donc voir comme cela était prévisible que les bandes de confiance de Hoem donnent un intervalle plus large. De plus, l'estimateur de Hoem permet d'obtenir les durées d'exposition exactes au risque de mortalité pour chaque âge. Pour ces raisons, dans la suite de l'étude nous utiliserons les taux bruts obtenus à l'aide de l'estimateur de Hoem.

La courbe de mortalité retenue pour la segmentation des hommes ayant souscrit un contrat ADP standard, gros capitaux ou prévoyance profesionnelle et dont le caractère fumeur n'est pas renseigné a été obtenue de la même manière que les précédentes. L'estimateur de Hoem est donc le suivant.

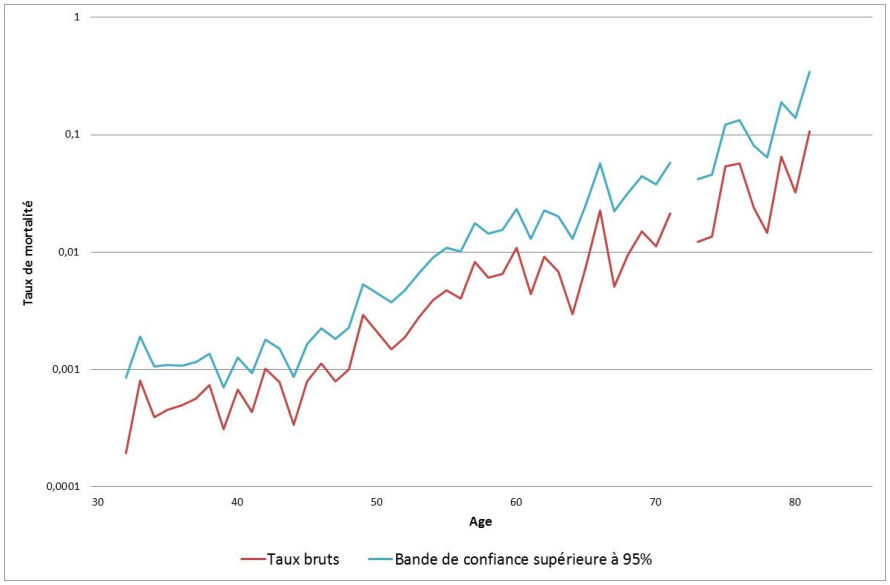


FIGURE 3.8 – Intervalles et bandes de confiance de Kaplan-Meier et de Hoem pour l'échantillon des hommes de référence

On n'a pas représenté les bandes de confiance inférieures car elles étaient très souvent égales à 0.

Comparaison de la mortalité chez les hommes et chez les femmes

FIGURE 3.9 – Comparaison des taux bruts de mortalité chez les hommes et chez les femmes

On sait que la mortalité des femmes est théoriquement deux fois moins élevée que celle des hommes. Nous avons donc tracé le graphique des taux bruts de mortalité en échelle logarithmique pour voir si cela se confirme sur la totalité du protefeuille.

On peut voir que la mortalité des hommes est bien supérieure à celle des femmes et dans des proportions à peu près similaires aux tables réglementaires. Cependant, on peut noter qu'entre 45 et 55 ans, la mortalité d'expérience du portefeuille chez les femmes est proche de celle des hommes.

Maintenant que les taux bruts ont été calculés et que l'estimateur de Hoem a été choisi, nous allons pouvoir effectuer des ajustements afin d'obtenir des courbes de mortalité en accord avec la réalité, c'est-à-dire une fonction croissante et sans cassure.

Chapitre 4

Ajustement des taux bruts

4.1 Choix des méthodes d'ajustement

Afin de lisser les courbes des taux bruts obtenues nous allons utiliser des méthodes d'ajustement et de lissage. Comme nous avons vu précédemment que nous ne disposions pas de données suffisantes en dehors d'une plage de 20 ou 30 années (selon la segmentation choisie), nous allons utiliser deux types de modèles pour obtenir des courbes lissées.

Ainsi, pour les âges où les données sont suffisantes, on peut utiliser les méthodes de lissage et d'ajustement à partir des taux bruts. Ces modèles sont divisés en deux catégories : les modèles paramétriques et ceux non paramétriques. Nous allons donc utiliser un modèle de chaque catégorie afin de pouvoir retenir celui qui est le plus adapté.

Parmi les méthodes non paramétriques, une des plus répandues est la celle de Whittaker-Henderson. Elle présente l'avantage de pouvoir modifier facilement des paramètres de régularité et de fidélité afin de s'adapter au mieux aux données brutes. Nous utiliserons ainsi cette méthode.

Parmi les modèles paramétriques, le plus couremment utilisé pour modéliser les taux de mortalité est celui de Makeham. Cependant, le modèle de Makeham suppose que la courbe $Ln(|q_{x+1} - q_x|)$ présente une tendance linéaire, nous avons donc tracé le graphe de cette courbe dont voici le résultat pour les femmes.

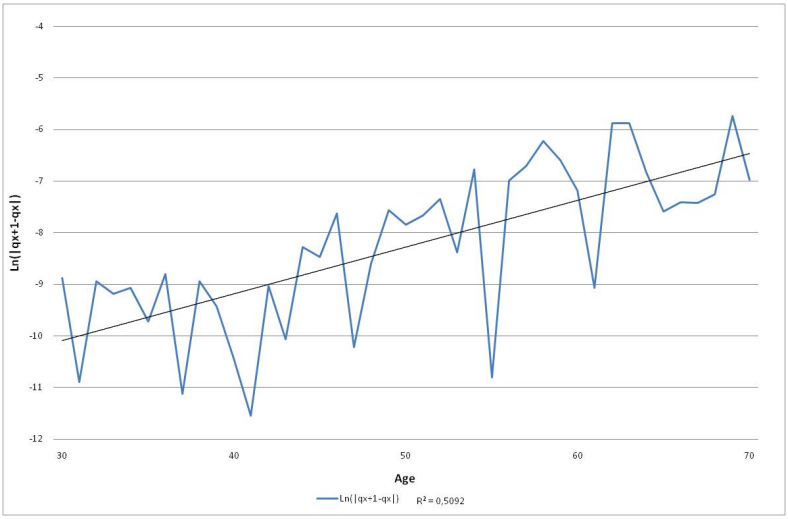


FIGURE 4.1 – Courbe des $Ln(|q_{x+1} - q_x|)$ pour les femmes servant à vérifier l'hypothèse du modèle de Makeham

On peut voir que la tendance linéaire de la courbe n'est pas forte ($R^2 = 0,52$) et ce même entre 35 et

55 ans où les données sont suffisantes. On a donc décidé de ne pas utiliser le modèle de Makeham pour les âges intermédiaires. Par conséquent, un autre modèle a été choisi, il s'agit de l'ajustement logistique qui permet de s'adapter aux changements d'incidence des courbes et qui sera présenté par la suite.

Le second type de modèle dans le cas où les données sont insuffisantes est celui d'ajustement par référence externe. Le principe consiste à positionner la courbe de mortalité d'expérience par rapport à une courbe de référence. Plusieurs méthodes existent afin de réaliser cela et nous en avons retenu deux : une méthode de régression des logits des taux bruts sur les logits d'une table de référence dont le modèle mathématique est basé sur celui des ajustements logistiques et le calcul d'un coefficient d'abattement sur une table de référence. Il faut noter qu'il existe aussi des méthodes d'extrapolation aux grands âges ne se basant pas sur des références externes. Cependant, ces méthodes sont adaptées pour des âges supérieurs à 80 ans alors que nous voulons obtenir des taux dès 60 ans. De plus, le portefeuille ne donnant pas lieu à des rentes (assurances emprunteurs en majorité ou assurances temporaires décès), le nombre de personnes exposées au risque au-delà de 80 ans ainsi que les capitaux sous risque sont très faibles comparés à l'ensemble du portefeuille.

Nous allons maintenant présenter les modèles mathématiques de ces différentes méthodes.

4.2 Méthode de lissage de Whittaker-Henderson

Le principe du modèle est de combiner linéairement deux critères de régularité et de fidélité de façon à minimiser la somme des deux.

Ainsi, le critère de fidélité est défini comme suit

$$F = \sum_{x=x_0}^{x_0+n} w_x (q_x - \hat{q}_x)^2 \, .$$

Avec n la plage d'âges sur laquelle le lissage des taux bruts est effectué et (w_x) les poids correspondants à chaque âge.

Les poids attribués le sont classiquement de deux manières. La première consiste à attribuer un poids équivalent de 1 à chaque âge. La seconde consiste à définir le poids en fonction des effectifs présents à chaque âge. On a alors $w_x = \frac{N_x}{\bar{N}}$, avec $\bar{N}$ l'effectif moyen sous risque dans la période considérée.

Le critère de régularité s'écrit de la façon suivante

$$S = \sum_{x=x_0}^{x_0+n-z} (\Delta^z q_x)^2 \, .$$

Avec z un paramètre du modèle permettant de jouer sur la régularité du lissage effectué et Δ une différence avant qui est définie comme $\Delta f(x) = f(x+1) - f(x)$.

Finalement le modèle de Whittaker-Henderson s'écrit

$$M = F + h \times S \, .$$

h étant un autre paramètre du modèle permettant de privilégier la régularité ou la fidélité.

Il s'agit donc d'un problème d'optimisation qui va être résolu grâce à la condition

$$\forall x \in [\![x_0, x_0 + n]\!], \; \frac{\partial M}{\partial q_x} = 0 \, .$$

Afin de résoudre les équations matricielles on définit les vecteurs et matrices suivants : $q = (q_x)_{x_0 \leq x \leq x_0+n}$, $\hat{q} = (\hat{q}_x)_{x_0 \leq x \leq x_0+n}$, $w = diag(w_x)_{x_0 \leq x \leq x_0+n}$ et $\Delta^z q = (\Delta^z q_x)_{x_0 \leq x \leq x_0+n-z}$. On définit alors la matrice K_z telle que $\Delta^z q = K_z q$. La matrice K_z ainsi définie est de taille $(n+1-z, n+1)$ et ses termes sont les coefficients binomiaux d'ordre z. Afin d'illustrer cela, voici les matrices K_2 et K_3 pour $n = 5$ qui seront utilisées lors de l'application de la méthode de Whittaker-Henderson :

$$K_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix} K_3 = \begin{pmatrix} -1 & 3 & -3 & 1 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 \end{pmatrix} .$$

Avec les matrices et vecteurs définis précédemment on peut donc écrire M sous la forme :

$$M =^t (q-\hat{q})w(q-\hat{q}) \ + \ h\,{}^t(K_z q)(K_z q) =^t qwq \ - \ 2\,{}^t qw\hat{q} \ + \ {}^t\hat{q}w\hat{q} \ + \ 2h\,{}^t K_z K_z q\,.$$

D'après la condition de résolution de ce problème d'optimisation, on a donc :

$$\frac{\partial M}{\partial q} = 2(w + h\,{}^t K_z K_z)q - w\hat{q} = 0\,.$$

La matrice $A = w + h\,{}^t K_z K_z$ étant inversible, on peut donc finalement obtenir le vecteur des taux ajustés

$$q = (w + h\,{}^t K_z K_z)^{-1} w\hat{q}\,.$$

4.3 Méthode d'ajustement logistique

La méthode d'ajustement logistique est basée sur l'utilisation du logit qui est défini de la manière suivante :

$$lg(q_x) = Ln\left(\frac{q_x}{1-q_x}\right)\,.$$

L'intérêt de cette transformation est d'obtenir des valeurs de $lg(x)$ qui peuvent prendre toutes les valeurs dans $\mathbb{R}$ contrairement aux q_x qui ne prennent leurs valeurs qu'entre 0 et 1. Ceci permet ensuite d'utiliser les méthodes de régression linéaire.

Ainsi le modèle d'ajustement logistique de base s'écrit

$$lg(\hat{q}_x) = ax + b + \epsilon$$

Ce modèle suppose donc que les logits des taux de mortalité ont une tendance linéaire sur la plage d'âges considérée. Or, cela n'est pas toujours vrai car en général on observe à des âges charnières x_c des modifications (qui sont des accélérations dans le cas d'un modèle de mortalité) de l'incidence de la courbe. Dans ce cas, une adaptation du modèle précédent a été proposée et est de la forme :

$$lg(\hat{q}_x) = ax + b + c(x - x_c)\mathbb{1}_{x>x_c} + \epsilon_x$$

avec ϵ_x un bruit blanc gaussien.

Ce modèle peut ensuite être généralisé, comme l'explique Daniel Serant en l'écrivant sous la forme :

$$lg(q_x) = ax + b + c(x - x_c)^{\lambda}\,\mathbb{1}_{x>x_c} + \epsilon_x$$

Afin de déterminer les paramètres, nous pouvons donc effectuer une régression avec les modèles linéaires généralisés. Il faut par conséquent vérifier que la modélisation est équivalente à celle présentée au-dessus.

Soit la variable aléatoire D_{xi} associée au décès de la personne i à l'âge x. Cette variable aléatoire prend la valeur 1 avec la probabilté q_x et 0 avec la probabilité $1-q_x$. Cette variable aléatoire suit donc une loi de Bernoulli de paramètre q_x. En considérent que les décès de toutes les personnes du portefeuille sont indépendants, la variable aléatoire D_x (dont la réalisation est notée d_x) du nombre de décès à l'âge x dans tout le protefeuille est une somme de variables aléatoires de Bernoulli indépendantes et identiquement distribuées. D_x suit donc une loi Binomiale $B(N_x, q_x)$ avec N_x l'effectif sous risque à l'âge x.

Or on peut montrer que la loi binomiale fait partie de la famille exponentielle car sa fonction de masse s'écrit

$$f(d_x) = \binom{n_x}{d_x}(q_x)^{d_x}(1-q_x)^{n_x-d_x}$$

et peut se réécrire sous la forme

$$f(d_x) = exp\left(d_x ln(\frac{q_x}{1-q_x}) + n\,ln(1-q_x) + ln\binom{n_x}{d_x}\right)\,.$$

On définit alors les fonctions suivantes : $\begin{cases} \theta = \ln(\frac{q_x}{1-q_x}) \\ b(\theta) = n\ln(1+exp(\theta)) \\ c(z,\phi) = \ln\binom{n_x}{d_x} \\ \phi = 1 \end{cases}$

On peut donc réécrire la fonction de masse de la loi binomiale sous la forme

$$f(d_x|\theta,\phi) = \exp\left(\frac{d_x\theta - b(\theta)}{\phi}\right) + c(d_x,\phi)\ .$$

On obtient donc bien la forme de la densité des lois de la famille exponentielle.

La fonction lien canonique associée à la loi binomiale étant la fonction logit, on peut, après avoir déterminé les paramètres, calculer les q_x lissés grâce à la fonction lien $\eta_x = \ln(\frac{q_x}{1-q_x})$:

$$q_x = \frac{e^{\eta_x}}{1+e^{\eta_x}}\ .$$

Remarque

Il faut noter que l'ajustement logistique est une méthode qui a tendance à sous-estimer les taux de décès, cela pouvant être à priori corrigé en effectuant une estimation par maximum de vraisemblance. Il faudra donc bien vérifier les résultats obtenus et éventuellement privilégier des taux de décès plus élevés par principe de prudence.

Nous avons donc présenté ici deux méthodes de lissage des taux bruts pour les âges intermédiaires où les données sont en quantité suffisante. Nous allons maintenant présenter les méthodes d'ajustements aux âges jeunes et élevés.

4.4 Méthodes d'ajustement par positionnement par rapport à une référence externe

4.4.1 Méthode de régression des logits ou modèle de Brass

Cette méthode se base sur l'ajustement logistique présenté juste au-dessus. Ainsi, on effectue ici une régression linéaire des taux bruts sur les logits des taux de la table de référence. Ainsi, le modèle est le suivant :

$$ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = a\ ln\left(\frac{q_x}{1-q_x}\right) + b + \epsilon_x\ .$$

Afin d'estimer les paramètres il suffit donc d'effectuer une estimation des moindres carrés. Ensuite, on peut calculer les taux ajustés de la même manière qu'avec l'ajustement logistique :

$$\tilde{q}_x = \frac{exp\left(a\ln\left(\frac{q_x}{1-q_x}\right)+b\right)}{1+exp\left(a\ln\left(\frac{q_x}{1-q_x}\right)+b\right)}\ .$$

$\tilde{q}_x$ désigne ici le taux estimé à l'aide de la régression.

4.4.2 Taux d'abattement sur une table de référence

Le modèle est ici simplifié par rapport au précédent car on cherche un coefficient a tel que

$$q_x = a\ q_x^{ref}$$

avec q_x le taux ajusté et q_x^{ref} le taux de la table de référence.

Pour estimer a plusieurs approches existent et nous avons décidé de retenir une approche de type χ^2. En effet, comme nous verrons que pour valider l'ajustement des modèles et les comparer on utilise une telle

statistique, il paraît naturel de la choisir. De plus, contrairement à d'autres méthodes qui raisonnent sur le nombre total de décès observés et prédits, cette statistique permet de raisonner en terme de distance âge par âge entre les taux observés et ceux ajustés. Ainsi, on estimera a en minimisant la distance suivante :

$$\chi^2(a) = \sum_{x=x_0}^{x_0+n} N_x \frac{(\hat{q}_x - aq_x)^2}{aq_x} .$$

La présentation théorique des méthodes d'ajustement et de positionnement par rapport à une référence externe est maintenant terminée. Nous allons donc pouvoir appliquer ces modèles à nos données de façon à obtenir des courbes de mortalité cohérentes. On rappelle que les méthodes de Whittaker-Henderson et des ajustements logistiques avec modèles linéaires généralisés vont être utilisées pour lisser les taux bruts aux âges intermédiaires (les données étant en quantité suffisante). Les méthodes de lissage par positionnement par rapport à une référence externe étant elles utilisées pour les âges jeunes et élevés.

4.5 Taux ajustés par la méthode de Whittaker-Henderson

Comme nous avons pu le voir dans la présentation théorique de la méthode, nous pouvons influer sur deux paramètres avec cette méthode afin de réaliser le meilleur ajustement en prenant en compte la régularité et la fidélité aux résultats. Cette méthode étant non paramétrique, nous avons calculé les taux lissés en essayant plusieurs valeurs de paramètres. Nous allons présenter la démarche pour choisir les paramètres qui correspondent le mieux avec la courbe de mortalité des hommes. On présentera donc uniquement les résultats retenus pour les femmes.

Méthode de choix des paramètres en se basant sur les hommes

La première étape consiste ainsi à choisir le paramètre z de la modélisation qui va permettre de jouer fortement sur la régularité de la courbe obtenue et donc d'impacter inversement la fidélité aux données brutes. Ainsi, si on choisit $z = 1$ nous obtenons une courbe qui suit très bien les données brutes mais qui par conséquent n'est pas suffisemment lissée, ce qui n'est pas réaliste pour modéliser la mortalité entre 30 et 60 ans. De même, une valeur de z égale à 4 donnera une courbe qui se rapprochera d'une droite, la fidélité aux données sera donc très mauvaise et la forme exponentielle de la droite non respectée. Les valeurs de z acceptables sont ainsi 2 et 3. Nous présentons ici le graphique permettant de choisir entre les 2 valeurs.

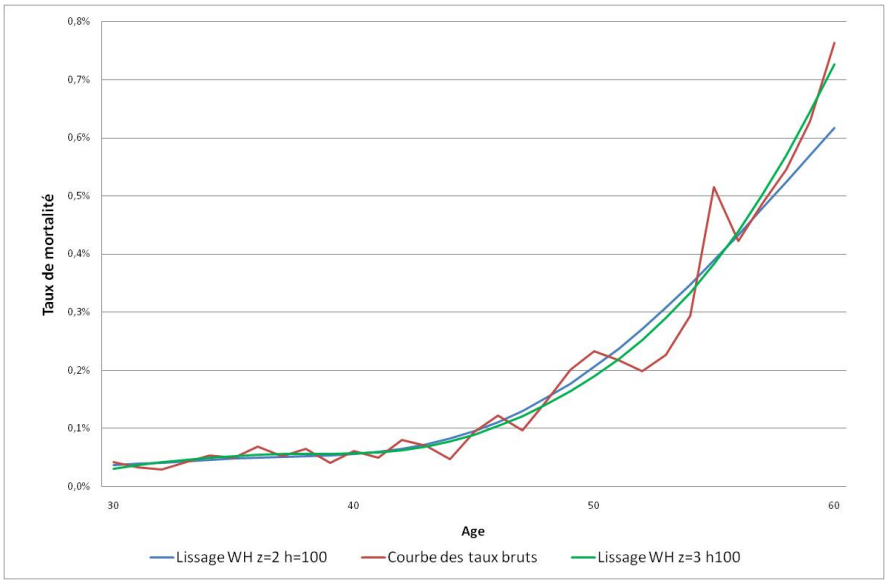


FIGURE 4.2 – Comparaison des résultats du lissage de Whittaker-Henderson selon le choix de z pour les hommes

On peut remarquer sur ce graphique que les courbes se croisent plusieurs fois sur la majorité de la plage tout en restant proches l'une de l'autre ainsi que des données. Le choix visuel paraît donc compliqué pour ces valeurs. Cependant, aux âges élevés, lorsque l'incidence de la courbe est élevée, on peut voir qu'avec $z = 2$ la courbe des taux lissés s'éloigne de façon non négligeable de la courbe des taux bruts, contrairement à la courbe des taux lissés en prenant $z = 3$. Pour cette raison nous utiliserons par la suite 3 comme valeur du paramètre z.

Le second paramètre à choisir est h. Celui-ci permet, une fois le paramètre z fixé, d'augmenter ou de diminuer l'importance du critère de régularité par rapport au critère de fidélité. Afin de choisir la valeur du paramètre nous avons testé le lissage avec des valeurs de h de 20, 100 et 500. La valeur 500 a tout d'abord été éliminée car les résultats obtenus étaient les plus extrêmes (c'est-à-dire qu'à chaque instant la courbe corespondait soit au taux le plus faible, soit au taux le plus élevé), notamment au niveau des âges les plus élevés. On a donc tracé le graphique avec des valeurs de h égales à 20 et 100.

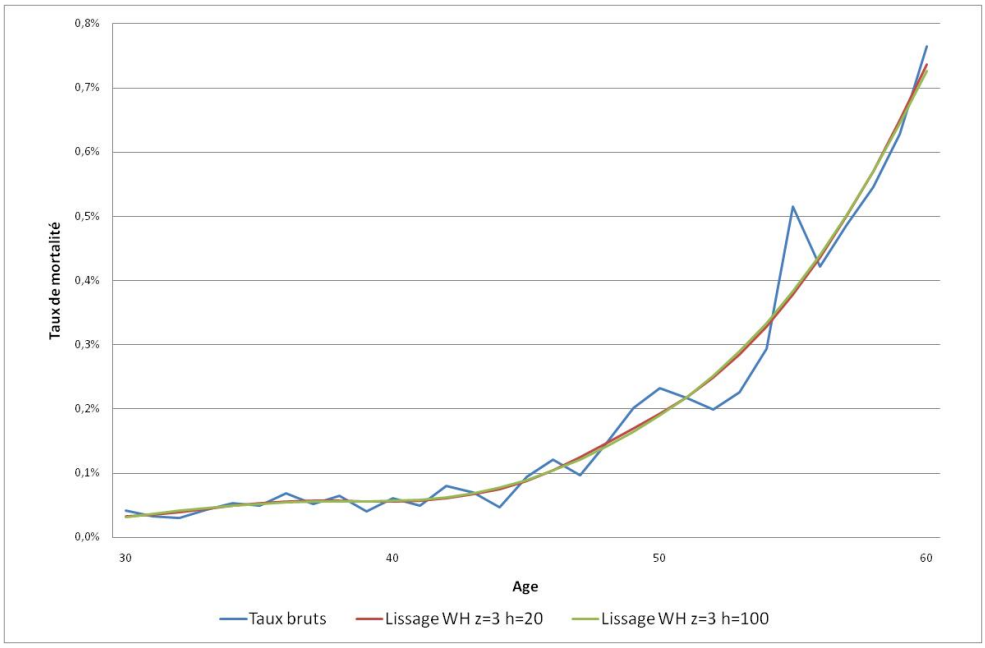


FIGURE 4.3 – Comparaison des résultats du lissage de Whittaker-Henderson selon le choix de h pour les hommes

On peut remarquer que les résultats obtenus sont très proches. De plus, à tous les âges les deux courbes semblent bien correspondre aus données brutes. Cependant, en regardant les taux de décès obtenus à chaque âge, on peut remarquer qu'avec le paramètre $h = 20$, les taux baissent très légèrement entre 37 et 40 ans ce qui ne correspond pas à la réalité. Ainsi, comme les taux stagnent au même âge avec le paramètre $h = 100$, on a décidé de choisir cette valeur de paramètre.

