ECOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ECONOMIQUE
3, avenue Pierre Larousse - 92245 Malakoff CEDEX, FRANCE

Mémoire d'Actuariat - Promotion 2008

CONSTRUCTION DE TABLES D'EXPERIENCE SEGMENTEES

Guillaume Alabergère et Jacky Phillips

MOTS-CLÉS : *Tables d'expérience, Contrats temporaires décès, Modèle à Hasard proportionnel de Cox, Modèle à hasard additif de Aalen, Tarification*

KEYWORDS: *Experience mortality tables, Term life insurance, Cox's proportionnal hazard model, Aalen additive hazard model, Ratemaking*

ENCADREMENT : Florence Faure (AXA France Solutions), Frédéric Planchet (IFSA)

CORRESPONDANT E.N.S.A.E. : Christian-Yann Robert

MEMOIRE CONFIDENTIEL

Table des matières

Remerciements

Nous remercions très chaleureusement notre tuteur Frédéric Planchet pour sa disponibilité et sa grande connaissance du milieu actuariel, ainsi que Christian-Yann Robert, notre correspondant à l'ENSAE, pour ses nombreux conseils.

Merci également à Marie Maedler et Frédéric Pourbaix pour leur aide concernant le traitement de la base de données et leur expérience du milieu professionnel de l'assurance.

Enfin, merci à Florence Faure pour nous avoir permis de travailler sur des données réelles, bien que non représentatives : un échantillon du portefeuille temporaire décès du groupe Axa.

Introduction

Les nouvelles normes comptables IFRS Phase II et la directive européenne Solvabilité 2 encouragent les assureurs à utiliser des hypothèses réalistes et prudentes, pour calculer leurs provisions et leurs tarifs. Il s'agit en effet d'aboutir à un nouveau système comptable plus adapté à la réalité des risques auxquels sont soumises les compagnies d'assurance, d'un point de vue à la fois comptable et prudentiel. Ainsi, concernant les contrats d'assurance sur la vie, ces dernières ont la possibilité de se baser sur des tables d'expérience certifiées par un actuaire indépendant plutôt que sur des tables réglementaires. Il s'agit par là même de mieux cerner les risques pour aboutir à un pilotage efficace et performant de l'activité d'assurance.

Parallèlement, lorsqu'une table de mortalité d'expérience se construit, le risque d'antisélection peut être limité par une recherche de segmentation en fonction de certaines caractéristiques de l'assuré comme le sexe, la CSP, l'aspect fumeur/non fumeur. Modéliser le comportement de chaque sous-population demande alors des bases de données suffisantes en termes d'effectifs, ce dont dispose rarement l'assureur. Pour limiter ce problème, on peut se tourner notamment vers des modèles de durée qui prennent en compte des variables explicatives et qui considèrent la population dans son ensemble. Alors, l'étude de tels modèles revient à rechercher l'effet de chaque variable caractéristique sur la mortalité de la population.

L'objet du mémoire consiste en la mise en parallèle de deux modèles de durée pour l'étude de la mortalité d'un portefeuille particulier d'assurés : d'une part le classique modèle de Cox à hasard proportionnel, permettant de positionner des individus par rapport à une référence, d'autre part le modèle additif d'Aalen plus long et difficile à mettre en oeuvre. Il s'agit de déterminer l'influence des caractéristiques des assurés sur leur mortalité et, in fine, de construire des tables de mortalité segmentées et valider leur pertinence suivant le modèle de Cox et le modèle d'Aalen.

Dans un premier temps, nous positionnons le problème en décrivant la théorie des modèles de durée et en cadrant notre étude par rapport à l'actualité comptable et prudentielle assurantielle. Ces bases étant posées, nous effectuons alors une étude poussée du portefeuille dont nous disposons pour établir l'influence de chaque caractéristique sur la mortalité suivant le modèle de Cox et le modèle d'Aalen, afin d'obtenir les tables segmentées. Enfin, nous appliquons les tables obtenues à la tarification du portefeuille. Nous dressons un comparatif entre les tables d'expérience et les tables réglementaires d'une part, les tables d'expérience du moment et prospectives d'autre part. Finalement, nous pourrons conclure sur l'intérêt des tables de mortalité d'expérience segmentées.

Chapitre 1

Présentation générale des modèles de durée pour l'étude de la mortalité

1.1 Les tables de mortalité

1.1.1 Définition des tables

Nous commençons par rappeler quelques généralités sur les tables de mortalité.

Le principe général d'une table de mortalité est de considérer un groupe fermé (c'est-à-dire ne prenant en compte aucun nouvel entrant et dont les seules sorties possibles sont causées par les décès des membres de ce groupe et les éventuelles résiliations) et de donner le nombre de survivants de ce groupe pour tous les âges. Pour construire une telle table, une étude statistique poussée d'une population doit être menée qui permet une modélisation des évolutions démographiques de cette population. Les tables de mortalité apparaissent donc comme l'outil privilégié de l'actuaire ou du démographe.

Intéressons-nous maintenant aux informations livrées par de telles tables. Elles donnent le nombre de survivants par âge, le nombre de décès par âge, le taux annuel de décès par âge, l'espérance de vie par âge, et indirectement, les probabilités de survie d'un âge à un autre et le taux de mortalité instantané.

1.1.2 Notations usuelles

Le nombre de survivants par âge renvoie aux personnes vivant encore à l'âge donné, dans un groupe fermé, pour 100 000 naissances. Dans la littérature actuarielle, il est noté l_x.

Le nombre de décès par âge s'assimile à la diminution de la taille de la population fermée entre deux âges consécutifs. Il est noté d_x .

$$d_x = l_{x-1} - l_x$$

Le taux annuel de décès par âge est écrit q_x

$$q_x = d_x/l_x$$

On définit d'autre part le supplémentaire, à savoir la probabilité annuelle de survie entre l'âge x et l'âge $x+1$:

$$p_x = 1 - q_x$$

La probabilité de survie d'un âge x à l'âge $x+n$, en utilisant les notations précédentes, est donnée par :

$$p_{xn} = l_{x+n}/l_x$$

L'espérance de vie à l'âge x est :

$$E_x = \sum_{i=1}^{\infty} l_{x+i}/l_x$$

Le taux de mortalité instantané est écrit sous la forme d'un taux annuel :

$$\mu_x \ = -(\frac{d}{dx}((l_x))/l_x$$

Quels sont les différents types de tables de mortalité utilisées par les compagnies d'assurance ?

1.1.3 Tables usuelles

Les tables du moment. Les tables de mortalité du moment sont réalisées à un moment précis dans le temps pour un groupe d'individus. Ainsi, toutes les informations contenues dans ces tables renvoient aux conditions de mortalité valables au moment choisi pour établir cette table.

Les tables prospectives. Il s'agit en particuliers de tables de mortalité propres aux rentes viagères qui permettent de prendre en compte l'augmentation de la durée de vie habituellement constatée depuis plusieurs années. Ainsi, elles présentent deux entrées : l'âge de l'assuré et l'année calendaire et permettent de calculer la probabilité de décéder au cours d'une certaine année pour chaque âge. Les assureurs sont maintenant dans l'obligation de se servir de ces tables pour provisionner et tarifer tous leurs contrats de rentes viagères, afin d'éviter une surestimation de la mortalité et donc finalement des pertes techniques. D'autre part, ces tables doivent être régulièrement révisées, l'évolution de la mortalité étant difficile à prévoir sur le long terme.

Les tables réglementaires. Ce sont des tables officielles fournies par des organismes ou des institutions tels que l'INSEE et auxquelles le Code des Assurances offre un rôle privilégié. Chaque table présente une origine et une utilisation spécifique à l'activité de l'assureur. Par exemple, la table TD 88/90 est une table du moment présentant la mortalité masculine française pour la période 1988-1990 et est

utilisée pour tarifer et provisionner les garanties décès, il s'agissait de la table de référence jusqu'à la fin de l'année 2005. Les assureurs utilisent maintenant les tables TH00-02 et TF00-02 pour de tels contrats.

Les tables d'expérience. A la différence des tables réglementaires, elles sont établies de manière spécifique à un portefeuille d'assurés. Le nombre de survivants à chaque âge est donc logiquement différent de celui donné par les tables réglementaires. Les tables d'expérience doivent être construites de manière prudente et être certifiées par un actuaire indépendant : décret sur les tables de mortalité, ARTICLE A335-1 du Code des Assurances.

Pourquoi utiliser des tables d'expérience ? En quoi ces dernières permettent-elles de mieux cerner le risque auquel est confronté l'assureur et de limiter le risque d'antisélection ?

1.1.4 Intérêt des tables d'expérience

Cerner le risque. Pourquoi établir des tables de mortalité d'expérience spécifiques au portefeuille et segmentées ? A priori, l'assureur n'a pas la possibilité de faire la différence entre les assurés selon leur probabilité de survenance du sinistre. Cela constitue le problème crucial de l'anti-sélection. En effet, s'il choisit de fixer un niveau de prime moyen pour l'ensemble de son portefeuille de contrats, les " bons risques " peuvent alors subventionner les " mauvais risques ". On peut même imaginer la situation où un assureur concurrent parvient lui à distinguer les assurés et proposer un tarif différencié. Le premier assureur subit une montée de la proportion de " mauvais risques " au sein de son portefeuille ce qui peut entraîner à terme une sous-tarification, voire des résultats négatifs, puisque les " bons risques " sont finalement incités à quitter cet assureur pour le concurrent. La segmentation du portefeuille selon quelques variables caractéristiques apparaît comme une solution naturelle pour faire face de manière efficace à ce problème d'anti-sélection. Il s'agit alors de fixer le niveau des primes pour chaque classe de risque, ou chaque segment de clientèle, plutôt que de manière globale. Le choix des variables de segmentation doit faire l'objet de tests de significativité statistique, afin de ne garder que les caractéristiques pertinentes pour le sinistre et le type de contrat concerné. Ainsi, le risque sera défini de la façon la plus juste pour chaque individu ou classe d'individus. Effectivement, une bonne appréciation du profil de sinistre est la clé d'un pilotage performant et efficace du risque d'assurance. Cerner le risque, et dans une plus large mesure, le maîtriser, constituent une des problématiques des nouvelles normes comptables IFRS Phase II harmonisées avec la directive européenne Solvabilité II.

Les nouvelles normes comptables et prudentielles. Les normes IFRS II et le projet Solvabilité II conduisent les assureurs à utiliser des hypothèses réalistes afin de prendre en compte de manière la plus complète possible l'expérience du portefeuille d'assurés. Il s'agit effectivement à travers une table de mortalité d'expérience de positionner la mortalité spécifique au portefeuille par rapport à une référence, telle les tables réglementaires par exemple. Ainsi, la construction d'une telle table passe par une exploitation des données fournies sur les assurés afin de dégager une loi de mortalité la plus fidèle possible pour la clientèle, le risque et le type de contrat

concerné. Il est à noter d'autre part que cette table doit être établie de manière raisonnablement prudente. Si les nouvelles normes comptables cherchent à dégager une image fidèle et nette de la situation économique de l'entreprise, il apparaît naturel de prendre en compte l'information complète fournie sur les assurés, plutôt que de se fier uniquement à des références externes, voire indépendantes. Il s'agit d'évoluer alors d'une évaluation prudentielle à une évaluation réaliste du risque, et donc des engagements de l'assureur, ce qui demande de ce fait le calcul d'une marge de risque explicite afin de pallier aux éventuelles erreurs d'estimation. Finalement, les tables d'expérience semblent être un outil privilégié dans ce contexte des prochaines normes, comme outil d'évaluation du Best Estimate des engagements de l'assureur pour les contrats décès.

Nous présentons par la suite la théorie nécessaire pour appréhender correctement les modèles de durée.

1.2 Principes généraux des modèles de durée - Modélisation sans variables exogènes

Les modèles de durée, comme leur nom l'indique en partie, ont pour but d'étudier plus précisément la durée d'un "état", d'un "statut", en dégageant notamment les facteurs caractéristiques influençant sa disparition ou son changement. Considérant la durée de cet état comme une variable aléatoire, cette modélisation est fréquemment utilisée par les biométriciens désireux d'étudier par exemple la durée d'une maladie en fonction d'un traitement. On retrouve également des applications dans les domaines démographiques (durée entre deux naissances...), économiques (durée d'une période de chômage). Mais ils sont aussi utilisables pour l'étude de la durée de vie humaine, et donc pour la construction des tables de mortalité. Ils permettent notamment de fournir un cadre d'étude statistique pour le taux instantané de décès μ à partir duquel on peut estimer les taux de décès par âge. Dans cette section et la suivante nous évoquerons des résultats généraux sur ces modèles de durée que nous réutiliserons par la suite sur notre portefeuille d'assurés. Avant d'évoquer la prise en compte de l'hétérogénéité, commençons par décrire ces modèles de façon globale.

1.2.1 Fonctions caractéristiques des modèles de durée

Dans les modèles de durée, cette dernière, au sens de durée de vie, est représentée par une variable aléatoire à valeurs dans $\mathbb{R}^+$, que l'on notera T. Outre la façon classique de caractériser et d'étudier la loi de cette variable aléatoire par l'intermédiaire de sa densité $f(t)$, il existe d'autres représentations utilisées fréquemment dans cette méthode de modélisation :

- La fonction de répartition

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

qui désigne la probabilité que la durée de vie de "l'état" considéré soit plus petite qu'une valeur donnée t.

– La fonction de survie

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

complémentaire à un de la fonction de répartition, qui désigne donc la probabilité la plus souvent pertinente dans les problématiques de durée de vie, celle d'être en vie au delà d'une date t. Par construction, S est bien évidemment décroissante.

– La fonction de hasard

$$h(t) = \lim_{s \to 0} \frac{1}{s} P\left(t \leq T \leq \ t + s | T \geq t\right)$$

qui représente le taux instantané de sortie, ou autrement dit la densité de probabilité de décéder à t sachant qu'on est vivant à t. On l'appelle également la fonction de risque.
On peut la relier aisément à la densité $f(t)$ et à la survie $S(t)$ précédemment introduites puisqu'on peut réécrire

$$h(t) = \lim_{s \to 0} \frac{1}{s} \frac{P(t \leq T \leq t + s)}{P(T \geq t)}$$

$$\text{avec } f(t) = \lim_{s \to +\infty} \frac{1}{s} P(t \leq T \leq t + s) \text{ et } S(t) = \int_t^{+\infty} f(u)du$$

d'où

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

D'où la remarque suivante : plus le hasard est élevé, plus la survie décroît rapidement

– La fonction de hasard intégré

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

Des calculs simples mènent aisément à

$$H(t) = -\ln S(t) \quad S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}$$

– La durée moyenne restante

$$r(t) = E(T - t | T > t)$$

– La survie conditionnelle. Introduisons enfin la survie conditionnelle$S_u(t)$, qui correspond à la probabilité de survie d'un individu après un instant t sachant qu'il est vivant en u. On a alors

$$S_u(t) = P\left(T > u + t | T > u\right) = \frac{P\left(T > u + t\right)}{P\left(T > u\right)} = \frac{S(u + t)}{S(u)}$$

Ces fonctions caractéristiques étant posées, quelles sont les lois usuelles utilisées pour étudier les modèles de durée ?

1.2.2 Lois usuelles

Plusieurs types de densité $f(t)$ sont souvent utilisés lorsqu'il s'agit d'étudier un problème de modèle de durée. On trouve ainsi :

- La loi exponentielle pour laquelle on a :

$$f(t) = \lambda e^{-\lambda t} \quad S(t) = e^{-\lambda t} \quad h(t) = \lambda \quad r(t) = \frac{1}{\lambda}$$

  La durée moyenne de vie restante reste donc constante dans le temps. On dit que "la loi exponentielle n'a pas de mémoire". Cependant, n'ayant qu'un paramètre, cette dernière n'est pas très flexible.

- La loi gamma, dont la densité s'exprime par $f(t) = a^p t^{p-1} e^{-at} \Gamma(p)^{-1}$, a elle une fonction de hasard monotone, croissante si p>1 (on parle alors de "vieillissement"), décroissante sinon (on parle alors de "rodage").

- La loi lognormale, dont le hasard présente un mode, avec une densité : $f(t) = \frac{1}{\sigma t \sqrt{2\Pi}} e^{\frac{1}{2\sigma^2}(\ln t - m)^2}$
- La loi de Weibull, caractérisée par la densité suivante :

$$f(t) = abt^{b-1} e^{-at^b}$$

  Compte tenu de leur lien avec la densité, les autres fonctions caractéristiques ont pour expression dans un modèle Weibull :

$$h(t) = abt^{b-1} \quad H(t) = at^b \quad S(t) = e^{-at^b}$$

  On remarque donc que le hasard y est monotone. Si b>1, le hasard est croissant : on parle de dépendance temporelle positive, c'est-à-dire que plus le temps passe, plus les chances de survie s'amenuisent vite, conformément à la logique de "durée de vie" (vieillissement). Si b<1, le hasard est décroissant, la dépendance temporelle négative (rôdage). Enfin si b=1, on retrouve une loi exponentielle à hasard constant.
- Nous terminerons cette liste d'exemples par le modèle de Gompertz-Makeham, couramment utilisé pour la construction de tables de mortalité ou de tables de maintien en arrêt de travail. Il se caractérise par la fonction de hasard suivante :

$$h(t) = a + bc^t$$

  La forme de cette fonction de hasard a une explication "démographique" : le paramètre a représente un taux de sortie ("décès") accidentel indépendant de l'âge, le terme en bc^t expliquant lui le "vieillissement". A noter que par rapport à un modèle de type Weibull, le vieillissement est plus rapide avec Gompertz-Makeham, puisqu'exponentiel.

1.2.3 Censure-Troncature

Nous continuons cette présentation générale des modèles de durée en évoquant un problème courant posé par les observations : celles-ci sont la plupart du temps incomplètes. En effet, en ce qui concerne notre portefeuille d'assurés dont nous voulons étudier la mortalité, nous n'allons pas avoir la durée de vie réelle des individus présents dans la base de données. D'une part nous n'observerons que la vraie durée de vie des personnes qui seront décédées durant la période du contrat, n'obtenant pour les autres qu'une borne inférieure de leur durée de vie. On appelle ce phénomène censure à droite.

D'autre part, ces individus ne seront pas observés depuis leur naissance, mais simplement depuis la souscription de leur contrat. Ainsi nous disposerons pour chacun d'eux d'une date d'entrée en observation. Ce phénomène portant le nom de troncature.

Afin de fixer quelques notations que nous réutiliserons par la suite, présentons de façon plus formelle ces deux phénomènes tels que nous y ferons face dans notre jeu de données.

Soient un échantillon de durées de survie $(T_1, \ldots, T_n)$ et un second échantillon indépendant composé de variables positives $(C_1, \ldots, C_n)$. Dans notre contexte, au lieu d'observer directement $(T_1, \ldots, T_2)$, on observe en réalité $(X_1, d_1), \ldots, (X_n, d_n)$ avec

$$X_i = min(T_i, C_i) \text{ et } d_i = \begin{cases} 1 \text{ si } T_i \leq C_i \\ 0 \text{ si } T_i > C_i \end{cases}$$

De plus les individus ne sont pas observés depuis l'origine mais depuis l'age atteint en début de la période d'observation que l'on notera E_i.

1.2.4 Les processus ponctuels

Une autre façon d'étudier le comportement de la variable aléatoire T consiste à raisonner avec le processus ponctuel qui lui est naturellement associé. Ce processus est défini par

$$N(t) = \mathbf{1}_{T<t}$$

Il vaut donc 0 tant qu'il n'y a pas eu de sortie (décès), et 1 après.

Nous reprenons ici la présentation de ce processus et des mécanismes mis en jeu dans une problématique de durée effectuée par Planchet et Thérond [2006]. Cette présentation se veut succinte et a surtout pour but d'introduire quelques concepts mathématiques et le vocabulaire associé qui nous seront utiles notamment lors de la description du modèle de Aalen.

Définition 1 Un *processus ponctuel* $(N(t), t \geq 0)$ est un processus adapté à une filtration $(F_t)_{t\geq 0}$ tel que $N(0) = 0$, $N(t) < +\infty$ presque sûrement et tel que ses trajectoires soient continues à droite, constantes par morceaux et ne présentent que des sauts d'amplitude +1.

On note $(F_t)_{t\geq 0}$ la filtration naturelle associée à ce processus $(F_t = \sigma\left(N(u), 0 \leq u \leq t\right))$.

Définition 2 Un processus prévisible est une variable aléatoire mesurable définie sur l'espace produit $]0, +\infty] \times \Omega$ muni de la tribu P engendrée par les ensembles de la forme $]s, t] \times \Gamma$, avec $\Gamma \in F_s$.

Rappelons enfin cette proposition, qui permet d'introduire la notion de compensateur et de martingale pour les processus ponctuels :

Proposition Soit $(N(t), t \geq 0)$ un processus ponctuel adapté à la filtration $(F_t)_{t\geq 0}$ tel que $\mathbb{E}[N(t)] < +\infty$. Il existe alors un unique processus croissant continu à droite Λ tel que $\Lambda(0) = 0$ et $\mathbb{E}[\Lambda(t)] < +\infty$ et tel $M(t) = N(t) - \Lambda(t)$ soit une martingale. En outre, lorsque Λ peut se mettre sous la forme

$$\Lambda(t) = \int_0^t \lambda(u)du \,,$$

le processus λ s'appelle l'intensité du processus ponctuel.

D'un point de vue heuristique, ce résultat exprime que le processus N "oscille" autour d'une tendance prévisible Λ (le compensateur), de sorte que la différence entre le processus N et cette tendance soit assimilable à un "résidu" dont on maîtrise les variations.

Dans le cadre des modèles de durée on s'intéresse alors au comportement de N(t) qui dépend de la durée de vie T. Notons tout d'abord $N(t^-)$ la limite à gauche de $N(t)$. Soit de plus $dN_t = N(t + dt) - N(t)$ variable aléatoire qui ne peut prendre que les valeurs 0 et 1, dt étant une grandeur infinitésimale. On s'intéresse alors à la loi de la variable aléatoire $\mathbb{P}(dN_t = 1|N(t^-))$ qui correspond à la probabilité de décès de l'individu entre les instants t et $t + dt$ connaissant son état juste avant t. Cet état étant incertain, cette grandeur est donc bien une variable aléatoire. Conformément aux définitions du hasard h et de la survie S introduits précédemment, on peut affirmer que l'on a :

$$\mathbb{P}(dN_t = 1|N(t^-)) = h(t)dt \text{ si } t \leq T, \text{ ie avec la probabilité } S(t)$$

$$\mathbb{P}(dN_t = 1|N(t^-)) = 0 \text{ si } t > T, \text{ ie avec la probabilité } 1 - S(t)$$

En effet, si l'individu est déjà mort, ie que le processus N a déjà sauté avant la date t, ce qui arrive avec la probabilité $1 - S(t)$, dN_t ne peut plus valoir 1. De la même façon, pour que dN_t puisse valoir 1 il faut que $N(t^-) = 0$ ce qui arrive avec une probabilité $S(t)$, auquel cas la probabilité de saut aux instants considérés est de $h(t)dt$ par définition du hasard.

On pose alors $\lambda(t) = h(t)\mathbf{1}_{T\geq t}$. Compte tenu des résultats précédents, on peut alors affirmer que M définie par $M(t) = N(t) - \int_0^t \lambda(u)du$ est une martingale. En effet, l'espérance de $M(t)$ est bien évidemment finie, et on a en outre pour $t \geq 0$:

$$\mathbb{E}(dM_t|F_{t^-}) = \mathbb{E}(dN_t - \lambda(t)dt|F_{t^-}) = \mathbb{E}(dN_t|F_{t^-}) - \lambda(t)dt$$

$$\mathbb{E}(dM_t|F_{t^-}) = 1.\mathbb{P}(dN_t = 1|F_{t^-}) + 0.\mathbb{P}(dN_t = 0|F_{t^-}) - \lambda(t)dt$$

$$\mathbb{E}(dM_t|F_{t^-}) = \mathbf{1}_{t\leq T}h(t)dt - \lambda(t)dt = 0$$

$\lambda(t)$ est donc l'intensité du processus ponctuel $N(t)$ qui se définit donc à partir de la fonction de hasard et d'un indicateur de présence à risque. C'est un processus prévisible.

Dans le cas où l'on s'intéresse au processus des événements non censurés $N^1(t) = \mathbf{1}_{\{X<t,d=1\}}$, le compensateur s'écrit

$$\lambda^1(t) = \int_t^0 R(u)h(u)du$$

avec $R(t) = \mathbf{1}_{X\geq t}$ indicatrice de présence à risque avant t (ie la fonction valant 0 si l'individu est sorti du cadre d'observation pour quelque raison que ce soit : censure, décès ...) Dans le cas d'une population dont on suppose que tous les individus ont la même fonction de hasard h, on associe à tous les individus un processus d'événements non censurés $N_i^1(t) = \mathbf{1}_{\{X_i<t,d_i=1\}}$ ainsi que l'indicatrice de présence sous risque, comptabilisant les individus ni morts ni censurés $R_i(t) = \mathbf{1}_{X_i\geq t}$ et on peut construire des processus agrégés

$$\overline{R}(t) = \sum_{i=1}^{n} R_i(t) \text{ et } \overline{N}^1(t) = \sum_{i=1}^{n} N_i^1(t)$$

qui comptabilisent respectivement l'effectif sous risque et le nombre de décès réellement observés (non censurés).

