Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Madame Clémentine Espitalié

Titre du mémoire : Construction d'une segmentation et tarification d'une garantie incapacité en prévoyance individuelle

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

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

Membres présents du jury de la filière : Signature :	Entreprise : Nom : AXA France Signature :
	Directeur de mémoire en entreprise Nom : Hugo Bernard-Brunel Signature :
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise : Signature du candidat :

RÉSUMÉ

Afin de se prémunir contre le risque d'arrêt de travail, de nombreux individus souscrivent à un contrat de prévoyance individuelle. Ceci est d'autant plus vrai pour les travailleurs non salariés qui sont souvent plus sensibilisés aux problématiques de prévoyance, dans la mesure où le risque personnel met directement en péril leurs revenus. L'étude réalisée dans ce mémoire portant sur un produit destiné au marché des travailleurs non salariés, constitue donc un véritable enjeu pour Axa France. Afin de rester compétitif sur ce marché, il est primordial pour chaque assureur de trouver une segmentation tarifaire la plus adaptée possible à son portefeuille et capable de capter la complexité du risque incapacité. Pour se faire, des modèles de clustering ainsi que la construction de lois d'entrée en incapacité peuvent être utilisés afin de modéliser finement le risque incapacité. Une fois cette maîtrise du risque acquise, de nouvelles classes tarifaires reflétant au mieux ce dernier, pourront alors être définies.

Mots clés : *Prévoyance individuelle, incapacité, tarification, segmentation tarifaire, modélisation de l'incidence, clustering, K-means, GLM poisson, lissage par splines*

ABSTRACT

In order to protect themselves against the risk of work disability, many individuals underwrite an individual provident contract. This is especially true for self-employed workers, who are often more aware of provident issues, since personal risk directly affects their income. The study carried out in this thesis on a product aimed at the market of self-employed workers is therefore a real challenge for Axa France. In order to remain competitive on this market, it is essential for each insurer to find a price segmentation that is as adapted as possible to its portfolio and able to capture the complexity of disability risk. To do so, clustering models and the construction of disability entry distributions can be used to analyze the risk in detail. Once the risk has been mastered, new tariff classes that best reflect the risk can be defined.

Key words : *health insurance, disability risk, pricing, clustering, K-means, GLM poisson, smoothing splines*

Table des matières

REMERCIEMENTS

Au préalable, je souhaiterais remercier l'ensemble des personnes qui ont contribué au bon déroulement de mon année d'alternance et de mon mémoire.

Plus particulièrement, je tiens à remercier et à témoigner ma reconnaissance à Hugo Bernard-Brunel, mon responsable d'alternance et tuteur de mémoire, pour ses conseils et son engagement dans le suivi de mon mémoire. Je le remercie également pour son implication, ses conseils avisés et sa disponibilité.

Je souhaite également remercier Valérie Gabot, responsable technique Prévoyance et Dépendance Individuelles d'Axa France, pour l'opportunité qui m'a été donnée de travailler dans une équipe qui s'efforce à garantir l'intérêt et la diversité des sujets de travail et de recherche.

J'adresse mes chaleureux remerciements aux membres de la direction Métier Prévoyance et Dépendance Individuelles pour leur accueil et pour l'expérience enrichissante et pleine d'intérêt qu'ils m'ont fait vivre.

Je souhaite aussi remercier mon tuteur académique, Guillaume Biessy, pour m'avoir guidée dans l'élaboration de ce mémoire grâce à ses conseils techniques.

Enfin, j'adresse un dernier remerciement à ma famille ainsi qu'à Marceau, qui m'ont toujours soutenue durant mes études.

NOTE DE SYNTHÈSE

Connaitre et maitriser son risque est essentiel pour toute activité d'assurance. Le risque d'incapacité temporaire totale de travail est un risque particulièrement différent des autres risques de prévoyance, comme le décès ou l'invalidité, puisqu'il s'agit d'un risque multiple et propre à la période d'activité professionnelle des assurés. L'étude réalisée dans ce mémoire porte sur un produit destiné au marché des travailleurs non salariés (TNS), elle constitue donc un véritable enjeu pour Axa France. En effet, les travailleurs non salariés sont souvent plus sensibilisés aux problématiques de prévoyance, dans la mesure où le risque personnel met directement en péril leurs revenus. De plus, ces derniers sont moins bien couverts par le régime obligatoire, que les salariés, notamment en ce qui concerne les risques décès, incapacité et invalidité.

Le risque d'incapacité peut être subdivisé en deux : le risque relatif à la fréquence des arrêts de travail et le risque relatif à leurs durées. Seule la fréquence des arrêts de travail, appelée incidence en incapacité, sera étudié ici. Ce mémoire a ainsi pour but de construire un tarif plus adapté et plus compétitif sur le risque arrêt de travail. L'incapacité étant un risque particulier, l'enjeu de ce mémoire sera de trouver une segmentation tarifaire capable de capter la complexité de ce risque. En effet, les principales causes d'entrée en incapacité étant la maladie et l'accident, les facteurs expliquant la fréquence des arrêts de travail sont différents selon la cause considérée.

Les données et leur analyse

Afin de mieux cerner les facteurs pouvant expliquer l'incidence en incapacité pour l'accident et pour la maladie, ce mémoire débutera par une étude de statistiques descriptives.

Les principaux points ressortant de cette analyse descriptive sont les suivants :

- L'incidence pour la cause accident semble plutôt stable au cours du temps, et est constamment inférieure au taux d'incidence pour la cause maladie ;
- Le taux d'incidence pour la cause maladie est en constante augmentation sur la période d'observation ;
- La profession semble être le facteur le plus discriminant pouvant expliquer l'incidence en accident ;
- Avant 40 ans, le sexe semble être le principal facteur explicatif pour la cause maladie. Après 40 ans, l'âge semble être le facteur le plus discriminant.

L'objectif de ce mémoire sera d'obtenir une segmentation tarifaire plus fine, plus juste et plus adaptée au portefeuille actuel d'Axa France. Les facteurs ayant un impact significatif sur les taux d'incidence, étant différents pour la cause accident et la cause maladie, la suite de l'étude sera divisée en deux parties, correspondant à une approche propre à chacune des deux causes.

Modélisation de l'incapacité pour la cause accident

L'objectif est ici de confirmer par des modèles, les hypothèses faites suite à l'étude de statistiques descriptives. Les effets des différents paramètres ayant un impact notable sur l'incidence en incapacité seront étudiés à l'aide d'un modèle linéaire généralisé (GLM). La variable CSP est une variable à 230 modalités, dont certaines étant très peu exposées. Afin d'étudier sa significativité à l'aide d'un GLM, des groupes de CSP homogènes seront construits à l'aide d'un algorithme de *clustering*.

Avant de réaliser le *clustering* des différentes CSP, il est important de préparer les données. Pour cela, il a été décidé de retenir deux variables pour chacune des CSP :

- L'incidence, sur la période d'observation, pour la cause accident ;
- La proportion des sinistres dus à un accident, sur l'ensemble des sinistres ayant eu lieu durant la période d'observation, pour la CSP étudiée.

Un algorithme K-means, pondéré par le poids des professions dans le portefeuille, a été utilisé, afin de partitionner les données en K *clusters* distincts, et d'obtenir des groupes homogènes.

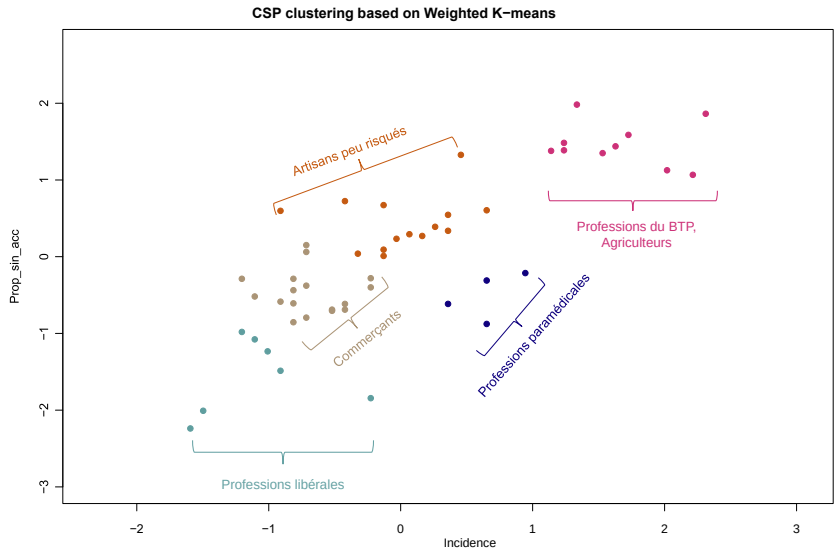

FIGURE 0.1 – Résultats obtenus suite à l'exécution de l'algorithme K-means

Sur la Figure 0.1, représentant les résultats du partionnement, les points de couleur rose en haut à droite, correspondent aux professions avec une incidence en incapacité pour la cause accident, et

une proportion de sinistres accidents, élevées. Il s'agit de professions que nous considérons comme risquées, par exemple, les professions du BTP.

Une fois les CSP du portefeuille regroupées en cinq *clusters*, un GLM de Poisson et une régression de Lasso permettent de démontrer que, comme conjecturé lors de l'analyse descriptive, le facteur le plus discriminant pour expliquer l'incapacité accident est la profession exercée par l'individu. L'objectif de ces modèles est double, puisqu'ils permettent également de valider les *clusters* de CSP, obtenus par l'algorithme *K-means*.

Les principales informations à retenir du GLM Poisson et de la régression de Lasso sont les suivantes :

- La variable la plus importante pour expliquer l'incidence en incapacité pour la cause accident est le *cluster* de CSP auquel appartient l'individu ;
- L'âge n'est pas une variable significative.

Ainsi, en ce qui concerne l'incapacité pour la cause accident, les hypothèses faites suite à la partie d'analyse descriptive ont pu être validées, et cinq groupes de CSP, homogènes en termes de risque accident ont été construits.

Cependant, les facteurs expliquant l'incapacité pour la cause accident et pour la cause maladie étant différents, il n'est pas possible se contenter d'utiliser ces cinq groupes de CSP obtenus en tant que nouvelles classes tarifaires. En effet, procéder ainsi reviendrait à reproduire la segmentation actuellement en vigueur, à savoir une segmentation de CSP par risque accidentogène croissant.

Il est donc nécessaire d'étudier le risque incapacité pour la cause maladie dans une partie qui lui est propre, avant d'aboutir aux classes tarifaires finales.

Modélisation de l'incapacité pour la cause maladie

L'étude de statistiques descriptives a montré que l'incidence en incapacité, pour la cause maladie, pouvait être expliquée par deux facteurs : l'âge et le sexe. Dans cette partie, l'incidence en incapacité pour la cause maladie sera modélisée de façon traditionnelle, à partir de lois segmentées par sexe et par âge. Puis, une segmentation des différentes professions, adaptée au risque incapacité pour la cause maladie sera réalisée.

Le modèle de tarification retenu, pour tarifier notre garantie IJ toute cause, étant de type fréquence $\times$ coût, il est donc nécessaire d'avoir des lois, permettant d'obtenir la probabilité pour un assuré d'âge x, d'entrer en incapacité. Afin de construire de telles lois, il est donc indispensable d'avoir un estimateur quantifiant le nombre moyen de sinistres par assuré d'âge x. La loi traditionnellement utilisée pour quantifier ce type d'évènement est une loi de Poisson. Des taux bruts d'entrée en incapacité, sont ainsi estimés chez les deux sexes séparément, pour les franchises 15 et 30 jours.

Cependant, les lois brutes précédemment obtenues nécessitent d'être lissées avant de pouvoir être exploitées. En effet, celles-ci présentent des irrégularités sur certains âges, ainsi que certaines valeurs extrêmes, parfois dues à trop peu de données sur l'âge en question. Un lissage par *splines* a donc été réalisé.

La construction des lois d'entrée en incapacité permet de confirmer que le sexe et l'âge ont tous les deux un impact considérable sur l'incidence en incapacité pour la cause maladie, comme le montrent les figures ci-dessous :

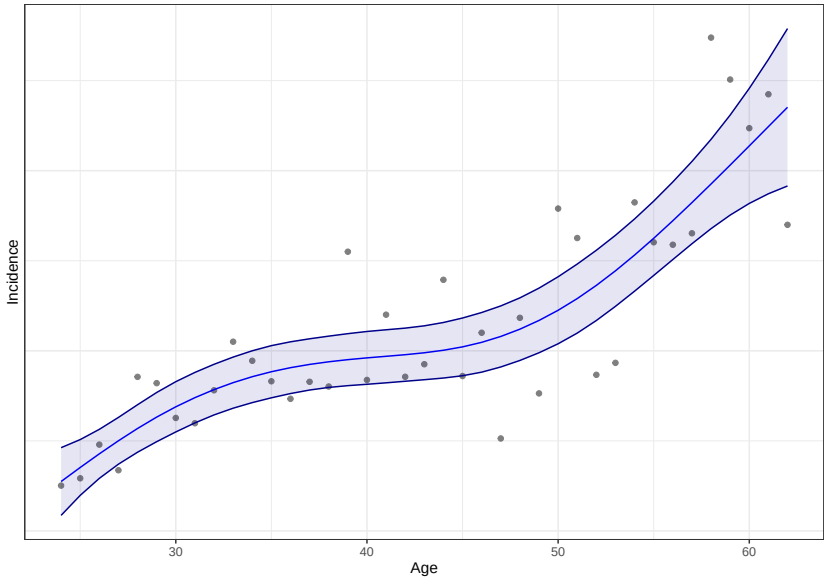

FIGURE 0.2 – Loi d'incidence en incapacité des femmes pour la franchise 30 jours

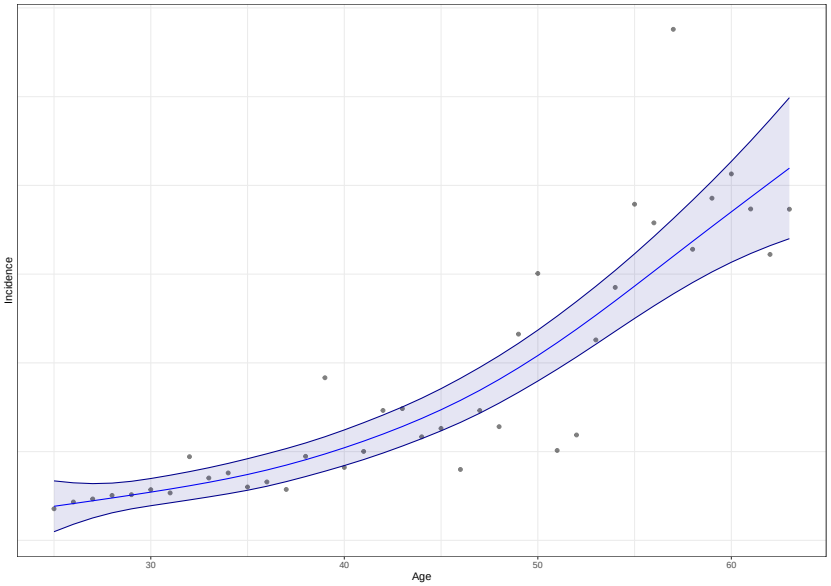

FIGURE 0.3 – Loi d'incidence en incapacité des hommes pour la franchise 30 jours

Ainsi, afin d'obtenir des groupes homogènes en termes de risque incapacité sur la cause maladie, il est nécessaire d'avoir des groupes homogènes en termes de répartition hommes/femmes. En effet,

la dimension âge est déjà prise en compte dans le tarif actuel, puisque ce dernier est différent selon l'âge de l'assuré.

Une autre dimension qu'il est nécessaire d'intégrer aux groupes de professions, construits sur la partie maladie, est la proportion de sinistres maladie au sein de chaque profession. Cependant, il est important de noter que les groupes obtenus sur la partie maladie n'ont pas du tout le même objectif que ceux construits sur la partie accident. En effet, pour la cause accident, l'objectif était d'expliquer l'incidence, et de démontrer que la profession était le principal facteur pouvant expliquer un taux d'incidence différent entre deux individus. En revanche, sur la maladie, les facteurs explicatifs sont déjà identifiés et confirmés par des lois, le but est donc de créer des groupes homogènes pour ces facteurs.

Pour créer ces groupes, l'algorithme *K-means* pondéré sera de nouveau utilisé et sera exécuté sur les mêmes CSP que celles sélectionnées sur la partie accident.

Les résultats obtenus sont représentés dans la figure 0.4.

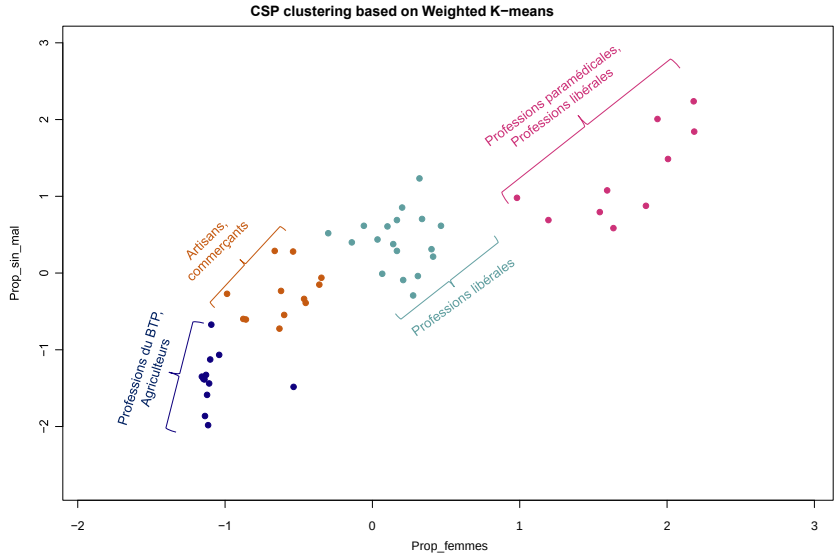

FIGURE 0.4 – Clustering des CSP pour la cause maladie

Des groupes de CSP, ont donc été construits indépendamment pour le risque accident et pour le risque maladie. Cependant le tarif de l'IJ faisait l'objet d'une refonte tarifaire, est une addition de la prime pure accident et de la prime pure maladie. Il est donc nécessaire de trouver comment concilier ces groupes de professions, construits indépendamment sur ces deux risques, afin de pouvoir proposer un tarif adapté et cohérent à chaque profession, à la fois pour la cause accident et la cause maladie.

Elaboration des nouvelles classes tarifaires

Les groupes de professions, obtenus pour la cause accident et la cause maladie ont été construits séparément, et sur des critères différents, mais propre à chacune des deux causes. Afin de traiter conjointement ces deux causes au sein d'une même classe tarifaire, il a été choisi de partir des

groupes de professions obtenus pour la cause maladie, et de regarder, pour chacune des professions de chaque groupe, à quel cluster accident celle-ci appartenait. En procédant de la sorte, onze classes tarifaires, au sein desquelles les professions sont similaires sur la dimension accident, et la dimension maladie du risque incapacité, ont pu être élaborées.

Conclusion

Dans ce mémoire, une démarche basée sur la modélisation de l'incidence en incapacité a été menée, dans le cadre de la refonte tarifaire d'un produit de prévoyance. L'étude réalisée aura permis la construction de onze classes tarifaires.

En vue d'approfondir l'étude réalisée lors de ce mémoire, les pistes de perspectives, et axes d'amélioration suivants seraient envisageables :

- Mener une étude de rentabilité afin de confronter le tarif actuel et celui obtenu avec la nouvelle segmentation tarifaire proposée ;
- Trouver une façon d'enrichir les classes tarifaires construites sur 52 CSP, avec les CSP restantes dans notre portefeuille.

SUMMARY

Knowing and controlling its risk is essential for any insurance activity. The risk of temporary total disability is a risk that is particularly different from other insurance risks, such as death, since it is a multiple risk specific to the period of professional activity of the insured. The study carried out in this paper concerns a product aimed at the market of self-employed workers, and is therefore a real challenge for Axa France. Indeed, self-employed workers are often more aware of provident issues, insofar as the personal risk directly affects their income. Moreover, the latter are less well covered by the public health system than employees, particularly as regards the risks of death or disability.

The risk of disability to work can be subdivided into two : the risk relating to the frequency of work interruptions and the risk relating to their duration. Only the frequency of work interruptions, known as the incidence of disability to work, will be studied here. Thus, the aim of this paper is to build a more adapted and competitive price on the risk of work disability. As disability is a specific risk, the challenge of this thesis will be to find a price segmentation capable of capturing the complexity of this risk. Indeed, the main causes of incapacity being illness and accident, the factors explaining the frequency of work interruptions are different depending on the cause considered.

The data and its analysis

In order to better understand the factors that may explain the incidence of disability for accident and illness, this paper will begin with a study of descriptive statistics.

The main points emerging from this descriptive analysis are :

- The incidence for the accident cause seems rather stable over time, and is consistently lower than the incidence rate for the disease cause ;
- The incidence rate for the disease cause is constantly increasing over the observation period ;
- Profession seems to be the most discriminating factor that can explain the incidence in accidents ;
- Before the age of 40, gender seems to be the main explanatory factor for the disease cause. After the age of 40, age seems to be the most discriminating factor.

The objective of this study will be to obtain a finer, fairer and more appropriate tariff segmentation for Axa France's current portfolio. As the factors having a significant impact on the incidence

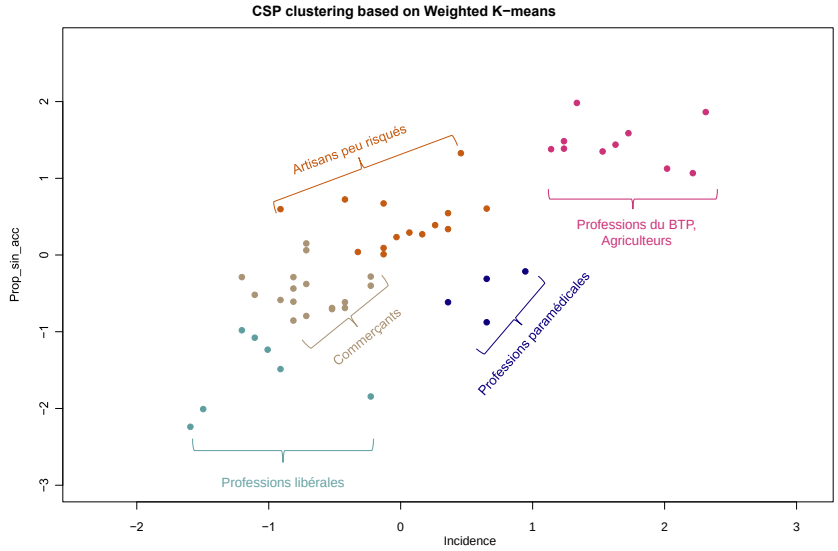

FIGURE 0.5 – Results obtained following the execution of the K-means algorithm

rates are different for accident and illness causes, the rest of the study will be divided into two parts, corresponding to a specific approach for each of the two causes.

As the factors having a significant impact on the incidence rates are different for the accident cause and the disease cause, the rest of the study will be divided into two parts, corresponding to an approach specific to each of the two causes.

Modelling disability for the accident cause

The objective here is to confirm, through models, the hypotheses made following the study of descriptive statistics. The effects of the different parameters having a significant impact on the incidence of disability will be studied using a generalized linear model. The profession variable is a 230-modality variable, some of which have very little exposure. In order to study its significance using a GLM, groups of homogeneous professions will be constructed using a clustering algorithm.

Before clustering the different professions, it is important to prepare the data. For this, it was decided to retain two variables for each of the professions :

- The incidence, over the observation period, for the accident cause ;
- The proportion of claims due to an accident, out of all claims that occurred during the observation period, for the profession studied.

A K-means algorithm, weighted by the importance of the professions in the portfolio, was used in order to partition the data into K distinct clusters, and to obtain homogeneous groups.

In the figure 0.5 showing the results of the clustering, the pink dots in the upper right-hand corner correspond to professions with a high incidence of disability due to accidents, and a high proportion of accident claims. These are professions that we consider to be risky, for example, construction professions.

Once the professions in the portfolio have been grouped into five clusters, a Poisson GLM and a Lasso regression make it possible to demonstrate that, as conjectured during the descriptive analysis, the most discriminating factor in explaining accidental disability is the the profession exercised. The objective of these models is twofold, since they also allow us to validate the professional clusters obtained by the K-means algorithm.

The main takeaways from the Poisson GLM and Lasso regression are the following :

- The most important variable to explain the incidence in disability for the accident cause is the profession exercised ;
- Age is not a significant variable.

Thus, as far as disability due to accidents is concerned, the hypotheses made following the descriptive analysis were validated, and five groups of homogeneous professions in terms of accident risk were constructed.

However, since the factors explaining disability for accident and illness causes are different, it is not possible to simply use these five profession clusters obtained as new price classes. In fact, we would be reproducing the segmentation currently in force, i.e. a segmentation of professions by increasing accident risk.

It is therefore necessary to study the disability risk for the disease cause in its own part, before arriving at the final price classes.

Modeling disability for the disease cause

The study of descriptive statistics showed that the incidence of disability for disease causes could be explained by two factors : age and sex. In this section, the incidence of disability due to disease will be modelled in the traditional way, using distributions segmented by sex and age. Then, a segmentation of the different professions, adapted to the risk of disability due to disease, will be performed.

Since the pricing model used to rate our all-cause daily benefit is of the frequency $\times$ cost type, it is necessary to have distributions that allow us to obtain the probability of an insured person of age x becoming disabled. In order to construct such distributions, it is therefore essential to have an estimator quantifying the average number of claims per insured of age x. The distribution traditionally used to quantify this type of event is a Poisson distribution. Crude rates of disability entry are thus estimated for both sexes separately, for the 15 and 30 day deductibles.

However, the raw distributions obtained above need to be smoothed before they can be used. Indeed, they present irregularities on some ages, as well as some extreme values, sometimes due to too little data on the age in question. A smoothing by splines was thus carried out.

The construction of disability incidence distributions confirms that both gender and age have a considerable impact on the incidence of disability to work due to illness, as shown in the figures below :

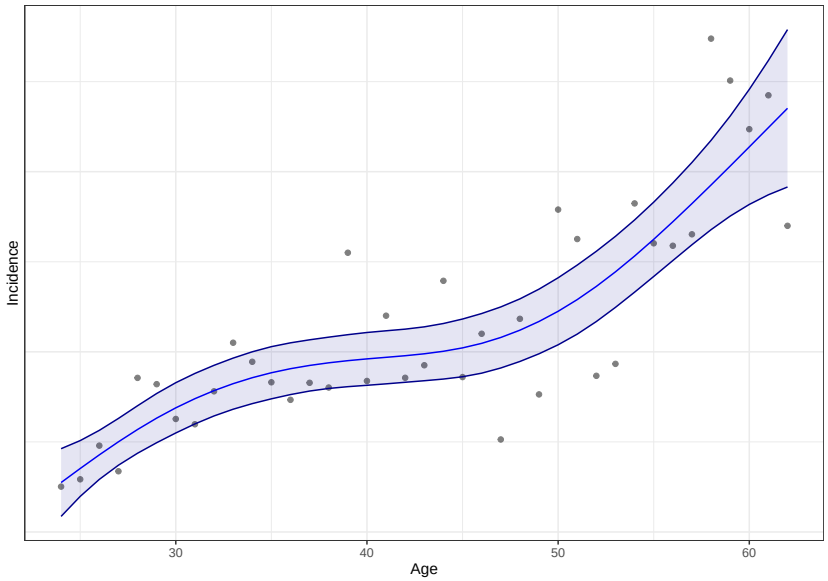

FIGURE 0.6 – Incidence of disability distribution for women for the 30-day deductible

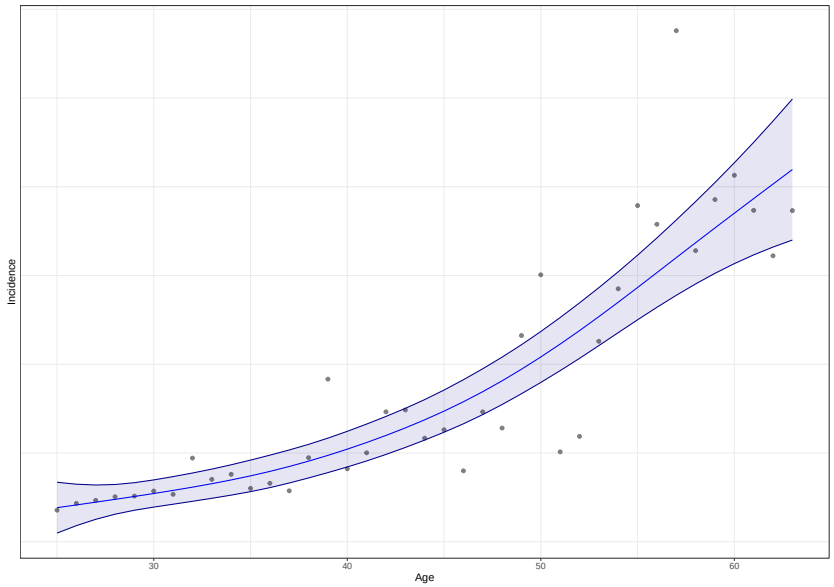

FIGURE 0.7 – Incidence of disability distribution for men for the 30-day deductible

Thus, in order to obtain homogeneous groups in terms of risk of disability due to illness, it is necessary to have homogeneous groups in terms of male/female distribution. In fact, the age

dimension is already taken into account in the current price, since it is different according to the age of the insured.

Another dimension that needs to be incorporated into the professions groups constructed on the disease part is the proportion of disease claims within each profession. However, it is important to note that the groups obtained on the disease part do not have at all the same objective as those constructed on the accident part. Indeed, for the accident part, the objective was to explain the incidence, and to demonstrate that the profession was the main factor that could explain a different incidence rate between two individuals. On the other hand, for the disease, the explanatory factors have already been identified and confirmed by distributions, so the aim is to create homogeneous groups for these factors.

To create these groups, the weighted K-means algorithm will be used again and will be run on the same professions as those selected on the accident part.

The results obtained are shown in the figure 0.8.

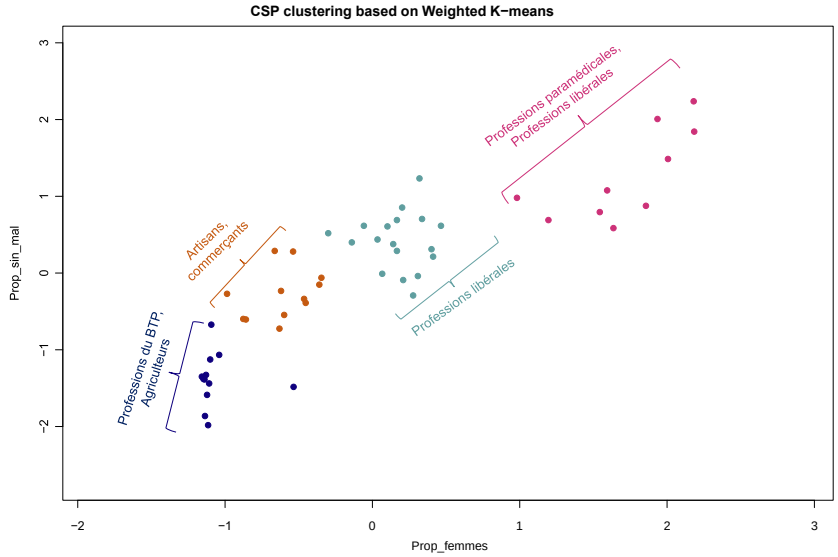

FIGURE 0.8 – Clustering of professions for the disease cause

Professions groups were therefore constructed independently for accident and sickness risk. However, the daily benefit premium, which is the subject of a price recast, is an addition of the pure accident premium and the pure illness premium. It is therefore necessary to find a way to reconcile these groups of professions, built independently on these two risks, in order to be able to propose a price adapted to each profession, for both the accident and the disease cause.

