Mémoire présenté le :

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

Par : EZZAKRAOUI Nassim

Titre : Tarification d'un contrat Santé Collective et estimation de l'effet d'une surcomplémentaire

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

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

Membres présents du jury de l'IA *Signature*

Entreprise :
AG2R La Mondiale
Signature :

Membres présents du jury de l'ISFA

Directeur de mémoire en entreprise :
Odile KUSNIK
Signature :

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

Signature du responsable entreprise

Signature du candidat

Secrétariat :
Mme Christine DRIGUZZI
Bibliothèque :
Mme Patricia BARTOLO

REMERCIEMENTS

Je tiens tout d’abord à remercier ma tutrice de stage Odile Kusnik, qui m’a permis d’avancer avec méthodologie et précision dans mon travail. Je souhaite remercier de même mes collègues Jimmy Drapeau et Suzanne Michel pour leur disponibilité, m’apportant tout au long de mon travail les éléments nécessaires à mon étude et me faisant partager leur expérience.
De plus, j’adresse mes remerciements à mon responsable pédagogique Didier Rullière pour ses remarques et l’attention qu’il a porté à mon mémoire.
Enfin, je remercie mes proches qui me soutiennent et me permettent d’avancer sereinement dans mon travail.

Sommaire

Mots clés

Sécurité Sociale, assurance complémentaire, mutualisation, segmentation, Accord National Interprofessionnel, Contrat de base, surcomplémentaire, tarification, primes pures, variables explicatives, aléa moral, anti sélection, algorithme CART, Modèles Linéaires Généralisés

---

Résumé

La santé est l'une des préoccupations majeures des individus. En France, comme dans la plupart des pays développés, cette préoccupation est accentuée par l'évolution croissante de l'espérance de vie et des progrès médicaux. Cependant, bien que le système de santé français soit particulièrement performant, les dépenses excessives engagées en santé par la population française entrainent un désengagement progressif de l'état. Ce retrait de l'assurance maladie dans les remboursements initie une forte implication des assureurs complémentaires, qui récupèrent de plus en plus d'assurés dont les profils de risques peuvent être complètement différents. Ces organismes assureurs doivent alors améliorer leurs méthodes d'identification des risques et de tarification. Nécessité de recherche et d'innovation encouragée par les mesures futures dictées par l'Accord National Interprofessionnel.

C'est pourquoi l'objectif de ce mémoire est double. Le premier objectif est de proposer une méthode de tarification adaptée au portefeuille du groupe AG2R La Mondiale, et qui se concentre principalement sur le comportement de l'assuré en fonction du montage contrat de base - surcomplémentaire souscrit. La méthode retenue sera une modélisation par Modèles Linéaires Généralisés de la fréquence annuelle moyenne et des frais réels moyens, méthode s'adaptant le mieux à l'estimation de l'impact d'une variable. L'intérêt second de ce mémoire est de rechercher une solution innovante en tarification. Nous avons retenu une modélisation par algorithme CART, méthode de classification et de régression par arbre, faisant partie des méthodes par apprentissages statistiques en plein essor mais encore peu utilisées en assurance santé.

Dans ces deux modèles utilisés, nous avons réduit notre périmètre d'étude à trois Conventions Collectives Nationales ainsi qu'à neuf garanties différentes, représentant à la fois un chiffre d'affaire et un volume de consommation majeurs. De plus, nous avons décidé de modéliser non pas comme il se fait classiquement le montant engagé par l'assureur, mais plutôt la consommation réelle de l'assuré. Cette modélisation ne perdant aucune information par rapport à la première mais assurant au contraire une actualisation plus rapide des tarifs lors de différents changements. Sauf indication contraire, tous les résultats présentés tout au long de ce mémoire concerneront la consommation adulte uniquement.

Key Words

Social security, complementary health insurance, risk sharing, segmentation, inter-professional National Agreement, basic contract, supplementary health coverage, pricing, pure premiums, explanatory variables, moral hazard, anti – selection risk, CART algorithm, Generalized Linear Models

---

Abstract

Being healthy is one of the main concerns of the population. In France, like in all developed countries, this preoccupation is getting stronger with the life expectancy and the medical advances. Even though the Health French system is quite effective, the government tends to lower its involvement due to the massive health expenses engaged by the consumers. The setback from the health care system on the reimbursement procedures strengthen the role of the complementary insurers, bringing them more and more insured people with different risk profile. These insurance institutions must then improve its methods to identify the risks and to price. Besides, the need of research and innovation is encouraged by the upcoming regulations fixed by the inter-professional National Agreement.

That is why this thesis master has two main objectives. The first one is to propose a new pricing method adapted to AG2R portfolio, and that is focusing on the insured behaviour induced by the link basic - supplementary coverage subscribed. The one chosen is a modelling by a Generalized Linear Model, which is the best method to estimate the impact of a variable. The second interest of this thesis is to seek an innovative pricing solution. We chose a modelling by the CART algorithm, classification and regression tree, which belongs to the statistical learnings, in a real boom but yet not used in the health insurance.

In these two models chosen, we reduced our study to three National Collective Agreement and nine health benefits, representing respectively a major turnover and volume of consumption. Moreover, we decided to model the real consumption and not the usual reimbursement paid out by the insurer. This kind of modelling does not lose any information but actually enables a quickest update of the prices in the case of different modifications. If we do not precise anything, each result that is showed in this thesis concerns the adult consumption.

I. Introduction

Le monde de l'assurance est en constante évolution, de nouvelles règlementations telles que Solvabilité II imposant des règles prudentielles aux assureurs. En santé, ce contexte évolutif est encore plus marquant en raison notamment de l'intervention de l'état par l'intermédiaire de la Sécurité Sociale. Chaque année est d'ailleurs établie la loi de financement de la Sécurité Sociale qui dresse un bilan des dépenses et recettes de l'année, pouvant modifier certains éléments tels que le Plafond Mensuel de la Sécurité Sociale ou encore l'engagement de celle-ci dans la couverture des risques santé. Actuellement, le secteur santé est en plein bouleversement avec la prochaine mise en place de l'ANI. Ainsi, dans un environnement déjà fortement concurrentiel pour les complémentaires Santé, celles-ci doivent s'adapter aux changements constants. Afin de s'adapter à ceux-ci et de continuer leur activité rentablement, il est donc primordial qu'elles analysent parfaitement les profils de risques auxquels elles sont exposées. Cela passe par une identification précise des assurés couverts et de leurs caractéristiques. La tarification est donc un enjeu majeur pour l'assureur à la fois pour rentabiliser son activité mais aussi pour s'assurer une marge de solvabilité suffisante respectant les règles prudentielles.

Les tarifs établis par l'assureur doivent, tout en séduisant les assurés par rapport à la concurrence, couvrir les engagements futurs auxquels il fera face. La problématique est alors de définir quel modèle doit être utilisé, selon quel paramétrage et pour quelle raison ce modèle est plus adapté qu'un autre. En effet, le choix du modèle peut ne pas être justifié uniquement par le calcul de la prime pure mais aussi par son pouvoir explicatif. Ceci peut être le cas pour un assureur désirant connaître les effets d'un facteur en particulier, afin de pouvoir ajuster ses tarifs lorsque ce facteur intervient. L'effet du niveau de couverture sur la consommation peut être un tel facteur intéressant à identifier pour l'assureur. Par exemple si la sortie d'une nouvelle gamme de produits est prévue, son tarif pourra être facilement réévalué sans avoir à refaire tous les calculs.

De plus, afin d'établir le meilleur tarif en terme de ratio compétitivité/rentabilité, il est indispensable que l'assureur étudie la consommation de ses assurés en fonction des critères qui les définissent. Le but est alors de définir les critères les plus significatifs et d'en déterminer l'impact sur la consommation. Or, les mesures actuellement édictées par l'ANI incitent la population à consommer plus raisonnablement et à avoir une approche plus préventive. La définition de contrats dits responsables, avantageux fiscalement, est l'outil principal de l'Etat dans sa démarche de responsabilisation. Ces contrats proposeront ainsi des produits de base relativement bas afin que le reste à charge dissuade l'assuré de trop consommer. De nombreuses études et mémoires de tarification ont été réalisés et ont déterminé l'influence du niveau de couverture du produit sur la consommation de l'assuré. Néanmoins, peu de travaux ont été menés concernant les liens bases-surcomplémentaires et leurs niveaux de couvertures respectifs. Or, pour compenser cette baisse de niveau de couverture, les assurés désirant mieux se couvrir auront nécessairement recours aux surcomplémentaires proposées par l'organisme assureur. Ce type de montage base-surcomplémentaire sera prochainement le montage récurrent chez chaque assuré. C'est pourquoi dans notre étude, le principal critère dont nous cherchons à évaluer l'impact est la souscription d'une surcomplémentaire par l'assuré. Notre objectif est de déterminer la consommation de l'assuré en fonction du montage base-surcomplémentaire dont il dispose. Les deux méthodes qui nous ont paru adaptées à cette étude sont l'algorithme CART et les Modèles Linéaire Généralisés.
La première partie de ce mémoire permettra au lecteur de se familiariser avec le secteur de la protection sociale et de la santé en particulier. Le but de cette partie sera d'introduire les mécanismes propres à ce marché et les perspectives pour les différents acteurs.

La seconde partie a pour objectif de définir le risque santé et de réaliser une revue des méthodes de tarification principalement utilisées en assurance santé. Afin d'orienter le choix du modèle à utiliser, nous dresserons une liste d'avantages et d'inconvénients pour chacune des méthodes énoncées. Dans cette partie, nous détaillerons plus en profondeur les méthodes par algorithme CART et les Modèles Linéaires Généralisés.
Ensuite, la partie suivante du mémoire consistera en l'analyse de la consommation en santé des assurés d'AG2R La Mondiale et illustrera l'influence des différentes variables étudiées par le biais de statistiques descriptives.
La quatrième partie illustrera les résultats obtenus après application des modèles retenus et fournira une comparaison de ces résultats afin de déterminer la méthode la plus adaptée au portefeuille du groupe.
Enfin, nous finirons ce mémoire par une analyse critique sur la méthode de tarification utilisée par le groupe pour sa nouvelle gamme de produits dans le cadre de l'ANI et du respect des contrats responsables.

II. Présentation de l'assurance santé

A. La Protection Sociale en France

Les régimes de protection sociale répondent à un besoin universel : se protéger du risque. Et bien que le système actuel nous apparaisse comme acquis, sa mise en place a été le fruit d'un long processus ponctué de nouvelles lois, de débats et d'aménagements afin de satisfaire tous et chacun. Ce n'est que le 19 Octobre 1945 que l'on voit naître la Sécurité Sociale en France. Son objectif est triple. Le premier est l'unicité de la sécurité sociale. Le second a pour but la généralisation de cette couverture du risque à tout le monde. Et enfin, le troisième a pour ambition d'étendre le nombre de risques assurés. À ce jour, les deux derniers objectifs ont donc été atteints et permettent à la France de bénéficier d'un des meilleurs systèmes de protection sociale d'Europe.
La protection sociale est aujourd'hui assurée par l'ensemble des organismes qui permettent à la population de se prémunir des pertes financières liées aux risques sociaux. Risques sociaux composés des risques Vieillesse, Maladie, Incapacité/Invalidité, Chômage, Accident de travail/Maladie professionnelle et Maternité. On peut distinguer deux principes fondamentaux dans le système de protection sociale français qui sont les suivants :

- L'assurance contre une perte de revenu entrainée par le chômage ou autres risques sociaux. Néanmoins, seuls les individus cotisant – sur leurs salaires – bénéficient de cette couverture. Nous pouvons en quelque sorte parler de redistribution dans la mesure où les prestations seront souvent proportionnelles aux revenus de l'individu.

- La possibilité d'avoir un niveau de vie acceptable. Ce principe de solidarité permet aux concernés, sous condition d'un seuil de ressources, de recevoir un revenu minimum et de se prémunir contre la pauvreté. Le Revenu de Solidarité Active est un des principaux moyens d'entraide mis en œuvre.

- La protection sociale, qui se concentre uniquement sur certaines catégories de dépenses prédéfinies. Dans ce cas de protection, la couverture est identique pour tous les individus. Les Allocations Familiales en sont un bon exemple.

B. Situation actuelle : l'ANI

Le 11 janvier 2013, des organisations patronales (MEDEF, UPA, CGPME) et trois organisations syndicales ont signé un nouvel accord national interprofessionnel (l'A.N.I). L'ANI apporte de nouvelles règlementations dans le monde professionnel et ouvre la porte à un nouveau modèle économique et social. Cet accord pointe particulièrement sur deux volets distincts : le parcours professionnel et la flexibilité. Nous nous concentrons sur le parcours professionnel, volet concernant notre domaine de la santé. Qu'apporte donc l'ANI à ce niveau ? En premier lieu, une généralisation de la couverture complémentaire santé pour tous les salariés. Cette généralisation devra être effective au 01 Janvier 2016, avec une participation de la part de l'employeur d'au moins 50% de la cotisation totale, et un panier de soins minimum que devra couvrir chaque contrat. Celle-ci sera donc responsable d'une diminution des contrats individuels au profit des contrats collectifs. En deuxième lieu, une suppression de l'ancienne clause de

désignation. Cette clause permettait aux branches professionnelles de choisir un organisme assureur et de l'imposer à toutes les entreprises de ce même secteur. Or, l'ANI abroge cette clause, jugée contraire à la liberté d'entreprendre et à la concurrence. De nombreux acteurs ont l'opportunité d'entrer dans la course, les instituts de prévoyance étant jusque-là les leaders du marché santé collective. Plus d'informations à ce sujet sont disponibles sur le site internet www.argusdelassurance.com. Enfin, afin de limiter les dépenses engagées par les assurés, et par implication celles de la sécurité sociale, l'ANI entreprend une démarche de responsabilisation des assurés. Le principe est de pénaliser les contrats jugés non responsables. Les contrats dits responsables étant ceux proposant des garanties et des niveaux de couverture limités, proscrivant ainsi les frais réels et accentuant sur un minimum d'actes de prévention. Ce procédé se fonde sur l'observation que l'assuré maximise l'utilité qu'il peut tirer du produit souscrit et donc modifie son comportement en terme de consommation santé, en fonction du niveau de couverture.

C. Les Acteurs : complémentarité et concurrence

Aujourd'hui, la protection sociale est répartie en quatre niveaux distincts que nous détaillons ci – dessous.

a) La Sécurité Sociale

La Sécurité Sociale, bien que réservée dans un premier temps aux salariés permanents, s'est étendue à l'ensemble de la population au fil des années. Aussi bien qu'aujourd'hui, elle est le principal assureur obligatoire en France. Avec la mise en place de la Couverture Maladie Universelle, depuis le 01 Janvier 2000, la Sécurité Sociale assure même la totalité (99.9%) de la population française. Elle fournit la couverture « de base » de différents risques dont les principaux types sont les risques Maladie, Incapacité/Invalidité, Décès et Retraite. Ces branches de risque sont gérées quant à elles par différentes caisses qui composent donc le socle de la Sécurité Sociale.
En ce qui concerne notre périmètre, à savoir la branche Maladie, il s'agit de la Caisse Nationale d'Assurance Maladie des Travailleurs Salariés, plus connue sous le nom CNAM. Les autres branches sont celles figurant sur le graphique ci – dessous.

Figure 1. Dépenses de Sécurité Sociale du régime général

S’ajoute aussi à cette séparation des caisses une séparation des régimes. Le régime général en est le principal régime, couvre environ - selon les chiffres affichés sur www.securite-sociale.fr et www.amelie.fr - 80% de la population (la majorité des salariés, étudiants…) et représente plus de la moitié des dépenses en France.
Les régimes spéciaux qui couvrent les salariés non protégés par le régime général et bénéficiant d’un régime souvent propre à leur professions (corps professoral, SNCF…).
Les régimes des non-salariés non agricoles qui assurent principalement les artisans, commerçants et les professions libérales.
Et enfin le régime agricole, qui dépend du ministère de l’agriculture et répond aux besoins des exploitants et salariés agricoles.

b) Les régimes complémentaires

Ces régimes fournissent une protection supplémentaire aux risques assurés par la sécurité sociale. En effet, la couverture assurée par la sécurité sociale n’est pas illimitée. D’une part, la sécurité sociale ne rembourse presque jamais la totalité d’un acte et laisse une partie du coût de l’acte à la charge de l’individu : c’est ce qu’on appelle le « reste à charge ». D’autre part, certains praticiens proposent des tarifs supérieurs à la base de remboursement, et ces dépassements d’honoraires impliquent un reste à charge plus élevé pour le patient. C’est dans cette optique que les complémentaires interviennent, afin de permettre à la population de se protéger pleinement et contre un plus grand nombre de risques. Néanmoins, certaines réglementations, que nous verrons dans la suite de ce mémoire, imposent aux complémentaires des contraintes dans leur remboursement afin de limiter les abus de consommation en santé. Les trois principaux organismes complémentaires sont listés ci-dessous.

Les mutuelles

Ce sont des sociétés à but non lucratif et sont la principale source de complémentaire en termes de nombre de personnes assurées. En effet, en plus du nombre important de petites mutuelles, on retrouve souvent les grandes mutuelles associées à un corps professionnel telles que la MGEN pour l’éducation ou encore la MNH pour le service hospitalier. Elles sont régies par le code de la mutualité et s’organise autour du principe de solidarité et de mutualisation. C’est pourquoi leurs fonds proviennent majoritairement des cotisations des assurés et leurs assemblées générales sont composées des assurés eux – mêmes. Les excédents générés sur l’année sont réinvestis au service des adhérents, pouvant servir par exemple à prendre en charge de nouveaux traitements. Leur chiffre d’affaire est en grande partie généré sur les contrats santé individuels.

Les compagnies d’assurance

Elles suivent le champ d’application du code des assurances. La majorité d’entre elles correspond à des sociétés de capitaux, gouvernées par des actionnaires et donc soumises à des exigences de rentabilité. Une partie d’entre elles est aussi constituée des sociétés d’assurances mutuelles qui elles, sont à but non lucratif. Les compagnies d’assurances se concentrent principalement sur la couverture individuelle mais intègre de plus en plus le périmètre des contrats collectifs.

Les instituts de prévoyance

Les instituts de prévoyance sont en quelque sorte une combinaison des deux types de sociétés vus plus haut. Elles répondent au code de la Sécurité Sociale et se rapprochent des mutuelles dans la mesure où elles sont à but non lucratif. Leur système de gouvernance est dit paritaire car il est composé des représentants des partenaires sociaux (employeurs, employés, entreprises adhérentes). Les instituts de prévoyance se focalisent surtout sur les Conventions Collectives Nationales, et proposent une couverture

standard pour l'ensemble des entreprises affiliées à chacune des CCN. Notre étude portera donc principalement sur ces contrats « CCN ». Contrairement aux sociétés d'assurance et aux mutuelles, les instituts de prévoyance se concentrent principalement sur les contrats collectifs.

c) Bilan Financier des acteurs

1. Le financement de l'activité

La sécurité sociale perçoit en grande partie ses ressources par le biais des cotisations sociales. Par sociales, on entend toutes les cotisations versées par les employeurs, employés, et non-salariés afin d'obtenir les droits aux prestations de la sécurité sociale. Elles sont basées sur des montants plafonds tels que le Plafond Mensuel de la Sécurité Sociale et peuvent être minorées selon certaines conditions de ressources. La Sécurité Sociale perçoit aussi des revenus grâces aux taxes imposées sur les cotisations reçues par les complémentaires, et grâce aux impôts et autres contributions publiques. Celles-ci augmentent d'ailleurs au fur et à mesure au profit des cotisations sociales. Le régime compte 9.6 millions de cotisants en 2012, lui permettant d'obtenir 441 milliards d'euros de recettes, soit environ 30% du PIB. Chaque année est établie la Loi de Financement de la Sécurité Sociale, qui définit les résultats du dernier exercice clos et les dispositions relatives aux dépenses et recettes du prochain exercice. La figure ci-dessous nous montre la structure des recettes de la Sécurité Sociale.

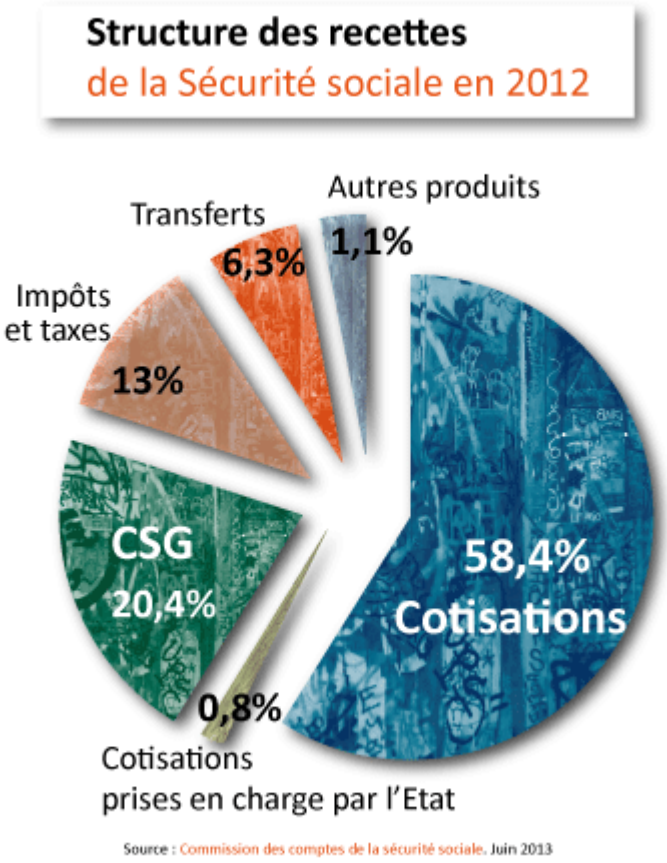


Figure 2. Structure des recettes de la Sécurité Sociale en 2012

Les complémentaires se financent par les cotisations versées par les assurés directement pour les contrats individuels et par l'entreprise lorsqu'il s'agit de contrats collectifs. Les assureurs et instituts de prévoyance se constituent un actif composé de titres souverains. Ces derniers ont aussi la possibilité de détenir des prêts de la part des banques, et de plus en plus de partenariat sont formés, en conséquence notamment des règles imposées par Solvabilité II.

2. Les dépenses et recettes

Le régime général

Les recettes 2012 ont progressé de 4.4% par rapport à 2011 tandis que les dépenses, elles, ont augmenté de 2.9%. Cette progression des recettes plus importante est due en particulier aux bonnes dispositions de la LFSS 2012 qui a ainsi permis de réduire le déficit de -17.4 à -13.3 milliards d'euros. Néanmoins, comme le traduisent ces chiffres, la sécurité sociale est en net déficit et ce depuis 2002. Comme énoncé plus haut, les recettes de la sécurité sociale proviennent en grande partie des cotisations sociales, et donc dépendent de la masse salariale. Or, les crises successives provoquent une détérioration du niveau de l'emploi salarié, et ainsi une diminution des recettes. Tandis que les dépenses, elles, continuent d'augmenter. On parle d'effet ciseau, le montant des charges dépassant celui des produits.
Les branches Maladie et Vieillesse en sont particulièrement responsables. Avec 58.8 millions de bénéficiaires, soit 90% de la population française, et prenant à sa charge 86% des dépenses en soins, la branche Maladie termine l'exercice 2012 avec un déficit de 5.8 milliards d'euros. Il est d'environ 6 milliards d'euros en 2012. Les raisons en sont variées. Entre autres, nous pourrions pointer la croissance des dépenses qui dépasse la croissance économique moyenne. En effet, dans les pays développés tels que la France, la santé est un « bien supérieur » : la demande pour un individu augmente plus vite que ses revenus. De plus, les progrès médicaux et l'espérance de vie entrainent des traitements non seulement plus onéreux mais aussi plus longs. Enfin, nous pouvons ajouter que le système de soins peine à trouver une régulation efficace, d'où l'apparition constante de nouvelles réformes comme l'ANI actuellement.

Les régimes complémentaires

La part des organismes complémentaires dans les dépenses de santé s'élève environ à 25 milliards d'euros, représentant 14 % de la consommation totale de soins et de biens médicaux. Ce sont les mutuelles qui arrivent en tête avec un chiffre d'affaires de 17 milliards, soit 56% du marché santé. Les sociétés d'assurance arrivent en seconde place, chiffrant environ 8 milliards d'euros et enfin les instituts de prévoyance totalisent 5 milliards.

3. Perspectives de croissance

Au vu des chiffres présentés ci-dessus, il apparait sans surprise que les organismes complémentaires sont les deuxièmes financeurs des dépenses en santé. Cette participation des complémentaires, comme le montre la figure 3, évolue petit à petit mais de manière croissante.

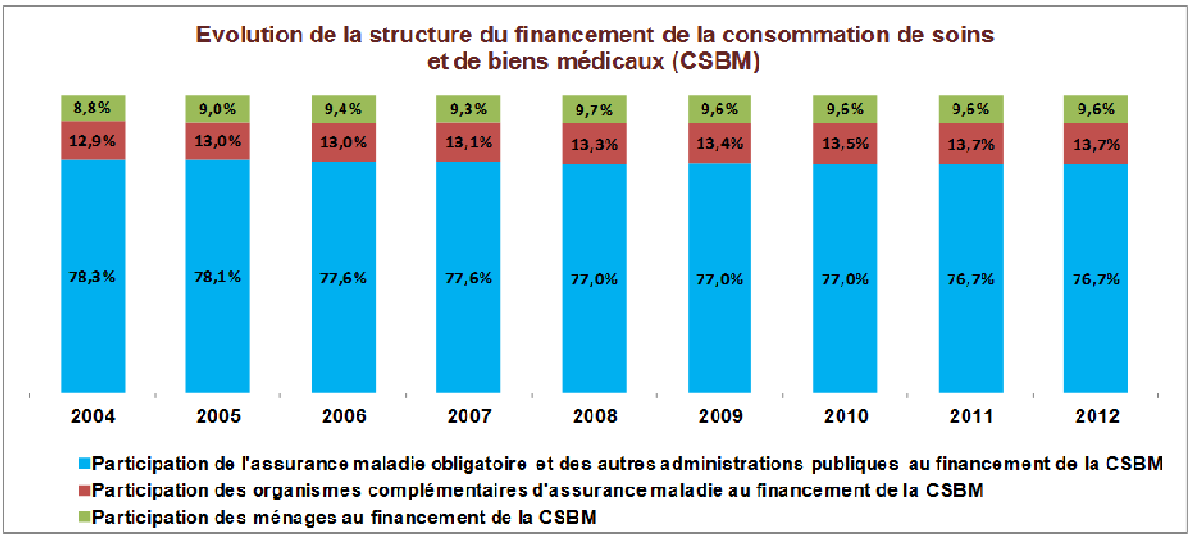


Figure 3. Evolution de la structure du financement de la consommation de soins et de biens médicaux

Entre 2004 et 2012, la part des complémentaires dans les dépenses santé a augmenté d'environ 1%. Cette augmentation est due au retrait progressif de l'assurance maladie et des autres administrations publiques. Retrait progressif qui est en conséquence du déficit croissant de la sécurité sociale. Plusieurs étapes conduisent à ce retrait et nous pourrions citer comme exemple l'instauration de la franchise médicale sur les médicaments et autres soins depuis 2008. Ce contexte législatif en constante évolution entraine donc un rôle de plus en plus significatif de la part des complémentaires, permettant de limiter le reste à charge d'un individu consommant des soins. En plus de ce retrait, s'ajoute l'espérance de vie croissante de la population et un niveau de vie plus élevé entrainant non seulement plus de demande de guérison mais aussi une demande de bien - être. Enfin, comme nous l'avons vu précédemment, les complémentaires ont un rôle très important sur deux postes : le dentaire et l'optique. En effet sur ces deux postes, le remboursement de la sécurité sociale n'est pas assez conséquent. Pourtant le besoin en dentaire couplé aux progrès médicaux (appareils dentaires performants, implants dentaires…) augmente significativement. De même la demande en optique continue d'augmenter fortement. D'une part car la population est de plus en plus sujette à des problèmes de vue, plus de technologie entrainant aussi plus d'heures passées devant l'écran pour ne citer que cet exemple. D'autre part car des produits tels que les lunettes ou les lentilles, utilisées auparavant uniquement par besoin, sont dorénavant utilisés par souci d'esthétisme et de style. Le graphique présenté ci-après nous montre que les complémentaires, à l'inverse de la Sécurité Sociale, prennent une place majeure dans les dépenses de ces postes.

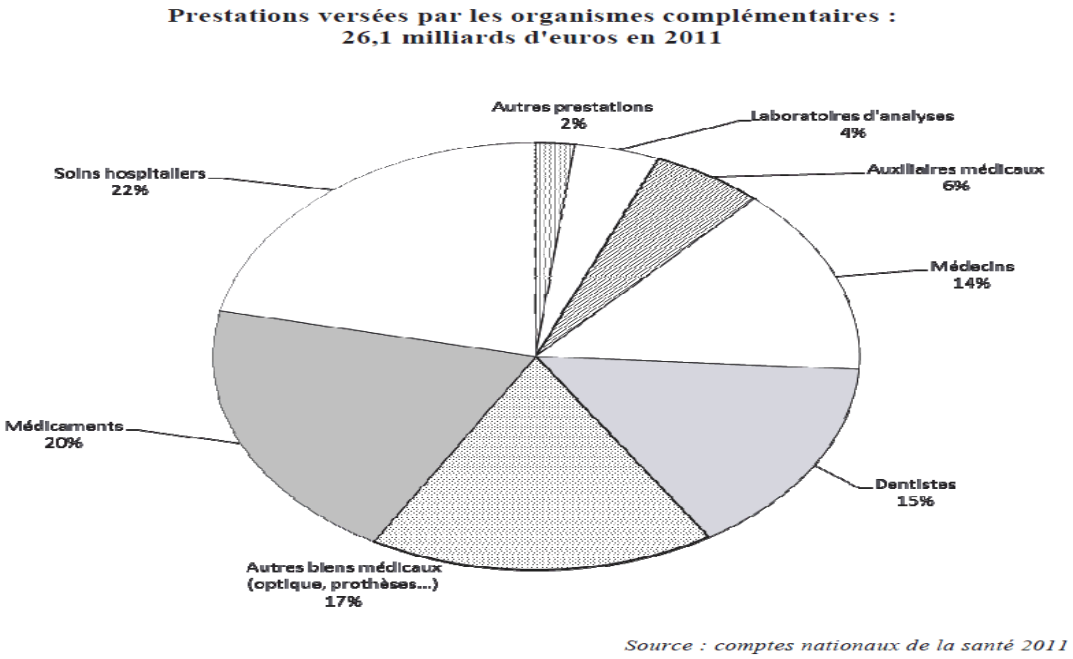


Figure 4. Répartition par poste des prestations versées par les complémentaires

4. Montée en puissance des IP

Les instituts de prévoyance progressent fortement dans le secteur de la santé, et ce depuis déjà les années 1990. Elles bénéficient d'un avantage non négligeable par rapport aux autres organismes car elles touchent à toutes les branches professionnelles, par le biais des Conventions Collectives Nationales notamment. Grâce à ces portefeuilles en collectifs, les instituts de prévoyance disposent d'une imposante base de données sur leurs assurés et donc ont une meilleure vision et interprétation de la sinistralité en France. Ceci leur confère un net avantage par rapport à leurs concurrents. Néanmoins, les nouvelles règlementations et obligations telles que Solvabilité II obligent les instituts de prévoyance à former des partenariats et/ou fusion afin de bénéficier d'une main d'œuvre qualifiée (actuaires, ingénieurs) et rivaliser contre les grosses compagnies d'assurance.