Résultats retenus pour les femmes et la segmentation des hommes de référence

Le lissage de la courbe des femmes a permis de retenir le paramètre $z = 3$ car comme pour les hommes le taux de mortalité aux âges les plus importants était sous-estimé par rapport aux données observées. Nous avons aussi retenu $h = 100$ car la plus grande volatilité des taux bruts pour les femmes fait que le lissage avec $h = 20$ sous-estime légèrement les taux de mortalité aux âges les plus jeunes et les plus élevés (le problème de baisse des taux de mortalité n'étant pas ici présent). Or, cette courbe des taux lissés sera ensuite utilisée et prolongée aux âges jeunes et élevés tout en sachant que les taux ne devront pas être décroissants entre deux âges consécutifs. Il y aurait donc un risque de sous-estimer les taux de mortalité sur plusieurs âges consécutifs où l'effectif sous risque est important.

Pour la segmentation des hommes de référence le paramètre $z = 3$ a été retenu pour les mêmes raisons que les hommes et les femmes (sous-estimation du taux de mortalité aux âges les plus élevés

pour $z = 2$). Par contre, pour le paramètre h nous avons retenu la valeur 500 car les trois courbes avec différentes valeurs du paramètre étaient très proches sur la quasi totalité des données exceptées pour les tous derniers âges modélisés où la courbe choisie donnait un taux légèrement inférieur qui paraissaît plus cohérent.

4.6 Méthode des ajustements logistiques et modèles linéaires généralisés

Afin d'appliquer le modèle mathématique défini et donc d'estimer les paramètres à l'aide des modèles linéaires généralisés, nous avons utilisé le logiciel SAS. La procédure GENMOD permet ainsi de calculer les coefficients à l'aide du maximum de vraisemblance et donne de nombreuses informations sur la qualité de l'ajustement du modèle.

Les critères utilisés pour définir le meilleur modèle sont la maximisation de la vraisemblance, la minimisation de la déviance ainsi que les critères AIC et BIC.

Ajustement sur les taux bruts des hommes

Pour les hommes, l'âge charnière qui a été retenu est 44 ans (plusieurs ajustements ont été effectués de façon à retenir le meilleur âge charnière). Ainsi, en comparant les sorties et en faisant varier la valeur de λ ($\lambda = 0$ correspond à $c = 0$) le meilleur modèle selon les critères définis précédemment est :

$$lg(q_x) = ax + b + c(x - x_c)\mathbb{1}_{x>x_c} .$$

Il faut noter que tous les critères précédents étaient les meilleurs pour ce modèle, il a donc été évident de choisir cette modélisation dont voici l'extrait de la sortie SAS à propos des variables explicatives et des coefficients du modèle (un exemple de sortie complète est présenté en annexes).

Algorithm converged.

Paramètres estimés par l'analyse du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée	Erreur type	Intervalle de confiance de Wald à 95 %		Khi-2 de Wald	Pr > Khi-2
Intercept	1	-9.3253	0.4662	-10.2390	-8.4115	400.10	<.0001
age	1	0.0484	0.0117	0.0254	0.0714	17.03	<.0001
agesp44	1	0.0985	0.0177	0.0639	0.1332	31.05	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

FIGURE 4.4 – Résultats de sortie de la procédure GENMOD pour les hommes

Sur ce tableau *agesp*44 correspond à la variable des âges supérieurs à 44 ans. On peut voir dans la dernière colonne que les p-value du test du Khi 2 sont inférieures à 0.01 et donc inférieures à 0.05. Ceci signifie que ces variables sont significatives et justifie donc qu'elles soient toutes gardées. Dans la seconde colonne nous obtenons ainsi les valeurs de b, a et c (dans l'ordre) qui permettent d'obtenir le graphique des taux ajustés pour les hommes.

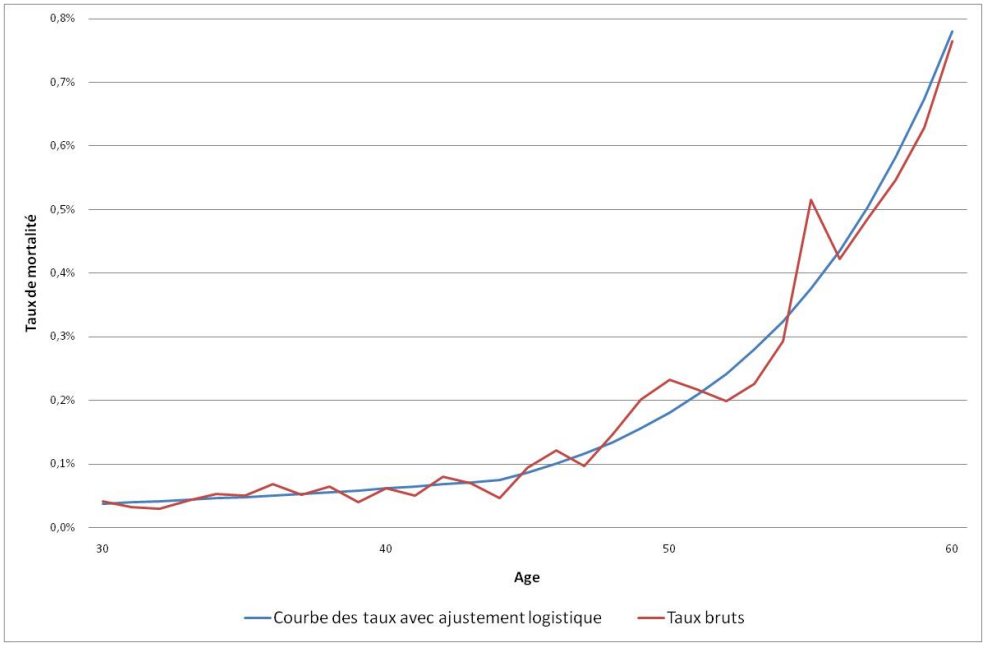


FIGURE 4.5 – Ajustement logistique avec modèles linéaires généralisés pour les hommes aux âges intermédiaires

On peut remarquer que la fidélité par rapport aux données semble bonne, cependant aux âges les plus élevés la mortalité semble être légèrement sur-estimée comparativement à la courbe des taux bruts qui est plutôt lisse à cet endroit là.

Pour la segmentation des hommes de référence on est arrivé aux mêmes conclusions et le modèle final retenu est présenté ci-dessous (on ne présente pas le graphique afin de ne pas surcharger le rapport mais il est visible en annexe 2).

$$lg(q_x) = 0,0683x - 10,127 + 0,1141(x-47)\mathbb{1}_{x>47} \, .$$

Ajustement sur les taux bruts des femmes

Nous avons effectué la même étude pour les femmes. L'âge charnière retenu est ici 45 ans après plusieurs tests. Cependant, les résultats des sortie SAS ne nous ont pas permis de retenir de façon évidente un modèle. En effet le modèle ressortant avec la meilleure déviance et la meilleure vraisemblance est

$$lg(q_x) = ax + b + c(x - x_c)\mathbb{1}_{x>x_c} \, .$$

Cependant, voici la sortie SAS que l'on obtient pour les variables.

Algorithm converged.

Paramètres estimés par l'analyse du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée	Erreur type	Intervalle de confiance de Wald à 95 %		Khi-2 de Wald	Pr > Khi-2
Intercept	1	-11.7419	0.7544	-13.2206	-10.2633	242.25	<.0001
age	1	0.0960	0.0189	0.0590	0.1330	25.89	<.0001
agesp45	1	0.0578	0.0382	-0.0171	0.1326	2.29	0.1303
Scale	0	1.0000	0.0000	1.0000	1.0000		

FIGURE 4.6 – Résultats de sortie de la procédure GENMOD pour les femmes

On voit ainsi que la p-value du Khi 2 pour la variable des âges au-dessus de 45 ans est supérieure à

0.05 ce qui signifie qu'elle n'est pas significative. Par conséquent nous avons décidé de tracer les courbes des taux ajustés des deux modèles que l'on pourrait retenir (le second étant $lg(q_x) = ax + b$) dont voici le graphique.

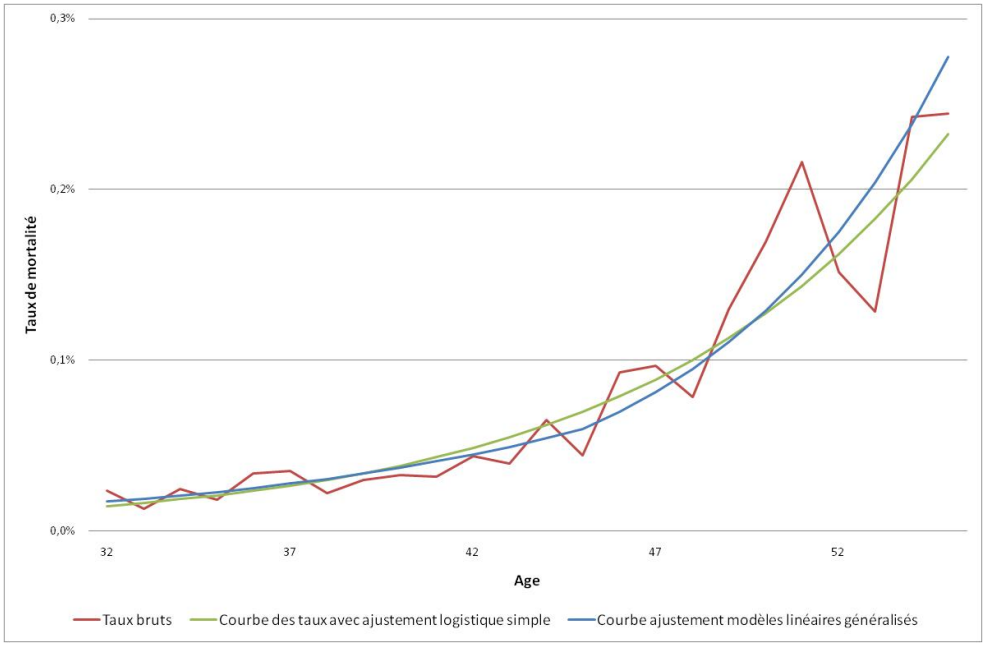


FIGURE 4.7 – Ajustements logistiques avec modèles linéaires généralisés pour les femmes aux âges intermédiaires

Ce graphique montre que l'ajustement logistique avec âge charnière semble plus fidèle aux données sur une majorité de la plage d'âges (entre 32 et 50 ans). Cependant, au-delà de 50 ans il semble que la mortalité soit sur-estimée contrairement au modèle de régression plus simple. On décide donc pour l'instant de garder les deux modèles qui seront comparés ensuite avec d'autres critères mathématiques ainsi qu'avec la méthode de Whittaker-Henderson de façon à choisir les taux ajustés retenus.

4.7 Positionnement des tables d'expérience par rapport à une référence externe

Les tables de références utilisées pour effectuer ces positionnements sont les tables réglementaires TH 00-02 et TF 00-02 car les tables Insee plus récentes présentent des taux moins lissés et donc moins en accord avec la réalité.

Afin de réaliser les positionnements par rapport à une table de référence nous avons décidé de réaliser deux ajustements pour les âges jeunes et les âges élevés. En effet, une comparaison graphique des taux bruts de mortalité avec les taux de la table de référence permet de voir que l'abattement constaté aux âges jeunes est plus élevé que l'abattement aux âges élevés. De plus, cela est aussi justifié par le fait que les âges intermédiaires sont modélisés indépendemment d'une référence externe. Ainsi, en analysant le nombre de décès ainsi que l'effectif sous risque pour chaque population observée, nous avons décidé de retenir des positionnements par rapport aux âges 30-50 ans / 55-85 ans pour les hommes, 26-50 ans / 55-82 ans pour les femmes et 32-50 ans / 55-80 ans pour la segmentation des hommes de référence. Enfin, pour la segmentation des femmes de référence, nous la ferons sur la courbe finale retenue pour les femmes sur la tranche d'âge 32-52 ans (années où les données ne sont pas trop volatiles et où la courbe des taux bruts suit une forme classique de taux de mortalité).

Nous allons présenter ici uniquement les résultats graphiques obtenus pour les hommes afin de ne pas surcharger le lecteur avec des répétitions. De plus, il faut noter qu'il a été choisi arbitrairement la courbe des taux lissés grâce à la méthode Whittaker-Henderson pour les âges intermédiaires non modélisés à

l'aide de ces méthodes.

Régression des logits sur une table de référence

Après avoir effectué une régression linéaire des logits des taux bruts sur les logits de la table de référence nous avons obtenu les modèles suivants pour les âges jeunes et les âges élevés respectivement :

$$\ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = 0,8888 \ln\left(\frac{q_x}{1-q_x}\right) - 1,965$$

$$\ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = 0,9206 \ln\left(\frac{q_x}{1-q_x}\right) - 0,9492\,.$$

Cela conduit à obtenir le graphique suivant des taux de mortalité (présenté avec une échelle logarithmique).

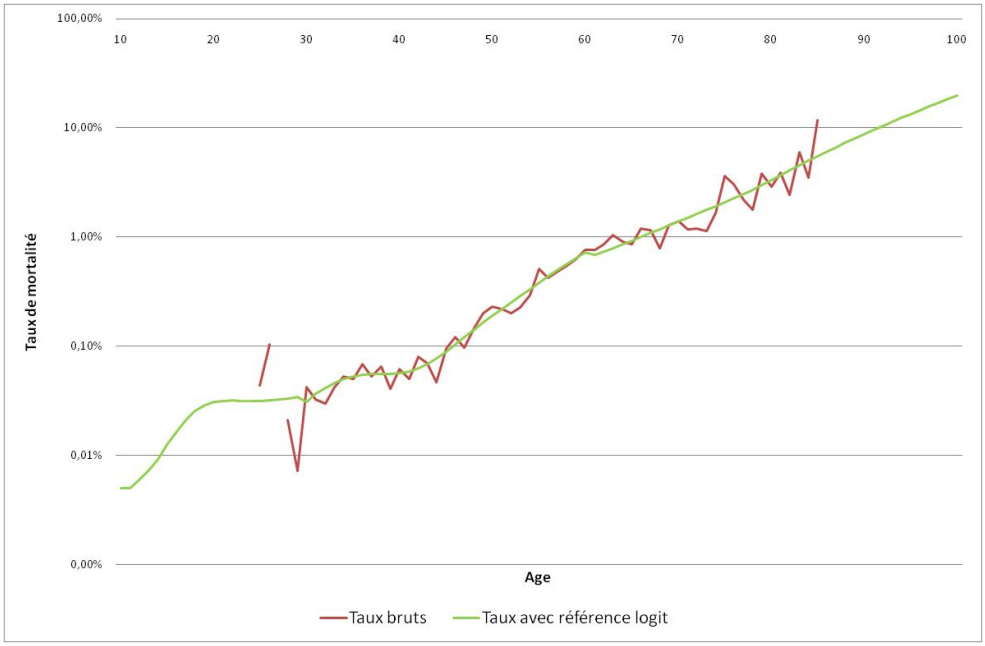


FIGURE 4.8 – Courbe des taux de décès des hommes avec régression des logits de la table TH 00-02

On peut voir sur ce graphiques des "marches" au niveau des raccordements avec les taux bruts lissés. Pour les âges jeunes, étant donné le peu de décès observés, on ne peut pas donner de conclusion pour le moment. Au niveau des âges élevés, on observe que la mortalité semble sous-estimée au niveau de l'âge de raccordement (60 ans) pour ensuite être en accord avec les données observées. Au-delà de 85 ans on ne dispose plus de données pour pouvoir estimer l'adéquation du positionnement, mais les très faibles capitaux à ces âges impliquent qu'une modélisation fine n'est pas nécessaire.

Pour les femmes les modèles retenus pour les âges jeunes et élevés sont respectivement :

$$\ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = 1,0198 \ln\left(\frac{q_x}{1-q_x}\right) - 0,8817$$

$$\ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = 0,9897 \ln\left(\frac{q_x}{1-q_x}\right) - 0,6896\,.$$

Pour la segmentation des hommes de référence nous avons obtenu les modèles suivants aux âges jeunes et élevés :

$$\ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = 0,8908 \ln\left(\frac{q_x}{1-q_x}\right) - 2,0239$$

$$\ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = 1,0152 \ln\left(\frac{q_x}{1-q_x}\right) - 0,5634\,.$$

Enfin, pour la population de référence des femmes le modèle retenu est (en rappelant que la courbe de référence est celle des femmes que l'on obtient à le fin de ce chapitre) :

$$\ln\left(\frac{\hat{q}_x}{1-\hat{q}_x}\right) = 0,9592 \ln\left(\frac{q_x}{1-q_x}\right) - 0,376\,.$$

Positionnement par rapport à une référence externe avec un coefficient d'abattement

Pour estimer au mieux le coefficient d'abattement se basant sur une méthode du Khi 2 nous avons tracé la courbe de la fonction χ^2 en fonction du paramètre a et ce pour les 2 tranches d'âges retenues.

Nous avons ainsi retenu les résultats suivants pour les âges jeunes et ceux élevés respectivement :

$$q_x^{jeunes} = 0,28\, q_x^{ref} \qquad q_x^{âgés} = 0,56\, q_x^{ref}\,.$$

Le graphique des taux ainsi calculés est le suivant.

FIGURE 4.9 – Courbe des taux de décès des hommes par calcul d'un coefficient d'abattement sur la table TH 00-02

On peut remarquer ici que les taux de décès aux âges jeunes se raccordent bien aux taux des âges intermédiaires calculés précédemment. Pour les âges élevés, on remarque aussi sur ce graphique une sous-estimation de la mortalité à 60 ans mais qui est plus faible qu'avec la méthode des logits. Ensuite, la courbe des taux estimés semble en adéquation avec celle des taux bruts avec à priori une légère surestimation des taux de décès.

Pour les femmes les valeurs des coefficients d' abattement retenus sont les suivants :

$$q_x^{jeunes} = 0,365\, q_x^{ref} \qquad q_x^{âgés} = 0,595\, q_x^{ref}\,.$$

On peut remarquer que ces coefficients sont plus élevés que ceux utilisés pour les hommes. Ainsi, on peut en conclure que l'écart relatif de mortalité entre les femmes et les hommes sur notre portefeuille est

plus faible que pour la population de référence (population française). Ceci confirme l'impression visuelle du graphique des taux bruts où les taux de référence sont aussi tracés (à la fin du chapitre précédent).

Les coefficients d'abattement des segmentations hommes et femmes de référence sont respectivement par rapport aux q_x de base utilisés :

$$q_x^{h\ jeunes} = 0,28\ q_x^{ref} \qquad q_x^{h\ âgés} = 0,645\ q_x^{base}$$
$$q_x^{f} = 1,075\ q_x^{base}\ .$$

Maintenant que nous avons obtenus les résultats de lissage, ajustement et positionnement, nous allons pouvoir valider ou non les modèles ainsi que les comparer entre eux de façon à choisir celui qui sera le plus adapté afin d'effectuer le calcul Best Estimate des provisions.

4.8 Validation et comparaison des modèles

Afin de choisir les taux de mortalité finaux, nous allons mettre en place différents tests de façon à valider l'adéquation des taux aux données ainsi qu'à choisir ce qu'on estimera être le meilleur modèle. Nous présenterons uniquement cette analyse pour les hommes et les femmes car les procédés sont les mêmes pour les segmentations de référence. Nous nous contenterons de présenter à la fin les deux courbes retenues et les choix qui ont été effectués.

4.8.1 Les critères de choix des modèles

Nous allons utiliser 3 critères afin de choisir le meilleur modèle pour chaque courbe de mortalité :
- la statistique du Khi 2
- l'analyse des résidus
- l'analyse graphique.

La statistique du Khi 2

Cette statistique permet de déterminer la distance entre les taux lissés ou ajustés et les taux bruts. Elle se définit mathématiquement de la sorte :

$$\chi^2_{modèle} = \sum_{x=x_0}^{x_0+n} \frac{(N_x\ \hat{q}_x - N_x\ q_x)^2}{N_x\ q_x}\ ,$$

$\hat{q}_x$ correspondant aux taux bruts et q_x aux taux ajustés.

Afin de savoir si le modèle est acceptable, il faut comparer cette valeur à une valeur seuil qui correspond à la distribution asymptotique du $\chi^2_{modèle}$. Dans le cadre de l'ajustement logistique qui est un ajustement paramétrique par maximum de vraisemblance, cette distribution asymptotique est un $\chi^2(n-r)$ où r représente le nombre de paramètres du modèle. Dans le cas du lissage de Whittaker-Henderson, le nombre de degrés de liberté est compliqué à estimer, par conséquent nous vérifierons l'ajustement en considérant un grand nombre de degrés de liberté (nous utiliserons 10).

Afin de valider ou non un ajustement, il faut de plus définir un degré de confiance ou p-value pour ce seuil. Il a été décidé de le fixer à 5% comme cela est classiquement utilisé pour de tels tests.

Enfin, il faut noter que cette statistique permet de comparer les modèles entre eux sur des plages définies. La plus petite valeur de $\chi^2_{modèle}$ signifiant que la courbe des taux lissés ou ajustés est la plus proche de celle des taux bruts.

L'analyse des résidus

Ce critère consiste à tracer le graphique des résidus définis par $r_x = q_x - \hat{q}_x$ et à regarder si les signes de ceux-ci sont alternés ou toujours identiques. Des résidus de même signe indiquent que les taux ajustés grâce à un modéle ont tendance à sous-estimer ou à surestimer la mortalité réellement observée. Il faut

aussi noter que ce test permet de voir si les résidus augmentent avec l'âge ou non (les résidus tels qu'ils sont définis sont en valeur absolue, ils ont donc à priori tendance à augmenter avec l'âge).

L'analyse graphique

Cette analyse peut être utilisée à la fois pour valider un modèle (à l'aide des bandes de confiance) ou pour en comparer plusieurs en regardant visuellement si certaines courbes ne sont pas acceptables sur certaines plages de données.

4.8.2 Les modèles de Whittaker-Henderson et d'ajustement logistique aux âges intermédiaires

Comme dans les parties précédentes nous ne présenteront pas tous les graphiques permettant de comparer les modèles mais nous expliquerons ce qui a été réalisé. Le test du Khi 2 pour les femmes et les hommes a donné les résultats suivants.

		Valeur X^2	Valeur seuil
Hommes	Whittaker-Henderson	24,46	31,41
	Ajustement logistique	26,74	40,11
Femmes	Whittaker-Henderson	11,40	22,36
	Ajustement logistique (y=ax+b)	16,89	32,67
	Ajustement logistique (y=ax+b+c(x-x_c))	14,36	31,41

FIGURE 4.10 – Valeurs de la statistique du khi 2

On peut donc remarquer que tous les modèles sont acceptables pour une p-value de 5%. On ne rejette donc pas l'hypothèse selon laquelle les lois des taux bruts et taux lissés ou ajustés suivent la même distribution. De plus, on peut remarquer que les valeurs de cette statistique sont meilleures pour les modèles de Whittaker-Henderson hommes et femmes. Voici à la suite le résultat pour les femmes.

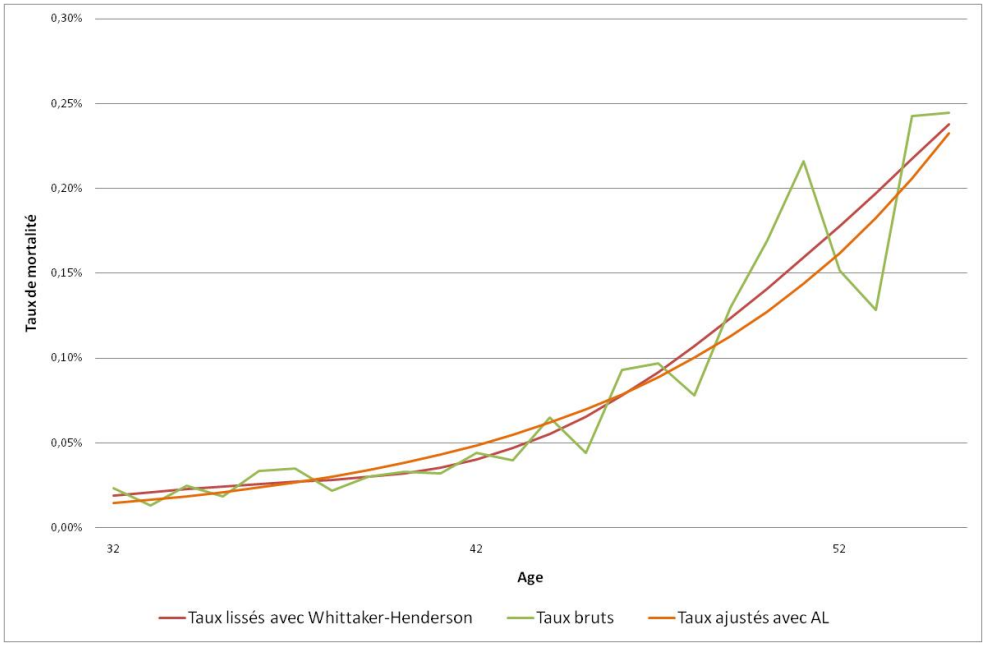


FIGURE 4.11 – Comparaison des modèles d'ajustement logistique et de Whittaker-Henderson pour les femmes

Tout d'abord, il faut noter que l'analyse graphique pour les hommes donnant des courbes proches, nous ne l'avons pas présentée. Pour les femmes, en retenant le modèle $y = ax + b$ pour l'ajustement logistique qui donne des résultats jugés plus concluants, on obtient le graphique précédent.