Development of new price classes

The professions groups obtained for the accident and disease causes were constructed separately, and on different criteria, but specific to each of the two causes. In order to treat these two causes jointly within the same price class, it was decided to start from the professional groups obtained for the disease cause, and to look, for each of the professions in each group, at which accident

cluster it belonged. By proceeding in this way, eleven price classes, within which the professions are similar on the accident dimension, and the disease dimension of the disability risk, could be elaborated.

Conclusion

In this paper, an approach based on the modelling of the incidence of disability was carried out, in the context of the pricing reform of a disability insurance product. The study led to the construction of eleven price classes.

In order to deepen the study carried out during this thesis, the following perspectives and axes of improvement could be considered :

- Conduct a profitability study to compare the current rate with the one obtained with the proposed new rate segmentation ;
- Find a way to enrich the classes built on 52 professions, with the remaining professions in our portfolio.

INTRODUCTION

Connaitre et maitriser son risque est essentiel pour toute activité d'assurance. Le risque d'incapacité temporaire totale de travail [1] est un risque particulièrement différent des autres risques de prévoyance, comme le décès ou l'invalidité, puisqu'il s'agit d'un risque multiple et propre à la période d'activité professionnelle des assurés. L'étude réalisée dans ce mémoire porte sur un produit destiné au marché des travailleurs non salariés (TNS), elle constitue donc un véritable enjeu pour Axa France. En effet, les travailleurs non salariés sont souvent plus sensibilisés aux problématiques de prévoyance, dans la mesure où le risque personnel met directement en péril leurs revenus. De plus, ces derniers sont moins bien couverts par le régime obligatoire, que les salariés, notamment en ce qui concerne les risques décès, incapacité et invalidité. Pour pallier ces insuffisances et permettre aux travailleurs non salariés de bénéficier du même niveau de revenu en cas d'arrêt de travail, il y a donc nécessité de proposer une couverture complémentaire. Enfin, il s'agit également d'un marché porteur puisque des dispositifs tels que la loi Madelin permettent aux travailleurs non salariés de déduire les cotisations versées, de leurs bénéfices imposables.

Ce mémoire répond ainsi à une demande d'Axa France, leader sur le marché de la prévoyance individuelle, qui souhaite avoir un tarif plus adapté et plus compétitif.

Le risque d'incapacité peut être subdivisé en deux : le risque relatif à la fréquence des arrêts de travail et le risque relatif à leurs durées. Seule la fréquence des arrêts de travail, appelée incidence en incapacité, sera étudié ici. L'incapacité étant un risque particulier, l'objectif de ce mémoire sera de trouver une segmentation tarifaire capable de capter la complexité de ce risque. En effet, les principales causes d'entrée en incapacité étant la maladie et l'accident, les facteurs expliquant la fréquence des arrêts de travail sont différents selon la cause considérée. L'enjeu de ce mémoire sera donc de construire des classes tarifaires, homogènes en termes de risque incapacité, pour la cause accident et la cause maladie.

Pour cela, afin d'identifier les différents facteurs ayant un impact notable sur les niveaux de taux d'incidence, et de mieux cerner le risque et le portefeuille étudiés, l'étude débutera par un chapitre consacré à la construction de la base de données et son analyse descriptive. Puis, une fois les facteurs d'importance identifiés, pour chacune des deux causes, l'étude sera divisée en plusieurs parties.

1. L'assuré est considéré en incapacité temporaire totale de travail lorsque son état de santé, médicalement constaté, l'oblige à arrêter totalement et temporairement l'exercice de ses activités professionnelles par suite d'une maladie ou d'un accident.

Tout d'abord, le modèle de tarification envisagé sera présenté. Puis, la modélisation du risque incapacité sera étudiée dans deux parties indépendantes, l'une consacrée à l'étude de l'incidence en incapacité sur la cause accident, et l'autre sur la cause maladie. Dans chacune de ces parties, des classes tarifaires de professions homogènes seront construites, sur des critères propres à chacune des deux causes.

Enfin, dans la dernière partie, nous tenterons de concilier les classes tarifaires obtenues indépendamment sur la dimension accident et sur la dimension maladie, afin de proposer un tarif le plus adapté possible au profil de risque de chaque profession.

PARTIE

1

CONTEXTE DE L'ÉTUDE

1.1 Contexte interne à l'entreprise

1.1.1 L'entité AXA France

C'est en 1985 que la marque AXA naît sous l'impulsion de Claude Bébéar, qui entreprend une démarche d'internationalisation de l'entreprise. Plusieurs années plus tard, AXA est sur le podium des premières marques d'assurance mondiales, et devient la deuxième marque d'assurance la plus cotée à l'échelle européenne. Bien implanté sur les marchés d'Europe, d'Amérique du Nord et d'Asie Pacifique, au sein de 64 pays, le Groupe, réalise en 2020, un chiffre d'affaire s'élevant à 96,7 milliards d'euros.

Les activités d'AXA portent principalement sur ses cinq pôles d'activité : Vie, Épargne, Retraite, Dommages, Banques, Santé ainsi que sur le pôle Gestion d'actifs. En 2020, le plus haut chiffre d'affaires était réalisé par le pôle Dommages.

Plus importante filiale du groupe AXA, AXA France rassemble près de 33 000 collaborateurs au service de 6,3 millions de clients et 26,2 milliards d'euros de chiffre d'affaires en 2019, soit 25% du chiffre d'affaire du Groupe AXA. L'entité possède un résultat opérationnel de 1,42 milliards d'euros dont 20% est réalisé en Prévoyance et en Santé. Les autres domaines d'activités assurantiel sont le Dommage (45% du résultat opérationnel), l'épargne en fond euro (19% du résultat opérationnel) et l'épargne en unité de compte (correspondant aux 16% restants du résultat opérationnel).

L'entité AXA France contient de nombreuses entités organisationnelles et entreprises lui étant dépendantes, dont AXA France IARD, AXA France Vie, AXA Banque. L'entité dans laquelle s'est déroulée l'étude faisant l'objet de ce mémoire est AXA Particulier et IARD d'Entreprises [1], l'entité d'AXA France chargée du marché des particuliers, des professionnels et de l'IARD d'entreprise.

1. Egalement appeléée AXA PIE

1.1.2 Branche prévoyance individuelle au sein d'AXA France

Au sein d'AXA France PIE, la direction « Epargne et Prévoyance » a pour mission l'élaboration des offres et produits de prévoyance, épargne et retraite individuelle, la réalisation des actes de gestion, le pilotage de la maitrise des coûts et des stratégies de transformation du marché.

La direction « Prévoyance Individuelle » en est une sous-branche et est elle-même divisée en quatre services : le Service Client, l'Expertise Médicale, la direction « Stratégie, Transformation et Marketing Offres » et la direction technique, au sein de laquelle s'est déroulée cette étude. Cette dernière a pour mission d'assurer le *leadership* technique de la prévoyance individuelle au sein d'AXA France à travers :

- La conception de produits, leur tarification, le suivi des évolutions contractuelles et le lancement de nouvelles offres ;
- Le suivi du cycle de vie du produit, incluant le suivi des souscriptions et de la sinistralité des produits en portefeuille.

Maintenant que le contexte interne de l'entreprise a été présenté, il est important d'appréhender l'environnement juridique et social du risque "arrêt de travail".

1.2 Environnement réglementaire en prévoyance

1.2.1 Les régimes obligatoires

Présentation de la Sécurité Sociale

La Sécurité Sociale est un ensemble de régimes qui ont pour fonction de protéger les individus des conséquences de divers événements ou situations, généralement qualifiés de risques sociaux. Elle est destinée à assister financièrement ses bénéficiaires qui rencontrent différents événements coûteux de la vie.

L'ordonnance du 4 octobre 1945 instaure un système de Sécurité Sociale en France, basé sur un réseau coordonné de caisses, se substituant à de multiples organismes existants à cette date. Le système actuel de sécurité sociale, La Sécurité sociale, est composé de différents régimes regroupant les assurés sociaux selon leur activité professionnelle :

- Le régime général concerne les salariés du secteur privé ainsi que les travailleurs indépendants et couvre 88% de la population française ;
- Le régime agricole accompagne les exploitants, les salariés agricoles et les entreprises agricoles. Il couvre 5% de la population française ;
- Les régimes spéciaux regroupent les fonctionnaires, les employés et clercs de notaires, les mines, les cultes, etc. Ces régimes spéciaux sont au nombre de 27 et couvrent 7% de la population française.

D'un point de vue fonctionnel, l'organisation actuelle de la Sécurité Sociale résulte de l'ordonnance de 1967 qui instaure la séparation de la Sécurité Sociale en 4 branches autonomes. Chaque branche est alors responsable de ses ressources et de ses dépenses :

- Branche « Maladie » (maladie, maternité, paternité, invalidité, décès) ;
- Branche « Accidents du travail et Maladies professionnelles » ;
- Branche « Vieillesse et veuvage » (retraite) ;

- Branche « Famille » (dont handicap, logement, RMI...).

En cas d'arrêt de travail pour maladie, la Sécurité Sociale verse des indemnités journalières aux assurés, afin de compenser la perte de revenus liée à l'arrêt de travail. Ces indemnités journalières sont versées sous conditions de cotisations avec un délai de carence, et un montant dépendant du salaire.

Présentation de la sécurité sociale des indépendants (ex-RSI)

La sécurité sociale des indépendants (SSI) est un système d'organisation de la gestion de la protection sociale des travailleurs indépendants. A noter qu'il existe également d'autres régimes complémentaires pour les indépendants, comme par exemple CARPIMKO (Caisse autonome de retraite et de prévoyance des infirmiers, masseurs-kinesithérapeutes, pédicures-podologues, orthophonistes et orthoptistes), CARCDSF (Caisse Autonome de Retraite des Chirurgiens Dentistes et des Sages-Femmes), (Caisse Autonome de Retraite des Médecins de France), CIPAV (Caisse interprofessionnelle de prévoyance et d'assurance vieillesse des professions libérales) etc. La SSI couvre ainsi les artisans, commerçants et certains professionnels libéraux et gère leur protection sociale. Elle fait partie intégrante du régime général de la sécurité sociale, et remplace le régime social des indépendants (RSI), organisme de droit privé, depuis 2018. Cependant, bien que désormais rattachés au régime général de la sécurité sociale, les travailleurs non salariés conservent leurs propres règles en matière de cotisations et de prestations.

En cas d'arrêt de travail pour maladie ou accident, les artisans, commerçants ou industriels et certains professionnels libéraux non réglementés, bénéficient tous comme les salariés d'indemnités journalières. Néanmoins, les travailleurs non salariés sont moins bien couverts, par le régime obligatoire, que les salariés, notamment en ce qui concerne les risques décès, incapacité et invalidité.

Ainsi, la restriction des conditions d'indemnisation et la baisse importante des revenus, expliquent la nécessité pour un individu de souscrire une couverture supplémentaire. Lorsque l'assuré est salarié au sein d'une moyenne ou grande entreprise, cette couverture supplémentaire peut lui être fournie par son employeur, via un contrat de prévoyance collective.

En revanche, s'il s'agit d'un travailleur non salarié, les revenus pris en compte pour le calcul des prestations liées à l'arrêt de travail sont ceux des trois dernières années, et ceux des dix meilleures années en cas d'invalidité. Ces revenus sont pris en compte dans la limite du plafond annuel de la sécurité sociale en vigueur au jour du constat médical de l'incapacité de travail, soit 41 136 € bruts (au 1er janvier 2021). Ces revenus ne reflètent donc pas la situation actuelle du travailleur.

Par conséquent, il est d'autant plus important, pour un travailleur non salarié ou un salarié d'une petite entreprise, de souscrire un contrat de prévoyance individuelle, afin de limiter la perte des revenus liée à l'arrêt de travail.

1.2.2 Les régimes complémentaires aux régimes obligatoires

Les contrats d'assurances prévoyance ont pour objectif de protéger l'assuré et ses proches contre les risques tels que l'incapacité, l'invalidité, le décès. Afin de faire face à un imprévu pouvant notamment engendrer une baisse de revenus, ces contrats apportent ainsi une protection complémentaire à celle existante (i.e. le régime obligatoire de la sécurité sociale ou assimilé).

Selon les niveaux de garanties souscrits, cette assurance peut compenser la perte de revenus subie par l'assuré et ses proches sous différentes formes de prestations.

On cite par exemple :

- Le versement d'un capital en cas de décès ou d'invalidité ;
- Des indemnités journalières en cas d'incapacité temporaire totale de travail ;
- Des rentes diverses (éducation, conjoint, invalidité) ;
- L'exonération de paiement des primes.

La loi Evin du 31 décembre 1989 crée le premier ensemble de règles portant sur les garanties prévoyance, applicables à toutes les familles d'assureurs. Elle distingue trois types d'organismes habilités à mettre en œuvre une protection sociale complémentaire :

- **Les mutuelles** : il s'agit de sociétés de personnes à but non lucratif, régies par le Code de la mutualité, reposant sur le principe de solidarité entre l'ensemble des adhérents et dont les fonds proviennent principalement des cotisations des membres.
- **Les institutions de prévoyance** : il s'agit d'organismes qui gèrent des contrats collectifs d'assurance de personnes, souscrits par les entreprises au bénéfice des salariés, dans le cadre des accords d'entreprises ou de branches professionnelles. Sociétés de personnes, de droit privé, et à but non lucratif, les institutions de prévoyances sont régies par le Code de la sécurité sociale et relèvent des directives européennes sur l'assurance.
- **Les entreprises d'assurance** : elles sont quant à elles majoritairement à objet commercial et sont soumises au code des assurances, ce qui implique qu'elles doivent respecter des normes prudentielles importantes (marge de solvabilité, fonds de garantie, capital social, provisions, placements financiers règlementés). Elles peuvent être de deux formes :
  - Sociétés anonymes : il s'agit d'organismes ayant pour vocation de faire des bénéfices et de les redistribuer aux actionnaires.
  - Sociétés mutuelles : les sociétés d'assurances mutuelles sont quant à elles à but non lucratif et fonctionnent sans capital social. Leur activité est pour l'essentiel dans le domaine des assurances dommages.

En France, le marché de la prévoyance est dominé par les contrats collectifs. Ces contrats sont gérés à parts presque égales par les institutions de prévoyance et les sociétés d'assurance. En revanche, les cotisations des contrats de prévoyance individuelle sont collectées, pour la quasi-totalité, par les sociétés d'assurance qui couvrent trois types de risques : le décès, l'incapacité et l'invalidité.

1.3 Les garanties proposées pour le risque arrêt de travail par les contrats de prévoyance

1.3.1 Quelques définitions

Incapacité temporaire totale de travail

L'incapacité de travail est l'arrêt momentané de l'activité professionnelle pour cause de maladie ou d'accident. La période d'incapacité est limitée dans le temps à 3 ans. Au-delà de 3 ans, et sous réserve que l'état de santé de l'assuré soit consolidé, il pourra éventuellement y avoir classement en invalidité

Invalidité permanente partielle ou totale

Une personne est déclarée invalide si sa capacité de travail est réduite de manière permanente à la suite d'un accident ou à une maladie.

L'état de consolidation en invalidité peut être reconnu, soit par le médecin de la sécurité sociale ou assimilé pour le régime obligatoire, soit par le médecin de l'assureur pour le versement des prestations invalidité.

Notion de Franchise

La franchise correspond à la période entre la survenance du sinistre ouvrant droit aux indemnités journalières (IJ) et le début de cette indemnisation. En effet, pour un assuré en arrêt de travail, le régime obligatoire verse des indemnités journalières après un délai de carence. Ce délai de carence varie selon la cause. Si l'assuré a souscrit des garanties complémentaires, en cas d'incapacité, sur un contrat de prévoyance individuelle, d'autres indemnités journalières viendront s'ajouter aux prestations versées par le régime obligatoire. Ces dernières permettent à l'assuré de maintenir son salaire durant son arrêt de travail, et sont versées après la période de franchise, définie dans le contrat.

1.3.2 Garanties en cas d'incapacité temporaire de travail

Dans un contrat d'assurance prévoyance, le versement des prestations des garanties en cas d'incapacité temporaire de travail se fait en général sous forme d'indemnités journalières. Ces dernières viennent en complément des indemnités versées par la Sécurité Sociale ou assimilé, afin de permettre à l'assuré de maintenir ses revenus en cas d'arrêt de travail. Les IJ versées par un régime obligatoire ne permettent en général pas d'avoir un maintien des revenus à 100% en cas d'arrêt de travail. C'est la raison pour laquelle les assurés souscrivent un contrat de prévoyance, avec une couverture sur le risque incapacité, pour compléter les IJ versées par le régime obligatoire.

Une garantie d'exonération de cotisations en cas d'incapacité peut également être souscrite. Celle-ci prévoit le remboursement des cotisations de l'ensemble des garanties en cours pendant la période d'indemnisation.

Pour le produit étudié, pour chaque arrêt de travail, le paiement des prestations débute au lendemain de l'expiration de la période de franchise et l'indemnisation cesse à la plus proche des dates suivantes :

- A la reprise totale d'activité professionnelle ou date de consolidation d'une invalidité permanente ;
- A partir du 1095e jour d'arrêt du travail, qui correspond à la durée maximale de l'état d'incapacité ;
- A la liquidation des droits au titre du régime obligatoire de retraite dont l'assuré relève selon sa profession.
- L'année d'assurance au cours de laquelle l'assuré est âgé de 67 ans.

Ces notions sont illustrées dans l'exemple suivant :

Un assuré a souscrit un contrat de prévoyance avec des garanties prévoyant le versement d'indemnités journalières en cas d'arrêt de travail. Cet assuré est pharmacien, et est rattaché au régime CAVP. La franchise associée au contrat est une franchise 0/15/15. Ce qui signifie qu'elle a une

durée de 0 jour pour un arrêt de travail faisant suite à une hospitalisation, et 15 jours pour un arrêt faisant suite à un accident ou une maladie.
Cet assuré entre en arrêt de travail à la suite d'une maladie.

Les différentes indemnités qu'il va percevoir durant sa période d'incapacité sont donc les suivantes :

- 0 - 3e jour : délai de carence, aucune indemnité perçue ;
- 4e jour : début du versement des indemnités de la CAVP à hauteur de 50% du salaire brut ;
- A partir du 16e jour : intervention du régime de prévoyance. Versement d'indemnités journalières en complément des prestations versées par la CAVP et à hauteur du montant souscrit ;
- A partir du 90e jour, fin du versement desprestations versées par la CAVP

Ces notions sont illustrées dans la figure 1.1 ci-dessous :

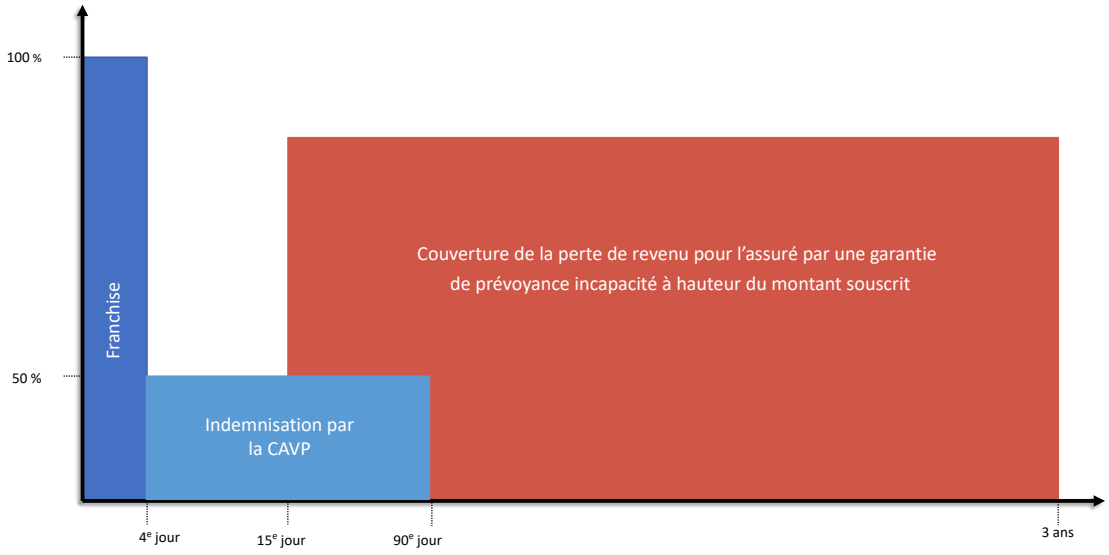

FIGURE 1.1 – Représentation de la prise en charge de la perte de revenus d'un assuré lors d'un arrêt de travail

1.3.3 Garanties en cas de décès ou d'invalidité

En cas de décès, l'assurance prévoyance permet de préserver l'indépendance financière des proches de l'assuré, alors que les garanties invalidité permettent le maintien du niveau de vie de l'assuré en cas d'invalidité. Cette protection s'effectue sous forme de rentes ou de capital, versés au(x) bénéficiaire(s) désigné(s) dans le contrat, selon le niveau d'invalidité ou en cas de décès.

Ce mémoire ayant pour objet la refonte tarifaire de la garantie incapacité du produit étudié, seul les risque incapacité sera désormais étudié.

1.4 La modélisation du risque incapacité

1.4.1 Les caractéristiques du risque incapacité

Le risque incapacité possède plusieurs particularités qu'il est nécessaire de présenter.

Tout d'abord, la population observée est celles des individus en âge d'être en activité professionnelle, c'est-à-dire entre 18 et 67 ans. En effet, les garanties prennent fin lors du départ à la retraite de l'assuré, ou au plus tard, à 67 ans. Passé cet âge, on ne parle plus de situations d'incapacité ou d'invalidité. Le statut de l'assuré est alors remplacé par celui de retraité et/ou de personne dépendante.

La deuxième spécificité du risque incapacité est son caractère provisoire. Il s'agit d'un état pouvant durer au maximum 3 ans. Si l'état se consolide et dure plus de 3 ans, l'assuré est alors en état d'invalidité. Cette spécificité du risque implique la nécessité de s'intéresser non seulement à la probabilité d'entrer en incapacité, mais également à la durée de maintien en incapacité des assurés.

Enfin, une autre particularité du risque incapacité concerne son caractère répétitif. En effet, contrairement au décès et à l'invalidité, l'incapacité n'est pas un état irréversible et peut se répéter plusieurs fois au cours de la durée de vie du contrat de l'assuré.

1.4.2 Cycle de vie d'un produit de prévoyance et tarification

Un produit de prévoyance possède plusieurs particularités qui le différencient des autres types de produits d'assurance. Premièrement, les garanties couvrant l'incapacité temporaire totale de travail ne s'appliquent que si l'assuré exerce une activité professionnelle effective. Par ailleurs, l'adhésion est souscrite pour une durée d'un an et est automatiquement reconduite chaque année pour la durée d'un an supplémentaire. Néanmoins, l'adhérent a la possibilité de résilier son contrat annuellement à la date anniversaire du contrat ou en cas de non-acceptation par ce dernier, des nouvelles évolutions tarifaires faites par l'assureur. L'assureur a quant à lui la possibilité de radier l'adhésion de l'assuré aux garanties de prévoyance dans les deux ans suivant la prise d'effet du contrat, sans avoir à se justifier. En revanche, passée l'expiration d'un délai de deux ans suivant l'adhésion, l'assureur ne peut plus résilier l'adhésion tant que l'assuré n'a pas atteint les limites d'âge prévues au contrat ou dans des circonstances bien précises [2].

Enfin, lors de l'adhésion, l'adhérent devra déclarer sa profession exacte, ainsi que répondre à des questionnaires médicaux, sportifs et/ou professionnels. On parle alors de sélection médicale, financière et/ ou sportive. En effet, la pratique d'un sport et/ou d'une activité à risque peut générer des conditions d'acceptation particulières (exclusion, surprime...). De même, la sélection médicale dès la souscription, permet au client d'avoir un contrat et une tarification adaptés à sa situation, tenant compte de son passé médical. Ainsi, après sélection médicale, un contrat pourra être accepté au tarif normal, avec une surprime et/ou une exclusion, être refusé ou ajourné.

Le tarif est ensuite construit en fonction de différents critères. Rappelons que, la tarification d'un contrat d'assurance a pour principe fondamental l'égalité des engagements de l'assureur et de l'assuré. La prime pure correspond donc au juste prix du contrat d'assurance, prix permettant d'égaliser l'engagement de l'assuré (le paiement des primes) et celui de l'assureur (le versement d'indemnité en cas de sinistre).

Etant données les particularités du risque incapacité, deux éléments sont indispensables à sa tarification et à son provisionnement. Il s'agit des tables d'entrée et de maintien en incapacité.

2. Loi 89-1009 du 31 décembre 1989, dite "loi Evin", art. 29 III : champ d'application de l'article 6.

Leur utilisation, par les entreprises régies par le code des assurances, est rendue obligatoire par deux textes essentiels : la Loi Evin et l'arrêté du 28 mars 1996 [3].

1.5 Caractéristiques du produit faisant l'objet de ce mémoire

1.5.1 Description des garanties

Le produit étudié est destiné aux professionnels indépendants (artisans, commerçants, professions libérales exploitants agricoles, et conjoints collaborateurs). Il a pour objet de garantir, en fonction des garanties souscrites, des prestations en cas de décès ou d'invalidité permanente totale (IPT) de l'assuré, d'invalidité permanente partielle ou totale (IPPT) de l'assuré, d'incapacité temporaire totale de travail (ITT) de l'assuré survenant pendant la période de couverture de ce risque. Selon les garanties souscrites, il est prévu le versement d'un capital et/ou d'une rente au(x) bénéficiaire(s) désigné(s), ou le versement d'indemnités journalières.

La refonte tarifaire du produit faisant l'objet de ce mémoire, portant sur le risque incapacité, les différentes garanties proposées pour les risques décès et invalidité ne seront pas étudiées.

De plus, plusieurs garanties de type indemnités journalières sont proposées par le produit, mais ce mémoire portera uniquement sur la garantie IJ toutes causes, qui permet le versement d'indemnités journalières des suites d'une maladie ou d'un accident, avec ou sans hospitalisation.

1.5.2 L'acceptation des risques et la tarification

L'enjeu de ce mémoire portant sur la tarification d'un produit de prévoyance individuelle, il est important d'avoir en tête les critères de cotisations auxquels sont soumis les assurés.

La première étape, avant de construire le tarif, est, comme expliqué précedemment, la sélection des risques. Cette évaluation des risques est établie à partir de questionnaires médicaux, sportifs et/ou professionnels.

Puis, après cette étape de sélection des risques, le tarif est construit en fonction de différents critères selon lesquels sont calculées les cotisations :

- L'âge de l'assuré : à chaque âge correspond un tarif spécifique pour chacune des garanties décès, invalidité et incapacité ;
- La franchise ainsi que le niveau des garanties souscrites ;
- Le groupe tarifaire auquel l'assuré appartient selon sa profession.

La refonte tarifaire proposé dans ce mémoire, portera sur ce dernier critère. Le but de la segmentation d'un portefeuille en classes tarifaires, est de créer des groupes les plus homogènes possibles, afin de pouvoir calculer une prime proche de chacun des risques qui composent la classe.

L'enjeu de ce mémoire sera donc, dans un premier temps d'étudier les comportements des assurés du portefeuille face au risque incapacité, afin de construire une nouvelle segmentation tarifaire plus adaptée au profil de ce dernier.

3. Arrêté du 28 mars 1996 fixant les règles de provisionnement des garanties d'invalidité et d'incapacité

PARTIE

2

LES DONNÉES ET LEUR ANALYSE

2.1 Présentation des données

2.1.1 Période d'observation

Choix de la période d'observation

Le choix de la période d'observation est primordial puisqu'il va conditionner la pertinence et la robustesse des taux d'incidence calculés.

Dans cette étude, le produit étudié a été commercialisé au cours de l'année 2013. Ainsi, afin que la sinistralité observée durant l'année du lancement du produit, ait moins d'impact sur les résultats, il a été décidé de ne pas retenir l'année 2013. En effet, durant l'année de lancement du produit, tous les assurés en portefeuille, sont très proches de la sélection médicale. La connaissance du risque est donc très bonne durant cette année là. En effet, plus l'année d'observation s'éloigne de l'année de souscription, plus la connaissance du risque se dégrade. S'agissant d'un produit relativement récent, il est donc normal d'observer une évolution de la sinistralité sur les dernières années. En effet l'ensemble des assurés s'éloigne progressivement de la sélection médicale et l'âge moyen du portefeuille vieillit.

Par ailleurs, la crise sanitaire et économique engendrée par la pandémie de Covid-19 ayant provoqué une sinistralité atypique durant l'année 2020, il a été décidé de ne pas intégrer cette année à la période d'observation, afin de ne pas déformer les résultats en incluant des comportements non représentatifs. Il s'agit de plus, d'une année d'observation non consolidée au moment de la création de la base de données.

Ainsi, cette étude s'appuiera sur l'observation des années 2014 à 2019, soit sur une période de six ans.

Censures et troncatures

Le choix de la période d'observation implique une perte d'une partie de l'information de l'échantillon de données. Il est donc important de définir deux notions primordiales des modèles de durées :

la censure et la troncature.

On parle de censure lorsque le phénomène observé n'est pas observé de manière complète. Par exemple, pour les assurés ayant souscrit un contrat d'assurance avant le début de la période d'observation, il se peut qu'un sinistre ait eu lieu durant cette période. On parle alors de censure à gauche. On sait alors seulement que la date de début du sinistre est antérieure au début de notre observation. Pour cette étude, étant donné que seuls les sinistres survenus durant la période d'observation, sont conservés, ce type de censure n'aura pas d'impact.

On rencontre le phénomène inverse lorsque l'individu ne peut pas être observé jusqu'à la date de fin de son sinistre. On parle alors de censure à droite lorsque le début de l'incapacité d'un individu est observé, mais que ce dernier sort de l'observation pour une autre cause que la fin de son arrêt de travail. Ce phénomène peut être rencontré lorsque l'assuré sort de l'exposition car la date de fin d'effet de son contrat a eu lieu, en cas de résiliation, de départ à la retraite ou bien lorsque la date de fin d'observation est atteinte.

Ces phénomènes de censures ont notamment un impact lorsque la variable étudiée est la durée du sinistre. Dans cette étude, puisque seule l'incidence en incapacité est étudiée, l'impact des phénomènes de censure sera donc moins important.

En revanche, les phénomènes de troncature auront un impact sur l'incidence observée. En effet, l'information concernant les sinistres dont la durée est inférieure à celle de franchise est inconnue. Seuls les sinistres dépassant la durée de la franchise du contrat sont donc disponnibles. Dans ce cas, l'information sur les observations est perdue, contrairement aux sinistres censurés, pour lesquels l'information est incomplète.