On dit alors que le processus de "comptage" $\overline{N}^1(t)$ possède une intensité qui se met sous la forme $\lambda(t) = \overline{R}(t)h(t)$ avec h fonction de hasard qu'il reste à estimer.

1.2.5 Estimation

Les fonctions représentatives d'un modèle de durée ayant été introduites, intéressons-nous à la façon pratique de les obtenir à partir d'un échantillon de données. Comme il est d'usage, on cherche en effet à trouver le modèle qui explique le mieux la répartition des données. Une double approche est alors possible. Une estimation non paramétrique à partir des données brutes, ou bien une estimation paramètrique en faisant l'hypothèse de l'existence d'une loi du type de celles présentées ci-avant avec des paramètres inconnus que l'on va estimer à l'aide des données.

Méthode non paramétrique

Estimateur de Kaplan-Meier

On cherche tout d'abord à estimer la fonction de survie, qui donnera une idée sur la loi suivie par les durées observées. Kaplan et Meier (1958) proposent alors l'estimateur suivant de la survie, en supposant une seule sortie non-censurée par date de l'échantillon :

$$\hat{S}_n(t) = \prod_{\substack{t_i<t \\ t_i \text{ non censurée}}} \left(1 - \frac{1}{r_i}\right)$$

où r_i désigne le cardinal de l'ensemble à risque en t_i, c'est-à-dire le nombre d'individus dont la durée de vie est connue supérieure à t_i. On peut même remarquer

que l'estimateur de la survie écrit sous cette forme apparaît comme un produit de probabilités conditionnelles.

$$\hat{S}_n(t) = \hat{P}_n(T > t) = \prod_{t_i < t} \hat{P}_n(T > t_i | T > t_{i-1})$$

En effet, un estimateur empirique naturel des probabilités présentes dans ce produit est bien évidemment $1 - \frac{1}{\text{cardinal sous risque}}$ si l'on suppose une seule mort par date d'observation. C'est d'ailleurs de cette façon que l'on peut justifier heuristiquement l'allure de cet estimateur, construit par conditionnements successifs de la probabilité définissant la survie en fonction des dates d'observation des sinistres.

Il est à noter que l'arrivée en cours d'observation de nouveaux individus ne change pas l'allure de l'estimateur, puisqu'il suffit d'incorporer ces nouveaux arrivants dans le calcul de r_i. Ceci est important dans notre contexte où l'on n'observe pas les individus depuis l'origine (troncature).

Dans le cas des "départs" simultanés, D_i départs à la date t_i, la formule d'estimation empirique de la survie devient

$$\hat{S}_n(t) = \prod_{\substack{t_i < t \\ t_i \text{ non censurée}}} \left(1 - \frac{D_i}{r_i}\right)$$

Il s'agit donc d'une fonction en escalier décroissante, avec des sauts au niveau de chaque sortie non censurée.

Bien que biaisé à distance finie, on peut montrer que cet estimateur est convergent de distribution connue (asymptotiquement Gaussien cf Gill [1980]).

Un estimateur empirique de sa variance connu sous le nom de Greenwood peut-être donné par

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{\substack{t_i < t \\ t_i \text{ non censurée}}} \frac{D_i}{r_i(1 - r_i)}$$

Il permet, avec la normalité asymptotique de l'estimateur de Kaplan-Meier, de calculer des intervalles de confiance (asymptotiques) dont les bornes sont en t_i (sortie non censurée)

$$S(t_i)\left(1 \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{D_1}{r_1(1-r_1)} + \frac{D_2}{r_2(1-r_2)} + \ldots + \frac{D_i}{r_i(1-r_i)}}\right)$$

où $u_{1-\frac{\alpha}{2}}$ désigne le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

La principale limite de cet estimateur reste tout de même qu'on suppose implicitement qu'une même loi se cache derrière les observations. Cet estimateur suppose donc qu'il n'y a aucune hétérogénéité.

Méthode paramétrique

Les techniques de régression linéaire (moindres carrés ordinaires) ne sont pas utilisées, en général, dans l'estimation de modèles de durée. On pourrait par exemple penser à écrire $\ln T = X.b + u$ où T est la durée de vie, X des variables exogènes et u une perturbation. Mais la censure des données rend la méthode des MCO non convergente. La méthode retenue dans le cas paramétrique, c'est-à-dire lorsqu'un modèle spécifiquement défini par un jeu de paramètres veut être estimé, est le plus souvent la méthode du maximum de vraisemblance. Toujours dans ce souci de présenter le formalisme propre aux modèles de durée, écrivons cette vraisemblance dans le cas où l'on dispose d'un jeu de n observations (x_i, d_i) où d_i est une variable d'indication de la censure : d_i vaut 1 si la donnée est non-censurée et d_i vaut 0 sinon.

Conditionnellement à la censure, la contribution à la vraisemblance d'une observation non-censurée est $f(x_i)$ où f désigne la densité. Une observation censurée contribue elle pour $S(x_i)$ à la vraisemblance. On a donc

$$L(x_1...x_n|d_1...d_n) = \prod_{i=1}^{n} f(x_i)^{d_i} S(x_i)^{1-d_i}$$

la log vraisemblance se met donc sous la forme

$$\ln L(x_1...x_n|d_1...d_n) = \sum_{i=1}^{n} d_i \ln f(x_i) + \sum_{i=1}^{n} (1-d_i) \ln S(x_i)$$

$$\ln L(x_1...x_n|d_1...d_n) = \sum_{i=1}^{n} d_i \ln h(x_i) + \sum_{i=1}^{n} \ln S(x_i)$$

$$\ln L(x_1...x_n|d_1...d_n) = \sum_{i=1}^{n} d_i \ln h(x_i) - \sum_{i=1}^{n} H(x_i) \tag{1.1}$$

où h et H sont respectivement les fonctions de hasard et de hasard intégré.
On peut alors trouver un estimateur des paramètres spécifiant la loi par la méthode classique de maximisation de la vraisemblance.
Un exemple analytique "simple" : Pour un hasard de type Weibull $h(t) = abt^{b-1}$, on trouve donc

$$\ln L(x_1...x_n|d_1...d_n) = \sum_{i=1}^{n} d_i(\ln a + \ln b + (b-1)\ln x_i) - a\sum_{i=1}^{n} x_i^b$$

d'où des estimateurs vérifiant

$$\begin{cases} \hat{a} = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} x_i^b} \\ \\ \sum_{i=1}^{n} d_i + \hat{b}\sum_{i=1}^{n} d_i \ln x_i - \hat{a}\hat{b}\sum_{i=1}^{n} \ln x_i x_i^{\hat{b}} = 0 \end{cases}$$

Nous nous intéressons maintenant à l'hétérogénéité de la population d'individus constituant le portefeuille de l'assureur.

1.3 Prise en compte des variables exogènes

Les outils et méthodes de la section précédente permettent donc de déterminer et d'estimer la survie d'un ensemble d'individus. Toutefois les populations étudiées ne sont jamais homogènes et de nombreux modèles ont cherché à prendre en compte l'hétérogénéité des individus dans l'estimation de la survie. Pour en revenir au sujet de ce mémoire, les outils et méthodes présentés ci-après pourront nous permettre d'évaluer quantitativement l'influence de caractéristiques individuelles sur les taux de décès des assurés de notre portefeuille tout en évitant de devoir segmenter au préalable la base de données en fonction de ses caractéristiques. Après avoir mis en évidence l'intérêt de la prise en compte de l'hétérogénéité, nous introduirons deux modélisations la prenant en compte directement à partir d'un échantillon de données : le modèle à hasard proportionnel de Cox et le modèle additif de Aalen.

1.3.1 Remarque sur l'hétérogénéité

L'oubli de variables exogènes dans la spécification d'un modèle se traduit souvent par des biais sur les coefficients estimés. C'est aussi le cas dans les modèles de durée.

Un exemple classique sur une idée de Lancaster (79) illustre ce propos.
Considérons une population partagée en deux groupes. Comme chacun des groupes est homogène, dans le premier, le hasard vaut 1, et dans le second, il vaut 2. La fonction de survie associée à un individu du premier groupe (resp. deuxième) est donc de e^{-t} (resp. e^{-2t}). Supposons qu'au début de l'étude (t=0), les deux groupes sont de tailles égales. A un instant t, la probabilité qu'un individu appartienne au premier (resp. deuxième) groupe est

$$p_1(t) = \frac{e^{-t}}{e^{-t} + e^{-2t}} \quad \left(\text{resp.} \frac{e^{-2t}}{e^{-t} + e^{-2t}}\right)$$

Le rapport de ces deux probabilités est e^{-t} : à t=0 ces probabilités sont égales, puis $p_2(t)$ diminue alors que $p_1(t)$ augmente. La proportion d'individu dont le hasard est 1 augmente donc dans l'échantillon, ce qui est normal, puisqu'ils en sortent moins vite que ceux dont le hasard est 2. Si on réalise une estimation en supposant la population homogène, on va en fait estimer la moyenne sur la population, des hasards individuels soit :

$$E_t(\lambda(t)) = \frac{e^{-t}}{e^{-t} + e^{-2t}} + 2\frac{e^{-2t}}{e^{-t} + e^{-2t}}$$

Cette fonction est décroissante : cela correspond bien au fait que plus t augmente, moins il reste, relativement, d'individus dont le hasard est grand.

Ainsi dans cet exemple, les hasards au niveau individuel sont constants, alors que le hasard agrégé observé est décroissant. Ce phénomène est souvent appelé effet "mover stayer".

Plus généralement, l'effet de l'hétérogénéité non contrôlée est que l'on surestime la décroissance de la fonction de hasard, au sens où le hasard observé au niveau global

à tendance à décroître plus rapidement que la moyenne des hasards individuels. Formellement, supposons que la fonction de hasard associée à un individu soit $h(t, \theta)$ θ étant distribué suivant la loi $\pi_t(\theta)$ à l'instant t. Fourgeaud-Gourièroux-Pradel [1987] ont alors montré que :

$$\frac{\partial}{\partial t}\mathbb{E}_t(h(t,\theta)) = \mathbb{E}_t\left(\frac{\partial}{\partial t}h(t,\theta)\right) - \mathbb{V}_t(h(t,\theta))$$

où les espérances et variances sont prises par rapport à $\pi_t(\theta)$.

Les diverses figures qui suivent illustrent les différences de survie lorsqu'on ne prend pas en compte l'hétérogénéité. Nous avons effectué sur une base de données, où la variable sexe était renseignée pour chaque individu, plusieurs estimations de la survie par la méthode Kaplan-Meier. L'une a été effectuée sur l'ensemble des données sans distinguer les personnes d'une quelconque manière. Dans un second temps nous avons séparé dans la base hommes et femmes. On remarque alors (benchmark à 60 ans sur les figures) que du point de vue de la survie, ces sous groupes de populations se distinguent entre eux et de la population globale.

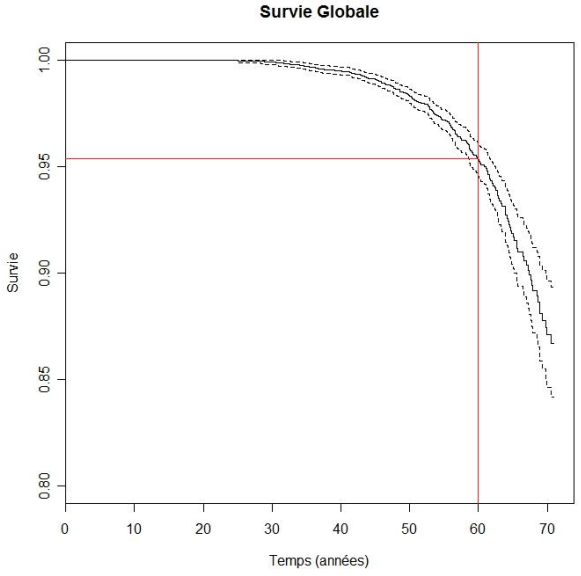


FIG. 1.1 – Survie estimée à partir d'un échantillon hétérogène

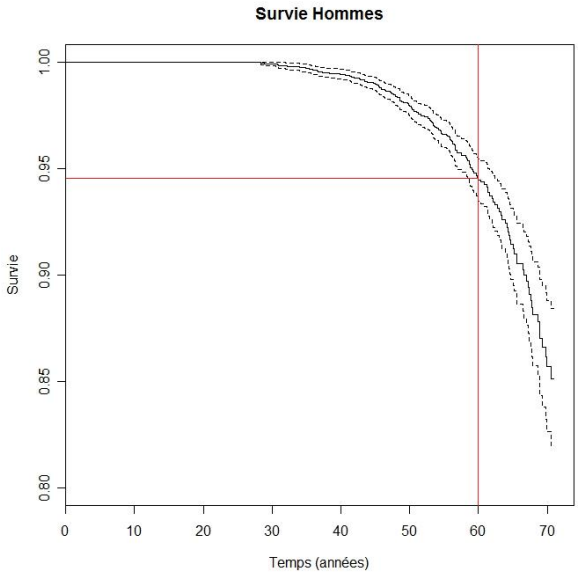


FIG. 1.2 – survie estimée des hommes

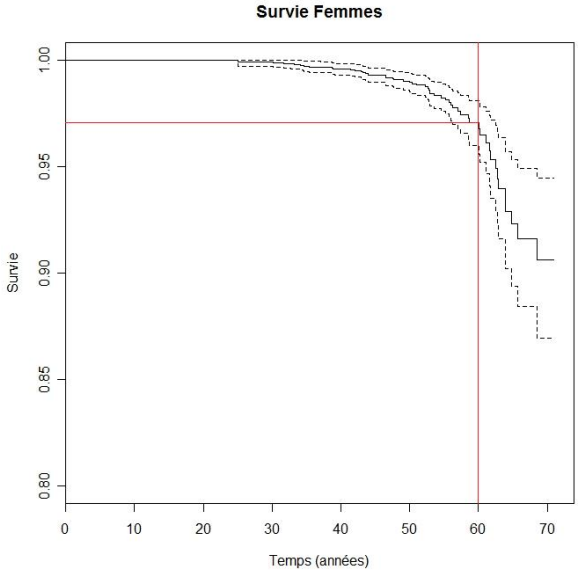


FIG. 1.3 – survie estimée des femmes

Comment prendre en compte cette hétérogénéité des individus dans l'estimation des données de survie, sans pour autant segmenter la base de données initiale ? C'est ce que nous allons voir en présentant deux modèles qui prennent en compte l'influence de variables explicatives dans la modélisation de durée.

1.3.2 Modèle à hasard proportionnel - Cox

C'est sans nul doute la modélisation la plus utilisée, introduite par Cox en 1972, lorsqu'il s'agit de prendre en compte l'hétérogénéité de l'échantillon de données afin d'estimer la durée de vie des individus en fonction de leurs caractéristiques. On le nomme d'ailleurs aussi modèle de Cox.

Modélisation Cox

La fonction de hasard y est exprimée en tant que produit d'une fonction de hasard dite de base $h_0(t)$ et d'une fonction des variables exogènes $\Psi(Z, \beta)$

$$h(t, Z, \beta) = h_0(t)\Psi(Z, \beta)$$

La fonction de survie associée s'écrit

$$S(t, Z, \beta) = S_0(t)^{\Psi(Z,\beta)}$$

Le plus souvent on choisit pour $\Psi(Z, \beta)$ une expression du type $\Psi(Z, \beta) = e^{Z\beta}$.

On peut donc remarquer que dans ce modèle, les covariables ont un effet multiplicatif par rapport à une fonction de hasard de base.

Estimation des coefficients

Dans ce cas, pour estimer les paramètres introduits et notamment les coefficients β liés aux variables exogènes Z, on peut utiliser la méthode paramétrique présentée précédemment en spécifiant explicitement $h_0(t)$, et en écrivant la vraisemblance résultante.

Toutefois, un des attraits du modèle de Cox est de permettre d'estimer les coefficients β sans avoir à estimer, et donc sans avoir à spécifier, $h_0(t)$, en introduisant une vraisemblance partielle associée aux observations.

Supposons que l'on dispose des observations de n individus aux caractéristiques Z_i, auxquelles on associe une durée de fin d'observation X_i qui correspond ou non à une durée censurée. On note leur fonction de hasard $h_i(t) = h_0(t)e^{Z_i\beta}$.

Parmi la suite des X_i on note $\tau_1 < ... < \tau_d$ les durées correspondant à des sorties non censurées, ordonnées par ordre croissant.

On note $I_j = i$ si et seulement si $X_i = \tau_j$. Soit $R(\tau_j) = \{i | X_i \geq \tau_j\}$ l'ensemble à risque à τ_j.

Soit H_j l'histoire du processus en τ_j, c'est-à-dire l'ensemble des durées de vie et des données censurées avant τ_j, l'information I_j exceptée.
La probabilité conditionnelle pour que $I_j = i$, compte tenu de l'histoire H_j, est la probabilité que l'individu i "meurt" en τ_j sachant qu'une personne de l'ensemble à risque $R(\tau_j)$ "meurt" en τ_j. D'après la définition du hasard, c'est donc simplement :

$$p(I_j = i|H_j) = \frac{h_i(\tau_j)}{\sum_{k\in R(\tau_j)} h_k(\tau_j)}$$

La fonction de hasard de base présente dans chaque terme de cette expression disparaît donc, et on peut exprimer cette probabilité conditionnelle uniquement en fonction de β :

$$p_j(i|H_j) \stackrel{def}{=} p(I_j = i|H_j) = \frac{e^{Z_i\beta}}{\sum_{k\in R(\tau_j)} e^{Z_k\beta}} \tag{1.2}$$

Or la vraisemblance totale peut s'écrire en supposant une censure non informative indépendante des sorties (cf Cox-Oakes [1984]) :

$$L = g_{d+1}(H_{+\infty}|H_d, I_d) \prod_{j=1}^{d} [g_j(\tau_j, H_j|H_{j-1}, I_{j-1}) p_j(i|H_j)]$$

où g_j désigne la densité jointe conditionnelle de la jème durée de vie observée, et de chaque donnée censurée contenue dans l'intervalle $[\tau_{j-1}, \tau_j]$. Autrement dit les g_j contiennent toute l'information pour la répartition de ce qui se passe entre deux sorties observées, connaissant "l'histoire" jusqu'alors.

L'idée de Cox est alors de sélectionner uniquement le produit

$$\prod_{j=1}^{d} p_j(i|H_j) = \prod_{j=1}^{d} \frac{e^{Z_i\beta}}{\sum_{k\in R(\tau_j)} e^{Z_k\beta}}$$

qu'on appelle vraisemblance partielle, puisque ce produit en possède des propriétés similaires, tout en ne dépendant que du seul paramètre β du modèle d'après (1.2). A noter que cela revient en fait à ne pas prendre en compte la contribution à la vraisemblance des observations censurées, puisqu'il s'agit d'une écriture avec les seules sorties effectives.

En maximisant en β cette vraisemblance partielle, on peut montrer (Andersen et Gill [1982]) que l'on obtient un estimateur $\hat{\beta}$ convergent et asymptotiquement normal, et semi-paramétriquement efficace, c'est-à-dire de variance minimale si on cherche un estimateur de β sans spécifier $H_0(t)$. Cox [1984], a d'ailleurs étudié la différence de cette estimation avec celle obtenue par une maximisation de la vraisemblance totale. Celui-ci en conclut que les résultats sont significativement proches à condition que la vraie valeur du paramètre β soit très différente de 0 et que la répartition des données censurées soit indépendante des variables exogènes.

Plus précisément, ce comportement asymptotique de $\hat{\beta}$, en notant β_0 la vraie valeur du paramètre (vecteur de taille p : $\beta_0 = (\beta_0^1, \ldots, \beta_0^p)$, constitué d'un paramètre β_0^i par covariable Z_i), peut être décrit par :

$$\frac{\hat{\beta} - \beta_0}{\sqrt{I^{-1}(\beta_0)}} \sim \mathbf{N}(0, I_p)$$

où

$$I(\beta_0) = -\frac{\partial^2 \ln L_{par}(\beta)}{\partial\beta\partial\beta'}|_{\beta=\beta_0}$$

désigne l'information de Fisher associée à cette vraisemblance partielle.

En pratique, comme il est difficile d'évaluer $I(\beta_0)$ on préfère lui substituer $\hat{I} = I(\hat{\beta})$.

Fort de cette distribution, il est alors possible d'utiliser les tests classiques de significativité de la totalité ou d'une partie des coefficients.

Considérons une partition du vecteur $\beta = (\beta_1', \beta_2')$ où β_1 est le vecteur (taille q×1) constitué des coefficients dont on veut tester la significativité, et β_2 vecteur constitué des autres coefficients.

On note alors

$$I(\beta) = \left(-\frac{\partial^2 \ln L_{par}(\beta)}{\partial\beta\partial\beta'}\right) = \begin{pmatrix} I_{11}(\beta) & I_{12}(\beta) \\ I_{21}(\beta) & I_{22}(\beta) \end{pmatrix}$$

avec donc notamment $I_{11}(\beta)$ matrice de taille q×q correspondant aux dérivations par rapport aux coefficients de β_1, $I_{22}(\beta)$ matrice correspondant aux dérivations par rapport aux coefficients de β_2 etc ...
On note par ailleurs l'inverse de cette matrice d'information sous la forme bloc

$$I^{-1}(\beta) = \begin{pmatrix} J_{11}(\beta) & J_{12}(\beta) \\ J_{21}(\beta) & J_{22}(\beta) \end{pmatrix}.$$

Pour tester l'hypothèse nulle

$$H_0 : \ \beta_1 = 0$$

3 tests classiques peuvent être mis en place, en notant naturellement $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ l'estimateur de β.

- Test de Wald : de statistique $\chi_w = \hat{\beta_1}' \left[J_{11}(\hat{\beta})\right]^{-1} \hat{\beta_1}$
- Test du rapport de vraisemblance : de statistique $\chi_{rv} = 2\left(\ln L_{par}(\hat{\beta}) - \ln L_{par}((0, \hat{\beta}_2(0)')')\right)$ où $\hat{\beta}_2(0)$ désigne l'estimateur de β_2 sous la contrainte $\beta_1 = 0$
- Test du score : de statistique $\chi_s = U_1((0, \hat{\beta}_2(0)')')' \left[J_{11}((0, \hat{\beta}_2(0)')')\right] U_1((0, \hat{\beta}_2(0)')')$ où l'on note le vecteur score

$$U(\beta) = \frac{\partial}{\partial\beta} \ln L_{par}(\beta) = \begin{pmatrix} U_1(\beta) = \frac{\partial}{\partial\beta_1} \ln L_{par}(\beta) \\ U_2(\beta) = \frac{\partial}{\partial\beta_2} \ln L_{par}(\beta) \end{pmatrix}$$

En effet, asymptotiquement, ces trois statistiques ont approximativement une distribution de type khi-deux à q degrés de liberté lorsque l'hypothèse nulle est vérifiée, compte tenu du comportement asymptotique de $\hat{\beta}$.

1.3.3 Modèle additif de Aalen

Le modèle de Aalen ou modèle à hasard additif a été introduit par Aalen en 1980. Contrairement à la modélisation de Cox semi-paramétrique, ce modèle est non-paramétrique et permet d'estimer directement de façon absolue le hasard tout en s'attachant à mesurer l'impact de différentes covariables sur celui-ci. Ce modèle est donc une alternative théorique à celui de Cox dans notre objectif d'estimation de la survie de façon segmentée. Toutefois, en raison du manque de logiciels dédiés à son implémentation, il est encore relativement peu utilisé en pratique, au contraire du modèle de Cox beaucoup plus présent dans les logiciels.

En ce qui concerne la théorie, nous nous sommes appuyés essentiellement sur l'ouvrage de Martinussen et Scheike [2006], *Dynamic Regression Models for Survival Data* et plus précisément sur le chapitre 5 sur les modèles à hasard additif, auquel nous renvoyons le lecteur pour un approfondissement des résultats présentés ici.

Dans cette modélisation, le hasard est supposé s'écrire

$$h(t) = {}^tZ(t)\beta(t)$$

où ${}^tZ(t) = (Z_1(t), \ldots, Z_p(t))$ désigne un vecteur de variables explicatives (prévisibles) et $\beta(t)$ un processus p-dimensionnel localement intégrable. En se référant au vocabulaire introduit dans la partie sur les processus ponctuels, on peut dire de manière équivalente que l'intensité du processus de comptage sous-jacent s'écrit :

$$\lambda(t) = {}^tZ(t)\beta(t)R(t)$$

Estimation des coefficients

On dispose d'un ensemble d'observations $\left(N_i^1(t), R_i(t), Z^i(t)\right)_{1\leq i\leq n}$ et on cherche à estimer le vecteur $\beta(t)$. Toutefois, il apparaît en pratique qu'il est beaucoup plus aisé d'obtenir un estimateur du vecteur de coefficients dit cumulés qui correspondent à $B(t) = \int_0^t \beta(u)du$.

En effet, notons $\lambda(t) = {}^t(\lambda_1(t), \ldots, \lambda_n(t))$, $N^1(t) = \left(N_1^1(t), \ldots, N_1^1(t)\right)$, et enfin $Z(t) = \left(R_1(t)Z^1(t), \ldots, R_n(t)Z^n(t)\right)$ matrice de taille $n \times p$. Avec ces notations, on a, en désignant par $\Lambda(t) = \int_0^t \lambda(u)du$ le processus vectoriel de taille n des intensités cumulées, $M(t) = N^1(t) - \Lambda(t)$ qui est une martingale. En observant alors que

$$dN^1(t) = Z(t)\beta(t)dt + dM(t) = Z(t)dB(t) + dM(t) \tag{1.3}$$

comme le terme en dM(t) est centré et que les incréments de la martingale sont non corrélés, on peut chercher à estimer les incréments $dB(t)$ par des techniques classiques de régression linéaire.