On peut remarquer sur ce graphique que le modèle de Whittaker-Henderson semble plus proche des données sur la plage 35-45 ans. Sur la plage 45-55 ans, le modèle d'ajustement logistique présente une courbe à tendance exponentielle plus marquée, cependant, la volatilité des données étant plus importante, la courbe de Whittaker-Henderson plus prudente ne peut être écartée.

Pour finir, l'analyse des résidus donne sensiblement les mêmes résultats pour les hommes et pour les femmes. Nous présentons ainsi le graphique des résidus des femmes.

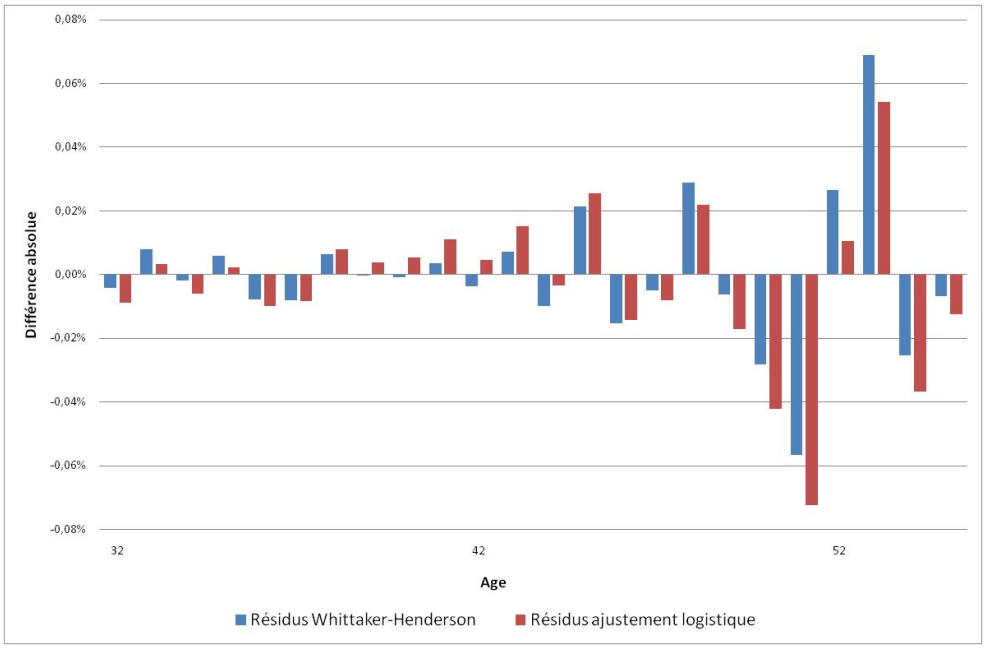


FIGURE 4.12 – Comparaison des résidus suivant les modèles pour les femmes

La première remarque que l'on peut faire est que les résidus semblent être assez proches pour les 2 modèles. De plus, la valeur absolue des résidus augmente avec l'âge ce qui est cohérent avec l'augmentation des taux de mortalité. Enfin, le signe des résidus ne reste pas identique pendant plus de 5 ans ce qui tend à valider l'adéquation des modèles aux données. Le nombre de changements de signe est de 16 pour la méthode de Whittaker-Henderson et de 12 pour celle des ajustements logistiques. Cela tend donc à privilégier le premier modèle pour ce critère.

Pour les hommes, le nombre de changements de signe est de 18 pour la méthode Whittaker-Henderson et de 13 pour l' ajustement logistique.

Grâce à ces 3 critères, nous avons décidé de retenir la méthode de Whittaker-Henderson pour le lissage des courbes des taux bruts des femmes et des hommes. En effet, pour chacune des courbes au moins 2 des critères étaient clairement meilleurs pour le modèle de Whittaker-Henderson, ce qui a incité à choisir celui-ci.

4.8.3 Choix des modèles pour les âges jeunes et élevés

Afin de choisir le meilleur modèle pour les âges où les données sont peu nombreuses voire inexistantes nous avons tracé sur un même graphique les courbes hommes et femmes avec à chaque fois les 2 modèles de positionnement (utilisation du modèle de Brass avec les logits et calcul d'un coefficient d'abattement). Il faut noter que nous avions vu précédemment (lors de l'application des modèles) que les résultats étaient acceptables dans les deux cas).

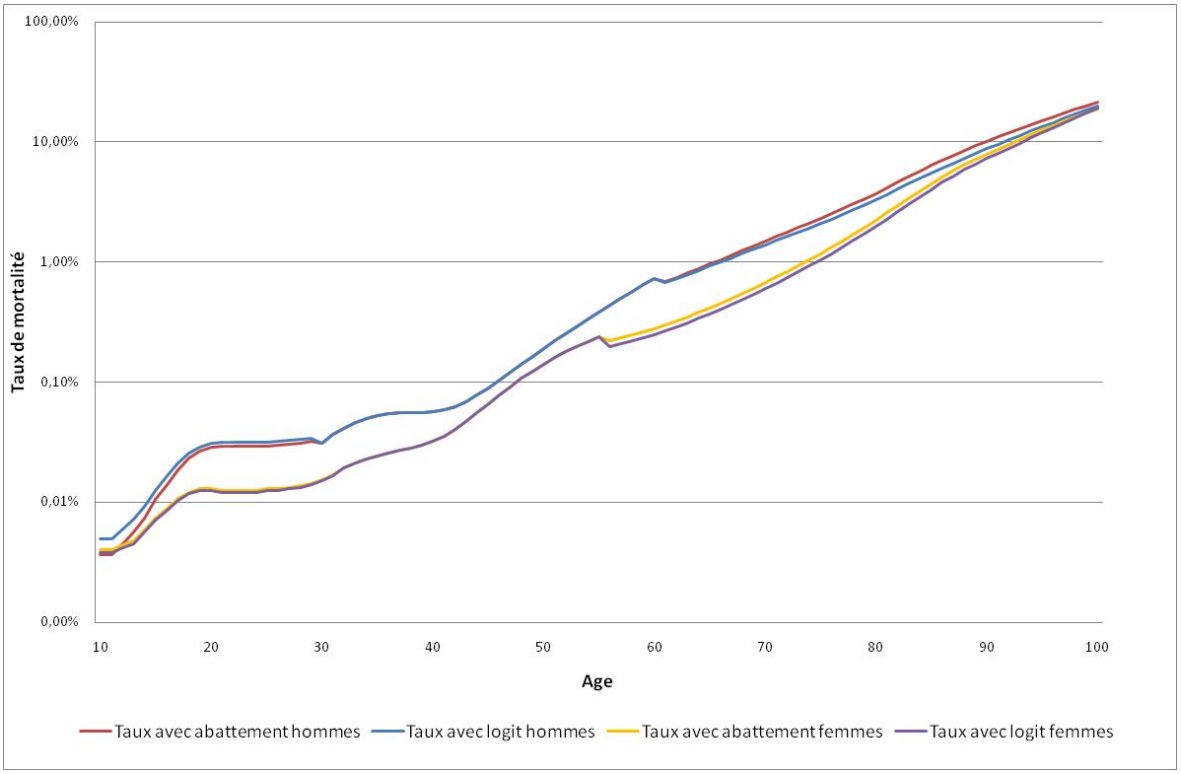


FIGURE 4.13 – Comparaison des modèles de positionnement par rapport à une référence externe

Nous pouvons voir avec ce graphique que pour les femmes, le calcul d'un coefficient d'abattement donne des mortalités plus élevées qu'avec le modèle de Brass. Pour les hommes, le modèle de Brass conduit à une mortalité plus élevée pour les âges jeunes mais plus faible aux âges élevés. Par application du principe de prudence (les données étant lacunaires), nous allons choisir les modèles qui donnent les taux de mortalité les plus élevés.

Ainsi, pour les femmes ce seront les résultats avec coefficient d'abattement qui seront retenus. Pour les hommes, nous allons aussi retenir le modèle avec taux d'abattement pour les âges jeunes et élevés. En effet, même si aux âges jeunes la mortalité est plus faible qu'avec le modèle de Brass, on peut voir qu'il existe une marche à l'endroit du raccordement. Ainsi, la mortalité est plus élevée que celle définie par lissage, et comme nous voulons faire une estimation Best Estimate des provisions, surestimer de façon plus prononcée les taux de mortalité bruts ne semble pas adéquat.

Remarque
Afin de lisser les "marches" que l'on obtient aux raccordements, on utilise une moyenne mobile aux grands âges sans modifier les taux de Whittaker-Henderson, tout en utilisant une moyenne simple des deux valeurs l'entourant pour le premier taux du raccordement. Pour les âges jeunes, les femmes n'ont pas besoin de retraitement et pour les hommes, on modifie la première valeur de Whittaker-Henderson en utilisant une moyenne glissante sur 5 ans (le changement de cette valeur est effectué par principe de prudence car il aurait fallu diminuer les taux de mortalité des 5 années précédentes dans le cas contraire).

4.8.4 Tables finales d'expérience pour les hommes et les femmes

Les courbe d'expérience hommes et femmes finalement retenues pour le calcul des provisions sont les suivantes.

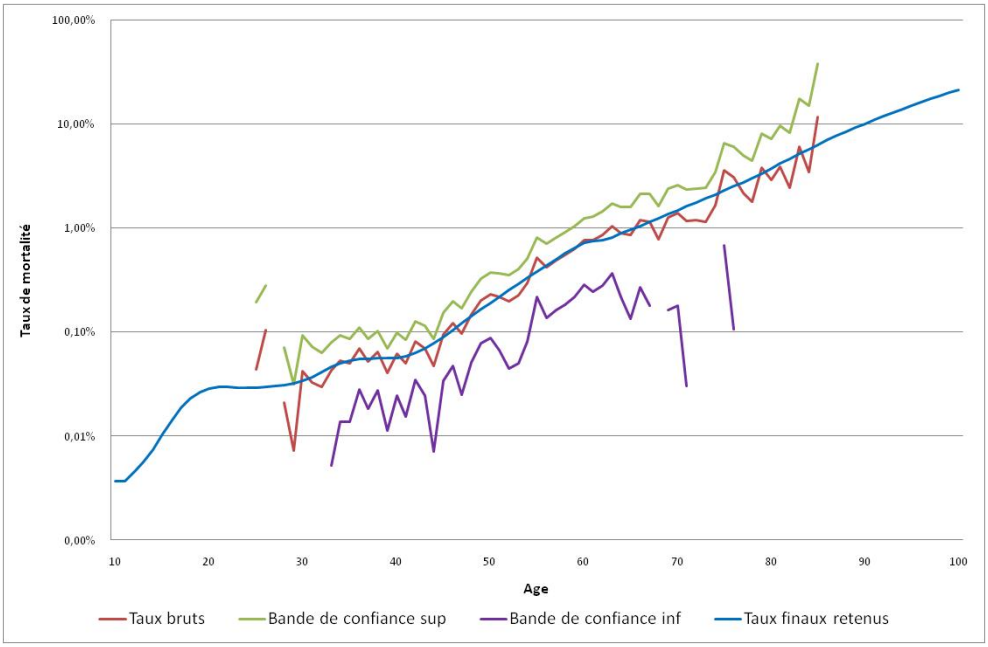


FIGURE 4.14 – Courbe de mortalité d'expérience hommes

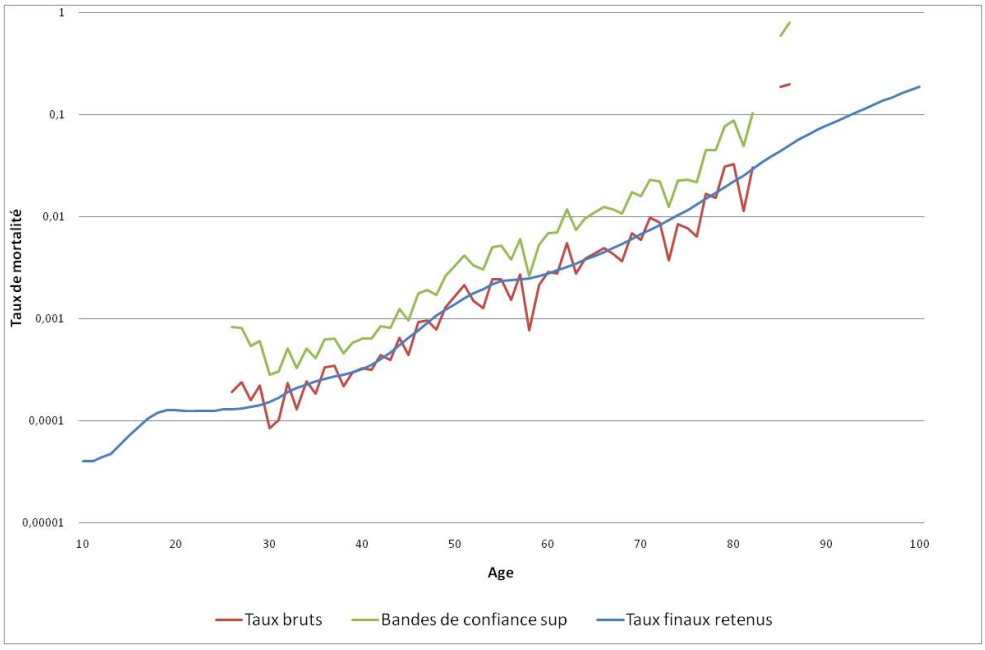


FIGURE 4.15 – Courbe de mortalité d'expérience femmes

Pour la segmentation des hommes de référence, nous avons également retenu la méthode de Whittaker-Henderson pour les âges intermédiaires. Pour les jeunes âges, le positionnement à l'aide des logit a été choisi tandis que le modèle de positionnement par abattement a été privilégié pour les âges plus élevés. Enfin, la segmentation des femmes de référence a été modélisée à l'aide de la méthode calculant un abattement sur la courbe des femmes.

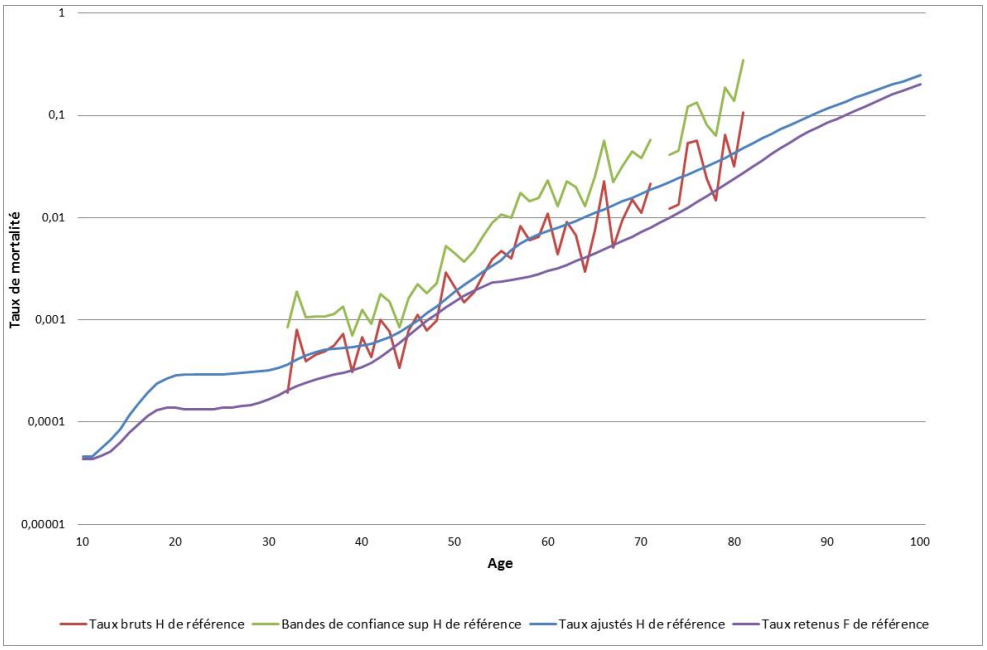


FIGURE 4.16 – Courbes de mortalité d'expérience pour les segmentations de référence

Ce graphique montre que les résultats sont cohérents car la courbe des femmes se trouve toujours en-dessous de la courbe des hommes.

Nous avons maintenant obtenu des courbes de mortalité générales. Cependant, comme on a pu le voir dans le chapitre 2, les données sont hétérogènes. Nous allons donc étudier cela dans le chapitre 5.

Chapitre 5

Modélisation de l'hétérogénéité des données

Dans les chapitres précédents, nous avons obtenu des tables de mortalité hommes et femmes regroupant tout le portefeuille. Cependant, les produits inclus dans ce portefeuille ne s'adressent pas tous aux mêmes clients. Il y a en effet des produits de prévoyance qui peuvent s'adresser à un large public alors que les produits emprunteurs sont en général vendus à des personnes ayant des revenus ou un patrimoine supérieurs à la moyenne. De même, au sein des produits emprunteurs, on a vu que certains s'adressent spécifiquement à l'assurance de gros capitaux ou aux personnes présentant un risque aggravé. Nous allons donc essayer dans ce chapitre de modéliser cette hétérogénéité des données et donc d'obtenir des courbes de mortalité plus segmentées que la séparation classique hommes et femmes.

5.1 Choix des variables explicatives avec le test du log-rank

Avant de choisir des modèles permettant de calculer des tables de mortalité segmentées, il faut définir une segmentation pertinente en utilisant des méthodes empiriques et mathématiques.

5.1.1 Choix empirique des variables à tester

Afin de faire un choix des variables que l'on va tester mathématiquement, on s'est appuyé sur les critères qui semblent explicatifs soit par connaisance globale des risques de mortalité de groupes d'individus, soit grâce à la connaisance du portefeuille d'assurés de la société.

Ainsi, la variable sexe sera testée et permettra de valider en partie la méthode mathématique utilisée car la mortalité des hommes et des femmes doit être très différente comme on l'a vu précédemment. Ensuite, on testera la variable fumeur (avec les modalités fumeur, non fumeur et non renseigné) car de nombreuses études scientifiques soulignent la diminution de l'espérance de vie des fumeurs et ce critère est largement utilisé comme élément ayant un impact tarifaire pour les contrats d'assurance garantisant le décès. Pour finir, nous nous intéresserons à la variable produits (emprunteur standard, emprunteur gros capitaux, emprunteur risque aggravé, prévoyance individuelle, prévoyance professionnelle) car les tarifs et la mortalité observés au cours des années sont disparates.

5.1.2 Test du Log-rank

Pour tester mathématiquement l'égalité ou non des distributions de mortalité de deux échantillons, de nombreux tests existent. On peut ainsi utiliser des tests paramétriques en choisissant donc une hypothèse de loi de distribution pour les mortalités. Mais il faut dans ce cas respecter l'hypothèse de normalité ce qui n'est pas le cas lorsqu'on étudie la mortalité comme on a pu le voir avec les courbes hommes et femmes obtenues dans les chapitres précédents. On va donc utiliser un type de test non paramétrique couramment utilisé et qui convient à l'étude de la mortalité : un test de rang. Celui-ci consiste à classer

les observations des deux échantillons X_a et X_b de la sorte

$$x_1 < x_2 < ... < x_n \ ,$$

puis à attribuer un score à chaque observation, avec $n = n_1 + n_2$ le nombre d'observations total des deux échantillons et les x_i représentant les instants de décès. L'hypothèse qui est ensuite testée est l'égalité de distribution des observations dans les deux échantillons. On peut noter qu'il ne doit pas y avoir d'égalité dans ce test, cependant s'il y a plusieurs décès pour un même âge la pondération sera plus forte et cela sera donc pris en compte. De plus, les échantillons (dépendant des modalités) comparés ne présentent pas de nombreux décès et la précision étant de l'ordre du jour, il y a peu de chance d'y avoir des ex-aequo.

Les test de rang les plus connus sont ceux de Wilcoxon et de Savage mais ne peuvent être utilisés en cas de censure (ils sont surtout utilisés dans le domaine de la santé). Comme dans notre cas les personnes peuvent entrer ou sortir des échantillons à tout âge, il faut utiliser un modèle prenant en compte la censure. Pour cela, on considère qu'à chaque instant de décès x_i, on a d_i décès pour un effectif e_i de personnes présentes dans le portfeuille. Si on considère que les deux échantillons étudiés a et b ont la même distribution de mortalité on doit avoir à chaque instant i :

$$d_i = d_{ia} + d_{ib} = d_i \frac{e_{ia}}{e_i} + d_i \frac{e_{ib}}{e_i} \ .$$

Ainsi, dans ce cas $d_{ia]}$ et d_{ib} ont une distribution suivant respectivement des lois hypergéométriques de paramètres $(e_i, d_i, e_{ia}/e_i)$ et $(e_i, d_i, e_{ib}/e_i)$. On peut donc en déduire leur espérance et leur variance (on ne présente les résultats que pour le groupe a) :

$$\mathbb{E}(d_{ia}) = d_i \frac{e_{ia}}{e_i}$$

$$V(d_{ia}) = d_i \frac{e_i - d_i}{e_i - 1} \frac{e_{ia} e_{ib}}{e_i^2} \ .$$

En utilisant le fait que d_{ia} suit une loi hypergéométrique on peut donc construire une statistique qui suit asymptotiquement un χ^2 à 1 degré de liberté en utilisant une pondération des $d_{ia} - \mathbb{E}(d_{ia})$. Le test du log-rank consiste à prendre 1 comme valeur de pondération. Ce test est couramment utilisé et cela se justifie dans le cas de notre étude car il y a le plus de décès dans le tranche d'âges 30-55 ans. En effet, mettre un poids plus important sur les décès à des âges jeunes par exemple pourrait nous faire conclure que les échantillons a et b ont des mortalités différentes alors que cela n'est vrai que sur une petite plage d'âges où le nombre de personnes sous risque est faible et la volatilité du nombre de décès importante. Finalement, on obtient donc la valeur de statistique suivante pour le test du log-rank :

$$X^2 = \frac{\left[\sum_{i=1}^{n} d_{ia} - \mathbb{E}(d_{ia})\right]^2}{\sum_{i=1}^{n} V(d_{ia})} \ .$$

On construit ainsi 2 statistiques pour les échantillons a et b. Comme elles doivent toutes les deux tendrent vers un χ^2 à 1 degré de liberté, on peut choisir de calculer la valeur de la statistique uniquement pour un seul des deux échantillons qui sera en pratique celui le plus petit.

5.1.3 Mise en place du test sur les données

Pour obtenir la valeur de la statistique pour chaque échantillon utilisé, il faut donc obtenir à chaque âge de décès, classés dans l'ordre croissant :

- Le nombre de décès observés dans les échantillons a et b
- Le nombre de personnes présentes dans les échantillons a et b
- Le nombre de décès théoriques dans chaque échantillon obtenu à l'aide des 2 informations précédentes.

Pour cela, on a utilisé la même méthode que pour les estimateurs des taux bruts afin de se ramener à des données où figurent uniquement les instants et le nombre de décès. On a ainsi pu calculer pour chaque i les valeurs de l'espérance et de la variance de d_{ia} et ainsi obtenir la valeur de la statistique.

5.1.4 Résultats du test et choix de segmentation

On présente ici les résultats du test d'égalité des distributions entre les échantillons dont les segmentations sont indiquées dans le tableau. Comme il y a beaucoup de segmentations différentes qui ont été testées, on ne présente ici que celles où l'on ne peut pas rejeter l'hypothèse d'égalité des distributions ainsi que les comparaisons hommes/femmes et fumeurs/non fumeurs où l'hypothèse devrait être rejetée. L'ensemble des résultats est présenté en annexe 4.