La partie suivante sera consacrée aux traitements effectués sur les données brutes, ayant permis la construction de la base de données de finale. La construction d'une première base de données brutes s'est faite en trois temps : récupération des données contrats dans une première base de données, récupération des données sinistres dans une deuxième base, puis jointures des deux tables.

2.1.2 Extraction des données contrats

Récupération des données

La première étape consiste à récupérer dans des bases techniques internes, les données relatives aux assurés et à leur contrat. Dans ces bases de données, un contrat est représenté par autant de lignes que de garanties souscrites. Un premier filtre permet de récupérer tous les contrats possédant des garanties versant des prestations de type indemnités journalières. Les variables suivantes sont ensuite conservées :

- Numéro de contrat ;
- Date de naissance de l'assuré ;
- Sexe de l'assuré ;
- Code CSP ;
- Classe tarifaire ;
- Date d'effet du contrat ;
- Code garantie ;

- Date d'effet de la garantie ;
- Date de fin d'effet de la garantie ;
- Statut du contrat ;
- Date de résiliation dans le cas des contrats résiliés ;
- Franchise.

La variable **code CSP** disponible dans la base de données est une variable composée de quatre caractères, permettant de classer les professions selon leur domaine d'activité.

Traitement des données

Pour rappel, chaque assuré a la possibilité de souscrire différentes garanties de types indemnités journalières pour un même contrat. Selon la durée du sinistre, les garanties s'enclenchent successivement en commençant par celle ayant la franchise la plus courte.

Dans la base sinistre une ligne sinistre est créée pour chacune garantie. Ainsi, pour un même sinistre, autant de lignes que de garanties déclenchées, seront enregistrées. Afin de pouvoir regrouper les garanties qui s'enclenchent successivement, certains traitements doivent être effectués sur la base contrats. L'objectif de ces traitements est d'obtenir, pour un même contrat, une ligne par exposition au risque.

La variable d'intérêt étant l'incidence et non le maintien en arrêt de travail, pour chaque contrat, les garanties IJ toutes causes s'enclenchant successivement sont regroupées sous un même numéro de sinistre. En effet, pour compléter les prestations du régime obligatoire, un montage d'IJ complémentaires est réalisé, permettant de couvrir l'assuré à 100% de ses revenus. Dans le cas où plusieurs garanties s'enclenchent successivement, la garantie ayant la franchise qui s'enclenche la première est retenue pour la variable franchise.

L'exemple suivant illustre ces nations. Un assuré a souscrit trois types d'IJ toute cause. La première, IJ1, a une franchise 15/15 et une durée de 90 jours. La seconde, IJ2, a une franchise 90/90 et une durée de 365 jours. Enfin, la troisième, IJ3, a une franchise 365/365 et une durée de 1095 jours. Initialement, cet assuré est représenté par trois lignes dans la base de données :

Numéro contrat	IJ	Franchise	Durée de la garantie
XXX	IJ1	15/15	90
XXX	IJ2	90/90	365
XXX	IJ3	365/365	1095

Après traitement, une seule ligne est retenue pour cet assuré :

Numéro contrat	IJ	Franchise	Durée de la garantie
XXX	IJ	15/15	1095

Ces traitements permettent de ne pas surestimer l'incidence, et de compter comme trois sinistres distincts, un sinistre qui s'enclenche sur trois garanties successivement.

2.1.3 Extraction des données sinistres

Récupération des données

Les données relatives aux sinistres sont ensuite extraites des bases annuelles de 2014 à 2019. Un premier filtre est ensuite effectué, comme pour la base contrats, afin de conserver les sinistres ayant reçu une prestation de type indemnités journalières.

Les variables suivantes sont conservées :

- Numéro de contrat ;
- Numéro de sinistre ;
- Code de la garantie enclenchée pour l'indemnisation du sinistre ;
- Date de début d'arrêt ;
- Date de fin d'arrêt ;
- Cause du sinistre.

Traitements des données

Dans les bases sinistres, pour chaque sinistre, plusieurs lignes sont présentes : une ligne ouverture, et des lignes prolongations. A noter que parfois, les prolongations d'un sinistre sont enregistrées sous un numéro de sinistre différent. Ces erreurs de traitements peuvent avoir un impact important sur la mesure de l'incidence en arrêt de travail. Afin d'éviter qu'un même sinistre soit compté comme plusieurs sinistres distincts, il est donc primordial de traiter ces anomalies.

Pour traiter les creux entre deux sinistres, il a été décidé de retenir comme critère celui utilisé dans les tables d'expérience de maintien en incapacité. Ainsi, on considère comme un creux ou une rechute, deux sinistres sur un même contrat, qui se suivent avec la même cause, mais avec un intervalle de temps, entre la fin du 1er arrêt et le début du second, inférieur à 90 jours. En effet, les conditions générales du produit stipulent qu'en cas de rechute dans les 90 jours suivant la reprise d'activité, le paiement des indemnités reprend immédiatement, sans nouveau délai de franchise, si le nouvel arrêt est dû aux mêmes causes que l'arrêt précédent. En revanche, en cas de rechute au-delà des 90 jours suivant la reprise d'activité, la franchise sera de nouveau appliquée.

FIGURE 2.1 – Représentation d'un creux

Un autre type d'anomalie est fréquemment rencontré sur les bases sinistres : les chevauchements. On appelle chevauchement, deux sinistres sur un même contrat, avec la même cause, mais dont le début du second arrêt précède la fin du premier.

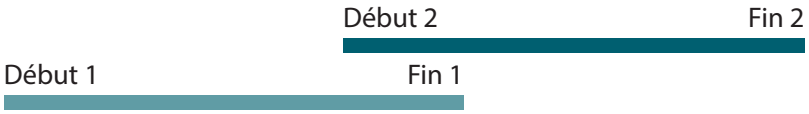

FIGURE 2.2 – Représentation d'un chevauchement

L'objectif du traitement de ces anomalies est d'obtenir, à partir de plusieurs sinistres, liés au même numéro de contrat, un seul sinistre. Le sinistre, une fois traité, aura donc comme caractéristiques :

- Le même numéro de contrat ;
- La même cause ;
- Une date de début correspondant à la date de début d'arrêt la plus ancienne des différents sinistres concernés ;
- Une date de fin correspondant à la date de fin d'arrêt la plus récente des différents sinistres concernés.

Pour les sinistres concernés par un chevauchement de dates, mais pour lesquels deux causes différentes sont associées à chaque sinistre, une recherche dans le système de gestion a été effectuée, afin de récupérer la cause exacte de l'arrêt. Pour dix sinistres il n'a pas été possible de récupérer la cause du sinistre. Ces sinistres ont donc dû être écartés de notre base de données.
Afin d'illustrer les traitements effectués, voici un exemple récapitulatif :

Avant traitements :

Numéro contrat	Garantie	Cause	Date de début	Date de fin
XXX	IJ	Maladie	01/01/2015	01/02/2015
XXX	IJ	Maladie	15/02/2015	30/04/2015
XXX	IJ	Maladie	01/04/2015	15/05/2015

Après traitements :

Numéro contrat	Garantie	Cause	Date de début	Date de fin
XXX	IJ	Maladie	01/01/2015	15/05/2015

2.1.4 Jointure

Une fois les traitements sur les bases contrats et sinistres effectués, une jointure entre ces deux tables est réalisée. Cette jointure se fait avec la clé *Numéro de contrat*. Ainsi, un contrat non sinistré sera représenté sur une ligne, et un contrat sinistré sera quant à lui représenté sur autant de lignes qu'il a de sinistres. Une fois la jointure effectuée, d'autres tests de cohérences des données ont été effectués :

- Age de souscription compris entre 18 et 67 ans ;
- Age de début du sinistre compris entre 18 et 67 ans ;
- Date d'effet des garanties $<$ date de début du sinistre $<$ date de fin du sinistre $<$ min(date de fin de la garantie, date de fin d'effet du contrat, date de résiliation) ;
- Valeurs manquantes (sexe, csp, date de naissance).

Finalement, presqu'aucun contrat n'était concerné par une des anomalies mentionnées ci-dessus. Le peu de contrats concernés ont pu être corrigés après consultation des systèmes de gestion et aucun contrat n'a dû être supprimé de notre base de données.

2.2 Création de variables à partir de la base de données

2.2.1 Transposition de la base de données

Comme expliqué dans la partie précédente, après jointure des bases contrats et sinistres, un contrat sinistré est représenté par autant de ligne que de sinistres. Afin de n'avoir qu'une ligne par exposition, la base de données a été transposée, de façon à n'avoir qu'une ligne par contrat.

Ainsi, avant transposition la base de données se présentait de la façon suivante :

Numéro contrat	Information contrat	Garantie	Informations sinistres
XXX	.....	IJ	Sinistre 1
XXX	.....	IJ	Sinistre 2
XXX	.....	IJ	Sinistre 3

Après transposition, elle se présente comme suit :

Numéro contrat	Infos contrat	Garantie	Infos sinistre 1	Infos sinistre 2	Infos sinistre 3
XXX	.....	IJ	Sinistre 1	Sinistre 2	Sinistre 3

2.2.2 Calcul de l'exposition

Afin d'étudier l'incidence du risque incapacité, il est primordial de calculer l'exposition de chaque contrat. L'exposition d'un assuré correspond aux intervalles de temps durant lesquels l'assureur est exposé au risque incapacité de l'individu observé, conformément aux périodes de validité des garanties de son contrat. Ainsi, pour chaque année d'observation, l'exposition de chacun des contrats est calculée.

La date de début d'exposition correspond à la date de début de la garantie, et la date de fin d'exposition correspond à la plus petite des dates suivantes :

- Date de changement de statut du contrat, si le contrat n'est plus en vigueur ;
- Date de fin de la garantie, si le contrat est toujours en vigueur.

Afin d'avoir une exposition par âge, l'exposition est donc calculée sur chaque âge durant lesquels l'assuré est présent dans le portfeuille.

Ainsi, pour un assuré né le 01/01/1960 ayant souscrit un contrat avec des garanties IJ prenant effet le 01/04/2015 et fin le 01/07/2019 on aura les caractéristiques suivantes :

N° contrat	Expo âge 55	Expo âge 56	Expo âge 57	Expo âge 58	Expo âge 59
XXX	0,75	1	1	1	0,5

L'exposition d'un contrat étant calculée de la façon suivante :

$$Expo_{Age_i} = \frac{\text{Nombre jours de présence sur l'âge } i}{365,25}$$

Dans cet exemple, l'assuré est exposé de ses 55 ans à ses 59 ans. La valeur de la variable **Expo** est donc nulle sur les âges autres, que ceux compris entre 55 et 59 ans.

Cependant, cette définition de l'exposition n'est pas suffisante pour mesurer le risque incapacité. En effet, lorsqu'un assuré est en arrêt de travail, l'assureur n'est de fait plus exposé au risque incapacité pour cet assuré. Ainsi, lorsqu'un assuré entre en arrêt de travail il sort de l'exposition, et lorsque son arrêt de travail prend fin, il est remis en exposition. Il est donc important de déduire de l'exposition calculée précédemment, les périodes d'arrêt de travail. Cette particularité et sa conséquence sur l'étude, seront analysées plus loin.

Maintenant que les différents traitements appliqués aux données, ont été expliqués, la partie suivante sera consacrée à une analyse descriptive du portefeuille, en termes de population et de sinistres.

Il s'agit d'une étape indispensable, qui permettra par la suite d'orienter cette étude et d'évaluer la fiabilité des modèles utilisés et de leurs résultats.

2.3 Analyse descriptive sur la base de données finale

L'analyse descriptive des données est divisée en deux parties : une partie consacrée à la l'étude du portefeuille au niveau des assurés, et l'autre consacrée à l'étude de la sinistralité.

Pour des raisons de confidentialités, les données chiffrées exactes associées aux graphes ne pourront être dévoilées, mais cela ne remet pas en cause les résultats obtenus.

2.3.1 Description du portefeuille au niveau des assurés

Avant d'analyser le risque incapacité sur le portefeuille, il est primordial d'effectuer une première description de ce dernier. Après s'être assurés que la répartition des sexes était la même sur chacune des années d'observation, les données ont été agrégées, afin de réaliser une répartition moyenne sur cette période.

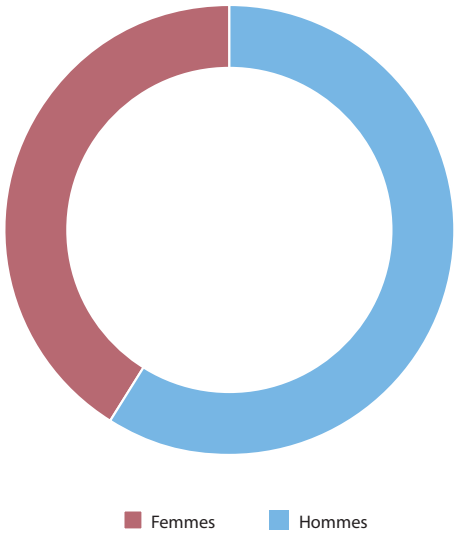

FIGURE 2.3 – Répartition des sexes dans le portefeuille d'assurés

Il s'agit donc d'un portefeuille constitué en grande majorité par des hommes.

L'âge étant également une dimension essentielle à l'étude du risque incapacité, il est intéressant d'étudier la répartition des sexes par âge sur le portefeuille.

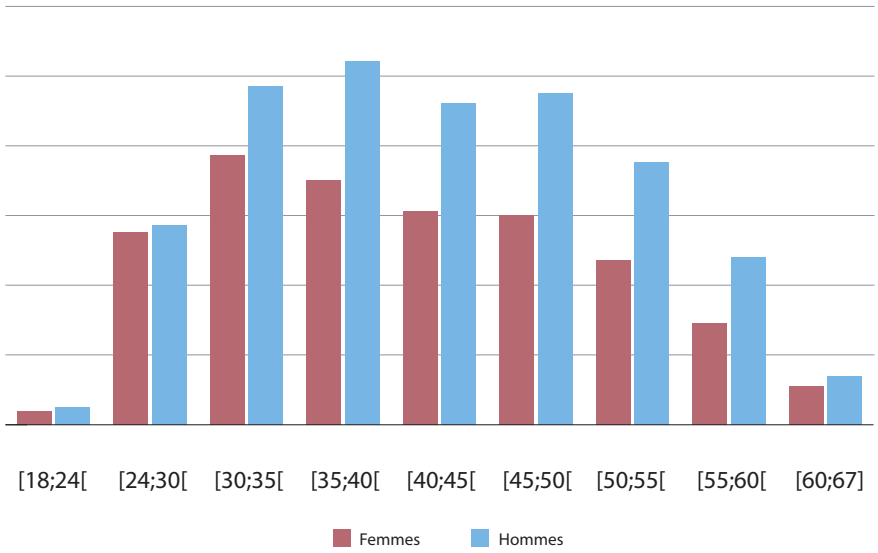

FIGURE 2.4 – Répartition des sexes par tranches d'âge dans le portefeuille d'assurés

Ce graphique montre que, sur chaque tranche, les hommes occupent une part plus importante que les femmes. En revanche, la répartition par tranche d'âge est différente entre les femmes et les hommes. La tranche [30 ;35[ans, est la tranche sur laquelle les femmes sont les plus présentes en proporition. Cette proportion ne fait ensuite que diminuer jusqu'à 67 ans. Pour les hommes leur répartition est plutôt équivalente entre 30 et 50 ans, puis une décroissance est observée jusqu'à 67 ans.

Il est également intéressant d'étudier le portefeuille au niveau des classes tarifaires, puisque ces dernières font l'objet de ce mémoire.

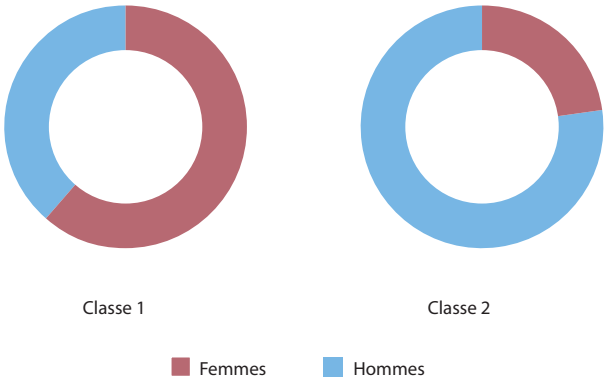

FIGURE 2.5 – Répartition des classes tarifaires par sexe dans le portefeuille d'assurés

Pour des raisons de confidentialité, afin de ne pas dévoiler le nombre réel de classes tarifaires ainsi que la structure tarifaire du produit, les classes tarifaires ont été agrégées. Ces traitements

ne remettent, néanmoins, pas en cause les résultats.

La figure 2.5 montre que la proportion d'hommes augmente fortement entre les classes 1 et 2. Ceci s'explique par le fait que la classe 2 regroupe les professions les plus accidentogènes, et que ces professions sont en majorité exercées par des hommes.

Il est également essentiel d'analyse le portefeuille au niveau de sa sinistralité.

2.3.2 Analyse descriptive du risque incapacité/description du portefeuille au niveau sinistralité

Introduction

Il existe plusieurs façons d'étudier le risque incapacité : il est possible d'étudier le risque relatif à la fréquence des arrêts de travail, et le risque relatif à leur durée. Dans ce mémoire, il a été décidé de prendre l'hypothèse selon laquelle le maintien en arrêt de travail est principalement déterminé par la cause de l'arrêt de travail, et l'âge à la survenance. Cela revient donc à supposer que les autres variables n'influent pas sur le maintien.

Ainsi, selon cette hypothèse, pour adapter la tarification au portefeuille de risque, il sera nécessaire d'étudier la fréquence des arrêts de travail, appelée incidence en incapacité. En ce qui concerne le maintien en incapacité, les tables d'expérience seront utilisées.

Cette partie sera donc consacrée à une première analyse descriptive de la sinistralité du portefeuille. L'enjeu ne sera pas d'établir un modèle pouvant rendre compte de l'influence d'une variable sur le taux d'incidence en incapacité, mais de se faire une première idée du comportement du portefeuille, afin de préparer les modélisations à venir et d'anticiper leur fiabilité.

Dans la suite, le taux d'incidence sera désigné de la manière suivante :

$$\text{Taux d'incidence} = \frac{\sum \text{Sinistres}}{\sum \text{Exposition}}$$

Analyse du taux d'incidence globale

Avant toute chose, il est important d'étudier l'évolution du taux d'incidence sur la période d'observation.

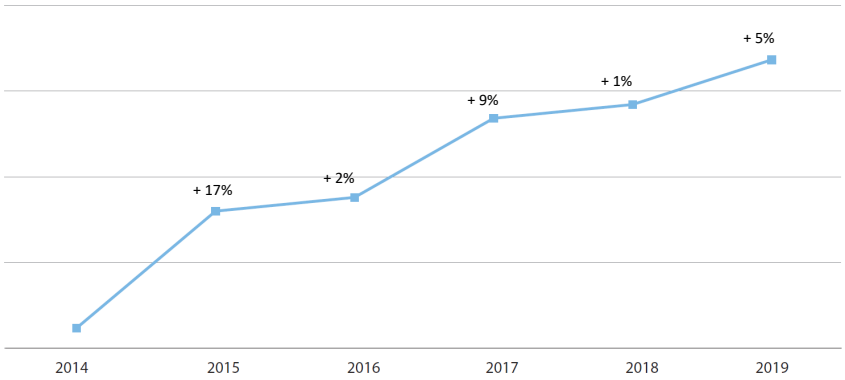

FIGURE 2.6 – Evolution de l'incidence du risque incapacité par année d'observation

Une constante augmentation de l'incidence est observée, entre 2014 et 2019. Les analyses suivantes auront donc pour objectif d'expliquer cette augmentation de l'incidence. Pour cela, la maille d'observation sera progressivement réduite, afin de capter des comportements ou critères, pouvant expliquer cette augmentation constante du taux d'incidence depuis 2014.

Analyse du taux d'incidence par cause

La variable cause renseigne sur le type d'évènement ayant conduit à l'arrêt de travail. Il s'agit d'une variable fondamentale, car elle va fixer, la franchise appliquée pour le sinistre.

Il est évident que cette variable a une grande influence sur la durée de l'arrêt de travail, puisque les tables d'expérience sont en premier lieu segmentées selon cette dernière.

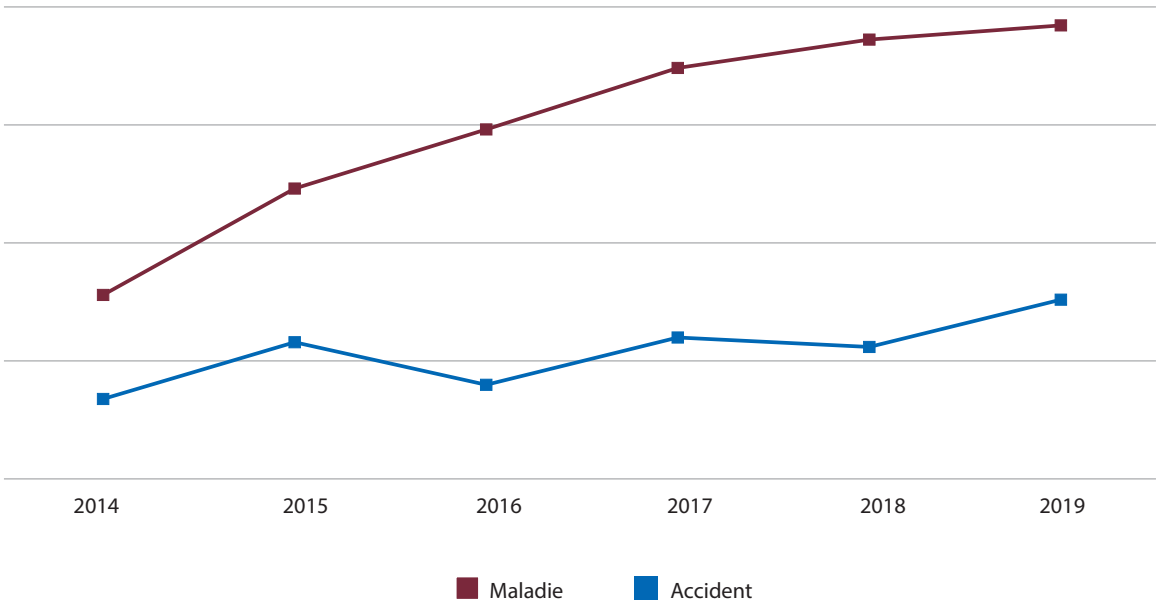

FIGURE 2.7 – Evolution de l'incidence du risque incapacité par année d'observation selon la cause du sinistre

La figure 2.7 montre que le taux d'incidence pour la cause accident semble plutôt constant au cours du temps, et est constamment inférieur au taux d'incidence pour la cause maladie.

En revanche, le taux d'incidence pour la cause maladie est en constante augmentation sur la période d'observation. L'augmentation du taux d'incidence global est donc fortement expliquée par l'augmentation du taux d'incidence pour la cause maladie.

A noter que le nombre de sinistre total sur la période d'observation est de 12 000, 40% étant des sinistres accident et 60% des sinistres maladie.

Une première hypothèse pouvant expliquer cette augmentation, serait le vieillissement du portefeuille en termes d'âge. Il est également possible de supposer que la connaissance du risque diminue au cours, du temps du fait de l'éloignement de la sélection médicale faite au moment de la souscription du contrat. Afin d'explorer d'autres pistes, cette analyse sera poursuivie avec l'étude de l'évolution du taux d'incidence par sexe.

Analyse du taux d'incidence par sexe

Le sexe constitue un élément central et discriminant de cette étude, il est donc nécessaire d'étudier l'évolution du taux d'incidence en fonction de ce paramètre.

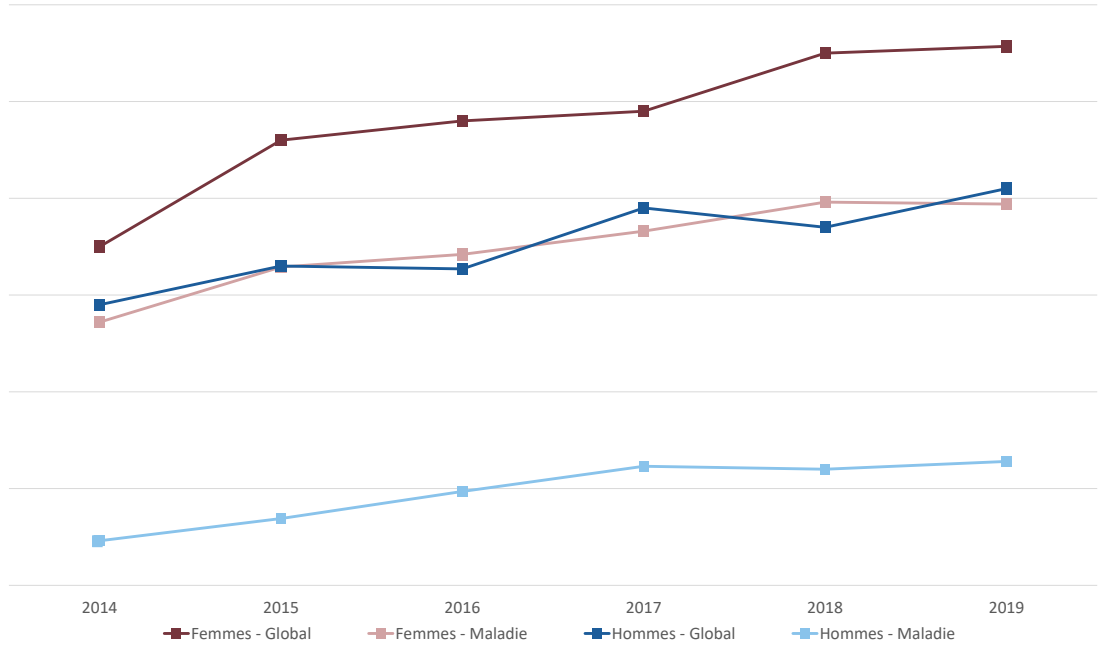

FIGURE 2.8 – Evolution de l'incidence du risque incapacité par année d'observation selon le sexe de l'assuré, au global et pour la cause maladie

La figure 2.8 révèle que le taux d'incidence global est en constante augmentation au cours du temps, pour les deux sexes. En revanche, ce taux augmente considérablement plus vite chez les femmes que chez les hommes.

Cette figure permet également de constater que, pour la cause maladie, l'incidence augmente également constamment chez les deux sexes, mais les taux sont largement supérieurs chez les femmes.

La figure suivante permet d'étudier cette évolution en fonction du sexe pour la cause accident.

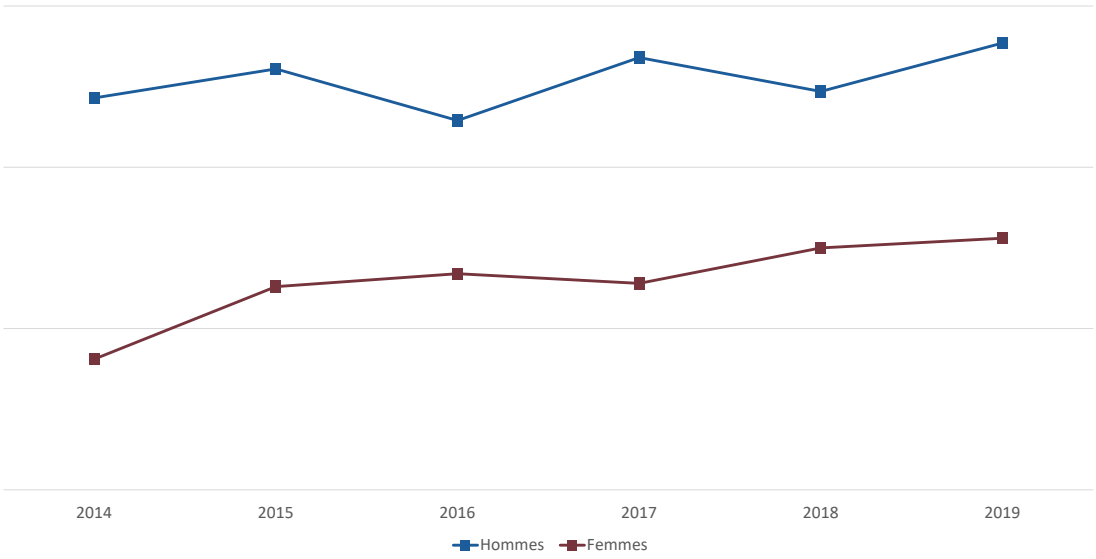

FIGURE 2.9 – Evolution de l'incidence du risque incapacité par année d'observation selon le sexe de l'assuré pour la cause accident

On note, grâce à la figure 2.9 que, pour la cause accident, la tendance est inversée, le taux d'incidence chez les hommes est supérieur à celui chez les femmes. Une possible explication à ce phénomène est que, les professions les plus accidentogènes sont principalement exercées par des hommes. En effet, la Figure 2.5 indique que la proportion d'hommes au sein d'une classe tarifaire augmente de la classe 1 à la classe 2, la classe 2 représentant la classe tarifaire des professions risquées. En revanche, ce taux d'incidence reste plutôt stable au cours du temps de manière globale, mais augmente légèrement chez les femmes.

Avant de poursuivre ces analyses, il serait intéressant d'étudier la proportion de sinistres par cause chez les femmes et chez les hommes.

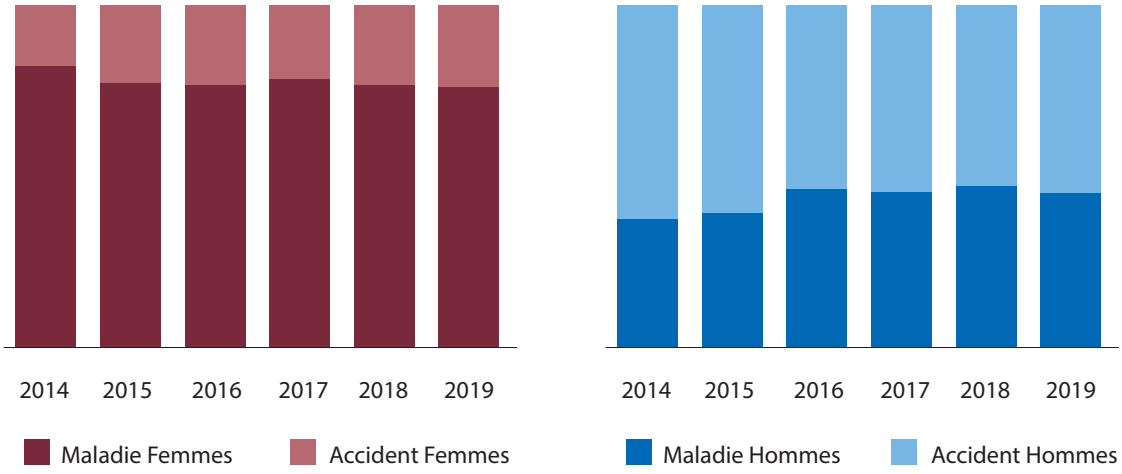

FIGURE 2.10 – Proportion de sinistres par cause chez les femmes et chez les hommes

La figure 2.10 permet de constater que, chez les femmes, une grande proportion des sinistres fait suite à une maladie. Au contraire, chez les hommes, ce ratio est inversé, et plus de la moitié des sinistres fait suite à un accident. Ainsi, à parts égales, les hommes ont plus de sinistres accident et les femmes plus de sinistres maladie. Cette figure permet également de constater que cette répartition est plutôt stable au cours du temps.