D. Le système de remboursement

a) Les mécanismes

Les dépenses de soins peuvent être remboursées à toute personne assurée du fait de sa propre activité professionnelle et à ses ayants – droits, à toute personne percevant des allocations (pour accident du travail, invalidité, chômeurs et autres). Ce remboursement correspond à ce qu'on appelle des prestations en nature. Si un individu ne dépend d'aucun régime obligatoire particulier, il est alors automatiquement affilié au régime général de la Sécurité Sociale. Cette affiliation peut être même gratuite, pour ceux ne disposant pas de ressources financières suffisantes, dans le cadre de la Couverture Maladie Universelle.
Ces prestations en nature dépendent de trois éléments majeurs qui sont le type de prestations, le tarif pratiqué et le taux de remboursement du régime. Elles sont calculées à partir de tarifs conventionnels, nommés Base de Remboursement. La différence entre cette base de remboursement (BR) et le

remboursement effectif de la sécurité sociale s'appelle le ticket modérateur. Ce dernier peut alors être pris en charge par la complémentaire santé à laquelle est adhérent l'individu.

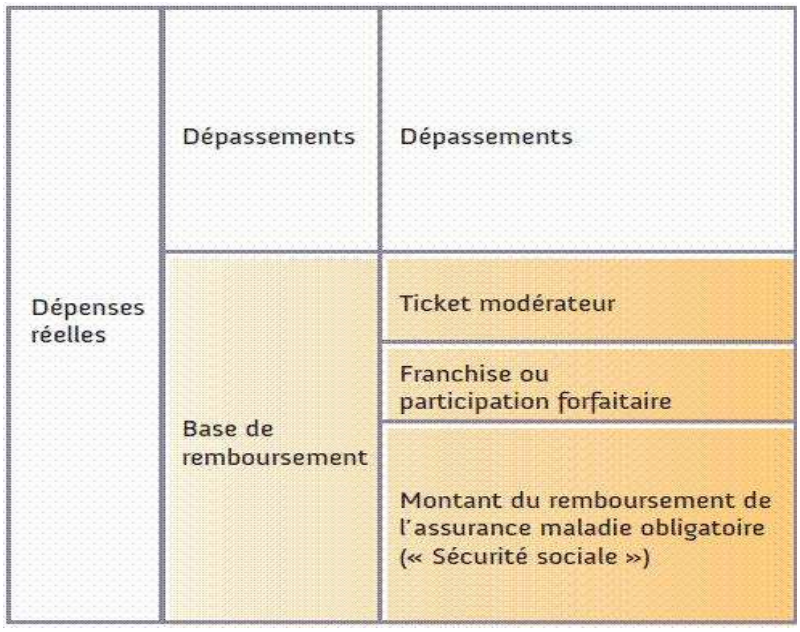


Figure 5. Structure du système de remboursement

Le schéma ci – dessus permet de bien saisir le fonctionnement de remboursement et dans quelle mesure chaque acteur intervient. Le remboursement proposé par la sécurité sociale s'échelonnant sur plusieurs taux, n'est que partiel et laisse donc à la charge de l'assuré une part plus ou moins importante de la dépense. Cette base de remboursement dépendra en particulier de plusieurs variables telles que le régime duquel dépend l'assuré (général, alsace/moselle ou autres), l'importance de l'acte ou encore l'état de santé de l'assuré (maladies graves, recours à des médicaments irremplaçables). C'est dans cet intervalle qu'interviennent les organismes complémentaires. Leurs objectifs est de limiter ce reste à charge pour l'assuré et même, pour des niveaux de garanties plus élevés, de prendre en charge les dépassements d'honoraires. Il est à noter que le remboursement total versé par tous les organismes ne peut jamais excéder les frais réels engagés par l'assuré. Ceci vérifiant le principe que l'assuré ne peut et ne doit pas s'enrichir grâce à sa consommation de soins.

b) La notation des remboursements de la Sécurité Sociale

Les notations permettant de définir le type de remboursement, et qui seront utilisées telles quelles dans la suite de ce mémoire, sont les suivantes :

- FR : Frais réels engagés par l'assuré
- BR : Base de Remboursement de la Sécurité Sociale (appelé Tarif de Convention pour les actes médicaux).
- TRSS : Taux de remboursement appliqué sur la BR par la Sécurité Sociale.
- RSS : Remboursement de la Sécurité Sociale (= TRSS * BR).

- TM : Ticket Modérateur (= BR – RSS)
- RC : Remboursement de la Complémentaire
- RAC : Reste à Charge (= FR – RSS – RC)

Le régime général verse la majorité de ses prestations en nature dans cinq postes différents : les soins de ville (consultations spécialistes, radiologie…), la pharmacie, l'hospitalisation (frais de séjour, honoraires médicaux et chirurgicaux…), le dentaire et l'optique. Les taux de remboursement de la Sécurité Sociale sont présentés dans la grille suivante.
La sécurité sociale utilise un taux de remboursement assez élevé puisqu'il se situe aux alentours de 65% pour pratiquement tous les actes. Néanmoins la BR, elle, n'est pas forcément élevée et le remboursement final ne sera donc pas très important. Ainsi, au niveau du dentaire et de l'optique, les montants engagés par l'assuré sont bien au-dessus de la BR. C'est pour cette raison que les assureurs essaient de proposer des garanties attractives dans ces deux familles d'actes.
Les figures 7 et 8 confirment d'ailleurs cette approche car on remarque que le principal des dépenses de la sécurité sociale se concentre sur les soins de villes (médecins généralistes, analyses…) et les frais d'hospitalisation.

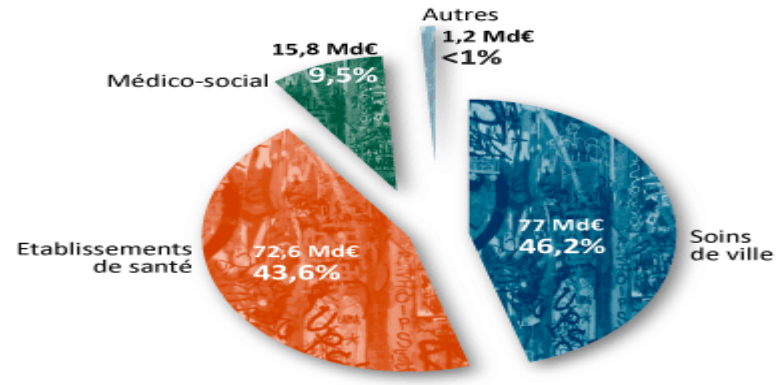
1er novembre 2013

Actes	Tous régimes (hors CMU)	Alsace / Moselle
Hospitalisation		
Soins, honoraires et frais de séjour	**80%** des tarifs conventionnés	**100%** des tarifs conventionnés
Forfait journalier	**0%** des tarifs conventionnés	
Médecine		
Consultation généraliste et spécialiste	**70%** de 23,00 € soit 15,10 €*	**90%** de 23,00 € soit 19,70 €*
Pharmacie		
Médicaments (vignette blanche barrée)	**100%** du prix de vente ou du TFR**	**100%** du prix de vente ou du TFR**
Médicaments (vignette blanche) et préparations magistrales (PMR)	**65%** du prix de vente ou du TFR**	**90%** du prix de vente ou du TFR**
Médicaments (vignette bleue)	**30%** du prix de vente ou du TFR**	**80%** du prix de vente ou du TFR**
Médicaments (vignette orange)	**15%** du prix de vente ou du TFR**	**15%** du prix de vente ou du TFR**
Préparations magistrales (PM4) et médicaments homéopatiques (PMH)	**30%** du prix de vente ou du TFR**	**80%** du prix de vente ou du TFR**
Dentaire		
Consultations chirurgien-dentiste	**70%** de 23,00 € soit 16,10 €	**90%** de 23,00 € soit 20,70 €
Soins	**70%** du tarif conventionnel du soin	**90%** du tarif conventionnel du soin
Prothèse	**70%** du tarif conventionnel de la prothèse	**90%** du tarif conventionnel de la prothèse
Optique		
Consultations ophtalmologue	**70%** de 23,00 € soit 15,10 €*	**90%** de 23,00 € soit 19,70 €*
Lunettes - Monture (adulte)	**60%** de 2,84 € soit 1,70 €	**90%** de 2,84 € soit 2,56 €
Lunettes - Verres	**60%** du tarif de base de remboursement	**90%** du tarif de base de remboursement

* Les montants remboursés indiqués tiennent compte de la participation forfaitaire de 1 € retenue sur chaque consultation ou acte réalisé par un médecin, sauf pour les personnes de moins de 18 ans, les femmes enceintes à partir du sixième mois de grossesse jusqu'à douze jours après l'accouchement, les bénéficiaires de la CMU complémentaire ou de l'aide médicale de l'État.

** TFR : Tarif Forfaitaire de Responsabilité (calculé à partir des médicaments génériques les moins chers).

Figure 7. Taux de remboursement du régime général et du régime Alsace Moselle



Source : Comité d'alerte sur l'évolution des dépenses d'assurance maladie. Mai 2012

Figure 8. Répartition par famille des dépenses des complémentaires

Le faible remboursement proposé par le régime général en optique et en dentaire entraine de forts dépassements d'honoraires dans ces postes. En dentaire, il s'agit particulièrement des soins en prothèses dentaires et en orthodontie.

c) La notation des remboursements des complémentaires

Les organismes complémentaires présentent à leurs futurs clients une grille de garanties, indiquant les taux de remboursement ou forfaits qui seront appliqués pour chaque acte médical. Ces taux et forfaits sont listés ainsi :

- % BR
- % PM
- % PMSS (Plafond Mensuel de la Sécurité Sociale)
- % RSS
- Forfait

Les grilles de garanties peuvent, au choix de l'assureur, proposer ces remboursements en compléments de la Sécurité Sociale ou y compris Sécurité Sociale. Dans les deux cas, cela doit être bien précisé afin que l'assuré ne soit pas induit en erreur.

E. L'assurance Collective

L'assurance collective est à différencier de l'assurance individuelle car, bien qu'elle propose les mêmes services, la mise en place et le système de règlementation sont différents.

a) Les contrats collectifs

Les contrats d'assurance collective s'adressent aux entreprises et couvrent un ensemble de personnes. Ils doivent être établis soit par convention collective de branche (CCN) ou accord collectif d'entreprise, soit par accord par référendum. Une troisième possibilité permet la mise en place du contrat et se traduit par une décision unilatérale de l'employeur, à condition que tous les bénéficiaires aient été informés par écrit. De plus l'entreprise souscriptrice peut choisir entre deux modes d'adhésion différents. Une adhésion facultative, qui laisse le choix aux salariés d'être couvert par l'organisme assureur ou non. Une adhésion obligatoire. Les salariés sont automatiquement affiliés à l'assureur défini par l'entreprise. Les contrats obligatoires doivent couvrir le collège salarial de manière objective, sans différencier les salariés par leur revenu, leur âge, leur sexe ou tout autre facteur discriminant. L'assureur doit proposer un tarif identique au sein d'une même catégorie professionnelle. Deux cadres de la même entreprise devront donc obligatoirement être couverts par le même contrat. Un cadre et un non cadre par contre pourront dépendre de deux contrats différents. L'assureur ne peut exclure personne du contrat, peu importe l'état de santé d'un individu en particulier ou tout autre facteur de risque. On parle donc de mutualisation au sein des « catégories objectives » de l'entreprise.
Lorsque l'entreprise souscrit un contrat en facultatif, le tarif proposé par l'assureur doit alors être étudié plus précisément. En effet, il apparait qu'un salarié souscrivant de lui-même au contrat, dans le cadre d'un contrat facultatif, consommera plus qu'un autre salarié étant obligatoire affilié au contrat. Il est ainsi important pour l'assureur de quantifier l'impact d'une souscription en facultatif sur la consommation. Cet impact est plus connu sous le nom de risque d'anti – sélection et sera plus détaillé dans la suite de ce mémoire. De plus, comme le montre le graphe ci-dessous, le ratio sinistre sur primes (S/P) est beaucoup plus élevé pour les contrats collectifs, et implique donc que l'analyse du risque n'est pas optimale.

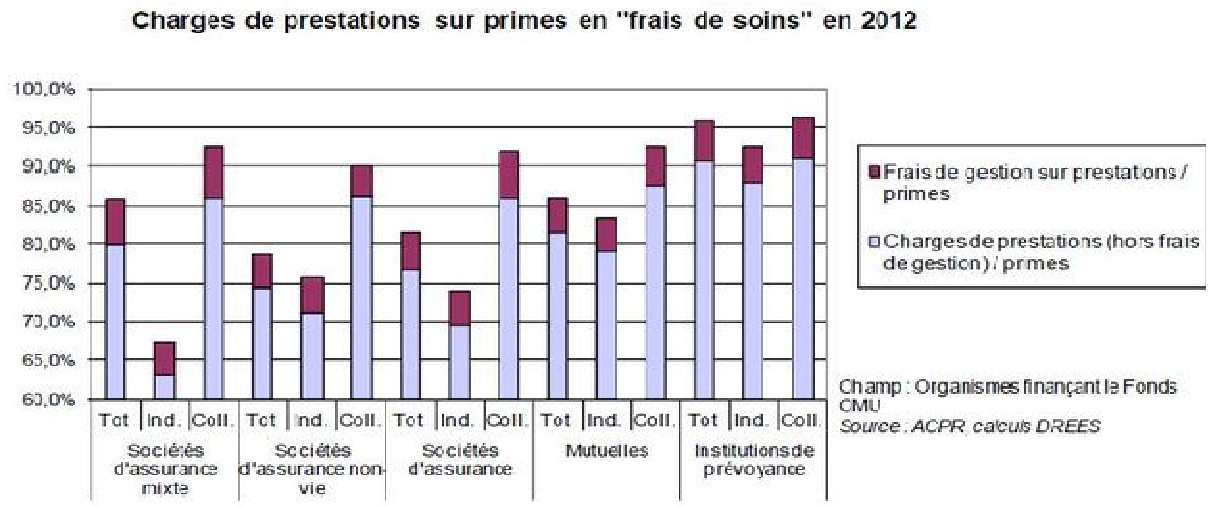


Figure 9. Ratios prestations sur primes

b) Les structures de cotisations

Une entreprise souscrivant à un contrat peut choisir le type de cotisation que le salarié devra payer. Les structures que l'on retrouve essentiellement sont les suivantes :

- La cotisation familiale : chaque salarié paie le même montant de cotisation, indexé sur le PMSS, sans tenir compte de sa composition familiale. La cotisation que verse l'assuré permet donc de couvrir toute sa famille. La famille à charge étant celle au sens de la sécurité sociale ainsi que le conjoint ou concubin. Ces derniers peuvent donc être doublement couvert et par cette couverture et par leur propre couverture s'ils en disposent d'une. Si l'entreprise ne propose que ce type de cotisation à ses salariés, on l'appelle alors une cotisation uniforme, car répartie également pour l'ensemble du collège.

- La cotisation « Famille à charge au sens de la Sécurité Sociale » : ce type de cotisation a pour principal objectif de limiter les dépenses de santé en empêchant la double couverture. En effet, avec cette cotisation le salarié couvre toujours sa famille excepté les conjoints et concubins qui ont déjà une couverture santé par la Sécurité Sociale.

- La cotisation « Isolé » : le salarié paie cette cotisation afin de ne couvrir que lui uniquement. Elle convient donc parfaitement aux célibataires.

- La cotisation « Par tête » : le salarié paie cette cotisation qui ne couvre que lui uniquement encore. Néanmoins, il a la possibilité de couvrir des personnes supplémentaires, enfants et/ou conjoint, en payant la cotisation supplémentaire adéquate. Les cotisations supplémentaires seront majorées par l'assureur par un coefficient d'anti sélection. En effet on peut penser que si le salarié tient à couvrir une tierce personne, c'est qu'il sait que celle – ci aura des dépenses santé dans l'année contre lesquelles se prémunir. Ces cotisations sont proposées dans une structure Adulte/Enfant.

- La cotisation différenciée : l'entreprise laisse au salarié la possibilité de choisir entre une cotisation « Isolé » ou « Famille ». Il s'agit d'une structure dite Isolé/Famille.

III. Tarification

A. La prime commerciale

Comme nous avons pu le comprendre dans les chapitres précédents, le rôle de la complémentaire est de fournir une couverture financière à l'assuré, ceci en contrepartie d'un versement périodique. Ce versement, appelé prime commerciale, doit être calculé avec grande précision car l'ensemble des primes acquises doivent être suffisantes pour couvrir tous les engagements pris par l'assureur envers ses clients. Cette cotisation peut être définie telle quelle :

Prime commerciale = Prime Pure + Frais de chargement (frais de gestion, frais d'acquisition, frais assureurs) + Taxes (TSCA, CMU) + Marge

Ces trois valeurs correspondent ainsi respectivement à la consommation moyenne de l'assuré en santé, le chargement et la marge commerciale que souhaite l'assureur. Nous nous posons alors la question : comment l'assureur peut – il déterminer cette consommation moyenne ? En effet cette question est tout à fait légitime car l'assureur n'a aucune connaissance de la nature, de la date et du montant des sinistres futurs. La modélisation du risque réel auquel il est confronté est donc indispensable de la part de l'assureur.

B. La connaissance du risque réel

Dans ce chapitre, nous mettons en relief l'importance d'une connaissance approfondie des risques couverts par l'assureur.

a) Identification des risques

L'assureur intervient dans le cycle de remboursement afin de limiter ou même éliminer les dépenses de l'individu. Le principe d'assurance est bâti sur le principe de mutualisation du risque. Comme pour le risque Auto par exemple, l'assureur se base sur le fait que tous ses assurés ne consommeront pas tous au même moment et pour des montants importants. Néanmoins, cette mutualisation ne peut être totale, sous peine de perdre certains bons risques au détriment de l'arrivée de mauvais risques : cas que l'on évoquera plus précisément dans les chapitres suivants. Il est ainsi important pour l'assureur de mieux segmenter son risque. Pour cela, ce travail doit nécessairement passer par une phase d'identification des risques. En effet, deux assurés d'âge différents n'auront surement pas besoin des mêmes types d'actes. Une personne myope n'ira pas chez l'ophtalmologue à la même fréquence qu'une autre non myope. D'autre part, la composition familiale diffère entre chaque individu. Or si l'on pratique une mutualisation totale, les personnes seules doivent alors financer les dépenses de santé des familles. Et de même pour les jeunes qui financent celles des personnes âgées.

b) Le risque d’anti – sélection

Il est moins ressenti en santé collective dans la mesure où le contrat couvre un ensemble de salariés. Ceux-ci ne sont pas forcément concernés par les mêmes besoins. Cependant il peut être identifié et analysé dans le cadre des adhésions facultatives (Contrat de base facultatif mis en place par l’entreprise, surcomplémentaire souscrite par le salarié). Cela se traduit par le fait que le salarié souscrit au contrat facultatif car il a conscience de ses futurs besoins. Un salarié n’ayant aucun problème de vue ne souscrira pas à une surcomplémentaire sur l’optique. De même si l’entreprise exige dans le contrat des garanties spécifiques, on peut prévoir dans une moindre mesure l’intégration d’une anti sélection. Car si une entreprise demande une garantie qui n’est généralement pas demandée, on peut supposer qu’il y a de fortes chances que l’employeur veuille couvrir ses salariés contre un risque auquel ils sont sujets. Idem si l’entreprise exige des niveaux de garanties élevés pour telle ou telle garantie.
Au final ce risque est donc dû à une dissymétrie d’information entre l’assureur et les assurés. L’assureur grâce à son historique, doit quantifier l’impact sur la consommation en santé généré par ce type d’assurés. Cette estimation des écarts est de prime importance. Imaginons le cas où l’assureur établirait un tarif sans segmenter par rapport à ce risque. Les bons risques jugeraient le tarif trop élevé par rapport à leur consommation et ne viendraient pas souscrire à notre contrat. À l’opposé les mauvais risques, sachant mieux leur probable consommation future, jugeraient ce tarif avantageux.

c) L’aléa moral

L’aléa moral peut être défini comme l’entend Adam Smith « la maximisation de l’intérêt individuel sans prise en compte des conséquences défavorables de la décision sur l’utilité collective ». C'est-à-dire que l’assuré peut agir très différemment lorsqu’il est isolé ou exposé au risque. Dans le cadre de l’Assurance Santé, cela se traduit donc par le fait que l’assuré n’adoptera pas le même comportement en fonction de son niveau de couverture ainsi que son niveau de revenu. En effet le premier diminue le comportement de prévention de l’assuré qui est certain d’être bien remboursé en cas de problème de santé. Le deuxième quant à lui incite l’assuré à consommer plus souvent car le reste à charge lui est moins contraignant.

d) Deux facteurs discriminants majeurs

Le sexe et l’âge sont deux variables que l’on retrouve classiquement lorsqu’il s’agit de modéliser la consommation d’un individu. On parle de critère de segmentation, sujet sur lequel nous reviendrons plus tard. Ces critères ont l’avantage d’être faciles à utiliser par les commerciaux car très vite appréhendés et compris par les clients. En effet, nous sommes tous conscients que les besoins en santé augmentent avec l’âge, et que ces besoins diffèrent entre hommes et femmes (gynécologue pour le sexe féminin).
L’assureur dispose bien souvent d’un historique assez conséquent lui permettant d’établir les bons critères de segmentation. Mais le principe de mutualisation, inhérent au domaine de l’assurance est alors à prendre en considération par l’assureur.

C. Segmentation vs Mutualisation

La mutualisation est sous-jacente au principe de solidarité, avec des tarifs similaires pour tous les assurés. Cependant le risque d'anti sélection est donc très fort car la population n'est pas nécessairement homogène en ce qui concerne sa consommation en soins médicaux. A l'inverse, la segmentation est sous-jacente au principe de rentabilité. L'assureur qui segmente sa clientèle établit des tarifs adaptés pour chaque profil et sécurise sa marge de solvabilité. Le danger est alors dans ce cas de proposer des tarifs trop élevés par rapport à la concurrence et de perdre en compétitivité.
Pour bien comprendre le principe de mutualisation et celui de segmentation, nous allons traiter d'un exemple classique, où deux assureurs se répartissent le marché, l'un choisissant une mutualisation totale et l'autre une segmentation totale.
Le marché est composé de P individus.
La population masculine est composée de P1 individus. La population féminine, elle, est composée de P2 individus. La population totale consulte un généraliste avec une fréquence F et un coût moyen C.
De même, on prévoit que la population masculine (respectivement féminine) consulte un généraliste avec une Fréquence F1 (respectivement F2) et pour un coût moyen C1 (respectivement C2). On suppose que pour ces personnes on a :

$$F_1<F<F_2 \text{ et } C_1<C<C_2$$

A étant l'organisme assureur privilégiant la mutualisation totale, la prime pure est unique et est établie telle quelle :

$$\text{Cotisation (A)} = F \times C$$

B étant l'organisme assureur segmentant totalement sa population, deux primes pures seront calculées cette fois :

$$\text{Cotisation } (B_1) = F_1 \times C_1 \rightarrow \text{ cotisation pour les hommes}$$
$$\text{Cotisation } (B_2) = F_2 \times C_2 \rightarrow \text{ cotisation pour les femmes}$$

Deux cas de figures peuvent se présenter :

- A est seul organisme assureur sur le marché et bénéficie d'aucune concurrence. Il récupère donc toute la population P en tant que clientèle. Néanmoins, son résultat technique est alors nul. Car le tarif moyen proposé correspond aux dépenses moyennes qu'il devra verser.

- B décide d'entrer aussi dans le marché et de concurrencer A. Chaque personne peut alors choisir entre ces deux sociétés puisqu'elles ne lui proposent pas les mêmes tarifs. On remarque sans difficulté que le tarif proposé pour les femmes chez la société A est plus faible que celui proposé par B : les femmes iront donc chez A. Et inversement pour les hommes qui iront, eux, souscrire chez la société B. Au final, A ne récupérant que les femmes, la sinistralité à laquelle elle fera face sera supérieur au montant des cotisations reçues. Son résultat sera ainsi déficitaire.

$$\text{Résultat technique} = \text{Primes} - \text{Sinistres} = F \times C - F_2 \times C_2 < 0$$

Pour la société B en revanche, le résultat est nul car la prime établie est celle couvrant exactement les sinistres futurs.

De cet exemple ressort qu'une mutualisation totale dans un environnement concurrentiel est impossible pour un assureur. Pour autant, une segmentation totale n'est pas non plus la solution idéale car elle fait perdre à l'entreprise de nombreux clients. Allier correctement mutualisation et segmentation est donc un enjeu considérable pour les complémentaires. Elles doivent définir les caractéristiques économiques, démographiques et sociales afin d'adapter au mieux leur ratio performance/compétitivité.

D. Segmenter la clientèle

Chaque assuré désire se couvrir contre les pertes financières engendrées par ses besoins en santé. Mais chaque assuré espère ne pas payer plus que les dépenses auxquelles il serait engagé sans assurance. L'assureur, afin d'optimiser la performance du groupe, tout en satisfaisant la volonté des clients, doit donc établir le tarif adéquat pour chaque client. Or, cela est bien entendu complexe compte tenu de la diversité de la population, de ses besoins, de son aversion au risque et sa capacité de financement. L'assureur va donc procéder à une segmentation de sa clientèle par rapport à des variables jugées, par la théorie et l'historique, pertinentes pour prévoir la consommation du client : ces variables sont aussi appelées facteurs discriminants. Une fois segmentée en différentes classes homogènes, la sinistralité est alors mutualisée au sein de ces classes.

a) L'âge

Comme énoncé plus haut, l'âge est un facteur discriminant communément utilisé en tarification. En santé, c'est le facteur le plus discriminant car plus une personne vieillit, plus sa santé se dégrade, et plus elle est susceptible de consommer en soins médicaux. Ce facteur est d'ailleurs encore plus important qu'il ne l'était auparavant avec pour principales raisons l'augmentation de l'espérance de vie et l'amélioration des techniques et services médicaux. Or, dans notre étude, nous traitons de contrats obligatoires, pour lesquels une différenciation sur critères non objectifs n'est pas autorisée par la règlementation. Comment l'assureur peut – il donc établir un tarif ajusté à l'âge ? Les équipes actuarielles ont pour mission de calculer des primes en fonction de l'âge puis de récupérer celles correspondant à l'âge moyen de l'entreprise. L'entreprise fournit donc cette information à l'assureur afin d'obtenir un tarif fixe pour tous les salariés d'un même collège.

b) Le sexe

Nous avons déjà donné les raisons qui poussent à retenir cette variable en santé. La suite de notre mémoire, notamment la partie portant sur l'analyse exploratoire de la consommation des assurés permettra de confirmer l'importance de ce facteur. Nous ajouterons simplement qu'à l'instar de l'âge, le tarif est déterminé en fonction de la répartition Hommes/Femmes du collège à assurer.

c) Le statut socioprofessionnel

La catégorie socioprofessionnelle est une variable fournissant une information conséquente sur la consommation de l'individu. Deux éléments influençant cette consommation sont facilement intuitifs : le

niveau de revenu et le niveau de couverture. Le niveau de revenu est, logiquement, source de disparité au niveau des dépenses engagées par l'assuré. En effet, le reste à charge sera plus ou moins un frein financier entre deux individus n'ayant pas les mêmes ressources. De plus, le revenu allant souvent avec le statut professionnel, le niveau de couverture proposé est aussi souvent supérieur. L'entreprise va proposer des garanties supérieures aux catégories professionnelles ayant plus de ressources financières. Nous verrons l'influence des niveaux de couverture dans le paragraphe qui suit.

d) Le niveau de garantie

Le niveau de garantie est un facteur déterminant car plus le niveau de garantie est élevé, plus le client peut consommer sans crainte de pertes financières. Pour un généraliste, il pourrait consulter un médecin du secteur trois (pratiquant des dépassements d'honoraires) si son niveau de couverture est élevé. Alors qu'avec un niveau plus faible il limiterait ses consultations uniquement aux généralistes du secteur un, ne pratiquant aucun dépassement. De plus, les soins médicaux consommés seront plus sophistiqués si la couverture est importante. L'assuré n'aura aucun frein financier et acceptera plus facilement une opération chirurgicale du genou par exemple, avec pour objectif de se rétablir au plus vite. Un autre assuré moins couvert privilégierait lui plus un repos prolongé, suivi de séances d'étirements personnels ou de kinésithérapie : moins coûteux mais plus long.

e) La surcomplémentaire

L'influence d'une surcomplémentaire sur la consommation d'un individu a été jusqu'à aujourd'hui très peu traitée. D'une part car les niveaux de couverture des contrats collectifs étant souvent particulièrement élevés, la souscription en surcomplémentaire restait très marginale. D'autre part car l'influence d'une surcomplémentaire est souvent déjà intégrée dans l'influence du niveau de garantie. Avec les nouvelles règlementations imposées par l'ANI, les organismes assureurs vont devoir proposer de plus en plus de contrats en montage Base plus Surcomplémentaire. En effet, ces contrats auront pour objectif de respecter les critères de « contrat responsable » se traduisant ainsi par une base relativement faible. Pour combler ce manque de couverture, les individus auront alors beaucoup plus recours aux surcomplémentaires mises à leur disposition par l'assureur. Plusieurs éléments de travail se présentent dorénavant aux actuaires. En effet, prédire l'impact d'une souscription en surcomplémentaire demande une analyse et une compréhension du comportement de l'individu en fonction du produit auquel il souscrit. C'est sur cet aspect que portera principalement notre étude.

f) Une liste non exhaustive

D'autres facteurs discriminants sont connus tels que la région, le secteur d'activité et le régime auquel appartient l'assuré. D'autres facteurs, plus ou moins intuitifs mais pas nécessairement utilisés, ont une influence sur la consommation. La pluviométrie ou la densité en médecins sont de tels facteurs. Enfin, certains facteurs, difficilement quantifiables et/ou perceptibles influent sur la consommation. Nous pourrions citer par exemple une politique de prévention spécifique dans chaque région.

E. Méthode de calcul des primes pures

La tarification consiste à calculer les primes pures pour chaque profil de risque. Ces primes pures représentent les dépenses moyennes couvertes par l'assureur. Le cycle de production d'un assureur étant inverse (paiements reçus avant livraison d'un service), il est difficile pour l'assuré de calculer ces primes pures avec précision et plusieurs méthodes existent actuellement chez les différents assureurs. Néanmoins l'équation que cherche à vérifier l'assureur reste identique, à savoir :
Λ =E(S), avec S étant la charge sinistre totale susceptible d'être payée par l'assureur pour un acte défini et Λ étant la prime pure que devrait en théorie payer l'assuré - sans tenir compte du remboursement de la sécurité sociale - afin que l'assureur soit capable d'effectuer les remboursements nécessaires.
La démonstration est décrite ci-dessous de manière assez brève car elle peut être faite facilement.
La prime pure devant couvrir les engagements futurs de l'assureur envers l'assuré, cela se traduit par la minimisation de l'écart quadratique moyen E $[(S - \Lambda)^2]$:

$$\begin{aligned} E\,[(S - \Lambda)^2] &= E[(S - E(S) + E(S) - \Lambda)^2] \\ &= E\,[(S - E(S)^2] + [(E(S) - \Lambda)^2] + 2 \times E\,[(S - E(S)] \times (E(S) - \Lambda) \\ &= E\,[(S - E(S)^2] + [(E(S) - \Lambda)^2] \end{aligned}$$

$E\,[(S - \Lambda)^2]$ *est donc minimale pour* $\Lambda = E(S)$.

Dans les chapitres suivants, nous proposons une brève présentation des méthodes qui peuvent être utilisées pour réaliser une tarification. Nous ne détaillerons réellement que les méthodes que nous utiliserons pour notre étude.

a) Méthode directe

La méthode directe est celle qui est actuellement utilisée par le service Santé Collective et sera donc expliquée dans les paragraphes qui suivent.