Segmentation	Valeur de la statistique	p-value	Valeur stat pour p-value
Hommes/Femmes	107,47	0,001	10,83
Fumeurs/Non fumeurs	142	0,001	10,83
ADP/ADP gros capitaux	1,58	0,21	1,57
ADP/Prev pro	0,12	0,73	0,12
ADP gros capitaux/Prev pro	3,1	0,078	3,11
ADP+Prev pro/ADP gros capitaux	1,6	0,21	1,57

FIGURE 5.1 – Tableau présentant les valeurs du test du log-rank

En choisissant un seuil de rejet pour le test à 5%, on peut donc conclure que l'on peut rejeter l'hypothèse d'égalité des distributions de mortalité pour les segmentations hommes/femmes et fumeurs/non fumeurs mais pas pour les autres qui sont présentées ici. Les résultats nous montrent donc qu'il n'est pas justifié d'effectuer une segmentation entre les produits ADP standard, prévoyance professionnelle et ADP gros capitaux. Pour les deux premiers critères, ce résultat semble cohérent car les personnes 'classiques' qui empruntent pour des montants entre 50k€ et 200k€ sont des actifs qui n'ont pas forcément des revenus très élevés. Le non rejet de l'hypothèse est par contre moins attendu pour le produit ADP gros capitaux car la sélection médicale est plus importante que pour les autres produits. Cependant, le nombre de contrats et donc de personnes sous risque et de décès est faible pour ce type d'assurance, le non rejet de l'hypothèse vient donc probablement du manque de données. Nous regrouperons donc l'ADP gros capitaux avec les deux autres produits car sinon la courbe de mortalité obtenue pour cette catégorie de personnes ne pourrait être considérée comme fiable car elle comporterait une incertitude et donc une volatilité trop importante à cause du manque de données.

Ce test nous permet donc de définir la segmentation suivante qui va être utilisée dans la suite de l'étude afin d'obtenir une courbe de mortalité pour chacune d'elles (on retient trois variables avec chacune plusieurs modalités) :

- La variable sexe : Femme / Homme
- La variable fumeur : Fumeur / Non fumeur / Non renseigné
- La variable produit : ADP standard, gros capitaux et prévoyance professionnelle / ADP risques aggravés / Prévoyance individuelle.

Nous allons donc maintenant pouvoir essayer d'estimer des courbes de mortalité pour chacune des segmentations définies précédemment.

5.2 Choix des modèles mis en oeuvre

On a maintenant défini les facteurs d'hétérogénéité que l'on voulait étudier. Il faut donc maintenant utiliser des modèles permettant de prendre en compte ces facteurs.

Une première idée aurait pu être de calculer des courbes de mortalité d'une façon identique à celles

utilisées précédemment. Cependant, ces méthodes basées sur l'estimation de taux bruts supposent d'avoir à disposition suffisamment de données (et en particulier suffisamment de décès) pour pouvoir les utiliser. Or, comme la segmantation est fine (les ADP risques aggravés ne représentent pas beaucoup de personnes comme cela a été montré dans le chapitre 2), ces méthodes ne pourront pas être utilisées.

La deuxième façon de procéder peut être de positionner les courbes de mortalité par rapport à une référence. Cela revient à calculer une courbe de mortalité pour une segmentation précise avec suffisamment de données à l'aide de la première méthode (hommes avec caractère fumeur non renseigné ayant souscrit un contrat ADP standard par exemple) puis à positionner les autre courbes par rapport à celle-ci. Nous allons donc mettre en oeuvre cette méthode à l'aide de modèles semi-paramétriques qui permettent de calculer des coefficients fixes permettant de modéliser l'influence de chaque variable tout en ayant l'avantage de ne pas définir de fonction de hasard de base (d'où la nécessité de calculer une courbe de mortalité de base qui correspond à ce que les coefficients des variables explicatives soient égaux à 0).

Comme on le verra dans la suite, cette deuxième méthode n'ayant pas permis d'obtenir tous les résultats souhaités nous utiliserons aussi une méthode paramétrique prenant en compte les variables explicatives et permettant donc d'obtenir des courbes de mortalité pour chaque segmentation. Il s'agit des modèles Accelerated Failure Time (AFT) utilisés assez couramment dans les modélisations de survie.

5.3 Modèle de Cox

5.3.1 Présentation du modèle

Le modèle de Cox est un modèle semi-paramétrique à hasard proportionnel. En effet, il consiste à calculer des coefficients fixes pour chacune des variales explicatives sans définir de fonction de hasard de base. Ceci est très utile pour étudier la mortalité car les courbes sont difficilements modélisables à l'aide d'un fonction de base classique, ce qui implique souvent un lissage trop important des particularités de l'évolution de la mortalité pour un portefeuille donné. Le terme de hasard proportionnel signifiant lui que les variables explicatives du modèle ont un impact multiplicatif sur la courbe de référence (les écarts seront donc proportionnels et non constants en valeur absolue). Ce terme s'oppose aux modèles dits additifs.

5.3.2 Théorie du modèle de Cox et hypothèses

Le modèle de Cox étant à hasard proportionnel, il s'écrit :

$$h(t|Z,\beta) = h_0(t)exp\left(\sum_{i=1}^{n}\beta_i Z_i\right)$$

où Z est le vecteur des variables explicatives, β le vecteur des coefficients de ces variables explicatives et h_0 la fonction de hasard de base que l'on aura calculée avec la courbe de mortalité de référence. On peut déjà remarquer qu'on peut écrire la fonction de hasard sous la forme :

$$Ln\frac{h(t|Z,\beta)}{h_0(t)}\,)= \beta^T Z\,.$$

Cela veut dire que le logarithme de la fonction de hasard calculée à l'aide du modèle de Cox divisée par la fonction de hasard de référence est une fonction linéaire dépendant des variables explicatives du modèle. Cela pourra se voir graphiquement avec le tracé logarithmique des courbes de mortalité obtenues dans la suite de l'étude.

Cette écriture du modèle implique aussi une hypothèse qui devra être vérifiée lors du calcul des coefficients des variables explicatives. Il s'agit de l'hypothèse de proportionnalité entre les taux de mortalité instantanés des personnes appartenant à des segmentations différentes. En effet le rapport

$$\frac{h_i(t|Z=Z_i,\beta}{h_j(t|Z_j,\beta} = exp(\beta^T(Z_i - Z_j))$$

ne dépend pas du temps, il faut donc que les formes des courbes de mortalité des différentes segmentations soient similaires, sinon le modèle de Cox ne pourra être utilisé en l'état.

5.3.3 Estimation des coefficients du modèle

Pour estimer les coefficients du modèle, le modèle de Cox consiste à calculer une vraisemblance partielle qui permet de ne pas prendre en compte h_0. Ceci permet de simplifier de façon très importante les calculs car h_0 n'est pas connu. Pour cela, on va calculer la probabilité qu'un individu i ait un événement (décès) en t_j sachant qu'un événement a eu lieu à cet instant, soit la probabilité conditionnelle :

$$\mathbb{P}(\textit{décès de i ent}_j|\textit{décès en t}_j) = \frac{\mathbb{P}(\textit{décès de i en t}_j)}{\mathbb{P}(\textit{décès en t}_j)} .$$

On suppose que $\delta_i = 1$ si l'individu i est présent au moment du décès et $\delta_i = 0$ si l'individu est censuré au moment du décès. De plus, on note R_{t_i} l'ensemble des individus présents en t_i. La vraisemblance globale du modèle avec censure s'écrit de la façon suivante (voir le cours Statistique des modèles paramétriques et semi-paramétriques de Mr PLANCHET pour le détail de l'obtention de ce résultat) :

$$L(\beta, h_0) = \prod_{i=1}^{n} [h_i(t_i)]^{\delta_i} S_i(t_i) = \prod_{i=1}^{n} \left(\left[\frac{h_i(t_i)}{\sum_{j \in R_{t_i}} h_j(t_i)} \right]^{\delta_i} \sum_{j \in R_{t_i}} h_j(t_i)^{\delta_i} S_i(t_i) \right) .$$

On montre alors que le terme $\left[h_i(t_i) / \sum_{j \in R_{t_i}} h_j(t_i) 1_{T_{ij}} \right]$ est la probabilité qu'un individu décède en t_i sachant qu'un des individus présents durant l'intervalle de temps contenant t_i est décédé. Il faut noter que pour cela on suppose qu'il n'y a pas de décès simultanés d'individus en t_i ce qui est acceptable dans le cas de notre étude. La probabilité de décéder de l'individu qui est effectivement mort en t_i est $h_i(t_i)\Delta t$ et la probabilité de n'importe quel individu présent en t_i de décéder à cet instant est $h_j(t_i)\Delta t$. Or le décès de chacun des individus est considéré comme étant indépendant, ainsi la probabilité que quelqu'un meurt en t_i est égale à la somme des probabilités des individus présents de décéder à cet instant. La probabilité de mourir en t_i sachant qu'une personne est morte est donc bien :

$$\frac{h_i(t_i)\Delta t}{\sum_{j \in R_{t_i}} h_j(t_i)\Delta t} = \frac{h_i(t_i)}{\sum_{j \in R_{t_i}} h_j(t_i)} .$$

La vraisemblance partielle du modèle de Cox ne prend ainsi en compte que le premier terme de la vraisemblance totale (le fait de ne pas connaître h_0 ne permet pas d'obtenir de maximum de vraisemblance) ce qui est justifié par ce que l'on vient de montrer et elle s'écrit :

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{h_i(t_i)}{\sum_{j \in R_{t_i}} h_j(t_i)} \right]^{\delta_i} = \prod_{i=1}^{n} \left[\frac{exp(\beta' Z_i)}{\sum_{j \in R_{t_i}} exp(\beta' Z_j)} \right]^{\delta_i} .$$

Une fois cette vraisemblance partielle obtenue, il faut résoudre le système d'équations $\frac{\partial}{\partial \beta_i} ln(L(\beta)) = 0$. Pour cela, des algorithmes numériques sont utilisés, le plus courant étant celui de Newton-Raphson qui est présenté en annexe 5.

5.3.4 Tests de validation du modèle

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

Test de significativité des coefficients

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

suivante :

$$\begin{cases} H_0 \ : \ \beta = 0 \\ H_1 \ : \ \beta \neq 0 \, . \end{cases} \tag{5.1}$$

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

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

On peut par conséquent obtenir un intervalle de confiance de niveau $1-\alpha$ (en pratique les logiciels donnent un intervalle à 95%) : $\hat{\beta} \pm u_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$.

Il faut noter que comme nous avons effectué un test du log-rank au début de ce chapitre pour choisir la segmentation adaptée, on a réduit le risque que nos variables ne soient pas significatives.

Tests de nullité globale des coefficients

Ces tests permettent de valider globalement le modèle et l'hypothèse H_0 est la nullité simulatnée de tous les coefficients. Il y a trois tests qui sont largement utilisés (et notamment par SAS que nous utiliserons dans la suite) : le rapport de vraisemblance, le test de Wald et le test du score. Ces tests sont asymptotiques et utilisent la vraisemblance. Les statistiques de ces tests sont présentées en annexe 6.

Tests de l'hypothèse des risques proportionnels

Pour vérifier la validité de cette hypothèse, on peut utiliser un test numérique ou des tests graphiques. Il en existe plusieurs mais les plus courants se basent sur les résidus de Schoenfeld. L'hypothèse de test pour chaque variable est $H_0 : \beta_i(t) = \beta_i$. En effet, l'hypothèse de proportionnalité implique que les coefficients des variables soient constants au cours du temps.

Les résidus de Schoenfeld calculent, pour chaque assuré i et chaque covariable j, à la date t_i de décès de l'individu i la différence entre la valeur d'une covariable j et la valeur attendue par le modèle qui peut s'écrire $z_{ij} - \bar{z}_{ij}(\hat{\beta}$. A partir de ces résidus, une méthode analytique permet de valider ou non cette hypothèse à l'aide d'un calcul de corrélation entre ces résidus et une fonction du temps. On peut aussi tracer les résidus de Schoenfeld normalisés sur un graphique pour chaque covariable, la courbe obtenue ne devant pas présenter de tendance temporelle.

5.3.5 Application du modèle de Cox

On a utilisé le logiciel SAS et la procédure PHREG afin de calculer les coefficients du modèle de Cox pour notre étude. On a tout d'abord choisit un modèle avec les variables sexe, produit et caractère fumeur. La fonction de hasard de base retenue a été présentée dans les chapitres 3 et 4 et il s'agit de la courbe de mortalité des individus présentant les caractéristiques suivantes :

- Homme
- Caractère fumeur non renseigné
- Produits ADP standard, gros capitaux et prévoyance professionnelle.

Cette modélisation nous a permis d'obtenir le modèle suivant :

$$\begin{aligned} h(t|Z,\beta) =& h_0(t) exp(-0,5691\mathbb{1}_{sexe=femme} + 1,1891\mathbb{1}_{produit=ADPrisquesaggravés} + 0,3341\mathbb{1}_{produit=prevind} \\ &+ 0,2511\mathbb{1}_{individu=fumeur} - 0,5631\mathbb{1}_{individu=nonfumeur}) \, . \end{aligned}$$

Nous allons commencer par regarder les résultats des tests pour ce modèle. Tout d'abord pour les tests de nullité globale des coefficients, la sortie SAS nous donne les informations ci-dessous.

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Likelihood Ratio	453.8440	5	<.0001
Score	533.2981	5	<.0001
Wald	487.4265	5	<.0001

FIGURE 5.2 – Valeurs des tests de nullité globale des coefficients

On peut donc voir que les trois tests donnent une p-value inférieure à 0,01 ce qui permet de rejeter l'hypothèse de nullité globale des coefficients (ces tests donnent de façon générale les mêmes résultats car ils sont tous basés sur la vraisemblance). Ce résultat n'est pas étonnant car on a déjà pu voir notamment dans le chapitre 2 de description des données et avec les courbes de mortalité hommes et femmes qu'il y avait des disparités importantes de taux de décès entre les populations.

Pour l'examen des coefficients du modèle, la sortie SAS nous donne les résultats suivants.

Estimations par l'analyse du maximum de vraisemblance									
Paramètre		DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque	Intervalle de confiance à 95 % du rapport de risque	Libellé
Sexe	F	1	-0.56855	0.05957	91.0886	<.0001	0.566	0.504 0.636	Sexe F
code_contrat	ADP_Solution	1	1.18939	0.08592	191.6123	<.0001	3.285	2.776 3.888	code_contrat ADP_Solution
code_contrat	Prev_ind	1	0.33418	0.07049	22.4732	<.0001	1.397	1.217 1.604	code_contrat Prev_ind
code_fumeur	Fumeur	1	0.25050	0.07680	10.6391	0.0011	1.285	1.105 1.493	code_fumeur Fumeur
code_fumeur	Non_fu	1	-0.56258	0.06111	84.7607	<.0001	0.570	0.505 0.642	code_fumeur Non_fu

FIGURE 5.3 – Sortie SAS avec la procédure PHREG pour les coefficients du modèle

La p-value du khi-2 pour toutes les variables nous indique que l'on peut rejeter l'hypothèse de non significativité de celles-ci. Cela confirme que le test du log-rank effectué précédemment nous a permis de suffisamment regrouper les variables. Il se peut par contre que le regroupement ne soit pas nécessaire pour ce modèle mais comme on l'a vu il y aurait eu un risque trop important de biais dans les résultats en particulier pour l'ADP gros capitaux. On remarque aussi que l'intervalle de confiance à 95% du rapport de risque est parfois important (de 1,2 à 1,6 pour la prévoyance individuelle par exemple) ce qui s'explique par le nombre limité de données dont nous disposons pour chaque variable. Cependant, on peut noter que ces rapports sont toujours clairement au-dessus ou en-dessous de 1 ce qui tend à montrer une cohérence des résultats selon les variables. On notera aussi que le rapport de risque de l'ADP risques aggravés est très élevé (de l'ordre de trois fois plus important), ce qui confirme que la cible de ces produits est spécifique. Enfin, le caractère fumeur à moins d'impact que celui non fumeur (augmentation du risque de 30% contre une baisse de plus de 40%), on peut donc supposer que la proportion de fumeurs parmi les personnes souscrivant un contrat depuis un dizaine d'années est moins importante que par le passé (ce qui est en accord avec la baisse globale du nombre de fumeurs en France).

La dernière validation à effectuer est l'hypothèse de proportionnalité (la plus contraignante en général). Pour obtenir les résultats des tests des résidus de Schoenfeld numériques et graphiques nous avons utilisé le logiciel R avec la fonction cox.zph qui permet de calculer les résidus de Schoenfeld normalisés puis d'obtenir soit le test numérique analytique, soit le graphique des résidus. Pour le test analytique, les résultats obtenus sont les suivants.

```
> cox.zph(out9)
                          rho   chisq     p
code_sexe           -0.00191 0.00533 0.942
code_contratADP_Solution -0.02546 0.94425 0.331
code_contratPrev_ind -0.03482 1.83126 0.176
code_fumeurFumeur    0.00309 0.01424 0.905
code_fumeurNon_fu    0.04053 2.33205 0.127
GLOBAL                    NA 5.78485 0.328
```

FIGURE 5.4 – Résultats du test analytique des résidus de Schoenfeld sous R

Ces résultats ne nous permettent pas de rejeter l'hypothèse de constance des coefficients au cours du temps avec un seuil classique à 5%. On peut donc considérer les modèle de Cox acceptable dans cette configuration. Pour les résultats graphiques, voici les sorties que nous obtenons sous R.

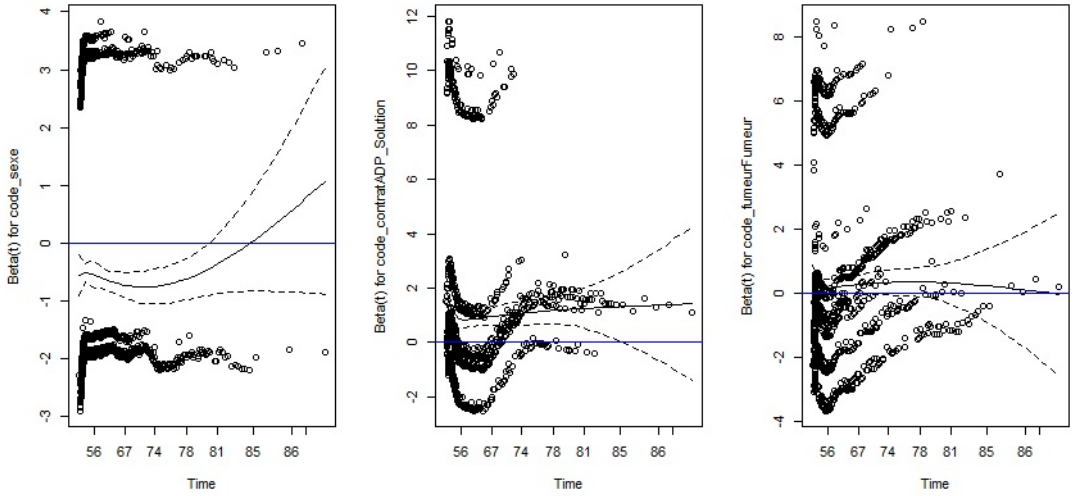


FIGURE 5.5 – Résultats du test graphique des résidus de Schoenfeld sous R

On peut tout de suite voir que pour la variable sexe une tendance temporelle semble se dessiner. Cela tendrait à rejeter le modèle car l'hypothèse de proportionnalité ne serait pas respectée. Cependant on peut voir que cette tendance n'apparaît que pour les âges élevés où il y a très peu de décès et donc un intervalle de confiance très évasé. De plus, le test numérique ne rejette pas l'hypothèse de proportionnalité.

Ainsi, même si ce modèle semble avoir des limites nous le comparerons avec un autre dans la suite. En effet, pour pallier le problème que peut poser cette hypothèse de proportionnalité on peut utiliser une variante du modèle de Cox : le modèle de Cox stratifié.

5.4 Modèle de Cox stratifié

5.4.1 Présentation du modèle

Le principe est d'ajouter des strates au modèle de Cox afin que l'hypothèse de proportionnalité n'ait plus besoin d'être respectée pour la variable sexe. Dans la pratique, cela signifie que la fonction de hasard de base ne sera pas la même selon que l'individu soit un homme ou une femme, tout en ayant les mêmes coefficients dans les deux strates pour les variables contrats et caractère fumeur. Le modèle s'écrit ainsi :

$$h(t|Z,\beta) = h_{0k}(t) exp\left(\sum_{i=1}^{n} \beta_i z_i\right) .$$

On comprend dans ce cas que l'hypothèse de proportionnalité pour la variable sexe n'a plus besoin d'être vérifiée car le modèle ne mesure pas son influence sur les autres variables dans ce cas. Cela signifie aussi qu'il faut réussir à obtenir des courbes pour deux sous-populations de base.

5.4.2 Application du modèle de Cox stratifié

Pour le modèle de Cox stratifié nous avons besoin d'une deuxième fonction de hasard de base et la population avec suffisamment de données pour obtenir une courbe de mortalité à l'aide des méthodes des chapitres 3 et 4 est la suivante :

– Femme
– Caractère fumeur non renseigné
– Produits ADP standard, gros capitaux et prévoyance professionnelle.

Toujours en utilisant la procédure PHREG sous SAS, nous avons obtenu le modèle de Cox stratifié suivant :

$$h(t|Z,\beta) = h_{0k}(t)exp(1,1891_{produit=ADPrisquesaggravés} + 0,3301_{produit=prevind} + 0,2481_{individu=fumeur} - 0,5621_{individu=nonfumeur}) .$$

On peut déjà remarquer que les coefficients ne diffèrent pas beaucoup du modèle de Cox non stratifié, cela signifiant que le sexe n'a pas d'interaction avec les autres variables. Cela peut aussi se voir en annexe 7 dans un tableau présentant la sortie SAS du modèle de Cox avec prise en compte des interactions entre les variables. Pour la validation du modèle, la sortie SAS est la suivante.

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Likelihood Ratio	336.7763	4	<.0001
Score	412.6582	4	<.0001
Wald	374.8540	4	<.0001

Estimations par l'analyse du maximum de vraisemblance										
Paramètre		DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque	Intervalle de confiance à 95 % du rapport de risque		Libellé
code_sexe		0	0	.	.	.	.	.	.	
code_contrat	ADP_Solution	1	1.18871	0.08595	191.2611	<.0001	3.283	2.774	3.885	code_contrat ADP_Solution
code_contrat	Prev_ind	1	0.33020	0.07062	21.8592	<.0001	1.391	1.211	1.598	code_contrat Prev_ind
code_fumeur	Fumeur	1	0.24816	0.07687	10.4225	0.0012	1.282	1.102	1.490	code_fumeur Fumeur
code_fumeur	Non_fu	1	-0.56159	0.06113	84.3983	<.0001	0.570	0.506	0.643	code_fumeur Non_fu

FIGURE 5.6 – Sortie de la procédure PHREG sous SAS pour le modèle de Cox stratifié

Les hypothèses de nullité globale des coefficients ainsi que de leur non significativité sont donc rejetées comme dans le cas précédent ce qui semble cohérent comme on enlève une variable. Pour l'hypothèse de proportionnalité entre variables, on ne présente que le résultat graphique qui est comparable à celui du modèle de Cox et est le suivant.

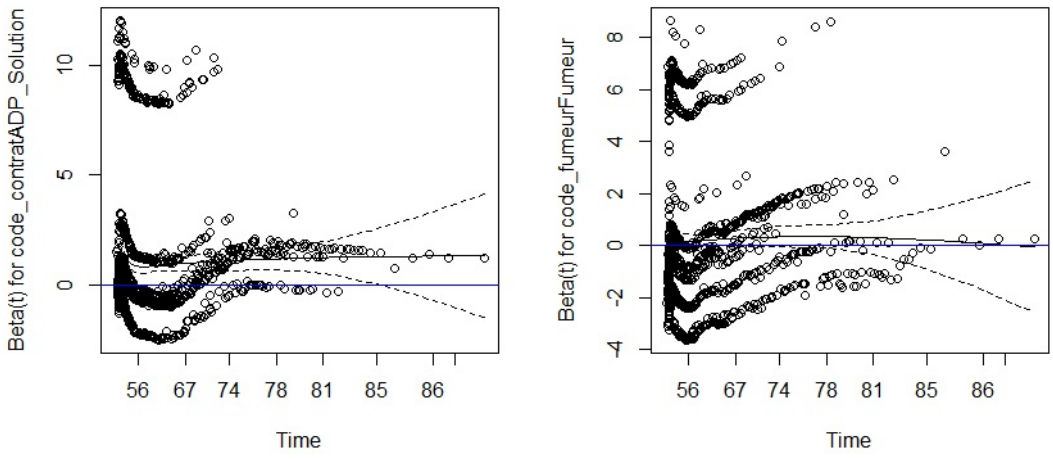


FIGURE 5.7 – Graphiques des résidus de Schoenfeld normalisés pour le modèle de Cox stratifié

Maintenant que les deux modèles ont été choisis, on va présenter les résultats puis les comparer afin de savoir lequel doit être conservé.