Analyse du taux d'incidence par âge

Maintenant que l'évolution du taux d'incidence par cause et par sexe a été étudiée, un paramètre déterminant dans l'explication du comportement vis-à-vis de l'arrêt de travail, reste à étudier. Il s'agit de l'âge. Il a été décidé de calculer des taux d'incidence par tranche d'âge de cinq ans, afin d'avoir des résultats plus robustes et d'éviter d'avoir des taux volatiles à cause d'une faible exposition sur un âge.

De plus, afin de mieux visualiser les résultats, il a été décidé de les agréger par année, après s'être assuré que les répartitions des âges et des taux d'incidence, étaient stables sur la période d'observation.

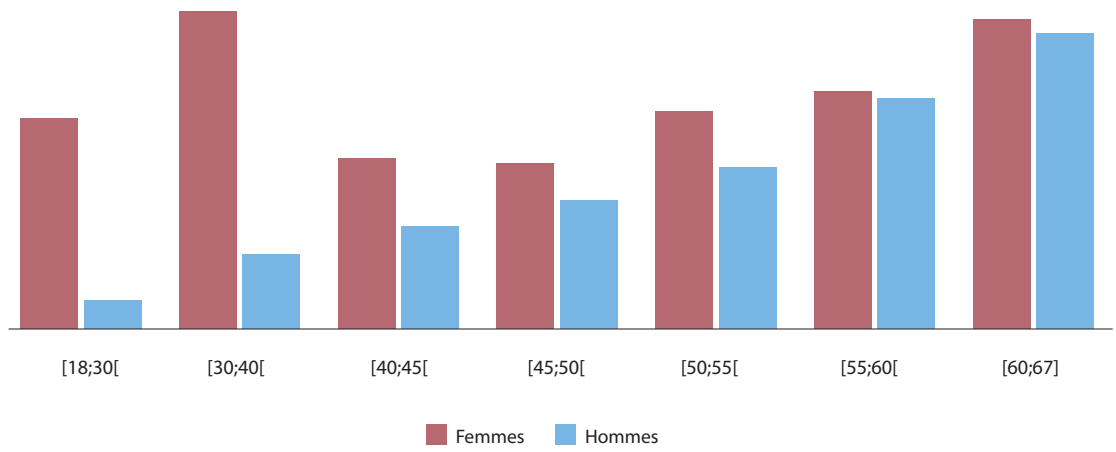

FIGURE 2.11 – Evolution du taux d'incidence par sexe et par tranche d'âge pour la cause maladie

La figure 2.11 met en évidence plusieurs choses. Premièrement, chez les hommes, l'augmentation de l'incidence en maladie semble corrélée positivement à l'âge, puisque, plus l'âge augmente, plus le taux d'incidence augmente également.

En revanche, chez les femmes, il semblerait que l'on puisse diviser le graphe en deux parties. En effet, les taux d'incidence sont très élevés jusqu'à 40 ans, avec un pic pour la tranche d'âge [30 ; 40[ans. Puis, à partir de 40 ans, il semblerait que, comme chez les hommes, l'incidence augmente de nouveau avec l'âge. Ce phénomène pourrait s'expliquer par de nombreux arrêts de travail dans le cadre des grossesses pathologiques chez les femmes avant 40 ans. Cette hypothèse est confirmée par des études de sinistralité internes, dans lesquelles les arrêts de travail sont étudiés au regard des pathologies déclarées.

Enfin, la figure 2.11 permet également de constater que les taux d'incidence chez les femmes et les hommes sont pratiquement les mêmes à partir de 55 ans.

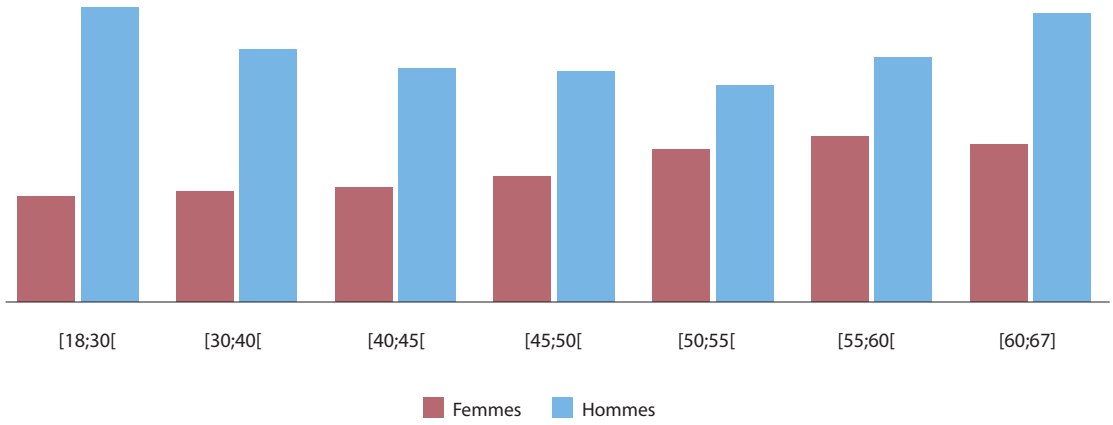

FIGURE 2.12 – Evolution du taux d'incidence par sexe et par tranche d'âge pour la cause accident

En ce qui concerne la cause accident, la figure 2.12 permet d'observer que la distribution n'est pas constante pour chaque âge. Chez les hommes, il semblerait que les âges jeunes et les plus âgés, aient des taux d'incidence plus importants que les âges moyens, et cela peu importe le sexe, mais sans savoir en expliquer la raison. Néanmoins, les taux d'incidence chez les hommes sont toujours supérieurs à ceux observés chez les femmes.

Ces premières analyses descriptives ont permis de mieux cerner les facteurs influant sur le taux d'incidence en arrêt de travail. Pour résumer, il semblerait que l'incidence pour la cause accident soit principalement déterminée par le sexe. En effet, celle-ci est nettement supérieure chez les hommes que chez les femmes. Cependant, comme vu en début de partie, les classes tarifaires regroupant les professions les plus accidentogènes sont principalement souscrites par des hommes. De plus, d'autres analyses de sinistralité internes, ont révélé qu'à profession égale, les taux d'incidence pour la cause accident étaient quasiment identiques chez les hommes et chez les femmes. Il est donc fortement possible que ce soit en fait la CSP qui soit un facteur discriminant pouvant expliquer l'incidence en accident.

En ce qui concerne la cause maladie, on distingue deux tendances différentes, avant et après 40 ans. En effet, avant 40 ans, il semblerait que le facteur le plus discriminant soit le sexe, car l'incidence chez les femmes est nettement supérieure à celle des hommes. Néanmoins, après 40 ans, l'incidence semble être corrélée positivement à l'âge chez les deux sexes, puisque celle-ci augmente avec l'âge.

Analyse du taux d'incidence par classe tarifaire

La refonte des classes tarifaires faisant l'objet de ce mémoire, il est important d'étudier l'évolution du taux d'incidence par classe tarifaire.

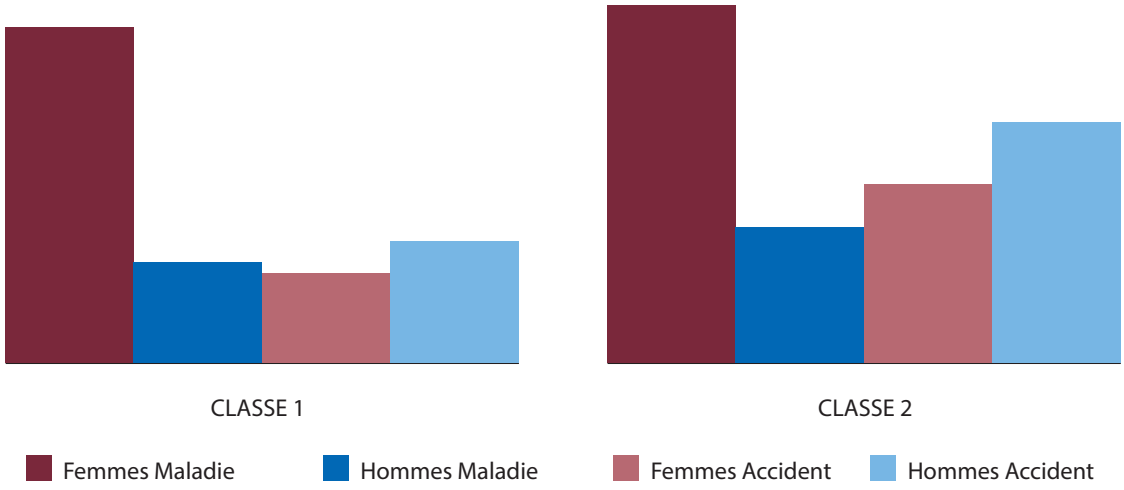

FIGURE 2.13 – Evolution du taux d'incidence par classes tarifaires

La figure 2.13 permet de constater que, l'incidence pour la cause maladie chez les deux sexes semble stable d'une classe tarifaire à l'autre. Cela renforce l'idée selon laquelle l'incidence pour la cause maladie serait en grande partie déterminée par le sexe et l'âge et non la profession.

En revanche, l'incidence pour la cause accident augmente chez les deux sexes, avec la classe tarifaire. Encore une fois, cela va dans le sens de l'hypothèse, selon laquelle l'incidence pour la cause accident, est fortement déterminée par le secteur professionnel. Il est également important de noter que, parmi les professions de la classe 2, on constate que les professions les moins accidentogènes ont une proportion de femmes plus importante, et que les professions les plus accidentogènes sont principalement constituées par des hommes.

D'autres analyses descriptives ont également été effectuées pour tenter de comprendre les différences de taux d'incidence d'une classe tarifaire à l'autre, notamment en analysant les distributions de franchises au sein de chaque classe tarifaire. En effet, la durée de la franchise a un impact considérable sur le taux d'incidence observé. Cependant, la distribution des franchises accident et maladie, est quasiment identique d'une classe à l'autre. Ce n'est donc pas un critère pouvant expliquer les différences de taux d'incidence entre nos classes tarifaires.

2.4 Bilan

Cette analyse descriptive aura permis d'avoir une première idée des facteurs influant sur le taux d'incidence selon la cause. Pour la cause accident, il semblerait que le facteur le plus discriminant soit la profession.

En revanche, pour la cause maladie, l'âge semble être le principal facteur pouvant expliquer une augmentation progressive de l'incidence, avec des taux quasiment similaires chez les hommes et les femmes à partir de 55 ans. Cependant, avant 40 ans, on observe des taux d'incidence beaucoup plus élevés chez les femmes que chez les hommes. Il existe donc, avant 40 ans, également un effet sexe. Le fait que les deux facteurs les plus discriminants, pour la cause maladie, soient l'âge et le

sexe semble tout à fait raisonnable. En effet, historiquement, les tables construites sont segmentées par sexe et sont fonction de l'âge.

Il a également été observé que le taux d'incidence était en constante augmentation pour la cause maladie, mais plutôt stable pour la cause accident. Une hypothèses pouvant expliquer cette augmentation pour la cause maladie, est le vieillissement du portefeuille ainsi que l'éloignement de la sélection médicale. En effet, au moment de la sélection médicale, la vision et la connaissance du portefeuille en termes de risques sont très bonnes. Néanmoins, avec le recul, cette vision se dégrade, et au fur et à mesure que le portefeuille vieillit. Il n'est donc pas étonnant de voir la sinistralité du portefeuille augmenter.

L'objectif de ce mémoire est d'obtenir une segmentation tarifaire plus fine, plus juste et plus adaptée au portefeuille actuel d'Axa France.

Pour cela, il est nécessaire de commencer par confirmer les intuitions faites à partir de l'analyse descriptive. A cette fin, l'étude sera divisée en deux parties. La première sera consacrée à la cause accident, et la deuxième à la cause maladie.

Mais tout d'abord, il est nécessaire de consacrer un court chapitre au modèle de tarification souhaité, afin de pouvoir orienter la suite de l'étude.

PARTIE

3

UNE TARIFICATION ADAPTÉE À L'ESTIMATION DU RISQUE

3.1 Tarification de l'IJ toute cause

Comme expliqué dans le premier chapitre, plusieurs garanties de type indemnités journalières sont proposées par le produit étudié, mais ce mémoire portera uniquement sur la garantie IJ toutes causes. Cette dernière permet le versement d'indemnités journalières des suites d'une maladie ou d'un accident, avec ou sans hospitalisation. L'équation de tarification de cette garantie est donc une somme de la prime pure de l'IJ accident, et de la prime pure de l'IJ maladie.

Aujourd'hui, l'équation de tarification de cette garantie est la suivante :

$$P_x = Pa_x + Pm_x \tag{3.1}$$

Avec :

- Pa_x : la prime pure de l'IJ accident pour un assuré d'âge x ;
- Pm_x : la prime pure de l'IJ maladie pour un assuré d'âge x.

Actuellement, chaque classe tarifaire possède un tarif qui augmente avec l'âge de l'assuré. La classe 1 possède le tarif le moins élevé, et la classe 2 le plus élevé.

Tout l'enjeu de ce mémoire va donc être de trouver une segmentation en classes tarifaires, permettant d'obtenir un tarif pour l'IJ toute cause, adapté à chaque profession. Les nouvelles classes tarifaires devront donc correspondre à des groupes de profession homogènes, à la fois face au risque accident, et au risque maladie.

Avant cela, afin de construire les nouvelles classes tarifaires, il est important de choisir l'approche permettant de calculer la prime pure.

3.2 Différents modèles de tarification

Dans son mémoire, Yann Fournier [8] propose différents modèles de tarification et explique leur impact sur la construction des lois d'entrée en incapacité. Ainsi, avant de choisir le modèle le plus approprié pour cette étude, deux des modèles de tarification envisagés seront présentés.

3.2.1 Tarification de type Assurance Vie

Dans ce modèle, la probabilité d'entrée en incapacité, est mesurée de façon analogue à la probabilité de décès en assurance vie, dans le sens où l'incapacité est vue comme un évènement ne pouvant se produire qu'une seule fois au cours d'une année. Cette approche implique donc de regrouper les sinistres par classes d'âge, sans distinction de leur ordre. On s'intéresse alors la probabilité pour un assuré d'entrer en incapacité sur l'année observée, sans accorder d'importance au nombre de sinistres ayant eu lieu, durant l'année en question. Le modèle de tarification se traduit alors, par l'équation suivante :

$$P_x = p_x.\overline{C_x}$$

Avec :

- P_x : la prime pure pour un assuré d'âge x
- p_x : la probabilité d'entrer au moins une fois en incapacité pour un assuré d'âge x
- $\overline{C_x}$: le coût moyen des sinistres à indemniser, cumulés sur la classe d'âge considérée

Ainsi, avec ce modèle, un assuré ayant eu deux sinistres de 2 mois et 3 mois durant l'année de ses 57 ans, sera traité comme un assuré ayant eu un sinistre de 5 mois. Il est tout à fait possible de créer des tables d'entrée en incapacité pour ce modèle de tarification, cependant, l'inconvénient majeur de cette méthode est le risque de surestimer le maintien en incapacité de nos sinistres, ainsi que leur provisionnement, et de minimiser l'incidence en arrêt de travail. En effet, les provisions des sinistres en cours sont estimées à l'aide des tables de maintien en incapacité. Ainsi, en évaluant les deux sinistres de 2 mois et 3 mois comme un sinistre de 5 mois, l'utilisation des tables de maintien en incapacité ne serait adaptée. Une solution à ce problème, serait de créer une nouvelle table de maintien en incapacité, afin d'avoir deux tables de maintien différentes pour le provisionnement et la tarification, ce qui est assez contraignant et couteux.

3.2.2 Tarification de type Assurance Non-Vie

Une autre approche envisagée est celle de tarification de type non-vie. Cette approche pouvant sembler, au premier abord, inadaptée pour tarifier un risque prévoyance, n'est finalement pas absurde. En effet, l'incapacité peut se répéter plusieurs fois au cours d'une année, chez un même assuré, et présente un caractère qui n'est pas forcément définitif, contrairement au décès. Il est donc possible d'imaginer, que cette dernière peut être assimilée à un risque de type non-vie. Le modèle de tarification classiquement utilisé en assurance non-vie est un modèle fréquence × coût.

Le modèle de tarification serait donc le suivant :

$$P_x = IJ \times E[N_x] \times \sum_{i=fra}^{min(\text{dmaxgar},\ \text{maintien}, fra+\text{maintien}-1, 1095)} \text{maintien}_x(i)$$

Avec :

- P_x : la prime pure pour un assuré d'âge x
- IJ : Le montant de l'indemnité journalière souscrite
- N_x : la variable aléatoire représentant le nombre de sinistres par assuré, survenus à l'âge x
- fra : durée de la franchise
- $dmaxgar$: durée maximale de la garantie, définie contractuellement
- maintien$_x(i)$: la probabilité de rester en incapacité jusqu'au i^{me} jour, pour un assuré d'âge x

Comme cette approche ne prend pas en compte la notion d'ordre entre les sinistres, il est nécessaire de vérifier l'indépendance entre les sinistres. Or, concernant le risque incapacité, ce dernier point n'est pas toujours vérifié. En effet, un assuré ayant été une première fois en incapacité des suites d'une maladie, peut tout à fait faire une rechute et entrer de nouveau en incapacité. Cependant, seulement 1,6% de notre portefeuille étant multisinistré, il est acceptable de prendre une approximation d'indépendance entre les sinistres.

Le choix d'une tarification de type assurance vie, impose la construction de deux tables de maintien en incapacité différentes, pour le provisionnement et la tarification. Il est donc préférable de choisir la méthode de tarification de types non-vie. Ce type de modèle s'inscrivant dans le cadre d'une approche fréquence $\times$ coût, il sera donc nécessaire de choisir un estimateur permettant de quantifier le nombre moyen de sinistres par assuré et par classe d'âge.

Maintenant que le modèle de tarification est établi, la suite de l'étude sera divisée en deux parties. La première partie sera consacrée à l'étude de l'incidence en incapacité pour la cause accident, et la seconde à son étude pour la cause maladie. Il est important de traiter ces deux causes de manières différentes, car comme vu lors de l'analyse descriptive, les facteurs expliquant l'incidence en incapacité ne sont pas les mêmes pour l'accident et pour la maladie.

PARTIE

4

MODÉLISATION DE L'INCIDENCE EN INCAPACITÉ POUR LA CAUSE ACCIDENT

L'objectif de ce chapitre est de confirmer par des modèles, les hypothèses faites lors de l'analyse descriptive. Pour rappel, il avait été émis l'hypothèse selon laquelle, le paramètre expliquant le plus l'incidence de la cause accident, serait la profession exercée. De plus selon cette hypothèse, il n'existerait pas de tendance particulière fonction de l'âge. Afin de vérifier cette hypothèse, les effets de ces différents paramètres sur l'incidence en incapacité seront étudiés à l'aide d'un modèle linéaire généralisé (GLM). La variable CSP est une variable ayant 230 modalités, dont certaines sont très peu exposées. Afin d'étudier sa significativité à l'aide d'un GLM, des groupes de CSP homogènes seront construits à l'aide d'un algorithme de *clustering*.

4.1 *Clustering*

4.1.1 Préparation des données

Avant de réaliser le *clustering* des différentes CSP, il est important de préparer les données. L'objectif est de capturer le risque incapacité pour la cause accident, afin de réaliser des groupes de professions, homogènes pour ce risque. Pour cela, il a été décidé de retenir deux variables pour chacune des CSP :

- L'incidence, sur la période d'observation, pour la cause accident ;
- La proportion des sinistres dus à un accident, sur l'ensemble des sinistres ayant eu lieu durant la période d'observation, pour la CSP étudiée.

Ces deux variables sont liées, mais cela n'est pas un problème, puisqu'elles apportent deux informations différentes, spécifiques au risque accident. Il aurait également été possible de prendre l'incidence globale, ainsi que la proportion de sinistres dus à un accident, puisque l'incidence pour la cause accident peut se déduire de ces deux variables. Cependant, en faisant ainsi, à la fois les risques pour la cause maladie et pour la cause accident, auraient été captés. En effet, les *clusters* avec une incidence élevée, mais une faible proportion de sinistres dus à un accident, correspondent

en fait à des sinistres qui ont une forte incidence et une forte proportion de sinistres dus à une maladie. Or, l'objectif est de créer des groupes homogènes pour la cause accident uniquement.

Avant de pouvoir réaliser ce *clustering*, une première sélection des CSP est effectuée. En effet, sur les 230 modalités de la variable CSP, certaines sont très peu représentées, et 50% du portefeuille est réparti sur 12 CSP uniquement. Ainsi, en sélectionnant une partie des CSP, une probable déformation des résultats avec les CSP les moins souscrites, est évitée. Par exemple, une CSP représentée par un seul contrat et sur laquelle un sinistre a été observé, aura une incidence très élevée, mais pas forcément représentative de la sinistralité qui aurait été observée sur un plus grand échantillon. De plus, avec 230 modalités, les résultats obtenus auraient été plus difficilement interprétables.

Deux critères sont donc fixés :

- Le nombre de sinistres observés sur la CSP doit être au minimum de 10 sinistres par an, en moyenne, sur la période d'observation ;
- Le volume de contrat doit représenter au moins 0,5% du portefeuille.

En respectant ces deux critères, 52 CSP, correspondant à 85% des souscriptions du portefeuille et 90% des sinistres observés, sur la période d'observation, sont ainsi retenues. Cette sélection de CSP permettra d'obtenir des résultats plus fiables, plus robustes et plus facilement interprétables. Une dernière variable, correspondant au poids de chacune des CSP, est également retenue, afin de corriger le biais de représentativité.

De plus, afin de s'assurer que l'échantillon retenu, est bien représentatif du portefeuille, il a été vérifié que la répartition des classes tarifaires dans l'échantillon, était équivalente à celle du portefeuille.

Afin de ne pas introduire de biais dans les données, la répartition des franchises déclenchées lors d'un sinistre, pour chacune des CSP, a également été étudiée. En effet, une CSP pour laquelle une franchise courte est sur-représentée par rapport aux autres CSP peut se traduire par une incidence plus élevée et déformer les résultats. Néanmoins, la répartition des franchises déclenchées est équivalente sur chacune des CSP présentes dans l'échantillon.

4.1.2 Introduction au *clustering*

La classification des données est une première étape importante de l'analyse exploratoire des données.

L'analyse des *clusters* est basée sur la minimisation des distances entre les coordonnées des centroïdes (points imaginaires représentant les centres d'un *cluster*) et les paires de coordonnées de chaque point au sein d'un *cluster*. Il existe plusieurs options pour calculer ces distances. Généralement, on utilise la distance euclidienne, après avoir centré et réduit les données, afin de les ramener à une même échelle.

D'une manière générale, l'analyse en *clusters* a pour but de construire des groupes d'individus, de telle manière que les individus au sein d'un groupe soient les plus similaires possibles entre eux (forte similarité intra-classe), et les plus différents possibles de ceux des autres groupes possible (faible similarité inter-classe).

La partie suivante sera consacrée à l'analyse en *clusters* par l'algorithme K-means.

4.1.3 L'algorithme *K-means*

Le *clustering* K-means est un algorithme permettant de partitionner les données en K *clusters* distincts.

Dans le *clustering* *K-means*, chaque *cluster* est représenté par son centre (i.e., centroïde) qui correspond à la moyenne des points assignés au *cluster*.

Il existe plusieurs algorithmes *K-means*. L'algorithme standard est celui de Hartigan-Wong [1], qui définit la variation totale au sein d'un *cluster* comme la somme des distances euclidiennes au carré entre les éléments et le centroïde correspondant :

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{4.1}$$

Où :

- x_i désigne un point appartenant au *cluster* C_k ;
- μ_k est la valeur moyenne des points assignés au *cluster* C_k.

Il est généralement recommandé de normaliser les variables avant le calcul de la matrice de distance. La normalisation rend les variables comparables, dans la situation où elles sont mesurées à des échelles différentes.

Chaque observation x_i est ensuite assignée à un *cluster* donné, de sorte que la somme des distances carrées de l'observation par rapport aux centres de *cluster* μ_k qui leur sont assignés soit minimale.

La variation totale au sein d'un *cluster* contenant n observations est définie comme suit :

$$V = \sum_{k=1}^{n} W(C_k) = \sum_{k=1}^{n} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

La somme totale des carrés à l'intérieur des *clusters* mesure la compacité (c'est-à-dire la qualité) du *clustering*, qui doit être aussi petite que possible.

L'algorithme *K-means* peut être résumé de la façon suivante :

- Tout d'abord, il faut spécifier le nombre de *clusters* K que l'on souhaite créer. Dans les parties suivantes, il sera expliqué comment déterminer le nombre optimal de *clusters* ;
- Puis, k objets seront ensuite sélectionnés de façon aléatoire, parmi l'ensemble des données, afin de former les centres des *clusters* initiaux ;
- Chacune des observations sera alors affectée à son centroïde le plus proche, en fonction de sa distance avec les centroïdes ;

1. HARTIGAN, John A. et WONG, Manchek A. Algorithm AS 136 : A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 1979, vol. 28, no 1, p. 100-108.

- Ensuite, pour chaque *cluster*, le centroïde sera mis à jour après avoir calculé les nouvelles moyennes de tous les points dans le *cluster*. Le centroïde d'un $K^{\text{ème}}$ *cluster* est un vecteur de longueur p contenant les moyennes de toutes les variables pour les observations du $K^{\text{ème}}$ *cluster*, où p est le nombre de variables ;
- Une fois les centroïdes mis à jour, on regarde de nouveau pour chaque observation quel est le centroïde le plus proche.

Cette étape est répétée de façon itérative jusqu'à ce que les affectations de *clusters* cessent de changer (l'algorithme converge), ou que le nombre maximal d'itérations soit atteint.

4.1.4 L'algorithme *K-means* pondéré

Afin de prendre en compte les différents poids de chaque CSP et de corriger les biais de représentativité, un algorithme K-means pondéré (*Weighted K-means*) sera utilisé. Le schéma ci-dessous permet de donner une interprétation de l'influence des poids sur l'algorithme :

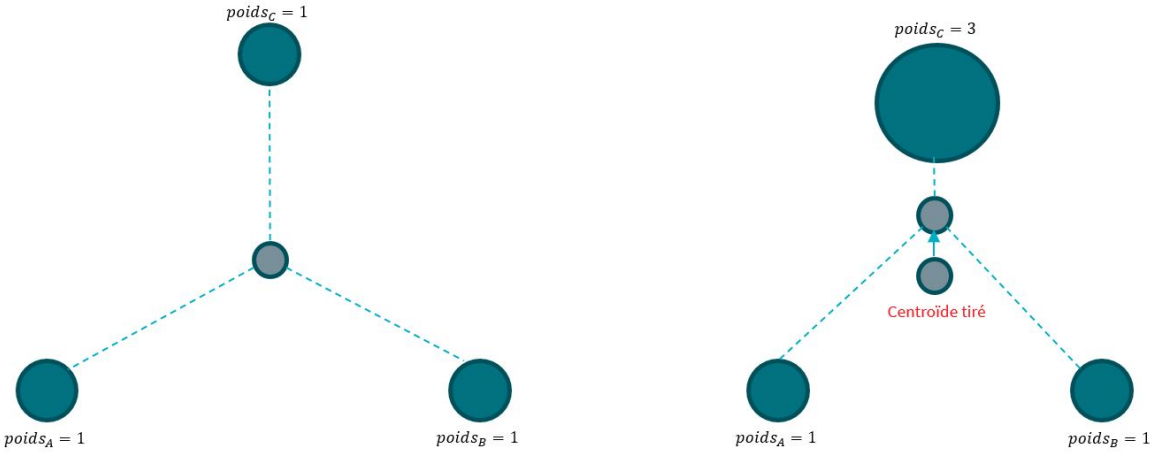

FIGURE 4.1 – Illustration d'un K-means standard (à gauche) contre un K-means pondéré (à droite) pour le calcul des centroïdes

Ainsi, l'équation (4.1) devient :

$$W(C_k) = \frac{\sum_{x_i \in C_k} w_i (x_i - \mu_k)^2}{\sum_i w_i}$$

Où :

- w_i est le poids associé à x_i
- x_i désigne un point appartenant au *cluster* C_k
- μ_k est la valeur moyenne des points assignés au *cluster* C_k

Cet algorithme présente l'avantage d'être simple et rapide, même sur un très grand nombre de données. En revanche, certains aspects limitent son efficacité. En effet, les résultats finaux obtenus sont sensibles à la sélection aléatoire initiale des centroïdes. Ainsi, pour chaque exécution différente de l'algorithme sur le même ensemble de données, les résultats de *clustering* peuvent être différents d'une fois à l'autre. Une solution à ce problème peut être d'exécuter l'algorithme plusieurs fois et de sélectionner le partitionnement pour lequel la somme totale des carrés l'intérieur des *clusters* est la plus petite, ou de sélectionner le partitionnement que l'on obtient le plus souvent.

4.1.5 Détermination du nombre optimal de *clusters*

Trouver le nombre optimal de *cluster* dans un échantillon de données et une question fondamentale dans le *clustering* de partitionnement. Malheureusement, il n'existe pas de réponse unique puisque le nombre de *clusters* optimal est subjectif et dépendant de la méthode utilisée.

Dans l'algorithme *K-means*, le nombre K de *clusters* est un hyperparamètre (c'est-à-dire un nombre prédéfini par l'utilisateur). En pratique, il est possible de sélectionner le K optimal en utilisant, par exemple, la méthode « du coude » [2].

Il s'agit d'une méthode consistant à optimiser un critère, tel que la somme des carrés à l'intérieur d'un groupe. En effet, il faut que la variation au sein d'un *cluster*, totale, soit la plus petite possible. Pour trouver le nombre de *clusters* optimal grâce à cette méthode, on fonctionne de la façon suivante :

1. On effectue un algorithme de *clustering* pour différentes valeurs de k, où k représente le nombre de *clusters* ;
2. Pour chacune des valeurs de k, la somme totale des carrés intra-*clusters* est calculée ;
3. La courbe des sommes totales des carrés intra-*clusters* pour chaque valeur de k, est ensuite tracée sur un graphique ;
4. Puis, on repère visuellement la valeur de k pour laquelle se situe le « coude » de la courbe. Il s'agit de la valeur de k pour laquelle la somme totale de carrés intra-*clusters* chute soudainement.

Pour chaque exécution de l'algorithme *K-means*, le graphe peut différer d'un tirage à l'autre. Cependant, dans le cadre de cette étude, pour chacune des exécutions de l'algorithme, le coude se trouve toujours au niveau de la même valeur. La figure ci-dessous illustre les résultats obtenus pour une des exécutions de l'algorithme.

2. NAINGGOLAN, Rena, PERANGIN-ANGIN, Resianta, SIMARMATA, Emma, *et al.* Improved the performance of the *K-means* cluster using the sum of squared error (SSE) optimized by using the elbow method. In : *Journal of Physics : Conference Series*. IOP Publishing, 2019. p. 012015.