1. Modélisation de la charge totale

Ce modèle correspond au modèle le plus simple à mettre en œuvre. Il se restreint à la formule de base énoncée dans le paragraphe précédent, et se calcule directement sur la somme totale déboursée par l'assureur et le nombre total d'assurés couverts. Généralement ce calcul va se faire pour chaque garantie et pour chaque niveau de couverture qu'offre l'assureur. L'actuaire, s'il se restreint uniquement à cette méthode, se focalise donc sur une forte mutualisation puisqu'au final il ne « segmente » sa population que par rapport au niveau de garantie auquel ont souscrit les assurés. La formule est donc de la sorte :

$$P_{ij} = \frac{X_{ij}}{N_{ij}}$$

Avec :

- i correspondant à la garantie i.
- j correspondant au niveau de couverture j.
- X_{ij} les prestations totales versées par l'assureur pour la garantie i au niveau de couverture j.
- N_{ij} le nombre d'assurés couverts sur la garantie i par le niveau de couverture j.
- P_{ij} la prime que doit payer un assuré souscrivant à la garantie i pour un niveau de couverture j.

2. Conclusion

L'avantage majeur de ce modèle est qu'il ne nécessite aucune hypothèse faite au préalable par l'assureur. C'est la méthode qu'a choisi AG2R La Mondiale pour tarifer sa dernière gamme de produits en 2011. Le service Santé Collective récupère pour chaque garantie les niveaux de couverture proposés par le groupe. Pour chaque niveau de couverture et pour chaque garantie, le rapport entre la somme des prestations et le nombre d'assurés est calculé, et constitue ainsi la prime pure. Néanmoins, l'assureur devra appliquer à cette prime des coefficients de réajustement (obtenus grâce à l'expérience et l'historique du portefeuille par exemple) afin d'affiner le tarif aux profils de risques couverts. En ce qui concerne notre problématique principale, à savoir l'influence d'une surcomplémentaire sur la consommation, cette méthode ne peut pas être appliquée pour plusieurs raisons. La première est qu'elle ne permet pas d'identifier la différence de comportement des assurés, mais uniquement le coût final auxquels ils sont sujets. La seconde est qu'il est difficile de tenir compte du temps pendant lequel on a couvert l'assuré, ce qui fausse les résultats. De plus, cette méthode nécessite une extraction de tous les niveaux de couverture, ce qui représente un temps de traitement des données énormes. En effet, dans notre base de données, nous disposerons d'un grand nombre de produits différents, différents pour chaque CCN étudiée. Enfin, même si ce travail de traitement de données peut être réalisé, il n'apportera pas une information claire car couvertures de base et couvertures de surcomplémentaire seront mélangées. Un contrat de base pouvant très bien proposer un niveau de couverture pour un certain acte, supérieur à celui proposé par un autre contrat de surcomplémentaire.

b) Modèle Probabiliste

Dans cette partie, nous traitons le cas d'une méthode assez intéressante dans son aspect théorique mais qui reste en pratique très peu utilisée.

1. Modélisation de la probabilité de consommer et du coût moyen associé

Notons R_{ik} le remboursement annuel probable versé par l'assureur à l'assuré i pour la garantie k; p_{ik} la probabilité que cet individu ait consommé au moins une fois cet acte de garantie k ; PR_{ik} la variable aléatoire modélisant le montant de la prime pure à payer, sachant que l'assuré a consommé au moins une fois cet acte ; et enfin, P_{ik} la prime pure finale. On peut traduire ce modèle par l'équation suivante :

$$PR_{ik} = R_{ik} \times 1\{i\ a\ consommé\ au\ moins\ une\ fois\ l'acte\ k\}$$

Ainsi :

$$P_{ik} = E\left(PR_{ik}\right) = E\left(R_{ik} \times 1\{i\ a\ consommé\ au\ moins\ une\ fois\ l'acte\ k\}\right) = E\left(R_{ik}\right) \times p_{ik}$$

Cette méthode n'étant pas celle que nous utiliserons, nous ne détaillerons pas tous les calculs.
Dans ce modèle, i étant l'individu de référence on pose :

$$E\,(R_{jk}) \;=\; \alpha_{jk} \;*\; E\,(R_{ik}),$$

R_{jk} ayant la même définition que R_{ik}, mais pour un assuré j, et α_{jk} étant le coefficient correctif.
De même :

$$p_{jk} \;=\; \beta jk \;\times\; p_{ik}$$

Au final, on obtient donc :

$$Pjk = \;(\alpha jk \;\times\; \beta jk) \;\times\; Pik$$

La signification de ces deux correctifs est la suivante :

- α_{jk} mesure les différences sur le coût moyen. Ces différences peuvent être dues à une différence de cout des actes. Par exemple, l'âge est souvent responsable de consommation en besoin plus coûteux. Elles peuvent être dues aussi à une fréquence plus élevée. Au niveau des consultations, le coût de l'acte reste sensiblement similaire, mais une fréquence plus forte provoquera donc une augmentation de la dépense moyenne versée par l'assureur.

- β_{jk} mesure les différences comportementales des assurés. Il permet d'identifier ceux ayant une plus forte probabilité de consommer. En dermatologie, les assurés de sexe féminin auront surement tendance à consommer plus que ceux de sexe masculin.

2. Conclusion

Les principaux avantages de ce modèle est qu'il ne nécessite pas d'hypothèse d'indépendance sur la survenance des sinistres, qu'il permet de prendre en compte assez facilement les éventuelles corrélations entre les différentes variables.
Les inconvénients du modèle concernent en particulier l'interprétation des résultats obtenus. Il ne délivre pas d'information sur ce qui influe réellement sur la consommation totale de l'individu. Nous ne pouvons dire si elle est impactée par une consommation plus fréquente ou par un coût moyen individuel par acte plus élevé. De plus, pour les garanties exprimées en pourcentage de la BR, un calcul additionnel est à réaliser. Au vu des nouvelles règlementations, et des garanties proposées pour la majorité d'entre elles en pourcentage de la BR, ce modèle risque d'être contraignant au niveau des calculs.

c) Modèle classique Fréquence × Coût

Il s'agit actuellement du modèle majoritairement retenu en assurance santé, et la recherche continue de développer des méthodes basées sur ce modèle de plus en plus élaborées.

1. Théorie du modèle

Nous détaillerons ce modèle plus en profondeur que le modèle précédent car c'est celui sur lequel nous allons travailler pour le calcul de nos primes pures. Définissons S la charge sinistre totale d'un assuré pour un acte donné, et Xi le montant du ieme sinistre survenu pour cette garantie. N le nombre de fois où l'assuré a consommé cet acte. La charge sinistre totale est alors donnée par :

$$S = \sum_{i=1}^{N} Xi$$

Afin de pouvoir déterminer l'espérance et la variance de la charge sinistre totale, le modèle nécessite deux hypothèses très fortes :

- Les X_i, qui représentent les coûts de sinistres, sont des variables aléatoires identiquement distribuées et indépendantes.
- L'indépendance entre les coûts de sinistre et la fréquence des sinistres. Cette hypothèse sera discutée plus amplement dans la partie D du chapitre V « Application des méthodes ».

La suite des calculs peut se faire aisément grâce aux hypothèses retenues.

$$E(S) = E_N[E(S|N = n)] = E_N\left[E\left(\sum_{i=1}^{n} X_i\right)\right] = \sum_{n} P(N = n) \times E\left(\sum_{i=1}^{n} X_i\right)$$

$$E(S) = \left(\sum_{n} P(N = n) \times n\right) \times E(X_1) = E(N) \times E(X_1)$$

La charge sinistre est ainsi exprimée très simplement et se calculera en multipliant le coût moyen de l'acte par son taux de recours.

En ce qui concerne la variance de la charge sinistre nous allons partir de l'expression de la variance sous sa forme décomposée conditionnellement à la variable N.

$$V(S) = V[E(S|N)] + E[V(S|N)] = V[N * E(X_1)] + E[V(\sum_{i=1}^{N} X_i|N)]$$

$$V(S) = E^2(X_1) * V(N) + E[N * V(X_1)] = E^2(X_1) * V(N) + E(N) * V(X_1)$$

Avec V étant la variance d'une variable aléatoire.
Le premier terme traduit l'incertitude sur la fréquence des sinistres. Le deuxième terme indique l'incertitude sur les montants de sinistres.

Dans ce modèle, nous avons deux axes d'interprétation :

1. Nous pouvons mesurer les différences de comportements des assurés. L'espérance de la variable N nous indiquera donc la tendance à consommer d'un assuré.

2. Nous pouvons identifier les écarts de coût moyen pour chaque acte.

Dans ce mémoire, le but étant de bien appréhender le comportement d'un assuré et sa consommation en fonction du produit souscrit, l'utilisation de ce modèle parait pertinente. Afin de bien comprendre la théorie de ce modèle et les hypothèses fortes qui en sont à la base, le lecteur peut se référer notamment aux différents travaux de CHARPENTIER Arthur sur sa plateforme de partage freakonometrics.

2. Approche empirique

2.1. Méthode de calcul

Cette méthode a l'avantage d'être simple d'utilisation et ressemble fortement à la première méthode énoncée. La principale différence est ici qu'au lieu de travailler sur la charge totale des sinistres, nous travaillons séparément sur le coût moyen et sur la fréquence. Il suffit de récupérer pour chaque acte, tous les montants de consommations, le nombre de sinistres et le nombre de risques années, appelée l'exposition et correspondant à la période de couverture.
Le coût moyen par acte est alors :

$$CM = \frac{Montant\ annuel\ des\ sinistres}{Nombre\ de\ sinistres}$$

La fréquence de consommation d'un acte est :

$$F = \frac{Nombre\ de\ sinistres}{Nombre\ de\ risques\ années}$$

L'expression de la prime pure est donc déterminée par :

$$C_M \times F = \frac{Montant\ annuel\ des\ sinistres}{Nombre\ de\ risques\ années}$$

2.2. Conclusion

Ces primes sont calculées pour des individus d'une même tranche d'âge et doivent ensuite être lissées de manière à être commercialement cohérentes vis-à-vis des clients.
Nous n'utiliserons pas cette méthode car elle nécessite un calcul empirique de la fréquence et du coût moyen, nous obligeant à constituer plusieurs bases distinctes en fonction des critères que l'on souhaite retenir. De plus, elle ne fournit aucun intervalle de confiance, aspect non négligeable dans une tarification.

3. Modèles linéaires généralisés

Les modèles linéaires généralisés (appelés MLG dans la suite du mémoire) ont été introduits par Nelder et Wedderburn dans les années 1970. Ils sont couramment utilisés en assurance IARD et sont une généralisation de la régression linéaire. Avant d'aller plus loin, nous allons donc procéder à un rappel du modèle linéaire gaussien.

3.1. Régression linéaire

Dans un modèle en régression linéaire, l'objectif de la modélisation est de prédire et expliquer par une relation de type linéaire, une variable dite variable expliquée en fonction d'un ensemble de variables dites variables explicatives. Ce modèle propose donc une équation de la forme :

$$Y = \alpha_0 + \sum_j^p \alpha_j X_j + \varepsilon$$

Avec :

- Y la variable à expliquer (dite aussi variable réponse)
- X_i la i^{eme} variable explicative
- p le nombre de variables explicatives
- α_o l'ordonnée à l'origine
- α_j les paramètres à estimer par le modèle. C'est-à-dire les coefficients dont la valeur détermine l'impact de chaque variable explicative sur la consommation de l'individu
- ε l'erreur associée à cette régression linéaire. Il s'agit de l'écart entre la valeur réelle observée de Y et sa valeur théorique obtenue par le modèle. Cette erreur ε est une variable aléatoire suivant une loi d'espérance nulle et de variance σ^2.

Les coefficients α et α_o sont estimés grâce à la méthode des moindres carrés, en minimisant donc l'expression :

$$\min_{a_j, a_0} \sum_{i=1}^{n} [y_j - (\alpha_j x_{ij} + \alpha_0)]^2$$

Avec y_j la i^{eme} observation de la variable Y et x_{ij} la i^{eme} observation de la variable explicative X_j.

Ce modèle impose donc une contrainte très forte : les $\varepsilon_i = y_j - a_j x_{ij} + \alpha_o$ sont des variables aléatoire indépendantes et identiquement distribuées selon une loi normale centrée et de variance σ^2. La variable expliquée Y suit une loi normale de moyenne $\alpha_0 + \sum_j^p \alpha_j X_j$ et de variance σ^2.
L'espérance de Y est donc sous forme très simplifiée puisqu'il s'agit encore d'une transformation linéaire. Néanmoins la contrainte sur la variable Y n'est plus adaptée face à la variété et à la complexité des événements touchés par le monde assurantiel. Cette condition est donc d'une part difficile à satisfaire mais nécessite un travail supplémentaire en effectuant des tests de normalité.

3.2. Principe des Modèles Linéaires Généralisés

3.2.1. Un modèle moins contraignant

L'objectif des MLG est de généraliser le modèle gaussien à un ensemble de lois plus large. Cet ensemble de lois qui permet l'application des MLG est la famille des lois exponentielles. Une variable expliquée Y appartient à cette famille si sa densité peut se mettre sous la forme suivante :

$$f(y, \theta, \varphi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\varphi)} + c(y, \varphi)\right\}$$

Avec :

- θ le paramètre canonique (appelé aussi paramètre naturel). C'est un paramètre de position
- φ le paramètre de dispersion. C'est un paramètre d'échelle
- a une fonction définie sur R et non nulle
- b une fonction définie sur R, deux fois dérivable, et sa dérivée première est injective
- c une fonction définie sur R^2

Les lois de Poisson, Binomiale – Négative, Gamma et Log – Normale font notamment partie de cette famille exponentielle.

Dans le modèle MLG, nous n'avons plus $Y \sim N(\mu, \sigma^2)$ mais :

- $Y \sim$ Loi (μ)
- $\mu = E(Y)$
- $E(Y) = g^{-1}(X^t\beta)$
- $V(Y) = a(\varphi) * V(\mu)$

Y et E(Y) ne se traduisent donc plus nécessairement par une transformation linéaire mais c'est g [E (Y)] qui est désormais exprimée par une relation linéaire ($X^t\beta$ dit le prédicteur linéaire) en fonction des variables explicatives. Deux points conséquents sont à noter ici. D'une part la prise en compte de l'hétéroscédasticité puisque Y suivant une loi dépendant de μ, il en est de même de sa variance. D'autre part, nous comprenons que le choix de la loi paramétrique de paramètre μ et de la fonction g^{-1} est le pilier d'une modélisation précise et fiable des risques encourus par l'assureur.
Dans un MLG, les variables explicatives pouvant maintenant être quantitatives ou qualitatives, il faudra parfois procéder à un retraitement des données. Une variable catégorielle X sera ainsi recodée en fonction de ses modalités (valeurs prises) de la façon suivante :

$$z_j = \begin{cases} 1 & si\ la\ modalité\ j+1\ est\ observée \\ 0 & sinon \end{cases}$$

Avec m le nombre de modalités de la variable explicative X, $j \in \{1, 2, \ldots, m-1\}$, et zj la jeme variable auxiliaire prenant ses valeurs en fonction de l'observation ou non de la modalité à laquelle elle est associée.

3.2.2. Propriétés spécifiques

Si Y la variable expliquée appartient à la famille de lois exponentielle alors on a les propriétés suivantes :

$$\begin{cases} E(Y) = b'(\theta) & (1) \\ Var(Y) = a(\varphi) * b''(\theta) & (2) \end{cases}$$

Ces deux propriétés établissent la relation entre μ et θ puisque nous avons donc $\mu = E(Y) = b'(\theta)$. De plus, elles permettent de décomposer la variance en un produit de deux fonctions. $b''(\theta)$, qui est appelée la fonction variance et qui dépend uniquement du paramètre canonique θ. $a(\varphi)$ que nous avons déjà décrite plus haut et qui est indépendante de θ. Enfin, d'après la relation établie entre μ et θ, il est donc justifié d'écrire Var (Y) sous la forme suivante : $Var(Y) = V(\mu) * a(\varphi)$

Démonstration

1. Soit Y une variable expliquée suivant une loi de la forme exponentielle telle qu'exprimée plus haut. L'intégrale de sa densité valant 1, grâce au lemme de Fatou et à la linéarité de l'intégrale nous avons la relation suivante :

$$\text{Lemme de Fatou} \Rightarrow \int_y \frac{\partial f_{\theta,\varphi}}{\partial \theta} dv = \frac{\partial \int_y f_{\theta,\varphi}\, dv}{\partial \theta} = 0$$

En dérivant l'expression de $f(y, \theta, \varphi)$ en fonction de θ et φ, on obtient :

$$\text{Linéarité de l'intégrale} => \int_y \frac{y - b'(\theta)}{a(\varphi)} f_{\theta,\varphi} dv = \frac{1}{a(\varphi)} \left[\int_y y f_{\theta,\varphi} dv - b'(\theta) \int_y f_{\theta,\varphi} dv = 0 \right]$$

D'où finalement : $E(Y) = b'(\theta)$

2. On réitère le même procédé et nous obtenons alors :

$$\int_y \frac{\partial^2 f_{\theta,\varphi}}{\partial^2 \theta} dv = \frac{\int_y \partial^2 f_{\theta,\varphi}\, dv}{\partial^2 \theta} = 0$$

Après dérivation :

$$\int_y \left[\left(\frac{y - b'(\theta)}{a(\varphi)} \right)^2 - \frac{b''(\theta)}{a(\varphi)} \right] f_{\theta,\varphi} dv = \frac{1}{a(\varphi)^2} \int_y [y - E(Y)]^2 f_{\theta,\varphi} dv - \frac{b''(\theta)}{a(\varphi)} = 0$$

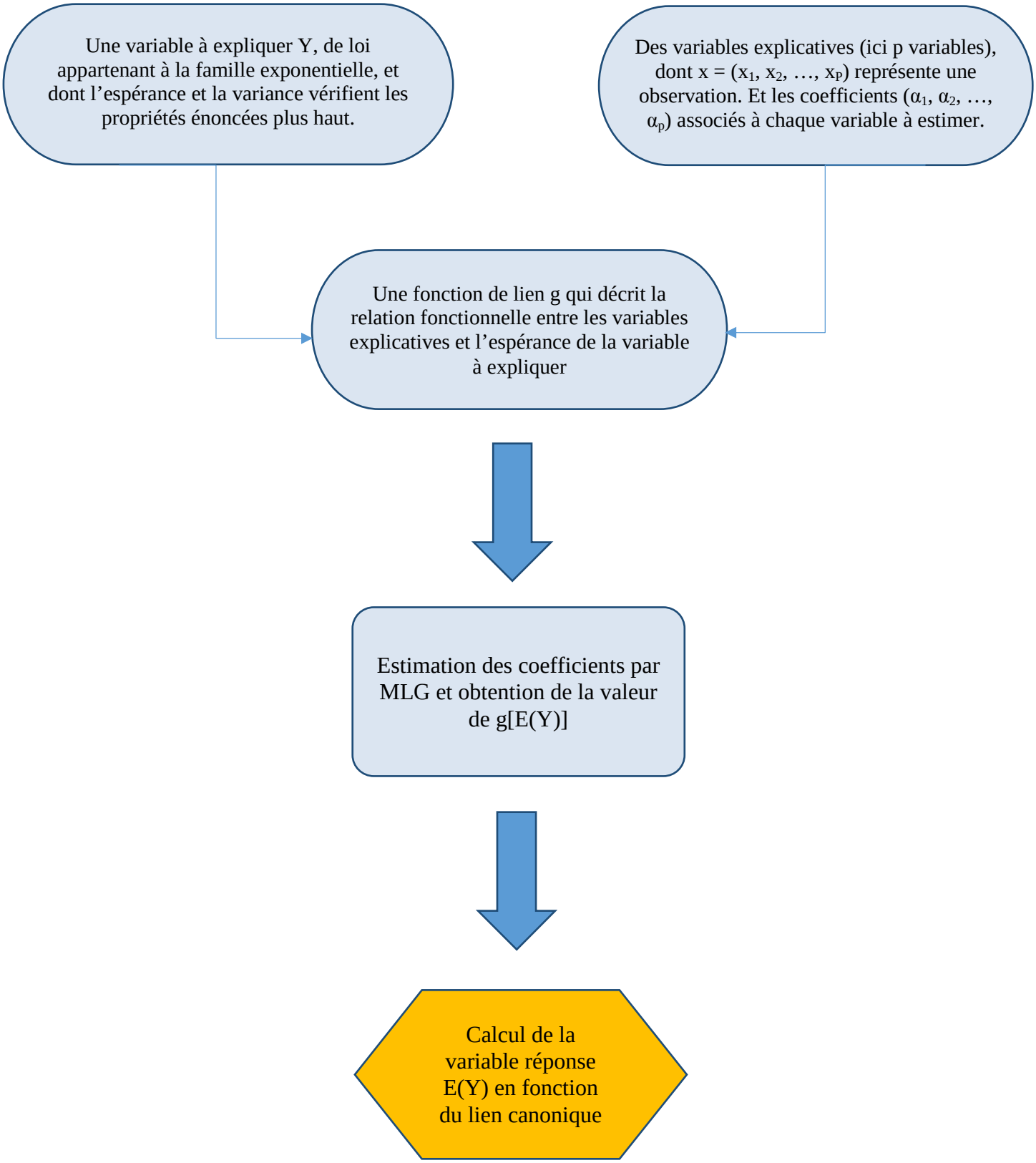
Et enfin, le résultat final :

$$Var(Y) = a(\varphi) * b''(\theta)$$

Comme nous l'avons fait remarquer précédemment, espérance et variance de Y dépendent de θ par l'intermédiaire des fonctions $b'(\theta)$ et $b''(\theta)$. La fonction g est appelée lien canonique lorsqu'elle vérifie la relation $g(\mu) = (b')^{-1}(\mu) = \theta$, avec $b'(\theta)$ injective et donc inversible. φ n'influe que sur la variance de Y par l'intermédiaire de la fonction a(.), qui est en général sous la forme $a(\varphi) = \frac{\varphi}{\omega}$, avec ω représentant le poids sur une observation. Nous fixerons $\omega = 1$ pour la suite de notre description du modèle

3.2.3. Schéma et résumé

Nous pouvons schématiser le processus d'un MLG et l'interaction des composantes principales par le schéma 1 présent sur la prochaine page.


Une variable à expliquer Y, de loi appartenant à la famille exponentielle, et dont l'espérance et la variance vérifient les propriétés énoncées plus haut.
Des variables explicatives (ici p variables), dont x = (x1, x2, …, xp) représente une observation. Et les coefficients (α1, α2, …, αp) associés à chaque variable à estimer.
Une fonction de lien g qui décrit la relation fonctionnelle entre les variables explicatives et l'espérance de la variable à expliquer
Estimation des coefficients par MLG et obtention de la valeur de g[E(Y)]
Calcul de la variable réponse E(Y) en fonction du lien canonique

Ce schéma nous montre encore une fois l'intérêt primordial d'effectuer un choix judicieux concernant la loi de Y et la fonction lien. La loi de Y sera déterminée par le type de la variable et l'historique de l'assureur concernant la survenance de ses sinistres. Il en est de même de la fonction lien. Nous aurons le plus souvent des liens logit ou probit pour une variable réponse binaire, un lien log pour une variable de comptage et les liens canoniques des lois Normales et Gamma - que nous présenterons plus bas dans le chapitre - pour les variables expliquées continues. Afin de simplifier les calculs, la fonction lien choisie peut être la fonction de lien canonique. En assurance, il apparait logique que la fonction de lien utilisée soit le plus souvent la fonction log. Les coefficients trouvés sont alors des coefficients multiplicatifs et permettent une meilleure lisibilité de la part des commerciaux. En effet, ces coefficients peuvent alors être interprétés facilement comme des taux de majoration ou de minoration par rapport à un individu de référence. Les autres fonctions de lien couramment utilisées sont présentées dans le tableau suivant :

Nom du lien	**Fonction de lien**
Identité	$g(\mu) = \mu$
Log	$g(\mu) = \ln(\mu)$
Cloglog	$g(\mu) = \ln(-\ln(1-\mu))$
Logit	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$
Probit	$g(\mu) = \phi(\mu), avec\ \phi\ la\ fonction\ inverse\ de\ la\ fonction\ de\ répartition\ d'une\ loi\ normale\ centrée\ réduite$
réciproque	$g(\mu) = -\frac{1}{\mu}$

Tableau 1. Fonctions de lien classiques

3.2.4. Lois usuelles de type exponentiel

Loi Normale

Utiliser une loi Normale dans un MLG correspond finalement à une régression linéaire telle que vue en début de chapitre, régression simple mais des hypothèses trop fortes pour la réalité. De plus, la loi Normale prenant aussi des valeurs négatives, elle n'est pas très adaptée dans la modélisation des fréquences et montants de sinistres.

Soit Y une variable aléatoire réelle suivant une loi normale d'espérance μ et de variance σ2. La densité de cette variable est donc :

$$f_Y(y) = \frac{1}{\sqrt{2\pi * \sigma^2}} * exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

Cette densité peut être mise sous la forme :

$$f_Y(y) = exp\left(\frac{y*\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{\frac{y^2}{\sigma^2} + \ln(2\pi*\sigma^2)}{2}\right)$$

La loi Normale appartient donc bien à la famille exponentielle avec :

- $\theta = \mu$
- $\varphi = \sigma^2$
- $a(\varphi) = \varphi$
- $b(\theta) = \frac{\theta^2}{2}$
- $c(y, \varphi) = -\frac{\frac{y^2}{\varphi} + \ln(2\pi * \varphi)}{2}$

Dans les exemples de lois suivants, nous présenterons uniquement la forme exponentielle de la densité sans expliciter les différentes fonctions et paramètres. Ceux - ci seront récapitulés dans un tableau à la fin de ce chapitre.

Loi de Poisson

La loi de Poisson est une loi de référence en assurance non vie. Elle dépend d'un unique paramètre, facilement interprétable et estimable. En effet, le paramètre λ correspond à la fréquence de survenance des sinistres et son estimation peut se faire aisément par moyenne empirique. De plus, sa propriété d'additivité permet de simplifier la modélisation. Néanmoins, le recours à d'autres lois s'impose car l'espérance et la variance d'une loi de Poisson sont égales, ce qui n'est pas en adéquation avec la réalité. En appliquant un MLG avec cette loi, il n'est pas rare d'observer un phénomène que l'on appelle surdispersion. Cette surdispersion est détectée lorsque les variances estimées à partir du modèle sont supérieures aux moyennes estimées. Une des raisons probables est la non homogénéité du portefeuille ou à un mauvais calibrage du modèle. Dans tous ces cas, la fonction de variance de la loi de Poisson est donc inadaptée et entraine une estimation biaisée des intervalles de confiances et même des coefficients eux - mêmes.

Soit Y une variable aléatoire à valeurs discrètes suivant une loi de Poisson de paramètre λ.
Sa fonction de densité est de la forme :

$$P(Y = y) = \exp(-\lambda) * \frac{\lambda^y}{y!}$$

Et peut être mise sous la forme exponentielle :

$$P(Y = y) = exp[y * \ln(\lambda) - \lambda - \ln(y!)]$$

Loi Gamma

Elle permet de modéliser des variables strictement positives, et de tenir compte de l'hétérogénéité des profils de risques. Les actuaires y ont donc particulièrement recours pour modéliser les montants de sinistres. Elle permet également l'étude de la fréquence de survenance des sinistres lorsqu'elle est utilisée pour estimer l'intensité du paramètre d'une loi de Poisson. La loi est alors une loi Poisson mélange appelée loi Poisson Gamma.

Soit Y une variable aléatoire réelle suivant une loi Gamma de moyenne μ et de variance ν-1. Sa fonction de densité est de la forme :

$$f_Y(y) = \frac{1}{\Gamma(\nu)} * \left(\frac{\nu}{\mu}\right)^{\nu} * y^{\nu-1} \exp\left(-\frac{\nu}{\mu} * y\right)$$

Avec $\Gamma(x) = \int_0^{\infty} e^{-u} * u^{x-1}\, du$

Et peut être mise sous la forme exponentielle :

$$f_Y(y) = exp\left[\nu * \ln\left(\frac{\nu}{\mu}\right) - \frac{\nu}{\mu} * y + (\nu - 1) * \ln(y) - \ln\left(\Gamma(\nu)\right)\right]$$

Loi Binomiale - Négative

Il s'agit d'une loi de Poisson mélange. Ce type de lois permet représenter la fréquence des sinistres en tenant bien compte de l'hétérogénéité du portefeuille. C'est-à-dire que le paramètre λ de la loi de poisson est désormais une variable aléatoire dépendant des caractéristiques de l'assuré. Lorsque la loi de ce paramètre est la loi Gamma, on retrouve alors la loi Binomiale Négative.
Soit Y une variable aléatoire suivant une loi binomiale négative de paramètres r
et p, avec r > 0 et p ϵ [0,1]. Sa fonction de densité est de la forme :

$$P(Y = y) = \binom{y + r - 1}{y} p^r * (1 - p)^y$$

Et peut être mise sous la forme exponentielle :

$$P(Y = y) = exp\left[y * \ln(1 - p) + r * \ln(p) + \ln\left(\frac{y^{r-1}}{\Gamma(r)}\right)\right]$$

Loi Log - Normale

La loi Log - Normale est celle qui est la plus utilisée en assurance non vie pour la modélisation des coûts des sinistres. C'est une loi dérivée de la loi normale car X suit une loi log normale si et seulement si ln(X) suit une loi normale. Cette loi sera donc utilisée dans le même cadre qu'une loi normale mais permet alors de n'avoir que des valeurs positives. De plus, elles interviennent lorsqu'il y a un effet multiplicatif d'un grand nombre de variables indépendantes entre elles. Nous comprenons donc pourquoi elle est très utilisée dans les MLG, les variables explicatives étant supposées indépendantes et ayant cet effet multiplicatif sur la variable réponse.

Soit Y une variable aléatoire réelle suivant une loi log - normale de paramètres μ et σ2.
Sa fonction de densité est :

$$f_Y(y) = \frac{1}{y * \sqrt{2\pi * \sigma^2}} * exp\left(\frac{-(\ln(y) - \mu)^2}{2\sigma^2}\right)$$

Cette variable ne suit pas une loi de type exponentiel et il n'est donc pas possible de lui appliquer un MLG. Cependant il est possible de se ramener à une loi de famille exponentielle en posant $X = ln(Y)$. La nouvelle variable aléatoire X ainsi créée suit alors une loi normale d'espérance μ et de variance σ^2.

Nous proposons sur la page suivante un tableau récapitulatif des lois et des paramètres adéquats.

Lois	$\mathbf{a(\varphi)}$	$\mathbf{b(\theta)}$	$\mathbf{c(y,\varphi)}$	$\mu = \mathbf{b'(\theta)}$	$\mathbf{V}(\mu) = \mathbf{b''(\theta)}$	$\mathbf{Var(Y)} = \mathbf{a(\varphi)} * \mathbf{V}(\mu)$	**Lien canonique** *Nom du lien*
Poisson P(μ)	1	$exp(\theta)$	$-\ln(y!)$	$exp(\theta)$	$exp(\theta)$	μ	$\ln(\mu)$ *log*
Normale $N(\mu,\sigma^2)$	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\varphi} + \ln(2\pi * \varphi)\right)$	θ	1	σ^2	μ *Identité*
Gamma G(μ,ν)	$\frac{1}{\nu}$	$-\ln(-\theta)$	$(\nu-1)*\ln(y) - ln[\Gamma(\nu)]$	$-\frac{1}{\theta}$	$\frac{1}{\theta^2} = \mu$	μ^2	$-\frac{1}{\mu}$ *inverse*
Binomiale Négative BN(r,p)	1	$-r * \ln(1 - e^{\theta})$	$\ln\left(\frac{y^{r-1}}{\Gamma(r)}\right)$	$r * \frac{e^{\theta}}{1-e^{\theta}}$	$r * \frac{e^{\theta}}{(1-e^{\theta})^2}$	$r * \frac{e^{\theta}}{(1-e^{\theta})^2}$	$ln\left(\frac{\alpha*\mu}{1+\alpha*\mu}\right)$ $avec\ \alpha = \frac{1-p}{\mu*p}$

Tableau 2. Fonctions de paramétrage associées aux distributions classiques

Nous avons présenté ici les notions fondamentales d'un MLG, et les différentes composantes qui interagissent dans ce modèle. Nous savons maintenant que l'objectif de ce modèle est d'estimer les coefficients associés aux variables explicatives. Mais comment se fait alors cette estimation ? C'est la question à laquelle nous allons répondre dans la partie suivante.