5.5 Résultats du modèle de Cox

Les coefficients obtenus avec le modèle de Cox vont permettre de calculer des taux de mortalité pour chaque segmentation. Pour cela on sait que :

$$q(x) = 1 - \frac{S(x+1)}{S(x)} .$$

En utilisant le fait que $h(t) = -\frac{d}{dt} ln(S(t))$, on obtient finalement en sortant les termes qui ne dépendent pas du temps de l'intégrale :

$$q_{Cox}(x) = 1 - (1 - q_0(x))^{exp(\sum_{i=1}^{n} \beta_i z_i)} .$$

On peut donc maintenant calculer les courbes de mortalité pour toutes les segmentations. Comme cela serait illisible de présenter toutes les courbes sur un même graphique on en a sélectionné certaines (qui peuvent être facilement différenciées sur un graphique) et on va présenter les résultats séparément pour les hommes et pour les femmes. A la fin de notre étude nous présenterons en annexes toutes les courbes retenues pour le calcul final des provisions. Commençons par les courbes de mortalité pour les hommes.

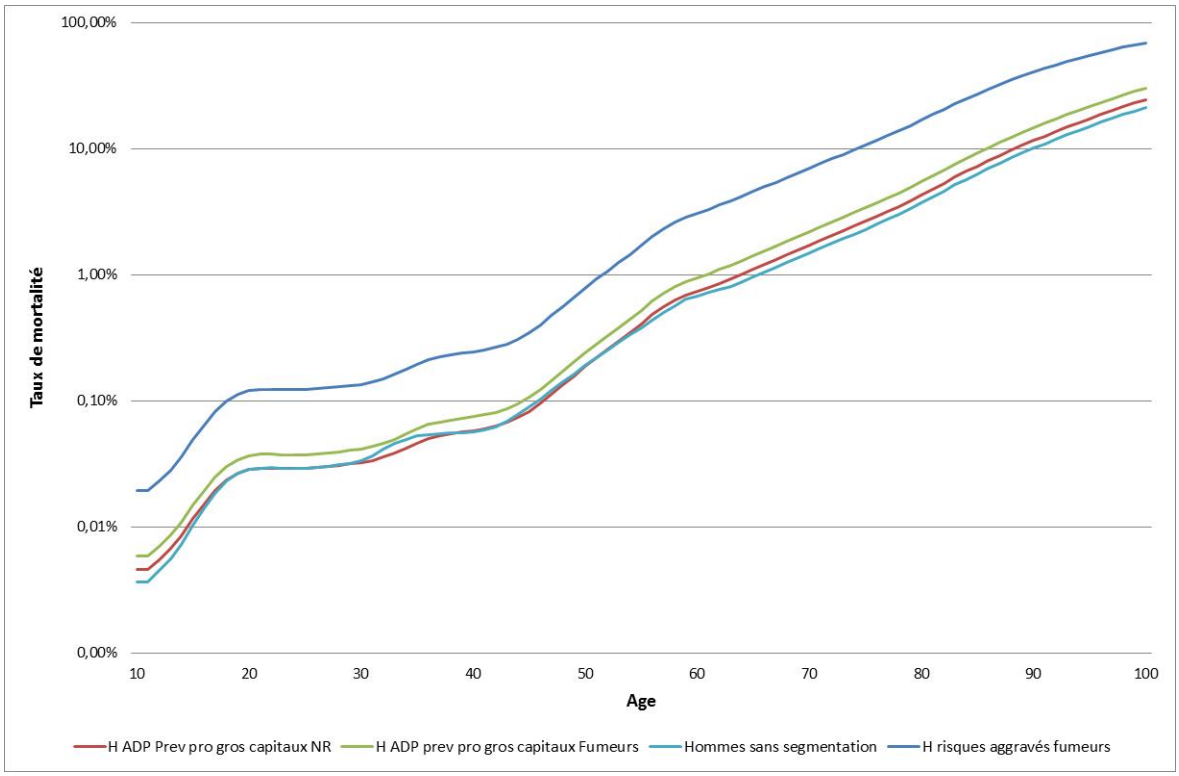


FIGURE 5.8 – Courbes de mortalité des hommes pour différentes segmentations obtenues avec le modèle de Cox

Sur ce graphique on peut déjà observer le fait que les courbes obtenues avec le modèle de Cox sont proportionnelles, ce qui normal puisqu'on fait une hypothèse de proportionnalité. Ensuite, la position des courbes est cohérente car les fumeurs ont plus de risques de décéder que les non fumeurs, de plus les hommes fumeurs et ayant un risque aggravé ont un taux de décès trois fois plus élevé que la population de référence ce qui ne semble pas incohérent au vu du rapport décès observés/décès théoriques du portefeuille global (0,29 contre 0,89), les décès théoriques provenant des tables TH-02 et TF-02.

Enfin, on peut voir que les courbes de mortalité des hommes sans segmentation et de celle de référence sont très proches sachant que la segmentation de référence regroupe un quart des assurés hommes étudiés. Ceci paraît cohérent car le caractère fumeur non renseigné est celui qui regroupe le plus de décès et se trouve être entre ceux fumeurs et non fumeurs pour le taux de mortalité.

Le graphique des résultats obtenus pour les femmes est le suivant.

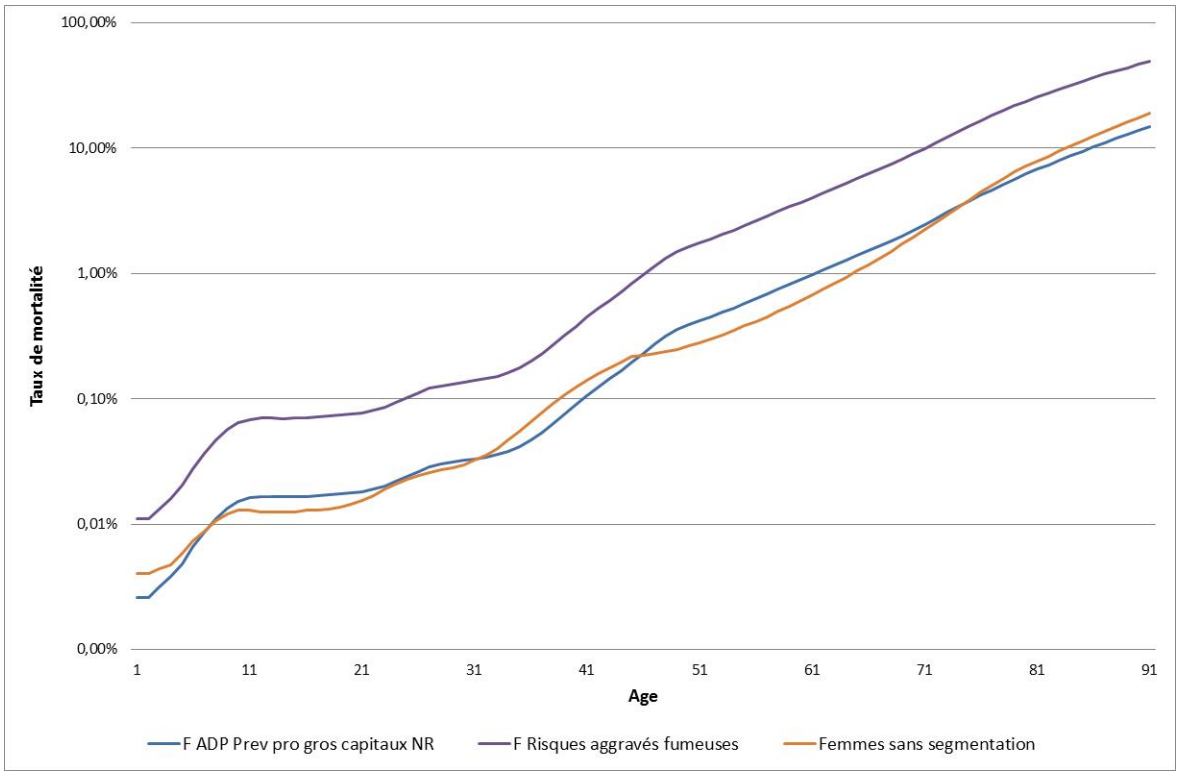


FIGURE 5.9 – Courbes de mortalité des femmes pour différentes segmentations obtenues avec le modèle de Cox

On peut effectuer le même commentaire que pour les hommes au niveau du positionnement des courbes car les variables qui augmentent le taux de mortalité sont les mêmes que pour les hommes. Pour ce qui est des courbes des femmes sans segmentation et des femmes ayant souscrit un contrat ADP standard ou gros capitaux ou encore une prévoyance professionnelle et dont le caractère fumeur n'est pas renseigné, on peut remarquer qu'elles sont similaires au niveau des taux mais diffèrent au niveau de la forme car se croisent plusieurs fois.

Comme il s'agit des mêmes segmentations que pour les hommes, on pourrait s'attendre à de courbes plus similaires. Cependant, comme la population de référence se base sur les hommes et que le modèle de Cox est à hasard proportionnel, la forme de la courbe est similaire à celle des hommes. Il sera donc intéressant de comparer ces résultats avec le modèle de Cox stratifié qui devrait corriger cela.

5.6 Résultats du modèle de Cox stratifié

Dans cette partie nous allons uniquement présenter les résultats pour les femmes car ceux des hommes sont quasiment identiques au modèle de Cox non stratifié comme la population de référence est la même.

En effet, on a pu remarquer que les coefficients des caractères fumeurs et des produits étaient presque égaux dans le modèle de Cox et le modèle de Cox stratifié. Ceci provenait de la non interaction entre les variables et nous permet donc de considérer les résultats identiques pour les hommes dans les deux modèles car la différence graphique n'est pratiquement pas visible.

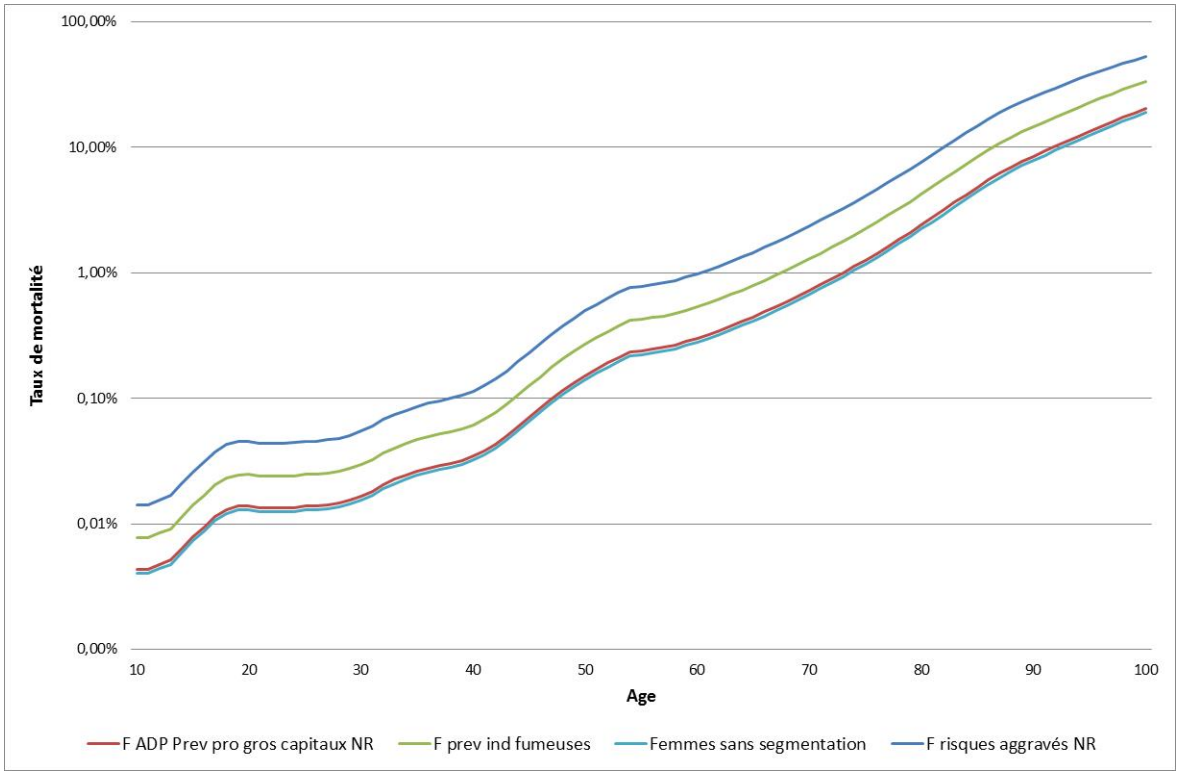


FIGURE 5.10 – Courbes de mortalité des femmes pour différentes segmentations obtenues avec le modèle de Cox stratifié

Les résultats sont bien toujours cohérents par rapport aux profils de risques. De plus, les courbes ont ici une forme presque identique à celle des femmes sans segmentation ce qui semble cohérent car la population de référence regroupe plus d'un tiers des femmes et plus de la moitié des décès. On peut noter que la courbe des femmes sans segmentation est celle qui a les taux de mortalité les moins élevés. Ceci est dû au fait que les femmes non fumeuses représentent la moitié des femmes du portefeuille, mais ont un taux de mortalité plus faible et influent donc sur la moyenne globale. Ceci permet notamment de confirmer l'intérêt d'effectuer des segmentations dans le portefeuille.

Nous venons de présenter les résultats obtenus pour les modèles de Cox et de Cox stratifié. Nous allons maintenant pouvoir les comparer afin de retenir le modèle qui sera le plus cohérent avec les données et les résultats auxquels on peut s'attendre.

5.7 Comparaison des modèles

Afin de choisir le modèle que l'on va retenir pour la suite de l'étude, on s'est intéressé aux taux modélisés pour les femmes (ceux des hommes étant identitques dans les deux cas). Ainsi, on a tracé les courbes de mortalité ainsi que la courbe des taux bruts pour la population de référence des femmes dans le modèle de Cox stratifié.

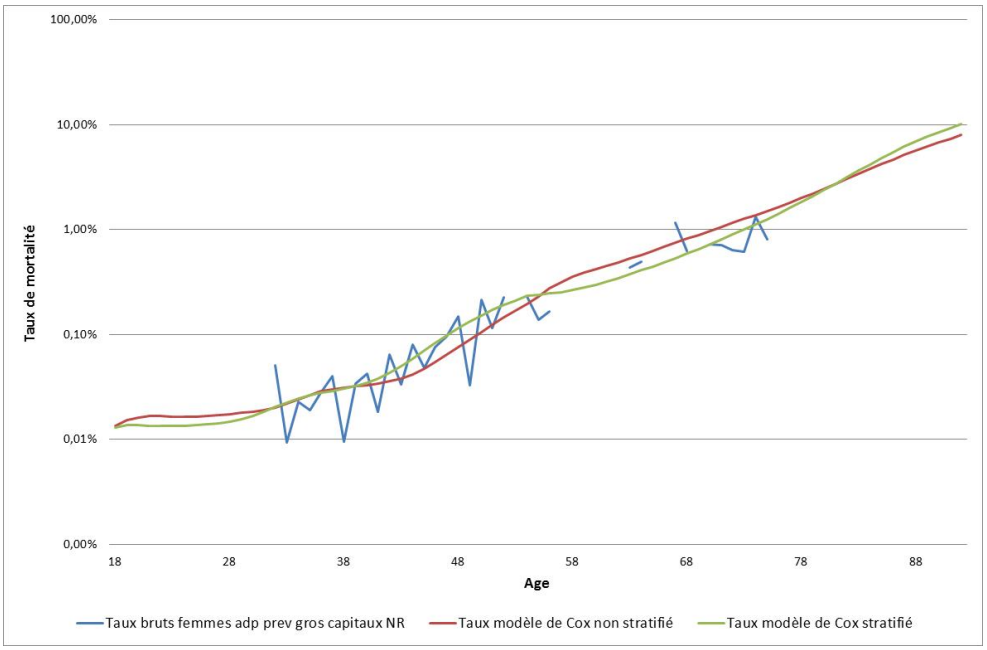


FIGURE 5.11 – Comparaison des courbes de mortalité obtenues avec les modèles de Cox et les taux bruts

Sur ce graphique on peut déjà remarquer que les deux courbes sont acceptables ce qui est intéressant car la courbe du modèle de Cox provient directement des méthodes d'estimation et de lissage des chapitres 3 et 4 alors que celle du modèle de Cox non stratifié est estimée à l'aide du modèle. Cependant, on peut voir que cette seconde courbe a tendance à sous-estimer le risque entre 40 et 55 ans, or ce sont des âges cibles pour notre étude, cela pourrait donc être problématique. Afin de pouvoir faire un choix final, nous avons tracé les courbes obtenues avec les modèles de Cox et les taux bruts pour les femmes ayant souscrit un contrat ADP standard, gros capitaux ou une prévoyance professionnelle et qui sont non fumeuses.

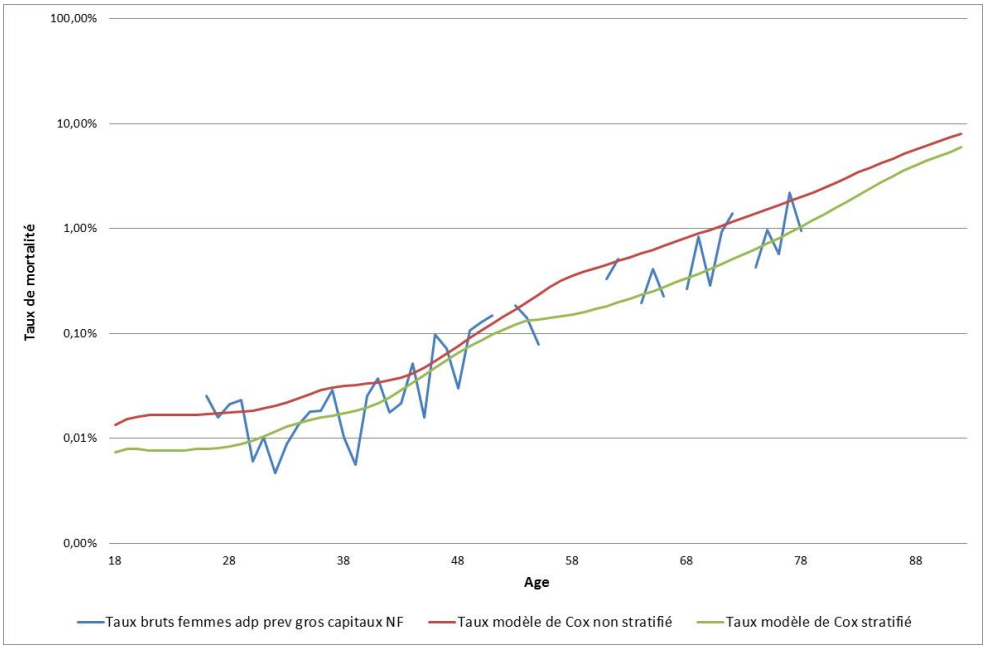


FIGURE 5.12 – Comparaison des courbes de mortalité obtenues avec les modèles de Cox et les taux bruts

Ces courbes montrent qu'hormis pour les âges inférieurs à 28 ans (où nous disposons de peu de données et très peu de décès), la courbe obtenue avec le modèle de Cox stratitifié semble très proche des taux

bruts alors que l'autre semble les surestimer, ce qui est confirmé par le test du khi 2 : la valeur de la statistique entre 30 et 51 ans, âges où les taux bruts suivent une tendance cohérente pour des taux de mortalité, est de 14,4 pour le modèle de Cox stratifié et de 25,0 pour celui qui ne l'est pas. On effectue une dernière comparaison qui peut nous orienter vers l'un ou l'autre des modèles. On compare les courbes de mortalité de la TF-002 avec celles des modèles de Cox pour la segmentation femmes avec un risque aggravé et non fumeuses.

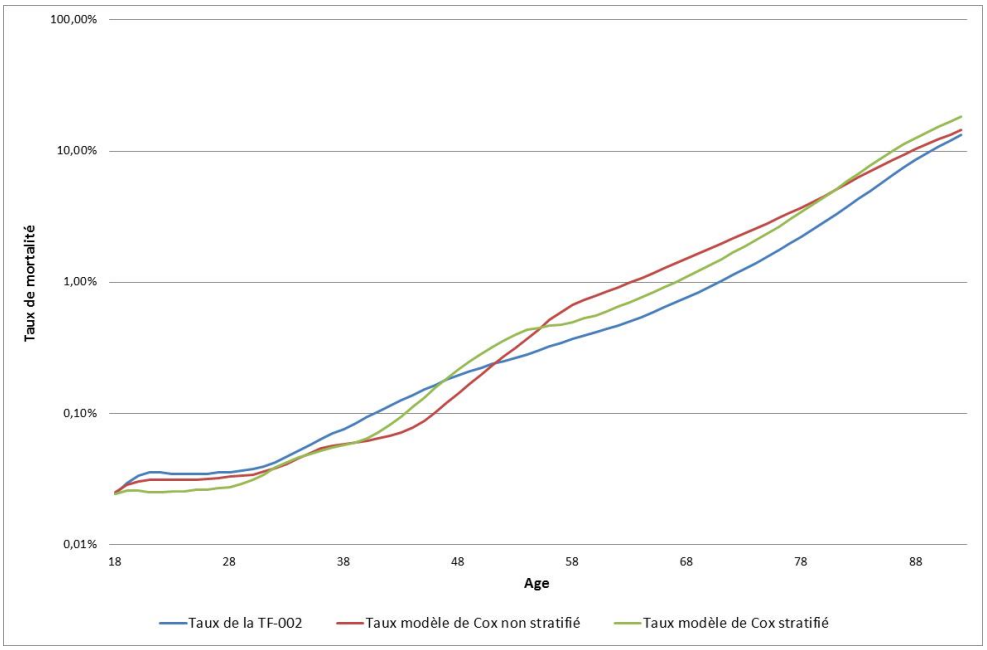


FIGURE 5.13 – Comparaison des courbes de mortalité de Cox pour la segmentation femmes avec risque aggravé et non fumeuses avec la TF-002

Ce graphique, en dehors du fait qu'il nous indique que le profil des courbes de mortalité du portefeuille est différent des courbes réglementaires, nous montre que de façon globale le modèle de Cox stratifié a un écart absolu avec la courbe réglementaire moins important que le modèle de Cox, même si ceci est assez limité (en effet la différence se fait surtout au-delà de 60 ans où la population sous risque est moins nombreuse).

Toutes ces observations nous permettent de choisir pour la suite de l'étude de ne conserver que le modèle de Cox stratifié qui donne des résultats plus proches pour chaque segmentation ce qui est cohérent par rapport au fait que l'on modélise de façon précise deux sous-populations de référence dans celui-ci.

Nous avons donc obtenu des courbes de mortalité en prenant en compte l'hétérogénéité avec une hypothèse forte de proportionnalité entre les variables. D'après les résultats cette hypothèse semble acceptable, cependant on peut se demander si, tout en gardant une part de proportionnalité entre les variables dans le modèle, on ne pourrait pas inclure une partie additive, c'est-à-dire que l'impact d'une variable soit plus constant en valeur absolue au cours du temps. Nous allons donc maintenant essayer de mettre en oeuvre le modèle de Cox-Aalen qui est un modèle semi-paramétrique additif et multiplicatif.

5.8 Le modèle de Cox-Aalen

Cox et Aalen ont développé des modèles semi-paramétriques aujourd'hui couramment utilisés dans les domaines de la santé et de l'actuariat notamment. En effet, comme on l'a vu précédemment en particulier avec le modèle de Cox, ils permettent de modéliser l'influence de plusieurs covariables de façon assez précise tout en étant moins contraignants que les modèles paramétriques. Le modèle de Cox est multiplicatif alors que celui d'Aalen est additif. Dès lors, on comprend que selon le type de données on préférera un modèle plutôt qu'un autre. Dans notre cas, de par la forme globalement exponentielle des

courbes de mortalité, on a préféré utiliser le modèle de Cox, un écart en valeur absolu de 1% aux âges jeunes ou élevés n'ayant pas du tout le même impact. Cependant, avec le modèle de Cox on est obligé d'obtenir des courbes parallèles du fait de l'hypothèse de proportionnalité. Il serait par conséquent sûrement intéressant de pouvoir implémenter de légères modifications de formes suivant les caractéristiques de la segmentation. D'où l'idée du modèle de Cox-Aalen introduit par Sheike et Zhang (2002) qui combine ces deux modèles additifs et multiplicatifs et s'écrit de la façon suivante :

$$h_i(t) = X_i(t)^T h(t) exp\left(\beta^T Z_i(t)\right) .$$

On voit bien dans cette hypothèse la séparation entre le vecteur des variables ayant un effet additif (X_i) et celui ayant un effet multiplicatif (Z_i). On a donc essayer d'implémenter ce modèle à l'aide du logiciel R qui proposait un algorithme (cox.aalen). Cependant, il semblerait que cet algorithme ne convenait pas à nos données, de plus son développement avait été effectué il y a quelques années (mis en ligne en 2008) et n'avait pas été suivi depuis. Ainsi, les résultats obtenus n'étaient pas satisfaisants du tout (la mortalité aux âges jeunes semblait bien modélisée mais augmentait très faiblement ensuite), nous avons décidé d'écarter ce modèle malgré l'intérêt qu'il peut susciter pour la prise en compte des disparités de répartition des décès suivant les segmentations.