Pour cela on introduit l'inverse généralisé de $Z(t)$ défini par :

$$Z^-(t) = \begin{cases} \left({}^tZ(t)Z(t)\right)^{-1}{}^tZ(t) \text{ si } \left({}^tZ(t)Z(t)\right) \text{ est inversible} \\ 0 \text{ sinon} \end{cases}$$

$Z^-(t)$ est donc une matrice de taille $p \times n$ vérifiant $Z^-(t)Z(t) = J(t)I_p$ avec J(t) qui vaut 1 si l'inverse existe, et 0 sinon. En pratique, lorsque $Z(t)$ est de plein rang,

${}^tZ(t)Z(t)$ est inversible et on a alors simplement $Z^-(t)Z(t) = I_p$.

On pose alors comme estimateur naturel, compte tenu de 1.3, le processus :

$$\widehat{B}(t) = \int_0^t Z^-(u)dN^1(u).$$

En effet, le fait que

$$\widehat{B}(t) = \int_0^t J(s)dB(s) + \int_0^t Z^-(s)dM(s)$$

assure que $\widehat{B}$ estime B essentiellement sans biais, et on peut montrer sous certaines conditions peu restrictives que $\sqrt{n}\left(\widehat{B} - B\right)$ converge en loi en tant que processus vers un processus gaussien centré, U, dont on peut de plus calculer la fonction de covariance, notée Φ(cf théorème 5.1.1 de l'ouvrage de Martinussen et Scheike cité en préambule de cette partie).

Notons que malgré son expression un peu compliquée puisque sous forme intégrale, le calcul de l'estimateur $\widehat{B}(t) = \int_0^t Z^-(u)dN^1(u)$ se ramène à des calculs de sommes discrètes aux instants du processus $N^1(t)$. De manière plus précise, on a $\widehat{B}(t)$ qui est un vecteur de taille p et :

$$\widehat{B}_j(t) = \sum_{i=1}^{n} \int_0^t Z_{ij}^-(u)dN_i^1(u)$$

Mais $N_i^1(t)$ saute au plus une fois à l'instant X_i et l'incrément à cet instant est de 1 (s'il y a eu saut). On en déduit l'expression suivante :

$$\widehat{B}_j(t) = \sum_{X_i \leq t} Z_{ij}^-(X_i) \times d_i.$$

(où X_i et d_i désignent toujours respectivement la date de sortie observée et l'indicateur de non-censure)
Le calcul de l'estimateur des coefficients cumulés nécessite donc la détermination de $Z^-(X_i) = \left({}^tZ(X_i)Z(X_i)\right)^{-1} {}^tZ(X_i)$ pour toutes les sorties non censurées.

Comme dans le modèle de Cox, il est possible dans le cadre de cette modélisation à hasard additif de se poser la question de la significativité des coefficients estimés, et donc par la-même de l'influence ou non de telle ou telle covariable sur la survie, en construisant des tests statistiques. Sans trop entrer dans le détail de ces tests présentés par Martinussen et Scheike, précisons toutefois celui que nous utiliserons en pratique lors de notre modélisation.

L'objectif est de tester l'hypothèse nulle H_0 $\beta_i = 0$ où i désigne l'une des covariables du modèle. Bien qu'il n'y ait pas d'équivalence, il est plus commode de s'intéresser à l'hypothèse similaire sur les coefficients cumulés $B_i = 0$, puisque ce sont eux que l'on estime en pratique. Rappelons aussi que dans ce cadre, les coefficients, cumulés ou non, sont des paramètres fonctionnels dont on veut tester la nullité.

Une méthode de test est alors de s'intéresser à la statistique suivante (τ désigne la fin de la période d'observation pour l'estimation) :

$$T_i = \sqrt{n} \sup_{t \in [0,\tau]} |\widehat{B}_i(t)|$$

dont le comportement est lié à celui du sup, sur la même période, de la ième composante de la martingale U, limite de $\sqrt{n}\left(\widehat{B}-B\right)$ dont on connaît la covariance par l'intermédiaire de la fonction Φ. En pratique, Martinussen et Scheike proposent un calcul des quantiles par des méthodes de réechantillonnage permettant d'approcher le comportement de Φ, ce qui leur permet de déterminer à la fois un intervalle de confiance autour de leurs estimations, et de déterminer une p-value concernant l'acceptation de H_0.

1.4 Quelques applications des modèles de durée

Nous venons de présenter les deux grands modèles que nous allons chercher à mettre en application sur un portefeuille d'assurés dont on cherche à modéliser la survie. Attardons nous quelques instants sur les domaines où l'on peut trouver également ces modèles.

En économie De nombreux articles scientifiques et autres mémoires traitent de la détermination des facteurs influençant la durée du chômage, étude permise par la disponibilité et la fiabilité des enquêtes de panel sur le statut d'activité. Effectivement, les taux de chômage et la proportion d'inactivité à long terme ont posé problème à beaucoup de pays européens depuis les années 1990, comme le montre la figure. En particuliers, celle-ci exprime que le chômage est particulièrement élevé dans les pays européens comparativement au Japon ou aux Etats-Unis par exemple. D'où la volonté des pouvoirs publics d'agir de manière efficace sur ce fléau.

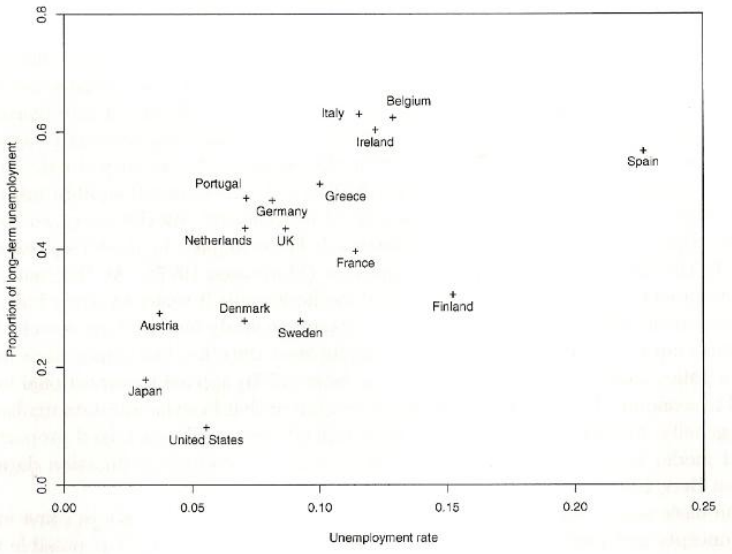


FIG. 1.4 – Chômage à long terme et taux de chômage dans des pays de l'OCDE en 1995

Olga Kupets a évoqué à ce sujet les éléments influençant le chômage en Ukraine (2005). Le modèle de hasard proportionnel de Cox et de la vraisemblance partielle sont utilisés pour trouver les facteurs pertinents expliquant la durée du chômage en Ukraine de 1997 à 2003 : sexe, statut marital, nombre d'enfants, éducation... Cette

démarche permet entre autre de déterminer l'effet positif, négatif ou nul des caractéristiques des individus à la base des estimations et d'en tirer les conclusions qui s'imposent.

En médecine De tels modèles peuvent être utilisés ici pour estimer la durée séparant l'apparition d'un symptôme et le décès, ou l'efficacité d'un traitement spécifique à une maladie... On peut imaginer ici que les facteurs d'influence peuvent être l'âge du patient, son sexe, ses antécédents familiaux. Huiling Cao (2005) a proposé une comparaison entre les modèles multiplicatif et additif dans quelques domaines biométriques. Ainsi, pour illustrer le modèle de Cox, elle s'intéresse au rejet de greffe de moëlle épinière en fonction de trois variables : risque bas de leucémie myélieuse, risque élevé de leucémie myélieuse, indicatrice d'apparition de rejet. Pour illustrer le modèle d'Aalen, elle s'intéresse à la dépendance de produits stupéfiants en fonction de l'âge de la première prise, d'un score de dépression, du nombre de traitements antécédents, de l'historique personnelle de toxicomanie, de la couleur de peau, de la durée du traitement, et du site. Cette démarche est riche d'enseignements puisque similaire à la nôtre dans un domaine certes totalement différent.

En actuariat Les modèles de durée se présentent comme un outil privilégié dans le monde de l'assurance puisqu'ils apparaissent dans de nombreux domaines tels que la durée de la vie humaine, de l'arrêt de travail, de la durée séparant des sinistres, avant la ruine... Pour l'assureur, il est essentiel de pointer les facteurs déterminant de telles durées pour pouvoir cerner correctement les risques auxquels il est exposé et ainsi gérer efficacement et prudemment son activité. Dans notre étude, nous nous basons sur une recherche de M. Choukroun (2008). Même si les objectifs diffèrent, la démarche et les méthodes usitées sont assez similaires. Dans un premier temps, l'influence des facteurs est donnée par le modèle de Cox qui permet de positionner les individus aux caractéristiques spécifiques par rapport à une référence. M. Choukroun montre cependant que ce modèle classique ne peut pas être toujours validé. Il propose alors un modèle alternatif plus original qui introduit un effet additif des covariables : le modèle d'Aalen, qui permet de prendre en compte la dépendance temporelle des variables et de visualiser l'effet de ces dernières à l'aide des courbes des coefficients cumulés, comme nous l'avons expliqué dans la partie théorique précédente. Il nous appartiendra de déterminer quel modèle est le plus adapté pour la construction d'une table de mortalité d'expérience à partir de notre portefeuille d'assurés particulier et spécifique.

Chapitre 2

Construction de la table d'expérience à partir d'un portefeuille AXA

La société Axa France et plus particulièrement la direction technique vie de celle-ci nous a fourni un ensemble de données correspondant à des contrats temporaires décès.

Pour toutes les observations dont nous disposons, il s'agit de contrats d'assurance qui garantissent le versement d'un capital au(x) bénéficiaire(s) si l'assuré décède ou est en invalidité absolue et définitive avant le terme du contrat. Ce sont des assurances à fonds perdus, au sens où rien n'est versé à terme s'il n'y a pas eu de sinistre. Trois types de produit différents sont représentés.

Précisons bien que cette base de données n'est qu'un échantillon du portefeuille temporaire décès du groupe Axa, et en cela il n'est aucunement représentatif. Les considérations chiffrées qui apparaîtront par la suite sont donc uniquement révélatrices du risque décès pour les personnes présentes dans cet échantillon. Ce dernier nous a surtout permis de mettre en oeuvre et d'illustrer une méthodologie, réutilisable pour qui voudrait construire une table de mortalité à l'aide de nos modèles. Dans un soucis de confidentialité, nous avons de plus masqué volontairement les graduations de l'axe des ordonnées de certains de nos graphiques

C'est donc à partir de ces contrats et des individus qui les ont souscrits que nous avons cherché à mettre en oeuvre les modèles de prise en compte de l'hétérogénéité introduits ci-avant, dans le but de construire diverses tables de mortalité prenant en compte certaines caractéristiques individuelles, qui serviront ensuite à tarifer des contrats temporaires décès de façon plus juste. Notre étude ne concernera que le risque décès.

2.1 Les données

2.1.1 Des fichiers initiaux à la base de données utilisable pour l'estimation

Pour chaque contrat, nous disposions alors des caractéristiques suivantes :
- le numéro du contrat
- le nom du produit d'assurance Axa associé.
- Les nom et prénom(s) de l'assuré.
- Le sexe de l'assuré.
- La date de naissance de l'assuré.
- La catégorie Socio-professionnelle de l'assuré.
- La situation actuelle du contrat.
- La date de souscription du contrat.
- La date de fin d'effet du contrat si celui-ci a pris fin pour une raison quelconque.
- La date de déclaration du décès en cas de décès.
- Capital sous risque.

L'objectif initial a pour nous été d'obtenir de façon "propre" pour chaque individu clairement identifié le quadruplet d'observations (X_i, d_i, E_i, Z_i) à partir de ces informations. Rappelons qu'il s'agit respectivement de l'âge en fin d'observation, de l'indicateur de censure, de l'âge d'entrée en observation, et des covariables dont on veut étudier l'impact sur la mortalité.

A ce sujet, conformément aux souhaits d'Axa, des remarques de Monsieur Planchet, et des données disponibles, nous avons retenus a priori 3 grandes catégories de covariables dont on présuppose l'influence sur la mortalité. Il s'agira du Sexe de l'individu, pour lequel on espère retrouver le phénomène bien connu par ailleurs et déjà intégré dans la segmentation réglementaire de surmortalité masculine. Nous avons retenu en outre deux autres variables : la catégorie Socio-Professionnelle de l'assuré révélatrice a priori d'un style de vie qui pourrait affecter d'une façon ou d'une autre la mortalité (sous-mortalité chez les cadres ou affiliés, aux conditions de vie présupposées favorables, sur-mortalité chez les non-cadres ou agriculteurs au cadre de vie plus difficile), et le montant du capital sous garantie, lui aussi en principe révélateur d'un niveau de vie (si l'on peut placer beaucoup de primes dans un contrat temporaire décès, ie avoir un capital garanti élevé, c'est que les choses vont bien par ailleurs).

Concrètement nous avons opéré chronologiquement de la façon suivante :

Elimination dans un premier temps des numéros de contrat doublons (un seul individu par contrat) afin d'éviter de considérer plusieurs fois la même ligne d'observation dans les statistiques.

Elimination ensuite de la base des contrats "sans effet" ou équivalents, correspondant à des individus qui n'ont pas en fait été exposés. Il s'agit pour la plupart de contrats où les garanties n'ont jamais été "actives". On trouve également des contrats pour lesquels le souscripteur a fait jouer sa clause réglementaire de renonciation (30

jours). En revanche commercialement, un dossier a été ouvert et ces individus apparaissaient dans la base. Au final on conserve donc dans la base tous les autres "statuts" de contrats pour lesquels l'assuré a donc en réalité été observé par Axa comme exposé au risque décès. Il s'agit donc des contrats toujours en cours à la date d'extraction des données, mais aussi des contrats clos pour de multiples raisons (sinistre décès, résiliation, annulation pour non paiement, sinistré en invalidité mais pas en décès, extinction ...) dans lesquels les garanties ont été actives sur une période qui sera notre période d'exposition au risque décès. Pour tous les contrats "sinistrés décès", nous avons affecté la date de déclaration de décès comme date de sortie du contrat, et nous avons créé un indicateur d'un tel statut (Donnée Non-censurée).

Elimination des contrats aux données individuelles manquantes ou mal renseignées (Capitaux sous risque, Catégorie Socio professionnelle, date de naissance). Nous avons ainsi en particulier enlevé des contrats dont la date de fin d'observation n'était pas présente (Non renseignée, erreur de saisie, date confidentielle pour le personnel Axa).

Prise en compte de la déshérence : affectation comme date de fin d'observation la date de 31/12/2006 à tous les contrats en portefeuille exposés à cette date là, et suppression des contrats souscrits au delà de cette date. En effet, même si la date d'extraction des données chez Axa datait de janvier 2008, rien n'assurait que les contrats en cours à la date d'extraction ne correspondait pas en fait à des personnes décédées dont les bénéficiaires du contrat n'avaient pas encore réclamé le capital garanti.

Gestion des multidétenteurs : il arrive qu'une personne soit assurée par plusieurs contrats temporaires décès différents. Afin de ne pas considérer un multidétenteur comme x individus différents, x correspondant au nombre de contrats à son nom, ce qui biaiserait les estimations de survie, nous avons conservé pour chacun d'eux parmi ses différents contrats la date minimale de souscription et la date maximale de sortie, ainsi que le capital maximal garanti (Rappelons l'objectif a priori d'étudier cette variable comme révélatrice d'un niveau de vie). En récupérant par ailleurs sa Catégorie Socio-Professionnelle (CSP), sa date de naissance et un indicateur de décès le cas échéant, nous avons remplacé dans la base ces contrats de multidétenteurs pas des contrats fictifs avec donc cette date de souscription la plus ancienne, la date de sortie la plus récente, le capital maximal garanti et les caractéristiques individuelles du multidétenteur. A noter que nous avons identifié dans la base ces personnes par la donnée combinée du sexe, de la date de naissance, et des noms et prénoms des individus, dont nous avons recherché les doublons.

Bilan 1 : A près tous ces retraitements, notre base contient des individus considérés comme différents, pour lesquels nous disposons du sexe, de la date de naissance, de la CSP, d'un niveau de capital garanti (variables qui nous permettront de définir un vecteur de caractéristiques Z_i par individu, pour reprendre les notations théoriques), d'une date de souscription que nous appellerons date d'entrée en exposition, d'une date de sortie d'exposition, et d'un indicateur précisant si la sortie est liée à un décès ou non (permettant de définir un indicateur de Non-censure d_i). Par un simple calcul à l'aide de la date de naissance, nous pouvons enfin remplacer les dates d'entrée et de sortie par des âges d'entrée et de sortie en exposition qui correspondront aux dates (X_i et E_i d'entrée et de sortie). Comme nous voulons construire

des tables du moment, seuls ces âges nous seront utiles à l'estimation. Nous savons en outre que tous ces individus ont comme caractéristique commune la souscription d'un contrat à garantie type temporaire décès chez Axa. On dénombre alors environ 45 000 individus avec un taux de sinistres décès d'environ $0,5\%$.

Remarque : Une dernière remarque est faite concernant le retraitement préalable de la base de données dans le but d'effectuer nos estimations, que nous souhaitons les plus robustes possibles. Il s'agit de la combinaison des âges de couverture possible dans ces contrats avec une observation rapide de la sinistralité par âge.

On remarque en effet que le premier sinistre observé dans notre base, résultant de la première partie des retraitements, l'est à 25 ans. On constate en outre qu'à partir de 71 ans, le nombre de personnes exposées au risque par tranche d'âge d'un an est de moins en moins important (< 200), pour un âge maximal d'observation de 80 ans, avec un nombre de sinistres par tranche de 0 ou de 1 . Afin d'obtenir des estimations pertinentes concernant la mortalité de notre portefeuille, nous avons donc choisi de restreindre notre fenêtre d'observation à la période 25-71 ans pour laquelle nous disposons de données a priori suffisantes aussi bien en effectif réel qu'en nombre de sinistres. Concrètement, cela nous a conduit à recoder la fenêtre d'exposition de chaque individu de la façon suivante (l'indice Bilan1 désignant la valeur d'une observation à l'issue de la phase de retraitement dont nous venons de présenter le bilan ci-avant) :

$$[E_i, X_i] = [max\left(25, E_i^{Bilan1}\right), min\left(71, X_i^{Bilan1}\right)]$$

et à modifier l'indicateur de sinistre selon la formule

$$d_i = d_i^{Bilan1} \times \mathbf{1}\left\{X_i^{Bilan1} < 71\right\}$$

Ceci nous a donc amenés à enlever de la base de travail les individus dont la fenêtre d'exposition ainsi recodée est nulle (environ 1500 personnes).

Ce choix de fenêtre d'exposition s'accorde avec la nature des contrats. En effet, ceux-ci comme la plupart des contrats d'assurance sont signés pour une durée d'un an avec tacite reconduction, avec un âge maximal de souscription initiale de 70 ans, la garantie pouvant s'étendre jusqu'à 80 ans.

Ainsi lorsqu'il s'agira de comptabiliser l'exposition au risque par âge entier, notée N_x pour l'année x, nous disposerons de valeurs pour x allant de 25 à 70 ans. Nous verrons alors que nous pourrons à la suite de diverses estimations leur associer des taux de sortie pas tranche d'âges estimés que nous noterons $\hat{q}_x$

Voici un résumé chiffré des ordres de grandeurs du retraitement de la base initiale que nous venons de présenter, pour passer d'environ 78000 observations initiales en entrée à environ 44000 assurés.

Etape de retraitement	Nombre d'observations après l'étape	Nombre de sinistres après l'étape
Initiale	78000	600
Contrats doublons	77000	550
Contrats "sans effet" ou équivalents	74000	550
Variables manquantes	51000	300
Déshérence	50000	250
Multidétenteurs	45000	250
25-70 ans	44000	240

Nous proposons par la suite une description statistique de la population d'assurés, afin de voir s'il n'est pas possible de mettre en évidence certaines caractéristiques pouvant influencer les estimations qui suivront.

2.1.2 Description de la base de données utilisable

Quels individus ?

Nous allons tenter ici d'effectuer quelques statistiques descriptives dans le but de mieux cerner la composition de la population à laquelle nous sommes finalement confrontés après ces multiples retraitements. Cette analyse, qui peut paraître a priori marginale, nous permet d'introduire en détail les variables explicatives que nous allons introduire dans le modèle additif et le modèle à hasard proportionnel, afin de tester leur éventuelle influence sur la survie, ce qui pourrait alors amener l'assureur à effectuer une tarification différenciée à partir d'elles.

Concernant le sexe, 59% des assurés sont des hommes et 41% des femmes. Cette répartition marquée est tout de même convenable puisque les effectifs de chaque sexe sont suffisamment importants et de même ordre pour les estimations qui suivent.

Concernant la CSP, 3.1% des individus forment la catégorie CSP1 (agriculteurs), 11.4% la seconde catégorie (CSP2, artisants commerçants), 6.4% la CSP3 (professions libérales et cadres), 71.2% la CSP4 (salariés), 2.7% sont des retraités (CSP5), 5.1% sont sans profession (CSP6). Deux remarques concernant cette classification. Tout d'abord les salariés sont des salariés non cadres, les ouvriers par exemple faisant partie de cette classe. D'ailleurs, cette catégorie CSP4 présente un fort poids qui pourrait poser problème lors des estimations. Nous avons cherché ensuite à effectuer des regroupements logiques et pertinents en vue de ne pas multiplier inutilement les modalités pour une variable. Ainsi nous avons choisi de rassembler cadres et professions libérales pour former la CSP3, bien qu'il existait une différence dans la base initiale. En effet nous avons considéré ces deux CSP assez proches en terme de niveau de vie a priori, et dans l'optique de détermination de variables influençant de façon significative la survie, on aimerait voir celle-ci se distinguer comme facteur améliorant cette dernière.

Concernant le montant de capitaux garantis, sans rentrer dans le détail de la répartition (notons simplement un montant minimal d'environ 700 euros et un montant maximal de plus d' 1,5 millions d'euros), il est important de noter que dans

certains types de contrats, ce n'est pas l'assuré qui choisit de façon "continue" le montant de capital garanti en cas de décès. Bien souvent, ce dernier se contente de choisir un contrat avec un montant proposé par l'assureur. Ainsi, dans notre base, on notera la présence de 3 montants de capitaux très représentés (15245, 30000 et 50000 euros) qui doivent correspondre à de tels contrats.

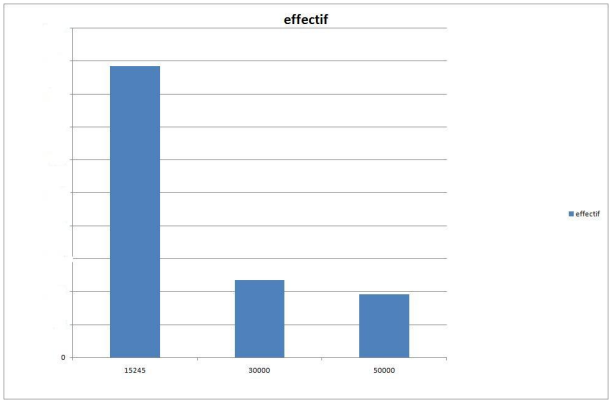


FIG. 2.1 – Montants de capitaux les plus représentés

Compte tenu de cette particularité dans la répartition de cette variable capital pourtant "quantitative" a priori, nous avons décidé de tester son influence sur la survie en la transformant en variable qualitative. Nous avons ainsi créé 4 classes de capitaux délimitées à l'aide des 3 montants singuliers évoqués

- capital1, pour des montants inférieurs ou égaux à 15245 euros
- capital2, pour des montants compris entre 15246 et 30000 euros inclus
- capital3, pour des montants compris entre 30001 et 50000 euros inclus
- capital4, pour des montants strictement supérieurs à 50000 euros

Les effectifs de ces classes sont alors équilibrés.