3.2.5. Log vraisemblance et équations

L'estimation des coefficients se fait par maximum de vraisemblance. La vraisemblance d'un échantillon est la probabilité d'observer cet échantillon. Un échantillon étant, pour un ensemble de n variables réponse Y_i, les n observations y_i. Il s'agit donc du produit des densités où la densité de y_i est dite la contribution de l'observation. Pour déterminer le maximum de vraisemblance, il faut trouver le paramètre qui annule sa dérivée première et qui rend négative sa dérivée seconde. De plus, la loi des variables réponses étant de type exponentiel, il est alors plus aisé de travailler sur le logarithme de la vraisemblance, de manière à se ramener au calcul d'une somme plus simple. Nous nommons cette fonction log vraisemblance. Enfin, maximiser la log vraisemblance équivaut à maximiser la vraisemblance car le logarithme est une fonction strictement croissante.

Les équations de log vraisemblance

Soit Y une variable à expliquer dont les observations pour n individus sont regroupés dans le vecteur (y_1, y_2, … y_n). Et (X_1, X_2, … X_p) les p variables explicatives retenues pour le MLG. Dans ce modèle, X_j peut représenter indifféremment l'ensemble des observations de la variable X_j ou la variable elle-même. $X_{i.}$ représente le vecteur ligne de la $i^{ème}$ observation de l'ensemble des variables explicatives. Enfin, X_{ij} correspond à la $i^{ème}$ observation de la variable explicative $X_{j.}$ Le vecteur $\beta = (\beta_1, \beta_2, \dots, \beta_p)$, vecteur des coefficients associés à chacune des p variables explicatives.

Les n variables réponses sont jugées indépendantes et nous pouvons alors déterminer la densité jointe de la manière suivante :

$$Vraisemlance \rightarrow L(y_1, y_2, \dots, y_n, \beta, \theta, \varphi) = \prod_{i=1}^{n} f(y_i, \theta, \varphi)$$

Et on obtient en passant au logarithme :

$$Log\ vraisemblance \rightarrow l(y_1, y_2, \dots, y_n, \beta, \theta, \varphi) = \sum_{i=1}^{n} \ln[f(y_i, \theta, \varphi)]$$

Et :

$$\ln[f(y_i,\beta,\theta,\varphi)] = \frac{y_i*\theta-b(\theta)}{\varphi} + c(y_i,\varphi) \text{ , avec } a(\varphi) = \frac{\varphi}{\omega} et\ \omega = 1$$

D'où :

$$l(y_1,y_2,\dots,y_n,\beta,\theta,\varphi) = \sum_{i=1}^{n} \frac{y_i*\theta-b(\theta)}{\varphi} + \sum_{i=1}^{n} c(y_i,\varphi)$$

L'objectif de la méthode par maximum de vraisemblance est de trouver les paramètres $\theta = \hat{\theta}$ et $\varphi = \hat{\varphi}$ qui maximisent cette fonction log vraisemblance. Le paramètre φ n'influence ni la prévision ni l'explication du modèle. Nous l'avons vu plus haut, le paramètre de dispersion n'a d'influence que sur la variance. Il sert en fait à ajuster la variance du modèle. Dans le modèle, il peut donc être supposé fixe et être estimé séparément ensuite. Nous nous apercevons que les coefficients β_j n'apparaissent pas explicitement dans l'expression ci-dessus. Pour pouvoir les estimer, il faudra donc avoir recours aux dérivées partielles. Ce procédé implique donc de définir trois notation μ_i,θ_i et η_i. La première correspond à l'espérance conditionnelle de Y par rapport à la $i^{ème}$ observation. La seconde correspond à la valeur du paramètre θ pour cette observation. Et enfin, η_i est le prédicteur linéaire de cette observation. Cette définition est tout à fait cohérente car nous l'avons vu plus haut, le prédicteur linéaire et θ dépendent totalement de μ. Nous posons aussi l_i la log vraisemblance pour la $i^{ème}$ observation.

$$dérivés\ partielles\ de\ l_i => \frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} * \frac{\partial \theta_i}{\partial \mu_i} * \frac{\partial \mu_i}{\partial \eta_i} * \frac{\partial \eta_i}{\partial \beta_j}$$

Calcul de chacun des dérivées partielles :

- $\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\varphi}$
- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{\frac{\partial b'(\theta_i)}{\partial \theta_i}} = b''(\theta_i)^{-1} = V(\mu_i)^{-1}$ en remarquant que $\mu_i = b'(\theta_i)$
- $\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial g(\mu_i)}{\partial \mu_i}} = \frac{1}{g'(\mu_i)}$ en remarquant que $\eta_i = g(\mu_i)$
- $\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \sum_{k=1}^{p} \beta_k X_{ik}}{\partial \beta_j} = x_{ij}$ en remarquant que $\eta_i = \sum_{k=1}^{p} \beta_k X_{ik}$

Nous obtenons alors la relation :

$$l_i = \left(\frac{y_i - b'(\theta_i)}{\varphi}\right) * \frac{1}{V(\mu_i)} * \frac{1}{g'(\mu_i)} * x_{ij}$$

Et enfin nous établissons l'expression finale de l sous la forme suivante :

$$l(y_1,y_2,\dots,y_n,\beta,\theta,\varphi) = \sum_{i=1}^{n} \frac{(y_i-\mu_i)*x_{ij}}{Var(y_i)*g'(\mu_i)} \quad Avec\ Var(y_i) = \varphi * V(\mu_i)$$

Le calcul de cette expression est encore simplifié lorsque la fonction de lien utilisée est la fonction de lien canonique car dans ce cas :

$$\begin{cases} g(\mu_i) = \theta_i \\ g(\mu_i) = \eta_i \end{cases} \Rightarrow \ \theta_i = \eta_i$$

Les dérivées partielles à calculer ne sont plus qu'au nombre de deux, $\frac{\partial l_i}{\partial \theta_i}$ et $\frac{\partial \eta_i}{\partial \beta_j}$.

L'expression obtenue est alors :

$$l\left(y_1, y_2, \dots, y_n, \beta, \theta, \varphi\right) = \sum_{i=1}^{n} \frac{\left(y_i - \mu_i\right) * x_{ij}}{\varphi}$$

La résolution de ces équations n'est pas pour autant triviale. μ_i est inconnu et dépendant de β. Elles seront résolues numériquement par le biais de méthodes itératives telles que Newton Raphson et l'algorithme du score de Fisher.

3.2.6. Adéquation et significativité du modèle

Tout comme pour les modèles de régression linéaire, des tests d'adéquation du modèle et de significativité des variables sont effectués de manière à proposer des tarifs fiables et précis. Dans un MLG, deux composantes permettent cette analyse. Il s'agit de la déviance et de la statistique de Pearson.

D'abord exprimons la log vraisemblance maximisée par le paramètre $\hat{\beta}$ en fonction de $\hat{\theta}_i$ puis en fonction de $\hat{\mu}_i$.

$$\varphi * l_{imax}(y_i) = y_i * \hat{\theta}_i - b\left(\hat{\theta}_i\right) + constante \quad (1)$$
$$\varphi * l_{imax}(y_i) = y_i * (b')^{-1}(\hat{\mu}_i) - b\left((b')^{-1}(\hat{\mu}_i)\right) + constante \quad (2)$$

Nous introduisons maintenant la notion de modèle parfait dit aussi modèle saturé. C'est-à-dire que nous avons autant de paramètres que d'observations distinctes. Dans un modèle saturé, nous avons donc l'information complète et la valeur de μ_i est ainsi égale à y_i.

L'expression de la log vraisemblance peut alors être mise sous la forme :

$$\varphi * l_{isaturé}(y_i) = y_i * (b')^{-1}(y_i) - b\left((b')^{-1}(y_i)\right) + constante \quad (3)$$

Nous pouvons alors définir l'expression de la déviance grâce aux expressions (2) et (3).

La déviance est l'écart entre la log vraisemblance résultant du modèle et celle obtenue pour le modèle saturé correspondant.

Elle s'écrit donc :

$$D = 2\varphi * \sum_{i=1}^{n} l_{isaturé} - l_{imax}$$

Le terme multiplicatif φ n'est pas toujours présent et dans ce cas nous avons plutôt la déviance standardisée $D^{std} = \frac{D}{\varphi}$

L'intérêt de la déviance est de distinguer si le modèle est adéquat ou non. Le modèle saturé correspondant au modèle parfait, plus la déviance sera importante plus notre modèle sera éloigné de la réalité et donc non adapté. Le but sera donc d'obtenir un modèle final avec la déviance la plus faible possible. Mais à partir de quel niveau pourrons-nous juger que notre modèle est adéquat ?

Il est possible d'établir un seuil de significativité à l'aide de la loi de Khi 2 afin de répondre à ce problème. En effet, sous l'hypothèse d'un modèle significatif, avec p variables explicatives et n observations, la déviance standardisée suit asymptotiquement une loi de à n-p degrés de liberté. Ceci nous permet alors d'établir un test de significativité du modèle de niveau α en comparant la déviance standardisée observée avec le quantile d'ordre 1- α de cette loi de χ_2.

3.2.7. Robustesse et précision du modèle

Lors d'une tarification, différents MLG peuvent être réalisés avec des paramétrages différents (utilisation d'autres variables explicatives, changement de la fonction de lien). Une fois les tests d'adéquation et de significativité du modèle réalisés il sera donc nécessaire de comparer deux modèles significatifs afin d'en retenir le meilleur. Car si nous disposons de deux modèles dont la log vraisemblance est jugée satisfaisante, comment savoir lequel des deux est réellement le mieux ? La valeur élevée de la log vraisemblance n'est pas suffisante pour faire ce choix. Le modèle saturé est considéré comme parfait car il dispose de toute l'information explicative possible. En fait, plus nous ajoutons de variables explicatives, plus la log vraisemblance augmente, jusqu'à atteindre la valeur de celle du modèle saturé. Or, cela n'est pas utile pour l'assureur dont le but est de restreindre son modèle aux variables explicatives les plus significatives en ne perdant qu'une faible partie de l'information. Ce choix est justifié car certes, plus un modèle a de variables explicatives, plus il est précis. Mais l'inconvénient est que plus il a de variables explicatives, moins il est robuste. Pour illustrer cette notion un exemple simple peut être énoncé : si un groupe de personnes est décrit entièrement (taille, vêtements, couleur des yeux…) il sera très facile de l'identifier. Or, si une personne parmi ce groupe se présente avec une coupe de cheveux différente, il ne nous sera alors pas possible de l'identifier et nous perdrions tout simplement cette personne. Certains critères moins nombreux tels que leur âge, leur couleur de peau, auraient suffi à retrouver toutes les personnes de ce groupe. Nous voyons donc dans cet exemple qu'un nombre élevé de caractéristiques ne s'adapte pas à un changement de situation et provoque au contraire une réponse incomplète. Nous pouvons rapprocher cette situation d'une segmentation en tarification. L'assureur désire que chacun de ses individus puisse toujours être regroupé dans une classe, même en cas de changement de l'une de ses caractéristiques. Il s'agit d'un compromis pour l'assureur entre le souhait que son modèle soit de taille raisonnable et le souhait qu'il ait un fort pouvoir explicatif.

L'une des réponses les plus reconnues apportée à cette problématique est la minimisation d'un critère appelé AIC (Akaïke Information Criterion). L'AIC est une mesure de la qualité d'un modèle. Il contient un facteur pénalisant en fonction du nombre de paramètres et permet ainsi de satisfaire le principe de parcimonie propre aux modèles prédictifs. Le modèle ayant la plus petite valeur pour l'AIC sera donc celui retenu.

L'expression du critère AIC est la suivante :

$$AIC = -2L + 2p$$

Lors d'une tarification par MLG, il sera donc intéressant de déterminer quelle est la combinaison de variables explicative la plus appropriée afin de satisfaire le compromis robustesse/précision.

3.2.8. Les tests de validation de modèle

Nous utiliserons les méthodes du type Forward et Backward de SAS pour sélectionner les combinaisons de variables. Nous allons donc expliquer le principe de ces trois méthodes. Avec la méthode Forward, les variables explicatives sont introduites une à une. Elle consiste en fait à rechercher la variable la plus significative au sens du R^2 (ou de la déviance) Nous commençons par un modèle à une variable, ensuite nous ajoutons à chaque étape la variable qui, associée à la première, explique le mieux la sinistralité. Cette variable est donc celle qui augmente (diminue) le plus fortement le R^2 (la déviance). Le processus est arrêté lorsque l'introduction d'autres variables n'augmente (ne diminue) plus de manière conséquente le R^2 (la déviance).

Avec la méthode Backward, nous partons cette fois ci du modèle complet, c'est-à-dire avec les p variables explicatives. Puis nous retirons une par une les variables, en sélectionnant à chaque étape celle diminuant (augmentant) le moins le R^2 (la déviance).
La méthode Stepwise est un combiné des deux méthodes précédentes. Cette méthode est en fait une méthode Forward à laquelle on laisse possible la sortie, à chaque étape, d'une variable devenue alors non significative.

Validité d'un modèle

Il est nécessaire de vérifier si la combinaison de variables retenues est pertinente. Afin d'y arriver, il est possible de vérifier la significativité de l'apport d'une variable grâce aux tests de type 1 et 3 disponibles avec la procédure GENMOD de SAS. Cette procédure affiche les écarts de déviance observés entre les combinaisons de variables. Dans un test de type 1, les lignes correspondent à l'ordre d'entrée des variables explicatives. Dans un test de type 1, il est difficile de cerner réellement l'apport d'une variable par rapport à une autre car l'ordre a une influence non négligeable. Il est alors plus intéressant de procéder à un test de type 3 qui ne tient pas compte de l'ordre. Chaque ligne correspond alors à la différence de déviance entre le modèle avec la variable associée à la ligne et sans cette variable. Nous pouvons avec ce test d'une part voir l'apport réel qu'à une variable indépendamment de celles déjà présentes dans le modèle. Et d'autre part ce test nous permet de comparer ces écarts de déviance et de déterminer quelles variables ont le plus de pouvoir explicatif.

Variabilité

Bien que moins utilisée, une mesure de la qualité d'un modèle peut aussi être déterminée en fonction de la variabilité de la valeur estimée des variables explicatives. Par variabilité nous entendons l'amplitude entre la valeur la plus faible et la plus élevée de chaque variable. Cela s'interprète dans le même état d'esprit que l'exemple énoncé dans le sous - chapitre « Robustesse et précision du modèle ». L'intérêt pour l'assureur est de balayer un large périmètre avec les variables explicatives choisies. À moins que notre portefeuille soit focalisé sur une population très ciblée, si la variabilité observée est faible, nous pouvons craindre que certains individus ne soient pas bien pris en compte, et que notre variable n'ait pas un pouvoir significatif suffisant. Ainsi, il est tout à fait possible qu'une variable soit jugée significative par les tests de type 1 et 3, alors qu'elle est jugée ensuite non pertinente par analyse de sa variabilité.

Cohérence du modèle

Il est nécessaire de vérifier la cohérence d'un modèle afin de le valider. Pour ce faire, il faut procéder à l'analyse des résidus. Dans un MLG, il n'y a pas de décomposition qui permette d'expliciter les résidus comme nous l'avons vu pour la régression linéaire. Les résidus sont donc dans ce cas plus difficiles à analyser. Les résidus les plus utilisés sont les résidus de Pearson et les résidus de Déviance.

$$Résidus\ de\ Pearson \qquad r_{Pi} = \frac{\widehat{\varepsilon_i}}{\sqrt{Var(Y_i)}}$$

Avec ε_i l'erreur existante entre le modèle et la valeur réelle observée : $\widehat{\varepsilon_i} = y_i - \hat{\mu}_i$

$$Résidus\ de\ Déviance \qquad r_{Di} = signe(\widehat{\varepsilon_i}) * \sqrt{D_i}$$

Avec D_i correspondant à la contribution de l'observation i à la déviance D décrite plus haut.
Le modèle est alors jugé valide si le nuage de point est de forme cylindrique autour de l'axe des abscisses. L'interprétation est donc particulièrement subjective et l'intérêt de cette analyse reste restreint.

Validation croisée

Si le volume de données disponibles le permet, il peut être très intéressant de procéder à une validation croisée. Le principe de cette méthode est de calibrer le modèle sur une base dite base d'apprentissage puis de confronter ce modèle à une base dite base test.

3.2.9. Conclusion

Les modèles linéaires généralisés permettent d'étendre la méthode de régression linéaire à un ensemble de lois plus large et correspondant plus à la sinistralité réelle couverte par l'assureur. De plus, contrairement aux régressions linéaires il est désormais possible de traiter des variables catégorielles. Les MLG proposent les mêmes avantages que l'approche empirique décrite au début du chapitre concernant la méthode Fréquence - Coût, mais ne souffrent pas des inconvénients propres à cette dernière.
Ils permettent aussi non seulement d'estimer la valeur des paramètres mais fournissent aussi un intervalle de confiance pour ces estimateurs. Nous n'avons pas détaillé cette partie mais le lecteur pourra se référer au livre « An introduction to Generalized Linear Models » d'Annette J. DOBSON et Adrian G.BARNETT. Ces intervalles ont une importance non négligeable car ils permettent à l'assureur d'estimer sa marge de manœuvre lors d'une tarification sur mesure par exemple. Néanmoins les MLG nécessitent un développement informatique avec notamment une détection à priori des interactions, et ne proposent un résultat fiable que sous condition d'un volume de données suffisant. Si ce n'est pas le cas, il sera difficile d'ajuster un modèle fiable. De plus, l'inversion exponentielle de la fonction de lien peut provoquer des écarts d'ajustement des paramètres considérables.

Enfin, une extension aux MLG se développe fortement. Il s'agit des Modèles Additifs Généralisés (GAM). La principale différence concerne le prédicteur. Celui - ci n'est plus obligé d'être linéaire comme dans le cadre des MLG. S'ajoute à cela qu'il est composé d'une somme de fonctions non paramétriques. Un GAM sera donc sous la forme suivante :

$$g[E(Y)] = \alpha + \sum_{j=1}^{p} f_j(X_j)$$

Nous ne traiterons pas plus ces modèles mais l'intérêt principal de leur utilisation lors d'une tarification repose surtout sur leur possibilité de constituer des classes d'âge ou autres classe homogènes.
Pour plus d'information, le lecteur pourra se référer au livre suivant FARAWAY.J [2006].

4. Apprentissages statistiques

Le volume des données toujours grandissant couplé à un fort progrès des outils informatiques a conduit ces dernières années les assureurs à se tourner vers de nouvelles méthodes de modélisation : les algorithmes par apprentissages statistiques. Comme leur nom l'indique, ces méthodes combinent études statistiques et apprentissage machine. Contrairement aux modèles classiques qui définissent un modèle global, ces algorithmes ne retiennent qu'une hypothèse : les observations de la variable réponse peuvent être prédites par un processus unique. C'est ce processus qui sera alors modélisé par un algorithme.
Leur but est de rechercher l'information pertinente afin d'aider à la prévision et à la décision. L'intérêt pour l'assureur est alors tourné vers la qualité de la prévision au détriment du caractère interprétable et explicatif du modèle. Le modèle est dit « boite noire ». Ces méthodes sont encore en plein développement et non utilisées dans le domaine de la Santé. Pourtant leur utilisation serait judicieuse dans la mesure où le nombre de variables exogènes au risque Santé est particulièrement important et la dépendance entre les différentes garanties - encore négligées par les méthodes classiques - pourrait enfin être prise en compte. Nous traiterons dans ce mémoire de l'algorithme CART, Classification and Regression Tree. Pour une recherche plus approfondie sur les différentes méthodes par apprentissages statistiques, le lecteur pourra se référer au livre HASTIE.T & al. [2009].

4.1. Algorithme CART

L'algorithme CART est l'un des algorithmes par apprentissages statistiques le plus avancé. C'est une technique non paramétrique qui permet notamment d'identifier les interactions entre les variables explicatives, de sélectionner les plus pertinentes et d'établir la prévision de la variable réponse. Sa popularité est due à la simplicité de sa lecture puisque le résultat sorti est une représentation graphique hiérarchisée et intuitive, affichant de manière claire les variables discriminantes sous forme d'arbre. De plus, cet algorithme permet de traiter un large nombre de variables, qu'elles soient qualitative ou quantitatives et ce sans être affecté par des effets de colinéarité et d'hétérogénéité du portefeuille.

4.2. Principes de l'algorithme

Le principe de cet algorithme est d'utiliser les variables explicatives pour subdiviser le portefeuille en « régions » homogènes. Le CART débute d'abord avec la population entière hétérogène. Il définit alors la variable la plus discriminante qui sépare le mieux la population en deux sous - groupes, appelés nœuds. On nomme ce procédé une division binaire et ne porte que sur une condition du type :

$$\begin{cases} x_1 \in A \; si \; x \; est \; une \; variable \; qualitative \\ x_1 \leq \alpha \; si \; x \; est \; une \; variable \; quantitative \\ Avec \; la \; variable \; explicative \; x = (x_1, \dots x_m) \end{cases}$$

Si ces deux nœuds comportent chacun au moins deux observations, ils sont alors de nouveau divisés en deux nœuds. A cette étape, nous nous retrouvons donc avec six nœuds et quatre « chemins » différents. Ce processus de division binaire est alors réitéré pour chaque nœud et s'arrête lorsque le volume n'est plus suffisant ou lorsque l'on ne constate plus d'amélioration. Ainsi l'enjeu de cette méthode est double. Il faudra en effet définir comment choisir les variables séparatrices et leurs seuils respectifs, et spécifier quand le procédé de séparation devra s'arrêter. Les nœuds finaux sont appelés les nœuds terminaux ou plus communément feuilles. Chaque valeur finale obtenue pour la variable réponse correspond à une classe. Et l'arbre est alors appelé soit un arbre de classification pour une variable réponse qualitative, soit un arbre de régression pour une variable quantitative.

4.3. Notations et relations

Ci-dessous nous énonçons la définition de chaque notation et explicitons les différentes relations à connaître pour la suite du chapitre.

- R, la racine. Soit le nœud primaire contenant la *population* totale.
- n, taille de l'échantillon
- K le nombre de classes
- N_k le nombre d'observations de la classe k
- N(t) le nombre total d'observations dans le nœud t
- $N_k(t)$ le nombre d'observations de la classe k en nœud t
- A l'arbre obtenu par l'algorithme CART
- A_t l'arbre issu du nœud t. C'est-à-dire l'ensemble des nœuds descendants directement du nœud t
- A_f l'ensemble des feuilles obtenues

La probabilité à priori de la classe k est notée π_k et s'exprime de la façon suivante : $\pi_k = \frac{N_k}{N}$

La probabilité à postériori de la classe k en nœud t est la proportion d'observations à partir du nœud t qui sont affectées à la classe k par l'arbre. C'est-à-dire la probabilité d'appartenir à la classe k sachant que l'on est issu du nœud t. Elle est notée p (k | t) et est exprimée de la manière suivante : $p(k|t) = \frac{N_k(t)}{N(t)}$

On a donc pour le nœud t, $p(t) = [p(1|t), \dots, p(K|t)]$

4.4. Critères de l'algorithme

Comme nous l'avons expliqué plus haut, l'objectif est de créer une partition de E en K classes. Avec E l'ensemble des observations des variables explicatives. Il nous reste cependant à définir certains critères indispensables à la construction de l'arbre.

Critère d'hétérogénéité

Nous allons dans un premier temps définir ce qu'est une fonction d'hétérogénéité car elle sera l'indicateur de la qualité de la division des nœuds.
Soit h une fonction définie sur un ensemble fini de probabilités discrètes et à valeurs réelles :

$$h : (p1, \ldots, pK) \rightarrow h\,(p1, \ldots, pK)$$

h est une fonction d'hétérogénéité si elle vérifie les propriétés suivantes :

- Elle est symétrique en $\mathrm{p}_1, \ldots, \mathrm{p}_\mathrm{K}$
- Son maximum est atteint par l'équiprobabilité

$$\arg\max(\mathrm{p}_1, \ldots, \mathrm{p}_\mathrm{K}) = (1/\mathrm{K}, \ldots . 1/\mathrm{K})$$

- Son minimum est atteint par les éléments de la base canonique

$$\arg\min(\mathrm{p}_1, \ldots, \mathrm{p}_\mathrm{K}) = (\mathrm{e}_1, \ldots . \mathrm{e}_\mathrm{K})$$

L'hétérogénéité ou impureté d'un nœud peut alors être défini tel quel :

$$\mathrm{Imp(t)} = \mathrm{h[p(t)]}$$

D'après la définition d'une fonction d'hétérogénéité, l'impureté d'un nœud est maximale lorsque les classes sont mélangées à part égales. En fait, moins le nœud comporte de classes différentes, plus il sera jugé pur.
De l'expression précédente on en déduit l'impureté d'un arbre A :

$$Imp(A) = \sum_{t \in A_f} p(t) \times Imp(t)$$

Le critère d'hétérogénéité propre à l'algorithme CART est l'indice d'inégalité de Gini défini de la façon suivante :

$$Imp(t) = 1 - \sum_{i=1}^{K} p(i|t)^2$$

Nous avons énoncé le critère de pureté d'un nœud mais il nous reste à définir de quelle manière l'algorithme utilise ce critère pour diviser un nœud en deux sous parties.
Nous avons bien compris que le but est, après division du nœud, d'obtenir deux sous - groupes plus pur que leur groupe parent. La division, appelée aussi la coupe, qui sera retenue est donc celle qui maximisera le décroissement d'hétérogénéité.
La fonction de diminution d'hétérogénéité étant :

$$\Delta Imp(t) = Imp(t) - \pi_1 Imp(t_1) - \pi_2 Imp(t_2)$$
$$Avec\ t1\ et\ t2\ les\ deux\ nœuds\ issus\ du\ nœud,$$
$$et\ \pi_j la\ proportion\ des\ observations\ dans\ les\ noeuds\ t\ et\ j$$

À chaque nœud, l'algorithme considère chaque variable explicative une par une et choisit alors le couple (variable explicative, seuil) qui présente la meilleure coupe au sens du critère choisi. En effet, en maximisant la diminution d'impureté à chaque nœud, l'algorithme maximise aussi la diminution d'impureté entre l'arbre initial et l'arbre crée à la suite de la division.

Critères d'arrêt

La construction de l'arbre est effectuée par coupes successives. Cependant, ce processus pourrait être effectué indéfiniment dans la limite des observations totales. Il est donc nécessaire d'établir un critère d'arrêt. Il existe différents critères dont nous faisons la liste ci - dessous :

- *La profondeur de l'arbre dépasse une limite fixée. C'est-à-dire le nombre de variables explicatives est supérieur au nombre maximum désiré.*
- *Le nombre de feuilles dépasse une limite fixée.*
- *Le nombre d'observations contenu dans chaque nœud est en dessous du seuil fixé.*
- *La qualité de l'arbre est jugée suffisante.*
- *La décroissance de l'impureté par division binaire n'est plus significative.*

On note donc qu'en fonction de ces critères le nombre de feuilles sera plus ou moins grand. Par exemple, si nous fixons un nombre d'observations minimum dans chaque nœud relativement faible, alors les divisions pourront continuer plus longtemps, aboutissant sur un nombre important de feuilles.

4.5. Erreur de classification

Soit A_{max} l'arbre finalement obtenu après application du critère d'arrêt. Chaque feuille peut contenir plusieurs classes. Aussi, pour chaque feuille t, les individus sont répartis selon la règle d'affectation suivante :

$$C(t) = \arg\max\{p(k|t), k = 1, \dots, K\}$$

Cette règle d'affectation est la règle de Bayes.
Puisque $p(k|t)$ est la proportion d'individus dans la feuille t appartenant à la classe k, cette règle se traduit par le fait qu'à chaque feuille est attribuée la classe dans laquelle il y a le plus d'individus. De par cette définition découle donc la notion d'erreur de classification car certains individus d'une même feuille devraient en fait appartenir aux classes non retenues et sont ainsi mal regroupés.
Cette erreur est exprimée ainsi :

$$R(t) = \sum_{k \neq C(t)}^{K} p(k|t)$$

Et l'erreur de prédiction sur l'arbre final est donc :

$$R(A_{max}) = \sum_{t \in A_f}^{K} p(t) \times R(t)$$

Exemple :

La classe attribuée à la feuille 1 sera donc « Individus non sportifs » car la proportion de cette classe est de :

$$p(Individus\ non\ sportifs) = \frac{40}{60} = \frac{2}{3}$$

Tandis que la proportion des individus sportifs est égale à :

$$p(Individus\ sportifs) = \frac{20}{60} = \frac{1}{3}$$

Pour la feuille 2, la classe attribuée, selon le même procédé sera la classe « Individus sportifs ».
L'erreur de classification sera ainsi égale à $\frac{1}{3}$ pour la feuille 1. Elle sera égale à $\frac{1}{4}$ pour la feuille 2.
Et l'estimation de l'erreur de prédiction de l'arbre - composé de ces deux feuilles - aura pour valeur :

$$R(feuille1 + feuille2) = \frac{60}{100} \times \frac{1}{3} + \frac{40}{100} \times \frac{1}{4} = \frac{3}{10}$$

4.6. Élagage

Selon le même raisonnement qu'établi plus haut dans le cadre des MLG, l'assureur désire limiter le nombre de feuilles obtenues par l'algorithme. En effet, nous pourrions construire un arbre composés uniquement de feuilles pures, c'est-à-dire de feuilles ne contenant qu'une seule classe. Mais la robustesse du modèle serait alors moindre. Le principe d'élagage est une solution efficace pour résoudre ce problème. Avant d'expliquer plus en détail ce que représente l'élagage d'un arbre, nous devons définir ce que nous appelons le paramètre de complexité noté α.
Ce paramètre permet de pénaliser, tout comme le critère AIC vu auparavant, un nombre de feuilles trop important. Car pour deux arbres A_1 et A_2, si $A_1 < A_2$ alors $R(A_1) > R(A_2)$. Ceci est tout à fait normal car plus il y a de feuilles, plus il y a de classes d'individus représentés, et moins l'erreur est importante.
Soit :

$$card\ (At) = le\ nombre\ de\ feuilles\ de\ l'arbre\ issus\ du\ nœud\ t$$

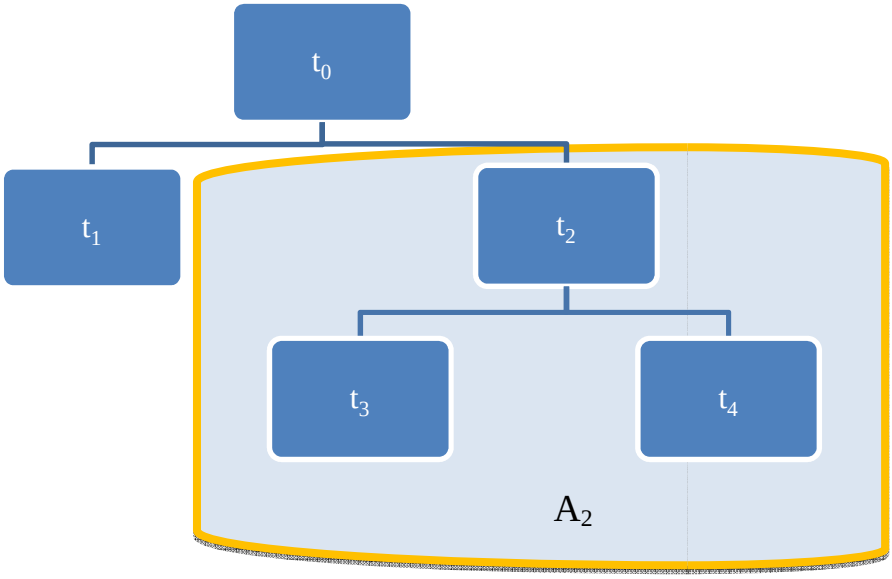
Le paramètre de complexité s'exprime par la relation suivante :

$$R_\alpha(A_t) = R(A_t) + \alpha \times card(A_t)$$

On en déduit que pour un nœud t, l'expression devient :

$$R_\alpha(t) = R(t) + \alpha$$

Prenons l'exemple de cette coupe :



L'arbre A_2 est composé de deux feuilles et nous avons donc :

$$R_\alpha(A_2) = R(A_2) + 2\alpha$$

Pour le nœud t_2, nous avons :

$$R_\alpha(t_2) = R(t_2) + \alpha$$

Élaguer ce sous arbre A_2 consisterait à supprimer l'arbre issus du nœud t_2. C'est-à-dire que seul le nœud t_2 serait retenu, les nœuds t_3 et t_4 étant enlevés du résultat final. Ainsi, puisque le but est de minimiser l'erreur de prédiction de l'arbre, le choix d'élaguer ou non se fera après comparaison de $R_\alpha(A_2)$ et $R_\alpha(t_2)$:

$$Si\ R_\alpha(t_2) < R_\alpha(A_2)\ alors\ on\ élague$$

La relation déduite de cette inégalité est la suivante :

$$El(A_2, t_2) = \frac{R_\alpha(A_2) - R_\alpha(t_2)}{card(A_2) - 1} \leq \alpha$$

Cette fonction $El(A_t, t)$ peut ainsi être calculée pour chaque nœud de l'arbre et permet de procéder à ce que nous nommons l'algorithme de coupe du maillon faible :

1. *Création de l'arbre maximum T : toutes les feuilles sont pures.*

2. *Calcul de la fonction d'élagage décrite plus haut pour chaque nœud interne. On choisit t_1 tel que :*
$$t_1 = \arg\min\{El(t, T), t \in T\} \text{ et } \alpha_1 = g(t_1, T).$$

3. *Obtention de l'arbre T_2 après suppression du sous arbre issu de t_1.Réitérer le procédé pour T_2, en fixant un coefficient de pénalisation $\alpha_2 > \alpha_1$. Réitérer ce procédé jusqu'à ce que l'élagage ne réduise plus l'erreur de prédiction.*

4. *Obtention d'une suite décroissante d'arbres emboités et une suite croissante d'α_i associées.*

4.7. Sélection finale

Le choix de l'arbre optimal se fait par validation sur un échantillon test. L'arbre élagué retenu est celui pour lequel le taux d'erreur estimé sur l'échantillon test est le plus bas. Cela se traduit par l'expression suivante :

$$R^{El}(A_{opt}) = min\{R^{El}(A_i), i = 1 \dots L\},$$
$$avec\ L\ le\ nombre\ de\ sous\ arbres\ emboités$$

Où :

$$\begin{cases} R^{El}(j) = \sum_i 1_{\{i \neq j\}} \frac{M_{i,j}}{M_j} \\ R^{El}(A_l) = \sum_{i,j} 1_{\{i \neq j\}} \frac{M_{i,j}}{M} \end{cases}$$

M est le nombre d'observations total de l'échantillon test. M_j est le nombre d'observations de cet échantillon appartenant à la classe j. Et enfin, $M_{i,j}$ est le nombre d'observations de l'échantillon regroupées dans la classe i par l'arbre alors qu'elles devraient appartenir à la classe j. De ces définitions, on en déduit que $R^{El}(j)$ décrit la proportion d'invidus de la classe j mal classés. Et que $R^{El}(A_l)$ décrit la proportion des observations de l'échantillon test mal classées.