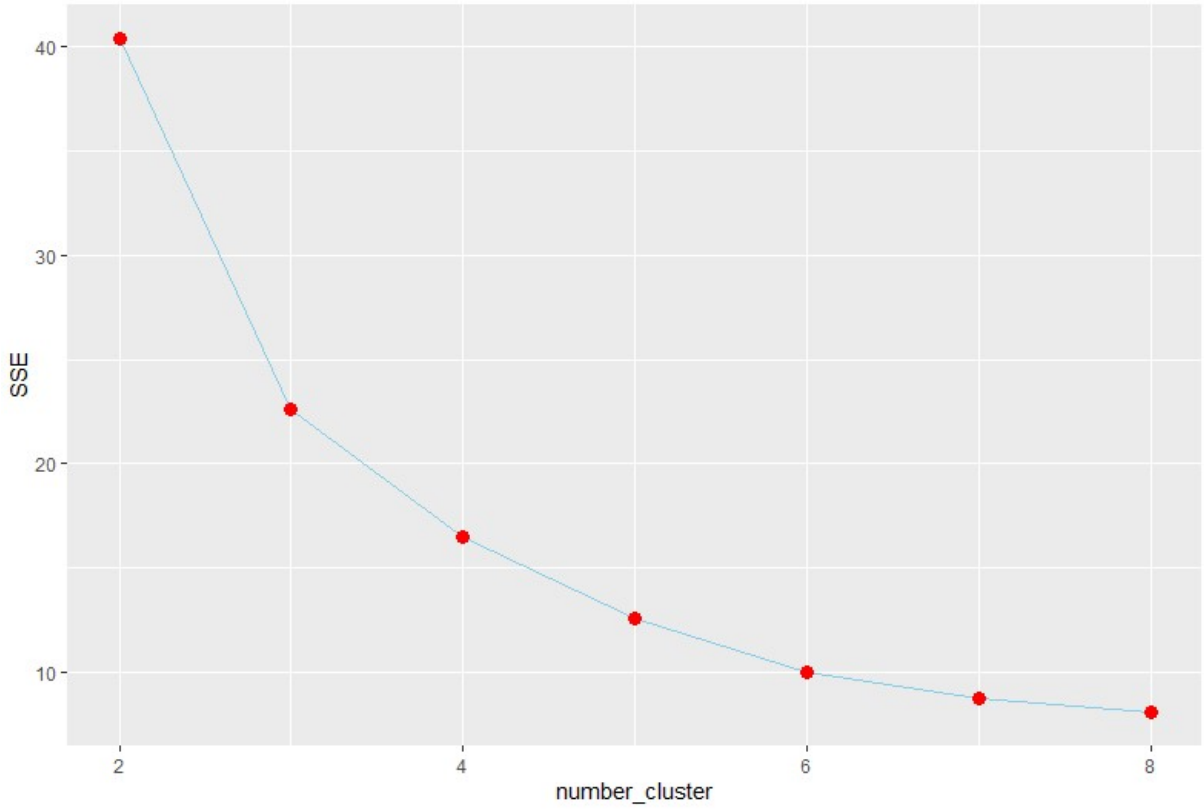

FIGURE 4.2 – Graphe obtenu pour la méthode du coude

La figure 4.2 montre la quantité de variation que chaque dimension peut expliquer.

La question critique dans le *clustering* est de savoir combien de dimensions doivent être modélisées. Si toutes les dimensions sont modélisées, toute la variation peut être expliquée par un grand nombre de *clusters*. Dans ce cas, l'ajustement est alors excessif. Les modèles statistiques avec trop de paramètres sont toujours plus difficiles à interpréter et à expliquer.

L'analyse de la figure 4.2, s'opère en partant du coin supérieur gauche. Le segment entre les points 1 et 2 explique une grande partie de la variation, près de 50%. Ce segment peut donc être considéré comme la première composante principale, ou la première dimension de l'ensemble des données.

Le segment entre les points 2 et 3 explique également une grande partie de la variation, mais moins que celle expliquée par le premier segment. Environ 25% de la variation semble être expliquée par cette deuxième composante principale, ou deuxième dimension.

De même, le segment entre les points 3 et 4 explique environ 13% de la variation globale. Il s'agit de la troisième dimension.

Le segment entre les points 5 et 6 explique beaucoup moins de variation, tout comme les segments

passant par chacun des points successifs. En d'autres termes, modéliser plus de *clusters* devient une sorte de sur-mesure. On peut donc en déduire que les données sont probablement expliquées par 5 *clusters*. L'algorithme a été exécuté une quinzaine de fois, et sur chacune des exécutions, le coude se situe au niveau de la 5e ou de la 6e valeur. Dans la partie suivante, les groupes obtenus, pour un nombre de *clusters* fixé à 5, seront analysés.

4.1.6 Visualisation des *clusters*

Comme mentionné précédemment, l'un des inconvénients de l'algorithme *K-means* est que, pour chaque exécution différente de l'algorithme sur le même ensemble de données, les résultats de *clustering* peuvent être différents d'une fois à l'autre.

Pour pallier à ce problème, l'algorithme a été exécuté une vingtaine de fois. Pour quinze de ces exécutions, les résultats étaient identiques, et pour les cinq autres, seuls les points proches d'une frontières entre deux *clusters*, étaient attribués à un autre centroïde.

Les résultats obtenus pour les quinze exécutions identiques sont représentés dans la figure ci-dessous :

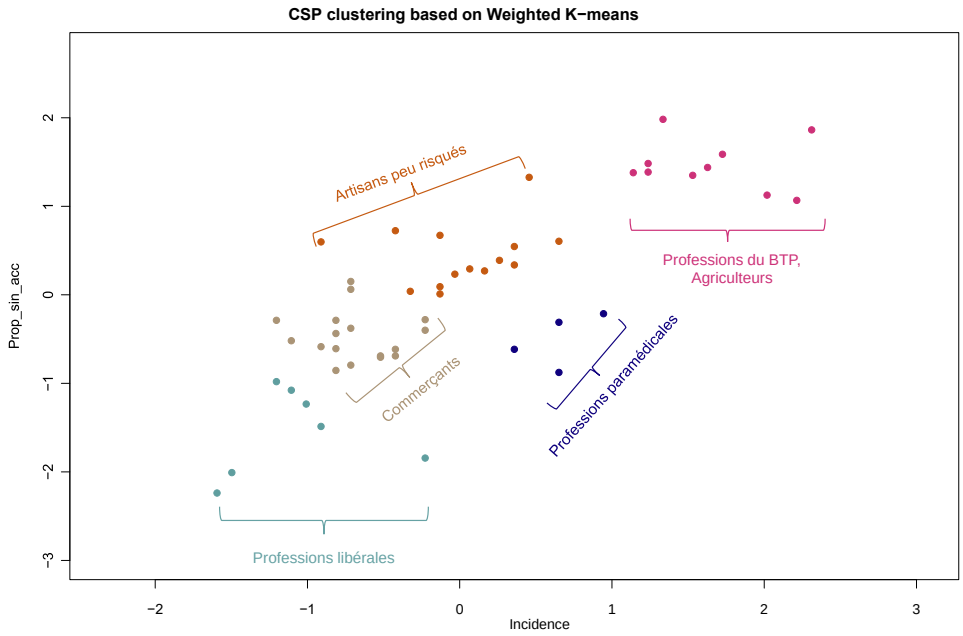

FIGURE 4.3 – Résultats obtenus pour une des exécutions de l'algorithme K-means

Pour des raisons de confidentialité, les libellés des CSP obtenues pour chaque *cluster* ne pourront être dévoilés de manière exhaustive. Néanmoins, les professions majoritaires dans chacun des *clusters* sont reportées sur la figure 4.3. Cette dernière permet de constater que les points de couleur rose, en haut à droite, correspondent aux professions avec une incidence en incapacité pour la cause accident, et une proportion de sinistres accidents, élevées. Il s'agit de professions que considérées comme risquées, et appartenant à la classe tarifaire 2, par exemple, les professions du BTP.

Dans cette étude, il a été choisi d'utiliser l'algorithme *K-means* mais il existe d'autres algorithmes de *clustering*, répartis en deux grandes approches : le *clustering* par partitionnement, et le *clustering* hiérarchique.

Les approches de *clustering* par partitionnement, dont l'algorithme *K-means* fait partie, subdivisent les ensembles de données en un ensemble de k groupes, où k est le nombre de groupes pré-spécifié par l'analyste. Il aurait également pu être envisagé d'utilisé une autre approche, comme l'algorithme *K-meloid*. Il s'agit d'une approche de *clustering* proche de l'algorithme *K-means*, utilisée pour partitionner un ensemble de données en k groupes ou clusters. Contrairement à la méthode *K-means* dans laquelle chaque cluster est représenté par un centre (*i.e.*, centroïde) qui correspond à la moyenne des points assignés au cluster, ici les centres sont représentés par l'un des points du cluster, appelé méloïdes. Les méloïdes sont des objets pour lesquels leur dissimilarité moyenne avec les autres individus du groupe, est minimale. Il correspond au point le plus central du cluster.

La méthode de *clustering* hiérarchique agglomératif, est quant à elle une approche alternative au *clustering* par partitionnement pour identifier des groupes dans un ensemble de données. Cette méthode ne nécessite pas de pré-spécifier le nombre de clusters à générer. Le résultat du *clustering* hiérarchique est une représentation arborescente des objets, également connue sous le nom de dendrogramme.

Dans ce mémoire, il a été choisi d'utiliser uniquement l'algorithme *K-means*, mais le *clustering* des données aurait également pu être réalisé à l'aide d'autres méthodes. Cette approche a été choisie car l'algorithme K-means présente de nombreux avantages, notamment le fait de :

- Garantir la convergence ;
- Se généraliser facilement aux clusters de différentes formes et tailles, tels que les clusters elliptiques ;
- Permettre l'intégration de pondération.

Cependant, il présente également certains inconvénients. En effet, le choix de k, qui correspond au nombre de *clusters* doit se faire manuellement. De plus, les résultats finaux obtenus sont sensibles à la sélection aléatoire initiale des centroïdes. Enfin, un dernier inconvénient de cet algorithme est sa sensibilité aux valeurs aberrantes, mais dans le cadre de ce mémoire, l'impact de ce point a été limité en sélectionnant les CSP dans le portefeuille.

Afin d'approfondir cette étude, il pourrait être intéressant de comparer les résultats obtenus avec les autres algorithmes de clustering cités plus haut.

Dans la partie suivante, un modèle linéaire généralisé a été réalisé, afin de valider les *clusters* précédemment construits. Celui-ci aura pour but de confirmer les hypothèses faites à l'issue de la partie d'analyse descriptive, et de tester la significativité des groupes obtenus par l'algorithme *K-means*.

4.2 Modèle linéaire généralisé

L'objectif cette partie est double. Le premier est de tester la significativité des groupes de CSP, obtenus dans la partie précédente. Le second est d'étudier l'influence des variables explicatives sur l'incidence en incapacité pour la cause accident. Pour cela, une régression de Poisson sera réalisée dans un premier temps, et sera complétée par une régression de Lasso.

4.2.1 Généralités sur les modèles linéaires généralisés

L'approche GLM est un exemple de modélisation de régression, ou procédure d'apprentissage supervisé. Un modèle de régression vise à expliquer certaines caractéristiques d'une variable réponse,

à l'aide de caractéristiques agissant comme des variables explicatives.

Dans un modèle linéaire standard, les observations sont supposées être normalement distribuées autour d'une moyenne qui est une fonction linéaire des paramètres et des covariables. Dans les modèles linéaires généralisés, les variables aléatoires impliquées ne doivent pas nécessairement être normales, et l'échelle dans laquelle les moyennes sont linéaires par rapport aux covariables peut également varier. Par exemple, elle peut être log-linéaire. De plus, les modèles linéaires généralisés peuvent être utilisés lorsque les réponses ne sont pas de type numérique continue. Ils sont principalement utilisés lorsque la variable réponse est une donnée de type comptage, et lorsque les données sont de type binaire.

Un GLM se compose de trois éléments :

- **Un prédicteur linéaire** : le prédicteur linéaire suppose que le GLM va prédire les réponses à partir d'une combinaison linéaire de variables explicatives

$$\eta = \sum_{j=1}^{p} \beta_j X_{ij}$$

- **Une fonction de lien** : cette fonction fournit la relation entre le prédicteur linéaire et la moyenne de la fonction de distribution. Cela signifie que les valeurs du prédicteur linéaire sont obtenues en transformant préalablement les valeurs observées par la fonction de lien.
  On a alors les équations suivantes :

$$\eta = g(\mu_y) = \sum_{j=1}^{p} \beta_j X_{ij}$$

  Par exemple pour les données de comptage on a :

$$log(\mu_y) = \sum_{j=1}^{p} \beta_j X_{ij}$$

  Ainsi, pour obtenir la prédiction moyenne, il est donc nécessaire d'appliquer la fonction de lien inverse :

$$\mu_y = g^{-1}(\eta)$$

- **Une composante aléatoire** : La composante aléatoire est la variable réponse Y, à laquelle une loi de probabilité est associée. Par exemple, les données de comptage, représentées sous forme d'entiers avec une valeur minimale de zéro, ont des erreurs distribuées selon la méthode de Poisson. Les proportions, qui vont de zéro à un, ont des erreurs distribuées de manière binomiale. Les données avec un coefficient de variation constant, où l'écart-type est un multiple de la moyenne, ont des erreurs de distribution gamma. Les données sur le temps jusqu'à la mort, comme l'analyse de survie, ont des erreurs qui suivent une distribution exponentielle.

4.2.2 Régression de Poisson

Généralités

La distribution de Poisson est le candidat naturel pour modéliser le nombre de sinistres déclarés par les assurés. La fréquence moyenne conditionnelle des sinistres, peut alors être écrite comme une fonction exponentielle d'une partition linéaire, dont les coefficients doivent être estimés à partir

des données. Ainsi, si la variable aléatoire discrète Y_i (fréquence des sinistres ou nombre observé de sinistres), conditionnée par le vecteur de variables explicatives X_i (caractéristiques de l'assuré), suit une loi de Poisson, la fonction de densité de probabilité de Y_i est :

$$f(y_i|x_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{4.2}$$

Ainsi, l'équation (4.2) représente la probabilité que la variable aléatoire Y_i prenne la valeur y_i ($y_i \in \mathbb{N}$), compte tenu des caractéristiques des assurés.

Une des principales charactéristiques de la loi de Poisson est l'égalité entre sa moyenne conditionnelle et sa variance. On a alors :

$$\mathbb{E}[y_i|x_i] = \mathbb{V}[y_i|x_i] = \lambda_i$$

Puisque λ doit être un nombre positif, la fonction logarithme est utilisée comme fonction de lien avec le prédicteur linéaire :

$$log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

où p correspond au nombre de variables.

En inversant la fonction logarithme, on obtient une relation exponentielle entre la réponse moyenne λ et les prédicteurs :

$$\hat{y_i} = \lambda_i = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} \tag{4.3}$$

$$\hat{y_i} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} ...$$

Les paramètres du modèle sont ensuite estimés par la méthode du maximum de vraisemblance. Afin de trouver le maximum de vraisemblance de l'équation (4.2), la vraisemblance est définie de la façon suivante :

$$log\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x_i\beta}}(e^{x_i\beta})^{y_i}}{y_i!}$$

Puis, en appliquant la fonction logarithme à l'équation précédente, on obtient la fonction de log-vraisemblance :

$$log\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i log(\lambda_i) - \lambda_i - log(y_i!)$$

ou encore

$$log\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i x_i \beta - exp(x_i\beta) - log(y_i!)$$

Lorsque la variable à expliquer dans le cas d'un modèle linéaire généralisé, dépend également linéairement d'une autre variable, ce qui est le cas avec la variable exposition, cette dernière est déclarée en offset. En effet, pour prendre en compte l'exposition dans le modèle, on voudrait un modèle de la forme :

$$y_i \sim \mathcal{P}(\lambda_i.E_i) \text{ avec } \lambda_i = exp(\beta_0 + \beta_1 x_{1,i})$$

ou encore

$$y_i \sim \mathcal{P}(\widetilde{\lambda_i}) \text{ avec } \widetilde{\lambda_i} = E_i.exp(\beta_0 + \beta_1 x_{1,i}) = exp(\beta_0 + \beta_1 x_{1,i} + log(E_i))$$

L'exposition intervient alors comme une variable de la régression, avec un coefficient fixé à 1.

Conditions de validité d'un GLM Poisson

Les conditions de validité d'un modèle de régression de Poisson sont les suivantes :

- Il doit y avoir indépendance entre les réponses ;
- Les réponses doivent être dispersées selon une loi de Poisson de paramètre λ ;
- Il ne doit pas y avoir de surdispersion.

Les deux premières conditions de validité du modèle seront détaillées plus loin, lors de l'application de la régression de Poisson aux données.

En ce qui concerne l'absence de surdispersion, il est important d'avoir à l'esprit qu'une des principales propriétés d'un GLM Poisson est l'égalité de la variance des réponses à la moyenne. On parle alors de surdispersion lorsque la variance réelle est supérieure à la variance théorique.

Pour détecter la présence d'une éventuelle surdispersion, il est possible d'utiliser une statistique du χ^2. Cette statistique est calculée à partir des écarts au carré, entre les valeurs observées y et les valeurs attendues $\hat{y}$, normalisés par la valeur attendue, pour chacun des n points du jeu de données. Si les données suivent effectivement une distribution de Poisson, alors la valeur moyenne du χ^2 est égale au nombre de degrés de liberté résiduels du modèle (ddl $= n - p$ où p est le nombre de paramètres estimés).

Le paramètre ϕ permet d'estimer si la variance réelle des observations n'est pas trop éloignée de la variance théorique. En pratique, ϕ est estimé par le ratio de la déviance résiduelle sur le nombre de degrés de liberté du modèle :

$$\hat{\phi} = \frac{\text{déviance résiduelle}}{\text{ddl}}$$

Lorsque ϕ est supérieur à 1, il y a alors surdispersion. Dans ce cas, il est nécessaire d'utiliser une autre structure d'erreur, telles que les structures Poisson-mélange.

Analyse de la déviance pour mesurer la qualité de l'ajustement

Comme pour les modèles linéaires, une fois que les paramètres ont été estimés, il est important de vérifier que le modèle reflète bien la réalité et qu'il est valide. Cependant, dans le cas des modèles linéaires généralisés, on a peu d'information sur les résidus. En particulier, ils n'ont aucune raison d'avoir la même variance et on ne connaît pas leur distribution.

Lorsque l'on travaille avec des GLM, il est donc utile de disposer d'une quantité qui peut être interprétée de manière similaire à la somme résiduelle des carrés, dans la modélisation linéaire ordinaire.

Formellement, la déviance est définie par la différence des log-vraisemblances entre le modèle ajusté, $l(\hat{\beta})$ et le modèle saturé l_s. Le modèle saturé correspond au modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. Si l'on utilise la fonction de liaison canonique on a $\theta_i = g(Y_i)$, et la déviance est définie de la façon suivante :

$$D = -2(l(\hat{\beta} - l_s))\phi$$

Où ϕ est le paramètre de surdispersion, qui vaut 1 dans le cas d'une régression de Poisson.

La log-vraisemblance $l(\hat{\beta})$ étant toujours plus petite que l_s, la déviance est toujours supérieure ou égale à zéro, n'étant nulle que si l'ajustement du modèle est parfait.

Si la fonction de lien canonique est employée, la déviance peut être exprimée comme suit :

$$D = -\frac{2}{a(\phi)} \sum_{i=1}^{p} (Y_i \hat{\theta}_i - b(\hat{\theta}_i) - Y_i g(Y_i) + b(g(Y_i)))\phi$$

$$D = -\frac{2\phi}{a(\phi)} \sum_{i=1}^{p} (Y_i(Y_i - \hat{\theta}_i) - b(g(Y_i)) + b(\hat{\theta}_i)) \tag{4.4}$$

Dans la plupart des cas, $a(\phi) \propto \phi$ donc la déviance ne dépend pas de ϕ.

Il s'agit donc d'une mesure de la distance entre un modèle particulier et les données observées définies au moyen du modèle saturé. De la même manière que la somme des erreurs quadratiques, elle quantifie les variations des données qui ne sont pas expliquées par le modèle considéré.

Comme le modèle saturé doit s'adapter aux données au moins aussi bien que tout autre modèle, la déviance résiduelle n'est jamais négative. Plus la déviance est grande, plus les différences entre les données réelles et les valeurs ajustées sont importantes.

Une référence pour évaluer l'ampleur de la déviance est la déviance nulle :

$$D_0 = -2[l(\hat{\beta}_0 - l_s)]\phi$$

Il s'agit de la déviance du modèle sans prédicteurs, le modèle ne comportant qu'un intercept, par rapport au modèle parfait.

L'utilisation de l'équation (4.4) permet de constater que la déviance nulle est une généralisation de la somme totale des carrés du modèle linéaire :

$$D_0 = \sum_{i=1}^{p} (Y_i - \hat{\eta}_i)^2 = \sum_{i=1}^{p} (Y_i - \hat{\beta}_0)^2 = \text{SST}$$

Puisqu'il n'y a pas de prédicteurs, $\hat{\beta}_0 = \hat{Y}$.

Ces notions peuvent être résumées dans le schéma suivant :

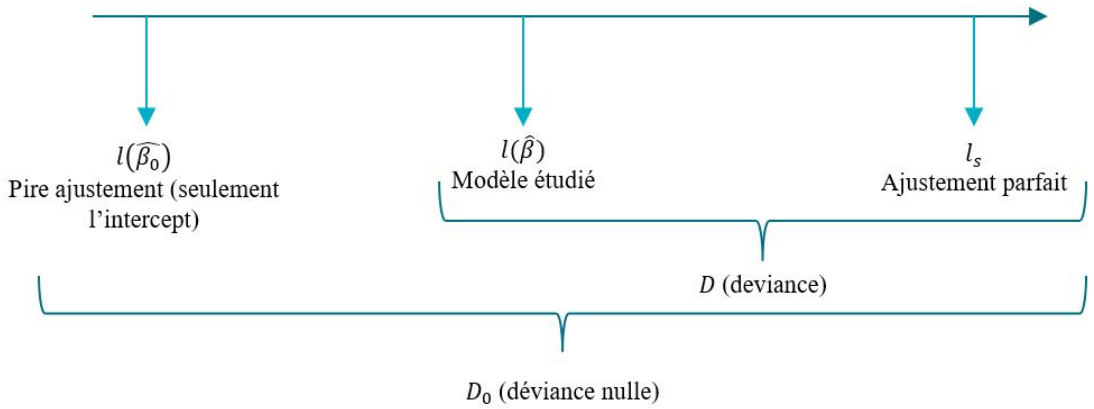

FIGURE 4.4 – Représentation de la déviance et de la déviance nulle

En utilisant la déviance et la déviance nulle, il est possible de mesure de combien le modèle s'est amélioré en ajoutant les prédicteurs $X_1, \ldots, X_p$, et quantifier le pourcentage de déviance expliquée. Ceci peut être fait au moyen de la méthode R^2 qui est une généralisation du coefficient de détermination pour la régression linéaire :

$$R^2 = 1 - \frac{D}{D_0}$$

Le R^2 des modèles linéaires généralisés est une mesure qui partage la même philosophie que le coefficient de détermination de la régression linéaire : il s'agit d'une proportion de la qualité de l'ajustement du modèle. S'il est parfait, $D = 0$ et $R^2 = 1$.

4.2.3 Sélection de modèles et de variables

Le critère AIC

L'idée derrière le critère d'information d'Akaike (AIC) est d'examiner la complexité du modèle ainsi que la qualité de son ajustement aux données de l'échantillon, et de produire une mesure qui équilibre les deux. Un modèle avec de nombreux paramètres fournira un très bon ajustement aux données, mais aura peu de degrés de liberté et sera d'une utilité limitée. Cette approche équilibrée décourage l'ajustement excessif et encourage la parcimonie. Le modèle préféré est celui qui présente la valeur AIC la plus faible. L'AIC est la valeur négative de deux fois la log-vraisemblance plus deux fois le nombre de paramètres linéaires et d'échelle, c'est à dire :

$$AIC = -2l + 2k$$

Où k est le nombre de paramètres, et l est la logvraisemblance.

L'AIC étant une pondération entre la vraisemblance statistiques du modèle, et son nombre de paramètres, il se doit d'être le plus petit possible.

Méthodes *stepwise*

La méthode *stepwise*, procédure itérative classique permettant d'élaguer les variables d'un modèle les moins pertinentes, peut prendre plusieurs formes.

Dans la méthode ascendante (*forward*), on part d'un modèle avec seulement la constante, et, à chaque étape, on ajoute la variable qui conduit à la plus forte baisse de l'AIC. Ainsi, à chaque itération, l'algorithme sélectionne la variable la plus pertinente, et s'arrête une fois que le pouvoir explicatif du modèle n'augmente plus.

La méthode descendante (*backward*), correspond à la même chose dans le sens inverse. L'algorithme part d'un modèle avec toutes les variables et effectue des suppressions à chaque itération.

La régression de Lasso

Il s'agit d'une méthode puissante qui effectue deux tâches principales : la régularisation et la sélection des variables. En effet, lorsque de multiples variables sont présentes dans un modèle de régression logistique, il peut être utile de trouver un ensemble réduit de variables résultant en un modèle à performance optimale. La régression logistique pénalisée impose une pénalité au modèle logistique pour avoir trop de variables. Cela a pour conséquence de réduire à zéro les coefficients des variables les moins contributives. Ceci est également connu sous le nom de régularisation.

Puis, pendant le processus de sélection des caractéristiques, les variables qui ont encore un coefficient non nul après le processus de réduction sont sélectionnées pour faire partie du modèle. L'objectif de ce processus est de minimiser l'erreur de prédiction.

La méthode de Lasso permet donc de palier à un des inconvénients majeurs des modèles de régression linéaire ou logistique, qui est le caractère très erratique de l'estimation des paramètres β, dû à une variance élevée.

Pour rappel, on évalue la qualité de prédiction du modèle, en mesurant à l'aide de l'espérance, l'écart au carré entre la prédiction $\hat{y}$ et la vraie valeur de y. On a alors :

$$\mathbb{E}[(y - \hat{y})^2] = \sigma^2 + \text{Biais}^2 + \text{Variance}$$

L'idée est donc d'améliorer la qualité de prédiction, en acceptant une légère augmentation du biais, mais en diminuant significativement la variance. En effet, accroitre la complexité d'un modèle va en général accroitre sa variance et réduire son biais. Donc, pour réduire la complexité du modèle, il est nécessaire de réduire sa variance et accepter une augmentation du biais. Pour se faire, n des contraintes seront imposées aux paramètres estimés (les β_i), afin de maitriser leur amplitude.
Cela est particulièrement utile lorsque les variables sont très corrélées, ce qui fausse souvent la résolution numérique.

Pour cela, on écrit la régression de Lasso de la manière suivante :

$$\min_{\beta_1,\ldots,\beta_p} \sum_{i=1}^{p} (y_i - \sum_{j=1}^{p} \beta_j z_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Où

- $\lambda (\lambda \geq 0)$ est le coefficient de pénaliter à fixer, permettant de contrôler l'impact de la pénalité ;

- $\sum_{j=1}^{p} |\beta_j|$ est la fonction de pénalité ;
- Les z_{ij} sont les x_{ij} centrés réduits, permettant de limiter l'influence des variables à trop forte variance.

En pratique, le paramètre d'ajustement λ, qui contrôle la force de la pénalité, prend une grande importance. En effet, lorsque λ est suffisamment grand, les coefficients sont forcés d'être exactement égaux à zéro, ce qui permet de réduire la dimensionnalité. Plus le paramètre λ est grand, plus le nombre de coefficients réduits à zéro est important. D'autre part, si $\lambda = 0$, il s'agit alors d'une régression MCO (moindres carrés ordinaires).

Il est possible d'obtenir le paramètre d'ajustement λ par validation croisée. Cette technique consiste à diviser la base d'apprentissage en k blocs, puis on sélectionne un des k blocs comme ensemble de validation pendant que les k autres échantillons constituent l'ensemble d'apprentissage. On répète ensuite l'opération en sélectionnant un autre échantillon de validation parmi les blocs prédéfinis. À l'issue de la procédure k scores de performances, un par bloc, sont ainsi obtenus. La moyenne et l'écart type des k scores de performances peuvent être calculés pour estimer le biais et la variance de la performance de validation.

L'un des inconvénients majeurs de la régression de Lasso, face à la régression de Ridge (équivalente à une régression de Lasso en norme $L2$) est que, parmi un groupe de variables corrélées, la régression de Lasso en choisit une seule, celle qui est la plus liée à la variable cible, et masque l'influence des autres. Néanmoins, cet inconvénient n'en est pas un dans le cadre de cette étude, puisque une forte corrélation entre les variables CSP et sexe, est déjà suspectée. La régression Lasso pourra alors permettre d'identifier, laquelle de ces deux variables a le plus d'influence sur la variable réponse. Ainsi, avant de passer à l'application de ces modèles, la partie suivante aura pour but d'étudier les corrélations entre les différentes variables explicatives.

4.3 Etude de la corrélation entre les variables

Avant de passer à la phase d'application, il est important de réaliser une étude des corrélations entre les variables. En effet, une corrélation trop importante entre deux variables pourrait biaiser le modèle de fréquence et le complexifier inutilement. Pour se faire, le V de Cramer, qui permet d'affecter une corrélation entre deux variables qualitatives, en se basant sur le test d'indépendance du χ^2, sera utilisé.

Le χ^2 indique qu'il existe une relation significative entre les variables, mais il ne dit pas à quel point cette relation est significative et importante. Le V de Cramer donne cette information supplémentaire, en normalisant la valeur du χ^2.

Soit X et Y deux variables catégorielles, prenant leurs valeurs respectivement dans $a_1, \ldots, a_I$ et $b_1, \ldots, b_J$, sur n observations. Pour une combinaison possible des valeurs (a_i, b_j), on note n_{ij} le nombre de fois où cette combinaison est retrouvée dans l'échantillon. On a alors $n = \sum_{i,j} n_{ij}$. Les sommes des lignes et des colonnes sont définies de la façon suivante :

$$n_{i.} = \sum_{j=1}^{J} n_{ij} \text{ et } n_{.j} = \sum_{i=1}^{I} n_{ij}$$

Puis, avec ces notations, la statistique du χ^2 est définie comme suit :

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.} n_{j.}}{n})^2}{\frac{n_{i.} n_{j.}}{n}}$$

Enfin, la statistique du V de Cramer est simplement une transformation de la statistique du χ^2 :

$$V = \sqrt{\frac{\chi^2/n}{\text{min(I,J)-1}}}$$

Le V de Cramer varie entre 0 et 1. Proche de 0, il indique une faible association entre les variables. Proche de 1, il indique une forte association.

En calculant les statistiques V de Cramer pour chacune des combinaisons de variables, on obtient les résultats suivants :

	Franchise	**Cluster de CSP**	**Sexe**	**Âge**
Franchise	1	0,06	0,08	0,04
Cluster de CSP	0,06	1	**0,85**	0,09
Sexe	0,08	**0,85**	1	0,07
Âge	0,04	0,09	0,07	1

Rappelons que l'objectif de cette analyse est d'identifier les variables explicatives les plus significatives pour prédire la variable réponse, c'est à dire le nombre de sinistres accidents. Cependant, comme le montre le tableau ci-dessus, deux variables sont fortement corrélées entre elles. Il s'agit des variables sexe et *cluster* de CSP. Cette corrélation entre ces deux variables n'est pas étonnante. En effet, les professions les plus accidentogènes sont principalement exercées par les hommes. Ce point avait été relevé lors de l'analyse descriptive (Figure 2.5), qui montrait que la classe tarifaire la plus risquée, était composée à 95% par des hommes. Les modèles de régression de Poisson et de Lasso auront donc pour objectif d'identifier la variable la plus discriminante, entre ces deux variables, en termes d'entrée en incapacité, afin de valider l'hypothèse faite lors de la partie d'analyse descriptive.