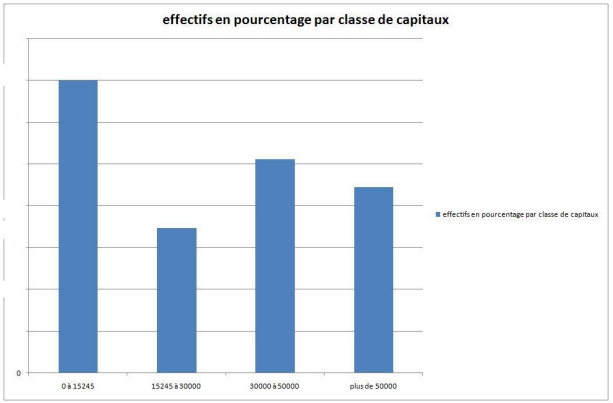


FIG. 2.2 – Effectif par classe de capitaux

Enfin, quelle est la répartition par âge de ce portefeuille ? Le graphique suivant permet d'y répondre, en prenant garde au fait qu'il s'agit de celui des expositions au risque par tranche d'âge, c'est-à-dire du nombre de personnes de ces âges exposés au risque décès dans notre portefeuille. Nous notons N_x l'effectif de notre portefeuille exposé au risque entre les âges x et $x+1$, et nous avons alors posé :

$$N_x = \sum_i \max\left(0; \min(x+1; X_i) - \max(x, E_i)\right)$$

afin de comptabiliser les personnes sur chaque âge en fonction de leur fenêtre d'observation, pas forcément "entière". Ainsi une personne présente en portefeuille de $34,2$ ans à $36,5$ ans contribue pour $0,8$ à N_{34}, pour 1 à N_{35} et pour $0,5$ à N_{36}. On obtient alors la répartition suivante

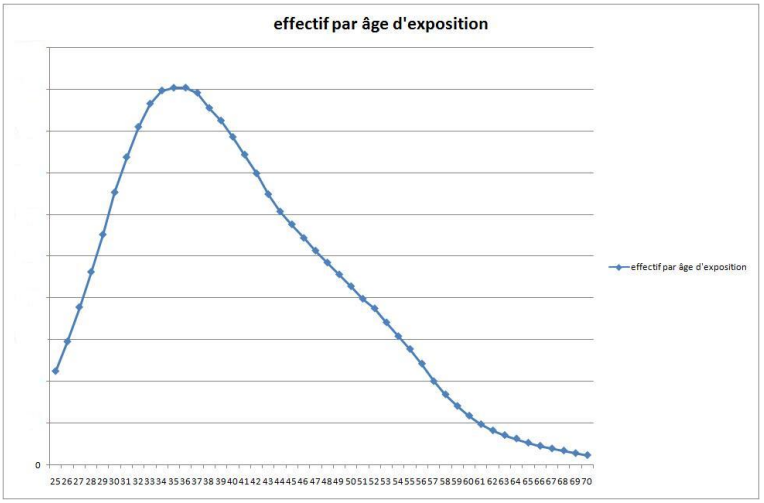


FIG. 2.3 – Effectif N_x par âge d'exposition

36 ans correspond à l'âge d'exposition le plus représenté. Les âges extrêmes sont quant à eux logiquement peu représentés.

Précisons d'autre part que la plus ancienne date de souscription correspond au 6 Mars 1995, la plus récente au 27 Décembre 2006. La fenêtre d'observation de nos contrats est donc relativement importante, et nous sommes donc à la limite de pouvoir parler de tables du moment pour nos tables d'expérience. Toutefois, nous ne disposons pas d'assez de données pour estimer l'influence de la génération et envisager une construction prospective. Ainsi, compte tenu du phénomène bien connu de l'augmentation de l'espérance de vie, nous allons classiquement surestimer la mortalité avec cette construction à partir de données espacées. De plus, d'un point de vue prudentiel, l'utilisation de ces tables du moment pour tarifer des contrats à venir sera également une bonne chose, pour la même raison d'augmentation de l'espérance de vie. Nous essaierons néanmoins d'introduire cet aspect générationnel propre aux tables prospectives de façon ad-hoc en fin de mémoire lors de l'exemple de tarification.

Evoquons maintenant la sinistralité empirique de notre portefeuille.

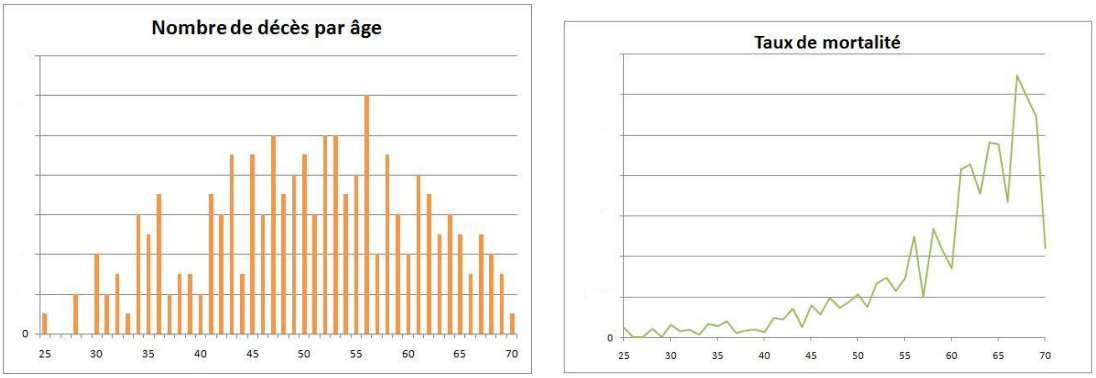


FIG. 2.4 – Décès et Taux de mortalité sur l'intégralité du portefeuille

Sinistralité du portefeuille

En regard de l'exposition au risque par âge, voici (figure 2.4) la répartition du nombre de décès D_x par classe d'âge x, ce qui nous permet d'en déduire des premiers taux de mortalité bruts empiriques $\hat{q}_x = \frac{D_x}{N_x}$ pour l'ensemble de notre portefeuille.

On constate donc un nombre relativement faible de décès pour les classes d'âges les plus exposés, avec au contraire de nombreux décès pour les âges moins exposés. Il s'ensuit une courbe de taux empiriques relativement peu régulière malgré la tendance classique de croissance exponentielle qui semble se dégager. Toujours est-il que l'insuffisance ou l'abondance de décès à certains âges exposés, liée à notre petite base de données, semble rendre difficile un travail d'estimation pertinent.

Comme nous allons tenter de déterminer l'influence des covariables sur ces $\hat{q}_x$, il nous a paru intéressant d'évoquer à titre illustratif le taux de sinistres pour chaque modalité de nos classes de variables.

0.7% des hommes et 0.3% des femmes sont sinistrés, ce qui indiquerait à première vue que le facteur $sexe$ influence considérablement la mortalité, puisque par ailleurs, les expositions au risque sont similaires chez les hommes et les femmes.

Concernant la CSP, le graphique suivant permet de donner l'influence empirique de ce facteur, toutes choses égales par ailleurs. Les retraités souffrent naturellement d'une plus forte mortalité. Concernant les autres classes, on ne remarque pas de différences significatives.

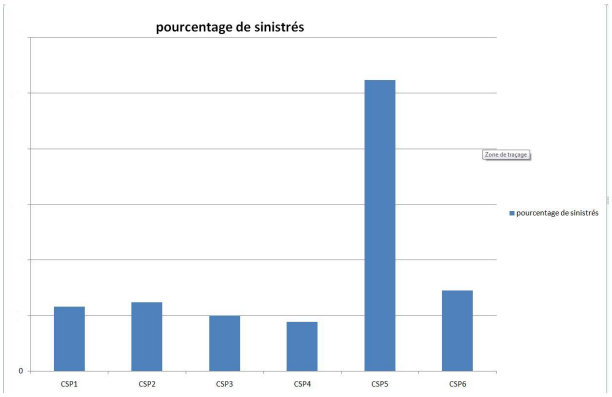


FIG. 2.5 – pourcentages de sinistralité par CSP

Concernant les capitaux assurés, il semble difficile de mettre en évidence quelques intuitions. Remarquons simplement à ce stade que la première classe (montants inférieurs à 15245 euros) semble se détacher par sa sinistralité relativement élevée.

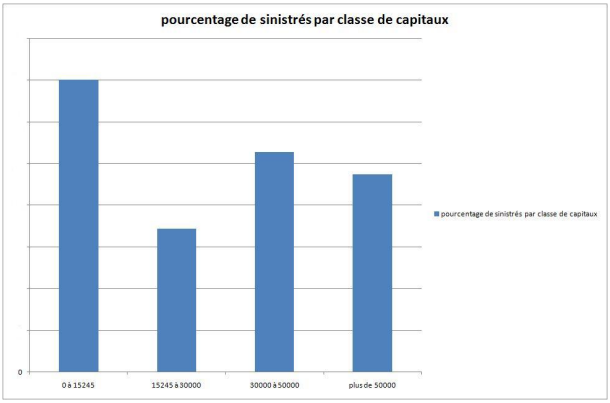


FIG. 2.6 – Pourcentages de sinistralité par classe de capitaux

Mais gardons à l'esprit que croiser la sinistralité avec ces modalités ne suffit de toute façon pas à conclure de façon définitive sur l'influence de ces dernière sur la mortalité des sous-populations concernées. Ces graphiques sont surtout étudiés pour décrire notre portefeuille en amont des premières estimations. En effet, comme on peut l'illustrer avec la catégorie "retraités" qui présente une forte sinistralité, il est important de croiser la mortalité observée dans la classe avec l'âge d'exposition au risque de cette classe. Justement, pour ce qui est des retraités, on est en présence d'une catégorie de population d'âge élevé pour laquelle il est normal d'observer une sinistralité supérieure à la moyenne, par le simple effet âge. Nos modèles, par construction, permettront de tester l'influence sur la mortalité de ces covariables au delà de l'effet naturel de l'âge.

Choix des covariables

Intéressons-nous maintenant à la corrélation éventuelle entre ces variables. Nous devons en effet supprimer les variables redondantes pour ne pas nuire à nos estimations à venir. Pour faire un premier tri parmi ces dernières, effectuons des tests d'indépendance du Khi-2 sur base de tableau de contingence. Nous utilisons la fonction chisq.test sous R.

	Chi2	Degrés de liberté	p-value
Capital-CSP	1144	15	<2.2E-16
Sexe-CSP	2556	5	<2.2E-16
Capital-Sexe	257	3	<2.2E-16

Nous pouvons donc rejeter l'hypothèse de dépendance entre les variables caractéristiques, comme le montrent les valeurs des p-values données par R. Nous retenons donc pour les estimations l'ensemble des covariables indiquées au début de ce paragraphe.

Au terme de cette analyse, nous retenons donc les variables suivantes avec leur nombre de modalités : 4 classes de capitaux, 6 CSP, 2 sexes. D'un point de vue pratique une indicatrice de chaque modalité de chaque variable est associée à tous les individus. Ce sont ces indicatrices qui constitueront les covariables $Z_i = (Z_i^1, \ldots, Z_i^p)$ des modèles de durée introduits dans le premier chapitre.

Pour pallier à tout problème d'identification dans nos estimations, on choisit une référence a priori(1 modalité par variable : Homme - Salarié - capital1) dont les indicatrices ne seront pas introduites dans les modèles. Le choix de cette catégorie n'est pas anodin : les statistiques descriptives ont montré que cette catégorie d'individus est la plus présente dans notre portefeuille d'assurés, et nous verrons que le choix d'une référence avec un nombre maximal d'observations est important. C'est donc par rapport à cette classe de population que les effets des autres covariables pourront être interprétés. Concrètement pour le modèle de Cox, cela signifiera que le hasard de base $h_0(t)$ correspondra à cette référence. Dans le cadre du modèle de Aalen, $h(t) = \sum Z_k(t)\beta_k(t)$, en plus de ces indicatrices $Z_k(t)$, on introduira une covariable Z_0 constante égale à 1, et le coefficient $\beta_0(t)$ associé servira à représenter la contribution au hasard de cette classe de référence.

2.2 Estimation de l'influence des covariables sur la survie

2.2.1 Modélisation Cox

Choix du logiciel

Nous travaillons sous R à l'aide du package $survival$ et des fonctions $coxph$ et $survfit$

Résultats

On cherche ici à estimer les paramètres β pour un hasard modélisé par

$$h(t) = h_0(t)exp(Z\beta)$$

avec

$Z\beta = \mathbf{1}_{Femme}\beta_{Femme} + \mathbf{1}_{CSP1}\beta_{CSP1} + \mathbf{1}_{CSP2}\beta_{CSP2} + \mathbf{1}_{CSP3}\beta_{CSP3} + \mathbf{1}_{CSP5}\beta_{CSP5} + \mathbf{1}_{CSP6}\beta_{CSP6} + \mathbf{1}_{capital2}\beta_{capital2} + \mathbf{1}_{capital3}\beta_{capital3} + \mathbf{1}_{capital4}\beta_{capital4}$

Voici les résultats de la régression effectuée sous R :

```
            coef     exp(coef)  se(coef)  z       p
Sexebin   -0.6192     0.538     0.160 -3.868 0.00011
CSP1      -0.4913     0.612     0.365 -1.345 0.18000
CSP2      -0.3785     0.685     0.203 -1.868 0.06200
CSP3      -0.6560     0.519     0.286 -2.296 0.02200
CSP5      -0.0525     0.949     0.265 -0.198 0.84000
CSP6       0.4053     1.500     0.275  1.475 0.14000
capital_2 -0.2062     0.814     0.219 -0.940 0.35000
capital_3 -0.0761     0.927     0.164 -0.464 0.64000
capital_4  0.1142     1.121     0.182  0.627 0.53000

          exp(coef) exp(-coef) lower .95 upper .95
Sexebin       0.538      1.857     0.393     0.737
CSP1          0.612      1.634     0.299     1.252
CSP2          0.685      1.460     0.460     1.019
CSP3          0.519      1.927     0.296     0.908
CSP5          0.949      1.054     0.564     1.595
CSP6          1.500      0.667     0.875     2.570
capital_2     0.814      1.229     0.529     1.251
capital_3     0.927      1.079     0.672     1.278
capital_4     1.121      0.892     0.785     1.602

Rsquare= 0.001   (max possible= 0.083 )
Likelihood ratio test= 26.2  on 9 df,   p=0.00187
Wald test            = 24.6  on 9 df,   p=0.00349
Score (logrank) test = 24.9  on 9 df,   p=0.00306
```

En analysant les valeurs des p-values, il semble que Sexe et CSP3 constituent les seules variables influençant le hasard proportionnel de Cox. En effet, seules ces dernières se placent sous le seuil de 5% pour le test de significativité des coefficients. Après "backward selection", méthode qui consiste à retirer du modèle la variable la moins significative, nous avons retiré progressivement du modèle les variables non pertinentes. Aucune variable autre que Sexe et CSP3 ne s'est révélée pertinente au regard des p-values pour le test de nullité des coefficients. En conservant uniquement ces 2 variables les résultats sont les suivants :

```
          coef   exp(coef) se(coef)     z       p
Sexebin -0.529     0.589    0.154    -3.44    0.00059
CSP3    -0.542     0.581    0.277    -1.96    0.05000

        exp(coef)   exp(-coef)  lower .95  upper .95
Sexebin     0.589       1.70      0.436     0.797
CSP3        0.581       1.72      0.338     1.000

Rsquare= 0   (max possible= 0.083 )
Likelihood ratio test= 16.5  on 2 df,   p=0.000266
Wald test            = 15.0  on 2 df,   p=0.000543
Score (logrank) test = 15.4  on 2 df,   p=0.000462
```

Les variables Sexe et CSP3 restent donc pertinentes au vue des p-values pour un seuil à 5%, même si CSP3 se situe à la limite du rejet de l'hypothèse d'indépendance avec le hasard de Cox. La nullité jointe des paramètres est elle rejeté par les 3 tests classiques présentés dans la théorie.

On ne conserve donc au final que la modélisation du hasard suivante :

$$h(t) = h_0(t) exp\left(\mathbf{1}_{Femme}\beta_{Femme} + \mathbf{1}_{CSP3}\beta_{CSP3}\right)$$

qui permet de différencier 4 sous-groupes de population en terme de mortalité. Les "Hommes Cadres", les "Femmes Cadres", les "Femmes Non-Cadres" et les "Hommes Non-Cadres", cette dernière sous-population faisant alors office de référence puisque correspondant à un vecteur de covariables Z nul.

Survies

Voici une représentation graphique de l'estimation de la fonction de survie de base du modèle $\hat{S}_0$, estimée à l'aide du programme $survfit$ sous R :

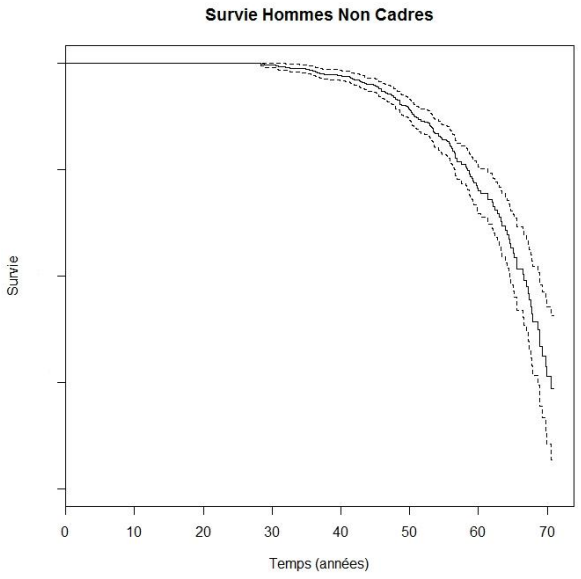


FIG. 2.7 – Cox - Survie de base/S_0

La survie de base de Cox se présente donc comme une fonction décroissante de l'âge, et conformément à la méthode de type Kaplan-Meier utilisée par survfit pour estimer la survie, on obtient une fonction en escalier avec des sauts à chaque date de sortie T_i non censurée présente en portefeuille, puisque le logiciel calcule chaque $\hat{S}_0(T_i)$. Autour de la courbe sont représentés les intervalles de confiance à 95% issus de l'estimation.

La clé maintenant réside en l'influence des deux variables (Sexe et CSP3) sur la survie des individus.

Rappelons tout d'abord le lien entre Survie de la base et survie de chaque classe.

$$S(t) = e^{\int_0^t h(u)du} \text{ et } h(t) = h_0(t)exp(Z\beta)$$

on en déduit donc, puisque nos covariables sont indépendantes du temps, que :

$$S(t) = (S_0(t))^{exp(Z\beta)}$$

Ainsi on obtient la survie d'une des 4 sous-populations à l'aide de la formule de passage

$$\hat{S}(T_i) = \hat{S}_0(T_i)^{exp(Z\beta)}$$

en faisant varier Z selon la catégorie de population voulue.

Le graphe suivant représentant les survies de chaque groupe permet d'élucider l'impact de chacune de ces variables Sexe ou CSP3 sur la mortalité :

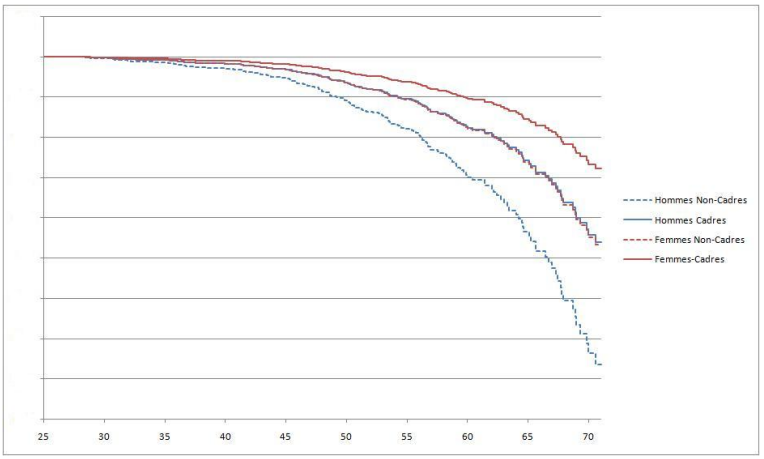


FIG. 2.8 – Cox - Survie des différentes classes

Pour cela, il nous faut travailler toutes choses égales par ailleurs. Tout d'abord, les femmes cadres présentent une survie plus élevée que les hommes cadres, la même remarque étant faite pour les individus non cadres. Cela montre la surmortalité masculine caractéristique, mise en évidence lors des statistiques descriptives du portefeuille d'assurés. D'autre part, les hommes cadres disposent d'une survie plus élevée par rapport aux hommes non cadres, de même pour les femmes. Cela s'explique peut-être par les conditions de travail et le niveau de vie plus confortables pour cette catégorie. Ce graphique montre surtout que les caractéristiques que nous avons retenues permettent une différenciation et une segmentation de la survie, et a fortiori nous le verrons, des quotients de mortalité.

Qu'en est-il pour le modèle additif d'Aalen ?

2.2.2 Modélisation Aalen

Choix du logiciel

Nous travaillons sous R, à l'aide du package $timereg$ et des fonctions $aalen$ et $plot.aalen$. Précisons que ce choix n'en est pas un, car il n'existe pas à notre connaissance d'autres logiciels permettant de mettre en application le modèle additif. Ce package a d'ailleurs été mis en place par T.H. Sheike, et on trouvera un tutorial de son utilisation dans l'ouvrage dont il est le co-auteur et déjà cité dans ce mémoire. Notons cependant que puisque l'estimation des coefficients dans le modèle de Aalen nécessite l'inversion d'une matrice de taille liée à la taille de l'individu, et que de nombreux tests dans ce cadre théorique nécessitent l'utilisation de méthodes de rééchantillonnage, l'espace mémoire dédié au Calcul de R est vite saturé, et il nous a été impossible d'exploiter avec nos données l'essentiel des possibilités de ce package.

Résultats

Dans cette modélisation, rappelons que notre hasard s'écrit sous forme additive

$$h(t) = \beta_0(t) + \sum Z_k \beta_k(t)$$

avec Z_k représentant toutes les indicatrices des modalités de nos 3 variables "Sexe" "CSP" et "capitaux", exceptées les indicatrices des modalités de la sous population de référence dont la contribution au hasard est captée par le paramètre $\beta_0(t)$

Nous avons vu en outre qu'il était plus commode dans cette modélisation de travailler avec les coefficients cumulés

$$B_k(t) = \int_0^t \beta_k(u)du$$

dont on connait un estimateur. C'est la significativité de ces coefficients qui est testée par la fonction $aalen$ du package, et ils nous permettront de déduire un éventuel impact des covariables sur la mortalité par rapport à notre catégorie de référence.

De la même façon que pour le modèle de Cox, incorporons d'abord toutes les variables dans la régression . Voici la sortie R de la régression obtenue :

```
        Additive Aalen Model

Test for nonparametric terms

Test for non-significant effects
                   sup| hat B(t)/SD(t) |        p-value H_0: B(t)=0
(Intercept)                       7.83                      0.000
Sexebin                           4.24                      0.001
CSP1                              4.22                      0.001
CSP2                              2.31                      0.271
CSP3                              3.69                      0.003
CSP6                              2.71                      0.097
CSP5                              8.17                      0.000
capital_2                         2.53                      0.177
capital_3                         1.59                      0.733
```

```
capital_4                    2.12                    0.474
```

En observant les p-values, il existe a priori plus de variables pertinentes que dans le modèle de Cox, à savoir le Sexe, CSP1, CSP3, CSP5. En supprimant également progressivement les modalités sans incidence par "backward selection", nous obtenons les résultats suivants :

```
        Additive Aalen Model

Test for nonparametric terms

Test for non-significant effects
                 sup| hat B(t)/SD(t) |       p-value H_0: B(t)=0
(Intercept)                   10.10                  0.000
Sexebin                        3.95                  0.002
CSP1                           5.58                  0.000
CSP3                           3.90                  0.004
CSP5                           9.46                  0.000
```

L'ensemble des modalités retenues restent donc significatives d'après les p-values.

Or, d'un point de vue théorique, le lien coefficients cumulés/Survie est le suivant : lorsque les covariables sont indépendantes du temps on peut exprimer facilement le hasard intégré $H(t) = \int_0^t h(u)du$ en fonction des coefficients cumulés $B_k(t)$ compte tenu de la forme simple du hasard. On a en effet $H(t) = B_0(t) + \sum_k Z_k \times B_k(t)$. Ainsi le lien entre la survie et ces coefficients cumulés est :

$$S(t) = exp\left(-B_0(t) - \sum_k Z_k B_k(t)\right)$$

Fort de cette écriture, intéressons-nous maintenant aux estimations, à chaque date de sortie non censurée, des coefficients cumulés afin de dégager l'action sur la mortalité de ces modalités considérées comme significatives. Rappelons qu'à ce stade des estimations, il ne reste plus dans le modèle que les indicatrices des CSP 1, 3 et 5, ainsi que l'indicatrice de Sexe de l'individu (1 si Femme, 0 si Homme). La catégorie de référence correspondant donc, par opposition, aux Hommes de CSP 2, 4 ou 6. Les allures des coefficients cumulés obtenues sont les suivantes :

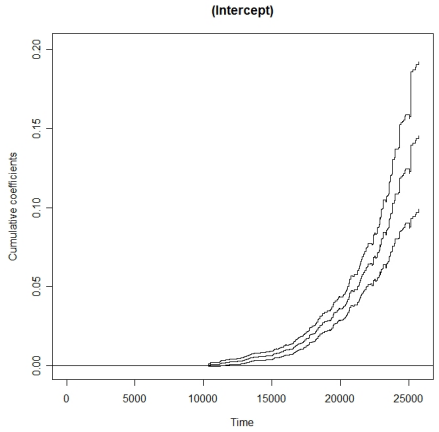


FIG. 2.9 – Coefficient cumulé pour la population de référence

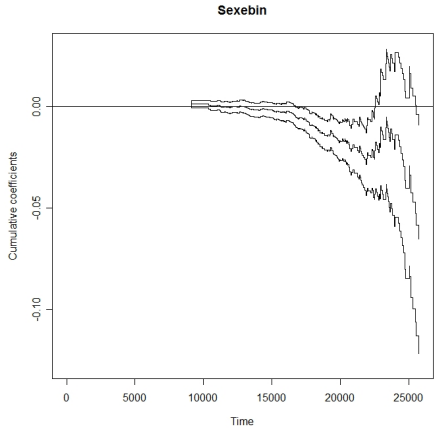


FIG. 2.10 – Coefficient cumulé pour le sexe

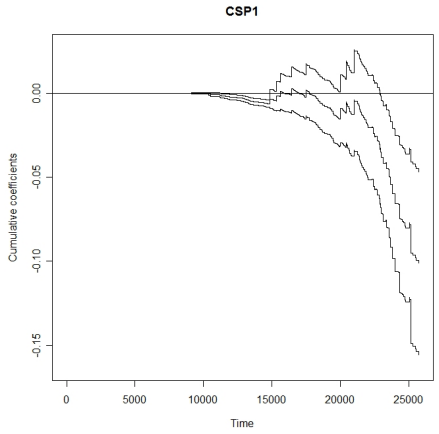


FIG. 2.11 – Coefficient cumulé pour la CSP1

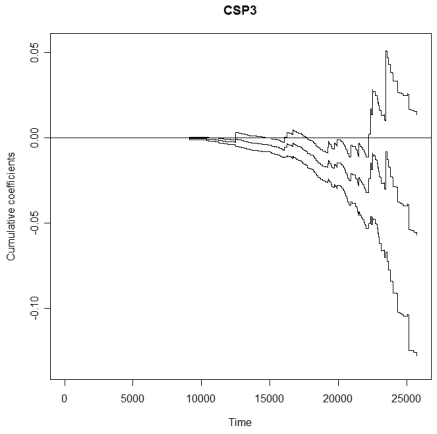


FIG. 2.12 – Coefficient cumulé pour la CSP3

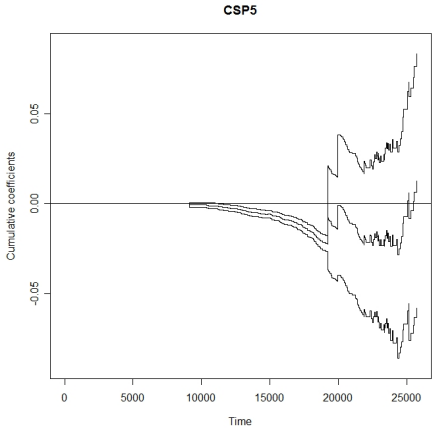


FIG. 2.13 – Coefficient cumulé pour la CSP5

Deux remarques générales sont à faire sur ces graphiques.