4.8. Conclusion

L'algorithme CART bénéficie de plusieurs avantages non négligeables qui justifient son développement dans le secteur actuariel et plus précisément dans le domaine de la santé. En effet, cet algorithme permet d'une part de prendre en compte un nombre considérable de variables explicatives, qu'elles soient continues, discrètes ou catégorielles. Et d'autre part, il impose très peu d'hypothèses au préalable. Cet aspect non paramétrique est un avantage significatif par rapport aux MLG pour lesquels il faut définir une structure globale et même définir les interactions entre variables explicatives. De plus, ce modèle est très peu sensible aux valeurs extrêmes et aux valeurs manquantes. Nous en expliquerons la raison par la suite lorsque nous décrirons les fonctionnalités de l'algorithme sur R. Enfin, le modèle propose une sortie très simple à interpréter grâce à sa représentation graphique par arbre et son code est facilement programmable pour un portefeuille différent puisque la structure globale n'est pas fixée à priori.

Néanmoins, l'algorithme CART présente deux inconvénients principaux. Ce type de modèle est très sensible aux données, notamment dans le cas de surapprentissage. D'où la nécessité d'un élagage bien effectué. De plus, cette méthode est sujette à un « effet papillon » car si nous changeons une seule variable, tout l'arbre est alors modifié.

IV. Traitement et Analyse de données

A. Traitement des données

Le traitement de données est essentiel lors d'une tarification car les modèles utilisés dépendent totalement des données du portefeuille. Ainsi une base d'étude présentant des erreurs entrainera nécessairement une modélisation fausse de la variable étudiée. Afin de vérifier et d'enlever toute ligne semblant erronée, l'actuaire doit donc au préalable réfléchir sur toutes les erreurs possibles. De plus, un travail de réflexion supplémentaire doit être fait sur les informations traitées. Par exemple, définir et repérer les valeurs extrêmes. Pour ces raisons, le traitement des données a été la partie la plus longue de notre travail.

Nous disposions initialement sur SAS de quatre principales bases de données :

- Un fichier « Contrats » présentant tous les contrats, santé et prévoyance, d'AG2R La Mondiale.
- Un fichier « Prestations » détaillant l'historique de consommation de tous les assurés. Chaque ligne correspondant à une demande de remboursement pour un acte donné à une date précise.
- Un fichier « Bénéficiaires » regroupant les informations pour chaque bénéficiaire couvert par AG2R.

Nous avons eu besoin d'utiliser d'autres bases de données afin de récupérer toutes les informations nécessaires à notre étude mais celles-ci ne sont pas présentées ici.

Les variables dans le fichier « Contrats » utilisées pour l'étude sont les suivantes :

- Le numéro de contrat.
- La catégorie de personnel.
- Le produit souscrit.
- La variable indiquant s'il s'agit d'une surcomplémentaire ou d'une base.
- Le nom de la CCN à laquelle est rattaché le contrat.
- La date d'effet du contrat.

De ce fichier, nous ne gardons que les contrats appartenant aux CCN « Boulangerie », « Patisserie » et « Association des Fédérations en Fruits et Légumes, Epicerie, Crémerie (AFFLEC) ».

Les variables indispensables que nous trouvons dans le fichier « Prestations » sont celles présentées ci - dessous :

- L'identifiant du bénéficiaire.
- La catégorie de personnel du bénéficiaire.
- Le numéro de contrat sous lequel il est rattaché.
- Le produit souscrit.
- Le code correspondant à l'établissement dans lequel il est employé.
- La garantie consommée.
- Le nombre d'actes consommés pour cette garantie (presque toujours égal à 1 sauf dans le cas de certaines garanties telles que la pharmacie ou la radiologie).
- La date de soin.
- Le montant de frais réels en centimes d'euros.
- Le montant de remboursement versé par AG2R La Mondiale en centimes d'euros.

Le traitement effectué sur cette base a consisté à ne récupérer uniquement que les actes de consommation pour neuf garanties ; garanties qui seront présentées dans la partie B de ce chapitre. Nous avons aussi limité le périmètre d'étude aux actes consommés entre le 1er Janvier 2011 et le 31 Décembre 2013. De plus, nous avons retiré de l'étude les bénéficiaires n'ayant pas souscrit aux produits présents dans la base « Contrats ». Enfin, les montants ont été remis en euros et multipliés chaque année par le taux d'inflation associé. La prise en compte du taux d'inflation était nécessaire afin que chaque année ait le même poids car 1 euro en 2011 n'avait pas la même valeur qu'1 euro en 2013. Nous obtenons un fichier d'une taille de 15,6 millions de lignes, correspondant chacune à un acte de consommation, auquel on supprime toutes les lignes où le montant de frais réels est nul ou inférieurs au montant de remboursement. De même, nous avons supprimé toutes les lignes où le montant de remboursement d'AG2R La Mondiale était négatif car cela correspondait à une régularisation.

Dans la base de données concernant les bénéficiaires nous travaillons sur les variables suivantes :

- L'identifiant du bénéficiaire.
- Le type de bénéficiaire dont les valeurs possible sont adhérent, conjoint ou enfant.
- Le sexe du bénéficiaire.
- La date de naissance du bénéficiaire.
- Le nom du bénéficiaire.
- Le prénom du bénéficiaire.
- Le numéro de sécurité sociale du bénéficiaire.
- La situation familiale du bénéficiaire.
- La date d'adhésion du bénéficiaire. Cette date correspond à l'adhésion à un produit et non au contrat ; Cela se traduit par le fait qu'un même individu peut avoir deux dates d'adhésion différentes s'il a changé de produit une fois.
- La date de fin d'adhésion du bénéficiaire.
- La variable durée indiquant pour chaque année combien de temps le bénéficiaire a été couvert par le produit associé.

Dans un premier temps, nous avons créé une nouvelle variable en utilisant le numéro de sécurité sociale, le prénom et le nom de chaque bénéficiaire. Un assuré peut avoir plusieurs identifiants de bénéficiaire s'il a souscrit à plusieurs produits, notamment dans le cas d'une souscription à la base et à une option surcomplémentaire. Afin d'identifier cet individu comme une seule et unique personne, il était donc nécessaire d'établir la correspondance entre son identifiant et son numéro de sécurité sociale. Ensuite, nous avons récupéré uniquement les individus dont le contrat avait été soit rompu après 2011 soit pas

encore rompu. Pour les individus ayant souscrit avant 2011, nous avons fixé la date d'adhésion au 1er Janvier 2011, et pour ceux ayant des contrats encore ouvert après 2013 nous avons fixé leur date de sortie au 31 décembre 2013. Pour les bénéficiaires ayant changé de produit au cours de leur adhésion au contrat, nous avons décidé de ne garder qu'un produit par année pour chaque bénéficiaire en gardant celui pour lequel la durée était la plus longue. Monsieur X ayant souscrit à une option 1 durant les quatre premiers mois de l'année 2011, et à une option 2 durant les huit mois restants serait donc associé à l'option 2. Enfin nous avons calculé l'âge de l'individu pour chaque année de l'étude nous permettant de savoir son âge pour chaque ligne de consommation. Nous avons supprimé les individus présentant des âges et des temps de couverture inférieurs à zéro.

a) Découpage par niveaux

Pour chaque produit, nous avons déterminé quel était le niveau de couverture associés à chacun des quatre postes de consommation. Travail qui nous a ainsi permis de définir le niveau de couverture pour chaque ligne de consommation de notre fichier final. Ce découpage a été réalisé grâce à l'étude du Bureau d'Information et de Prévisions Économiques (BIPE). Cette étude après comparaison des offres du marché a pu définir trois niveaux - Bas, Moyen et Haut - pour chaque produit en fonction du remboursement proposé. Les postes étudiés sont les postes « Hospitalisation », « Actes courants (ou actes médicaux) », « Dentaire » et « Optique ». Les niveaux des produits sont donc déterminés par poste et non pas par garantie. Néanmoins, chaque poste a été étudié en ne travaillant que sur la garantie la plus discriminante. Pour illustrer cette méthode par exemple nous pouvons traiter le cas du poste « Hospitalisation ». Sur ce poste, la plupart des mutuelles proposant des remboursements similaires sur chaque acte, excepté en « Chambre particulière », c'est cette dernière qui a été retenue dans l'étude BIPE.
Les garanties ayant servi de référence pour ce découpage sont les suivantes :

- Chambre particulière pour le poste « Hospitalisation »
- Consultations généralistes pour le poste « Actes médicaux »
- Prothèses dentaires remboursées par la SS pour le poste « Dentaire »
- Forfait optique (verre + monture) pour le poste « Optique »

Pour l'entrée de gamme, soit un produit de niveau bas, les remboursements proposés sont présentés ci-dessous :

Poste	Entrée de gamme
1. Consultation	100 % BR
2. Chambre particulière en chirurgie	0 € / jour
3. Prothèse dentaire*	32 € (TM)
4. Optique*	0 €

Figure 11. Découpage pour l'entrée de gamme

Pour le moyen de gamme, les remboursements sont ceux affichés dans le tableau ci-dessous :

Poste	Niveau de garantie
1. Consultation	130 % BR
2. Chambre particulière en chirurgie	50€ / jour
3. Prothèse dentaire	300€
4. Optique*	300€

Figure 12. Découpage pour le milieu de gamme

Enfin, le tableau suivant indique les remboursements correspondant à un niveau élevé :

Poste	Haut de gamme
1. Consultation	300 % BR
2. Chambre particulière en chirurgie	80 € / jour
3. Prothèse dentaire	430 €
4. Optique*	600 €

Figure 13. Découpage pour le haut de gamme

Dans notre découpage, les bornes constituent l'entrée dans le niveau supérieur. Si un produit propose un remboursement de 430 euros en prothèses dentaires, il est alors considéré comme un produit haut de gamme pour le poste « Dentaire ».
Pour l'instant, nous avons découpé par niveaux de garantie uniquement les contrats de base. Or, les surcomplémentaires doivent aussi être découpées par niveaux. Nous ne pouvons additionner les montants de remboursement de la base avec ceux de la surcomplémentaire. Si nous procédons ainsi, nous nous ramenons alors un niveau de couverture global. Le tarif obtenu sera un tarif pour le montage total et ne permettra pas de différencier le tarif réellement dû à la base et celui dû au niveau de la surcomplémentaire. Nous avons donc décidé de définir les niveaux de surcomplémentaire en se basant toujours sur le découpage du BIPE, mais en divisant chaque borne de passage par deux. Pour illustrer notre méthode, prenons l'exemple de la garantie « Chambre particulière ». Une base est haut de gamme si elle propose un niveau de remboursement supérieur ou égal à 80 euros. Une surcomplémentaire sera quant à elle dans le haut de gamme si elle propose un remboursement supérieur ou égal à 40 euros.

b) Récupération des informations et ajout de variables

Les trois bases de données présentées précédemment ne permettent pas d'avoir toutes les informations nécessaires afin de bien segmenter la clientèle.

Nous avons récupéré les variables suivantes pour chaque bénéficiaire :

- Le régime. Les individus ayant deux régimes différents ont été supprimés

- Le département et la région. Ces informations ont pu être retrouvées en fonction de l'établissement où était salarié l'individu. Les individus étant sur deux régions différentes ont été supprimés de l'étude. En effet la région étant un facteur discriminant récurrent, il est important d'avoir une information fiable à ce niveau.

- Le statut professionnel.

- Le type d'adhésion. Tous les bénéficiaires étaient en obligatoire sur la base.

- Le statut du bénéficiaire, prenant les valeurs « Adulte » ou « Enfant ».

- Le zonier établi par Analyse Factorielle de Données Mixtes.

- L'effectif - par tranches - de l'entreprise dans laquelle le bénéficiaire est salarié.

- La classe d'âge. Nous avons conservé les mêmes classes d'âge établies cette année par notre équipe pour la tarification de l'offre ANI.

Les modalités de chacune des variables sont observables en annexe-I.
Le dernier retraitement de données a consisté à supprimer les valeurs extrêmes pour les montants de frais réels et les remboursements versés par AG2R La Mondiale. En effet, les valeurs extrêmes peuvent correspondre à des erreurs d'enregistrement. Et lorsque ce n'est pas le cas, elles correspondent alors à des actes dont le nombre d'occurrence est trop faible pour être intégrés dans nos modèles. En effet, un MLG et un algorithme CART nécessite un volume de données non négligeable pour que la modélisation soit proche de la réalité. Ces valeurs extrêmes sont souvent déterminées par l'actuaire et son expérience du secteur. Dans notre cas, nous avons travaillé en fonction du premier et troisième quartile représentant respectivement 25% et 75% des observations. Nous avons ainsi conservé les observations dont les montants étaient dans l'intervalle suivant :

$$[Q_1 - 1{,}5 \times (Q_3 - Q_1), Q_3 + 1{,}5 \times (Q_3 - Q_1)]$$

B. Analyse exploratoire de la consommation

Cette partie présente tout le travail effectué afin d'apprécier la consommation en santé de notre portefeuille et d'orienter nos choix dans la modélisation des frais réels et de la fréquence.

a) Statistiques descriptives

Une étude statistique est toujours nécessaire car elle permet de définir le périmètre d'étude et les critères qui devront déjà être écartés de l'étude en fonction par exemple de l'observation d'un volume de données insuffisant.

L'objectif principal de notre étude étant d'estimer la surconsommation des assurés ayant une surcomplémentaire, nous avons dû cibler les contrats disposant d'un montage base – surcomplémentaire. Dans notre portefeuille, la majorité des contrats sous ce type de montage étaient présents dans les contrats CCN. Nous disposions au départ de données sur 14 CCN puis nous avons décidé de restreindre notre périmètre d'étude. D'une part, afin de réduire le temps de traitement des données, et d'autre part car nous avons remarqué une nette différence de comportement des assurés entre les différentes CCN. Afin d'obtenir des résultats cohérents, il était nécessaire de travailler sur des CCN présentant sensiblement les mêmes caractéristiques. Notre choix s'est basé sur deux critères de sélection. Le premier étant la présence de la CCN dans notre portefeuille dès 2011 car notre étude portera sur trois années, de 2011 à 2013. Le deuxième étant le volume d'affaires représentant au moins 60% du CA pour les CCN sélectionnées. Finalement, nous avons retenu les trois CCN suivantes :

- la CCN Boulangerie
- la CCN Afflec (Alimentation)
- la CCN Patisserie

Au total, l'étude porte sur 229 410 bénéficiaires, dont 228 972 adhérents.
Dans la même optique d'optimisation du temps, nous avons limité le nombre de garanties sur lesquelles travaillé. L'ensemble des prestations concernant 32 garanties, et le but étant de voir si notre méthode peut être fiable, il n'était pas nécessaire de tarifer chacune d'elles. Là encore, nous nous sommes limités aux garanties représentant le plus gros volume en termes de prestations, tout en veillant à conserver les garanties servant de référence à l'étude du BIPE. Étude nous ayant servi pour le découpage par niveaux de couverture et qui sera expliqué dans la suite du mémoire. Pour chaque poste – hospitalisation, actes médicaux, dentaire et optique – nous avons retenu au moins deux garanties différentes. La méthode a été la suivante :

1. Sélection des premières garanties correspondant à au moins 60% du volume total des montants engagés par les assurés. Dans cette sélection sont déjà présentes une garantie du poste Hospitalisation (frais de séjour), une garantie du poste dentaire (prothèses dentaires remboursées par la SS), une garantie du poste optique (verres) et trois garanties du poste actes médicaux (consultations généralistes, consultations spécialistes et pharmacie).

2. Sélection d'une garantie supplémentaire pour chacun des postes hospitalisation, actes médicaux dentaire et optique. Les trois garanties – radiologie, soins dentaires et monture – coïncident aussi avec les garanties représentant le plus gros volume de consommation directement après celles déjà sélectionnées. Ce qui rend notre choix plus pertinent. La quatrième garantie chambre particulière (poste hospitalisation) a été rajoutée car elle est une des garanties utilisées par l'étude du BIPE.

Les garanties que nous avons choisies, ainsi que leur part dans le volume total des prestations sont représentées sous le graphique qui suit :

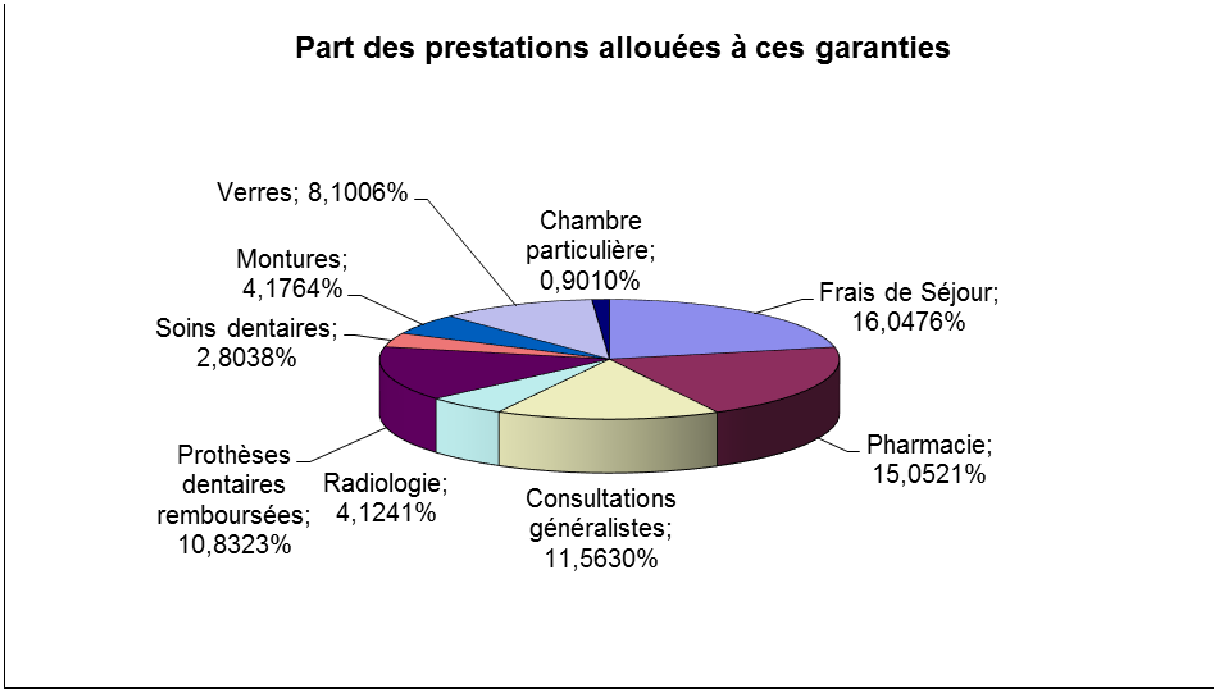


Figure 14. Part des prestations allouées aux garanties sélectionnées

Près de 75% de la consommation des assurés est répartie entre ces neuf garanties. Elles représentent aussi 60% des actes en termes de fréquence de consommation.

Nous présentons ici d'autres statistiques pertinentes concernant le portefeuille sur lequel notre étude se porte.

Sexe

Sexe	**Nombre de bénéficiaires**
Masculin	104371
Féminin	125039

Tableau 3. Nombre de bénéficiaires par sexe

Le sexe étant un facteur discriminant, il est donc indispensable pour l'assureur de savoir quel est le nombre de bénéficiaires masculins et féminins couverts. Le portefeuille de notre étude est très équilibré au niveau de sa répartition Homme/Femmes (45/55), enlevant ainsi toute crainte d'une sur représentation d'un des deux sexes. La première idée était de séparer notre étude en deux bases, hommes et femmes, afin d'éliminer les effets d'interaction possible entre le sexe et le type de montage base - surcomplémentaire. L'intérêt étant alors une meilleure estimation de l'effet dû uniquement au montage souscrit par l'assuré. Cette idée aurait pu être menée jusqu'au bout car au vu de ce tableau nous aurions disposé de données suffisantes pour les deux bases. Néanmoins, la création de ces deux bases auraient alors nécessité de doubler toutes les bases pour chaque garantie aussi, entrainement un temps de traitement conséquent. Cette séparation des bases pourra être envisagée par la suite, une fois nos méthodes de tarification validées. Finalement, pour cette première application de nos méthodes, nous avons décidé d'établir un tarif adulte unique sans différenciation du sexe. Ce tarif étant calculé à partir d'une base totalement équilibrée nous pourrons alors déterminer quel sexe est susceptible d'impacter réellement ce tarif en fonction de l'historique de consommation chez AG2R La Mondiale ou en fonction de statistiques nationales. Le tarif Adulte obtenu pourra alors être pondéré par ce rapport de consommation pour déterminer le tarif associé à chaque sexe.

Exposition moyenne des assurés

Le temps d'exposition moyen dans notre portefeuille pour un individu est de 18 mois. Le temps minimum est de moins d'un mois. Cette différence au niveau du temps d'exposition pourra poser un problème de surestimation de la prime pour ces individus restés couverts peu de temps. Cet aspect sera plus développé dans les prochains paragraphes.

Âge moyen

Statistique	**Âge**
Min	0
Max	109
Moyen	35,6
Médian	35

Tableau 4. Statistiques sur l'âge du portefeuille

L'âge moyen étant approximativement de 35,6 ans, nous avons donc un échantillon relativement jeune. La population de référence étant généralement fixée avec un âge moyen proche de 40 ans, il faudra envisager un redressement par rapport aux résultats que nous obtiendrons. L'âge moyen du portefeuille total est encore plus bas puisqu'il est alors égal à 34 ans. Cet âge très jeune est une des raisons qui rendent la consommation santé des CCN très spécifique.

Statut professionnel

Statut professionnel	**Nombre de bénéficiaires adhérents**
Ensemble de Personnel	226598
Cadres	227
Non Cadres	1711
Travailleur Non Salarié	186
Loi Evin	250

Tableau 5. Nombre de bénéficiaires adhérents par statut professionnel

Le statut professionnel étant un critère tarifaire d'importance, nous avons décidé de s'intéresser aux catégories concernées par notre périmètre d'étude. Les trois CCN contractent majoritairement des contrats pour des catégories « Ensemble de Personnel », sans distinguer donc les « Cadres » des « Non Cadres ». Les contrats EVIN concernent les retraités continuant de bénéficier de l'assurance complémentaire à laquelle ils avaient souscrite encore salarié. La Loi Evin, qui ne sera pas détaillée dans ce mémoire, impose aux assureurs de proposer aux salariés partant en retraite une couverture égale à celle souscrite en tant que salarié, avec une majoration d'au plus 50%.

L'assureur perd de l'information sur ces bénéficiaires, il est donc difficile d'établir un tarif adéquat. Un exemple simple concerne la zone géographique. Elle ne peut plus être un critère car nous ne savons pas où ces retraités décident de s'installer. De plus, à la retraite, leur consommation peut être totalement différente. Portée à la hausse par manque d'activité et par une vie privée plus tranquille leur permettant de s'occuper de leur santé. Ou tout aussi bien portée à la baisse par une meilleure prévention et un stress disparaissant. Nous les gardons tout de même dans l'étude, car les coefficients de pondération obtenus pourront être utiles dans le cadre d'études futures sur leur consommation.

Régime de sécurité sociale

Régime	Nombre de bénéficiaires adhérents
Général	219105
Alsace Moselle	8272
Travailleur Non Salarié	1591
Autres	4

Tableau 6. Nombre de bénéficiaires adhérents par régime de sécurité sociale

Nous avons décidé d'enlever de l'étude les individus concernés par le régime « Autres » car leur nombre était insuffisant pour que nos résultats soient fiables.

Option souscrite

Option	Nombre de bénéficiaires adhérents
Base	191586
Surcomplémentaire 1	12713
Surcomplémentaire 2	23039
Surcomplémentaire 3	1634

Tableau 7. Nombre de bénéficiaires adhérents par option

La majorité des adhérents ne souscrivent à aucune option car près de 84% d'entre eux ne sont couverts que par la base. Cela pourra fragiliser notre calcul par MLG car les effectifs sont disproportionnés entre ces quatre catégories de produits.

Mariés - Option

Option	Pourcentage de mariés ou union libre
Base	5,1%
Surcomplémentaire 1	5,9%
Surcomplémentaire 2	6,5%
Surcomplémentaire 3	9,9%

Tableau 8. Proportions de bénéficiaires adhérents mariés selon l'option souscrite

Cette statistique est intéressante car elle permet d'identifier les personnes susceptibles de souscrire à des surcomplémentaires. On remarque ainsi dans le tableau ci-dessus qu'un individu ayant souscrit une option 3 a presque deux fois plus de chances de vivre en concubinage qu'une personne ne souscrivant qu'à la base proposée par l'entreprise. Nous pouvons en déduire qu'un assuré ayant une famille aura plus tendance à mieux se couvrir qu'un célibataire n'ayant ni concubine ni enfants.

Âge - Option

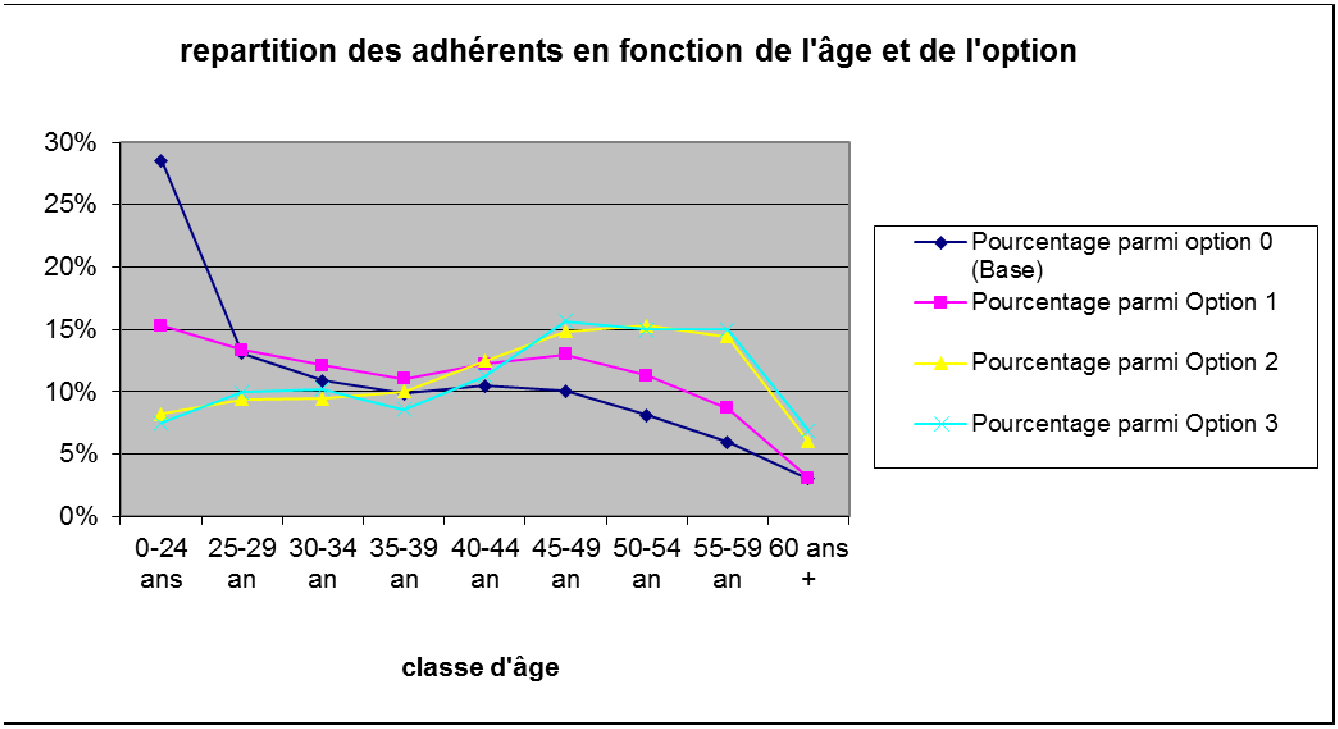


Figure 14. Répartition des adhérents en fonction de l'âge et de l'option souscrite

Nous représentons dans ce graphique le pourcentage d'adhérents par classe d'âge et par option souscrite. Ceci nous permet de remarquer les tranches d'âge les plus enclines à mieux se couvrir. On remarque ainsi que les individus entre 45 et 49 ans sont la catégorie d'assurés la plus représentative parmi les souscripteurs de surcomplémentaires. Leur poids augmente avec le niveau de surcomplémentaire choisie. A l'inverse, le poids des jeunes entre 0 et 29 ans diminue au fur et à mesure que l'on monte en gamme. Cela est tout à fait cohérent dans la mesure où les assurés entre 45 et 49 ans vont correspondre à des profils cadres à revenus plus élevés et vivant majoritairement en famille. Tandis que les jeunes représentent une population moins fortunée, vivant encore pour la plupart seuls et se préoccupant moins de leur santé.

b) Tri à plat

Nous avons complété notre analyse exploratoire par un tri à plat. Cette étude nous permet d'avoir un premier aperçu d'une surconsommation due à une classe tarifaire. Pour chaque garantie et pour chaque variable explicative, nous déterminons la fréquence et le coût moyen en fonction des différentes modalités de la variable. Ces deux valeurs sont rapportées à la fréquence moyenne et aux frais réels moyens sur l'ensemble du portfeuille afin d'appréhender plus facilement les coefficients de majoration et de minoration. Nous présentons ci-dessous les sorties SAS obtenues pour les garanties « généralistes » et « optique », et la variable Montage. Nous ne traitons ici que de la consommation adulte. La variable montage correspondant au niveau de couverture de l'ensemble Base + Surcomplémentaire.