4.4 Application de la régression

Avant de passer à la phase d'application, il est important de préciser l'objectif de cette analyse. En effet, contrairement à une tarification de type non-vie, seules peu de variables sont ici testées dans le modèle. De plus, le but n'est pas de revoir entièrement le modèle de tarification, mais plutôt de revoir la segmentation des classes tarifaires et s'assurer de la qualité des groupes de CSP créés. Ainsi, peu d'importance sera accordée au pouvoir prédictif du modèle, mais l'objectif sera de :

- S'assurer que la variable qui a le plus d'influence sur l'incidence en incapacité est la profession exercée, et non le sexe de l'individu ;
- S'assurer que chacun des *clusters* est bien significatif ;
- S'assurer que, comme conjecturé à la suite de la phase de statistiques descriptives, l'âge n'est pas un facteur explicatif de l'incidence en incapacité pour la cause accident ;
- Récupérer les taux d'incidence propres à chacun des clusters obtenus par l'algorithme *K-means*.

4.4.1 Préparation des données

La régression de Poisson a pour objectif de modéliser la fréquence des sinistres pour la cause accident. Mais avant tout, il est nécessaire de préparer les données. Pour rappel, l'objectif est d'identifier les variables qui ont le plus d'impact sur l'incidence en incapacité pour la cause accident, et de tester la significativité des *clusters*, précédemment créés.

Pour rappel, dans la base de données, les données sont organisées avec les informations suivantes sur chaque ligne :

- Numéro de contrat ;
- Sexe ;
- CSP ;
- Numéro du *cluster* de CSP ;
- Age_i : correspond aux âges sur lequel l'assuré observé est exposé ;
- Expo_i : correspond à l'exposition de l'assuré observé sur l'âge i ;
- Franchise ;
- Date survenance_i : correspond à la date de survenance du i^{me} sinistre ;
- Cause_i : correspond à la cause du i^{me} sinistre.

Ces notions sont illustrées dans l'exemple suivant. Un assuré, de sexe masculin, né le 01/01/1975 a souscrit un contrat de prévoyance le 01/06/2015 (l'année de ses 40 ans sur laquelle il est exposé 6 mois), et l'a résilié le 01/09/2019 (l'année de ses 44 ans, sur laquelle il est exposé 9 mois).

La franchise associée au contrat est une franchise 3/30 (3 jours en cas d'accident, et 30 jours en cas de maladie).

Sur cette période, cet assuré a eu deux sinistres des suites d'un accident : un premier le 01/03/2016 (l'année de ses 41 ans) et un le 01/08/2017 (l'année de ses 42 ans).

Les périodes d'exposition sont calculées seulement à partir des dates de début et de fin de contrat. Ainsi, dans cet exemple, on considère que cet assuré n'est pas totalement exposé sur les âges des années de souscription et de résiliation, mais qu'il est entièrement exposé entre ces deux dates. Il est important de souligner ici que, la définition de l'exposition, en ce qui concerne l'incapacité n'est pas exactement la même qu'en assurance non-vie. En effet lorsqu'un assuré est en incapacité, celui-ci n'est dès lors plus exposé au risque. Cependant, cette décision a été prise afin d'être en phase avec les hypothèses utilisées pour construire les lois d'incidence dans la partie suivante. Par conséquent, l'exposition du portefeuille est 1,4% plus élevée que si ce périodes d'incapacité avaient été retirées. Ainsi, si une étude de rentabilité devait être menée, cette sur-exposition au risque devrait être prise en compte.

Pour se ramener au cadre de la régression de Poisson, il est nécessaire de se ramener à un vecteur d'observations supposé suivre une loi de Poisson. Pour cela, il faut donc agréger les données, car pour un seul individu, l'approximation de la loi de Poisson n'est pas pertinente.

Le but est donc d'obtenir un vecteur dont la taille est le nombre d'âge présents dans le jeu de données.

En reprenant l'exemple précédent, la matrice suivante est obtenue :

Sexe	Groupe de CSP	Franchise	Âge	Exposition	Nombre de sinistres accident
Homme	Cluster 4	3/30	40	0,5	0
Homme	Cluster 4	3/30	41	1	1
Homme	Cluster 4	3/30	42	1	1
Homme	Cluster 4	3/30	43	1	0
Homme	Cluster 4	3/30	44	0,75	0

Cette opération est itéré pour chaque assuré. On obtient alors une matrice de 175 000 lignes, et 6 colonnes.

Dans ce modèle, on notera donc :

- Y_i : le nombre de sinistres pour la ligne i ;
- $X_{i,1}$: le sexe de l'assuré sur la ligne i (Variable factorielle à 2 modalités) ;
- $X_{i,2}$: le cluster de CSP de l'assuré sur la ligne i (Variable factorielle à 5 modalités) ;
- $X_{i,3}$: la franchise souscrite par l'assuré sur la ligne i (Variable factorielle à 7 modalités) ;
- $X_{i,4}$: l'âge de l'assuré sur la ligne i (Variable numérique allant de 18 à 67 ans) ;
- E_i : l'exposition de l'assuré sur la ligne i (Variable numérique allant de 0 à 1).

4.4.2 Vérification des hypothèses de validité du GLM de Poisson

L'objectif est à présent de vérifier que les hypothèses de validité du GLM sont bien vérifiées.

Indépendance entre les réponses

La première hypothèse à vérifier est l'indépendance entre les réponses. Cependant, il est important de souligner le fait que, pour le risque incapacité, un sinistre peut tout à fait entrainer une rechute quelques mois ou années plus tard. Néanmoins, cette situation se rencontre principalement pour la cause maladie. Or, l'objectif ici est de modéliser l'incidence pour la cause accident. Supposer l'indépendance entre les réponses dans ce genre de situation, est une hypothèse assez naturelle, sauf dans le cadre des modèles de fragilité. Il est donc possible de supposer que cette hypothèse d'indépendance entre les réponses est vérifiée dans le cadre de cette étude.

Distribution des réponses selon une loi de Poisson

A présent, en ce qui concerne la distribution de la variable réponse selon une loi de Poisson de paramètre $\lambda_i.e_i$, on obtient sur R le graphique suivant :

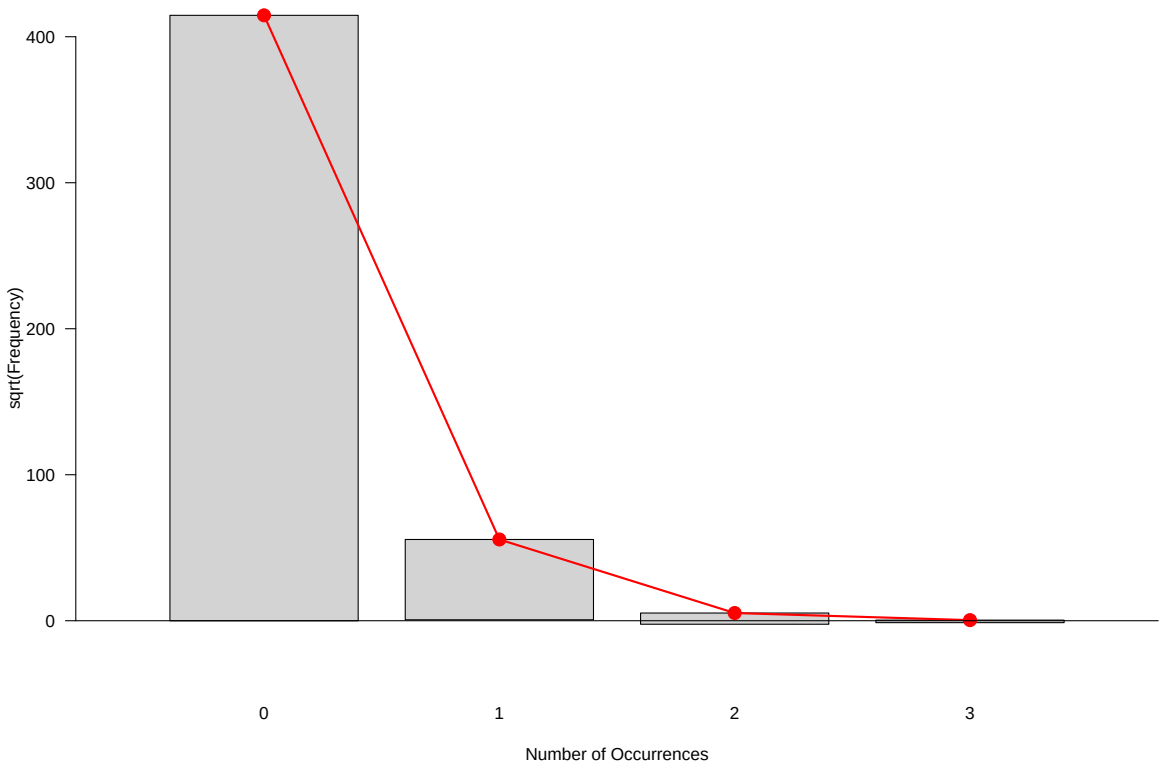

FIGURE 4.5 – Modélisation de la fréquence - globale - des sinistres par une loi de Poisson

La figure 4.5 compare les fréquences empiriques, en points rouges, aux fréquences théoriques obtenues avec une loi de Poisson, en bâton. Cette figure permet donc de valider la deuxième hypothèse de validité du modèle, à savoir, selon laquelle le nombre de sinistres pour la cause accident suit une loi de Poisson de paramètre $\lambda_i.e_i$.

On a alors, $\in [1; 175000]$:

$$Y_i \sim \mathcal{P}(E_i.exp(X_i'\beta))$$

ou de façon équivalente :

$$Y_i \sim \mathcal{P}(exp(log(E_i) + X_i'\beta)$$

Absence de surdispersion

Comme cela a été expliqué précédemment, afin de s'assurer de l'absence de surdispersion dans les données, il est essentiel de vérifier que :

$$\hat{\phi} = \frac{\text{déviance résiduelle}}{\text{ddl}} < 1$$

On obtient, pour le modèle contenant toutes les variables, un ratio de 0.13, donc largement inférieur à 1. Cette absence de surdispersion permet d'utiliser une modèle de Poisson, et non un modèle quasi-Poisson, comme cela aurait été le cas si une surdispersion était présente dans les données.

4.4.3 Sélection de variable

Sélection à l'aide de la p-value et de la méthode *stepwise*

Pour voir quelles variables explicatives ont un effet sur la variable de réponse, il est important d'analyser les p-values. Si la p-value est inférieure à un seuil α, alors la variable explicative sur laquelle est calculée la p-value a un effet sur la variable de réponse. Dans la litérature, α est souvent égal à 5%. Ce seuil est donc choisi dans cette étude. En effectuant la régression de Poisson sur toutes les variables explicatives, les résultats suivants sont obtenus :

```
Call:
glm(formula = Y ~ Age + Sexe + Cluster + Franchise + offset(log(Expo)),
    family = poisson, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3064  -0.2241  -0.1669  -0.1213   4.7845

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.134049 -29.028  < 2e-16 ***
Age                        0.002082   0.939 0.347785
SexeM                      0.059681   4.869 1.12e-06 ***
Cluster2                   0.075503  -3.556 0.000376 ***
Cluster3                   0.053072 -10.875  < 2e-16 ***
Cluster4                   0.067904 -14.697  < 2e-16 ***
Cluster5                   0.090073 -12.751  < 2e-16 ***
Franchise180/180          87.827726  -0.114 0.909499
Franchise3/3               0.081930   4.990 6.04e-07 ***
Franchise30/30             0.110617  -4.970 6.70e-07 ***
Franchise365/365           1.003150  -3.046 0.002322 **
Franchise60/60             0.386216  -1.495 0.134820
Franchise90/90             0.415944  -3.746 0.000180 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 19926  on 140055  degrees of freedom
Residual deviance: 19018  on 140043  degrees of freedom
AIC: 24032

Number of Fisher Scoring iterations: 12
```

FIGURE 4.6 – Résultats R de la régression de Poisson effectuée sur toutes les variables explicatives

Les résultats dans la figure 4.6 permettent de déduire qu'à l'exception de la variable âge, et de certaines modalités de la variable franchise, toutes les variables ont une p-value inférieure à 5%, ce qui signifie qu'elles ont un effet significatif sur le nombre de sinistres.

Ce premier modèle va dans le sens dans l'analyse descriptive précédente, à savoir que l'âge n'est pas une variable qui a un effet significatif sur le nombre de sinistres. Il est également possible de

visualiser cela à l'aide du graphe suivant, qui représente l'incidence moyenne en fonction de l'âge :

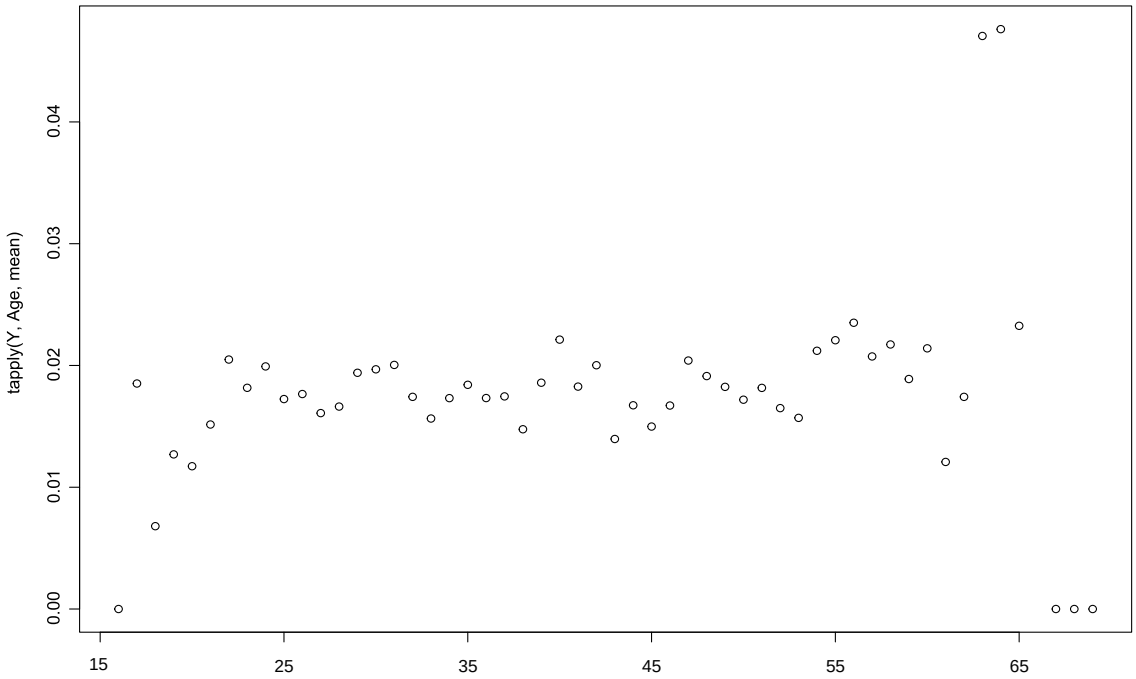

FIGURE 4.7 – Incidence moyenne en fonction de l'âge

En appliquant la méthode backward, on retient également un modèle ou seules les variables Sexe, Cluster et Franchise sont significatives.

Les résultats obtenus par le GLM Poisson sont résumés dans le tableau ci-dessous :

Sexe	Déviance nulle	Déviance résiduelle	AIC
Âge + Sexe + Cluster + Franchise	19 926	19 018	24 032
Sexe + Cluster + Franchise		19 020	24 031
Cluster		19 336	24 334
Sexe		19 695	24 687
Franchise		19 655	24 657

Ces résultats montrent que le modèle pour lequel la variable âge est supprimée possède un AIC plus faible, et donc que ce modèle est meilleur que celui obtenu avec toutes les variables.

Différents modèles avec une seule variable explicative ont également été réalisés. En effet, pour rappel, l'objectif ici, est de déterminer, laquelle des variables explicatives, entre le sexe et la CSP

(regroupée en *clusters*), est la plus discriminante pour expliquer l'incidence en incapacité. Le modèle a variable unique qui possède l'AIC le plus faible est celui obtenu avec la variable « Cluster ».

Afin de confirmer ces résultats, une régression de Lasso va à présent être utilisée pour sélectionner les variables significatives.

Régression de Lasso

La figure ci-dessous montre, pour une séquence de $log(\lambda)$, les valeurs des coefficients du modèle. En augmentant λ, les coefficients sont pénalisés et finissent par s'annuler. Cette figure permet ainsi de voir, pour les différentes valeurs de λ, quand chaque variable entre dans le modèle et dans quelle mesure elle influence la variable réponse.

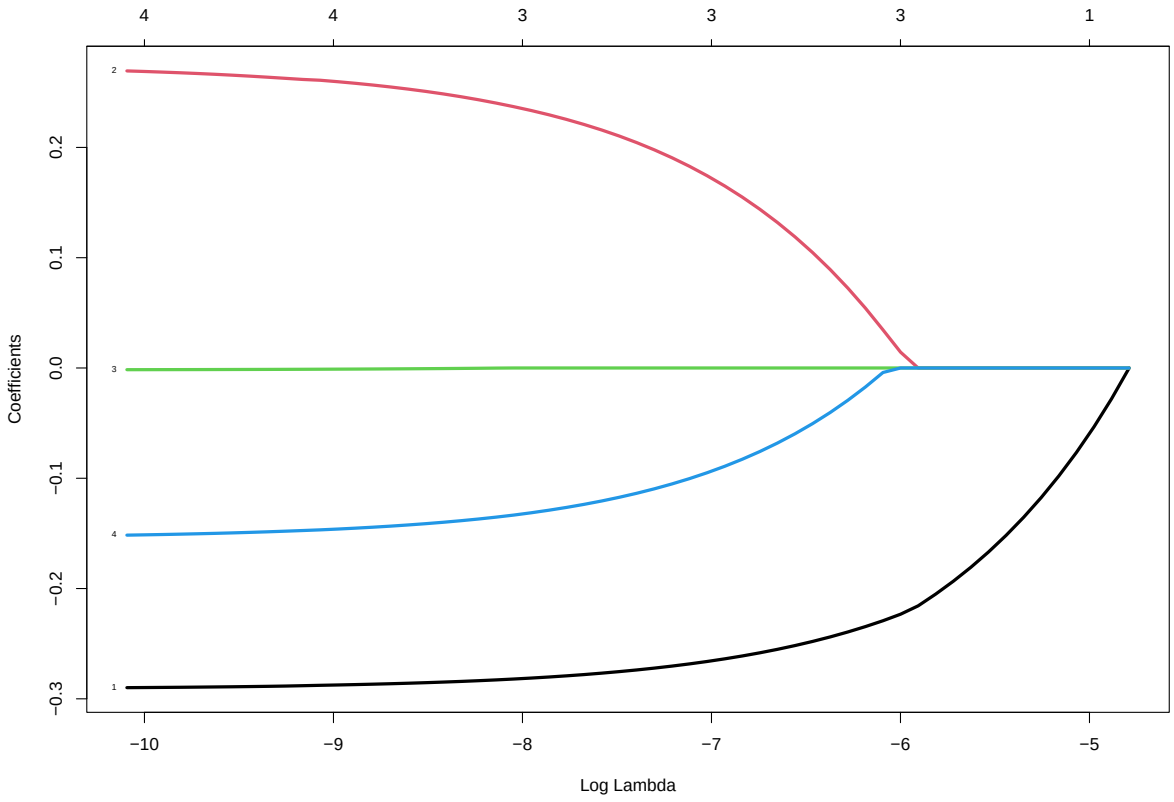

FIGURE 4.8 – Valeurs des coefficients du modèle en fonction des différentes valeurs de lambda

Chaque courbe du graphique correspond à une variable. Le coefficient correspondant à cette variable est représenté pour chaque valeur de λ (on utilise ici la norme L_1). La courbe noire correspond à la variable cluster, la courbe rouge à la variable sexe, la courbe bleue à la variable franchise, et enfin, la courbe verte à la variable âge. L'axe x supérieur montre le nombre décroissant de coefficients du modèle avec des valeurs croissantes de λ. La figure 4.8 permet de bien visualiser la sélection des variables via la méthode LASSO. La Figure 4.8, permet de constater que la variable qui influence le plus le modèle est la variable *cluster* de CSP, car c'est la dernière pour laquelle le coefficient s'annule. Ces résultats sont en phase avec ceux obtenus précédemment, qui montraient

également que la variable la plus influente sur l'incidence en incapacité pour la cause accident, étant la variable correspondant au groupe de CSP.

La deuxième variable la plus importante est la variable sexe, qui s'annule pratiquement en même temps que la variable franchise.

De plus, comme avec la régression de Poisson, la régression de Lasso montre que la variable Age n'est pas une variable significative pour expliquer la variable réponse.

4.5 Bilan

Les principales informations à retenir du GLM Poisson et de la régression de Lasso sont les suivantes :

- La variable la plus importante pour expliquer l'incidence en incapacité pour la cause accident est le groupe de CSP auquel appartient l'individu ;
- L'âge n'est pas une variable significative à retenir.

Ainsi, il ne sera donc pas nécessaire de créer des lois d'incidence par sexe et par âge pour modéliser l'incidence en incapacité pour la cause accident. En effet, comme démontré précédemment, la variable sexe est très corrélée à la variable *cluster*, et la variable *cluster* de CSP explique mieux l'incidence en incapacité que la variable sexe. C'est donc la profession qui est le principal facteur expliquant l'incidence en arrêt de travail.

Il est néanmoins important de souligner le fait que lors de la création des *clusters* de CSP, la dimension sexe est implicitement prise en compte. En effet, les clusters ont été construits sur deux dimensions, l'une d'elles étant l'incidence pour la cause accident au sein de la CSP. Or, comme le montre la figure 2.12, les taux d'incidence sont toujours beaucoup plus élevés chez les hommes que chez les femmes. Ainsi, la notion de sexe, qui apporte beaucoup d'information, n'est pas vraiment écartée du modèle, puisque celle-ci est implicitement prise en compte dans les *clusters* de CSP.

Le GLM Poisson permet également de récupérer l'incidence moyenne propre à chaque *cluster* de CSP, nécessaire à l'équation de tarification. Pour cela, la fonction exponentielle est appliquée aux coefficients issus du GLM à variable unique (*Cluster*), comme l'illustre l'équation (4.3).

Les résultats obtenus montrent également que les groupes de CSP sont chacun significativement différents les uns des autres, et révèlent une notion de croissance de la fréquence des sinistres du *cluster* 5 au *cluster* 1.

Ainsi, en ce qui concerne l'incapacité pour la cause accident, le GLM de Poisson et la régression de Lasso ont permis de valider les hypothèses faites suite à la partie d'analyse descriptive, ainsi que les *clusters* précedemment construits. Le *clustering* effectué par l'algorithme *K-means* a quant à lui permis d'obtenir cinq groupes de CSP homogènes pour le risque accident.

Cependant, vu lors du Chapitre 2., les facteurs expliquant l'incapacité pour la cause accident et pour la cause maladie sont différents. Il n'est donc pas possible de se contenter d'utiliser ces cinq groupes de CSP obtenus comme nouvelles classes tarifaires.

Il est donc nécessaire d'étudier le risque incapacité pour la cause maladie dans une partie qui lui est propre, avant d'aboutir aux classes tarifaires finales.

PARTIE

5

MODÉLISATION DE L'INCIDENCE EN INCAPACITÉ POUR LA CAUSE MALADIE

Dans le chapitre précédent, il a été montré que l'incidence pour la cause accident était en grande partie déterminée par la profession. Cinq groupes de CSP ont ainsi été construits. Pour ce qui est de la maladie, l'étude de statistiques descriptives a montré que l'incidence en incapacité pouvait être expliquée par deux facteurs : l'âge et le sexe. Cela n'a rien d'étonnant, en effet, les assureurs utilisent classiquement des lois d'expérience dans lesquelles ces deux variables interviennent, bien que la loi leur interdise d'avoir un tarif différent par sexe. Dans cette partie, l'incidence en incapacité pour la cause maladie sera donc modélisée de façon traditionnelle, à partir de lois segmentées par sexe et par âge. Puis, une segmentation des différentes professions, adaptée au risque incapacité pour la cause maladie, sera effectuée.

5.1 Construction de tables d'entrée en incapacité pour la cause maladie

Comme expliqué dans le Chapitre 3, il a été décidé d'utiliser un modèle fréquence $\times$ coût pour tarifier la garantie IJ toute cause. Ce choix de modèle de tarification impose d'avoir des lois, permettant d'obtenir la probabilité pour un assuré d'âge x, d'entrer en incapacité. Afin de construire de telles tables, il est donc nécessaire d'avoir un estimateur quantifiant le nombre moyen de sinistres par assuré d'âge x. La loi traditionnellement utilisée pour quantifier ce type d'évènement est une loi de Poisson.

5.1.1 Construction et propriétés de l'estimateur

Construction

On considère traditionnellement que, la variable N_x, quantifiant le nombre moyen de sinistres par assuré d'âge x sur la classe d'âge $[x; x+1]$, suit une loi de Poisson dont le paramètre est le produit entre un paramètre λ_x inconnu et l'exposition. Pour rappel, la loi de Poisson sert à modéliser un nombre entier d'évènement, et sa loi discrète dépend d'un unique paramètre λ, qui représente à la fois la moyenne et la variance de la loi. L'exposition E_i d'un individu i, sur la classe d'âge $[x; x+1]$ correspond quant à elle au nombre de jours sur lesquels l'individu i a été exposé au risque sur la

classe d'âge, divisé par 365. On a donc $e_i \leq 1$. Il est ainsi raisonnable de postuler que l'espérance du nombre de sinistres, et donc le paramètre de la loi est proportionnel à cette quantité.

L'estimateur du paramètre de cette loi a l'avantage de pouvoir prendre en compte les censures et troncatures présentes dans le jeu de données, si l'on considère que le paramètre λ varie linéairement avec la durée d'exposition au sein d'une même classe d'âge. Par conséquent, pour utiliser cet estimateur, il sera donc nécessaire vérifier que :

- Les sinistres survenus annuellement pour chaque assuré peuvent être modélisés par une loi de Poisson ;
- L'estimateur de λ est une fonction linéaire de la durée d'exposition au risque.

Ainsi, si l'on souhaite modéliser le nombre de sinistres N_i survenu pour un assuré i d'âge x, sur la période $[a;b]$ (avec $[a;b] \subset [x;x+1]$, on considère que cette variable N_{xi} suit une loi de Poisson de paramètre $\lambda_{xi} = \lambda_x e_{xi}$, où $e_{xi} = \frac{b-a}{365,25}$ est l'exposition de l'assuré i sur l'âge x.

Si la variable aléatoire N_x suit une loi de Poisson de paramètre λ_{xi}, on a alors :

$$\mathbb{P}(N_{xi} = n_{xi}) = \frac{\lambda_{x_i}^{n_{xi}}}{n_{xi}!} e^{-\lambda_{xi}}$$

Avec n_{xi} est la réalisation de la variable aléatoire N_{xi}, soit le nombre de sinistres par assuré au cours de l'âge x.

Il est ensuite possible de déterminer l'estimateur du maximum de vraisemblance de λ_{xi}, noté $\widehat{\lambda_{xi}}$, en partant de la fonction de vraisemblance qui s'écrit :

$$\mathcal{L} = \prod_{i=1}^{s_x} \mathbb{P}(N_{xi} = n_{xi}) = \prod_{i=1}^{s_x} \frac{\lambda_{x_i}^{n_{xi}}}{n_{xi}!} e^{-\lambda_{xi}}$$

s_x est le nombre d'individus à l'âge x.

En tant que produit de probabilités, L est toujours supérieur à 0, c'est pourquoi il est possible d'utiliser la fonction $y \to \ln(x)$, pour $x \in]0;+\infty[$ afin d'obtenir la log-vraisemblance :

$$log\mathcal{L} = \log\left\{\prod_{i=1}^{s_x} \mathbb{P}(N_{xi} = n_{xi})\right\} = \log\left\{\prod_{i=1}^{s_x} \frac{\lambda_{x_i}^{n_{xi}}}{n_{xi}!} e^{-\lambda_{xi}}\right\}$$

Puis en utilisant l'hypothèse fondamentale : $\lambda_{xi} = \lambda_x e_{x_i}$, on a :

$$l = \log\left\{\prod_{i=1}^{s_x} e^{-\lambda_x e_{x_i}} \frac{(\lambda_x e_{x_i})^{n_{xi}}}{n_{xi}!}\right\} = \sum_{i=1}^{s_x} \{-\lambda_x e_{x_i} + n_{xi} \log \lambda_x + n_{xi} \log\, e_{x_i} - \log(n_{xi}!)\}$$

Il s'agit d'une somme de quatre termes, dont les deux derniers de dépendent pas de λ_x, et donc des paramètres du modèle.
En notant :

- $E_x = \sum_{i=1}^{s_x} e_{x_i}$, la somme des expositions pour l'âge x
- $D_x = \sum_{i=1}^{s_x} N_{x_i}$, le nombre de sinistres apparus à l'âge x et d_x sa réalisation

On obtient :

$$l = -\lambda_x E_x + ln(\lambda_x) d_x + \sum_{i=1}^{s_x} \{n_{xi} \log\, e_{x_i} - \log(n_{xi}!)\}$$

Maximiser cette équation revient alors à trouver λ_x tel que :

$$\frac{\partial l}{\partial \lambda_x} = -E_x + \frac{d_x}{\lambda_x} = 0$$

On en déduit l'estimateur du maximum de vraisemblance :

$$\widehat{\lambda_x} = \frac{D_x}{E_x}$$

Cet estimateur correspond alors à une moyenne pondérée des sinistres, par le temps passé dans la période d'observation.

Propriétés

L'estimateur $\widehat{\lambda_x}$ possède les propriétés suivantes :

Espérance :

$$\mathbb{E}[\widehat{\lambda_x}] = \frac{\sum_{i=1}^{s_x} N_{x_i}}{E_x} = \frac{\sum_{i=1}^{s_x} \lambda_x e_{x_i}}{\sum_{i=1}^{s_x} e_{x_i}} = \lambda_x$$

Il s'agit donc d'un estimateur sans biais.