On observe tout d'abord une certaine régularité pour le coefficient $B_0(t)$ (intercept) dont l'aspect décroissant positif au cours du temps permet d'affirmer qu'il engendre une survie $S = exp(-B_0(t)$ décroissante plus petite que 1 assez régulière, ce qui semble normal. Ce coefficient étant présent si l'on s'intéresse à la survie de n'importe quel sous-groupe de personnes ($S = exp\left(-B_0(t) - \sum_k Z_k B_k(t)\right)$) , on peut s'attendre donc à ce qu'il soit responsable de l'allure générale de la survie de chacun de ces groupes, les autres coefficients cumulés ne permettant qu'un ajustement par rapport à cette référence. Toutefois, et c'est là l'objet de la seconde remarque, l'allure des courbes aux âges élevés pour les autres coefficients cumulés, et notamment les intervalles de confiance associés (traits pointillés) laisse transparaître un certain manque de précision des estimations. Il semble donc difficile de mener une quelconque interprétation fiable quant à l'influence de ces coefficients à partir des ces seuls graphiques. Retenons juste une tendance à la négativité des coefficients cumulés autres que $B_0(t)$, avec une tendance à la décroissance, ce qui permettrait de conclure que les covariables dont ils représentent l'effet ont donc tendance à améliorer la survie par rapport à la catégorie de référence.

Afin de clarifier cette tentative d'interprétation, nous avons tracé les graphes des survies estimées correspondant à différents sous-groupes de population afin de déterminer d'éventuelles influences remarquables de ces covariables sur la survie. Nous avons utilisé pour cela l'estimation naturelle suivante, évaluée à chaque date de sortie non censurée pour lesquelles nous avions accès à la valeur des coefficients cumulés.

$$\hat{S}(t) = exp\left(-\hat{B}_0(t) - \sum_k Z_k \hat{B}_k(t)\right)$$

Nous nous interrogeons tout d'abord sur l'influence de la variable Sexe :

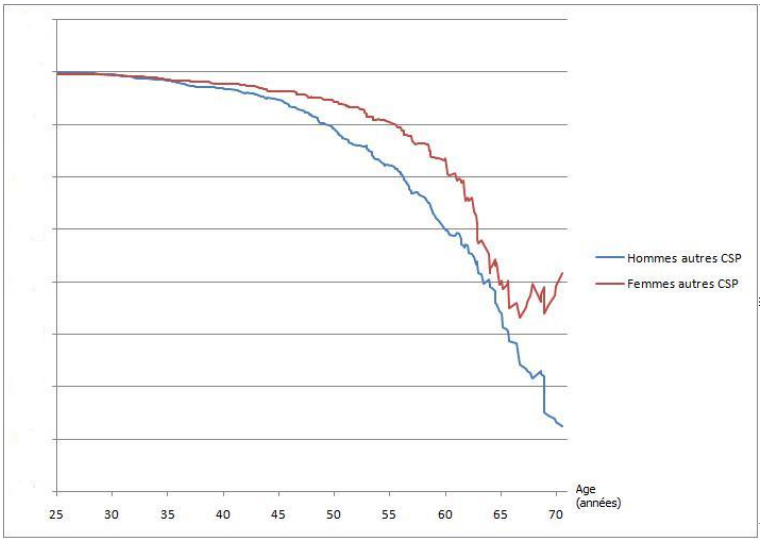


FIG. 2.14 – Survies estimées - Sexe autres CSP

On s'interroge sur l'influence de la variable CSP :

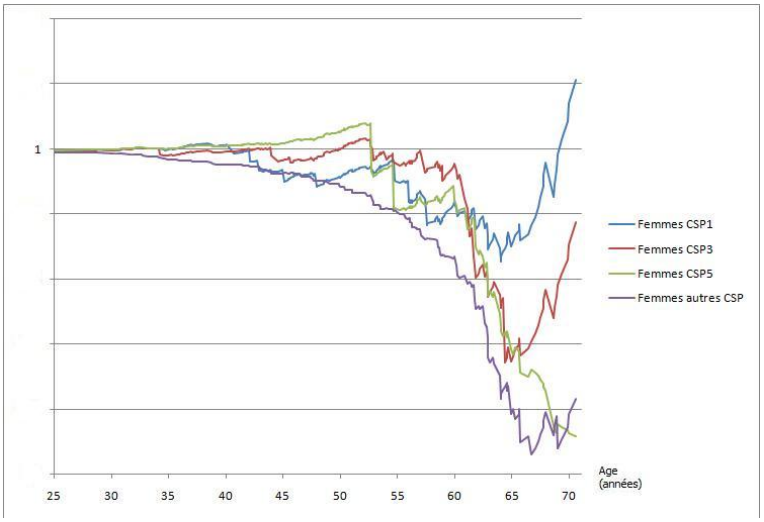


FIG. 2.15 – Survies estimées - Femmes toute CSP

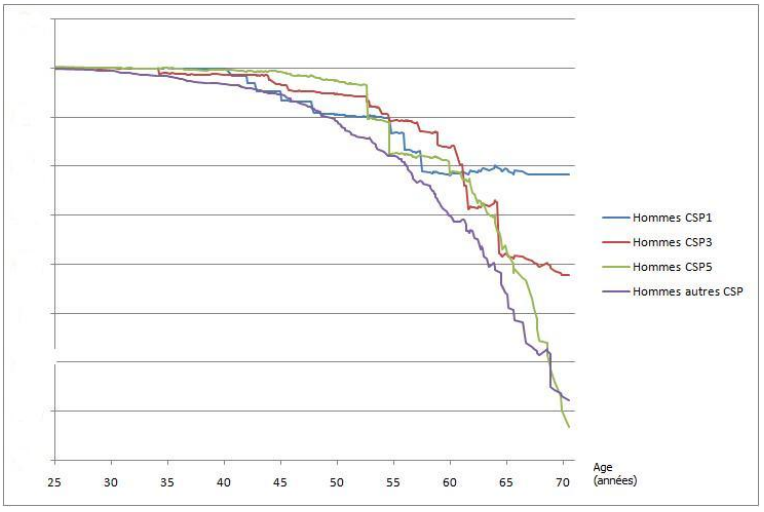


FIG. 2.16 – Survies estimées - Hommes toute CSP

De prime abord un problème pratique se pose : la fonction de survie dépasse parfois la limite théorique 1. Cela est rendu possible car les estimateurs des coefficients cumulés obtenus via la modélisation de Aalen sont non contraints. Ainsi lorsque que l'on reconstitue la survie à partir de ces coefficients cumulés selon la formule précisée plus haut, il arrive que la survie dépasse la valeur 1. En pratique ce problème peut être résolue en considérant pour chaque date T_i par la formule

$$\hat{S}(T_i) = \min\left(exp\left(-\hat{B}_0(T_i) - \sum_k Z_k \hat{B}_k(T_i)\right); 1\right)$$

D'autre part, certaines courbes de survie ont tendance à croître pour les âges élevés, ce problème résultant sûrement des soucis d'estimation mis en évidence aux grands âges pour les coefficients cumulés qui peuvent ne pas être robustes à cause de trop faibles effectifs. Nous reviendrons sur ce point très vite.

De manière générale, il est difficile de dégager quelques caractéristiques de l'influence de la CSP sur la survie. Remarquons simplement que la catégorie $autre\ csp$ (les CSP non-significatives mises en référence) se détache par une survie inférieure.

Afin de dégager des estimations "correctes", nous avons décidé de nous baser uniquement sur les résultats de significativité obtenus par la modélisation de Cox, puisque les estimateurs obtenus par le modèle de Aalen sont non contraints et ne semblent pas satisfaisants. Le modèle d'Aalen nous sert juste ici à estimer l'effet des variables CSP3 et Femme par rapport à la référence, de façon alternative à la modélisation de Cox, et non plus à mettre en évidence un éventuel impact sur la mortalité .

Voici le résultat de la nouvelle régression effectuée sur R :

```
Additive Aalen Model

Test for nonparametric terms

Test for non-significant effects
                  sup| hat B(t)/SD(t) |      p-value H_0: B(t)=0
(Intercept)                     10.40                      0.000
Sexebin                          3.97                      0.003
CSP3                             3.81                      0.002
```

Les p-values sont significatives. Nous remarquons alors que les coefficients cumulés de CSP3 et de l'indicatrice de Sexe féminin ont logiquement la même allure que dans la précédente régression, les effectifs de ces classes n'ayant pas été modifiés, au contraire de la population de référence qui a elle changé. Le graphe pour l'$intercept$ est donc lui légèrement modifié, mais son allure laisse transparaître une survie associée à l'allure classique de décroissance.

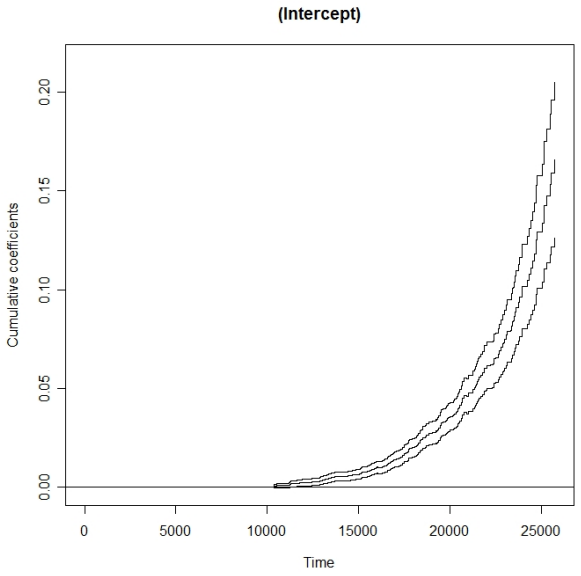


FIG. 2.17 – Coefficient cumulé pour la population de référence

En ce qui concerne les survies, on les estime toujours dans ce cadre à l'aide de la formule suivante :

$$\hat{S}(T_i) = \min\left(exp\left(-\hat{B}_0(T_i) - \mathbf{1}_{femmes}\hat{B}_{femmes}(T_i) - \mathbf{1}_{CSP3}\hat{B}_{CSP3}(T_i)\right); 1\right)$$

On obtient alors comme allure :

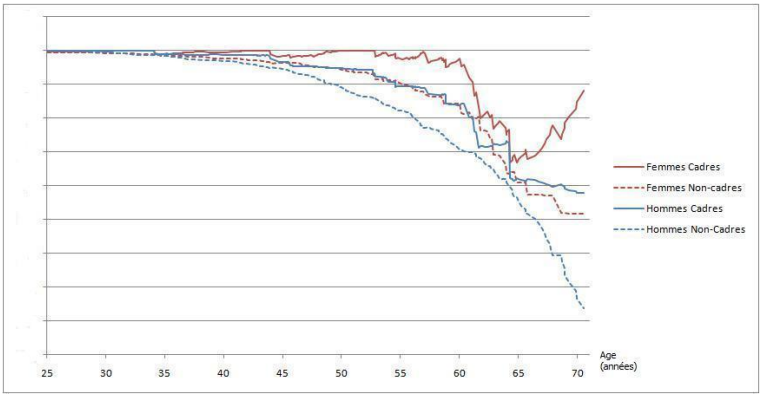


FIG. 2.18 – Survie $\hat{S}$ des différentes classes - Aalen

Malgré cette diminution du nombre de paramètres à estimer, les courbes de survie issues des valeurs de coefficients cumulés ne sont toujours pas satisfaisantes. Outre le problème de survie supérieure à 1, c'est le caractère localement croissant de certaines courbes qui pose un véritable problème (cf : Femmes/Cadres aux alentours de 65 ans). C'est la raison pour laquelle nous avons ajouté à la construction de ces survies la contrainte suivante :

$$\tilde{S}(T_i) = \min(\hat{S}(T_i), \hat{S}(T_{i-1})$$

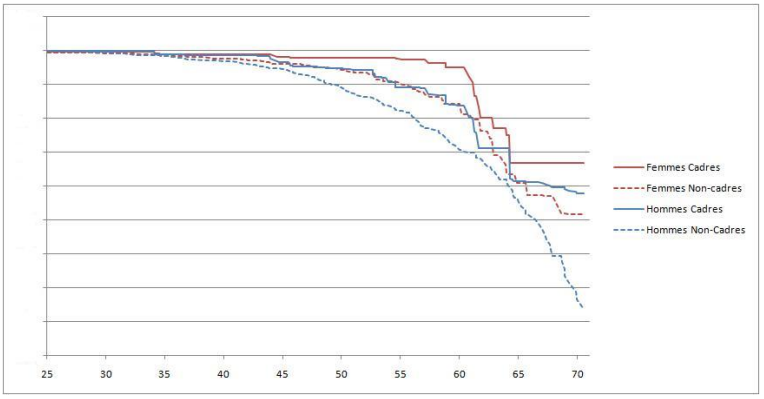


FIG. 2.19 – Survie $\tilde{S}$ des différentes classes - Aalen

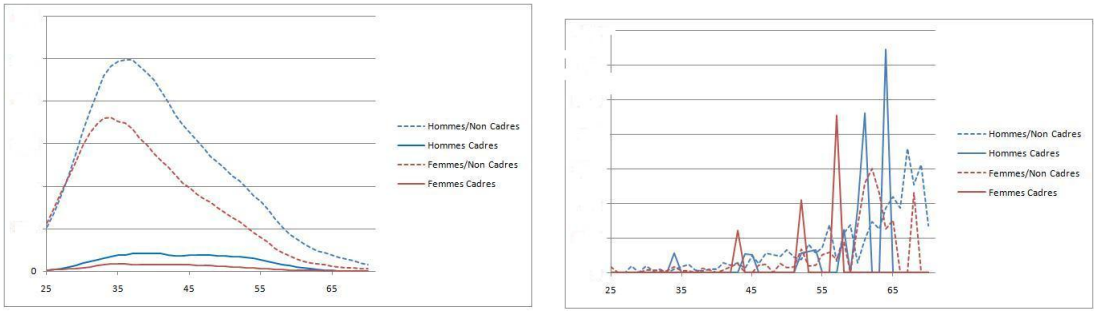
de telle sorte que les courbes de survie présentent alors l'aspect :

Cette nouvelle contrainte prise en compte, pour une même CSP, nous remarquons alors une surmortalité masculine. De plus, les individus cadres présentent une survie plus élevée que les non cadres de même sexe, ce qui corrobore les résultats obtenus avec la modélisation de Cox. On retrouve notamment la similarité des courbes de survie entre les catégories Hommes/Cadres et Femmes/NonCadres mises en évidence dans le modèle de Cox, ceci étant toutefois moins marqué ici.

2.2.3 Bilan de l'estimation de la survie

4 classes de risques

Le premier bilan de ces estimations est donc la mise en évidence d'un effet "Sexe" et "Cadre" sur la mortalité. Pour la suite du mémoire, nous évoquerons donc les 4 classes de risques associées : Hommes/NonCadres - Hommes/Cadres - Femmes/NonCadres - Femmes/Cadres. C'est pour ces 4 classes que nous chercherons à construire les tables de mortalité. Il est alors intéressant d'ajouter à l'étude descriptive présentée ci avant une précision concernant l'exposition au risque de ces classes et les taux de mortalité empiriques observés compte tenu de notre portefeuille. En effet, nous verrons par la suite que la composition même de ces classes rendra difficile une validation satisfaisante de la table, en raison du faible effectif de certaines classes (classes de type Cadres notamment) duquel découle le faible nombre de décès observés et donc des taux de mortalité empiriques très peu réguliers qu'il faudra pourtant estimer et ajuster (cf figure 2.20).

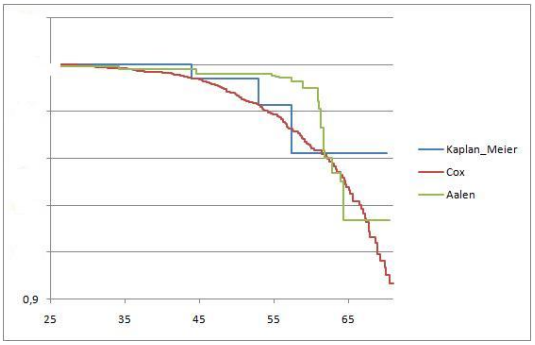


(a) (b)

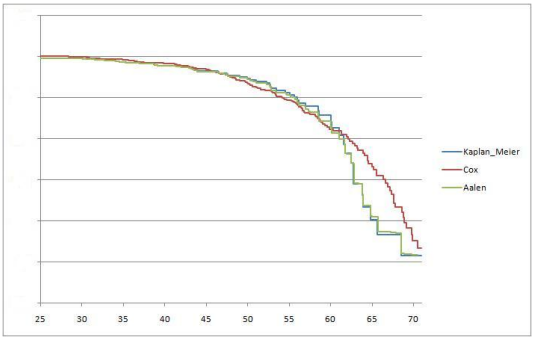
FIG. 2.20 – (a) :Exposition au risque par âge (b) :Taux bruts de mortalité associés

Aalen/Cox vs Kaplan-Meier

Rappelons ensuite qu'un des objectifs de ce mémoire est de s'interroger sur la pertinence de l'utilisation des modèles de Cox et de Aalen dans le but d'éviter une segmentation préalable des données lorsqu'il s'agit de différencier les individus selon leurs caractéristiques lors de la construction de tables de mortalité. Dans le but d'illustrer cette pertinence, il nous a semblé intéressant de comparer graphiquement les survies obtenues par ces méthodes pour nos 4 classes de risque, à celle d'une estimation de type Kaplan-Meier réalisée à partir d'une segmentation initiale de la base de données. (cf figure 2.21 et 2.22)

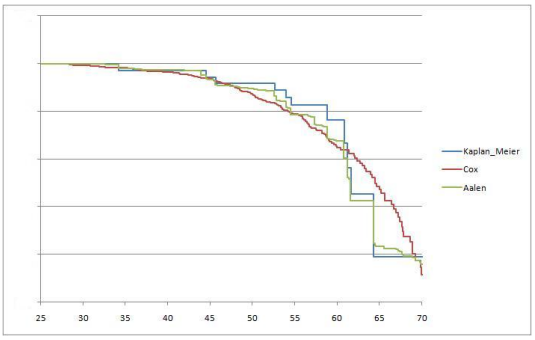


(a)

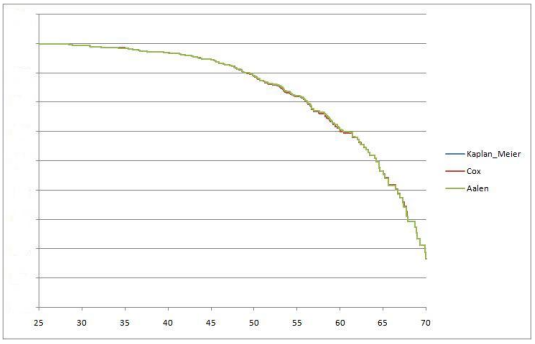


(b)

FIG. 2.21 – (a) :Femmes/Cadres (b) :Femmes/Non Cadres



(c)



(d)

FIG. 2.22 – (c) :Hommes/Cadres (d) :Hommes/Non Cadres

On retrouve alors, pour ce qui est du modèle de Cox, des effets attendus compte tenu de la théorie : les courbes "Cox" et "Kaplan-Meier" sont identiques pour la classe de référence Hommes/Non Cadres, ce qui est logique puisque justement la survie est estimée par un Kaplan-Meier pour cette sous population. Ensuite, les autres courbes pour les autres classes ne sont qu'un ajustement de cette dernière à partir des coefficients issus de la maximisation de la vraisemblance partielle. C'est la raison pour laquelle on retrouve des survies à l'allure plutôt lisse, au contraire des estimations Kaplan-Meier pour lesquelles on n'observe des sauts que lorsqu'il y a eu une sortie non censurée dans la population concernée. Ainsi le modèle de Cox, en ayant choisi au préalable une population de référence aux effectifs conséquents (ce qui a motivé notre choix des hommes non cadres comme référence), permet d'ajuster aux autres classes significatives une survie plus régulière que l'estimation Kaplan-Meier, et ainsi s'affranchir d'un éventuel problème de sous-effectif de la classe considérée.

En ce qui concerne les survies issues du modèle à hasard additif de Aalen, on remarque en revanche que les estimations de la survie semblent plus proches de celles obtenues par Kaplan-Meier. En effet, en théorie, l'estimation dans la modélisation additive permet d'accéder directement à la survie en proposant une estimation des coefficients de hasard cumulés à chaque date de sortie. Et, contrairement à la modélisation proportionnelle où les coefficiens ne dépendent pas du temps et où il est nécessaire d'obtenir la dépendance temporelle du hasard par une méthode supplémentaire, cette dépendance dans le temps de la survie va être contenue directement dans l'estimation de ces coefficients pour le modèle additif. Ceci présente un

avantage indéniable par rapport à notre utilisation du modèle de Cox où l'on n'estime la survie "temporellement" qu'une seule fois sur une population, estimation que l'on ajuste ensuite par un simple coefficient. En effet l'allure générale temporelle de la survie de cette population de référence va être répercutée sur les autres sous-populations, alors que le modèle additif permet d'ajuster sortie par sortie la dépendance temporelle de la survie afin d'être le plus fidèle possible à la survie de la sous-classe considérée. Pour résumer, le modèle de Aalen permet de mieux prendre en compte l'hétérogénéité de la survie des sous-classes.

Mais immédiatement, notre base de données nous a permis de mettre en évidence le défaut pratique de ce modèle. Puisqu'il est construit pour être très fidèle au comportement de chaque sous-classe, il semble de fait très sensible à la taille de ces dernières. Ainsi les estimations des survies pour des sous-populations avec un nombre très faible de sorties censurées sont relativement erratiques et il nous a alors semblé dangereux d'utiliser ce modèle pour déterminer quelles caractéristiques individuelles influencent la survie. Et même en ne l'utilisant par la suite qu'en tant que simple estimateur de survie pour des classes de risques déterminées par ailleurs (Cox), le modèle additif de Aalen présente ce défaut. Toutefois au prix de quelques contraintes ex-post on retrouve tout de même son avantage théorique sur le modèle à hasard proportionnel en ce qui concerne sa fidélité de prise en compte de l'hétérogénéité comme l'attestent ces graphiques comparatifs.

Pourtant la simple estimation de la survie, même si elle permet une conclusion partielle sur la pertinence des modèles, n'est pas l'objectif final de notre étude qui a pour but de déterminer des tables de mortalité segmentées pour les catégories de risque indentifées comme pertinentes. Et notamment, la construction qui va suivre des taux de mortalité par âge à partir de ces survies sera l'occasion pour nous d'effectuer d'autres conclusions sur la pertinence pratique de nos deux modélisations.

2.3 Tables de mortalité segmentées

Nous venons donc de mettre en évidence dans la partie précédente l'impact de certaines caractéristiques individuelles sur la mortalité : le Sexe et le fait d'être considéré comme Cadre ou non. Nous avons pu en déduire pour chacune des sous-populations concernées une estimation quantitative de la survie aux âges de sortie, par deux méthodes, de Cox ou de Aalen, pour lesquelles nous venons de discuter leur pertinence théorique et pratique. Nous sommes conscients qu'un éventuel lecteur "statisticien" souffrira de l'absence d'intervalles de confiance autour de ces estimations, ou tout du moins de l'absence de considérations à leur sujet, notamment dans le cadre de l'estimation de Aalen. En effet, nous n'avons pas traité ce point car notre objectif reste in fine la construction de tables de mortalité segmentées dont la pertinence sera testée moins au fil des étapes menant à sa construction, qui constitue une sorte de "boîte noire" , que lors de la confrontation des taux de sortie présents dans ces tables aux données réellement observées. Et dans le cadre de la construction de ces tables sous leur forme classique, c'est-à-dire l'obtention par tranche d'âge d'exposition au risque d'un taux de mortalité q_x, seule la valeur des survies issues de ces estimations nous sera utile, puisque c'est au travers de la pertinence de ces

q_x face à la réalité que nous validerons ou non l'ensemble des estimations qui nous amènent à les déterminer.