Les six modalités de cette variable sont définies de la sorte :

- BN : Base Basse et Surcomplémentaire Nulle
- BB : Base Basse et Surcomplémentaire Basse
- BH : Base Basse et Surcomplémentaire Haute
- MN : Base Moyenne et Surcomplémentaire Nulle
- MB : Base Moyenne et Surcomplémentaire Basse
- MM : Base Moyenne et Surcomplémentaire Moyenne

Consultations généralistes

Montage	Fréquence moyenne	Frais réels moyens	Impact sur la Fréquence moyenne (%)	Impact sur les Frais réels moyens (%)	Impact sur Consommation totale (%)
MN	12.0929	23.2912	97.163	100.009	97.172
MM	12.0929	23.2912	100.876	100.068	100.945
BN	12.0929	23.2912	102.830	100.015	102.846
BB	12.0929	23.2912	96.192	99.889	96.085
BH	12.0929	23.2912	97.650	99.700	97.357
MB	12.0929	23.2912	112.751	100.082	112.844

Tableau 9. Impact du montage sur la consommation sur l'acte de santé « Généralistes »

Les résultats obtenus semblent peu cohérents par rapport à ceux auxquels nous nous attendions. Les individus ayant une base basse et une surcomplémentaire nulle prennent plus souvent rendez-vous avec un généraliste que ceux bénéficiant d'une base similaire et d'une meilleure surcomplémentaire. Même constat en ce qui concerne les frais réels. Ces résultats peuvent cependant se justifier par le fait qu'une grande majorité des généralistes sont en secteur conventionné 1. Ces médecins ne pratiquent pas de dépassement d'honoraires. Le tableau indique d'ailleurs que les frais réels moyens pour cet acte sont de 23,3 euros, soit approximativement 100% de la base de remboursement. Ainsi, le niveau de couverture a peu d'impact car dans tous les cas l'assuré n'aura pas de reste à charge.

Consommation sur le poste « Optique »

Montage	Fréquence moyenne	Frais réels moyens	Impact sur la Fréquence moyenne (%)	Impact sur les Frais réels moyens (%)	Impact sur la Consommation totale (%)
MN	5.40696	145.790	123.637	95.613	118.213
MM	5.40696	145.790	107.834	113.466	122.355
MB	5.40696	145.790	103.649	97.934	101.508
HB	5.40696	145.790	85.237	107.554	91.676
HN	5.40696	145.790	85.163	96.441	82.133
MH	5.40696	145.790	103.871	120.337	124.996

Tableau 10. Impact du montage sur la consommation de l'acte de santé « Optique »

Les résultats obtenus pour le poste « Optique » confirment un peu mieux ce que nous attendions, mais uniquement au niveau des frais réels. Le montant moyen engagé par les assurés ayant une base moyenne augment avec le niveau de la surcomplémentaire. On observe la même évolution pour les assurés ayant une base de niveau élevé. Ces résultats ne sont pas retrouvés en ce qui concerne la fréquence de consommation mais cela s'explique assez aisément. En effet, l'optique est un poste spécifique car il est souvent nécessaire de passer d'abord par un spécialiste, en l'occurrence un ophtalmologue. Ce qui oblige en quelque sorte l'assuré à se rendre chez un opticien uniquement lorsque le besoin est réel. Ce sera donc - dans la plupart des cas - plus le besoin et l'ordonnance délivrée par un spécialiste qui pousseront l'assuré à se rende chez un opticien que son envie de consommer. Ainsi, le niveau de couverture ne sera pas vraiment ce qui influe sur le fait qu'il se rendre plusieurs fois chez un opticien. Par contre le niveau de couverture lui permettra de s'offrir une monture plus esthétique, de meilleure marque, et donc plus chère.

Les résultats pour les autres garanties sont assez semblables à ceux présentés ci-dessus, avec des fréquences de consommation ne vérifiant pas toujours ce à quoi nous devrions observer en fonction du niveau des montages. L'étude a été menée sur d'autres variables - sexe, âge, situation familiale, zonier, régime, statut professionnel - mais les résultats ne sont pas présentés ici. En effet, ce sont des facteurs discriminants classiques et leur effet a bien été vérifié.

c) Courbes de consommation

Dans cette partie, nous essayons de percevoir l'allure des courbes de consommation, au niveau de la fréquence et des frais réels engagés par l'ensemble des assurés. Cette étude statistique permet à l'actuaire de visualiser la tendance de consommation du portefeuille et de déterminer une loi usuelle la décrivant au mieux. L'intérêt de réaliser ce travail est donc justifié car il nous indique le paramétrage à choisir dans la programmation du MLG, notamment sur le choix de la loi à utiliser pour la variable réponse. Pour réaliser ce travail, nous avons travaillé sur SAS pour les frais réels et sur R pour les fréquences. Les codes programmés sont présentés en annexe-II.

Frais réels - Optique

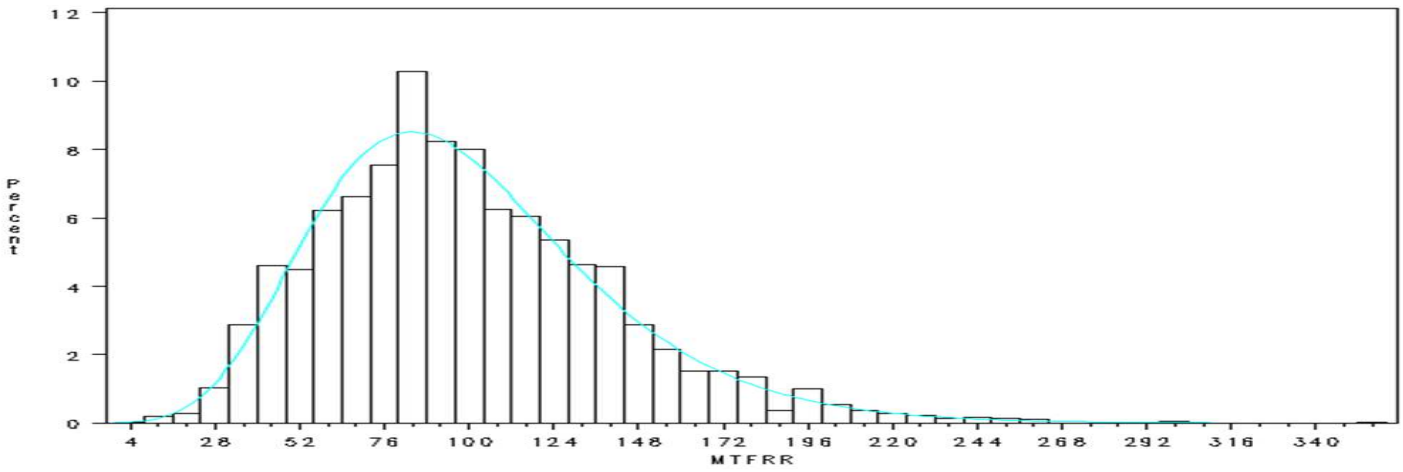


Figure 15. Histogramme des frais réels pour la garantie « Optique »

La courbe en bleue correspond à une loi log - normale. Nous remarquons que cette loi semble très bien s'ajuster à l'histogramme de consommation.

Fréquence - Optique

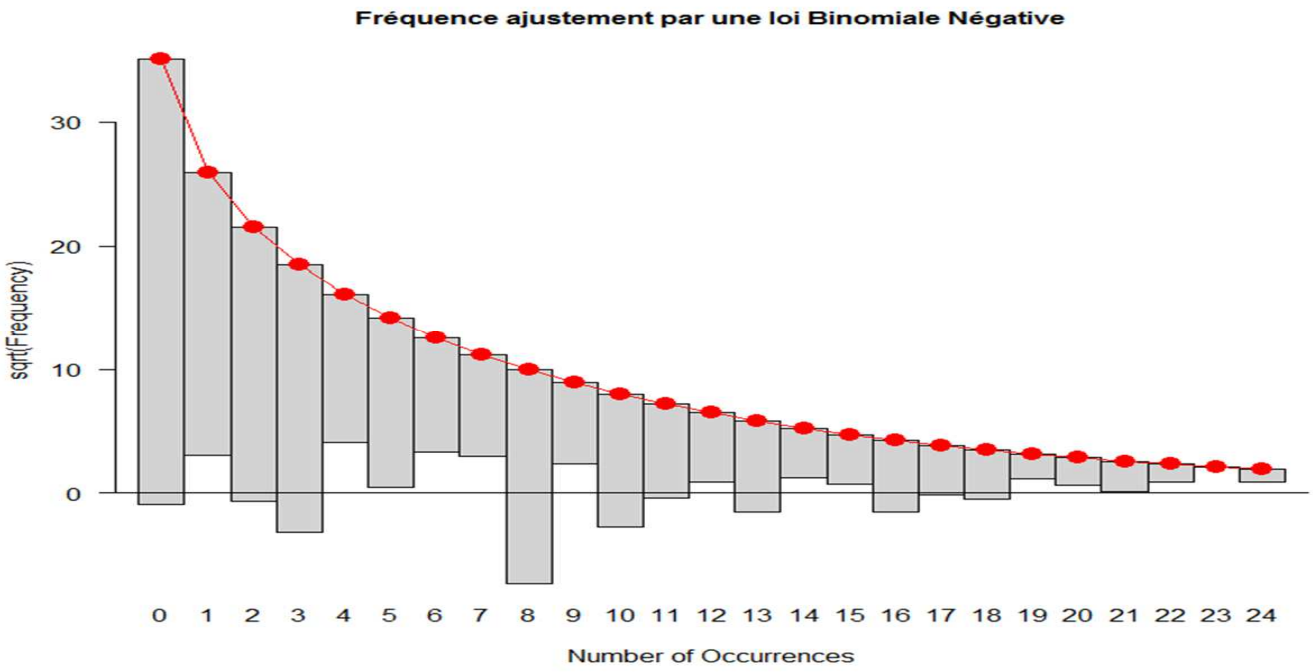


Figure 16. Histogramme de la fréquence pour la garantie « Optique »

Ici, nous essayons d'ajuster une loi binomiale au taux de recours pour l'acte « Chambre particulière ». Plus la base d'un histogramme s'écarte de l'axe des abscisses, moins fiable est l'ajustement. Pour cette garantie, l'ajustement semble donc raisonnablement justifié.

La loi log - normale est la loi qui semblait la plus adapté pour chacune des neuf garanties, au niveau des frais réels. En ce qui concerne la fréquence de consommation, la loi binomiale négative est celle qui ajustait constamment le mieux les histogrammes de consommation.
De manière générale, l'ajustement de loi semblait plus satisfaisant pour les frais réels que pour les fréquences.

V. Application des méthodes

Dans cette section nous présentons les méthodes de modélisation utilisés et leurs principales étapes ayant permis d'aboutir aux primes pures. Une partie des résultats sera illustrée dans ce mémoire.

A. Application du Modèle Linéaire Généralisé

Cette section a pour objectif d'expliquer la mise en place de notre MLG et les résultats obtenus. L'application de ce modèle a suivi les étapes que nous avons présentées précédemment dans le chapitre réservé à cette méthode de modélisation.

a) Création des bases

Pour chaque garantie, nous avons d'abord séparé les adultes des enfants. Ensuite, nous avons dû constituer deux bases distinctes : une base servant à la modélisation de la fréquence et l'autre servant à la modélisation des frais réels. En effet, pour calculer une fréquence il faut avoir à disposition tous les bénéficiaires ayant consommé ou non. Tandis que pour calculer des frais réels moyens, il faut travailler uniquement sur la population consommatrice.

Les tables pour la fréquence

Dans un premier temps nous avons créé la variable « lnexposition » qui est le logarithme de l'exposition de chaque assuré. Cette variable est celle qui servira d'offset à la procédure GENMOD de SAS. La notion d'offset sera expliquée plus en détail lors de la description de la procédure GENMOD implémentée dans notre programme.
Ensuite, nous avons créé la variable « A_consommé ». Pour expliquer la démarche prenons le cas de l'acte « Soins dentaires ». Pour chaque ligne de consommation, la variable « A_consommé » prend la valeur 1 si la garantie consommée est « Soins dentaires » ou 0 le cas contraire. Nous obtenons donc une base avec toutes les lignes de consommation et la valeur 0 ou 1 pour la variable « A_consommé ». Ceci nous permet alors de calculer pour chaque individu le nombre de fois qu'il a consommé l'acte « Soins dentaires ». Ce processus est reproduit pour chacune des neuf garanties étudiées.

Les tables pour les frais réels

Nous récupérons pour chaque garantie, uniquement les bénéficiaires ayant consommé l'acte étudié. Deux choix sont possibles. Le premier consiste à agréger les montants et le nombre d'actes par bénéficiaire.
Dans ce cas- ci, nous nous retrouvons alors avec une base constituée d'une ligne unique par bénéficiaire indiquant le montant total engagé et le nombre de fois que le bénéficiaire a consommé l'acte étudié. Dans la procédure GENMOD, il sera alors nécessaire d'imposer le nombre d'actes en offset. Le deuxième choix, pour lequel nous avons opté, est de laisser toutes les lignes de consommation sans aucune agrégation. La raison est simple : nous voulons garder l'information sur le montant engagé exact pour chaque acte consommé.

b) Test d'indépendance du Khi 2

Le test d'indépendance du Khi 2 permet de voir pour chaque garantie, quelles sont les variables qualitatives influentes sur la fréquence de consommation. Pour la fréquence, nous avons utilisé la variable « A_consommé ». Cela n'est pas forcément le plus approprié étant donné que sur trois années, il y a de fortes chances que chaque assuré ait eu recours à chacune des garanties. L'idéal serait de pouvoir utiliser comme variable le nombre d'actes consommés. Cela n'a pu être retenu car le croisement des deux variables pouvaient donner des effectifs inférieurs à 5, enlevant alors toute significativité au test. Nous ne présentons ici que les résultats pour la variable « Montage » présentée plus haut, car les autres variables explicatives sont classiquement utilisées et nous savons qu'elles influent fortement sur la consommation. Pour chaque garantie, nous nous référons alors à la table de Khi 2, afin de savoir si la valeur obtenue dépasse le seuil de significativité et si la p value associée permet d'accepter ou rejeter l'hypothèse d'indépendance.

Prothèses dentaires

Statistique	DDL	Valeur	P - value
Khi 2	10	32462.5855	<.0001

Tableau 11. Test du Khi 2 sur les variables « A consommé » et « Montage »

Au vu du résultat obtenu, nous pouvons supposer que le montage influe la fréquence de consommation des assurés.

Radiologie - Test du Khi 2 sur les variables « Frais réels » et « Montage »

Statistique	DDL	Valeur	P - value
Khi 2	10	4708.1660	<.0001

Tableau 12. Test du Khi 2 sur les variables « A consommé » et « Montage »

Même constat pour la garantie radiologie. Nous obtenons un rejet de l'hypothèse d'indépendance pour chacune des garanties de l'étude. L'intérêt d'intégrer le niveau de la base et de la surcomplémentaire dans notre modélisation de la fréquence est donc justifiée.

Nous avons aussi réalisé un test d'indépendance entre les frais réels et la variable « Montage » en utilisant les coefficients de Spearman. Ceux-ci ont été préférés aux coefficients de Pearson car ils imposent moins de conditions, notamment au niveau de la linéarité des observations. Néanmoins, les résultats sont peu pertinents et n'influencent pas notre choix pour un premier tri de variables.

Prothèses dentaires

Statistique	Valeur	P - value
Coefficients de corrélation de Spearman	0,10361	<.0001

Tableau 13. Coefficients de Spearman sur les variables « Frais réels » et « Montage »

Nous regardons surtout la p - value car l'échantillon étant très grand il est difficile de fixer une valeur seuil. Mais 0,10 semble convenir au vu de la taille de notre effectif ayant consommé cet acte. La p - value indique un test très significatif où la variable « Montage » influe sur le montant des frais réels.

c) Variables pré sélectionnées

Les variables pré - sélectionnées sont celles que nous avons intégrées à l'étude en fonction des statistiques descriptives telles que le tri à plat et les tests d'indépendance réalisés.

Elles sont au nombre de dix :

- l'âge
- le sexe
- le niveau de la base
- le niveau de la surcomplémentaire
- le statut professionnel
- le secteur d'activité
- le zonier
- le régime de Sécurité Sociale
- la situation familiale de l'assuré
- l'effectif de l'entreprise

Les variables ci-dessus sont celles pré - sélectionnées pour l'étude de la consommation adulte. Pour la consommation enfant, nous n'avons pas gardé l'âge, le sexe, et la situation de l'assuré. En effet, pour l'âge et le sexe, ce sont des informations dont l'assuré n'a pas forcément accès au moment de la tarification.
Certaines de ces variables sont un peu moins interprétables que d'autres. C'est le cas de l'effectif de l'entreprise par exemple. En effet, il est étonnant de se dire que la consommation d'un assuré varie en fonction du nombre de personnes employées dans son établissement. Pourtant, c'est une variable fréquemment utilisée en tarification de contrats collectifs car il est observé une dérive de consommation sensiblement différente en fonction de la taille d'entreprise.
Certaines de ces variables sont aussi relativement corrélées entre elles. On parle alors d'interactions. Une interaction entre deux variables est le fait que les modalités de l'une influent sur les modalités de l'autre. Les tests d'indépendance du Khi 2 montrent une liaison non négligeable entre le niveau de la base et le niveau de la surcomplémentaire. Nous avons aussi, pour compléter cette idée, construit une table de corrélation avec SAS Guide, dont nous présentons aussi les résultats ci-dessous. Cette table de corrélation nous montre aussi une corrélation entre ces deux variables.

Optique

Statistique	**DDL**	**Valeur**	**P - value**
Khi 2	3	49816.3890	<.0001

Tableau 14. Test du Khi 2 sur les variables « Niveau Base » et « Niveau Surcomplémentaire »

Variable	age	sexe	zonier	statut Professionnel	situation familiale	Secteur d'activite	effectif	Niveau_Base	Niveau_Surcomp	regime
age	1,000	-0,079	0,000	0,030	0,115	0,051	0,149	-0,082	0,177	0,020
sexe	-0,079	1,000	0,030	0,023	-0,013	-0,014	-0,036	0,018	-0,041	0,016
zonier	0,000	0,030	1,000	0,015	0,001	0,023	-0,048	-0,023	-0,035	-0,135
statut Professionnel	0,030	0,023	0,015	1,000	0,067	0,091	-0,046	-0,101	0,049	0,196
situation familiale	0,115	-0,013	0,001	0,067	1,000	0,012	-0,033	-0,002	0,011	0,029
Secteur d'activite	0,051	-0,014	0,023	0,091	0,012	1,000	-0,148	-0,919	0,277	0,010
effectif	0,149	-0,036	-0,048	-0,046	-0,033	-0,148	1,000	0,111	0,665	-0,053
Niveau Base	-0,082	0,018	-0,023	-0,101	-0,002	-0,919	0,111	1,000	-0,290	0,013
Niveau Surcomp	0,177	-0,041	-0,035	0,049	0,011	0,277	0,665	-0,290	1,000	-0,004
regime	0,020	0,016	-0,135	0,196	0,029	0,010	-0,053	0,013	-0,004	1,000

Tableau 15. Table de corrélation

Pour l'acte « Optique », nous observons que le niveau de base est corrélé principalement avec deux variables, le secteur d'activité et le niveau de la surcomplémentaire. Cette corrélation avec le secteur d'activité est très forte et est en particulier due au fait que nous ne travaillons qu'avec trois CCN disposant chacune d'un seul niveau de base. Cette corrélation n'est donc pas forcément représentative de l'ensemble du portefeuille en santé collective.
La corrélation avec le niveau de surcomplémentaire est importante et négative. Cela se comprend naturellement car l'assuré disposant d'un contrat de base suffisant éprouve alors moins le besoin de se couvrir encore par une surcomplémentaire.
En ce qui concerne le niveau de surcomplémentaire, nous observons une corrélation forte avec l'effectif de l'entreprise. Cependant cette corrélation n'est pas observable pour toutes les garanties. Enfin, le niveau de surcomplémentaire semble être lié à l'âge de l'assuré. Ce qui appuie nos premières intuitions énoncées dans la section « Statistiques descriptives ».

Ces effets d'interactions entre variables peuvent être précisées dans un MLG afin qu'il prenne automatiquement en compte le croisement des variables. Le nombre de croisements possibles étant très élevé, nous avons choisi l'interaction à trois variables. Ces trois variables sont l'âge, le niveau de la base et le niveau de la surcomplémentaire. Cette interaction a été choisie car elle répondait à trois critères :

- *la corrélation observée entre ces variables était conséquente pour chacune des garanties*
- *l'interaction était facilement interprétable*
- *les tests de Fischer validaient la significativité de l'interaction pour la modélisation de la variable réponse*

Pour le dernier critère, la méthode a été de comparer le R^2 et de vérifier la statistique de Fisher pour les modèles suivants :

- *le modèle intégrant l'interaction des trois variables et toutes les autres variables explicatives*

- *le modèle excluant deux des trois variables étudiées (âge, Base et Surcomplémentaire) et retenant toutes les autres variables explicatives*

- *le modèle intégrant toutes les variables mais sans préciser aucun effet d'interactions*

Chambre particulire

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	78	357399.129	4582.040	6.05	<.0001
Error	9970	7549458.341	757.217		
Corrected Total	10048	7906857.470			

R-carré	Coef de Var	Racine MSE	MTFRR Moyenne
0.045201	50.47719	27.51759	54.51489

Figure 17. Sortie SAS Tests de Fisher Intéraction âge * Base * Surcomplémentaire intégrée

Nous avons obtenu pour chaque variable un test validant l'intégration de l'interaction à trois variables pour la consommation adulte. Pour la consommation enfant, nous avons intégré le croisement à deux variables entre le niveau de base et le niveau de surcomplémentaire Cela nous a donc permis de supprimer les effets d'interaction sur les deux variables nous important le plus dans cette étude, et de rendre notre résultat interprétable car il apparait logique qu'avec l'âge le besoin de se couvrir plus augmente.

d) Sélection des variables

Pour sélectionner nos variables, nous avons d'abord programmé une méthode -disponible en annexe-III - combinant les méthodes forward et backward. Nous réalisons d'abord un processus forward, ajoutant à chaque étape la variable la plus significative au sens du R^2 ajusté. Le modèle obtenu est alors soumis à un deuxième processus de sélection, de type backward cette fois. Les variables dont la p value est supérieur au seuil 0.05 sont supprimées du modèle et l'on regarde la variation en R^2 résultant de cette suppression. Les variables finalement sélectionnées sont alors intégrées au modèle dans une procédure GENMOD avec une option de type 3. Cette option permet de réaliser les tests de signification basés sur la déviance pour chaque variable du modèle. Elle nous permet donc aussi de valider notre sélection forward - backward. Validation qui n'a pas été vérifiée pour toutes les garanties, en particulier dans le cadre de la modélisation de la fréquence. En effet, nous avons obtenu pour certaines garanties une sortie SAS indiquant la non convergence de notre modèle vers un résultat fiable. Les tableaux 16 et 17 récapitulent les variables retenues ou non par notre programme :

Frais réels

Variable / Garantie	Age Niveau_Base Niveau_Surcomp	Sexe	Situation de l'assuré	Statut professionnel	Secteur d'activité	Zonier	Effectif	Régime
Généralistes	Oui	Non	Oui	Oui	Oui	Oui	Oui	Oui
Radiologie	Oui	Oui	Non	Oui	Oui	Oui	Oui	Oui
Spécialistes	Oui	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Optique	Oui	Oui	Non	Oui	Oui	Oui	Oui	Oui
Séjour	Oui	Oui	Non	Non	Non	Oui	Oui	Oui
Chambre particulière	Oui	Non	Non	Non	Non	Oui	Oui	Non
Soins dentaires	Oui	Oui	Non	Non	Oui	Oui	Oui	Oui
Pharmacie	Oui	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Prothèses dentaires	Oui	Non	Non	Oui	Oui	Oui	Oui	Oui

Tableau 16. Variables retenues pour les frais réels

Fréquence

Variable / Garantie	Age Niveau_Base Niveau_Surcomp	Sexe	Situation de l'assuré	Statut professionnel	Secteur d'activité	Zonier	Effectif	Régime
Généralistes	Oui	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Radiologie	Non	Non	Non	Oui	Oui	Oui	Oui	Oui
Spécialistes	Oui	Oui	Oui	Oui	Oui	Oui	Oui	Oui
Optique	Non	Non	Non	Oui	Oui	Oui	Oui	Oui
Séjour	Non	Non	Oui	Oui	Oui	Oui	Oui	Oui
Chambre particulière	Oui	Oui	Oui	Non	Oui	Oui	Oui	Non
Soins dentaires	Non	Non	Non	Non	Non	Non	Non	Non
Pharmacie	Oui	Oui	Non	Oui	Oui	Oui	Oui	Oui
Prothèses dentaires	Non	Non	Oui	Oui	Oui	Non	Oui	Non

Tableau 17. Variables retenues pour la fréquence de consommation

Nous remarquons que notre programme a rejeté toutes les variables pour la modélisation de la fréquence de consommation en soins dentaires. Il en est de même pour plusieurs garanties au niveau de la consommation enfant. Nous avons dans ces cas ci effectué sur SAS une PROC MEANS de manière à récupérer le nombre d'actes moyens et l'exposition moyenne en fonction des niveaux de base et de surcomplémentaire. La fréquence est alors calculée en faisant le rapport des deux.

Pour d'autres garanties, nous remarquons que l'interaction à trois variables prenant en compte l'âge, le niveau de base et le niveau de surcomplémentaire, n'a finalement pas été retenu. Dans ces cas ci, nous avons réitéré notre programme mais en spécifiant différentes interactions. Finalement, seul le croisement entre l'âge et le sexe, interaction souvent gardée en tarification, permettait d'avoir une convergence vers des estimateurs.

Les garanties concernées par cette modification sont celles présentées ci-dessous, pour lesquelles nous indiquons les variables finalement retenues.

Prothèses Dentaires - frais réels
Variables retenues :

- statut professionnel
- secteur d'activité
- effectif
- régime
- zonier
- niveau de la surcomplémentaire
- sexe et âge

Radiologie - fréquence
Variables retenues :

- statut professionnel
- secteur d'activité
- effectif
- régime
- zonier
- niveau de la surcomplémentaire
- sexe et âge

Optique - fréquence
Variables retenues :

- statut professionnel
- secteur d'activité
- effectif
- régime
- zonier
- niveau de la surcomplémentaire
- niveau de la base
- sexe et âge

Les variables retenues pour le MLG, il reste maintenant à définir la structure globale du comportement des assurés vis-à-vis de la consommation en santé.

e) Le paramétrage du modèle

Comme nous l'avons expliqué dans le chapitre concernant les MLG, l'actuaire doit spécifier deux paramètres indispensables lors d'une modélisation : la loi de la variable à expliquer et la fonction lien traduisant sa relation avec les variables explicatives.

1. Choix de la distribution

La distribution de la variable réponse a été choisie en fonction des statistiques descriptives de notre portefeuille. Nous avons remarqué que pour les frais réels, la courbe semblait suivre une loi Log Normale quel que soit la garantie. Au niveau de la fréquence de consommation, la loi Binomiale Négative semblait plus adaptée. Afin de s'assurer de la légitimité de notre choix pour les frais réels, nous avons tracé les graphes Q - Q plots. Pour la fréquence de consommation, le test d'ajustement, réalisé sur R, a déjà été présenté plus haut.

Optique

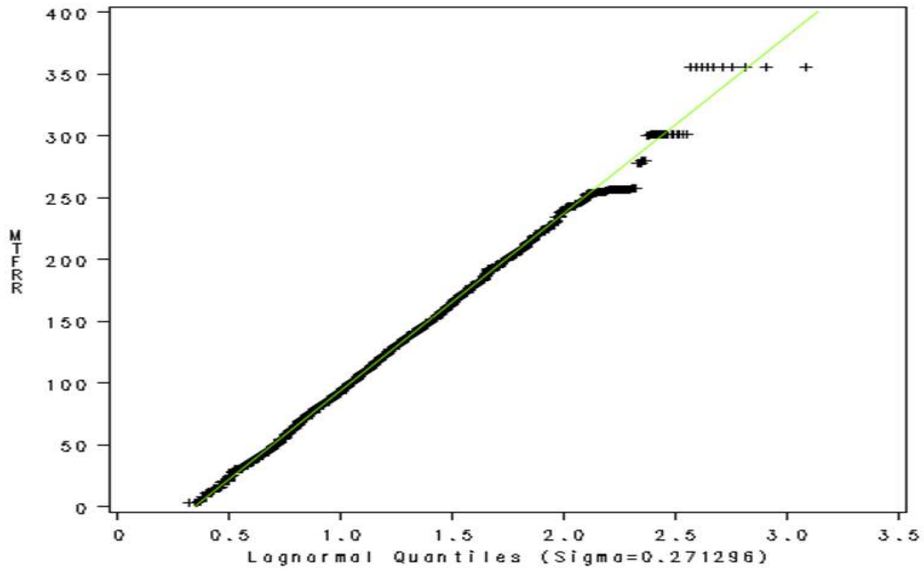


Figure 18. Q - Q plot ajustement des frais réels par une loi Log Normale

Nous reprenons l'exemple de l'acte « Optique ». Nous avions choisi une loi Log Normale pour simuler les montants de frais réels engagés par les consommateurs. Sur ce graphique, nous remarquons que les points sont relativement bien alignés avec la droite, validant ainsi notre choix.
Pour les garanties dont le choix était moins évident, nous réalisions deux MLG, et prenions la loi pour laquelle la déviance est la plus faible. La loi Log Normale a été finalement choisie pour les frais réels et pour toutes les garanties. La loi Binomiale négative a été choisie pour la modélisation de la fréquence de consommation. La loi Log Normale étant celle la plus fréquemment utilisée pour modéliser les coûts de sinistre en assurance non vie, notre choix semble tout à fait justifié. Il en est de même pour la loi Binomiale Négative qui est préférée à la loi de poisson car elle diminue fortement l'effet de surdispersion.
Au niveau des frais réels, le choix d'une loi Log Normale nous oblige à travailler alors avec le logarithme des frais réels et d'utiliser une loi normale pour cette nouvelle variable créée.