5.9 Les modèles Accelerated Failure Time

5.9.1 Présentation du modèle

Nous venons de présenter des modèles semi-paramétriques pour modéliser l'hétérogénéité des données. Cependant, ces modèles présentent des limites et notamment le fait qu'on ne puisse calculer de maximum de vraisemblance, ce qui est un facteur limitant de l'efficience du modèle. C'est pourquoi nous avons décidé d'étudier les modèles paramétriques qui ne présentent donc pas ce problème. Parmi ces modèles, nous avons choisi de nous intéresser à ceux appelés AFT ou à durée de vie accélérée. Nous les avons choisis car les covariables ont une influence directe sur la durée de survie contrairement aux modèles à hasard proportionnel qui comme leur nom l'indique vont jouer sur la fonction de risque. Les modèles AFT peuvent s'écrire mathématiquement sous la forme suivante :

$$S(t|X) = S_0\left(\frac{t}{\eta(X)}\right) .$$

Ici, $S_0(t)$ représente la fonction de survie de base, X le vecteur des covariables de dimension n et $\eta(X)$ est défini par $\eta(X) = \sum_{i=1}^{n} \alpha_i X_i$. Cette dernière fonction est ce qui explique le terme 'accélérée' dans le modèle car suivant que la valeur soit supérieure ou inférieure à 1, le temps est décéléré et accéléré. Dans le modèle AFT, on utilise des variables aléatoires de type position-échelle. Ainsi, le modèle est usuellement écrit sous la forme :

$$log(T_i) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sigma\epsilon_i .$$

Le paramètre d'échelle étant ainsi σ, ϵ_i étant une variable aléatoire qui est définie et dont on parlera ensuite. On voit bien avec cette expression que le modèle consiste donc à mesurer l'influence des variables explicatives sur le paramètre de position. Il faut aussi noter que c'est la variable aléatoire $log(T)$ qui est de type position-échelle et donc que T est de type log-position-échelle. De plus, cette écriture du modèle amène à écrire la fonction de survie à l'aide de ces termes :

$$S_i(t) = \mathbb{P}(T_i \geq t) = \mathbb{P}(log(T_i) \geq log(t)) \tag{5.2}$$

$$= \mathbb{P}(\epsilon_i \geq \frac{log(t) - \beta_0 - \sum_{i=1}^{n} \alpha_i X_i}{\sigma}) \tag{5.3}$$

$$= S_{\epsilon_i}\left(\frac{log(t) - \beta_0 - \sum_{i=1}^{n} \alpha_i X_i}{\sigma}\right) . \tag{5.4}$$

Estimation du modèle

Le modèle étant paramétrique, il s'estime de façon classique à l'aide du maximum de vraisemblance. Dans SAS, l'algorithme de Newton-Raphson est utilisé comme pour la vraisemblance partielle du modèle de Cox afin d'estimer les valeurs des paramètres. Dans notre cas, comme les données sont majoritairement censurées à droite, la vraisemblance du modèle pourra s'écrire, en considérant l'ensemble D des individus non censurés aux temps t_i et l'ensemble C des individus censurés à droite aux temps t_i :

$$L = \prod_{i \in D} f_i(t_i) \prod_{i \in C} S_i(t_i) \, .$$

Ceci est justifié par le fait que les individus censurés à droite aux temps t_i ont une vraisemblance égale à la probabilité de durée de censure soit $\mathbb{P}(T_i > t_i) = St(t_i)$.

Présentation des distributions étudiées

Comme il s'agit d'un modèle paramétrique avec une distribution de base nous allons présenter celles que nous allons étudier. Il faut noter que la dénomination des distributions que nous allons présenter font référence à la variable T. En effet, les distributions des lois de Weibull et log-normal ne sont pas de des distributions position-échelle. Cependant quand on passe au logarithme, on obtient des distributions normales et de Gumbel qui sont bien position-échelle. Nous allons ainsi présenter quatre distributions.

*La distribution log-normale

Si T a une distribution log-normale, alors $log(T)$ est une distribution normale. On a ainsi les densités, fonction de survie et de risque pour T suivantes :

$$f(t) = \frac{1}{2\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left[\frac{log(t)-\mu}{\sigma}\right]^2\right) \tag{5.5}$$

$$S(t) = 1 - \Phi\left(\frac{log(t)-\mu}{\sigma}\right) \tag{5.6}$$

$$h(t) = \frac{\frac{1}{t\sigma}\varphi\left[\frac{log(t)-\mu}{\sigma}\right]}{\Phi\left(-\frac{log(t)-\mu}{\sigma}\right)} \, . \tag{5.7}$$

φ et ϕ représentent les fonctions de densité et de répartition de la gaussienne standard. Ce modèle peut être utilisé à la fois pour des fonctions de survie monotones mais aussi croissantes puis décroissantes.

*La distribution de Weibull

Les fonctions de densité, de survie et de risque de cette distribution sont les suivantes :

$$f(t) = \gamma\alpha t^{\gamma-1} exp(-\alpha t^{\gamma}) \tag{5.8}$$

$$S(t) = exp(-\alpha t^{\gamma}) \tag{5.9}$$

$$h(t) = \gamma\alpha t^{\gamma-1} \, . \tag{5.10}$$

Selon la valeur de γ, la fonction de risque est monotone croissante, décroissante ou constante. Ce dernier cas particulier correspond à la distribution exponentielle que nous allons aussi étudier.

*La distribution log-logistique

Dans ce cas, de la même manière que pour la distribution log-normale, $log(T)$ est une distribution logistique. Les fonctions de densité, survie et risque se présentent de la façon suivante :

$$f(t) = \frac{\gamma\alpha t^{\gamma-1}}{(1+\alpha t^{\gamma})^2} \tag{5.11}$$

$$S(t) = \frac{1}{1+\alpha t^{\gamma}} \tag{5.12}$$

$$h(t) = \frac{\gamma\alpha t^{\gamma-1}}{1+\alpha t^{\gamma}} \,. \tag{5.13}$$

La fonction de risque est décroissante monotone pour une valeur de γ inférieure ou égale à 1 puis va être non monotone pour des valeurs supérieures.

5.9.2 Application des modèles AFT

Maintenant que nous avons présenté la théorie de ces modèles, nous allons utiliser SAS et la procédure LIFEREG afin de calculer des coefficients pour nos variables. Les données dont nous disposons n'ont pas permis de créer un modèle simple à l'aide de cette procédure. Par conséquent, nous avons listé tous les âges qui sont des covariables pour lesquelles la valeur 1 ou 0 est associée à chaque individu (1 s'il est présent et 0 sinon). De plus, nous avons indiqué si l'individu est censuré ou non. Ainsi, la procédure sous SAS nous donne des résultats qui estiment un coefficient à chaque âge ainsi que les coefficients des covariables qui nous intéressent pour la modélisation. Nous présenterons donc dans la suite uniquement les coefficients des covariables qui nous intéressent.

Comme indiqué plus haut, nous avons testé quatre distributions de base pour notre modélisation : la distribution exponentielle, celle de Weibul, la distribution log-normale et enfin celle log-logistique. Afin de choisir parmi ces modèles celui le plus adapté à nos données, nous avons tout d'abord comparé les critères AIC et BIC, les plus petites valeurs correspondant au meilleur ajustement du modèle.

	AIC	BIC
Log-normale	24109	25125
Weibull	24080	25095
Exponentielle	24084	25086
Log-logistique	24079	25095

FIGURE 5.14 – Valeurs AIC et BIC obtenues pour les modèles à l'aide de la procédure LIFEREG

On voit tout de suite que les deux critères ne sont pas minimisés par le même modèle. On va donc essayer de différencier les modèles en traçant les courbes obtenues à l'aide de ces modèles. Voici les sorties SAS que nous obtenons pour ces deux modèles qui nous donnent les valeurs des coefficients.

Paramètres estimés par l'analyse du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée	Erreur type	Intervalle de confiance à 95 %		Khi-2	Pr > Khi-2
Intercept	1	18.0531	5269.733	-10310.4	10346.54	0.00	0.9973
code_sexe	1	0.5685	0.0597	0.4515	0.6855	90.75	<.0001
contrat_adpsolution	1	-1.1718	0.0859	-1.3403	-1.0034	185.95	<.0001
contrat_prevind	1	-0.3217	0.0704	-0.4596	-0.1837	20.88	<.0001
fumeur_non_fumeur	1	0.5512	0.0611	0.4315	0.6710	81.39	<.0001
fumeur_fumeur	1	-0.2621	0.0769	-0.4127	-0.1115	11.63	0.0006

FIGURE 5.15 – Valeurs des coefficients des covariables pour la distribution exponentielle

Paramètres estimés par l'analyse du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée	Erreur type	Intervalle de confiance à 95 %		Khi-2	Pr > Khi-2
Intercept	1	17.7478	6264.615	-12260.7	12296.17	0.00	0.9977
code_sexe	1	0.5364	0.0579	0.4229	0.6499	85.85	<.0001
contrat_adpsolution	1	-1.1068	0.0858	-1.2749	-0.9387	166.51	<.0001
contrat_prevind	1	-0.3035	0.0670	-0.4347	-0.1722	20.53	<.0001
fumeur_non_fumeur	1	0.5201	0.0591	0.4043	0.6360	77.47	<.0001
fumeur_fumeur	1	-0.2471	0.0728	-0.3897	-0.1045	11.54	0.0007
Scale	1	0.9411	0.0238	0.8956	0.9890		

FIGURE 5.16 – Valeurs des coefficients des covariables pour la distribution log-logistique

On peut noter que dans les deux modèles, toutes les variables sont significatives car les p-value sont très faibles. Les intervalles de confiance à 95% des coefficients donnent des plages assez larges ce qui, comme pour les modèles précédents, est dû à la quantité de données qui est limitée. Cependant, on obtient bien des tendances nettes pour chaque coefficient ce qui nous permettra bien de modéliser l'hétérogénéité. Afin de valider que ces modèles ne donnent pas des résultats aberrants, nous présentons aussi le graphique donné par la sortie SAS pour la distribution log-logistique. Comme les âges sont des variables dans le modèles, le résultat donné nous donne une idée de la forme des courbes de mortalité que l'on peut obtenir. Or, on voit que celle-ci paraît classique.

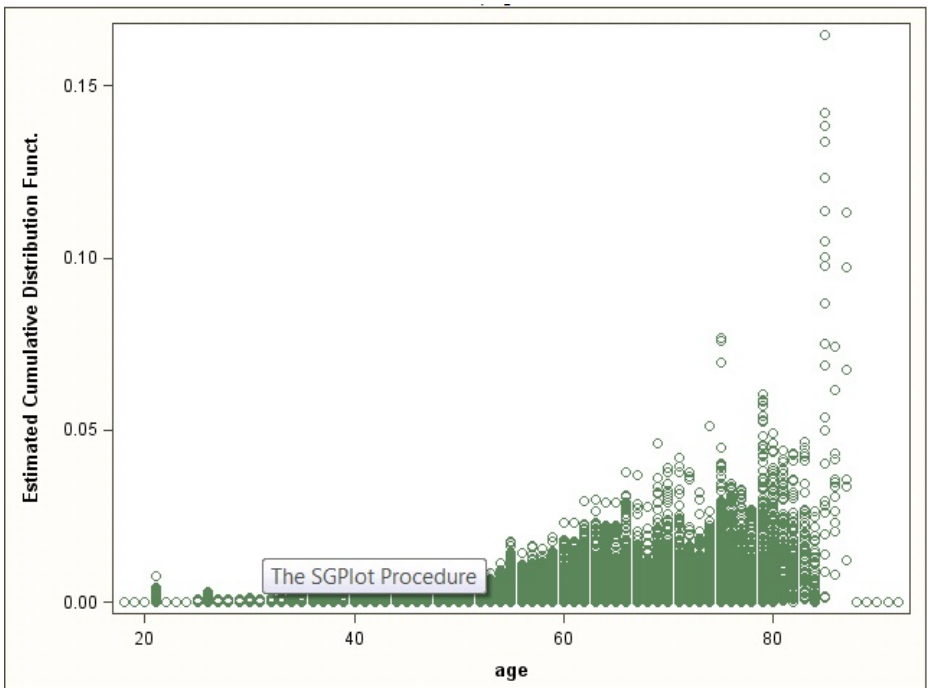


FIGURE 5.17 – Sortie SAS pour la distribution log-logistique donnant la forme des courbes que l'on obtiendra

Nous allons maintenant pouvoir présenter les résultats obtenus avec ces deux modèles de façon à choisir celui qui sera le mieux adapté à nos données. Pour cela, on a d'abord calculé la fonction de survie à chaque âge pour ensuite obtenir les taux de mortalité comme ces modèles permettent de calculer directement la survie. Nous avons donc tracé le graphe des courbes de mortalité de l'échantillon des hommes de référence ainsi que la courbe obtenue à la fin du chapitre 4. Les résultats sont les suivants.

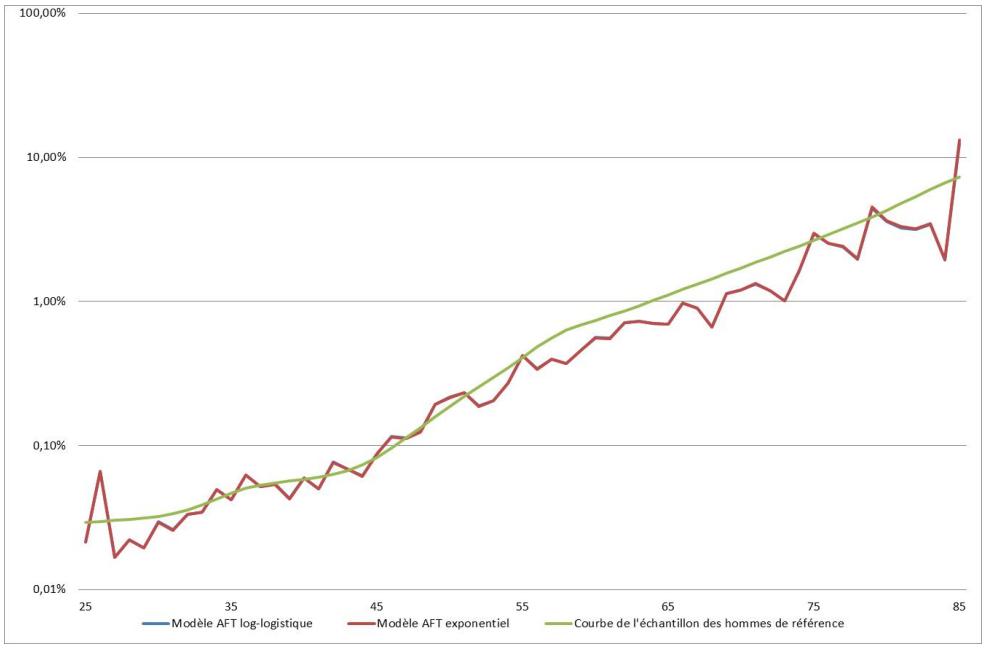


FIGURE 5.18 – Courbes de mortalité du modèle AFT comparées à la courbe obtenue dans le chapitre 4

La première observation que l'on peut faire est que les deux courbes des modèles AFT sont très proches (on ne distingue pas de différence visuelle sur le graphique même si les valeurs des taux de mortalité ne sont pas exactement identiques). Ceci provient du fait que notre modélisation considère les âges comme des covariables ce qui diminue l'écart entre les différentes distributions de base utilisées. On comprend alors mieux pourquoi il était difficile de faire un choix de modèle de façon claire avec les critères AIC et BIC. Cette prise en compte des âges comme des covariables implique aussi que les courbes obtenues ne sont pas lissées. Il faut donc ensuite utiliser les méthodes du chapitre 4 pour obtenir une courbe de mortalité lissée. Comme il y a un grand nombre de segmentations et qu'il faut lisser les courbes une par une, obtenir les résultats lissés est assez fastidieux.

La deuxième remarque que l'on peut faire est que jusqu'à 50 ans, les modèles AFT semblent bien correspondre à la courbe de mortalité de référence. Ensuite, il semble que la mortalité soit plus faible qu'avec la courbe obtenue à l'aide des méthodes du chapitre 4. Nous avons aussi tracé le même type de graphique pour l'échantillon des femmes de référence et dans ce cas, la mortalité était plus faible avec les modèles AFT sur la tranche d'âge 30-60 ans. Cette tranche représentant la majorité de la population que nous étudions, y sous-estimer la mortalité aurait un impact non négligeable. Nous allons donc voir dans la suite si cette sous-estimation semble systématique.

Comme d'après les précédents résultats nous n'avons pu définir un modèle AFT clairement meilleur qu'un autre, nous allons continuer avec les deux modèles afin de voir si certaines segmentations font apparaître une différence non négligeable. De plus, les courbes présentées dans la suite ne seront pas lissées dans l'immédiat, nous allons d'abord pouvoir tirer des enseignements avant de faire des estimations qui augmentent la probabilité de biais dans nos résultats.

5.10 Comparaison des modèles de Cox stratifié et AFT

Nous avons dans ce chapitre 5 présenté plusieurs méthodes pour modéliser l'hétérogénéité ainsi que plusieurs résultats. Le modèle de Cox stratifié a déjà été préféré à celui non stratifié, nous allons donc pouvoir comparer le premier avec les modèles AFT. Commençons par comparer directement les résultats obtenus pour différentes segmentations que nous avons modélisées.

FIGURE 5.19 – Comparaison des courbes de mortalité des modèles de Cox et AFT

Ce graphique nous donne de nombreuses informations. Tout d'abord, on peut voir que même pour une segmentation très différente de celle de base (hommes ayant souscrit un contrat ADP standard, gros capitaux ou prévoyance professionnelle et dont le caractère non fumeur n'est pas renseigné), les résultats des modèles AFT n'ont pas de différence graphique visible. Nous pourrons donc utiliser dans la suite indifféremment les deux modèles. Nous avons choisi de prendre celui log-logistique. Ensuite, la seconde observation est que pour les deux segmentations étudiées comme pour celles précédentes, les mortalités sont similaires sur de larges tranches d'âge mais il y en a toujours une où la courbe de mortalité avec le modèle AFT se trouve sous celle obtenue avec le modèle de Cox.

On peut d'ailleurs aussi remarquer que ces tranches d'âges sont toujours les hommes au-delà de 50 ans et les femmes entre 30 et 60 ans. Cette dernière remarque n'a rien d'étonnant car les formes des courbes du modèle AFT sont parallèles quelle que soit la segmentation (cela étant dû au fait que dans ce modèle les covariables jouent sur le paramètre de position et non d'échelle) et celles du modèle de Cox peuvent avoir deux formes différentes selon que les individus soient des femmes ou des hommes (du fait de la stratification avec la variable sexe et de l'hypothèse de proportionnalité).

Ceci nous amène à nous demander si c'est le modèle AFT qui sous-estime la mortalité ou alors le modèle de Cox qui la surestime. Pour cela, nous allons comparer les courbes obtenues grâce au modèle AFT pour les populations des hommes et des femmes de référence avec les courbes de la population totale des hommes et des femmes obtenues dans le chapitre 4. En effet, nous avons pu voir, notamment grâce aux méthodes de positionnement par rapport à une référence externe, que ces courbes présentent des mortalités proches.

Les résultats obtenus sont présentés dans le graphique suivant où les courbes non lissées sont celles obtenues avec le modèle à durée de vie accélérée et les autres sont les résultats hommes et femmes finaux présentés en fin de chapitre 4.

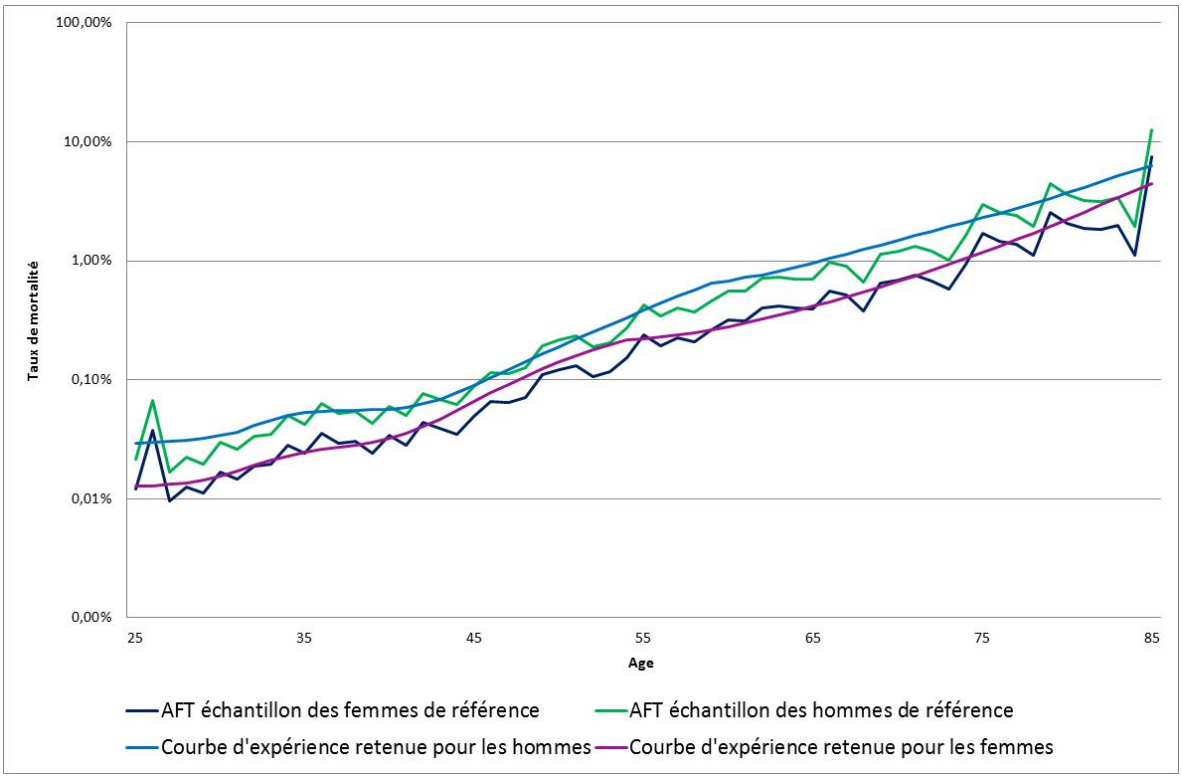


FIGURE 5.20 – Comparaison des courbes de mortalité du modèle AFT avec les courbes d'expérience des hommes et des femmes

Ce graphique montre que le modèle AFT semble bien sous-évaluer la mortalité observée du portefeuille sur certaines tranches d'âge. On pourrait supposer que ce sont les courbes d'expérience qui n'ont pas bien été calibrées, cependant pour les femmes la tranche d'âges 30-60 ans est celle où nous disposons du plus de données et de décès et donc d'une estimation de la mortalité assez précise. Pour les hommes, on estime aussi de façon précise la mortalité jusqu'à 60 ans, or on peut voir qu'à ces âges il y a déjà une sous-estimation.

De plus, même si les effectifs sont moins élevés au-delà de 60 ans, il y a encore suffisamment de décès (entre 10 et 25 jusqu'à 75 ans) pour que le positionnement par rapport à une référence externe ne surestime pas d'au moins 10% la mortalité. Pour ces raisons, nous avons donc finalement décidé de retenir le modèle de Cox stratifié pour produire des tables de mortalité pour toutes les segmentations. L'intégralité des courbes de mortalité obtenues avec le modèle de Cox stratifié sont présentées en annexe 8. Nous allons donc pouvoir utiliser ces tables dans le dernier chapitre afin de calculer des provisions pour le risque décès.

5.11 Discussion sur les modèles utilisés

Nous avons conclu avec le modèle AFT l'étude de la mortalité de notre portefeuille afin de pouvoir produire des tables. Pour cela, nous avons fait de nombreuses hypothèses et choisi des modèles parmi les nombreux qui existent. Il est donc intéressant de porter un regard critique sur les choix effectués.