Nous présentons dans cette partie l'ensemble des méthodologies et résultats liés à la détermination des q_x pour chacune des catégories concernées, en différenciant ceux s'appuyant sur le modèle à hasard proportionnel, et ceux s'appuyant sur le modèle à hasard additif.

2.3.1 Détermination des taux bruts pour les âges 25/70 ans et ajustement à une loi Makeham

Nous disposons donc, quelle que soit la situation ou la classe d'âge considérées, d'une estimation de la fonction de survie $\hat{S}$ pour chaque date de décès en portefeuille que nous notons $T_1 < \ldots < T_r$. D'un point de vue fonctionnel on peut alors poser pour tout instant t, $\hat{S}(t) = \hat{S}(T_i)$ avec i tel que $T_i \leq t < T_{i+1}$. Bien évidemment, comme on se raccorde à la valeur de la survie à gauche, par décroissance de la fonction de survie, cet estimateur a tendance à surestimer la survie, notamment aux grands âges où la décroissance est la plus forte. Dans une optique prudentielle de tarification, cela ne va toutefois pas dans le mauvais sens comme nous le verrons ci dessous.

Car, puisque la survie désigne la probabilité qu'un individu soit vivant à un instant t donnée $S(t) = \mathbb{P}(T > t)$ (T est la durée de vie de l'individu), on peut définir la probabilité de décéder à l'âge entier x (ie entre les instants x et $x + 1$) sachant que l'on est vivant en x, q_x, comme étant $q_x = 1 - \frac{S(x+1)}{S(x)}$. En effet, on a :

$$1-\frac{S(x+1)}{S(x)} = 1-\frac{\mathbb{P}(T > x+1)}{\mathbb{P}(T > x)} = 1-\frac{\mathbb{P}(T > x+1; T > x)}{\mathbb{P}(T > x)} = 1-P(T > x+1|T > x)$$

C'est-à-dire 1 moins la probabilité de rester en vie jusqu'à l'âge $x+1$ sachant que l'on est en vie à l'age x. On a donc bien défini q_x. D'un point de vue prudentiel, plus que la surestimation de la survie réelle inhérente au modèle comme nous l'avons vu, ce qui est important pour q_x est l'écart relatif entre la survie à l'âge $x + 1$ et la survie à l'âge x. Donc l'écart par rapport à la vraie valeur de q_x suite à la surestimation de la survie est ambigu.

On pose alors comme estimateur des taux bruts $\hat{q}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)}$ à partir de notre estimateur de la fonction de survie défini ci-avant.

Naturellement, la courbe des taux bruts ainsi estimés va présenter des irrégularités en fonction de l'âge en raison des irrégularités initiales des estimateurs de la fonction de survie. Or, on peut légitimement supposer que ces variations assez brusques ne sont pas dues à des variations de l'incidence réelle du risque, mais sont plutôt la conséquence d'une insuffisance de données.

C'est la raison pour laquelle il est nécessaire d'ajuster ces taux bruts à une fonction paramétrique censée modéliser le vrai risque sous-jacent afin d'obtenir une table de mortalité conforme et notamment des q_x strictement croissants et "réguliers".

Ajustement par maximum de vraisemblance discret

On suppose que la vraie probabilité de décéder à l'âge x notée $q_x(\theta)$ dépend

d'un paramètre θ vectoriel. Nous allons présenter ici une méthode permettant de le déterminer afin que cette réalité s'ajuste au mieux à nos taux bruts $\hat{q}_x$ estimés par ailleurs.

Considérons ainsi le nombre de décès de l'âge x, $\tilde{d}_x$, estimé à partir des $\hat{q}_x$: $\tilde{d}_x = \hat{q}_x \times N_x$ où N_x désigne le nombre de personnes exposées au risque sur la période $[x, x+1]$. En toute rigueur, $\tilde{d}_x$ suit une loi binomiale $\mathbb{B}(N_x, q_x(\theta))$, puisque $q_x(\theta)$ est supposé représenter la vraie probabilité de décès sur la période, et que l'on cherche à modéliser le nombre de réalisations de l'événement "décès" de probabilité $q_x(\theta)$ sur un échantillon de taille N_x. Utilisons alors l'approximation gaussienne d'une expérience binomiale (on rappelle le critère de Cochran validant cette approximation pour une binomiale $\mathbb{B}(n,p)$: $n \times p \geq 5$ et $n \times (1-p) \geq 5$). On a alors

$$\tilde{d}_x \approx \mathbf{N}\left(N_x q_x(\theta), N_x q_x(\theta)(1 - q_x(\theta))\right)$$

et donc par suite, comme $\tilde{d}_x = \hat{q}_x \times N_x$, la loi de $\hat{q}_x$ peut être approchée par

$$\hat{q}_x \approx \mathbf{N}\left(q_x(\theta), \sigma^2(\theta) = \frac{q_x(\theta)(1-q_x(\theta))}{N_x}\right)$$

Pour l'ensemble des observations $(\hat{q}_n, \ldots, \hat{q}_m)$, la fonction de vraisemblance dépendant du paramètre θ peut donc s'écrire :

$$L(\hat{q}_n, \ldots, \hat{q}_m | \theta) = \prod_{x=n}^{m} \frac{1}{\sigma(\theta)\sqrt{2\pi}} \exp\left(\frac{1}{2}\frac{(q_x(\theta) - \hat{q}_x)^2}{\sigma^2(\theta)}\right)$$

D'où une log-vraisemblance

$$\ln L(\hat{q}_n, \ldots, \hat{q}_m | \theta) = \sum_{x=n}^{m} \ln \frac{1}{\sigma(\theta)\sqrt{2\pi}} - \sum_{x=n}^{m} \frac{1}{2}\frac{(q_x(\theta) - \hat{q}_x)^2}{\sigma^2(\theta)}$$

et la maximisation de la vraisemblance en θ est alors équivalente à la minimisation de

$$\sum_{x=n}^{m} \frac{(q_x(\theta) - \hat{q}_x)^2}{\hat{\sigma}^2} = \sum_{x=n}^{m} \frac{N_x}{\hat{q}_x(1-\hat{q}_x)}(q_x(\theta) - \hat{q}_x)^2$$

(on a remplacé la variance de la loi des $\hat{q}_x$, $\sigma(\theta)$, par un estimateur empirique $\hat{\sigma}^2$).

A condition de bien choisir la valeur initiale du paramètre θ_0, de nombreux logiciels de calcul sont capables de converger vers un argmin pour cette fonction et donc de déterminer un $\hat{\theta}$ acceptable pour l'obtention de $q_x(\theta)$ et ainsi conclure l'ajustement.

Loi Makeham

Comme il est fait classiquement pour la construction de tables de mortalité, nous avons choisi d'ajuster nos taux bruts $\hat{q}_x$ par une loi Makeham suivant la méthode présentée ci-avant.

La loi Makeham suppose le hasard de la forme (paramétrique) suivante :

$$h(x) = a + b \times c^x$$

qui représente donc le taux instantané de décès à l'âge x. Rappelons que le paramètre a peut s'interpréter comme un taux instantané de décès accidentels $(a \geq 0)$; le

coefficient bc^x correspondant lui à un effet associé au vieillissement, qui fait croître le taux de décès de façon exponentielle. Compte tenu de la croissance des taux de décès avec l'âge, on a nécessairement $c \geq 1,\ b \geq 0$.

De plus, d'après les relations reliant q_x, hasard et survie, on peut montrer que

$$q_x = 1 - sg^{c^x(c-1)} \text{ avec } s = \exp(-a) \text{ et } g = \exp(-\frac{b}{\ln c})$$

C'est donc ce $q_x(\theta),\ \theta = (a, b, c)$ ainsi défini, que nous avons introduit dans la fonction

$$C(\theta) = \sum_x \frac{N_x}{\hat{q}_x(1-\hat{q}_x)}(q_x(\theta) - \hat{q}_x)^2,$$

que l'on se doit de minimiser pour ajuster les taux bruts construits à partir des survies issues des modèles de Cox et Aalen.

Ajustement Makeham pour le modèle de Cox

En pratique, dans le cadre de la modélisation Cox, nous avons effectué ce lissage à partir de la Survie S_0 estimée dans le modèle (Survie de la Classe "Hommes Non-Cadres"). Nous avons ainsi obtenu, à partir de cette survie, des estimateurs bruts théoriques $\hat{q}_x^0$ pour des âges de 25 à 70 ans, que nous avons ajustés à une loi Makeham suivant la méthode présentée ci-avant, à l'aide des expositions au risque par âge de notre portefeuille pour cette catégorie de référence. Conformément à la théorie, nous avons minimisé, à l'aide du solveur d'Excel, la fonction critère

$$C(\theta) = \sum_x \frac{N_x}{\hat{q}_x^0(1-\hat{q}_x^0)}(q_x^0(\theta) - \hat{q}_x^0)^2$$

sous les contraintes $a \geq 0\ b \geq 0\ c \geq 1$, pour $\theta = (a, b, c)$, en partant d'une valeur initiale du paramètre θ_0 égal à celui issu d'un ajustement Makeham sur la table réglementaire TF 00-02. Nous nommons alors $\hat{\theta} = Argmin\, C(\theta)$.

Nous avons alors obtenu comme résultats :

θ	a	b	c
θ_0	0.0015	4.2E-06	1,12
$\hat{\theta}$	2,1E-08	8,9E-06	1,10

Toutefois, la valeur du paramètre a, censé représenter le taux de décès accidentels dans un modèle de Makeham, et qui contribue pour l'essentiel à l'évaluation de la mortalité des jeunes, est très faible. Ceci est bien évidemment dû à la nature de notre portfeuille où l'on trouve à la fois peu de décès et une grande exposition au risque pour les jeunes classes d'âge. Ainsi, lors de l'ajustement, les premiers termes de la fonction critère que l'on minimise sont responsables en grande partie de la valeur des paramètres estimés, et l'ajustement va alors accorder les paramètres de sorte de faire coïncider la série des $q_x^0(\hat{\theta})$ avec les premiers termes, proches de zéro, de la série des estimateurs bruts $\hat{q}_x^0$.

Or l'expérience montre que pour des portefeuilles d'assureur de contrats temporaires décès (Source : F. Planchet), le taux accidentel de mortalité peut raisonnablement être fixé de l'odre de $a = 0,015\%$. C'est la raison pour laquelle nous avons relancé notre ajustement avec cette nouvelle contrainte. On trouve alors :

θ	a	b	c
θ_0	0,00015	8,9E-06	1,10
$\hat{\theta}$	0,00015	7,6E-06	1,11

ce qui nous fournit graphiquement l'ajustement makeham $q_x^0(\hat{\theta})$ suivant pour les taux de sortie $\hat{q}_x^0$ à l'âge x de la classe "Hommes Non-Cadres" sur la période 25-70 ans.

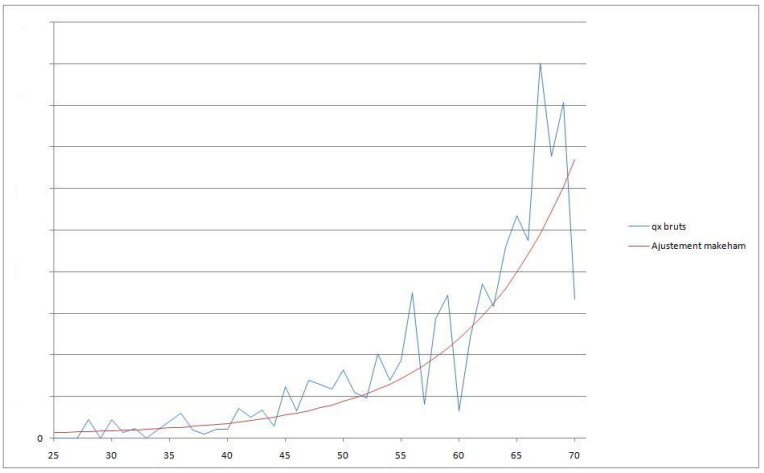


FIG. 2.23 – Ajustement Makeham des taux de décès par année $\hat{q}_x$ issus de la survie de la base

Une première constatation graphique nous amène à penser que malgré les précautions prises, les taux lissés sous-estiment globalement les taux bruts initialement estimés. Nous aurons l'occasion de revenir plus en détail sur ce point dans la partie consacrée à la validation par confrontation aux observations des taux ainsi obtenus.

Or ne perdons pas de vue que notre objectif est d'obtenir ces taux lissés non pas pour le seul hasard de base qui correspond dans le modèle de Cox à la classe des "Hommes Non-Cadres", mais pour chacune des 4 classes significatives mise en évidence. Pour cela, nous avons utilisé une formule de passage entre les taux $q_x^0(\hat{\theta})$ de la base et les taux q_x^Z de la classe "Z" à l'aide des coefficients β précédemment estimés. En effet, dans le cadre d'une modélisation à hasard proportionnel de type Cox, on a :

$$q_x^Z = 1 - \frac{S_Z(x+1)}{S_Z(x)} = 1 - \left(\frac{S_0(x+1)}{S_0(x)}\right)^{exp(Z\beta)} = 1 - (1 - q_x^0(\hat{\theta}))^{exp(Z\beta)}$$

et donc l'ajustement Makeham $q_x^0(\hat{\theta})$ des $\hat{q}_x^0$ nous permet de déterminer de façon "lisse" chacune des séries q_x^Z pour les 3 autres classes "Hommes Cadres" "Femmes Cadres" et "Femmes Non-Cadres"

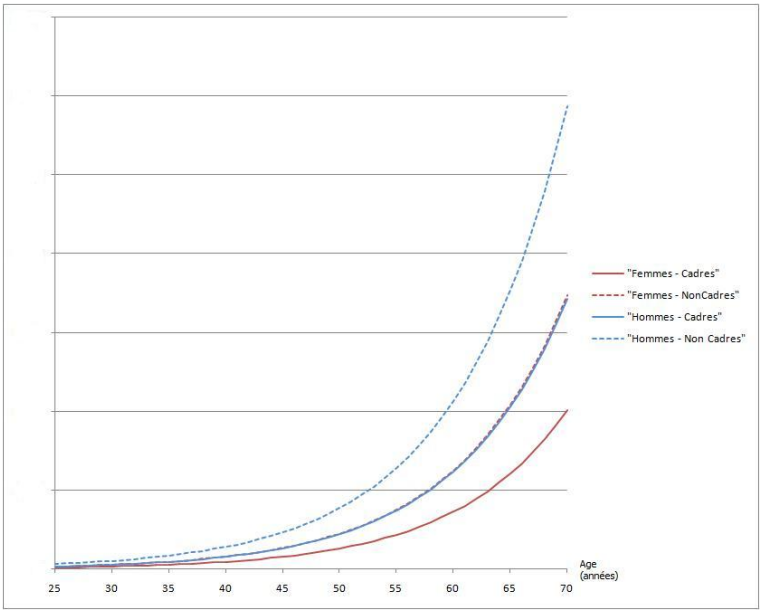


FIG. 2.24 – Taux annuels de décès par âge et par classe - Cox

En raison de la formule de passage ci-avant, et des valeurs des coefficients β issus de l'estimation, on trouve des taux de mortalité plus faibles de 41% chez les cadres par rapport aux non-cadres, et de 41% chez les femmes par rapport aux hommes, aux mêmes âges.

Ajustement Makeham pour le modèle de Aalen

Pour lisser les taux de mortalité issus des survies estimées à l'aide du modèle de Aalen, le subterfuge qui consiste à ne lisser que les taux bruts issus de la survie de la base n'est plus possible, car il n'y a pas de lien aussi simple entre les différentes survies. Nous avons donc appliqué les méthodes présentées précédemment pour chaque sous classe. En effet, disposant pour chacune d'elle d'une survie estimée par le modèle additif, nous avons pu introduire à nouveau la notion de taux bruts théoriques $\hat{q}_x^0$, que l'on a cherché à ajuster à un modèle de Makeham pour obtenir des taux théoriques lissés $q_x^0(\hat{\theta})$. Conformément à la remarque introduite dans le cadre proportionnel, notons que nous avons également contraint ici le taux accidentel à $a = 0,00015$.

On obtient alors

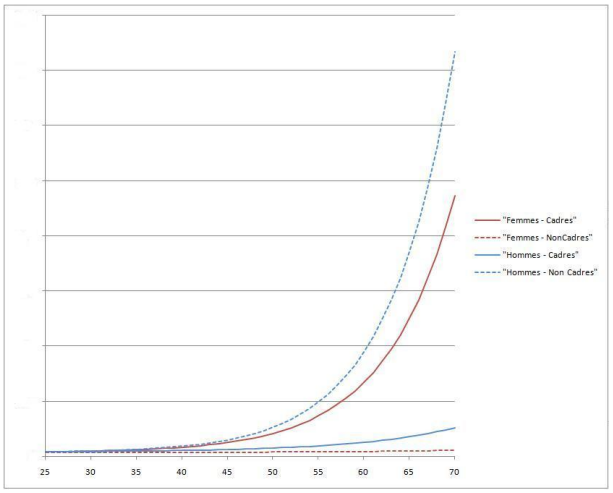


FIG. 2.25 – Taux annuels de décès par âge et par classe - Aalen

où l'on peut déjà noter une grande disparité avec les taux lissés du modèle de Cox, et notamment des taux bruts quasi nuls pour les Femmes non-cadres.

Rappelons une nouvelle fois que dans le cadre du modèle proportionnel on se contente d'ajuster les taux bruts par rapport à un seul lissage, et donc une fois les écueils de l'ajustement liés aux fluctuations d'échantillonnage de la classe de référence surmontés (on choisit d'ailleurs cette classe de référence dans le but d'avoir le plus d'observations possibles), il est simple de comparer les q_x entre eux. En revanche ici, les problèmes d'échantillonnage (peu de décès pour les plus grandes expositions au risque notamment) se rencontrent pour toutes les classes, et il devient alors très difficile d'ajuster des taux bruts très disparates (cf figure 2.26), proches des taux bruts empiriques (cf figure 2.20 (b)).

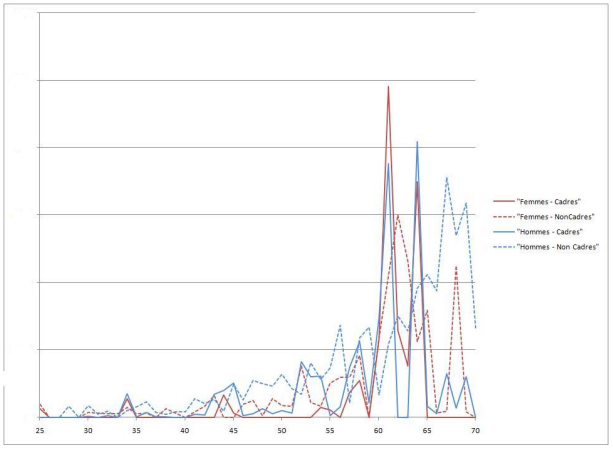


FIG. 2.26 – Taux annuels de décès bruts par âge et par classe - Aalen

Le lissage n'est donc pas satisfaisant dans le cadre du modèle de Aalen. Associée à la difficulté déjà mentionnée de déterminer la survie dans ce modèle, en raison des problèmes d'effectifs de notre échantillon, cette raison nous a poussés donc à aban-

donner le modèle de Aalen pour la suite de nos résultats. Nous nous concentrerons donc sur le modèle de Cox par la suite.

2.3.2 Validation des tables - Confrontation aux données réelles

1ère vérification

La confrontation des q_x estimés classes par classe, âge par âge, aux données réelles se heurte aussi au problème d'échantillonnage déjà évoqué : les taux bruts réellement observés sont très variables en raison du peu de données dont nous disposons, et en particulier, les taux de sortie empiriques à chaque âge, lorsqu'ils sont calculables (ie que l'on a effectivement observé un décès) ne sont pas forcément croissants et présentent de fortes irrégularités. De plus, le nombre de personnes exposées au risque par tranche d'âge est lui aussi fluctuant, avec notamment une sous-représentation des âges élevés. Ainsi, si l'on souhaite confronter nos taux issus des lissages à ces courbes, il est difficile de mener une bonne interprétation, puisque par construction, le lissage rend les q_x réguliers et croissants.

Afin de tenter de palier à ces problèmes d'échantillonnage, nous avons décidé d'adopter la méthode suivante pour confronter nos estimations à la réalité. Nous avons regroupé les classes d'âge, pour chaque catégorie, de façon à obtenir 10 classes (indicées par i par la suite) d'exposition au risque comparable. A chaque classe i correspond donc un certain nombre de classes d'âge x regroupées pour former cette nouvelle classe.

En pratique, sur notre base, cela revient à créer 10 groupes : les 25-30ans, 31-33ans, 34-35ans, 36-37ans, 38-39ans, 40-42ans, 43-45ans, 46-49ans, 50-54ans, et enfin les 55-70ans.

On cherche alors a comparer sur cette catégorie de population le nombre de décès observé empiriquement $\tilde{D}_i$ a celui obtenu théoriquement D_i^{th} à l'aide de nos q_x théoriques. Afin de voir si l'estimation est robuste on s'intéressera alors au nombre de décès prédits globalement dans chaque sous-population, ainsi qu'à la position de $\tilde{D}_i$ par rapport à D_i^{th} et à l'intervalle de confiance qui lui est associé pour chaque classe d'âge.

Détaillons ce dernier point. On pose $N_i = \sum_{x \in i} N_x$ l'effectif de la classe i, et q_i la probabilité de décès sur une année dans cette même classe. On a alors :

$$D_i \sim \mathbf{B}(N_i, q_i) \approx \mathbf{N}\left(N_i q_i, N_i q_i (1 - q_i)\right)$$

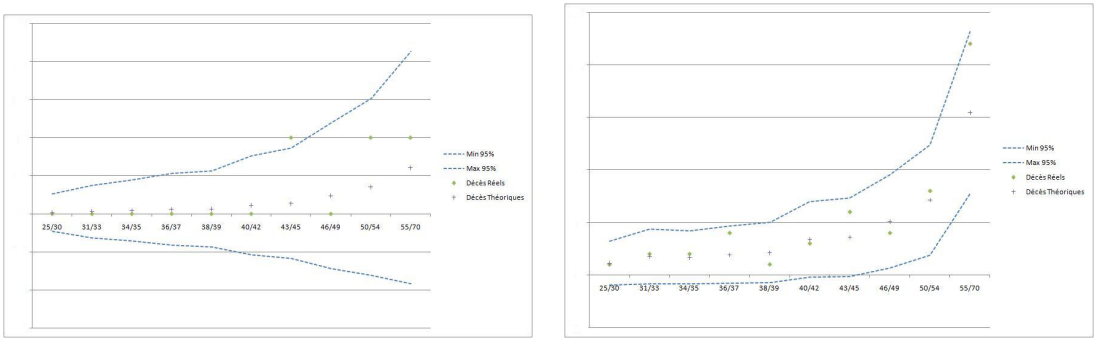
par approximation d'une loi binomiale par une loi normale comme déjà vu ci-avant.

Ainsi on peut dire que l'intervalle de confiance à 95% pour le nombre de décès réel D_i est :

$$\mathbf{P}\left(D_i \in \left[N_i q_i - 1,96\sqrt{(N_i q_i (1 - q_i))}, N_i q_i + 1,96\sqrt{(N_i q_i (1 - q_i))}\right]\right) = 95\%$$

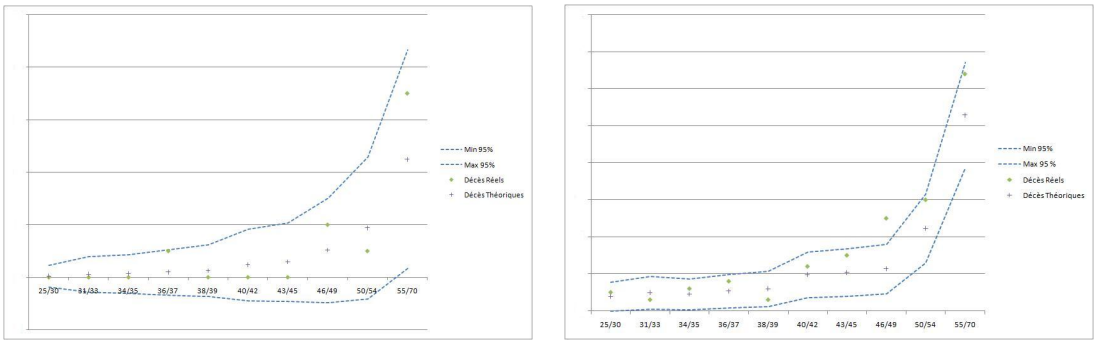
Or, un estimateur empirique de q_i peut-être donné par $\hat{q}_i = \frac{D_i^{th}}{N_i}$ de telle sorte que

$$\mathbf{P}\left(D_i \in \left[D_i^{th} - 1,96\sqrt{\left(D_i^{th}(1 - \frac{D_i^{th}}{N_i})\right)}, D_i^{th} + 1,96\sqrt{\left(D_i^{th}(1 - \frac{D_i^{th}}{N_i})\right)}\right]\right) = 95\%$$



(a) (b)

FIG. 2.27 – (a) :Femmes/Cadres (b) :Femmes/Non Cadres



(c) (d)

FIG. 2.28 – (c) :Hommes/Cadres (d) :Hommes/Non Cadres

Pour valider la qualité de nos modèles, il convient alors de placer le nombre de décès observés $\tilde{D}_i$, dans cet intervalle de confiance obtenu à l'aide des estimations théoriques du nombre de décès par classe $\tilde{D}_i$.