2. Choix de la fonction lien

La fonction logarithmique a été choisie comme fonction lien dans notre modèle. Le choix s'est porté sur cette fonction car elle permet d'avoir en sortie un modèle multiplicatif. Les coefficients étant au logarithme, le passage à l'exponentielle nous donne une multiplication des coefficients entre eux. Ce modèle multiplicatif s'adapte parfaitement à notre étude et au domaine de la tarification, car chaque estimation s'interprète aisément comme un coefficient de majoration ou de minoration par rapport à un individu de référence.

3. Intégration d'un offset

Nous intégrons dans le modèle ce que l'on appelle un offset. Nous avons vu auparavant que dans un MLG, la variable à expliquer pouvait ne plus dépendre linéairement des variables explicatives. Or, nous pouvons tout de même avoir une relation linéaire entre la variable réponse et certaines variables explicatives. Ainsi, lorsque l'on a ce type de relation avec une variable explicative X, cette variable X est alors déclarée en offset dans le modèle. Dans notre étude, nous avons des individus ayant été couverts pendant des périodes différentes. Devons-nous alors considérer qu'un assuré A couvert depuis 12 mois et ayant été chez le généraliste six fois consomme plus de fois qu'un assuré B couvert depuis 6 mois et ayant

été chez le généraliste à quatre reprises ? Nous avons décidé de considérer que la fréquence est proportionnelle au temps de couverture d'un assuré. Cela veut donc dire que la fréquence annuelle de notre assuré A est de 6 tandis qu'elle est de 8 pour l'assuré B. Cette prise en compte de l'exposition n'est pas forcément la plus adapté et sera discutée dans la suite de ce mémoire.

f) Validité du modèle

Pour les frais réels, la modélisation par MLG est transcrite sous la forme suivante :

$$ln[E(LnFrais\ réels)] = \alpha_0 + \alpha_1 \times x_1 + \cdots + \alpha_p \times x_p$$

d'où :

$$E(LnFrais\ réels) = \exp(\alpha_0) \times \prod_{j=1}^{p} \exp(\alpha_j)$$

Avec : $\begin{cases} LnFrais\ réels\ suivant\ une\ loi\ Normale \\ \alpha_j\ les\ coefficients\ estimés \\ \alpha_0\ le\ coefficient\ correspondant\ \text{à}\ l'individude\ référence \end{cases}$

Enfin :

$$E(Frais\ réels) = \exp\left\{E(LnFrais\ réels) + \frac{Var(LnFrais\ réels)}{2}\right\}$$

Nous effectuons ce calcul car E [ln(.)] est différent de ln [E(.)]. Si Y suit une loi Log Normale, ln(Y) suit une loi normale et :

$$E(Y) = \exp\left[E\left(\ln(Y) + \frac{Var[\ln(Y)]}{2}\right]\right. = \exp\left[E(\ln(Y)] \times \left(1 + \frac{Var(Y)}{E(Y)^2}\right)\right.$$

Nous devrons donc pour chaque garantie calculer la variance de LnFrais réels.

Pour la fréquence, la modélisation par MLG est retranscrite sous la forme suivante :

$$ln\left[E\left(\frac{Nombre\ d'actes}{Exposition}\right)\right] = \alpha_0 + \alpha_1 \times x_1 + \cdots + \alpha_p \times x_p$$

d'où :

$$E\left(\frac{Nombre\ d'actes}{Exposition}\right) = \exp(\alpha_0) \times \prod_{j=1}^{p} \exp(\alpha_j)$$

Avec : $\begin{cases} Nombre\ d'actes\ suivant\ une\ loi\ Binomiale\ Négative \\ \alpha_j\ les\ coefficients\ estimés \\ \alpha_0\ le\ coefficient\ correspondant\ \text{à}\ l'individude\ référence \end{cases}$

Notre modèle a déjà été dans un premier temps validé par la PROC GENMOD de notre programme sur SAS. Nous procédons maintenant à l'analyse des résidus. Ce travail permet à la fois de renforcer notre

confiance portée sur le modèle mais aussi de repérer des valeurs atypiques et ainsi identifier la raison de certains écarts entre les tarifs trouvés et la consommation réelle.

Nous illustrons ici les deux graphes des résidus de déviance obtenus pour la garantie « Optique ».

Optique

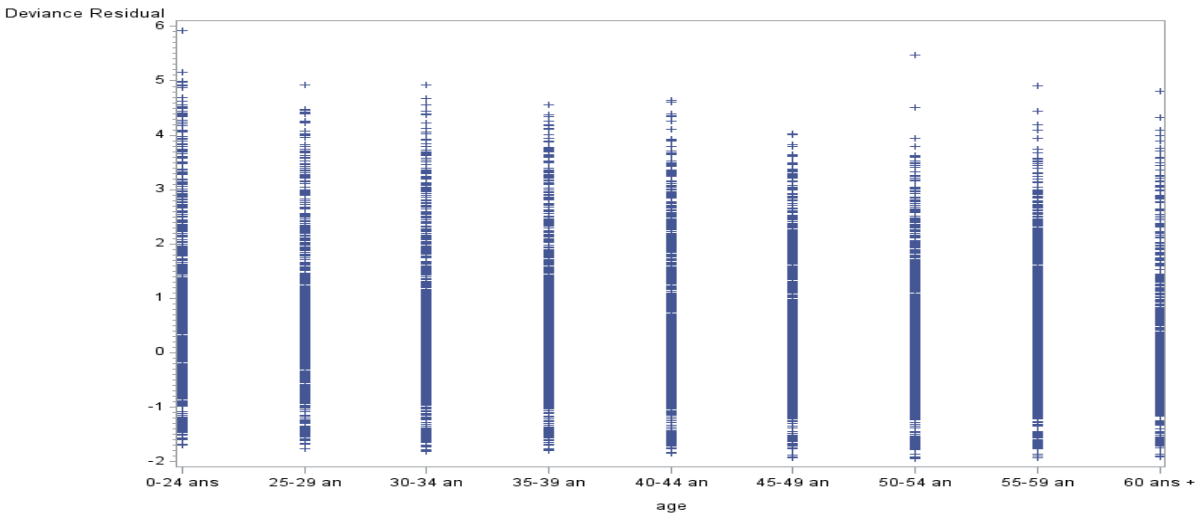


Figure 19. Résidus de déviance sur les logarithmes des Frais réels

Nous avons choisi de tracer ce graphique en fonction uniquement des classes d'âge pour le rendre plus lisible. Un graphique avec trop de variables et donc plus de modalités étant un peu moins interprétable. Notre modélisation des frais réels semble bonne car les résidus sont bien centrés autour de la valeur zéro et presque aucune valeur atypique n'est détectée.
Cette analyse a été satisfaisante pour l'ensemble des garanties.

Optique

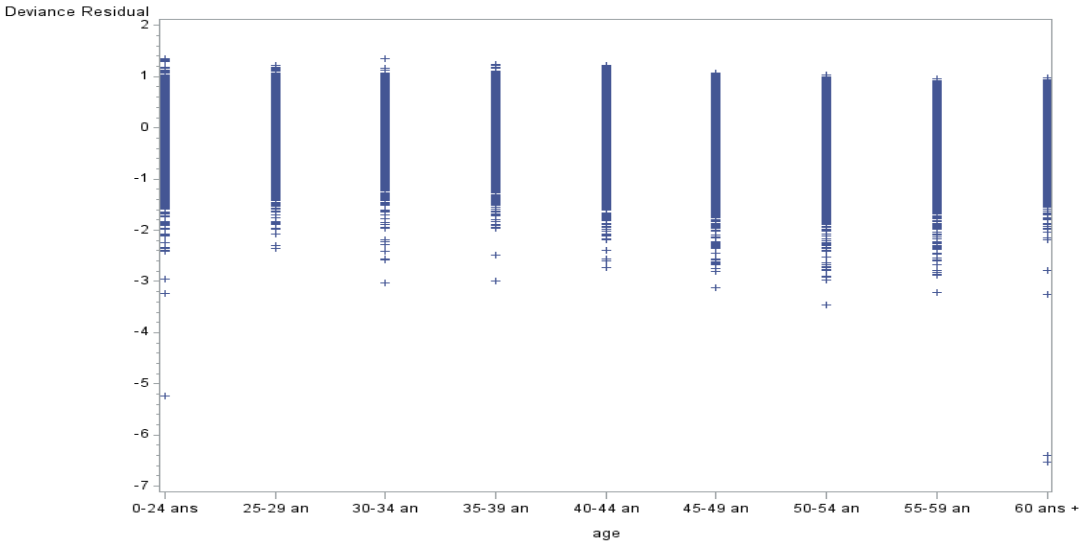


Figure 20. Résidus de déviance sur les fréquences de consommation

Au niveau de la fréquence, la cohérence de notre modèle semble être remise en cause car les résidus ne sont pas centrés sur zéro et s'en écarte relativement vers le haut. L'analyse des résidus de la déviance pour la majorité des garanties étudiées laisse supposer des erreurs de calibrage du modèle.

g) Traitement des résultats

Notre programme, nous l'avons expliqué plus haut, permet de ne garder que les variables significatives dans le modèle. Cependant certaines modalités de ces mêmes variables retenues peuvent n'apporter aucune diminution réelle de la déviance. Cela veut donc dire que ces modalités ne sont pas significatives. Nous aurions pu alors pour chaque variable, effectuer un regroupement de ces modalités et relancer la PROC GENMOD. Nous avons préféré imposer un coefficient égal à 1 pour ces modalités, traduisant ainsi le fait qu'elles ne sont pas significatives et n'impactent pas la consommation par rapport à celle de l'individu de référence.

h) Coefficients finaux obtenus

Dans cette partie, nous ne présentons les résultats que pour la garantie « Optique » et les modalités ayant été jugées non significatives n'apparaissent pas, tout comme les variables non significatives. De plus, afin de ne pas surcharger la lecture du mémoire, nous n'affichons pas la grille des coefficients pour les différentes modalités des interactions. Nous laissons le lecteur se référer à l'annexe-IV.

Optique

Bornes des coefficients	individu de référence
Borne inférieure	4,79
Coefficient obtenu	4,84
Borne supérieure	4,89

Tableau 18. Coefficients obtenus par MLG pour les frais réels de l'individu de référence

Bornes	Sexe		Secteur d'activité		Statut professionnel			Zonier				Classe d'effectif				Régime		
	Homme	Femme	CCN PATISSERIE	CCN BOULANGERIE	Non Cadre	EVIN	Ensemble de Personnel	Région Parisienne	Proche Paris - Lyon	Grand Sud	DOM	Entre 0 et 4	Moins de 20	Moins de 100	Plus de 100	REGIME GENERAL	REGIME ALS MOSELLE	REGIME TNS
Borne inférieure	0,96	1,00	0,96	1,00	1,00	0,82	0,87	1,04	1,00	0,96	1,21	1,00	1,01	1,01	1,00	0,96	1,00	0,91
Coefficient obtenu	0,96	1,00	0,97	1,00	1,00	0,88	0,90	1,05	1,00	0,97	1,33	1,01	1,01	1,03	1,00	0,97	1,00	0,94
Borne supérieure	0,97	1,00	0,97	1,00	1,00	0,94	0,94	1,06	1,00	0,98	1,46	1,02	1,02	1,04	1,00	0,97	1,00	0,96

Tableau 19. Grille des coefficients obtenus par MLG pour les frais réels

Le MLG réalisé sur le logarithme des frais réels pour la garantie « Optique » utilise comme individu de référence l'assuré ayant les caractéristiques qui suivent :

- *de sexe féminin*
- *travaillant dans le secteur de la boulangerie*
- *non cadre*
- *vivant dans le zonier Proche Paris - Lyon*
- *travaillant dans une entreprise de plus de 100 salariés*
- *bénéficiant du régime d'Alsace Moselle*
- *ayant moins de 24 ans et couvert par une base haute et une surcomplémentaire basse*

D'après les résultats que nous obtenons, nous interprétons les résultats ainsi :

- Les hommes dépensent moins que les femmes.

- Le secteur pâtissier est moins consommateur que les secteurs boulanger et AFFLEC (AFFLEC étant une modalité non significative).

- Les assurés « ensemble du personnel » et « cadres » (modalité non significative) engagent moins de frais réels. Ce résultat est assez étonnant car il suppose que les non cadres consomment plus que les cadres. Même constat d'ailleurs pour les assuré sous contrats « EVIN » qui sont supposés consommer plus, notamment en raison de leur âge plus avancé.

- Les DOM et la région parisienne engagent plus de dépenses que le reste. À l'inverse, les assurés résidant dans le Grand Sud et dans la zone Autres (modalité non significative) consomment moins.

- Les assurés employés de grosses entreprises - plus de 100 salariés - consomment moins que les autres.

- Les assurés étant sous le régime d'Alsace Moselle consomment plus fortement. Ce qui paraît normal car ce régime propose des taux de remboursement plus avantageux.

- Les frais réels engagés, de manière générale, augmentent avec l'âge, avec le niveau de la base et avec le niveau de la surcomplémentaire. Il est possible d'avoir des frais réels plus élevés pour une base plus faible mais avec un niveau de surcomplémentaire plus élevé.

Les coefficients affichés dans nos grilles ont déjà été remis à l'exponentiel. Ainsi, pour calculer les frais réels d'une personne vérifiant les premières modalités de chaque variable présente dans le tableau ci-dessus (et d'âge inférieur à 24 ans, avec une base haute et une surcomplémentaire basse) nous appliquons la méthode suivante :

$$E(LnFrais\ réels) = 4{,}84 \times 0{,}96 \times 0{,}97 \times 1 \times 1{,}05 \times 1{,}01 \times 0{,}97 \times 1 = 4{,}64$$

Et : $$Var(LnFrais\ réels) = 0{,}27$$

⇨ Frais réels $= \exp\left(4{,}64 + \frac{0{,}27}{2}\right) = 118{,}51\ €$

L'individu de référence, quant à lui, dépense un montant égal à exp (4,97), soit 144,03 euros.
L'assuré de notre exemple, lors d'une visite dans un magasin d'optique, dépense donc en moyenne 1,2 fois moins que l'assuré de référence.

Grille des coefficients pour la fréquence en Optique

Bornes des coefficients	individu de référence
Borne inférieure	0,29
Coefficient obtenu	0,38
Borne supérieure	0,49

Tableau 20. Coefficients obtenus par MLG pour la fréquence de consommation de l'individu de référence

Bornes	Situation familiale		Secteur d'activité			Statut professionnel				Niveau de Surcomplémentaire				Classe d'effectif				Régime	
	Seul	En couple	CCN PATISSERIE	CCN BOULANGERIE	CCN AFFLEC	Non Cadre	EVIN	Ensemble de Personnel	TNS	Nulle	Basse	Moyenne	Haute	Entre 0 et 4	Entre 5 et 19	Entre 20 et 100	Plus de 100	REGIME ALS MOSELLE	REGIME TNS
Borne inférieure	0,89	1,00	0,71	0,10	1,00	0,23	0,51	1,00	0,57	2,34	1,00	0,98	1,16	0,36	0,36	0,36	1,00	1,00	0,73
Coefficient obtenu	0,93	1,00	0,74	0,10	1,00	0,33	0,66	1,00	0,76	2,47	1,00	1,00	1,29	0,37	0,38	0,38	1,00	1,00	0,85
Borne supérieure	0,96	1,00	0,77	0,10	1,00	0,48	0,85	1,00	1,00	2,60	1,00	1,07	1,43	0,39	0,40	0,41	1,00	1,00	1,00

	Niveau de la base		
Bornes des coefficients	Basse	Moyenne	Haute
Borne inférieure	1	12,1848624	91,3532067
Coefficient obtenu	1	12,5647303	95,1567226
Borne supérieure	1	12,9564406	99,118599

Tableaux 21 et 22. Grille des coefficients obtenus par MLG pour la fréquence de consommation

Le MLG réalisé sur la fréquence pour la garantie « Optique » utilise comme individu de référence l'assuré ayant les caractéristiques qui suivent :

- *vivant en couple*
- *travaillant dans le secteur des fruits et légumes*
- *appartenant à la catégorie « Ensemble de personnel »*
- *ayant souscrit à une surcomplémentaire basse*
- *travaillant dans une entreprise de plus de 100 salariés*
- *bénéficiant du régime d'Alsace Moselle*
- *ayant une base de faible niveau*
- *ayant moins de 24 ans et étant de sexe féminin*

D’après les résultats que nous obtenons, nous interprétons les résultats ainsi :

- Les personnes vivant seules vont moins fréquemment chez l’opticien.

- Le secteur d’activité concerné par l’AFFLEC est plus consommateur que les autres, en termes de fréquence de consommation. Le coefficient trouvé pour le secteur boulanger est surprenant car il est excessivement bas. Cette mauvaise estimation est surement due à l’interaction entre la base et le secteur d’activité qui n’a pas été spécifié.

- Les assurés appartenant à la catégorie « Ensemble de personnel » ont une fréquence de consommation plus élevée. Cela compense alors les résultats étonnants que nous avions pour les frais réels.

- Plus le niveau de la surcomplémentaire augmente plus la fréquence de consommation augmente. Cependant nous voyons que cette fréquence est la plus élevée pour les assurés n’ayant aucune surcomplémentaire. Ce résultat ne semble pas cohérent. Néanmoins, la valeur élevée - compensant celle obtenue pour le secteur boulanger - vient confirmer l’idée que cela est dû à l’interaction forte entre les deux variables citée plus haut. En effet, le contrat de base de la CCN Boulangerie correspond à un niveau élevé.

- Les assurés employés de grosses entreprises - plus de 100 salariés - consomment plus souvent que les autres.

- Les assurés étant sous le régime d’Alsace Moselle et sous le régime général (modalité non significative) consomment plus souvent.

- Quel que soit l’âge, les hommes ont moins souvent recours à l’opticien. Au niveau des âges, la fréquence de consommation augmente avec l’âge à partir de 40 ans. Avant cet âge, elle reste assez équilibrée.

Pour calculer la fréquence de consommation d’une personne vérifiant les premières modalités de chaque variable présente dans le tableau ci-dessus (et de sexe féminin ayant moins de 24 ans) nous appliquons la méthode suivante :

$$Fréquence = 0{,}38 \times 0{,}93 \times 0{,}74 \times 0{,}33 \times 2{,}47 \times 0{,}37 \times 1 \times 1 \times 1 = 1{,}08$$

Pour l’individu de référence, le nombre d’actes est de 0,38 fois, soit presque trois fois moins que l’assuré étudié.

Les coefficients obtenus par MLG sont particulièrement satisfaisant au niveau des frais réels. Nous avons notamment des résultats cohérents en ce qui concerne l’impact sur la consommation du montage Base - Surcomplémentaire. Pour chacune des garanties étudiées, nous remarquons un impact à la hausse lorsque la base augmente à niveau de surcomplémentaire identique. Et idem pour les surcomplémentaires à niveau de base identique. Les résultats confirment ce que nous voulions montrer : deux surcomplémentaires de même niveau ne doivent pas nécessairement avoir le même tarif si la base associée est différente.
Pour les fréquences, les coefficients obtenus restent intéressants lorsque nous regardons l’influence du niveau d’une surcomplémentaire souscrite. Cependant, les assurés n’ayant souscrit à aucune surcomplémentaire ont une fréquence anormalement plus élevée. Ce problème peut être dû à la non

spécification de l'interaction entre la base et le secteur d'activité. En effet, la limitation de notre périmètre d'étude - à seulement trois CCN - a entrainé le fait que chaque base soit totalement associée à un secteur d'activité. Le nombre de contrat de base étant très élevé, ceci peut fortement influer sur la distribution du taux de recours aux garanties. Le graphique ci-dessous montre la distribution du nombre d'actes optiques lorsque l'on se restreint aux contrats proposant un niveau base haut.

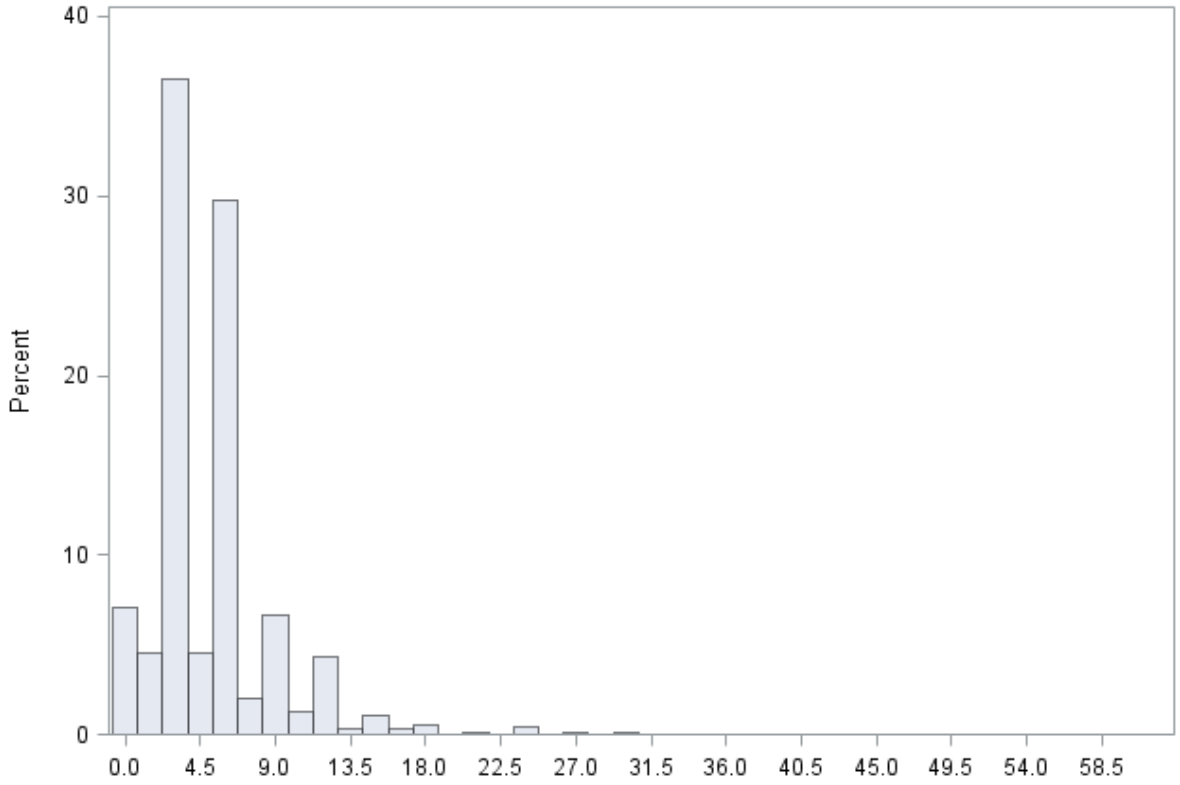


Figure 21. Histogramme de la fréquence pour la garantie « Optique » en base haute

On observe des résultats étonnants avec une masse particulièrement importante pour un nombre d'actes compris entre deux et neuf. Cette distribution semble s'éloigner de la réalité dans la mesure où d'une part cette garantie correspond uniquement aux verres et aux montures, et d'autre part car en moyenne une personne renouvelle ses lunettes seulement une fois tous les deux ans. Obtenir une probabilité avoisinant les 10% pour un nombre d'actes égal à neuf est ainsi très surprenant. Au vu de ces résultats, nous pouvons en déduire que notre base de données ne représente pas tout à fait l'ensemble de la population et reste très spécifique. C'est pour cette raison que les résultats de modélisation de la fréquence semblent moins corrects.
Seules les garanties « Pharmacie » et « Prothèses dentaires » ne sont pas sujettes à ce problème et obtiennent des coefficients tout à fait cohérents avec ce que nous pensions.

i) Calcul de la prime pure

Nous avons jusqu'à présent calculé les frais réels engagés ainsi que le taux de recours à un acte en santé. Le but étant de faire une tarification, nous devons procéder au calcul de la prime pure. C'est-à-dire établir le passage de la consommation réelle moyenne des assurés au remboursement moyen que devra verser l'assureur. Ce calcul va avant tout dépendre du type de remboursement proposé par l'organisme assureur et bien entendu, du niveau de couverture proposé. Nous décrivons dans les prochains paragraphes la méthode de calcul pour une base. La méthode de calcul pour une surcomplémentaire sera expliquée ensuite.

1. Calcul de la base

1.1. Remboursement sur les frais réels

Dans ce type de remboursement, l'assureur accepte de prendre à sa charge la totalité du montant engagé par l'assuré. La valeur du remboursement d'un acte est donc facile à retrouver puisqu'il correspond exactement au montant de consommation réel que nous avons obtenu par MLG. Auquel on enlève la participation forfaitaire (s'il y en a une et que l'on nommera Pf) et le remboursement de la Sécurité Sociale. Nous avons ainsi, avec Rbt le remboursement de l'assureur :

$$Rbt = E(frais\ réels) - RSS - Pf$$

⇨ $$\text{Prime pure} = \text{Fréquence} \times \text{Rbt}$$

1.2. Remboursement sur un montant fixé à priori

Dans ce cas de remboursement, l'assureur verse au maximum un montant égal à celui fixé dans la grille de garantie proposée à l'assuré. Nous trouvons donc dans cette section plusieurs type de remboursement. Nous avons principalement le remboursement sur la BR, avec un taux de remboursement fixé ; Et le remboursement par montant forfaitaire dit forfait par acte. Quel que soit la consommation de l'assuré, l'assureur ne versera pas plus que ce montant. Si l'assuré consomme moins que le montant proposé, alors la somme à verser par l'assureur est celle engagé par le consommateur. Il s'agit donc d'un plafond pour l'assuré au-delà duquel il engagera ses propres revenus. En prenant comme exemple le cas d'un remboursement sur la BR (en complément du RSS), pour chaque acte le remboursement de l'assureur est donc égal à :

$$Rbt = \min\left(E(frais\ réels) - RSS - Pf, K\%\ de\ la\ BR\right)$$

⇨ $$\text{Prime pure} = \text{Fréquence} \times \text{Rbt}$$

Le calcul est identique s'il s'agit d'un forfait par acte ou d'un autre remboursement du même type.
Théoriquement, il serait plus juste de modéliser E [min (Frais réels - RSS - Pf, K% de la BR)]. Cependant cela demanderait beaucoup plus de traitement car le MLG devrait alors être relancé pour chaque taux de remboursement proposé. Nous pourrions aussi, de manière à être certains de couvrir les engagements futurs de l'assuré, établir une prime égale à :

$$Rbt = Max(0, E(Frais\ réels) - RSS - Pf)$$

⇨ $$\text{Prime pure} = \text{Fréquence} \times \text{Rbt}$$

Avec ce calcul, la prime pure est sur estimée et plus prudente. Nous nous basons sur le fait que l'assuré consommera presque toujours la totalité du montant remboursé. Mais elle reflète moins bien la consommation réelle des assurés et pourrait donc faire perdre des parts de marché à l'assureur s'il propose un tarif trop élevé. Aussi, nous préférons déterminer la prime pure par la première méthode énoncée.

1.3. Remboursement sur un forfait par période

Le forfait par période impose une double contrainte. Il impose un plafond de remboursement comme dans le cas du forfait par acte. Mais ce plafond est étalé sur une période. Pour un forfait par période, l'assureur s'engage à rembourser totalement l'assuré à chaque acte de consommation tant que le montant cumulé engagé sur la période ne dépasse pas le montant fixé.
Prenons l'exemple d'un tel forfait par période pour la garantie « Optique » : 800 euros par année. L'assuré se rend chez l'opticien à quatre reprises où il consomme pour 200 euros les trois premières fois et 300 la

dernière fois. L'assuré n'aura aucun reste à charge pour les trois premiers actes car le montant total engagé de 600 euros est inférieur au forfait de 800 euros. Le dernier acte, par contre, entraine un dépassement du forfait car les montants cumulés sont alors de 900 euros. L'assureur devra donc pour cette année verser 800 euros à l'assuré, qui devra de son côté prendre à sa charge les 100 euros restants. Le calcul de la prime pure se fait dans ce cas-ci de la manière suivante :

$$Rbt = \min[E(Frais\ réels) - RSS - Pf, Forfait]$$
$$Car\ l'assuré\ peut\ dépasser\ le\ forfait\ en\ un\ seul\ acte\ de\ consommation.$$

⇨ $$\text{Prime pure} = \min(\text{Forfait}, \text{Fréquence moyenne} \times \text{Rbt})$$

En pratique, et en particulier avec les nouveaux contrats dans le cadre de l'ANI, ce type de forfait sera étalé sur deux années. Soit des forfaits bisannuels. Comme nous avons fixé une hypothèse de proportionnalité entre la fréquence de consommation et la durée de couverture, nous avons donc :

$$Fréquence\ bisannuelle = 2 \times Fréquence$$

⇨ $$\text{Prime pure} = \min(\text{Forfait}, 2 \times \text{Fréquence} \times \text{Rbt})$$

1.4. Remboursements sur un montant fixé à priori et en nombre limité

Nous rencontrons de plus en plus ce type de remboursement qui est dans la continuité du principe de responsabilisation de l'assuré. Le nombre d'acte de consommation est limité sur une période donnée. Si l'assureur propose un nombre d'actes de consommation maximum égal à N, alors l'assuré devra prendre en charge la totalité des dépenses à partir du $(N+1)^{ème}$ acte. Soit K le taux de remboursement sur la BR et N le nombre d'actes maximum pour lequel l'assureur s'engage à rembourser. La prime pure se calcule ainsi :

$$Rbt = \min(E(frais\ réels) - RSS - Pf, K\%\ de\ la\ BR)$$

$$Taux\ de\ recours = \min(Fréquence, N)$$

⇨ $$\text{Prime pure} = \text{Taux de recours} \times \text{Rbt}$$

Nous avons choisi un remboursement par montant fixé à priori. Mais le calcul reste identique dans le cadre d'un forfait par acte ou d'un forfait par période.

2. Calcul de la surcomplémentaire

Le calcul de la surcomplémentaire s'effectue dans la même logique que celui de la base. Le but est de calculer le montant moyen versé par l'assureur par acte de consommation. Ensuite de multiplier ce montant moyen par la fréquence moyenne de consommation. Le calcul de la fréquence est évidemment identique à celui présenté plus haut. Seul le calcul du montant diffère. Cette différence est due au mécanisme d'une surcomplémentaire. Lorsqu'un assuré souscrit à un tel produit pour une garantie, cela se traduit par le fait qu'il rajoute une couverture supplémentaire. Pourtant, cette couverture supplémentaire ne sera pas nécessairement toujours mise en jeu. Prenons l'exemple d'un montage tel quel :

- consultations généralistes => Base proposant 100% de la BR
- consultations généralistes => Surcomplémentaire proposant 50% de la BR

Cela veut dire que l'assuré pourra consommer à hauteur de 150% de la BR, soit un montant de 34,5 euros, sans rien avoir à sa charge (hormis la participation forfaitaire). Si cet assuré se rend en consultation chez

un généraliste tarifant 23 euros sa consultation, soit la BR, l'assuré sera donc remboursé 100% de la BR. Seule la base suffit au remboursement de cet assuré. Ainsi, bien qu'il dispose d'une surcomplémentaire, celle-ci n'est pas intervenu dans son remboursement. Ceci doit donc être pris en compte par l'assureur. Aussi, dans notre approche, nous calculons le remboursement moyen par acte pour le montage base et surcomplémentaire - de la même manière que pour le calcul de la base expliqué plus haut - auquel nous déduisons celui de la base uniquement. Cela se traduit sous la forme suivante :

$$Rbt\ (Surcomplémentaire) = Rbt(Base + Surcomplémentaire) - Rbt(Base)$$

Le reste du calcul aboutissant à la prime pure est ensuite identique.

j) Validation par backtesting

Afin de vérifier l'adéquation de nos primes pures avec la consommation réelle des assurés, nous avons choisi trois contrats différents pour lesquelles nous calculons l'ensemble des prestations versées chaque année par AG2R La Mondiale, en fonction de la garantie et du produit souscrit. C'est donc une validation en utilisant l'historique réel de consommation du groupe. La première étape consiste donc à récupérer pour chaque contrat, les informations nécessaires au calcul de la prime pure. C'est-à-dire qu'il faudra trouver les informations nous permettant de placer les assurés dans une des classes tarifaires propres à chaque garantie. Une fois ces informations récupérées, la fréquence de consommation ainsi que les montants de frais réels peuvent être déterminés. La seconde étape consiste alors à passer de la consommation réelle au remboursement effectué par le groupe assureur. Nous devons ainsi récupérer d'un côté les remboursements proposés par AG2R La Mondiale pour chaque produit et pour chaque garantie. Et de l'autre côté les remboursements de la SS pour ces mêmes garanties. Cette prime pure individuelle est ensuite multipliée par le nombre total de bénéficiaires. Il nous reste alors à calculer le rapport entre le total des primes pures acquises par le groupe et le montant total des prestations versées, et ce pour chaque garantie. La valeur trouvée est appelée le P/C. Le P/C global - c'est-à-dire sur l'ensemble des garanties - ne nous intéresse pas car le but est d'avoir une tarification juste sur chaque garantie. De plus cela n'aurait pas de sens réel car chaque contrat couvrant différentes garanties, le P/C pourrait très bien être bon (supérieur à 1) dans certains cas et être totalement mauvais (inférieur à 1) dans d'autres, en fonction des garanties couvertes. Nous gardons ici la garantie « Prothèses dentaires » pour illustrer les étapes du calcul. Cette garantie présente l'avantage d'être souvent concernée par l'augmentation de couverture en surcomplémentaire, contrairement à d'autres garanties comme « Radiologie » ou « Pharmacie ».