Variance :

$$\mathbb{V}[\widehat{\lambda_x}] = \mathbb{V}(\frac{D_x}{E_x}) = \frac{1}{E_x^2} \mathbb{V}(\sum_{i=1}^{s_x} N_{x_i})$$

Selon une des hypothèses, $\forall i \in [1; n_x]$, $N_{xi} \sim \mathcal{P}(\lambda_{x_i} = \lambda_x e_{x_i})$, où les N_{xi} sont indépendants et identiquement distribués. Ainsi,

$$\mathbb{V}[\widehat{\lambda_x}] = \frac{\lambda_x}{E_x^2} \sum_{i=1}^{s_x} e_{x_i} = \frac{\lambda_x}{E_x}$$

Estimation de l'intervalle de confiance et constitution de classes d'âge

Puisque l'estimateur $\widehat{\lambda_x}$ est sans biais, on a :

$$\frac{\sqrt{n}[\widehat{\lambda_x} - \lambda_x]}{\sqrt{\mathbb{V}[\widehat{\lambda_x}]}} \xrightarrow[n \to \infty]{} \mathcal{N}(0, 1)$$

On a donc l'intervalle de confiance, pour λ_x, de niveau asymptotique $(1 - \eta)$, suivant :

$$\left[\widehat{\lambda_x} - q_{1-\eta/2} \sqrt{\frac{\mathbb{V}[\widehat{\lambda_x}]}{n}}; \widehat{\lambda_x} + q_{1-\eta/2} \sqrt{\frac{\mathbb{V}[\widehat{\lambda_x}]}{n}}\right]$$

Avec $q_{1-\eta/2}$ le quantile d'ordre $1-\eta/2$ de la loi $\mathcal{N}(0,1)$.

Ainsi, l'intervalle de confiance 95% s'écrit :

$$\left[\widehat{\lambda_x} - 1.96\sqrt{\frac{\widehat{\lambda_x}}{n\sum_{i=1}^{s_x} e_{x_i}}}; \widehat{\lambda_x} + 1.96\sqrt{\frac{\widehat{\lambda_x}}{n\sum_{i=1}^{s_x} e_{x_i}}}\right]$$

Où n correspond au nombre d'assurés exposés au moins un jour sur la classe d'âge $[x; x+1]$.

La largeur de l'intervalle est ensuite obtenue de la façon suivante :

$$L = 2 \times 1.96\sqrt{\frac{\widehat{\lambda_x}}{n\sum_{i=1}^{s_x} e_{x_i}}}$$

L'objectif est de construire une table par franchise et par sexe. Les franchises 15 et 30 correspondent à 97.5% des franchises souscrites pour la maladie dans le portefeuille étudié. Afin d'avoir des volumes de données suffisants, il a donc été décidé de construire des lois sur ces deux franchises uniquement. Il faut donc ici construire quatre tables.

Comme $\widehat{\lambda_x}$ a une valeur différentes pour chaque classe d'âge, de chacune des tables, on calcule $\bar{\lambda_x}$, la moyenne pondérée par l'exposition des $\widehat{\lambda_x}$ d'une table.

Ainsi, pour une largeur L donnée, le volume n d'assurés exposés par classe d'âge est différent selon la table considérée.

Pour les assurés exposés entièrement sur la classe d'âge observée, donc $e_{xi} = 1$, on a alors :

$$L = 2 \times 1.96\sqrt{\frac{\widehat{\lambda_x}}{n}}$$

Afin d'avoir une largeur L la plus petite possible, et donc des résultats plus fiables, il est primordial d'avoir des volumes suffisants pour chaque classe d'âge. Le volume n d'assurés exposés sur une classe d'âge n'étant pas atteint pour chaque âge, il est donc parfois nécessaire de regrouper certaines classes. Pour cela, la méthode par glissement a été retenue. Cette dernière consiste à regrouper certains âges entre eux, dès lors que les contraintes de volumes de sont pas vérifiées.

Pour conserver un résultat par âge et de ne pas déformer les résultats, certaines règles ont été fixées. Premièrement, il a été décidé de regrouper des classes d'âges entre elles, seulement si l'écart avec la classe centrale est inférieur à deux ans. Secondement, le nouvel âge x, obtenu par moyenne pondérée par l'exposition, ne doit pas s'écarter de plus de 0,5 ans de l'âge initial.

Cette opération est ensuite répétée pour chaque âge où la contrainte de volume n'est pas vérifiée. Comme la contrainte de volume n'est souvent pas vérifiée sur les âges extrêmes, et qu'une des règles impose de ne pas s'écarter de plus de 0,5 ans de l'âge initial, cela a pour conséquence que les tables débutent aux altentours de 23 ans (au lieu de 18 ans), et finissent aux alentours de 63 ans (au lieu de 67 ans). Cependant ceci n'est pas un problème car, dans la tarification actuellement en vigueur, le tarif de l'IJ toute cause est identique pour les âges de 18 à 24 ans. En ce qui concerne le tarif des âges supérieurs à 63 ans, il est possible d'envisager d'appliquer les mêmes coefficients de passage que ceux actuellement en vigueur sur les âges de 63 à 67 ans.

5.1.2 Construction des lois brutes

Adéquation à la loi de Poisson

Afin de pouvoir utiliser l'estimateur $\widehat{\lambda_x}$, il est nécessaire de s'assurer que celui-ci suit bien une loi de Poisson, sur chaque âge et sur chacune des tables. Pour rappel, l'objectif est de construire des lois d'entrée en incapacité, segmentées par sexe et par franchise maladie. Cela revient donc à construire quatre tables distinctes, car seules les franchises 15 jours et 30 jours seront étudiées. En effet, pour rappel, ces deux franchises correspondent à 97,5% des franchises souscrites sur notre portefeuille sur la partie maladie. Il existe également des franchises 60, 90, 180 et 365 jours, mais les volumes sont par conséquent trop insuffisants pour obtenir des résultats fiables.

Pour vérifier que l'estimateur $\widehat{\lambda_x}$, suit bien une loi de Poisson , seuls les assurés exposés sur la totalité de la tranche considérée, ont été conservés. Ces derniers représentent 70% du jeu de données. Selon une des hypothèses fondamentales, l'estimateur est linéaire de l'exposition dans le cadre d'une loi de Poisson. Ainsi, si ce dernier suit bien une loi de Poisson sur les assurés exposés sur la totalité de l'année, alors il en sera de même chez les assurés exposés partiellement.

De plus, comme pour la partie accident, l'exposition d'un assuré sur la tranche d'âge $[x; x+1]$ est déterminée uniquement avec les dates de début et fin d'exposition, sans retirer les périodes d'incapacité. En effet, si les périodes d'arrêt de travail avait été déduite de la période d'exposition, alors tous les assurés exposés sur la totalité d'une année seraient des assurés non sinistrés, ce qui aurait eu un impact important sur les résultats.

Afin de vérifier l'adéquation de l'estimateur à la loi de Poisson, un test d'adéquation du χ^2 a été réalisé. Ce test a pour objectif de comparer le nombre de sinistres obtenus, au nombre de sinistres attendus, pour une variable N_x suivant une loi de Poisson de paramètre λ_x.
Pour cela, on utilise la statistique de test suivante :

$$Z^2 = \sum_{i=0}^{k} \frac{R_{x_i} - N_{x_i}}{N_{x_i}} \sim \chi^2(k)$$

Avec :

- R_{x_i} le nombre réel d'individus ayant eu i sinistres sur la classe d'âge $[x; x+1]$;
- N_{x_i} le nombre théorique d'individus ayant eu i sinistres sur la classe d'âge $[x; x+1]$;
- k le nombre maximum de sinistres par individu sur la classe d'âge $[x; x+1]$, qui correspond à 3 sur notre jeu de données.

On teste ainsi l'hypothèses nulle H_0, selon laquelle : $N_x \sim \mathcal{P}(\lambda_x)$.

Puis, on décide de rejeter ou non l'hypothèses nulle, en comparant la valeur Z^2 calculée avec la valeur z^2 coïncidant avec un χ^2 à 3 degrés de liberté. Si un seuil de confiance de 5% est fixé, cela revient alors à comparer la p-value associée, avec le seuil de 5%. Si cette p-value est inférieure à 5%, alors on rejette l'hypothèse nulle.

Afin de mesurer l'adéquation de l'estimateur à la loi de Poisson, on calcule pour chaque table le pourcentage d'adéquation. Pour cela, on calcule le rapport du nombre de classes d'âge pour lesquelles l'hypothèses nulle n'est pas rejetée, sur le nombre de classes d'âge total. Les résultats obtenus sont résumés dans le tableau ci-dessous :

Franchise	Femmes	Hommes
15	90%	85%
30	100%	92%

TABLE 5.1 – Taux d'adéquation à la loi de Poisson par table

Les résultats sont très satisfaisants, bien que l'hypothèse nulle soit rejetée sur certains âges. En effet, le modèle de Poisson surestime souvent la probabilité d'obtenir un unique sinistre, et sous-estime la probabilité d'en avoir deux.

Cependant, l'adéquation moyenne de 92% obtenue sur l'ensemble des tables, permet de valider la modélisation de la sinistralité, pour la cause maladie, par une loi de Poisson.

Linéarité de l'estimateur

Avant d'établir les lois brutes, il est nécessaire de vérifier notre deuxième hypothèse fondamentale, qui est celle de la linéarité du paramètre λ en fonction de la durée d'exposition.

Cependant, cette hypothèse est difficilement vérifiable pour chaque âge. En effet, pour cela, il faut commencer par créer des classes d'exposition en différenciant les assurés exposés entièrement sur la classe d'âge considérée, ceux exposés 9 mois, ceux exposés 6 mois, 3 mois, et moins de 3 mois. Puis, il faudrait construire les mêmes tables que celles réalisées sur les assurés exposés la totalité de l'année, et s'assurer, pour chaque sexe et chaque franchise, que l'estimateur $\widehat{\lambda}$ d'une classe d'âge donnée décroit linéairement avec la durée d'exposition. Cela reviendrait donc à tracer plus de 160 courbes.

De plus, une étude sur les durées d'exposition révèle que :

- 70% des assurés sont exposés sur la totalité de la classe d'âge considérée ;
- 9% sont exposés entre 9 et 12 mois ;
- 7% sont exposés entre 6 et 9 mois ;
- 6% sont exposés entre 3 et 6 mois ;
- 8% sont exposés moins de 3 mois.

Il est donc évident que pour certaines classes d'exposition, le volume d'assurés sur une classe d'âge donnée est insufisant pour obtenir des résultats robustes au niveau de la linéarité.

Néanmoins, cette hypothèse de linéarité a tout de même été vérifiée sur les classes d'âge ayant un volume de données suffisant pour le permettre. Pour ces classes d'âge, la linéarité du paramètre λ était bien vérifiée.

La linéarité du paramètre λ étant une hypothèse couramment utilisée en assurance, pour les modèles de comptage intégrant une notion d'exposition au risque, on considère que cette hypothèse peut être admise dans le cadre de cette étude.

Maintenant que les deux hypothèses fondamentales sont vérifiées, les lois brutes d'entrée en incapacité peuvent être construites.

5.1.3 Lissage des lois brutes

Les lois brutes précédemment obtenues nécessitent d'être lissées avant de pouvoir être exploitées. En effet, celles-ci présentent des irrégularités sur certains âges, ainsi que certaines valeurs extrêmes, parfois dues à trop peu de données sur l'âge en question.

Afin de pallier à ces problèmes, un lissage de ces lois brutes doit être effectué. Il a été décidé d'effectuer un lissage par *splines*.

Introduction aux modèles additifs généralisés

Dans certaines situations, la variables réponse ne peut être modélisée par une fonction linéaire. C'est souvent le cas de la variable âge, pour laquelle il n'est pas rare d'observer une augmentation du risque aux âges extrêmes, comme cela est le cas en assurance auto par exemple. Dans ce contexte, les GLM ne permettent pas de capter les effets non monotones, et une solution naturelle consiste à utiliser un modèle additif généralisé (GAM). En effet, les GAM s'affranchissent de l'hypothèse fondamentale de linéarité des GLM, tout en conservant leur seconde contrainte fondamentale, qui est l'additivité.

Les GAM sont donc une généralisation des GLM, permettant d'incorporer des formes non-linéaires des prédicteurs. On a alors :

$$g(\mathbb{E}[y_i|x_i]) = \sum_{j=1}^{p} f_j(x_{ij})$$

Où g est une fonction de liaison, les valeurs observées x_{ij} sont supposées appartenir à une famille exponentielle et les fonction f_j sont des fonctions indéfiniment dérivables des variables prédictives.

Plusieurs choix sont possibles en ce qui concerne les fonctions f_j. En effet, celles-ci peuvent notamment prendre la forme de fonctions polynomiales, de fonctions en escaliers ou de splines. Ce sont ces dernières quit ont été choisies dans le cadre de cette étude.

Les GAM offrent ainsi une solution intermédiaire : ils peuvent être adaptés à des relations complexes et non linéaires et faire de bonnes prédictions dans ces cas, tout en nous permettant de comprendre et d'expliciter la structure sous-jacente des modèles.

En effet les GAM permettent de capter non seulement les aspects linéaires de cette relation, mais aussi de nombreuses relations non linéaires, notamment grâce à la flexibilité des splines.

En général, lors de l'ajustement d'un modèle non linéaire, deux choses sont à équilibrer. En effet, ce dernier doit pouvoir capturer la relation en étant proche des données, tout en évitant de s'adapter au bruit, ou de le sur-ajuster.

La façon dont le GAM capture les modèles dans les données est mesurée par un terme appelé vraisemblance. Sa complexité, c'est-à-dire la mesure dans laquelle la courbe change de forme, est mesurée par l'ondulation. La clé d'un bon ajustement est le compromis entre les deux. Ce compromis est exprimé par cette équation simple, avec un paramètre de lissage, λ, qui contrôle l'équilibre.

$$\text{Ajustement} = \text{Vraisemblance} - \lambda \times \text{Ondulation}$$

Le paramètre λ contrôle ainsi le compromis entre le lissage et la fidélité du modèle. Un λ plus élevé conduira à un terme de pénalité plus élevé, et, par conséquent, à des valeurs ajustées plus lisses. Inversement, avec un petit λ, le terme non pénalisé gagne en importance et la courbe lisse sera plus proche des données.

Dans la suite, les fonctions *splines*, particulièrement adaptées à la modélisation de la variété de phénomènes non linéaires, seront introduites.

Introduction aux *splines*

Les *B-splines* Il existe plusieurs types de splines. Les *B-splines*, comme expliquées par Michael Price dans sa thèse [16] approximent des fonctions en utilisant des polynômes par morceaux.

Le théorème de Weierstrass [1] stipule qu'il existe toujours un polynôme arbitrairement proche d'une fonction continue. Cependant, le degré de ce polynôme peut être très grand. Afin d'abaisser le degré du polynôme, plusieurs polynômes peuvent être utilisés, chaque polynôme n'estimant qu'un petit segment, situé entre deux noeuds, de la fonction à approximer.

Ces polynômes sont connectés aux nœuds de manière à atteindre certains critères de régularité, notamment pour que la fonction d'approximation ait une régularité C_d, où d est le degré des polynômes par morceaux. Cela implique que la fonction est d-fois différentiable.

Une *B-spline* est une combinaison linéaire de fonctions de base. Soit K le nombre de nœuds dans un intervalle fermé et d le degré de la fonction de base de la *B-spline*. Définissons κ_k comme l'emplacement du $k^{\text{ième}}$ nœud, k prenant ses valeurs dans le vecteur suivant $(-d, \ldots, K+d+1)$. Nous avons donc le vecteur de noeuds suivant $(\kappa_{-d}, \ldots, \kappa_{K+d+1})$. La fonction de base de la *B-spline* est alors définie par la formule récursive de *Cox de Boor* :

Initialisation pour $s = 0$:

$$B_{k,0}(x) = \begin{cases} 1 & \text{si } \kappa_{k-1} \leq x < \kappa_k \\ 0 & \text{sinon} \end{cases}$$

Pour $s = 1, 2, \ldots, d$.

$$B_{k,s}(x) = \frac{x - \kappa_{k-1}}{\kappa_{k+s-1} - \kappa_{k-1}} B_{\kappa,s-1}(x) + \frac{\kappa_{k+s} - x}{\kappa_{k+s} - \kappa_k} B_{k+1,s-1}(x)$$

Une *B-spline* de degré d possède donc les propriétés suivantes :

- Elle est constituée de $d+1$ parties polynomiales de degré d ;
- Les morceaux polynomiaux se rejoignent au niveau de d nœuds internes ;
- Les dérivées jusqu'à l'ordre $d-1$ sont continues et positives sur un domaine couvert par $d+1$ nœuds.

Il est alors possible d'écrire la *B-spline* C comme combinaison linéaire de I fonctions de base de *B-splines* :

1. Théorème d'approximation de Weierstrass, 1885

$$C(x) = \sum_{i=1}^{I} a_i B_{i,d}(x)$$

Où I est le nombre total de fonction de base de *B-splines* utilisées. Les points de contrôle de la courbe *B-spline* sont représentés par les a_i. Dans notre cas, les points de contrôle a_i doivent être vus comme des scalaires (ou des poids) qui vont multiplier la i^{ime} fonction de base. Il est ainsi lié indirectement à l'intervalle où la fonction de base $B_{i,d}$ est non nulle ($[\kappa_i, \kappa_{i+d}]$)

La figure ci-dessous (5.1) illustre une *B-spline* de degré 1. Une *B-spline* de degré 1 consiste se compose de deux parties linéaires, centrées sur κ_k et s'étendant sur trois nœuds au total.

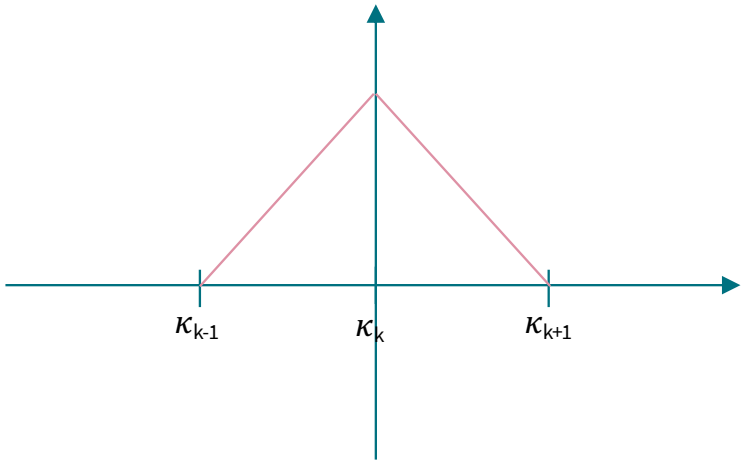

FIGURE 5.1 – Exemple d'un spline de degré 1

Une *B-spline* de degré 3 (figure 5.1) se compose quant à elle de quatre parties cubiques (ie de polynômes de degré 2), centrées sur κ_{k+1} et s'étendant sur cinq nœuds. Aux trois endroits où les morceaux cubiques se rencontrent ($\kappa_{k-1}, \kappa_k, \kappa_{k+1}$), non seulement les dérivées premières des deux parties sont égales, mais les dérivées secondes le sont aussi.

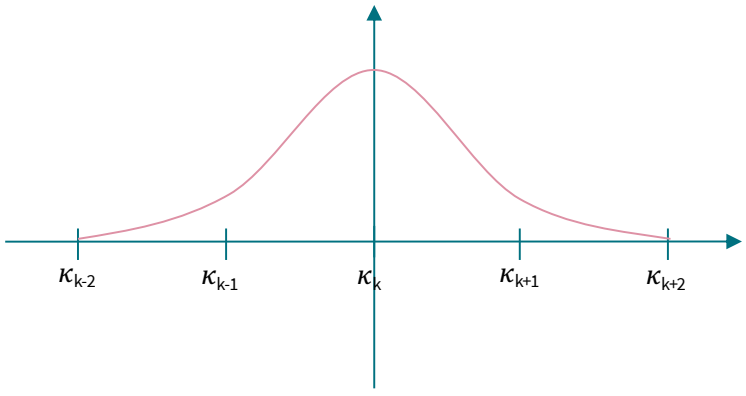

FIGURE 5.2 – Exemple d'un spline de degré 3

Lors de l'interpolation des données, la courbe est estimée par la méthode des moindres carrés pour trouver les valeurs optimales des points de contrôle.

En utilisant y_j comme données observées, la fonction objective à minimiser est la suivante :

$$\hat{a_i} = \text{argmin}_{a_i} \sum_{j=1}^{n} \left\{ y_j - \sum_{i=1}^{I} a_i B_{i,d}(x_j) \right\}^2$$

Ce qui donne la courbe de la *B-spline* ajustée :

$$\hat{c}(x) = \sum_{i=1}^{I} \hat{a_i} B_{i,d}(x)$$

Les P-splines Bien que les *B-splines* soient une bonne option pour l'estimation en termes de qualité d'ajustement, de facilité de calcul et de lissage, le choix du nombre et de la position optimaux des nœuds est une tâche complexe.

Si le nombre de nœuds intérieurs, k, devait être augmenté jusqu'à un nombre relativement élevé, la courbe ajustée $C(x)$ aurait une variation plus importante que ce qui est raisonnable compte tenu des données. Nous serions alors dans le cas d'un surajustement (ou *overfitting*). Un nombre insuffisant de nœuds peut quant à lui conduire à l'exclusion de caractéristiques importantes des données.

Afin de limiter ce surajustement, O'Sullivan [2] a créé une pénalité sur la dérivée seconde de la courbe ajustée. En utilisant la méthode d'O'Sullivan, les points de contrôle, a_i , sont trouvés en minimisant la fonction objective :

$$\hat{a_i} = \text{argmin}_{a_i} \sum_{j=1}^{n} \left\{ y_j - \sum_{i=1}^{I} a_i B_{i,d}(x_j) \right\}^2 + \lambda \int_{x_{min}}^{x_{max}} \left\{ \sum_{i=1}^{I} a_i B''_{i,d}(x) \right\}^2 dx$$

Où λ contrôle la régularité de l'ajustement (où l'ondulation comme introduit sur dans la partie précédente portant sur les GAM) et y_j est la donnée observée. Lorsque $\lambda = 0$, l'ajustement P-spline est le même que l'ajustement B-spline. Lorsque λ devient grand, l'ajustement est similaire à celui d'un polynôme de degré $d - 1$.
Eilers et Marx [6] ont modifié cette configuration en basant la pénalité sur les différences finies des coefficients des B-splines adjacentes, ce qui réduit la dimensionnalité du problème de n, le nombre d'observations, à I, le nombre de B-splines.

Les points de contrôle, a_i de la courbe ajustée en utilisant la méthode d' Eilers et Marx sont trouvés en minimisant la fonction objective :

$$\hat{a_i} = \text{argmin}_{a_i} \sum_{j=1}^{n} \left\{ y_j - \sum_{i=1}^{I} a_i B_{i,d}(x_j) \right\}^2 + \lambda \sum_{i=m+1}^{I} (\Delta^m a_i)^2$$

Où m est le degré de la pénalité, Δ l'opérateur de différence, et y_j les données observées.

2. O'Sullivan. A statistical perspective on ill-posed inverse problems (with discussion). *Statist. Sci.* 1986

5.1.4 Lissage de la loi d'incidence des femmes pour la franchise 30 jours

Afin de lisser les quatre lois d'incidence, des *P-splines* sont utilisées. Dans la suite de cette partie, la loi obtenue sur la franchise 15 jours chez les femmes sera utilisée à titre d'exemple pour illustrer le lissage par *splines*. En effet, on observe sur cette dernière un pic d'incidence plus marqué aux alentours de 30 ans, comme cela avait observé lors de l'analyse descriptive. Les taux bruts d'incidence en incapacité des femmes pour la franchise 15 jours sont représentés par des points sur la figure 5.3. La loi d'incidence en incapacité lissée et son intervalle de confiance associé sont également représentés.

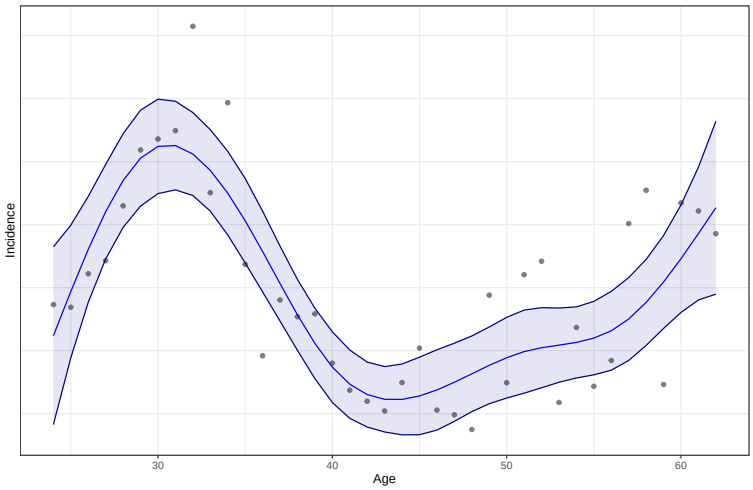

FIGURE 5.3 – Loi d'incidence en incapacité des femmes pour la franchise 15 jours

Cette courbe lissée a été obtenue en utilisant la méthode d'Eilers et Marx présenté précédemment et implémenté sous R avec le package package *MGCV*. Dans l'article d'Eilers and Marx, différentes méthodes sont proposées pour estimer le paramètre de régularité λ comme l'AIC, la validation croisée, ou encore le maximum de vraissemblance. Cependant comme présenté par Iain Currie et Maria Durban [4], la méthode du maximum de vraissemblance est une bonne alternative à l'AIC ou à la validation croisée qui peuvent parfois ne pas suffisament lisser les données en retenant un paramètre λ trop faible.

En reprenant l'exemple de la loi lissée d'incidence en incapacité des femmes ayant une franchise 15 jours, il est intéressant de comprendre comment la *P-Splines* se compose. La figure 5.4 représente les 8 fonctions de bases (et l'intercept représenté en rouge) de la *P-spline*. Pour rappel une *P-spline* (ou *B-spline*) peut s'écrire sous la forme

$$C(x) = \sum_{i=1}^{I} a_i B_{i,d}(x)$$

Ce sont donc les 9 fonctions de base $B_{i,d}$ qui sont représentées ci dessous (5.4).

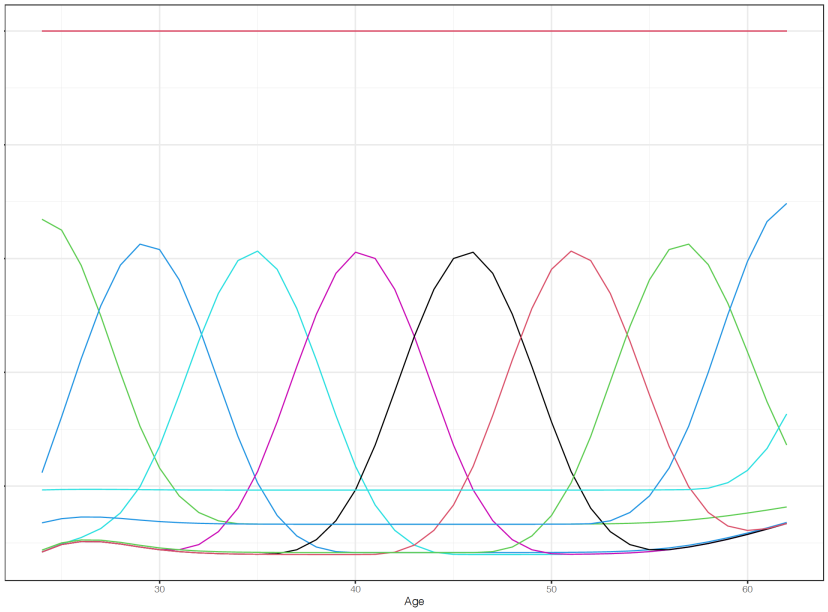

FIGURE 5.4 – Fonctions de base de la P-spline femmes franchise 15 jours

Ces fonctions sont ensuite pondérées par les points de contrôles a_i. La figure 5.5 représente les 9 fonctions de base pondérées par les points de contrôles soit les $a_i B_{i,d}$.

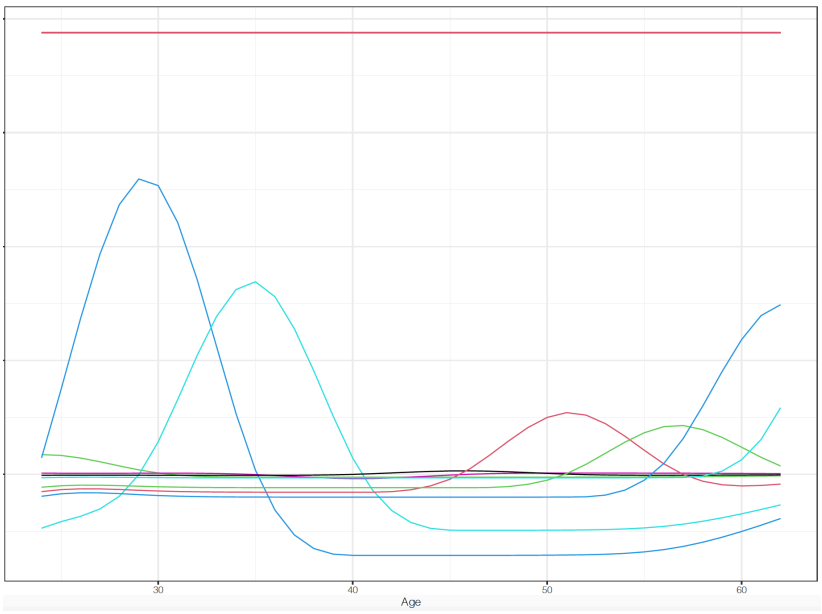

FIGURE 5.5 – Fonctions de bases pondérées de la P-spline femmes franchise 15 jours

En sommant ces différentes fonctions de bases pondérées nous obtenons la *P-spline* C définie par $C(x) = \sum_{i=1}^{I} a_i B_{i,d}(x)$ qui correspond à notre loi d'incidence lissée. Cette *P-spline* est représentée en rouge sur la figure 5.6.

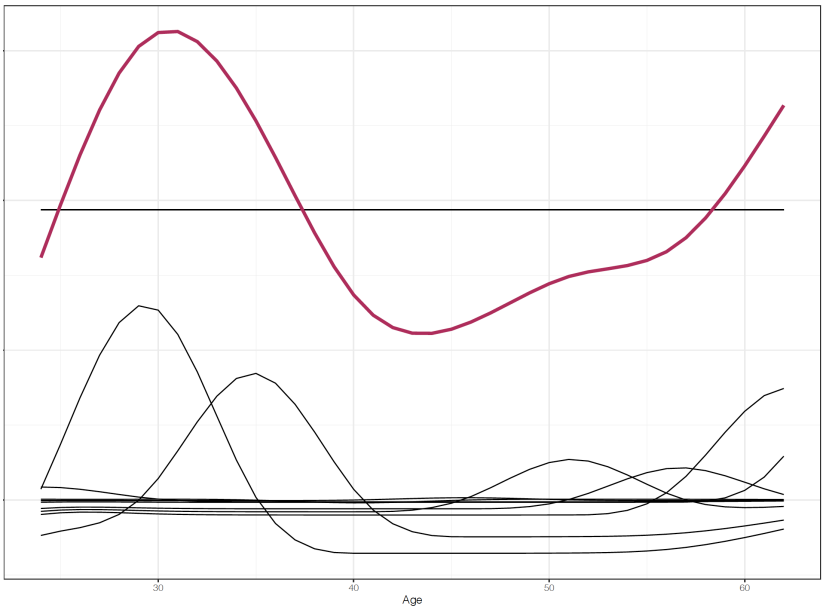

FIGURE 5.6 – Schéma splines

En observant les résidus normalisés du modèle présentés sur la figure 5.7, nous ne remarquons pas d'inadéquation du lissage par *P-splines*. En effet les résidus sont tous en valeurs absolus inférieurs à 2.