Classiquement, on utilise par ailleurs la formule

$$D_i^{th} = \sum_{x \in i} q_x^{th} N_x$$

pour déterminer ces décès théoriques à partir de nos taux q_x déterminés précédemment

En utilisant comme convenu les q_x de Cox utilisés précédemment, on trouve pour chaque classe les résultats présentés par les graphiques ci-dessus (fig 2.27, fig 2.28).

Et en terme de décès total :

	Total Décès Réels	Total Décès Théoriques
Hommes Non Cadres	171	131,2
Hommes Cadres	11	9,2
Femmes Non Cadres	53	43,4
Femmes Cadres	3	1,7

On constate donc que, bien que les décès observés soient pour chaque classe d'âge dans l'intervalle de confiance statistique à 95% des décès théoriques, ils sont relativement proches de la borne supérieure de ces derniers, notamment pour la classe de référence Hommes Non-Cadres pour laquelle l'ajustement était censé le mieux résister aux fluctuations d'échantillonnage. Par conséquent le nombre de décès global de chaque sous-population est ainsi sous-estimé avec nos taux théoriques ajustés, comme nous l'avions pressenti au vu de la figure 2.23 .

Cette double confrontation à la réalité, globale et classe par classe, nous pousse donc à rejeter cette méthode de construction de la table, car, par cette sous estimation, elle entraînerait évidemment une perte technique lors de la tarification.

Nous avons cherché à comprendre d'où pouvait provenir un tel écart.

La première idée fut pour nous de se pencher sur la pertinence ou non d'appliquer une modélisation à hasard proportionnel sur nos données. Pour cela nous nous sommes appuyés sur un test, dont le principe est étudié en détail par Therneau et Grambsch [2000], qui se basent sur les résidus de Schoenfeld.

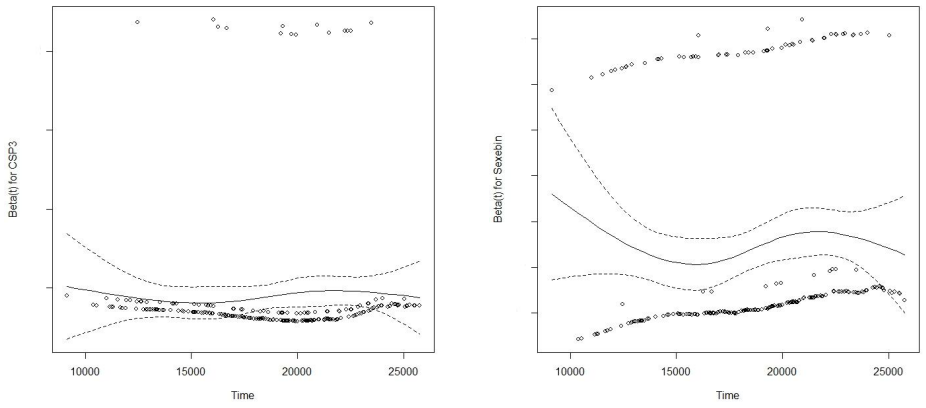
Ces derniers sont définis pour chaque individu $i = 1, \dots, n$ et chaque covariable Z_j, $j = 1, \dots, p$ comme la différence entre la valeur, à la date X_i de sortie de i, de la covariable pour cet individu Z_j^i et sa valeur attendue :

$$r_{ij} = d_i \left(Z_j^i - \frac{\sum_{k \in R(T_i)} e^{Z^k \beta} Z_j^k}{\sum_{k \in R(T_i)} e^{Z^k \beta}} \right)$$

En introduisant alors le produit de l'inverse de la matrice de variance-covariance des résidus de Schoenfeld pour l'individu i avec le vecteur de ces mêmes résidus, appelé résidu de Schoenfeld réduit, on peut montrer que la moyenne de ce dernier $m(X_i)$ vérifie la relation $\beta(X_i) = \beta + m(X_i)$ lorsque l'on suppose que les coefficients β dépendent du temps selon une relation $\beta(t) = \beta + g(t)$. Or, tester l'hypothèse de proportionnalité, c'est justement tester H_0, $g(t) = 0$, et en utilisant des méthodes de Monte Carlo, ces auteurs ont proposé un test basé sur la suite des $\beta(X_i)$, $i = 1, \dots, n$. Outre une p-value permettant de déterminer un seuil de rejet du test, la fonction cox.zph du package $survival$ de R permet d'en effectuer une interprétation graphique : sous l'hypothèse nulle, la courbe, lissée, des $\beta(X_i)$ ne doit pas laisser transparaître de tendance, et s'apparenter à une horizontale.

Et pour nos données, le test de proportionnalité est validé pour le seuil de 95% aussi bien globalement que pour les variables Sexe et CSP3 :

```
           rho chisq     p
Sexebin 0.0331 0.264 0.608
CSP3    0.0390 0.363 0.547
GLOBAL      NA 0.599 0.741
```



(a) (b)

FIG. 2.29 – (a) :Résidus de Schoenfeld réduits - CPS3 (b) :Résidus de Schoenfeld réduits - Sexe

Les graphiques des résidus concordent d'ailleurs bien avec l'absence de tendance temporelle (fig 2.29).

En plus de cette non désapprobation statistique de la modélisation à hasard proportionnel, une autre évidence graphique a pu nous mettre sur la piste de l'endroit où notre méthodologie semble montrer ses limites. En effet, lorsque l'on compare les taux bruts empiriques de la classe de référence avec les taux bruts issus de l'estimation de la survie par Kaplan-Meier, on se rend bien compte que l'adéquation est quasi-parfaite :

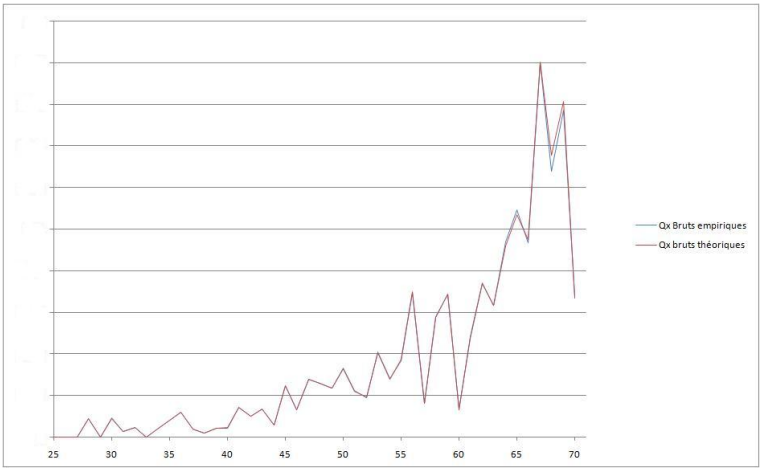


FIG. 2.30 – Taux annuel de décès par âge et par classe - Cox

il semblerait que c'est bien le lissage de ces taux bruts théoriques qui pose problème en raison des soucis d'échantillonnage déjà maintes fois évoqués et malgré les précautions prises à leurs égards.

Une nouvelle méthode de lissage

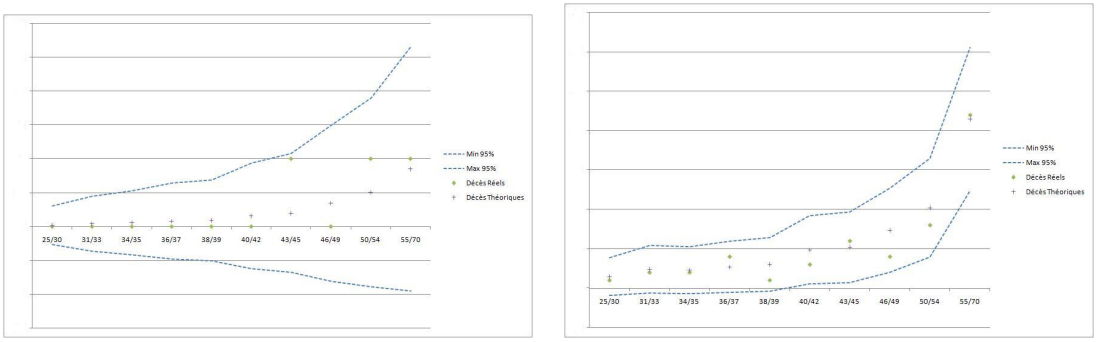
Afin de palier à ce souci de lissage, et obtenir en dernier lieu des taux bruts aboutissant à une tarification prudente, nous avons décidé d'adopter la méthode suivante :

Les taux bruts théoriques de la classe de référence issus de l'estimation de la survie, bien que fidèles à la réalité observée en raison de l'estimation Kaplan-Meier, sont soumis eux aussi aux fluctuations d'échantillonage lorsqu'il s'agit de faire le lien avec la vraie mortalité. On peut, donc autour de ces taux bruts estimés par la formule $q_x = 1 - \frac{S_{x+1}}{S_x}$, construire un intervalle de confiance à 95% du type

$$\mathbf{P}\left(q_x \in \left[q_x - 1,96\sqrt{\left(\frac{q_x}{N_x}(1-q_x)\right)}, q_x + 1,96\sqrt{\left(\frac{q_x}{N_x}(1-q_x)\right)}\right]\right) = 95\%$$

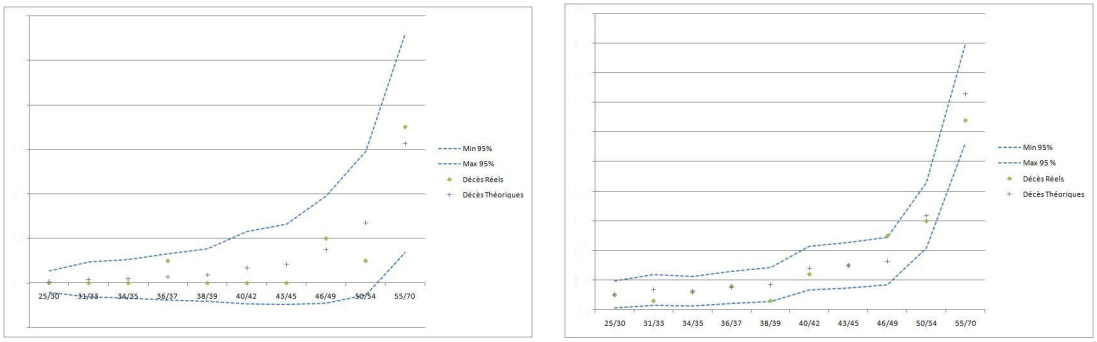
Statistiquement, toute suite de q_x dont les valeurs sont comprises pour chaque x entre les bornes de cet intervalle convient donc a priori pour estimer le vrai taux de mortalité, compte tenu de nos données, avec une confiance à 95%. Fort de cette remarque, nous avons alors cherché à lisser, non pas les points médians (les q_x) comme fait précédemment, et comme réalisé classiquement dans ce genre de cas, mais d'autres suites de q_x qui seraient plus susceptibles de rendre compte au final du nombre de décès réels de notre base. Nous avons pour cela utilisé diverses suites correspondant à des limites supérieures d'intervalles de confiance de seuils moins restrictifs que 95%, afin de s'assurer d'être à la fois dans cet intervalle, mais aussi au dessus des q_x médians, afin de résoudre ce problème de sous-estimation de la mortalité. La suite de q_x correspondant à la classe de référence que nous avons lissée puis adaptée aux différentes classes grâce au modèle à hasard proportionnel est la suite de q_x correspondant au seuil à 50%. Elle est retenue comme satisfaisante, puisqu'en lui appliquant la méthode de lissage Makeham pour obtenir des $q_x^0(\hat{\theta})$ comme introduit précédemment , la confrontation à la réalité des taux ainsi ajustés pour chaque classe donne les résultats suivants :

θ	a	b	c
θ_0	0,00015	8,9E-06	1,10
$\hat{\theta}$	0,00015	1,4E-05	1,11



(a) (b)

FIG. 2.31 – (a) :Femmes/Cadres (b) :Femmes/Non Cadres



(c) (d)

FIG. 2.32 – (c) :Hommes/Cadres (d) :Hommes/Non Cadres

	Total Décès Réels	Total Décès Théoriques
Hommes Non Cadres	171	184,5
Hommes Cadres	11	13,1
Femmes Non Cadres	53	61,2
Femmes Cadres	3	2,4

Ainsi c'est donc à partir d'une formule Makeham

$$q_x = 1 - sg^{c^x(c-1)} \text{ avec } s = \exp(-a) \text{ et } g = \exp(-\frac{b}{\ln c})$$

avec

θ	a	b	c
$\hat{\theta}$	0,00015	1,4E-05	1,11

et une formule de passage

$$q_x^Z = 1 - \frac{S_Z(x+1)}{S_Z(x)} = 1 - \left(\frac{S_0(x+1)}{S_0(x)}\right)^{exp(Z\beta)} = 1 - (1 - q_x^0(\hat{\theta}))^{exp(Z\beta)}$$

que nous avons déterminé les taux de mortalité par classe de risque sur la période 25-70 ans.

A noter qu'il est d'une part possible d'étendre ces taux à des âges à la fois plus et

moins élevés selon l'usage (reconduction possible des contrats jusqu'à 80 ans, souscription légalement possible à partir de 12 ans), en appliquant les formules précédentes avec un x adéquat.

De plus, notons qu'une autre façon d'échapper au problème d'échantillonnage, entraînant un lissage global hasardeux, aurait pu être d'effectuer des lissages Makeham différents sur certaines tranches d'âges et d'effectuer ensuite des raccordements. Nous lui avons préféré la méthode précédente qui permet en outre d'envisager une tarification prudente en surestimant les taux bruts empiriques.

2.3.3 Bilan

Résumons la méthodologie qui nous a permis d'aboutir à ces taux :
- Détermination des variables explicatives retenues comme significative par le modèle et estimation de leur influence
- Estimation de la survie par classe de risque
- Détermination de taux bruts à l'aide de ces survies
- Ajustement Makeham en vue d'un lissage des taux bruts
- Confrontation des décès estimés à l'aide des taux lissés à la réalité des décès observés
- Amélioration du lissage pour ajustement à la réalité

Compte tenu de la problématique de notre sujet, nous avons mis en évidence lors de cette construction un avantage certain du modèle à hasard proportionnel.

Si, en effet, il est possible dans les deux méthodes d'agréger les individus afin de déterminer statistiquement les effets de telle ou telle covariable sur la mortalité, la construction de tables segmentées qui s'en suit nécessite elle toutefois d'utiliser les effectifs (exposition et décès) d'au moins une sous-population afin d'effectuer le lissage sous-jacent. Le modèle à hasard proportionnel se contente d'une seule de ces sous-populations, qui peut être choisie de telle sorte à limiter le problème d'insuffisance de données, mais le modèle à hasard additif impose lui l'utilisation des tables de données segmentées, dont on voulait pourtant se prémunir de l'usage compte tenu de notre problématique.

De plus, la vérification de nos résultats sur les observations issues de notre base de données s'est elle aussi révélée difficile en raison des problèmes d'insuffisance de données. Bien qu'ayant eu le mérite de mettre en évidence un souci dans la méthode d'ajustement, (les effectifs de la classe de référence, bien que les plus importants, ne suffisent pas à déterminer une loi Makeham satisfaisante), et donc de nous faire réfléchir à d'autres méthodes, plus prudentes par ailleurs, il nous a semblé qu'il aurait été plus judicieux de confronter ces estimations à un portefeuille différent de celui qui a entraîné les estimations. Toutefois, nous ne disposions pas d'un tel échantillon test.

Chapitre 3

Application à la tarification

3.1 Tables réglementaires et tables d'expérience

Nous proposons ici une comparaison des taux de mortalité réglementaires et des taux obtenus par la méthode de Cox. Cette comparaison est pertinente puisque, d'après l'Article A335-1 du code des Assurances, l'assureur est en droit d'utiliser, par défaut, les tables réglementaires TH00-02 et TF00-02 pour tarifer les contrats temporaires décès. Il paraît donc naturel de mettre en parallèle ces deux outils ou bases de tarification.

Le graphe suivant présente cette comparaison.

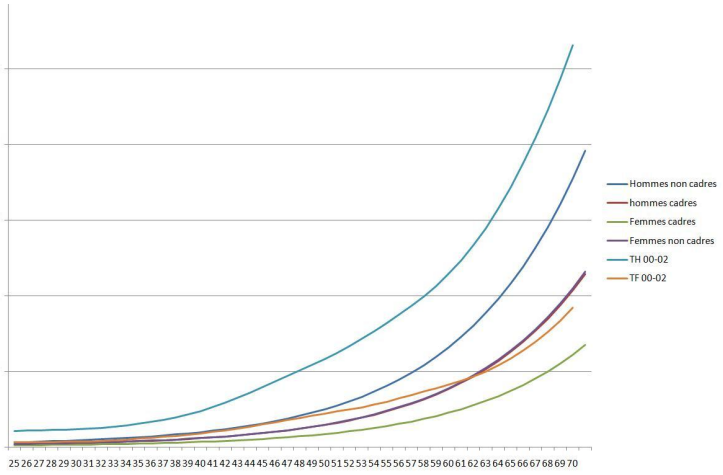


FIG. 3.1 – Comparaison des taux de mortalité issus des tables réglementaires et d'expérience

Concernant les hommes, toutes les catégories estimées précédemment présentent une mortalité inférieure à celle indiquée par la table réglementaire TH00-02. Notre population masculine d'assurés est donc caractérisée par une moindre mortalité par rapport à la population générale. Les tables estimées sont donc, a priori, moins prudentes que la table réglementaire.

Concernant les femmes, les cadres présentent elles aussi une mortalité inférieure. Pour les non cadres, la même remarque est faite jusqu'à l'âge de 62 ans, puis elles se caractérisent par une mortalité plus forte que ce que prévoit la table TF00-02. Globalement, les femmes elles aussi sont touchées par une mortalité plus faible que la

population générale féminine.

En terme de résultat, considérons la situation suivante : deux assureurs proposant les contrats temporaires décès présentés dans ce mémoire sont en concurrence et sont confrontés à la population d'assurés définie par le portefeuille décrit précédemment. Le premier assureur use des tables réglementaires pour tarifer, le second se sert des tables d'expérience segmentées. Majoritairement, les tarifs proposés par le second sont plus faibles, d'autre part, la segmentation permet de s'affranchir, certes partiellement, du problème d'antiselection qui se pose pour le premier assureur. Finalement, le second assureur jouit de résultats bien supérieurs au premier grâce à la construction et à l'utilisation de tables de mortalité d'expérience segmentées.

Précisons que des tables de mortalité d'expérience doivent être certifiées par un actuaire indépendant, d'après l'Article A335-1 du Code des Assurances. Il faut cependant rappeler que cette certification ne concerne que le provionnement et non la tarification. Effectivement, cette procédure est définie par l'Institut des Actuaires et demande l'accord de la Commission de Contrôle des Assurances. Concrètement, une certification initiale de la table d'expérience est complétée par un suivi annuel pour vérifier s'il est pertinent et prudent d'utiliser de telles tables. Au final, il s'agit de s'assurer que ces dernières permettent "la constitution de provisions suffisantes et prudentes". Dans une optique de tarification, les tarifs calculés à partir d'une table d'expérience doivent donc être eux aussi raisonnablement prudents. Le rapport final doit :

- "valider les données utilisées et leurs sources, quelles soient internes ou externes à l'entreprise,
- vérifier les hypothèses de travail et les modalités utilisées pour construire les tables de mortalité ou les lois de maintien en incapacité de travail ou en invalidité
- s'assurer que les principes de prudence communément admis ont été respectés, eu égard aux risques induits (en particuliers stabilité des tables ou des lois de maintien)
- définir précisément les conditions d'application et de validité des éléments certifiés, les statistiques ou tableaux de bord à préparer périodiquement par l'entreprise pour permettre le suivi des résultats d'expérience"

Les tables de mortalité d'expérience doivent être vérifiées tous les ans et sont valides pour une durée maximale de 5 ans. Finalement, la certification s'impose de manière particulière à une table d'expérience définie et construite à partir d'un groupe de contrats particuliers définissant un risque ou une catégorie de risque particuliers.

Dans l'optique d'un réalisme plus poussé, nous allons tenter dans la suite d'évaluer les dérives prospectives des taux de mortalité qui s'imposent naturellement et intuitivement.

3.2 Tables du moment et tables prospectives

3.2.1 L'augmentation de la durée de vie

L'espérance de vie, outil statistique, définit la durée de vie moyenne qu'on peut espérer atteindre, à la naissance, à un moment donné et dans un pays donné. Durant le XX siècle, l'espérance de vie a presque doublé en France, puisqu'elle est passée de 40 ans en 1900 à 78 ans en 2000, comme le montre le graphique. Cependant, elle n'a pas connu une croissance continue, notamment à cause des deux guerres mondiales.

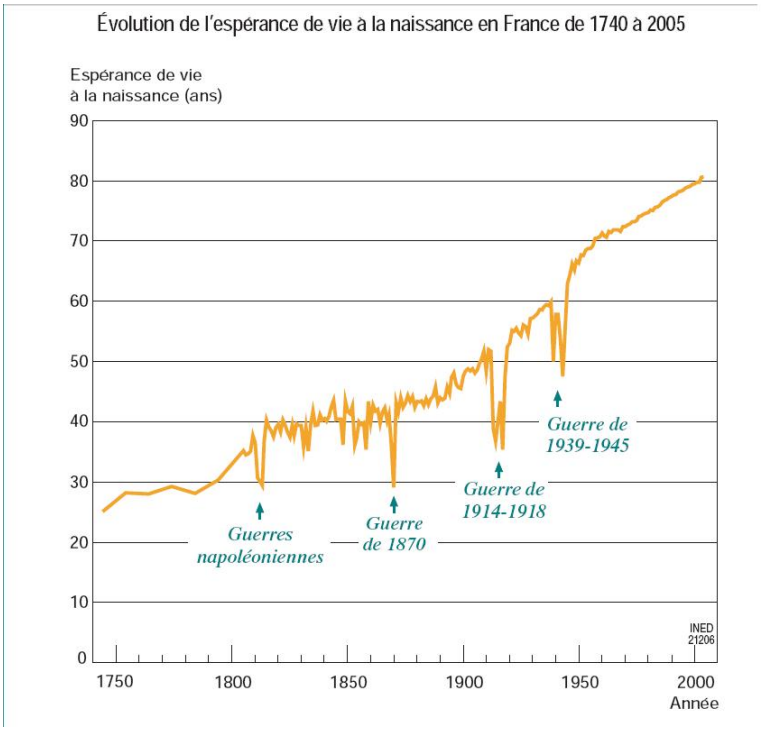


FIG. 3.2 – Evolution de l'espérance de vie en France

Cette évolution trouve son origine dans les progrès techniques, médicaux, dans les politiques de santé publique (lutte contre les mauvaises habitudes alimentaires, le tabac, l'alcool) et dans la hausse générale du niveau de vie. Il est à noter que cette tendance n'est nullement spécifique à la France, en particuliers pour les pays développés.
Ainsi, l'espérance de vie constitue une donnée en perpétuelle croissance et dont la modélisation d'évolution est difficile. Pour y faire face, l'actuaire sera amené, pour tarifer et provisionner des contrats s'étendant sur de longues durées, à utiliser des tables de mortalité prospectives.

3.2.2 Utilité des tables prospectives

L'analyse précédente peut être précisée en disant que l'augmentation de l'espérance de vie a d'abord été associée à une diminution de la mortalité infantile. Cependant, ces dernières décennies ont montré que le recul de la mortalité du troisième âge constitue le moteur de cette croissance. Ceci est d'autant plus important

à remarquer que cette catégorie de la population forme la plus grande partie de la clientèle concernant les contrats de rentes viagères.

Lors de la souscription d'un tel contrat, l'assureur propose un tarif et donc des primes qui seront fixés, figés à l'origine, alors que ces contrats se caractérisent par une durée de vie relativement longue. Reste alors à l'assureur d'évaluer au mieux l'espérance de vie de chaque souscripteur et de tenir compte à juste titre de l'évolution de cette dernière dans les années qui suivent la signature du contrat. C'est justement l'utilité des tables de mortalité prospectives.

Ces dernières se présentent comme un modèle bidimensionnel, à savoir l'âge des individus et le temps constituent les données qui interviennent. Il existe de nombreuses méthodes pour construire une table de mortalité prospective. Le lecteur pourra par exemple, pour plus de précisions techniques, se référer à la Notice de Présentation sur les tables de mortalité d'expérience sur un portefeuille de rentiers par Frédéric Planchet, produit en Avril 2007.

Concernant la réglementation, les nouvelles tables doivent être basées sur une analyse des portefeuilles d'assurés par les compagnies d'assurance. Deux tables générationnelles prospectives existent, il s'agit de TGF05 pour les femmes et TGH05 pour les hommes. Elles doivent être utilisées pour la tarification et le provisionnement des rentes viagères depuis le 1er janvier 2007. Pour plus de précisions, on pourra se rapporter à l'Article A 335-1 du Code des assurances, article modifié par l'arrêté du 08/01/2006.