Tout d'abord pour obtenir les taux bruts de notre portefeuille, nous avons utilisé des estimateurs couramment utilisés et dont l'efficience a été montrée par de nombreuses études, articles et thèses. On peut donc considérer que l'on a obtenu des résultats fiables si l'on a respecté les hypothèses. De ce point de vue là, les taux bruts estimés pour les hommes et les femmes respectent bien les hypothèses. Pour la population des hommes de référence, le nombre de décès à chaque âge est limité mais est suffisant sur la tranche d'âge qui a ensuite été modélisée le plus finement.

Ensuite, la méthode de Whittaker-Henderson utilisée et retenue pour le lissage est pratique car comme

elle est non paramétrique, elle a permis de rester le plus fidèle possible aux données tout en conservant une cohérence par rapport au risque modélisé : le risque de mortalité est croissant avec l'âge (excepté à la naissance mais cela ne nous intéressait pas dans notre étude). Par contre, nous ne disposons pas de critère mathématique tel que le calcul du maximum de vraisemblance pour pouvoir valider la modélisation qui est faite, il a donc fallu vérifier que les résultats étaient cohérents notamment par rapport aux tables de mortalité certifiées auxquelles on a accès. Enfin, dans le chapitre 4 les méthodes de positionnement par rapport à une référence externe présentent des risques limités car on calcule uniquement un coefficient global pour tous les âges, on a donc facilement un nombre de données important et un risque d'erreur limité.

La partie de l'étude la plus sujette à discussion concernant les risques d'erreur est bien évidemment la modélisation de l'hétérogénéité car le nombre de données pour une segmentation spécifique est parfois très faible (l'ADP risques aggravés représente moins de 20 000 individus au total). C'est pourquoi les modèles qui ont été finalement retenus pour effectuer la modélisation étaient des modèles permettant de jouer sur la position de la courbe de mortalité et non sa forme (le modèle de Cox-Aalen présenté aurait pu permettre cela sans augmenter trop fortement l'incertitude des résultats). Cependant, même en se limitant à la variation de la position de la courbe, on a pu voir qu'à chaque estimation des coefficients des modèles, on obtenait des intervalles de confiance assez importants (un paramètre ayant une effet proportionnel et ayant une valeur de 0,4 ou 0,6 ne donne pas du tout les mêmes résultats).

Les courbes de mortalité d'expérience obtenues dans le chapitre 4 et dans lesquelles on peut avoir confiance ont donc été utiles pour donner des points de repères sur les résultats et permettre de faire des validations visuelles de cohérence en plus de celles mathématiques. En effet, le respect de critères mathématiques est une indication mais la superposition de modèles pour arriver aux résultats finaux augmente graduellement le risque de biais, il est donc important de garder des points de repère et notamment de comparer ses résultats avec les courbes certifiées quand cela a un sens.

On peut d'ailleurs noter que dans le cadre du modèle de Cox stratifié, la courbe de référence pour les femmes n'a pas pu être obtenue directement à partir des taux bruts par manque de données, on ne peut donc pas être sûr que la forme de la courbe soit bien ajustée. Cependant, l'ajustement effectué pour les hommes ayant les mêmes caractéristiques a montré que la courbe de mortalité obtenue était proche de celle des hommes sans segmentation. De plus, les proportions d'effectifs dans chaque segmentation sont globalement comparables (et en particulier pour la segmentation de référence utilisée). On peut donc supposer que l'hypothèse d'effectuer un ajustement par référence externe à l'aide de la courbe de mortalité d'expérience des femmes est correcte.

Enfin, on peut noter que le modèle AFT, même s'il n'a pas été retenu (à cause de la sous-estimation de la mortalité qui s'observait pour chaque segmentation par rapport au modèle de Cox) nous permet de valider en partie le modèle de Cox car les courbes de mortalité pour les segmentations étaient similaires sur de larges tranches d'âges malgré des méthodes de modélisation différentes (même si le fait d'avoir introduit les âges comme covariables a réduit les écarts potentiels avec le modèle de Cox).

Chapitre 6

Impact des modélisations sur l'estimation de sinistralité décès

Nous avons construit dans les chapitres précédents des tables de mortalité pour différentes segmentations du portefeuille (hommes et femmes dans les chapitres 3 et 4 puis une segmentation suivant les variables sexe, produit et caractère fumeur dans le chapitre 5). Comme nous l'avons expliqué dans le début du mémoire, l'intérêt de la construction de ces nouvelles tables est de pouvoir estimer la sinistralité décès à horizon un an de notre portefeuille de façon plus précise. Nous allons donc estimer cette sinistralité.

6.1 Présentation des bases et des calculs

Pour estimer la sinistralité, nous avons deux bases pour 2013 à notre disposition : la base des contrats et celle des adhérents. La base des contrats contient une ligne par personne sous risque avec les montants garantis (montants de capitaux décès pour la prévoyance et montant de l'emprunt pour l'ADP), l'identifiant adhérent principal ou secondaire, le produit et la date de fin du contrat. La base des adhérents contient une ligne par numéro d'adhérent et donne les dates de naissance et le caractère fumeur de l'adhérent principal et éventuellement du second assuré ayant le même numéro d'adhérent (le conjoint en général) s'il existe.

La première étape de calcul a été d'obtenir un capital sous risque pour les contrats ADP car seul le montant initial du prêt garanti est indiqué. Pour cela, nous avons le nombre de mois de prêt à disposition ainsi que sa date de fin. Ainsi, nous avons pu calculer le capital restant dû au 01/01/2013 et au 01/01/2014 (ou à la date de fin de prêt si elle est antérieure). Ensuite, nous avons décidé de faire l'hypothèse que le capital sous risque durant toute la période considérée était la moyenne de ces deux montants. Effectuer un calcul mois par mois aurait complexifié de façon importante les calculs en ayant un impact très faible sur les résultats.

Une fois les capitaux sous risque estimés, nous avons lié les tables des contrats et des adhérents de façon à obtenir toutes les informations dont nous avions besoin (qui sont celles listées dans le premier paragraphe). On peut noter que comme ces tables ont déjà été fiabilisées précédemment pour effectuer des calculs dans la société (pour l'estimation de la sinistralité à un an notamment), il n'y avait de contrat sans adhérent correspondant. Il y avait seulement une cinquantaine d'individus dont l'identifiant adhérent principal ou secondaire était mal renseigné mais cela a été modifié manuellement et ne représente que 0,01% des données.Une fois que toutes les informations nécessaire au calcul du taux de mortalité ont été indiquées pour chaque personne sous risque, nous avons calculé les q_x pour chacun à l'aide de la même formule que celle utilisée en début de mémoire :

$$\forall 0 \leq t \leq 1,\ {}_tq_x = 1 - (1 - q_x)^t \,.$$

Ces coefficients de mortalité ont été calculés à partir de quatre tables de mortalité différentes de façon à pouvoir comparer les résultats : la table obtenue à l'aide du modèle de Cox, la table femme ou homme d'expérience du chapitre 4, la table utilisée jusqu'à présent qui est obtenue en appliquant un taux

d'abattement par rapport aux tables réglementaires (TH-002 et TF-002) suivant le sexe, le produit et le caractère fumeur et un critère de sélection médicale (les résultats obtenus ont été fournis pour cette table) et enfin la même table que la précédente en ne prenant pas en compte le dernier critère de sélection médicale. Pour finir, nous avons calculé par tête la sinistralité estimée théorique à l'aide de la formule $sinistralité\ estimée\ =\ capital\ sous\ risque\ *\ probabilité\ de\ décès$. On peut noter que cette sinistralité estimée tête par tête n'est cohérente qu'à travers l'ensemble du portefeuille car pour chaque individu, soit on paye la totalité du capital, soit on ne paye rien dans le cas du risque décès.

6.2 Sinistralité estimée sur un an

La sinistralité estimée avec les différents taux de mortalité, provenant des modèles étudiés et des hypothèses utilisées par la société aujourd'hui, est la suivante.

	Modèle de Cox stratifié	Courbe F/H d'expérience	Calcul avec taux d'abattement et sélection médicale	Calcul avec taux d'abattement
Sinistralité estimée	21 716 920 €	22 448 515 €	25 983 882 €	32 160 971 €

FIGURE 6.1 – Comparaison des la sinistralité estimée sur un an avec les différentes tables

On remarque tout de suite avec ces résultats que la sinistralité calculée à l'aide des tables d'expérience est bien moins élevée que celle calculée avec les hypothèses jusqu'alors utilisées. Nous avons chiffré en pourcentages ces différences et celles-ci sont résumées ci-dessous.

Abattement	Modèle de Cox stratifié	Courbe F/H d'expérience
Calcul avec taux d'abattement et sélection médicale	16,4%	13,6%
Calcul avec taux d'abattement	32,5%	30,2%
Courbe F/H d'expérience	3,3%	

FIGURE 6.2 – Différences relatives entre les modèles

La sinistralité globale serait donc surestimée actuellement d'après les résultats. On remarque que l'écart entre les nouvelles courbes obtenues avec le modèle de Cox stratifié et le calcul effectué jusqu'à présent est de l'ordre de 15% tandis qu'il grimpe à 30% quand on ne prend pas en compte le bénéfice de la sélection médicale. On peut donc d'ores et déjà considérer que le fait d'avoir intégré un facteur de sélection médicale dans le calcul actuel est indispensable car sinon la mortalité serait assez largement surestimée. Pour ce qui est de l'écart de 15% avec le calcul actuel, on peut considérer qu'il est significatif, il est donc intéressant de comprendre d'où il peut venir.

Tout d'abord, au niveau des données, la vérification de la qualité des données au début de l'étude nous permet d'écarter cette hypothèse. Ensuite, les méthodes utilisées ont déjà montré leur intérêt dans des recherches précédentes et les résultats comparatifs étaient à chaque fois similaires. Enfin, toujours dans le cadre de la validation des données, on peut noter que les courbes d'expérience hommes et femmes se sont basées sur des échantillons de plusieurs centaines de milliers d'individus, il y a donc un risque d'erreur d'estimation que l'on peut considérer faible par rapport au 15% d'écart que l'on a mis en évidence. L'écart provient donc des méthodes utilisées pour estimer la mortalité. Celle utilisée jusqu'à présent est une méthode prudente sur plusieurs points :

- le nombre de décès est estimé en cohérence avec les années précédentes, la diminution du nombre de têtes assurées due au run-off n'est donc pas prise en compte
- l'âge des assurés n'est pris en compte que globalement (même forme de courbe que celles réglementaires).

Cette étude ayant été demandée car on supposait une mortalité plus faible que celle modélisée, on peut juger que les résultats obtenus valident l'intérêt de l'avoir réalisée. On peut aussi noter que l'écart de 3% obtenu en utilisant le modèle de Cox plutôt que les courbes globales d'expérience hommes et femmes nous montre qu'une segmentation plus fine des individus nous permet d'abaisser encore le montant de sinistralité estimée car nous pouvons effectuer un calcul tête par tête plus juste. Cependant, comme les échantillons sont petits, dans la pratique ce seront plutôt les tables globales hommes et femmes qui seront utilisées pour les calculs de *best estimate* dans le cadre de l'obtention du SCR. Les tables de mortalité du modèle de Cox stratifié permettent en tout cas d'obtenir des résultats de sinistralité estimée tête par tête cohérents et sont utiles si l'on veut faire des études de sinistralité sur des divers échantillons fins du portefeuille et non pour un produit dans sa globalité.

Comme ces chiffres ont été obtenus à l'aide d'estimation, il est intéressant de faire un back-testing et donc de savoir si les courbes de mortalité obtenues ne sous-estiment pas la mortalité. Pour cela, nous avons utilisé le montant réel de prestations, PSAP et IBNR qui a été de 22,4 m€ pour l'année 2014. Ce chiffre est donc équivalent à celui obtenu avec les courbes de mortalité hommes et femmes sans segmentation particulière et 3% plus élevé qu'avec le modèle de Cox stratifié. Il faut noter que sur la partie du portefeuille étudié il y a eu une petite partie de nouvelles affaires qui n'ont donc pas été prises en compte pour estimer la mortalité sur un an. De plus, les PSAP et et les IBNR comportent aussi un potentiel biais.

Au vu de ces observations, on peut considérer que la charge de mortalité estimée sur une année est équivalente à celle réellement observée ce qui veut dire qu'on a bien calculé un *best estimate* sur le risque mortalité du portefeuille. On peut ajouter que si l'on veut privilégier une certaine prudence, il faudra mieux utiliser les courbes d'expérience globales pour les hommes et les femmes. L'utilisation des courbes obtenues avec le modèle de Cox stratifié permettront elles d'observer et mesurer des changements dans la répartition des assurés à l'intérieur du portefeuille global (les affaires nouvelles ne correspondant pas à la répartition actuelle).

En conclusion de ce chapitre, on peut donc dire que les résultats obtenus permettent de mettre en évidence une surestimation de la mortalité du portefeuille avec les hypothèses actuelles et que les courbes obtenues permettent de calculer un *best estimate* satisfaisant. On pourra donc à l'avenir, en utilisant les résultats, baisser l'estimation de la sinistralité future en donc le SCR et les besoins en fonds propres réglementaires.

Conclusion

Cette étude de la mortalité d'expérience du portefeuille de l'entreprise a permis de mettre en évidence plusieurs points dans ses différentes parties.

Tout d'abord, afin de mettre en place les tables de mortalité, deux estimateurs des taux bruts ont été utilisés et nous avons pu voir qu'ils donnaient des résultats très similaires. Ensuite, pour obtenir des tables hommes et femmes présentant des taux croissants et réguliers nous avons mis en place deux types de méthodes. La première se basant uniquement sur les taux bruts et permettant de les lisser ou de les ajuster pour les âges où les données étaient suffisantes. La seconde positionnant les courbes de mortalité par rapport à celles réglementaires. Nous avons ainsi pu voir que dans notre cas la méthode non paramétrique donnait des résultats plus proches mathématiquement et graphiquement des taux bruts et c'est pourquoi elle a été retenue dans le cadre de notre étude dont le but est de donner une vision Best Estimate.

Pour les méthodes de positionnement par rapport à une référence externe, la méthode la plus prudente a été retenue du fait de la quantité faible de données aux âges concernés, sans pour autant en choisir une semblant trop surestimer la mortalité. Les tables femmes et hommes obtenues ont permis de valider définitivement la nécessité pour Axéria Prévoyance de construire ses propres tables d'expérience car la forme des courbes diffère par rapport à celles réglementaires. Nous avons notamment pu voir que la mortalité d'expérience était très largement inférieure aux jeunes âges et qu'elle avait tendance à se rapprocher proportionnellement parlant de celle réglementaire avec l'augmentation de l'âge.

Ensuite, la mise en place des modèles prenant en compte l'hétérogénéité a été celle qui a demandé le plus de temps car elle demande de faire des choix justifiés du fait de la faible quantité de données. Nous avons donc pu voir que l'utilisation de modèles à hasards proportionnels était la plus cohérente dans notre cas et permettait d'obtenir des résultats prenant bien en compte les psécificités des échantillons. On peut par exemple noter que pour les échantillons avec les mortalités les plus extrêmes (les hommes présentant un risque aggravé et fumeurs contre les femmes ayant souscrit un contrat ADP standard, gros capitaux ou prévoyance professionnelle et fumeuses), il y a un rapport de taux de mortalité de 12 à 40 ans. Cette différence peut paraître trop élevée, cependant en mettant en relation le fait que de base la mortalité des femmes est deux fois moins importante que celle des hommes, celle des individus présentant un risque aggravé plus de trois fois plus élevée que les autres et enfin celle des fumeurs par rapport aux non fumeurs étant égale à plus du double, on comprend que ce rapport n'est pas forcément incohérent.

Le modèle AFT a lui aussi permis de bien mettre en évidence les spécifités des échantillons comme les comparaisons avec le modèle de Cox stratifié l'ont montrée. Cependant, la sous-estimation récurrente de la mortalité sur les échantillons ainsi que la seule forme de courbe de base nous ont décidés à ne pas retenir ce modèle. Pour finir, on peut noter que le modèle de Cox est celui le plus utilisé pour la prise en compte des variables explicatives et donc la modélisation de l'hétérogénéité dans le domaine de l'actuariat ce qui paraît maintenant cohérent du fait de sa relative simplicité à mettre en place et des possibilités qu'il permet notamment avec sa variante utilisant les strates.

Pour finir, les résultats finaux obtenus ont permis de mettre en évidence une surestimation de la

sinistralité estimée à un niveau significatif. La mise en place de cette étude reposant sur l'observation globale du portefeuille, les demandes des autorités de régulation et l'intuition, on peut donc dire qu'il a été utile de la réaliser malgré le temps que cela prend. En effet, ce genre d'études permet d'obtenir des résultats chiffrés plus précis et justifiés qui peuvent aider à différents niveaux dans l'entreprise. Ainsi, l'utilisation des résultats est bien sûr surtout axée sur la justification des hypothèses de calcul du *best estimate* mais peut permettre en suivant l'évolution ou en s'intéressant à des échantillons précis de piloter ou suivre la réalisation concrète de stratégies sur des produits en particulier ou les données sont insuffisantes pour effectuer une étude directement à partir de celles-ci.

Cette étude s'est donc arrêtée sur une estimation à l'aide de certaines variables mais une des pistes d'évolution pourrait aussi être de prendre en compte des critères sociologiques ou géographiques dans la modélisation. Il faudrait cependant faire des choix de variables car en ajouter au modèle actuel ne paraît pas possible afin d'obtenir des résultats cohérents. Il faudrait donc savoir si les critères précités sont aussi explicatifs que ceux que nous avons utilisés. Pour conclure ce mémoire, il est important de noter que cette étude de mortalité n'est pas prospective et demande donc une actualisation afin d'obtenir des résultats les plus justes possibles dans le futur.

Bibliographie

[1] Bagui H. (2013) *Refonte des lois de maintien en incapacité temporaire de travail*, Mémoire de master, ISFA, Université Claude Bernard Lyon 1.

[2] Cao H. (2005) *A comparison between the additive and multiplicative risk models*, Mémoire de maître des sciences, Faculté des sciences et de génie, Université Laval.

[3] Choukroun M. (2008) *Le modèle additif d'Aalen, une alternative au modèle de Cox dans le cadre de la construction d'une loi de maintien en incapacité de travail*, Bulletin Français d'Actuarait, vol 8, n°16, p. 107-138.

[4] Clement O. (2013) *Elaboration d'une table d'expérience : comparaison de métodes de lissage analytique et d'ajustement statistique*, Mémoire d'actuaire, EURIA, Université de Bretagne Occidentale.

[5] Colletaz G. (2012) *Modèles de survie*, Notes de cours, Master 2 ESA, Université d'Orléans, p.67-99.

[6] Dragomir A., *Estimateurs non-paramétriques*, Notes de cours, Université de Montréal.

[7] El Sanharawi M., Naudet F. (2013) *Comprendre la régression logistique*, Journal Français d'Ophtalmologie, Vol. 36, p.710-715.

[8] Fardel V., Gallic E. (2013) *Le modèle de Cox*, Notes de cours, Faculté des sciences économiques, Université Rennes 1.

[9] Insee (2014) *Table de mortalité* [en ligne], Insee, disponible sur : *http ://www.insee.fr/fr/methodes/default.asp ?page=definitions/table-de-mortalite.htm.*

[10] Jouannigot E. (2009) *Etude comparative des modèles de durée en présence de données hétérogènes et censurées*, Mémoire d'actuaire, EURIA, Université de Bretagne Occidentale.

[11] Kamega A., Planchet F. (2010) *Mesure du risque d'estimation associé à une table d'expérience*, Université Claude Bernard Lyon 1.

[12] Kamega A., Planchet F. (2011) *Construction de tables de mortalité sur un groupe restreint : mesure du risque d'estimation*, Université Claude Bernard Lyon 1.

[13] Kamega A., Planchet F. (2011) *Hétérogénéité : mesure du risque d'estimation dans le cas d'une modélisation intégrant des facteurs observables*, Bulletin Français d'Actuariat, Vol. 11, Nbr. 21, p. 99-129.

[14] Kraus D. (2004) *Goodness-of-fit inference for the Cox-Aalen additive-multiplicative regression model*, Department of probability and statistics, University Charles.

[15] *L'approche semi-paramétrique : le modèle ed Cox*, Notes de cours, Université d'Orléans, p. 51-76, disponible sur *http ://www.univ-orleans.fr/deg/masters/ESA/GC/sources/Survie%20semi*$_p$*arametrique.pdf*.

[16] Lelieur V. *Utilisation des méthodes de Lee-Carter et Log-Poisson pour l'ajustement de tables de mortalité dans le cas de petits échantillons*, Mémoire d'actuaire, ISFA, Université Claude Bernard Lyon 1.

[17] Lemler S. (2012) *Modèles de durée, analyse de survie*, Notes de cours, ENSIIE.

[18] Leste-Lasserre C. (2011) *Construction de tables de mortalité d'expérience*, Mémoire d'actuaire, ISUP, Université Pierre et Marie Curie.

[19] Leurent T. (2010) *Construction de tables d'expérience des risques incapacité et invalidité*, Mémoire d'actuaire, DUAS, Faculté des sciences économiques et de gestion de Strasbourg.

[20] Martel L., Provost M., Lebal A., Coulombe S., Sherk A. (2013) *Méthodologie des tables de mortalité pour le Canada, les provinces et les territoires* [en ligne], Statistique Canada, disponible sur : *http ://www.statcan.gc.ca/pub/84-538-x/84-538-x2013001-fra.htm.*

[21] Michel E., Jougla E., Hatton F., Chérié-Challine L. *Principaux indicateurs de mortalité* [en ligne], Inserm-CépiDc, disponible sur : *http ://www.cepidc.inserm.fr/inserm/html/pages/Principaux_Indicateurs_fr.htm.*

[22] Planchet F. (2013-2014) *Modèles de durée : Statistique des modèles paramétriques et semi-paramétriques*, Notes de cours, ISFA.

[23] Planchet F., Thérond P. (2006) *Modèles de durée, Applications actuarielles*, Assurance Audit Actuariat, Economica.

[24] Qi J. (2009) *Comparison of proportional hazards and accelerated failure time models*, Mémoire de master, Department of mathematics and statistics, University of Saskatchewan.

[25] Quashie A., Denuit M. (2005) *Modèles d'extrapolation de la mortalité aux grandes âges*, Institut des Sciences Actuarielles et Institut de Statistique, Université Catholique de Louvain.

[26] Roch G. *Estimation d'un taux de survie*, Notes de cours, Faculté de médecine d'Aix-Marseille.

[27] Rose N. (2009) *Provisionnement en assurance non-vie : Utilisation de modèles paramétriques censurés*, Mémoire d'actuaire, ISUP, Université Pierre et Marie Curie.

[28] Sallah K. *Méthode actuarielle d'estimation des courbes de survie : principe, différences avec la méthode de Kaplan-Meier*, DELBIM - Université de Lomé.

[29] Scheike T. (2011) *Modern regression methods for survival data*, Présentation de cours, Université de Copenhague.

[30] Spac actuaires (2011) *Table de mortalité* [en ligne], disponible sur : *http ://www.spac-actuaires.fr/glossaire/Table_de_mortalite.*

[31] Taupin M-L. (2011) *Durées de survie*, Présentation de cours, Université Paris Descartes.

[32] Zhang X. (2011) *Construction des tables de mortalité d'expérience en cas de décès*, Mémoire d'actuaire, ISFA, Université Claude Bernard Lyon 1.