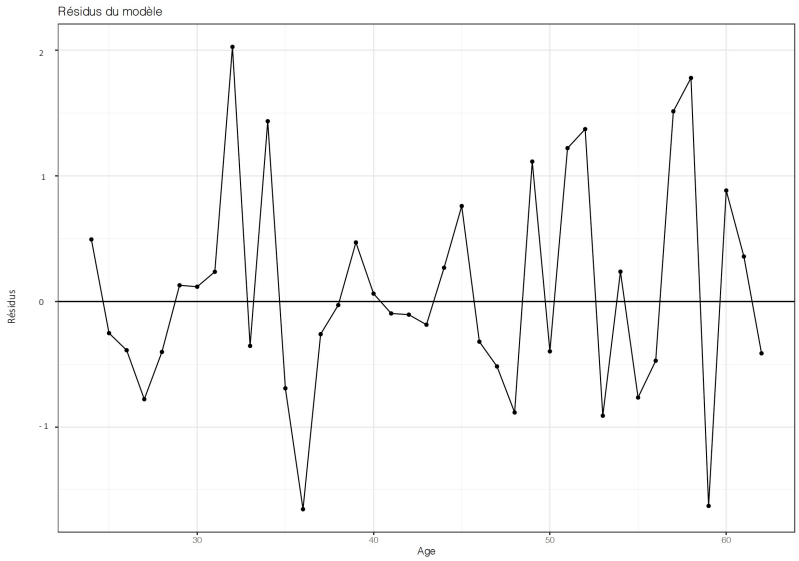

FIGURE 5.7 – Schéma splines

Courbes lissées

En plus de la loi d'incidence en incapacité maladie chez les femmes sur la franchise 15 jours, nous obtenons les lois suivantes sur la franchise 30 jours :

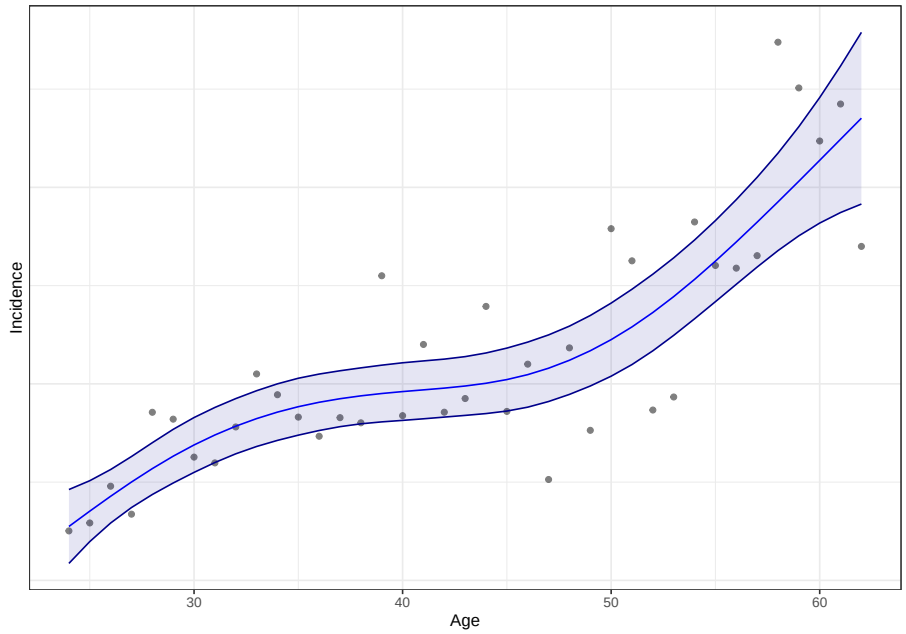

FIGURE 5.8 – Loi d'incidence en incapacité maladie, chez les femmes sur la franchise 30 jours

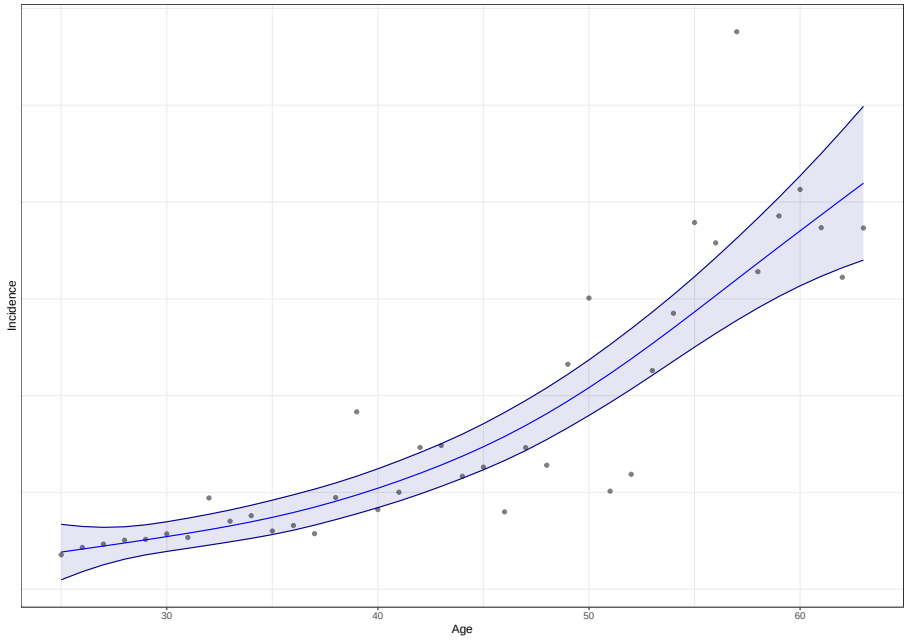

FIGURE 5.9 – Loi d'incidence en incapacité maladie, chez les hommes sur la franchise 30 jours

La construction des lois d'entrée en incapacité confirme que le sexe et l'âge ont tous les deux un impact considérable sur l'incidence en incapacité pour la cause maladie.

Chez les femmes, une bosse aux alentours de 30 ans est observée comme sur la figure Figure 2.11. Ce pic est en revanche moins marqué sur la franchise 30 jours que sur la franchise 15 jours.

Chez les hommes, on observe, comme lors de l'analyse descriptive, une incidence croissante avec l'âge.

5.2 Construction de groupes de professions homogènes pour le risque maladie

La construction des lois d'entrée en incapacité a démontré que le sexe et l'âge ont tous les deux un impact important sur l'incidence en incapacité pour la cause maladie.

Ainsi, afin obtenir des groupes de professions, homogènes en termes de risque incapacité sur la cause maladie, il faudrait avoir des groupes homogènes en termes de répartition hommes/femmes. En effet, la dimension âge est déjà prise en compte dans le tarif actuel, puisque ce dernier est différent selon l'âge de l'assuré.

Il a également été observé que chez les femmes, le pic d'incidence aux alentours de 30 ans était plus marqué sur la franchise 15 jours que sur la franchise 30 jours. Afin de limiter l'effet d'antisélection, les groupes de CSP, pourraient être construits sur un échantillon d'assurés exposées sur la franchise 30 jours. En effet, la franchise maladie 30 jours représente 60% des franchises souscrites sur notre jeu de données, et c'est pour cette franchise qu'une meilleure adéquation à une loi de Poisson (cf. Table 5.1) était obtenue. Mais avant cela, afin de ne pas introduire de biais dans les résultats, il est important de s'assurer que la répartition des franchises 15 jours et 30 jours est équivalente sur chacune des professions.

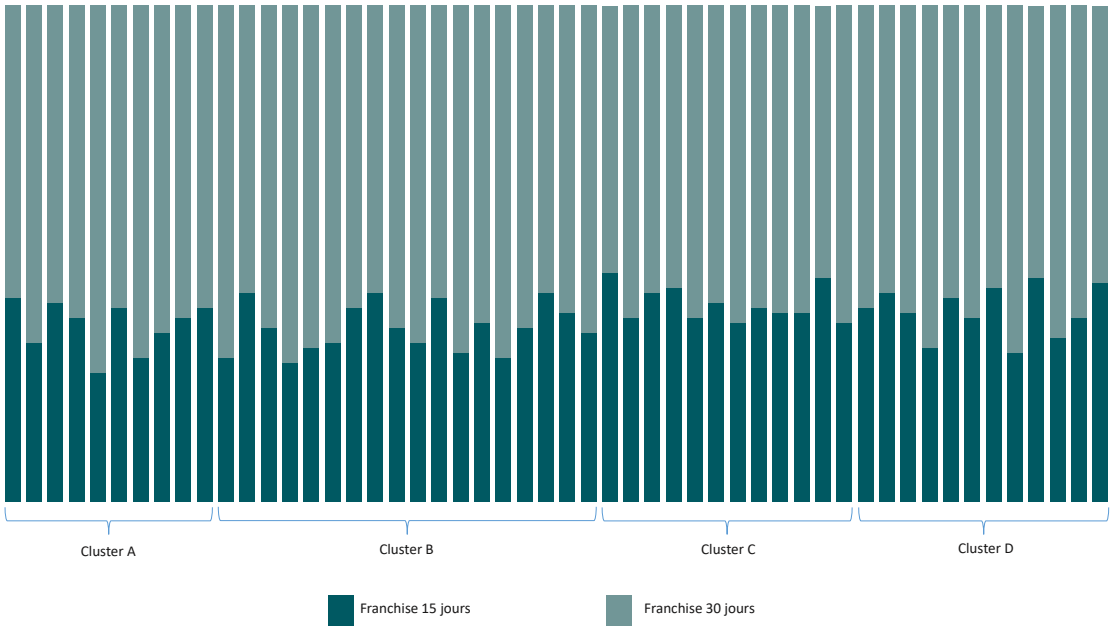

FIGURE 5.10 – Répartition des franchises 15 jours et 30 jours par CSP

La figure 5.10 permet de voir que la répartition des franchises est plutôt équivalente entre les différentes professions. Il n'existe donc pas de différence majeure de répartition des franchises maladie sur les différentes professions étudiées.

Une autre dimension qu'il est nécessaire d'intégrer aux groupes de professions, que nous souhaitons créer sur la partie maladie, est la proportion de sinistres maladie au sein de chaque profession. Pour rappel, cette dimension avait déjà été intégrée lors de la construction des *clusters* sur la partie accident, puisque ces derniers avaient été construits à partir de deux critères : l'incidence en incapacité sur la cause accident, et la proportion de sinistres accident, sur chaque CSP.

Cependant, il est important de noter que les groupes construits sur la partie accident n'ont pas le même objectif que ceux sur la partie maladie. En effet, pour la cause accident, le but était d'expliquer l'incidence, et de démontrer que la profession était le principal facteur pouvant expliquer un taux d'incidence différent entre deux individus. En revanche, sur la maladie, les facteurs explicatifs sont déjà identifiés, et l'objectif est de créer des groupes homogènes pour ces facteurs.

Néanmoins, pour créer ces groupes, le même algorithme de clustering que celui utilisé pour la cause accident sera utilisé : l'algorithme *K-means* pondéré. De même, l'algorithme sera exécuté sur les mêmes CSP que celles sélectionnées sur la partie accident. Pour rappel, 52 CSP avaient été sélectionnées sur deux critères :

- Le nombre de sinistres observés sur la CSP doit être au minimum de 10 sinistres par an, en moyenne, sur la période d'observation ;
- Le volume de contrat doit représenter au moins 0,5% du portefeuille.

Ces 52 CSP représentant 85% des souscriptions du portefeuille et 90% des sinistres sur la période d'observation, elles permettent ainsi d'obtenir des résultats fiables et robustes.

Ainsi, pour créer des groupes de professions homogènes en termes de risque incapacité sur la cause maladie, l'algorithme *K-means* pondéré sera exécuté, sur les 52 CSP séléctionnées sur la partie accident, en ne conservant que les individus exposés sur la franchise 30 jours, afin de prendre en compte l'effet de cette franchise sur l'incidence. Puis pour chacune de ces CSP, deux critères seront relevés :

- Le ratio hommes/femmes au sein de la profession considérée ;
- La proportion de sinistres maladie.

En choisissant ces deux critères, l'objectif est donc d'obtenir des groupes homogènes en termes de répartition hommes/femmes et de proportion de sinistres maladie. Il est important d'avoir à l'esprit que la notion d'âge est implicitement prise en compte dans le *clustering*, via la proportion de sinistres maladie. En effet, comme vu lors de la partie d'analyse descriptive, l'incidence en incapacité pour la cause maladie est fortement corrélée à l'âge.

En procédant ainsi, on obtient après plusieurs exécutions de l'algorithme K-means les groupes suivants :

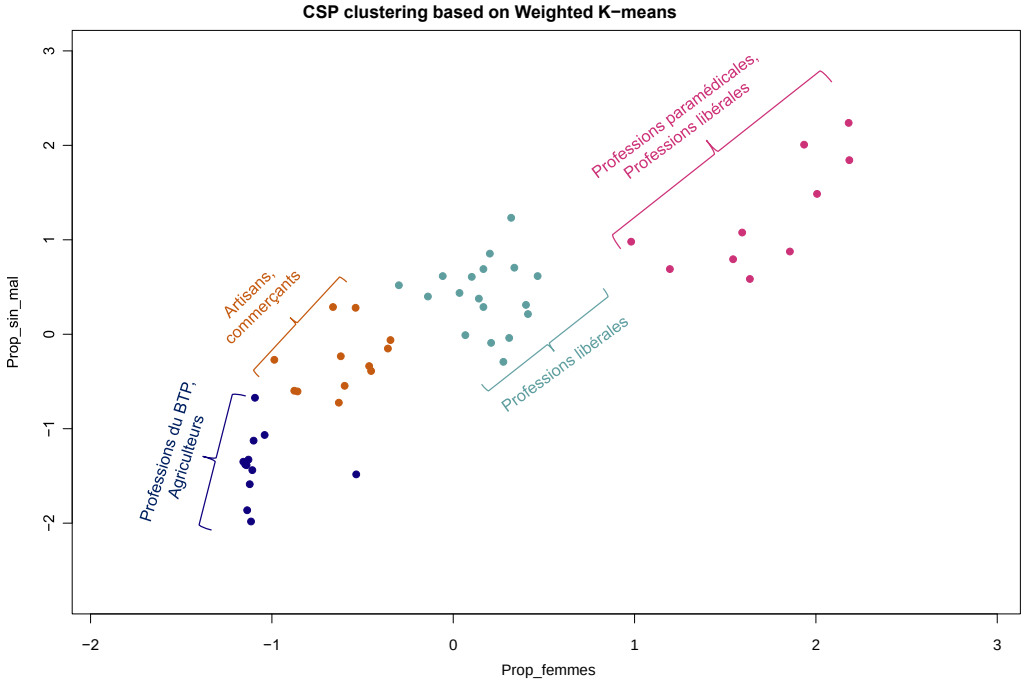

FIGURE 5.11 – Clustering des CSP pour le risque maladie

Après plusieurs exécutions de l'algorithme *K-means*, le partitionnement des CSP, le plus stable est celui représenté dans la figure 5.11. Comme pour la partie accident, les libellés de CSP ne sont pas dévoilés de manière exhaustive, mais de manière générale. Dans ce partitionnement, le groupe représenté en rose correspond aux CSP dans lesquelles la proportion de femmes est la plus importante, avec une proportion de sinistres maladie élevée. En revanche, les professions représentées en beige correspondent à des professions qui sont majoritairement exercées par des

hommes, avec une faible proportion de sinistres maladie (et donc une forte proportion de sinistres accident).

Les effectifs de chaque groupe sont résumés dans le tableau suivant :

Cluster	Effectifs
A	10
B	18
C	12
D	12

TABLE 5.2 – Effectifs des CSP par cluster

Il est essentiel de comprendre que ces groupes de CSP ne servent pas à construire un tarif sexué, mais bien à déterminer l'incidence moyenne au sein d'un groupe, en utilisant les lois d'incidence précédemment construite ainsi que la proportion d'hommes/femmes. Ainsi, une femme exerçant une profession qui se trouvent dans le cluster bleu, aura, à âge égal, exactement le même tarif qu'un homme exerçant cette même profession.

Des groupes de CSP, ont donc été construits indépendamment pour le risque accident et pour le risque maladie. Cependant, le tarif de l'IJ faisait l'objet d'une refonte tarifaire, est une addition de la prime pure accident et de la prime pure maladie. La prochaine étape va donc consister à trouver comment concilier ces groupes de professions, construits indépendamment sur ces deux risques, afin de pouvoir proposer un tarif adapté à chaque profession, à la fois pour la cause accident et la cause maladie.

PARTIE

6

CONSTRUCTION DES NOUVELLES CLASSES TARIFAIRES

Jusqu'à présent, les classes tarifaires en vigueur, sur le produit faisant l'objet de ce mémoire, ont été construites sur le caractère accidentogène de la profession. Ainsi, à âge égal, une personne souhaitant souscrire une garantie IJ toute cause sur ce produit, se verra appliquer un tarif différent selon sa profession. Or, le sexe est également un facteur influant sur le taux d'entrée en incapacité. Cependant, l'arrêt « Test Achats » de la Cour de Justice de l'Union Européenne, interdit aux assureurs de proposer des produits dont la tarification ou les prestations diffèrent en fonction du sexe des assurés.

L'objectif est donc ici d'affiner la tarification à l'aide des nouvelles classes tarifaires, et d'être capable de capter de façon équivalente, le risque accident et le risque maladie, au sein d'une même classe tarifaire.

6.1 Calcul de la prime pure

Pour rappel, comme présenté dans le chapitre 3 (Equation (3.1)), la prime pure de l'IJ toute cause est une somme de la prime pure accident et de la prime pure maladie :

$$P_x = Pa_x + Pm_x \tag{6.1}$$

Comme le modèle de tarification retenu est de type fréquence × coût, on a :

$$P_{cause_x} = E[N_x] \times C_x[IJ]$$

où

$$C_x[IJ] = IJ \times \sum_{i=fra}^{min(\text{dmaxgar},\ \text{maintien}, fra+\text{maintien}-1, 1095)} \text{maintien}_x(i) \tag{6.2}$$

Avec :

- Pa_x : la prime pure de l'IJ accident pour un assuré d'âge x ;
- Pm_x : la prime pure de l'IJ maladie pour un assuré d'âge x ;
- P_x : la prime pure pour un assuré d'âge x ;
- IJ : Le montant de l'indemnité journalière souscrite ;
- N_x : la variable aléatoire représentant le nombre de sinistres par assuré, survenus à l'âge x ;
- fra : durée de la franchise ;
- $dmaxgar$: durée maximale de la garantie, définie contractuellement ;
- maintien$_x(i)$: la probabilité de rester en incapacité jusqu'au i^{me} jour, pour un assuré d'âge x ;

On obtient alors $E[N_x]$ grâce aux tables d'entrée en incapacité, et $C_x[IJ]$ grâce aux tables d'expérience de maintien, qui donnent, pour un âge x, la probabilité de rester n jours en incapacité.

6.2 Elaboration des nouvelles classes tarifaires

Jusqu'à présent, cinq groupes de professions homogènes en termes de risque accident, et quatre pour le risque maladie, ont été construits indépendamment. Il est donc nécessaire de trouver façon de concilier ces deux partitionnements.

Les *clusters* précédemment construits sur la partie accident, et sur la partie maladie, possèdent les répartitions suivantes, en termes d'effectifs de CSP :

Clusters maladie	
Cluster	Effectifs
A	10
B	18
C	12
D	12

Clusters accident	
Cluster	Effectifs
1	10
2	4
3	14
4	17
5	7

Les groupes obtenus sur la partie maladie étant plutôt équivalents en termes de volumes, contrairement aux groupes obtenus sur la partie accident, il a été décidé de partir des groupes construits sur la partie maladie pour construire les nouvelles classes tarifaires.

Puis, une correspondance entre les professions des *clusters* maladie et des *clusters* accident est établie. Plusieurs sous-groupes, correspondant chacun à une classe tarifaire, ont ainsi été obtenus, comme cela est illustré dans l'exemple ci-dessous :

Cluster maladie	Profession	Cluster accident	Nouvelle classe tarifaire
A	Profession 1	Cluster 3	A3
	Profession 2	Cluster 1	A1
	Profession 3	Cluster 5	A5
	Profession 4	Cluster 3	A3
	Profession 5	Cluster 3	A3
	Profession 6	Cluster 1	A1
	Profession 7	Cluster 5	A5
	Profession 8	Cluster 3	A3
	Profession 9	Cluster 1	A1
	Profession 10	Cluster 3	A3

Le cluster A, obtenu sur la partie maladie, permet d'obtenir 3 classes tarifaires A1, A3 et A5, en faisant intervenir les *clusters* obtenus sur la dimension accident.

En procédant ainsi, sur le reste des *clusters* maladie, la segmentation tarifaire suivante est obtenue :

Cluster maladie	Cluster accident	Classe tarifaire	Nombre de CSP
A	1	A1	3
	3	A3	5
	5	A5	2
B	1	B1	7
	2	B2	2
	3	B3	7
	4	B4	2
C	2	C2	2
	4	C4	8
	5	C5	2
D	3	D3	2
	4	D4	7
	5	D5	3

Puis, afin de déterminer les paramètres de l'équation (6.2), on récupère pour chaque classe tarifaire :

- Le ratio hommes/femmes au sein de la classe tarifaire ;
- L'incidence moyenne pour l'incapacité en accident, correspondant à celle du *cluster* accident qui lui est assigné, obtenue à l'aide du GLM.

Pour ce qui est de la prime pure accident, l'incidence moyenne n'étant pas impactée par l'âge, on récupère donc l'incidence moyenne du *cluster* accident associé, obtenue par le GLM.

En revanche, en ce qui concerne la prime pure maladie, il a été montré que l'incidence était expliquée par le sexe et l'âge de l'assurée. Ainsi, pour un âge x donné, on récupère l'incidence en pondérant par le ratio hommes/femmes, l'incidence obtenue pour cet âge grâce aux lois d'entrée en incapacité.

Enfin, en ce qui concerne la partie maintien, les données sont récupérées dans les tables d'expérience.

En procédant ainsi, une segmentation tarifaire permettant à chaque profession d'avoir un tarif juste, a été construite. Cependant, les classes tarifaires ont été construites sur 52 CSP et non sur les 230 présentes dans le portefeuille. Pour rajouter les CSP manquantes, plusieurs options seraient possibles. Dans un premier temps, il serait envisageable de rattacher toutes les professions appartenant à une même catégorie ensemble. Par exemple, si dans l'échantillon la profession "Commerçant en librairie" est présente, alors il serait possible de rattacher d'autres types de commerçants similaires. Une autre solution envisagée serait de récupérer des données publiques, pour les CSP dont les volumes sont insuffisants dans notre portefeuille. Par exemple, il serait envisageable de récupérer le ratio hommes/femmes de cette profession, ou des données relatives aux types de sinistres, sur des sites tels que l'INSEE ou AMELI.

Par ailleurs, il est nécessaire d'avoir à l'esprit que, les tables d'incidence, utilisées dans l'équation tarifaire de prime pure, ont été construites avec une vision *Best Estimate*, alors qu'une marge de prudence est en réalité appliquée sur notre table d'expérience de maintien.

CONCLUSION ET PERSPECTIVES

Dans ce mémoire, une démarche basée sur la modélisation de l'incidence en incapacité a été menée, dans le cadre de la refonte tarifaire d'un produit de prévoyance, proposant des garanties de types indemnitées journalières pour faire face au risque d'incapacité. Cette étude constitue un véritable enjeu pour Axa France, leader sur le marché des travailleurs non salariés.

L'objectif de cette étude était de trouver une segmentation tarifaire plus adaptée au portefeuille, capable de prendre en compte à la fois la dimension accident et la dimension maladie du risque incapacité. En effet, les classes tarifaires actuellement en vigueur dans le tarif, reposent sur le caractère accidentogène de chaque profession, regroupant ainsi les professions en deux classes, de la moins risquée à la plus risquée. La dimension maladie est quant à elle, prise en compte via l'âge. Or, afin d'avoir un tarif plus adapté à la réalité du risque arrêt de travail, il a été décidé de revoir la segmentation tarifaire, en élaborant des classes tarifaires de professions, capables de capter de manière équivalente le risque accident et le risque maladie.

Pour cela, afin d'identifier les variables ayant un impact notable sur les niveaux de taux d'incidence en incapacité pour chacune des deux causes, ce mémoire a débuté par une étude de statistiques descriptives. L'objectif de cette dernière était de permettre de mieux cerner les facteurs pouvant expliquer l'incidence en incapacité pour l'accident et pour la maladie. Celle-ci avait également vocation à orienter la suite de l'étude et les choix en matière de modélisation du risque incapacité. Après avoir conjecturé que les facteurs ayant un impact significatif sur les taux d'incidence étaient différents pour la cause accident et la cause maladie, il a été décidé de diviser la suite de l'étude en deux parties.

La première partie était consacrée à la cause accident. Dans cette partie, cinq groupes de professions, homogènes pour le risque accident, ont été construits à l'aide d'un algorithme de clustering. Ces groupes de professions ont été construits sur deux critères :

- L'incidence de chaque profession en incapacité pour la cause accident ;
- La proportion de sinistres dus à un accident.

Puis, il a été démontré, à l'aide d'un GLM Poisson et d'une régression Lasso que, comme conjecturé lors de l'analyse descriptive, le facteur le plus discriminant pour expliquer l'incapacité accident était la profession exercée par l'individu. Aussi, il a été démontré que l'âge n'était pas un facteur explicatif, et que le sexe de l'individu était fortement corrélé à la profession exercée.

Puis, afin de ne pas reproduire des classes tarifaires similaires aux classes actuellement en vigueur, cette étude a été complétée par une partie consacrée à la modélisation du risque incapacité pour la cause maladie. Pour rappel, l'étude de statistiques descriptives avait conduit à conjecturer que les facteurs expliquant l'incapacité pour la cause maladie étaient le sexe et l'âge de l'individu. Une approche traditionnelle a donc été suivie afin de construire des lois d'incidence par âge, segmentée par franchise et sexe. Ces lois ont confirmé qu'il existait bien une tendance différente selon l'âge et le sexe de l'individu. De nouveaux groupes de professions, homogènes pour le risque maladie, ont alors été créés à l'aide d'un algorithme de *clustering*, en utilisant néanmoins des critères différents que ceux choisis pour la cause accident. En effet, le sexe de l'individu étant un facteur essentiel pour expliquer l'incidence pour la cause maladie, il a été choisi de construire des groupes à partir des critères suivant :

- Le ratio hommes/femmes au sein de la profession ;
- La proportion de sinistres maladie au sein de la profession.

Ainsi, les groupes de professions obtenus pour la cause accident et la cause maladie ont été construits séparément, sur des critères différents. Afin de traiter conjointement ces deux causes au sein d'une même classe tarifaire, il a été choisi de partir des groupes de professions obtenus pour la cause maladie, et de regarder, pour chacune des professions de chaque groupe, à quel cluster accident celle-ci appartenait. En procédant de la sorte, onze classes tarifaires ont pu être élaborées. Au sein de ces dernières, les professions sont à la fois similaires sur la dimension accident, et sur la dimension maladie du risque incapacité.

Cette étude comporte néanmoins certaines limites. En effet, lors de la phase de souscription, les assurés sont soumis à des sélections médicale et sportive, à l'issue desquelles, l'assureur peut décider d'exercer une majoration tarifaire, en cas de risque aggravé. Or, dans cette étude, tous les individus de même âge ont été traité indifféremment. Cependant, dans l'échantillon, certains individus présentaient un risque aggravé. En traitant indifféremment les individus du même âge, l'incidence en incapacité a donc été surestimée. Il est donc important de souligner qu'une marge de prudence a indirectement été appliquée ici.

Par ailleurs, il est nécessaire de garder à l'esprit que cette étude a été menée sur un produit relativement récent. Ainsi, la majorité des assurés présents dans le portefeuille se trouvent peu éloignés de la phase de sélection médicale. Au moment de l'étude, la connaissance du risque est donc relativement bonne. Néanmoins, il est important d'être conscient que, dans les années à venir, le portefeuille sera plus éloigné de la sélection médicale que le portefeuille ayant servi à construire ce tarif. Ainsi, il est fortement possible que le tarif, aujourd'hui *Best Estimate*, ne le soit plus dans les années à venir. Afin de faire face à cela, deux solutions sont envisageables. Il pourrait être décider de majorer directement le tarif actuel afin de prendre en compte les évolutions futures, ou bien, de décider de garder un tarif compétitif pour le moment, mais de prévoir des campagnes de majoration ou une révision du tarif, lorsque cela sera nécessaire.

En vue d'approfondir l'étude réalisée lors de ce mémoire, les pistes de perspectives, et axes d'amélioration suivants seraient envisageables :

- Mener une étude de rentabilité afin de confronter le tarif actuel et celui obtenu avec la nouvelle segmentation tarifaire proposée ;
- Trouver une façon d'enrichir les classes tarifaires construites sur 52 CSP, avec les CSP restantes dans notre portefeuille.

BIBLIOGRAPHIE

[1] Guillaume Biessy. *Modélisation pour l'Assurance Dépendance*. Cours, ISUP, 2021.

[2] Arthur Charpentier. *Modèles Linéaires Appliqués*. Cours, UQAM, 2020.

[3] Michael J Crawley. *The R book*. John Wiley & Sons, 2012.

[4] Iain D Currie et Maria Durban. "Flexible smoothing with P-splines : a unified approach". In : *Statistical Modelling* 2.4 (2002), p. 333-349.

[5] Piet De Jong, Gillian Z Heller et al. "Generalized linear models for insurance data". In : *Cambridge Books* (2008).

[6] Paul HC Eilers et Brian D Marx. "Flexible smoothing with B-splines and penalties". In : *Statistical science* 11.2 (1996), p. 89-121.

[7] Valeria Fonti et Eduard Belitser. "Feature selection using lasso". In : *VU Amsterdam Research Paper in Business Analytics* 30 (2017), p. 1-25.

[8] Yann Fournier. *Construction de tables d'entrée en incapacité et application à la tarification de produits Prévoyance*. Mémoire, ISUP, 2010.

[9] Aurélie Gaumet. *Construction de tables d'expérience pour l'entrée et le maintien en incapacité*. Mémoire, ISFA, 2001.

[10] Bilale Harouna Abassi. *Modélisation de l'incidence en incapacité sur un portefeuille de prévoyance individuelle*. Mémoire, ISUP, 2020.

[11] Trevor Hastie, Junyang Qian et Kenneth Tay. *An Introduction to glmnet*. 2016.

[12] Maxime Huttin. *Refonte tarifaire du produit Prévoyance Evolution, accompagnée du modèle de rentabilité*. Mémoire, EURIA, 2013.

[13] Pararawendy Indarjo. *Using Weighted K-Means Clustering to Determine Distribution Centres Locations*. 2020.

[14] Alboukadel Kassambara. *Practical guide to cluster analysis in R : Unsupervised machine learning*. T. 1. Sthda, 2017.

[15] Agence Centrale des Organismes de Sécurité Sociale (ACOSS). *Sécurité Sociale des indépendants*. https ://www.secu-independants.fr. Mis en ligne en 2019.

[16] Michael Joseph Price. “Penalized b-splines and their application with an in depth look at the bivariate tensor product penalized b-spline”. In : (2018).

[17] Noam Ross. *A Free, Interactive Course using mgcv*. https ://noamross.github.io/gams-in-r-course/.