Comment intégrer l'ensemble de ces informations à nos modèles de mortalité d'expérience ?

3.2.3 Construction de tables prospectives avec notre modèle

Nous allons essayer à présent d'imprimer une tendance prospective à notre table de mortalité segmentée obtenue à l'aide du modèle de Cox. Effectivement, cette dernière, que nous pouvons considérer comme valide pour l'année 2005, doit prendre en compte l'augmentation de la durée de vie au fil des années. A titre d'exemple, si nous considérons une personne de 50 ans en 2005, le taux de mortalité extrait de notre table lorsqu'il aura 60 ans est valable en 2005, mais sera probablement différent en 2015 : c'est toute la limite des tables du moment dont la pertinence se vérifie lors de leur construction puis s'affaiblit au fil du temps.

Nous proposons ici une calibration de notre table de mortalité segmentée sur les tables prospectives actuellement en vigueur, à savoir TGH05 et TGF05. Une question importante que l'on doit se poser concerne bien évidemment l'utilisation de ces tables prospectives : est-il pertinent, voire juste, d'en user puisqu'elles sont élaborées pour des produits de type rente et non décès ? Plus particulièrement, les évolutions observées sur de telles tables sont-elles du même type et du même ordre de grandeur pour notre population d'assurés ? Nous allons tenter d'y répondre, ou du moins de légitimer cette démarche

Le graphique suivant présente la mortalité issue de notre table pour les hommes non cadres (la base), les hommes cadres, et la mortalité tirée de la table TGH05 pour l'année 2002.

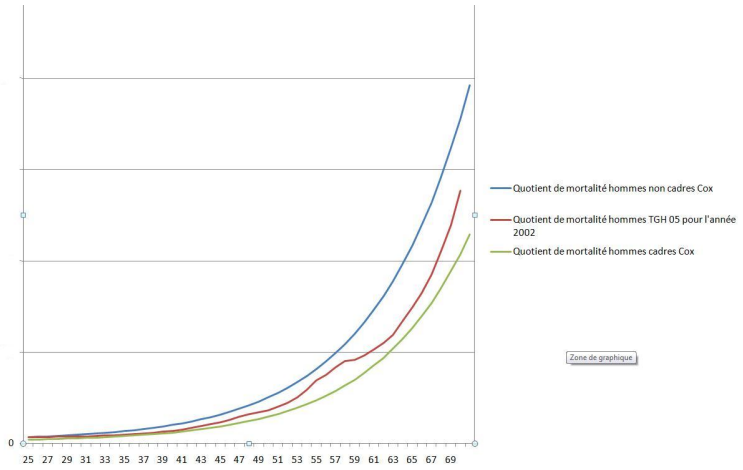


FIG. 3.3 – Comparaison de la mortalité masculine, en 2002, pour la table d'expérience et TGH05

Ce graphique, dont l'analyse peut certes paraître marginale, permet pourtant de visualiser une tendance identique pour les trois courbes, et indique que la mortalité tirée de la table TGH05 est contenue entre la mortalité des cadres et des non cadres et qu'elle est relativement proche. La même remarque peut être produite pour la mortalité féminine.

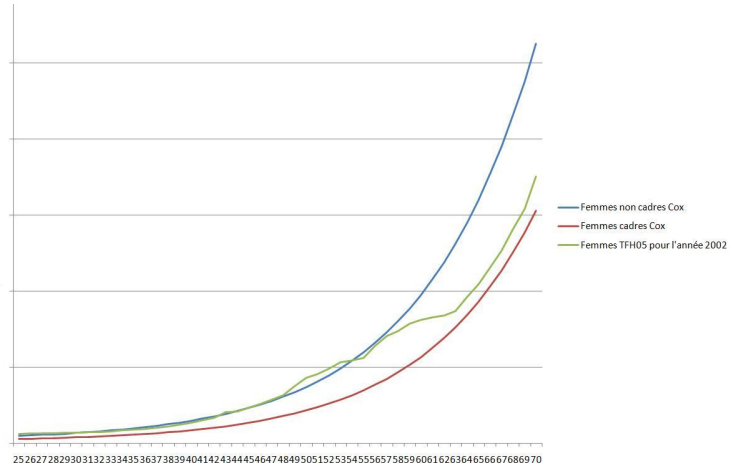


FIG. 3.4 – Comparaison de la mortalité féminine, en 2002, pour la table d'expérience et TGF05

Cette fois-ci, la mortalité féminine des non cadres correspond à celle tirée de la TGF05 en 2002 de 20 à 50 ans. En outre, une calibration prospective ne tient pas compte de la segmentation cadre/non cadre (les tables prospectives ne sont en effet segmentées que par sexe), mais ce problème semble être limité par le fait que, pour chaque sexe, la mortalité tirée des tables réglementaires prospectives en 2002 est inclue entre les mortalités des cadres et des non cadres. De plus, nous observons une sous-mortalité des catégories tracées sur ces graphiques par rapport aux taux de référence des tables réglementaires TH00-02 et TF00-02, traduction a priori de la mortalité spécifique des rentiers et de notre population d'assurés, inférieure à celle de la population générale.

Au final, le recours aux tables prospectives réglementaires peut certes paraître discutable, mais on cherche ici en priorité à se raccrocher à une référence, et cela est légitimé par le fait que notre population d'assurés présente une mortalité proche, en 2002, de celle donnée par les tables prospectives. On fait donc l'hypothèse que les tendances de long terme ne différeront que légèrement.

Comment imprimer un caractère prospectif à la table de mortalité d'expérience segmentée ? Il s'agit en particuliers de conserver deux tendances cruciales : une première spécifique à notre table bâtie à l'aide du modèle de Cox, une seconde caractéristique d'une certaine dérive prospective. Nous proposons la relation suivante :

$$q_{x+1,t_0+1} = q_{x+1} * q_{pros,x+1,t_0+1}/q_{pros,x+1,t_0}$$

Dans cette relation, q_{x,t_0} désigne le taux de mortalité à l 'âge x valable à l'année t_0 donc avec la dérive prospective, q_x désigne le taux de mortalité tirée de notre table d'expérience, q_{pros,x,t_0} désigne le taux de mortalité extrait de la table prospective réglementaire pour une personne d'âge x à l'année t_0, donc de la génération $t_0 - x$. Cette relation vérifie une certaine homogénéité, tant du point de vue de l'âge que de l'année. En réitérant le procédé, on obtient :

$$q_{x+y,t_0+y} = q_{x+y} * q_{pros,x+y,t_0+y}/q_{pros,x+y,t_0}$$

Ici, y est un entier et $x + y$ renvoie donc à l'âge dont on veut déterminer un taux de mortalité qui comporte une dérive prospective.

A titre d'exemple, voici ce que donne cette relation pour un homme non cadre de 50 ans en 2005 :

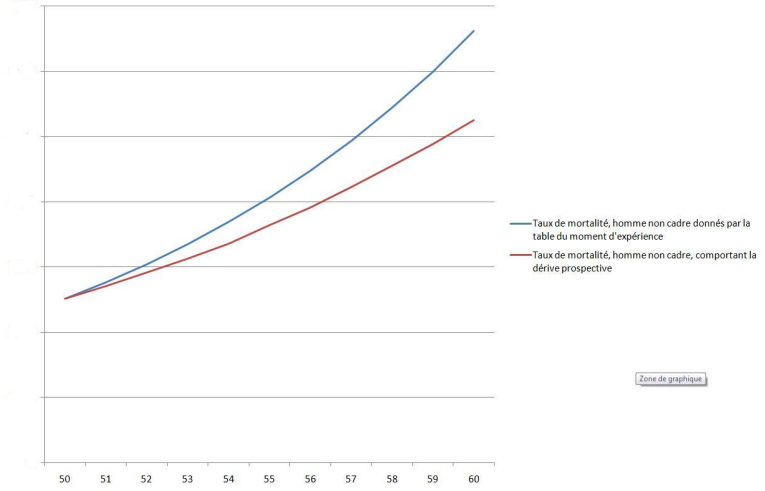


FIG. 3.5 – Dérive prospective pour un homme non cadre de 50 ans en 2005

Comme cela était prévisible, les taux de mortalité avec dérive prospective sont inférieurs aux taux du moment, et le décalage augmente logiquement au fil des ans. Théoriquement et intuitivement, les taux prospectifs d'expérience sont donc plus proches de la réalité que les taux du moment.

Comment se traduit cette dérive prospective en terme de tarification ?

3.2.4 Impact de la dérive prospective sur la tarification

Nous allons dans cette partie présenter les comptes de résultat technique de l'assureur utilisant les tables de mortalité d'expérience segmentées sur la population d'assurés et illustrer l'impact de la dérive prospective. Pour des raisons de confidentialité, l'échelle de l'axe des ordonnées du graphique sera arbitraire, il ne représentera nullement la réalité, mais nous permettra tout de même de dégager des tendances caractéristiques. Nous considérons pour la population d'assurés des contrats temporaires décès sur 10 ans avec prime payée au début de chaque année. Nous nous focaliserons donc sur une problématique de tarification plutôt que de provisionnement. D'autre part, nous ne prendrons pas en compte dans le calcul les éléments intervenant dans le compte technique, tels les produits des placements ou les frais d'acquisition et d'administration. Effectivement, l'objectif est de se focaliser sur des tendances plutôt que d'aboutir à une tarification réaliste.

Dans un premier temps, supposons que l'assureur tarifie chaque année en se basant sur les tables d'expérience segmentées estimées par le modèle de Cox. Le tarif proposé à chaque assuré augmente donc chaque année. La dérive prospective permet à l'assureur de dégager des résultats positifs tous les ans.

Ensuite, supposons que pour chaque assuré, le tarif est fixé à l'origine du contrat, l'assureur utilisant alors une moyenne des quotients de mortalité pour chaque client. Intuitivement, les résultats dégagés par l'assureur devraient être positifs durant les premières années, car pour chaque assuré, leur mortalité est alors sensiblement inférieure à la base de tarification. Puis les résultats deviennent naturellement négatifs. concrètement, chaque client paie trop au début du contrat, puis pas assez à la fin.

Le graphique suivant permet de comparer l'émergence de résultats sur les 10 ans suivant les deux tarifications :

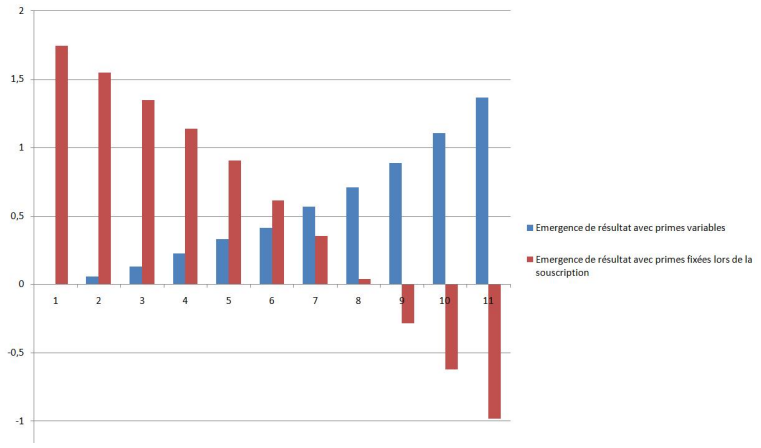


FIG. 3.6 – Emergence de résultats pour les deux tarifications

Des résultats positifs lorsque la prime varie tous les ans s'expliquent par la dérive prospective. Concernant la seconde tarification, avec prime identique durant les dix années, "'l'assuré moyen"' bénéficie d'un tarif moindre à partir de 7 ans.

Nous pouvons pousser un peu plus loin la réflexion en évoquant les possibilités de remboursement anticipé. Si un assureur choisit la seconde méthode de tarification, c'est-à-dire une prime annuelle fixe sur les 10 ans, cela permet de dégager plus de résultats voire de ne pas subir les résultats négatifs des dernières années du contrat.

Nous nous interrogeons maintenant sur le réalisme d'application de ces tables : la sélection ne parait-elle pas a priori trop forte ?

3.3 Une sélection trop forte ?

La figure 3.1 présente une segmentation très prononcée. A titre d'exemple, toutes choses égales par ailleurs, les hommes non cadres de 68 ans sont touchés par une mortalité trois fois plus élevée que les femmes cadres du même âge. La question que l'on est en droit de se poser est la suivante : un tel écart tarifaire est-il envisageable concrètement pour un assureur concernant deux individus qui ont pourtant le même âge ? Comment l'assureur peut-il diminuer de tels décalages ? Nous nous plaçons uniquement dans le cadre de contrats temporaires décès et dans une optique de tarification plutôt que de provisionnement.

Une première solution naturelle semble être un des principes de l'activité d'assurance : la mutualisation du risque. Cette dernière permet de faire supporter une partie du risque pour une personne très exposée (ici typiquement les hommes non cadres de 68 ans) par une personne moins exposée (ici typiquement une femme cadre de 68 ans). En clair, les personnes présentant le moins de risque paient un peu plus que ce que prévoient les estimations de mortalité, à la différence des personnes sous haut risque. Cela permet de réduire l'écart tarifaire conséquent. A noter que cette mutualisation particulière explicite n'est pas un mécanisme automatique ou un réflexe général de l'assureur. Il faut aussi garder à l'esprit qu'il existe deux types de mutualisation : la mutualisation dans une classe de risque homogène (on peut segmenter et faire payer le "'vrai prix"') qui est un critère technique objectivable et la revalorisation au sens du transfert de risque (mélanger des "'mauvais"' risques avec des "'bons"').

Cet écart tarifaire brut peut aussi être affecté par une "sélection a posteriori", et on peut citer à ce sujet l'exemple de la sélection médicale. On peut imaginer la situation où l'assuré est en parfait état de santé, du moins pour son âge, et il paiera alors la prime standard tirée directement de nos estimations. Si l'assuré présente des antécédents médicaux ou si son état de santé est altéré, l'assureur peut majorer la prime et dans le cas extrême, refuser de prendre l'individu en charge. Apparaît alors des majorations ou bien des abattements qui interviennent dans le calcul de la prime, tout comme les estimations de taux de mortalité, concernant les contrats temporaires décès. A titre d'exemple, le caractère fumeur/non fumeur que nous n'avons pas pris en compte dans ce mémoire peut apparaître a posteriori par une majoration de la prime estimée. Cela permet le calcul de primes plus spécifiques à l'assuré et de gommer la sélection a priori trop forte.

3.4 Conclusion

Cerner le risque auquel est soumis l'assureur et parer par là-même au problème de l'antisélection justifient la construction de tables de mortalité d'expérience segmentées. Pour palier à des problèmes d'insuffisance de données qu'entraînerait la méthode classique d'étudier des sous populations d'assurés en fonction de classes de risque déterminées a priori, ce mémoire avait pour objectif de présenter deux méthodes statistiques évitant cette segmentation préalable : la modélisation de la survie par un hasard proportionnel, dite de Cox, et celle d'un hasard additif, dite de Aalen.

Tout en ayant présenté théoriquement ces méthodes, nous avons mis en évidence les différentes étapes d'une construction de tables segmentées à partir d'elles : une première phase de préparation et de nettoyage des données conditionne des résultats précis et de bonne qualité. Il a s'agit en effet de procéder à des statistiques descriptives de la population d'assurés, de vérifier l'homogénéité des caractéristiques de chaque individu et la pertinence de ces dernières pour la mortalité. Puis, nous avons déterminé à l'aide de nos deux modèles la significavité et l'impact des variables tarifaires ainsi sélectionnées , pour en déduire des classes de risques sur lesquelles nous pouvions déterminer des taux de mortalité segmentés, que nous avons enfin ajustés et confrontés à la réalité. Lors de ces étapes, nous avons mis en évidence la difficulté d'utiliser le modèle de Aalen, à la fois par le manque de souplesse du peu de logiciels le mettant en oeuvre, mais aussi en raison de sa grande sensibilité à la taille des classes de risque utilisées, puisqu'il permet de faire, classe par classe, des estimations directes de survie. La modélisation proportionnelle de Cox, qui se contente d'estimer un "écart" entre les classes de risque qu'elle distingue, est moins sensible à la taille des échantillons et se révèle plus pratique en vue de s'affranchir d'une segmentation préalable des données. L'objectif final a donc été atteint grâce à ce modèle : réussir à obtenir des tables de mortalité segmentées conformes aux observations à l'aide d'une modélisation intégrant l'hétérogénéité des individus.

Rappelons par ailleurs que les données qui nous ont servi de base de travail forment un échantillon volontairement non représentatif du portefeuille temporaire décès de la société Axa. C'est donc bien sur la méthode, et non sur les résultats quantitatifs, que nous avons souhaité insisté.

Nous avons par la suite réalisé une application actuarielle concrète de nos tables d'expérience en les comparant d'abord aux tables réglementaires citées par l'Article A335-1 du Code des Assurances. Ces dernières présentent une surmortalité par rapport aux individus ayant servi pour nos estimations. D'autre part, dans une optique de tarification plutôt que de provisionnement, nous nous sommes penchés sur la dérive prospective de ces taux de mortalité due à l'augmentation de l'espérance de vie. Ceci nous a permis de tracer des profils d'émergence de résultats selon deux procédés distincts de tarification. Enfin, nous avons discuté quant à la force de cette sélection en précisant que d'autres critères interviennent a posteriori dans la tarification, ce qui permet peut-être de limiter les écarts de tarifs pour des individus de classes différentes.

A travers la méthodologie retenue, ce mémoire présente finalement un exemple d'aller-retour systématique entre la théorie et les contraintes de la pratique, et ex-

pose, par là-même, les difficultés qui surgissent lors de la construction d'un modèle interne pour les compagnies d'assurance. En effet, bâtir de tels modèles demande un travail de fond important et des capacités informatiques et techniques exigeantes. La construction de tables de mortalité d'expérience segmentées n'en constitue qu'une faible partie.

Bibliographie

[1] Aalen, O. O. [1980] *A model for non-parametric regression analysis of counting processes*. Mathematical Statistics and Probability. Lecture Notes in Statist. 2 1-25. Springer, New York.

[2] Andersen, P.K. et Gill, R.D. [1982] *Cox's regression model for counting processes : a large sample study*, Annals of Statistics, 10, 1100-1120

[3] Cao, H. [2005] *A Comparison Between the Additive and Multiplicative risk Models*, mémoire de maîtrise en statistique, Faculté des études supérieures de l'Université de Laval (Québec)

[4] 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", Préversion article pour le *Bulletin Français d'Actuariat à paraître en octobre 2008*

[5] Cox, D.R. et Oakes, D. [1984] *Analysis of survival Data*, Chapman et Hall

[6] Fourgeaud, C. et Gourieroux, P. et Pradel, J. [1987] *Heterogeneity and hazard dominance in duration data models*, CEPREMAP

[7] Gill, R.D. [1980] "Censoring and stochastic Integrals" *Mathematical Centre Tracts*, n°124, Amsterdam : Mathematische Centrum

[8] Kaplan, E.L. et Meier, P. [1958] "Non-Parametric estimation from incomplete observation" *Journal of the American Statistical Association* 53, 457-481

[9] Kupets, O. [2005] *Determinants of unemployment duration in Ukraine*

[10] Martinussen, T. et Scheike, T. H. [2006] *Dynamic Regression Models for Survival Data* New York : Springer-Verlag

[11] Moreau, A. [1989] *Econométrie des modèles de durée*, Note de synthèse INSEE

[12] Planchet, F. et Thérond, P. [2006] *Modèles de durée, applications actuarielles*, Economica

Annexe A

Code R

```
##librairie modèles de durée
library(survival)
library(timereg)


##extraction du fichier données sous R
donnees=read.csv2("Base25-70ans.csv",header=TRUE,sep=";")
t<-donnees



##Traitement des dates

##fonction transformant les dates des données en objet de classes
  "date" pour R
TraiteDate<-function(Table,NomChamp){

x0=strptime(Table[,NomChamp],"%d/%m/%Y")
Table=data.frame(Table,x0)
Table[,NomChamp]<-NULL
if(names(Table)[length(names(Table))]=="x0"){
names(Table)[length(names(Table))]=NomChamp
}
else{}
Table
}

##application aux dates utiles de la table t
t=TraiteDate(t,"date_fin")
t=TraiteDate(t,"date_début")
  t=TraiteDate(t,"date_naissance")

##Création de variables d'age à l'entrée et à la sortie en unité "jour"
t$AgeEntreeJours=difftime(t$date_début,t$date_naissance,"","days")
t$AgeSortieJours=difftime(t$date_fin,t$date_naissance,"","days")
```

```
##Création des variables de censure
t$non_censure=1-t$Censure

##inversion variable ancienneté de contrat: 1 = contrat jeune,
## 0 = contrat ancien
t$anciennete_ref=1-t$ancienneté.du.contrat_2ans


##test d'indépendance du chi deux entre variables
chisq.test(t$Classe_CSP,t$Sexebin)

##Application modèles de durée

##Kaplan Meier

wkm_tot=survfit(Surv(AgeEntreeJours,AgeSortieJours,non_censure,
type="counting"),data=t,type="kaplan-meier")
plot(wkm_tot, mark.time=FALSE, xscale=365.25, ymin=0.8,
xlab="Temps (années)",ylab="Survie", main="Survie Globale")

##exemple de Kaplan Meier sur une sous population

t_hnc<-subset(t, t$Sexebin==0 & t$CSP3==0)
length(t_hnc[,1])
wkm_hnc=survfit(Surv(AgeEntreeJours,AgeSortieJours,non_censure,
type="counting"),data=t_hnc,type="kaplan-meier")
plot(wkm_hnc, mark.time=FALSE, xscale=365.25, ymin=0.8,
xlab="Temps (années)",ylab="Survie", main="Survie Hommes Non Cadres")
Sur_hnc<-NULL
Sur_hnc$surv<-as.data.frame(wkm_hnc[["surv"]])
Sur_hnc$time<-as.data.frame(wkm_hnc[["time"]])
write.table(Sur_hnc,"KM_survie_HommesNonCadres.csv",sep=";",dec=",")

##Cox

wcox=coxph(Surv(AgeEntreeJours,AgeSortieJours,non_censure,
type="counting"~Sexebin+CSP1+CSP2+CSP3+CSP5+CSP6+capital_2+
capital_3+capital_4,data=t,singular.ok=TRUE)
summary(wcox)

wcox=coxph(Surv(AgeEntreeJours,AgeSortieJours,non_censure,
type="counting")~Sexebin+CSP3,data=t,singular.ok=TRUE)
summary(wcox)

##analyse des résidus de Schoenfeld
cox.zph(wcox)
plot(cox.zph(wcox),var=1)
```

```
##aalen

##marqueur des individus
t$identity<-1:length(t[,1])

##gestion des ties: la fonction aalen ne prends pas en compte
##les sorties simultanées: on perturbe donc les sorties par
##un bruit blanc
t$AgeSortieJoursBis<-t$AgeSortieJours+runif(length(t$AgeSortieJours),-1,1)

##augmentation de la taille mémoire de R pour pouvoir gérer
##45000 individus avec la fonction aalen
memory.limit(size=2500)

waalen<-aalen(Surv(AgeEntreeJours,AgeSortieJoursBis,non_censure,
type="counting")~Sexebin+CSP1+CSP2+CSP3+CSP6+CSP5+capital_2+
capital_3+capital_4,data=t,id=t$identity)
summary(waalen)

waalen<-aalen(Surv(AgeEntreeJours,AgeSortieJoursBis,non_censure,
type="counting")~Sexebin+CSP3,data=t,id=t$identity)
summary(waalen)
write.table(waalen$cum,"Aalen_CoeffCumvrai.csv",sep=";",dec=",")
```

Annexe B

Article A 335-1 du Code des Assurances

Les tarifs pratiqués par les entreprises d'assurance sur la vie et de capitalisation comprennent la rémunération de l'entreprise et sont établis d'après les éléments suivants :

1° Un taux d'intérêt technique fixé dans les conditions prévues à l'article A. 132-1.

2° Une des tables suivantes :

a) Tables homologuées par arrêté du ministre de l'économie et des finances, établies par sexe, sur la base de populations d'assurés pour les contrats de rente viagère, et sur la base de données publiées par l'Institut national de la statistique et des études économiques pour les autres contrats ;

b) Tables établies ou non par sexe par l'entreprise d'assurance et certifiées par un actuaire indépendant de cette entreprise, agréé à cet effet par l'une des associations d'actuaires reconnues par l'autorité mentionnée à l'article L. 310-12.

Les tables mentionnées au b sont établies d'après des données d'expérience de l'entreprise d'assurance, ou des données d'expérience démographiquement équivalentes.

Lorsque les tarifs sont établis d'après des tables mentionnées au a, et dès lors qu'est retenue une table unique pour tous les assurés, celle-ci correspond à la table appropriée conduisant au tarif le plus prudent.

Pour les contrats en cas de vie autres que les contrats de rente viagère, les tables mentionnées au a sont utilisées en corrigeant l'âge de l'assuré conformément aux décalages d'âge ci-annexés. (Annexes non reproduites, voir le fac-similé).

Pour les contrats de rentes viagères, en ce compris celles revêtant un caractère temporaire, et à l'exception des contrats relevant du chapitre III du titre IV du livre Ier, le tarif déterminé en utilisant les tables mentionnées au b ne peut être inférieur à

celui qui résulterait de l'utilisation des tables appropriées mentionnées au a.

Pour les contrats collectifs en cas de décès résiliables annuellement, le tarif peut être établi d'après les tables mentionnées au a avec une méthode forfaitaire si celle-ci est justifiable.