Annexes

Annexe 1 : Estimateur de la variance de Greenwood

L'estimateur de la variance de Greenwood se base sur le fait que le nombre de décès à l'âge x suit une loi binomiale :

$$d_x \ \sim \ B(N_x, q_x)\,.$$

Ainsi l'estimateur de la fonction de survie vérifie les égalités suivantes.

$$\begin{aligned}
Var(\hat{S}(t) &= Var\left[\prod_{i,t_i\leq t}\left(1-\frac{d_i}{N_i}\right)\right] \\
&= \mathbb{E}\left[\prod_{i,t_i\leq t}\left(1-\frac{d_i}{N_i}\right)^2\right] - \mathbb{E}^2\left[\prod_{i,t_i\leq t}\left(1-\frac{d_i}{N_i}\right)\right] \\
&= \prod_{i,t_i\leq t}\mathbb{E}\left[\left(1-\frac{d_i}{N_i}\right)^2\right] - \prod_{i,t_i\leq t}\mathbb{E}^2\left(1-\frac{d_i}{N_i}\right) \quad \text{(On utilise l'indépendance des évènements de décès)} \\
&= \prod_{i,t_i\leq t}\left[Var\left(1-\frac{d_i}{N_i}\right)+\mathbb{E}^2\left(1-\frac{d_i}{N_i}\right)\right] - \prod_{i,t_i\leq t} p_i^2 \quad \text{(l'espérance de survie est } p_i\text{)} \\
&= \prod_{i,t_i\leq t}\left(\frac{p_i(1-p_i)}{N_i}+p_i^2\right) - \prod_{i,t_i\leq t} p_i^2 \quad \text{(on utilise la variance de la loi binomiale)} \\
&= \prod_{i,t_i\leq t} p_i^2\left(\frac{(1-p_i)}{p_iN_i}+1\right) - \prod_{i,t_i\leq t} p_i^2
\end{aligned}$$

Or, on sait qu'avec l'estimateur de Kaplan-Meier $S(t) = \prod_{i,t_i\leq t} p_i$. On a donc :

$$\begin{aligned}
Var(\hat{S}(t) &= S^2(t)\prod_{i,t_i\leq t}\left(\frac{(1-p_i)}{p_iN_i}+1\right) - S^2(t) \\
&\approx S^2(t)\sum_{i,t_i\leq t}\left(\frac{(1-p_i)}{p_iN_i}\right) \text{(avec un développement limité en prenant le log)} \\
&\approx \hat{S}^2(t)\sum_{i,t_i\leq t}\left(\frac{(1-\hat{p}_i)}{\hat{p}_iN_i}\right) \\
&\approx \hat{S}^2(t)\sum_{i,t_i\leq t}\left(\frac{d_i}{N_i(N_i-d_i)}\right)\,.
\end{aligned}$$

On obtient donc bien la formule présentée pour la variance de l'estimateur de Kaplan-Meier avec $\sum_{i,x\leq t_i\leq x+1}\left(\frac{d_i}{N_i(N_i-d_i}\right) = \gamma^2(x)$

Annexe 2 : Exemple de sortie SAS de la procédure Genmod et graphique pour la segmentation des hommes de référence

The GENMOD Procedure

Informations sur le modèle		
Data Set	WORK.ESTIMATION_HOMMES	
Distribution	Binomial	
Link Function	Logit	
Response Variable (Events)	deces	deces
Response Variable (Trials)	exposition	exposition

Number of Observations Read	31
Number of Observations Used	31
Number of Events	753
Number of Trials	807295.4

Profil de réponse		
Valeur ordonnée	Binary Outcome	Fréquence totale
1	Event	753
2	Nonevent	806542.4

Critères d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Deviance	28	26.9642	0.9630
Scaled Deviance	28	26.9642	0.9630
Pearson Chi-Square	28	26.7266	0.9545
Scaled Pearson X2	28	26.7266	0.9545
Log Likelihood		-5744.0742	
Full Log Likelihood		-90.5287	
AIC (smaller is better)		187.0573	
AICC (smaller is better)		187.9462	
BIC (smaller is better)		191.3593	

Algorithm converged.

Paramètres estimés par l'analyse du maximum de vraisemblance							
Paramètre	DDL	Valeur estimée	Erreur type	Intervalle de confiance de Wald à 95 %		Khi-2 de Wald	Pr > Khi-2
Intercept	1	-9.3253	0.4662	-10.2390	-8.4115	400.10	<.0001
age	1	0.0484	0.0117	0.0254	0.0714	17.03	<.0001
agesp44	1	0.0985	0.0177	0.0639	0.1332	31.05	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

neter was held fixed.

Statistique LR pour Analyse de Type 3			
Source	DDL	Khi-2	Pr > Khi-2
age	1	17.74	<.0001
agesp44	1	30.45	<.0001

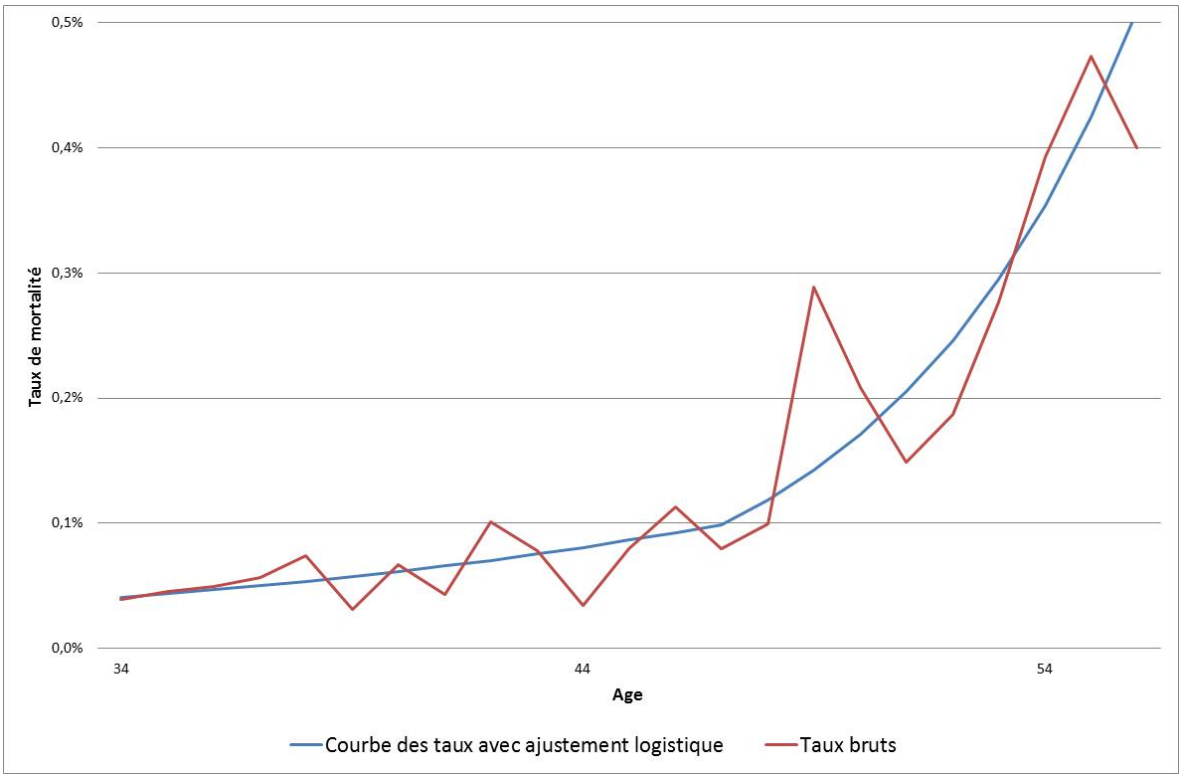


FIGURE 3 – Graphique de l'ajustement logistique de la segmentation des hommes de référence

Annexe 3 : Tables de mortalité d'expérience

Voici les valeurs des tables de mortalité d'expérience retenues pour les hommes et les femmes.

Age	Hommes	Femmes	Age	Hommes	Femmes
10	0,004%	0,004%	56	0,439%	0,240%
11	0,004%	0,004%	57	0,501%	0,243%
12	0,005%	0,004%	58	0,570%	0,248%
13	0,006%	0,005%	59	0,645%	0,263%
14	0,007%	0,006%	60	0,727%	0,281%
15	0,010%	0,007%	61	0,745%	0,299%
16	0,014%	0,009%	62	0,764%	0,323%
17	0,019%	0,011%	63	0,817%	0,350%
18	0,023%	0,012%	64	0,889%	0,381%
19	0,027%	0,013%	65	0,969%	0,415%
20	0,029%	0,013%	66	1,051%	0,454%
21	0,029%	0,013%	67	1,146%	0,499%
22	0,029%	0,013%	68	1,250%	0,549%
23	0,029%	0,013%	69	1,363%	0,607%
24	0,029%	0,013%	70	1,488%	0,674%
25	0,029%	0,013%	71	1,626%	0,749%
26	0,030%	0,013%	72	1,774%	0,834%
27	0,031%	0,013%	73	1,935%	0,930%
28	0,031%	0,014%	74	2,110%	1,042%
29	0,032%	0,014%	75	2,305%	1,172%
30	0,034%	0,015%	76	2,524%	1,326%
31	0,036%	0,017%	77	2,765%	1,504%
32	0,041%	0,019%	78	3,038%	1,710%
33	0,046%	0,021%	79	3,356%	1,949%
34	0,050%	0,023%	80	3,729%	2,231%
35	0,053%	0,024%	81	4,163%	2,561%
36	0,055%	0,026%	82	4,650%	2,949%
37	0,055%	0,027%	83	5,181%	3,397%
38	0,056%	0,028%	84	5,748%	3,906%
39	0,056%	0,030%	85	6,357%	4,473%
40	0,057%	0,032%	86	7,019%	5,097%
41	0,059%	0,036%	87	7,729%	5,768%
42	0,063%	0,040%	88	8,487%	6,465%
43	0,069%	0,047%	89	9,283%	7,171%
44	0,078%	0,055%	90	10,113%	7,890%
45	0,089%	0,066%	91	10,990%	8,655%
46	0,104%	0,078%	92	11,922%	9,499%
47	0,121%	0,092%	93	12,913%	10,431%
48	0,142%	0,107%	94	13,959%	11,441%
49	0,164%	0,124%	95	15,063%	12,531%
50	0,190%	0,141%	96	16,220%	13,688%
51	0,219%	0,159%	97	17,430%	14,902%
52	0,252%	0,178%	98	18,691%	16,189%
53	0,290%	0,197%	99	20,023%	17,536%
54	0,334%	0,217%	100	21,370%	18,943%
55	0,383%	0,238%			

Annexe 4 : Résultats du test du log-rank

Segmentation	Valeur de la statistique	p-value	Valeur stat pour p-value
Hommes/Femmes	107,47	0,001	10,83
Fumeurs/Non fumeurs	142	0,001	10,83
Fumeurs/Non renseigné	54	0,001	10,83
Non fumeurs/Non renseigné	110	0,001	10,83
ADP/ADP gros capitaux	1,58	0,21	1,57
ADP/ADP risques aggravés	281,9	0,001	10,83
ADP/Prev ind	35,4	0,001	10,83
ADP/Prev pro	0,12	0,73	0,12
ADP gros capitaux/ADP risques aggravés	79,8	0,001	10,83
ADP gros capitaux/ Prev ind	13,4	0,001	10,83
ADP gros capitaux/Prev pro	3,1	0,078	3,11
ADP risques aggravés/Prev ind	71,1	0,001	10,83
ADP risques aggravés/Prev pro	62,6	0,001	10,83
Prev ind/Prev pro	6,6	0,0185	5,55
ADP+Prev pro/ADP gros capitaux	1,6	0,21	1,57

FIGURE 4 – Tableau présentant les valeurs du test du log-rank

Annexe 5 : Algorithme de Newton-Raphson

Cette présentation est basée sur les notes de cours de Mr. Planchet (2013).

Afin de résoudre les équations faisant intervenir la vraisemblance, cet algorithme est très utilisé (en particulier dans les logiciels comme SAS et R). Le principe pour résoudre les équations de la forme $f(x_0) = 0$, x jouant ici le rôle de θ dans les équations de vraisemblance, est d'utiliser un développement de Taylor à l'order un qui s'écrit de la façon suivante :

$$f(x_{k+1} = f(x_k) + (x_{k+1} - x_k)\frac{df}{dx}(x_k) + o(x_{k+1} - x_k)\,.$$

Cela permet d'obtenir la récurrence suivante en utilisant le fait que cette expression doit être nulle :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}\,.$$

En remplaçant x par le paramètre du score et f par la dérivée de la log-vraisemblance par rapport au score, on obtient l'expression finale suivante que l'on doit faire converger :

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2}{\partial\theta\partial\theta'} lnL(y|z,c;\theta_k)\right]^{-1} \frac{\partial lnL(y|z,c;\theta_k)}{\partial\theta} .$$

Afin de faire converger cet algorithme et qu'il soit asymptotiquement efficace, ce qui permet d'obtenir une estimaion des paramètres, on doit donc définir une valeur initiale. Pour cela, on peut utiliser un estimateur convergent mais non asymptotiquement efficace.

Annexe 6 : Tests de vraisemblance pour le modèle de Cox

La présentation de ces tests est également basée sur les notes de cours de Mr. Planchet (2013).Dans notre cas, nous voulons tester l'hypothèse de nullité globale des coefficients. Pour cela on teste donc l'hypothèse $g(\beta) = 0$ contre $g(\beta) \neq 0$. On note $\hat{\beta}$ et $\hat{\beta}^0$ respectivement les estimateurs du maximum de vraismeblance dans le modèle non contraint et celui contraint. En considérant que $g(\beta)$ est un vecteur de dimension r, tous les tests qui vont suivre doivent tendre vers un $\chi^2(r)$ si l'hypotèse de nullité globale des coefficients est vérifiée.

La statistique du test du rapport de vraismeblance s'écrit :

$$T_R = 2\left(lnL(\hat{\beta}) - lnL(\hat{\beta}^0)\right) .$$

La statisqtique du test de Wald est :

$$T_W = ng'(\hat{\beta})\left[\frac{\partial g(\hat{\beta})}{\partial\beta'} I(\hat{\beta})^{-1} \frac{\partial g'(\hat{\beta})}{\partial\beta}\right]^{-1} g(\hat{\beta}) .$$

On précise ici que I est la matrice d'information de Fisher.

Enfin, la statistique du test du score est définie de la manière suivante :

$$T_S = \frac{1}{n} \frac{\partial lnL(\hat{\beta}^0)}{\partial\beta'} I(\hat{\beta}^0)^{-1} \frac{\partial lnL(\hat{\beta}^0)}{\partial\beta} .$$

Annexe 7 : Sortie SAS concernant les interactions entre variables dans le modèle de Cox

Estimations par l'analyse du maximum de vraisemblance

Paramètre			DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque	Intervalle de confiance à 95 % du rapport de risque		Libellé
Sexe	F		1	-0.62440	0.09966	39.2534	<.0001	0.536	0.441	0.651	Sexe F
code_contrat	ADP_Solution		1	1.23190	0.11033	124.6719	<.0001	.	.	.	code_contrat ADP_Solution
code_contrat	Prev_ind		1	0.47006	0.10758	19.0919	<.0001	.	.	.	code_contrat Prev_ind
code_fumeur	Fumeur		1	0.42443	0.10270	17.0776	<.0001	.	.	.	code_fumeur Fumeur
code_fumeur	Non_fu		1	-0.45615	0.08215	30.8350	<.0001	.	.	.	code_fumeur Non_fu
code_sexe*code_contr	ADP_Solution		1	0.44047	0.18885	5.4397	0.0197	.	.	.	code_contrat ADP_Solution * code_sexe
code_sexe*code_contr	Prev_ind		1	0.15692	0.14750	1.1317	0.2874	.	.	.	code_contrat Prev_ind * code_sexe
code_sexe*code_fumeu	Fumeur		1	-0.25138	0.18562	1.8340	0.1757	.	.	.	code_fumeur Fumeur * code_sexe
code_sexe*code_fumeu	Non_fu		1	0.03802	0.13529	0.0790	0.7787	.	.	.	code_fumeur Non_fu * code_sexe
code_cont*code_fumeu	ADP_Solution	Fumeur	1	-0.50756	0.42739	1.4103	0.2350	.	.	.	code_contrat ADP_Solution * code_fumeur Fumeur
code_cont*code_fumeu	ADP_Solution	Non_fu	1	-0.71404	0.26589	7.2115	0.0072	.	.	.	code_contrat ADP_Solution * code_fumeur Non_fu
code_cont*code_fumeu	Prev_ind	Fumeur	1	-0.30560	0.16494	3.4329	0.0639	.	.	.	code_contrat Prev_ind * code_fumeur Fumeur
code_cont*code_fumeu	Prev_ind	Non_fu	1	-0.31905	0.16150	3.9028	0.0482	.	.	.	code_contrat Prev_ind * code_fumeur Non_fu

FIGURE 5 – Sortie SAS de la procédure PHREG permettant de voir si des intercations entre variables existent

Annexe 8 : Présentation des courbes de mortalité obtenues avec le modèle de Cox stratifié

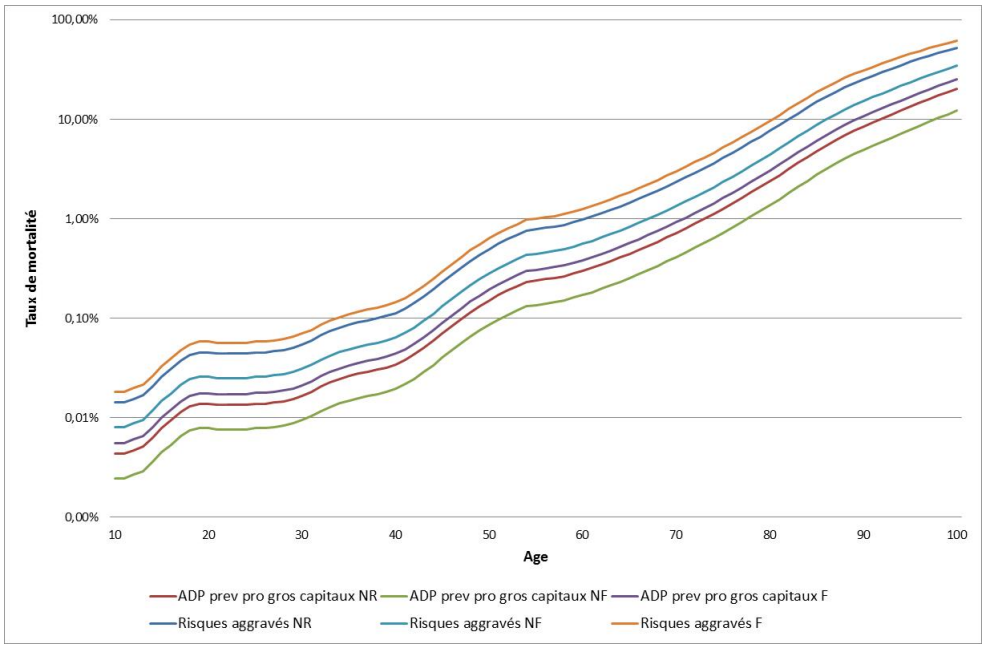


FIGURE 6 – Courbes de mortalité obtenues avec le modèle de Cox stratifié pour les femmes

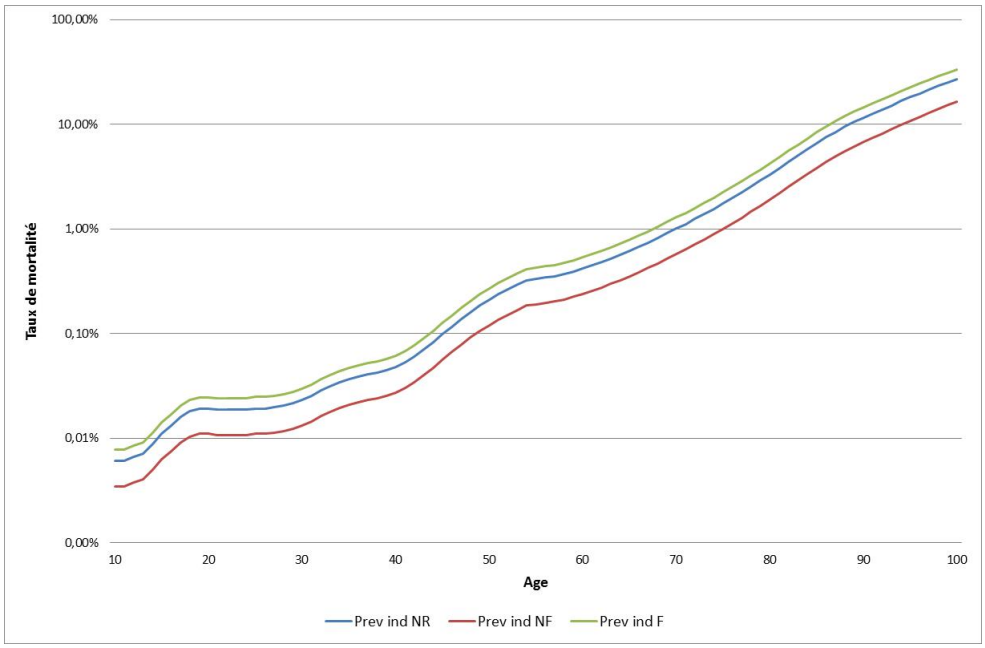


FIGURE 7 – Courbes de mortalité obtenues avec le modèle de Cox stratifié pour les femmes

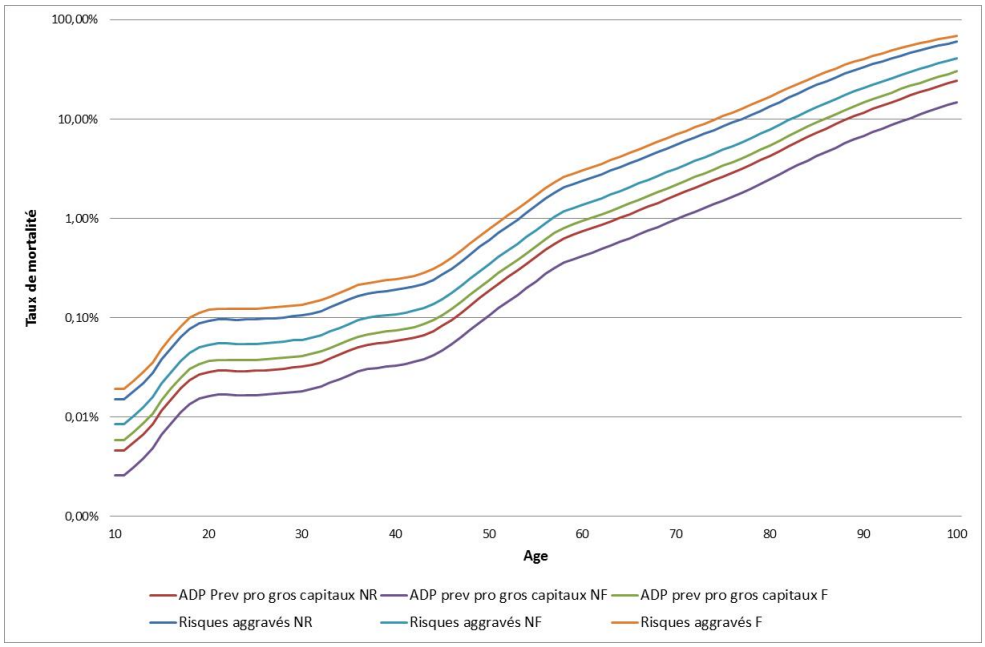


FIGURE 8 – Courbes de mortalité obtenues avec le modèle de Cox stratifié pour les hommes

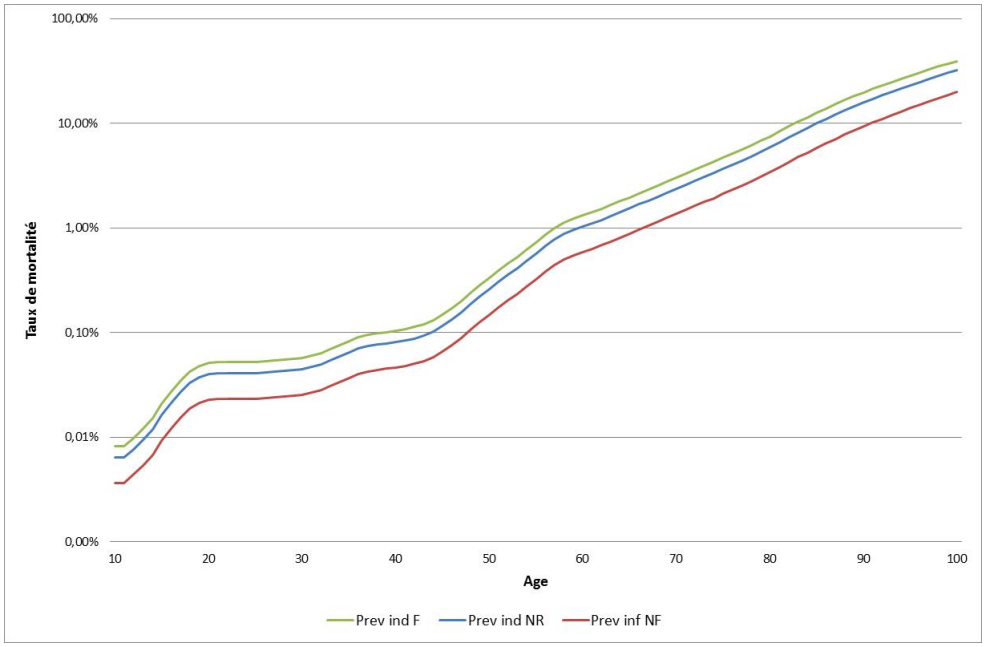


FIGURE 9 – Courbes de mortalité obtenues avec le modèle de Cox stratifié pour les hommes