Les classes tarifaires pour les frais réels retenues par MLG sont les suivantes :

- le croisement sexe et âge
- le niveau de surcomplémentaire
- le secteur d'activité
- le statut professionnel
- le zonier
- l'effectif
- le régime

Celles retenues pour la fréquence sont :

- le croisement Âge, niveau de base et niveau de surcomplémentaire
- le secteur d'activité
- le statut professionnel
- la situation familiale
- l'effectif

Nous avons donc au total 10 informations à récupérer.
Nous choisissons en premier un contrat A d'une entreprise appartenant à la CCN BOUCHERIE. Cette entreprise propose uniquement un contrat de base. Les informations concernant l'entreprise sont présentées dans les tableaux ci-après :

	Base	**Surcomplémentaire**	**Secteur d'activité**	**Statut professionnel**	**Situation familiale**	**Zonier**	**Régime**
Base	Moyenne	Nulle	Boucherie	Ensemble de personnel	Seul	Grand Sud	Général

Tableau 23. Caractéristiques du contrat A

	Age	**Hommes**	**Femmes**	**Nombre de bénéficiaires adulte**	**Nombre de bénéficiaires enfant**
2013	30	9	6	15	0
2012	32	9	6	15	0
2011	32	7	3	9	0

Tableau 24. Démographie de l'entreprise par année

Nous avons toutes les informations nécessaires pour obtenir la consommation réelle d'un assuré de cette entreprise. Nous présentons les résultats obtenus dans les tableaux ci-dessous. Le calcul de la fréquence et des frais réels n'est pas explicitée ici car il a déjà été expliqué plus haut dans ce mémoire.

Prothèses dentaires

Garantie	Type et Niveau de remboursement	Base de remboursement	Taux de remboursement Régime Général	Montant de remboursement SS	Ticket modérateur	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la base	Frais réels	Fréquence	Reste à rembourser	Prime pure
Prothèses dentaires	450% de la Base de Remboursement	107,50	0,70	75,25	32,25	3,80	408,50	Moyenne	498,52	0,34	408,50	138,83

Tableau 25. Prime pure homme de la garantie « Prothèses dentaires » sur la base

Garantie	Type et Niveau de remboursement	Base de remboursement	Taux de remboursement Régime Général	Montant de remboursement SS	Ticket modérateur	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la base	Frais réels	Fréquence	Reste à rembourser	Prime pure
Prothèses dentaires	450% de la Base de Remboursement	107,50	0,70	75,25	32,25	3,80	408,50	Moyenne	519,18	0,34	408,50	138,83

Tableau 26. Prime pure femme de la garantie « Prothèses dentaires » sur la base

Les tarifs Homme et Femme sont identiques. Ce qui semble assez cohérent car rien ne justifierait réellement une consommation différente au niveau de ce type de soins en fonction du sexe.

Nous récupérons maintenant l'ensemble des prestations versées pour chaque année. Et nous calculons le montant total de primes acquises par l'assureur en multipliant la prime pure par le nombre de bénéficiaires.

Résultats de validations

Année	Nombre de bénéficiaires	Pourcentage d'hommes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Prime Individuelle	Somme des primes	Somme des prestations	P/C
2011	10,00	0,70	32,00	138,83	138,83	138,83	1388,29		
Année	Nombre de bénéficiaires	Pourcentage d'hommes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Prime Individuelle	Somme des primes	Somme des prestations	P/C
2012	15,00	0,60	32,00	138,83	138,83	138,83	2082,44	382,25	545%
Année	Nombre de bénéficiaires	Pourcentage d'hommes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Prime Individuelle	Somme des primes	Somme des prestations	P/C
2013	15,00	0,60	30,00	138,83	138,83	138,83	2082,44	659,50	316%
Sur les 3 années									
	Somme des primes	Somme des prestations	P/C						
	5553,18	1041,75	533%						

Tableau 27. Ratios P/C pour la base

L'écart entre les primes reçues et les montants réellement versés est trop important. Notre tarif de la base pour cette garantie est trop élevé et suggère que notre modélisation ne peut être retenue.

Nous avons travaillé uniquement sur la base. Nous choisissons cette fois ci un deuxième contrat B proposant un contrat de base de niveau moyen pour la garantie « Prothèses dentaires » et deux options à niveaux bas et moyen pour cette même garantie. Nous calculons le tarif Homme ainsi que le tarif Femme des deux options.

	Base	Surcomplémentaire	Secteur d'activité	Statut professionnel	Situation familiale	Zonier	Régime
Base	Moyenne	Nulle	Boucherie	Ensemble de personnel	Seul	Grand Sud	Général
Option 1	Moyenne	Basse	Boucherie	Ensemble de personnel	Seul	Grand Sud	Général
Option 2	Moyenne	Moyenne	Boucherie	Ensemble de personnel	Seul	Grand Sud	Général

Tableau 28. Caractéristiques du contrat A

Option 1	Age	Hommes	Femmes	Nombre de bénéficiaires adulte	Nombre de bénéficiaires enfant
2013	40,2	2761	3230	5991	0
2012	40	2744	3247	5991	0
2011	40	2680	3209	5889	0

Tableau 29. Répartition par année des bénéficiaires de l'option 1

Option 2	Age	Hommes	Femmes	Nombre de bénéficiaires adulte	Nombre de bénéficiaires enfant
2013	40,2	4544	5315	9859	0
2012	40	4515	5343	9858	0
2011	40	4410	5280	9690	0

Tableau 30. Répartition par année des bénéficiaires de l'option 2

Prothèses dentaires

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la base	Frais réels	Fréquence	Reste à rembourser (Montage)	Reste à rembourser	Prime pure
Prothèses dentaires	75% de la Base de Remboursement	75,25	4,55	489,125	Basse	503,54	0,32	428,29	19,79	6,39

Tableau 31. Prime pure homme de la garantie « Prothèses dentaires » sur l'option 1

La prime pure Homme pour une couverture complémentaire en prothèses dentaires est donc égale à 6,39 euros.

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la surcomplémentaire	Frais réels	Fréquence	Reste à rembourser (Montage)	Reste à rembourser	Prime pure
Prothèses dentaires	75% de la Base de Remboursement	75,25	4,55	489,125	Basse	524,544	0,32	449,19	40,69	13,13

Tableau 32. Prime pure femme de la garantie « Prothèses dentaires » sur l'option 1

La prime pure Femme pour une couverture complémentaire en prothèses dentaires est donc égale à 13,13 euros.

De la même manière que précédemment, nous calculons le rapport P/C.

Résultats de validation

Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2011	2761,07	3229,76	40,00	6,39	13,13	60060,03	61710,93	97%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2012	2743,69	3246,77	40,00	6,39	13,13	60172,37	56485,63	107%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2013	2679,81	3208,59	40,20	6,39	13,13	59262,91	50796,20	117%
Sur les 3 années	Somme des primes	Somme des prestations	P/C					
	179495,31	168992,76	106%					

Tableau 33. Ratios P/C pour l'option 1

Nous obtenons des rapports très intéressants. Nous sur tarifons un peu la garantie « Prothèses dentaires » mais l'écart peu important nous rassure quant à notre modélisation.
Nous tarifons aussi l'option 2, et présentons les résultats obtenus dans les tableaux qui suivent.

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la surcomplémentaire	Frais réels	Fréquence	Reste à rembourser (Montage)	Reste à rembourser	Prime pure
Prothèses dentaires	125% de la Base de Remboursement	75,25	5,05	542,875	Moyenne	534,90	0,43	459,65	51,15	21,96

Tableau 34. Prime pure homme de la garantie « Prothèses dentaires » sur l'option 2

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la surcomplémentaire	Frais réels	Fréquence	Reste à rembourser (Montage)	Reste à rembourser	Prime pure
Prothèses dentaires	125% de la Base de Remboursement	75,25	5,05	542,875	Moyenne	557,33	0,43	482,08	73,58	31,59

Tableau 35. Prime pure femme de la garantie « Prothèses dentaires » sur l'option 2

Résultats de validation

Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2011	2761,07	5315,24	40,00	21,96	31,59	228540,22	272144,28	84%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2012	2743,69	5343,23	40,00	21,96	31,59	229042,62	220988,65	104%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2013	2679,81	5280,41	40,20	21,96	31,59	225655,22	191956,92	118%
Sur les 3 années	Somme des primes	Somme des prestations	P/C					
	683238,06	685089,85	99,7%					

Tableau 36. Ratios P/C pour l'option 2

Les tarifs que nous avons obtenus sont très satisfaisants. Ils recouvrent parfaitement l'année 2012 et 2013. Seule l'année 2011 est en déficit. Sur les trois ans, nous avons un rapport tout proche de 100% qui conforte notre méthode de modélisation.

Au vu de l'ensemble des résultats, nous pouvons considérer que nos tarifs sont justes lorsqu'il s'agit des surcomplémentaires. En ce qui concerne la base, le tarif est bien trop élevé pour pouvoir être proposé. Dans la section suivante, nous illustrons comment remédier à ce problème.

k) Intérêt des intervalles de confiance

Une caractéristique non négligeable des MLG est qu'ils permettent, en plus de modéliser une variable réponse, d'obtenir des intervalles de confiance. Pour chaque coefficient, et donc pour chaque modalité retenue par le MLG, un intervalle de confiance est obtenue en sortie. Ces intervalles sont de grande importance car ils fournissent à l'assureur une estimation du risque pris. En effet, même si le MLG converge vers une solution, l'actuaire garde en considération que son tarif ne peut pas parfaitement coller à la réalité. Il pourra ainsi juger de l'écart possible entre sa tarification et la consommation réelle. Ce sont donc de bons indicateurs de risque qui assurent une marge de pilotage par rapport au tarif établi. SAS fournit par défaut des intervalles de confiance avec un niveau de confiance à 95%. Dans les résultats qui vont être présentés, nous avons modifié le calcul de la fréquence de consommation car les montants par actes nous semblaient réalistes. La prime trouvée plus haut ayant été jugée trop élevée, nous avons cette fois opté pour un calcul basé sur les bornes inférieures des coefficients de la fréquence de consommation.

Prothèses dentaires

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la base	Frais réels	Fréquence	Reste à rembourser	Prime pure
Prothèses dentaires	450% de la Base de Remboursement	75,25	3,80	408,5	Moyenne	498,52	0,096	408,50	39,31

Tableau 37. Prime pure homme de la garantie « Prothèses dentaires » sur la base

Logiquement, la prime pure est plus basse que celle obtenue précédemment. Le tarif femme est identique.

Résultats de validation

Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2011	7	3	32,00	39,31	39,31	393,13		
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2012	9	6	32,00	39,31	39,31	589,70	382,25	154%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2013	9	6	30,00	39,31	39,31	589,70	659,50	89%
Sur les 3 années	Somme des primes	Somme des prestations	P/C					
	1572,52	1041,75	151%					

Tableau 38. Ratios P/C pour la base

La différence est flagrante. Alors que nous avions des écarts considérables mettant en cause notre modélisation, nous observons maintenant des tarifs semblant totalement pertinents. Nous recouvrons pleinement les engagements pour l'année 2012 et sommes en léger déficit pour l'année 2013. Nous voyons donc dans cet exemple toute l'importance des intervalles de confiance. L'actuaire, après obtention des coefficients par MLG, pourra jouer sur les intervalles en fonction de son expérience, de l'historique de consommation du groupe et du contexte commercial pour choisir le tarif adéquat.

l) Bilan des résultats

Pour chacune des garanties - hormis la chambre particulière et les frais de séjours - nous avons finalement obtenu une surestimation de la prime pure, notamment pour les bases. Celle-ci semble être entrainée par une mauvaise modélisation de la fréquence. Problème ressenti d'ailleurs par l'obtention de fréquence plus élevée chez les assurés n'ayant aucune surcomplémentaire que sur ceux en ayant une. En ce qui concerne les options, tout comme pour la garantie « Prothèses dentaires », les primes ont été légèrement sur estimées mais la modélisation a semblé plus acceptable. Pour les garanties « Chambre particulière » et « Frais de séjour », la prime pure est sous-estimée et semble être une conséquence du peu de données dont nous disposons pour ces actes.
Au niveau des primes pures Enfant, la modélisation semble être bien ajustée pour les frais réels de chacune des garanties. Idem pour les fréquences de consommation, excepté sur les garanties « Frais de séjour », « Chambre particulière » et « Soins dentaires ». Encore une fois nous rencontrons des problèmes de modélisation sur les deux premières garanties. Ceci était d'ailleurs prévisible car les enfants sont encore moins sujets que les adultes à ce type d'actes. Comme pour les adultes, la prime pure enfant résultante est sur évaluée pour l'ensemble des garanties, excepté les trois garanties énoncées plus haut.

m) Discussion des résultats

Comme nous l'avons indiqué dans la partie précédente, nous avons globalement sur tarifé nos garanties. Certes, la modélisation des frais réels semble relativement juste. Cependant, la modélisation de la fréquence fournit des fréquences trop élevées. Et plus important, elle aboutit sur des questions d'interprétations auxquelles nous n'aurions pas pensé. En effet, si nous nous basions sur les résultats obtenus, nous devrions en comprendre qu'un assuré à tendance à consommer plus souvent qu'un individu sans surcomplémentaire. Est-ce que cette interprétation est totalement incohérente ? Il est possible d'imaginer qu'un assuré disposant d'une surcomplémentaire consomme plus en une seule fois d'une garantie, de sorte à ce qu'il ait moins recours à cette même garantie par la suite. Ceci nous amène tout de même à supposer une mauvaise modélisation par notre modèle. Les raisons de cette modélisation peuvent être multiples.
Nous pouvons supposer que notre paramétrage n'est pas adéquat. Nous avons choisi une loi Binomiale Négative, associé à la fonction de lien logarithmique. Il aurait pu être plus approprié de choisir le lien canonique de la distribution Binomiale Négative. De plus, la fréquence de consommation est totalement dépendante du comportement propre à l'assuré vis-à-vis de sa santé. Ceci n'est pas aussi vrai en ce qui concerne les frais réels. Les frais réels sont avant tout fixés par les praticiens qui, au final, proposent approximativement une même tranche de tarifs. Il est donc plus facile d'estimer les montants engagés dans un acte de consommation -et donc la distribution à considérer- en fonction de la garantie et des acteurs proposant le service en santé. Or, en ce qui concerne les fréquences, nous pouvons légitiment avancer qu'il y a autant de comportements qu'il y a d'assurés. Une telle hétérogénéité est donc très difficile à modéliser par une distribution.
Toujours au niveau du paramétrage, les effets d'interactions peuvent être responsables de cette sur évaluation. Les niveaux de base et de surcomplémentaire étaient ici très corrélés avec le secteur d'activité,

le statut professionnel, et plus étonnamment avec l'effectif aussi. Nous aurions pu opter pour un modèle sans ses trois variables afin de ne cerner réellement que l'impact du montage Base - Surcomplémentaire.
La deuxième source d'erreur peut provenir d'un manque de données. Nous nous sommes restreints uniquement à trois CCN : Boulangerie, Pâtisserie et AFFLEC. Or, chaque CCN affiche une consommation bien spécifique, et il est ainsi probable que notre modélisation ne convienne pas parfaitement aux assurés de la CCN Boucherie. Une étude plus complète sur l'ensemble des CCN, ou même sur l'ensemble du portefeuille fournirait de meilleurs résultats. En particulier pour les garanties « Frais de séjour » et « Chambre particulière » qui sont des garanties dont le nombre d'actes est réduit et donc plus difficile à modéliser. Cette remarque sera d'ailleurs valable aussi lors de l'application de l'algorithme CART.
Enfin, il est probable que nous ayons sur évalué la fréquence de consommation par la prise en compte de l'exposition. Nous nous sommes basés sur un rapport de proportionnalité entre le nombre d'actes consommés et la durée pendant laquelle l'assuré est couvert par un produit. Or, ce raisonnement n'est pas tellement réaliste. D'une part, la consommation en santé répond souvent à un besoin ponctuel. Il n'est donc pas légitime de se dire qu'un assuré consommera de façon équilibrée tout au long de l'année. Prenons l'exemple suivant : un individu est assuré pour une durée d'un mois. Mois durant lequel il se rend chez un médecin généraliste. Selon notre modèle, cet assuré aurait alors été 12 fois chez un médecin généraliste sur l'année entière. Nous négligeons donc le fait que cette visite correspondait à un besoin ponctuel et que ce besoin ne reviendra pas chaque mois. Cette problématique est encore plus frappante si nous traitons de garanties de type optique ou dentaire. En effet, un assuré ayant acheté une paire de lunettes dans le mois n'ira surement plus chez l'opticien de toute l'année car il s'agit souvent d'un besoin unique satisfait lors de l'obtention des lunettes. Au final, nous pouvons affirmer que pour certaines garanties, c'est l'effet inverse qui se produit : le fait d'avoir déjà consommé nous laisse penser que l'assuré ne consommera pas plus.
De plus, nous manquons d'informations sur l'historique de consommation de l'individu avant qu'il n'adhère à la complémentaire du groupe. Nous ne savons pas s'il avait déjà consommé en santé et quels actes.
Plusieurs manières de prendre en compte l'exposition pourraient être mises en place, dépendant de l'historique de la consommation du portefeuille. L'actuaire pourrait par exemple fonctionner par semestre au lieu de fonctionner mois par mois. Cela se traduirait par le fait que si l'individu est assuré depuis moins de six mois, son nombre d'actes consommés est doublé. S'il est resté plus de six mois, son nombre d'actes reste inchangé. Ou alors, si l'assuré est resté couvert pendant une période reconnue comme une période de forte consommation, son nombre d'actes n'est pas augmenté. D'importantes statistiques doivent donc être réalisées sur la consommation d'un portefeuille de sorte à émettre des hypothèses pertinentes et indiquer la tendance de consommation d'un assuré afin de mieux prendre en compte l'exposition.
Finalement, nous pourrions même ajouter que l'exposition, en santé, n'est pas un paramètre si indispensable. En assurance Auto, l'exposition prend toute son importance, car un assuré ayant eu deux accidents en six mois, et par sa faute, peut être déterminé comme un assuré risqué et mauvais conducteur. Le nombre d'accidents associé à une période de couverture est donc un bon indicateur du comportement de l'assuré vis-à-vis de la sécurité routière. En santé, le comportement de l'assuré, nous l'avons dit, intervient aussi. Par exemple, certains assurés plus préventifs, ou plus axés médecines traditionnelles, auront certes moins recours aux praticiens. Néanmoins, le comportement n'est pas aussi déterminant qu'en assurance Auto. Comme nous l'avons expliqué précédemment, la consommation nait avant tout d'un réel besoin en santé, qui peut être dû à des variables indépendantes du comportement tel que l'âge ou le sexe.
Malgré cette modélisation jugée peu fiable au niveau des fréquences, nous avons pu tout de même obtenir une consommation en santé assez proche de la consommation réelle de notre portefeuille étudié. Et ceci par l'intermédiaire des intervalles de confiance fournis par le MLG en choisissant uniquement les bornes inférieures pour les fréquences. Lors d'une tarification par MLG, l'actuaire a donc la possibilité de réajuster ses tarifs grâce aux bornes supérieures et inférieurs. Ceci est un avantage considérable par rapport aux autres méthodes de modélisation qui ne permettent pas un tel pilotage du risque.

Pour finir cette discussion, le dernier point abordé est celui concernant la modélisation des frais réels au lieu des montants remboursés. Tout d'abord, les rapports P/C que nous trouvons sont satisfaisants, confirmant la possibilité d'appliquer notre méthode de passage de l'un à l'autre. De plus, avec l'ANI qui se profile, nous avons d'importantes évolutions règlementaires en santé, avec des garanties minimum et des types de remboursement variés. Ceux-ci ne sont pas encore fixés, si bien que les groupes assureurs ne peuvent pas encore proposer d'offres finales. Il est donc important que ces dernières puissent proposer rapidement un tarif en cas de changement de situation. Notre méthode de calcul est garante de cette spontanéité de réponse. De même, l'organisme assureur se doit de proposer constamment des nouveaux produits, des améliorations de garanties. Avec notre méthode de calcul, l'actualisation des tarifs se fait rapidement. Prenons l'exemple d'une gamme de deux produits proposant un forfait en optique de 300 euros et de 700 euros, pour lesquels les tarifs sont respectivement X et Y. Comment calculer le tarif d'un nouveau produit C proposant cette fois 500 euros en forfait optique ? Dans la plupart des cas, l'assureur va établir un rapport de proportionnalité entre les anciens forfaits et le nouveau. Avec notre modélisation et notre méthode de calcul, l'assureur pourra avoir le tarif adéquat pour le nouveau montant exact de forfait offert.
Cependant, deux points peuvent être soulevés. Le premier concerne le temps de calcul pour passer des frais réels aux remboursements exacts pour chaque nouveau produit émis. Le deuxième concerne les surcomplémentaires car une mauvaise modélisation peut aboutir sur une consommation réelle plus forte avec un montage « base sans surcomplémentaire » qu'avec un montage « base et surcomplémentaire ». Ceci obligerait alors à fixer une prime pure nulle pour le second montage.

Nous avons pu soulever dans cette partie les points essentiels - points forts, points faibles et problématiques - de notre modélisation. Nous allons dans la partie qui suit appliquer l'algorithme CART afin de voir la valeur ajoutée qu'il apporte réellement par rapport aux MLG.

B. Application de l'algorithme CART

Nous appliquerons l'algorithme CART à la garantie « Prothèses dentaire » car nous disposons d'un volume de données suffisant et car c'est aussi la garantie qui nous a servi d'exemple pour le MLG. Pour l'application de cette méthode nous utiliserons R sur lequel les packages nécessaires peuvent être facilement installés. Tout comme pour les MLG, il est nécessaire de préciser la variable à expliquer et les variables explicatives.

a) Création des bases

Les bases utilisées sont les mêmes que celles utilisées pour la modélisation par MLG. Nous avons néanmoins apporté quelques modifications. Pour la modélisation des frais réels, nous ne travaillons plus avec le logarithme puisque la loi n'est plus à spécifier. Au niveau de la fréquence de consommation, nous prenons toujours en compte l'exposition mais de manière directe. C'est-à-dire que pour chaque assuré, nous faisons le rapport :

$$\frac{Nombre\ d'actes\ consommés}{Exposition}$$

Ce rapport sera notre nouvelle variable « Fréquence » et sera donc celle estimée par l'algorithme.

Nous aurions pu aussi définir une nouvelle fonction d'hétérogénéité prenant en compte l'exposition dans un groupe d'assurés homogène. En effet, dans un tel groupe la fréquence peut être mise sous la forme suivante :

$$Fréquence = \frac{Nombre\ d'actes\ consommés\ dans\ le\ groupe\ G}{Exposition\ totale\ dans\ le\ groupe\ G} = \frac{\sum_{i=1}^{M} N_i}{\sum_{i=1}^{M} E_i}$$

Avec :

- M le nombre d'assurés dans le groupe homogène G
- N_i le nombre total de sinistres pour l'assuré i
- E_i la durée de couverture, soit l'exposition, de l'assuré i

La fonction d'hétérogénéité pour un nœud t aurait alors pu être définie de la sorte :

$$Imp(t) = \sum_{i \in t} (n_i - \hat{n}_i)^2$$

où n_i représente l'estimation de la fréquence par l'algorithme et Imp (t) est alors une minimisation de l'erreur quadratique. Nous n'avons pas opté pour ce choix car nous préférions effectuer le moins de paramétrage possible afin de ne pas fausser notre modélisation. Notre méthode entraine néanmoins une sous-estimation de la fréquence par rapport à la méthode tout juste énoncée. Or, une sur estimation technique est préférable à une sous-estimation dans la mesure où le tarif technique peut toujours ensuite subir des abattements avant d'être proposé commercialement. Il sera donc intéressant de développer un travail à ce niveau.

b) Calibrage du modèle

Nous avons présenté l'algorithme CART comme un modèle non paramétrique. Pourtant, nous devons tout de même fixer certains paramètres au départ de l'algorithme.
Ces paramètres sont expliqués ci-dessous :

- rpart => le package utilisé pour la mise en place de notre algorithme. C'est la fonction qui permettra la création d'un arbre.

- rpart.control => fonctionnalité permettant un certain calibrage du modèle. Les paramètres fixés dans cette fonction sont ceux qui suivent.

- xval => le nombre de validations croisées effectuées automatiquement par l'algorithme. Par défaut, la valeur est fixée à 10. Cela veut dire que la base initiale est divisée aléatoirement en dix parties, et la construction de l'arbre se fait à chaque étape sur seulement neuf bases. La base non sélectionnée sert de base de validation. Le modèle retenu est alors celui ayant fourni le meilleur résultat (erreur minimum) sur les dix tests. Ce nombre de 10 nous apparait trop grand car nous ne disposons pas d'un volume considérable de données, et car la variance semble assez importante. Il est donc plus judicieux d'abaisser cette valeur.

- minbucket => il s'agit d'un critère d'arrêt. Il détermine le nombre minimum

d'individus dans une classe. La valeur doit être fixée grâce à l'expérience de l'actuaire. Une valeur trop élevée conduira à une perte d'information sur certains assurés. Une valeur trop faible conduira à l'obtention d'un arbre sur-apprenti. Bien que ce phénomène soit compensé par le principal critère d'arrêt (minimisation de l'erreur sur la base de validation), il reste néanmoins perturbant pour la modélisation car il empêche la création de nœuds plus significatifs. Nous l'avons fixé à 1000 pour les frais réels car les frais, bien que dépendant des praticiens, restent sensiblement dans la même tranche. Nous pouvons ainsi estimer qu'en dessous de ce seuil, il s'agit uniquement de rares individus que l'on peut intégrer à une classe plus significative. Pour la fréquence nous sommes restés relativement petit (300) car le comportement d'un assuré est plus variable.

- minsplit => ce paramètre détermine le nombre minimum d'individu dans chaque nœud pour que la division binaire soit réalisée. Il est en général égal à trois fois la valeur de minbucket. Nous l'avons fixé à 2000 pour les frais réels et à 1000 pour la fréquence.

- maxsurrogate => indique le nombre de variables de substitutions. Pour certains individus, des valeurs peuvent manquer pour une ou plusieurs variables. Si cela se produit sur une variable de séparation d'un nœud (celle sur laquelle s'est basée la division binaire), l'algorithme va alors choisir les deux variables qui classifieraient l'individu le plus similairement.

- cp => il s'agit du coefficient de pénalisation. Il est d'abord fixé à zéro afin d'obtenir l'arbre complet.

- plotcp => c'est la fonctionnalité fournissant la courbe des erreurs de prédiction en fonction de cp. Le coefficient de pénalisation optimal est alors celui qui minimise l'erreur de prédiction.

- prune => fonctionnalité assurant l'élagage de l'arbre complet, avec en entrée de paramètre le coefficient de pénalisation optimal.

Le paramètre pour la fonction d'hétérogénéité n'est pas précisé ici car la fonction choisie sera celle utilisée par défaut dans rpart. Il s'agit de l'indice de GINI. Nous l'avons choisi pour les raisons évoquées précédemment.

c) Modélisation

Nous avons expliqué dans le chapitre correspondant à l'algorithme CART que celui-ci était très sensible aux changements de données. De plus, le nombre de feuilles de l'arbre est très important et moins facilement lisible si sa construction est réalisée sur une large base de données. Aussi, nous avons décidé de construire plusieurs arbres sur différents échantillons tirés aléatoirement dans nos données. Ces échantillons aléatoires sont dits des échantillons d'apprentissage. À chaque échantillon d'apprentissage est associé un échantillon témoin, qui n'est rien d'autre que son échantillon complémentaire. L'intérêt des échantillons témoins est qu'ils permettent alors de tester et de sélectionner le meilleur arbre au sens de l'erreur de prédiction. Nous avons décidé de construire dix arbres différents sur dix échantillons d'apprentissages. Ce nombre a été retenu en rappel au nombre de validations croisées effectuées automatiquement par l'algorithme lors de la création d'un arbre.

Le processus de construction sera le même pour chacun des échantillons et est décrit ci-après :

1. Création de l'échantillon d'apprentissage i et de l'échantillon témoin i. Les deux échantillons sont de taille égale et forment une partition sur l'ensemble des données.

2. Construction de l'arbre complet $(A_{max})_i$ sur l'échantillon d'apprentissage i. Le coefficient de pénalisation initial est non nul et fixé à 0,001 car il permet d'avoir une courbe des erreurs de prédiction lisible.

3. Obtention du coefficient de pénalisation optimal.

4. Élagage de l'arbre $(A_{max})_i$.

5. Prédiction sur l'échantillon témoin de la variable réponse grâce à l'arbre élagué i.

6. Calcul de l'erreur de prédiction moyenne.

Une fois les dix arbres construits et l'erreur de prédiction moyenne calculée pour chacun des échantillons témoins associés, nous sélectionnons l'arbre dit optimal. Cet arbre optimal est donc tout simplement celui dont l'erreur de prédiction moyenne sera la plus faible.
Puisque nous allons répéter les mêmes étapes dix fois, nous pouvons présenter à titre d'exemple les résultats obtenus uniquement pour les arbres numéro 4 et numéro 5. Nous avons choisi ces arbres car ils sont respectivement les arbres optimaux pour la fréquence et les frais réels.

1. Modélisation de la fréquence

Nous modélisons donc la variable fréquence en fonction de toutes les variables explicatives qui sont celles présentées ci-dessous :

- *l'âge*
- *le sexe*
- *la situation familiale*
- *le secteur d'activité*
- *le statut professionnel*
- *le régime de la Sécurité Sociale*
- *l'effectif*
- *le zonier*
- *le niveau de la base*
- *le niveau de la surcomplémentaire*

Nous avons remarqué lors de nos premiers résultats que certains individus avaient une fréquence très supérieure à la moyenne. Leur taux de recours était estimé à 1,9 fois par an. Ce qui est beaucoup trop élevé pour un type acte tel que les prothèses dentaires. Ces individus formaient une classe composée uniquement d'une centaine d'assurés. Nous avons donc considéré qu'il s'agissait de cas particulièrement rares et qu'ils faussaient notre validation des tarifs. C'est pourquoi nous avons calibré les valeurs du minbucket et du minsplit à la hausse.
La courbe des erreurs de prédiction pour l'arbre numéro 4 est présentée ci-dessous :

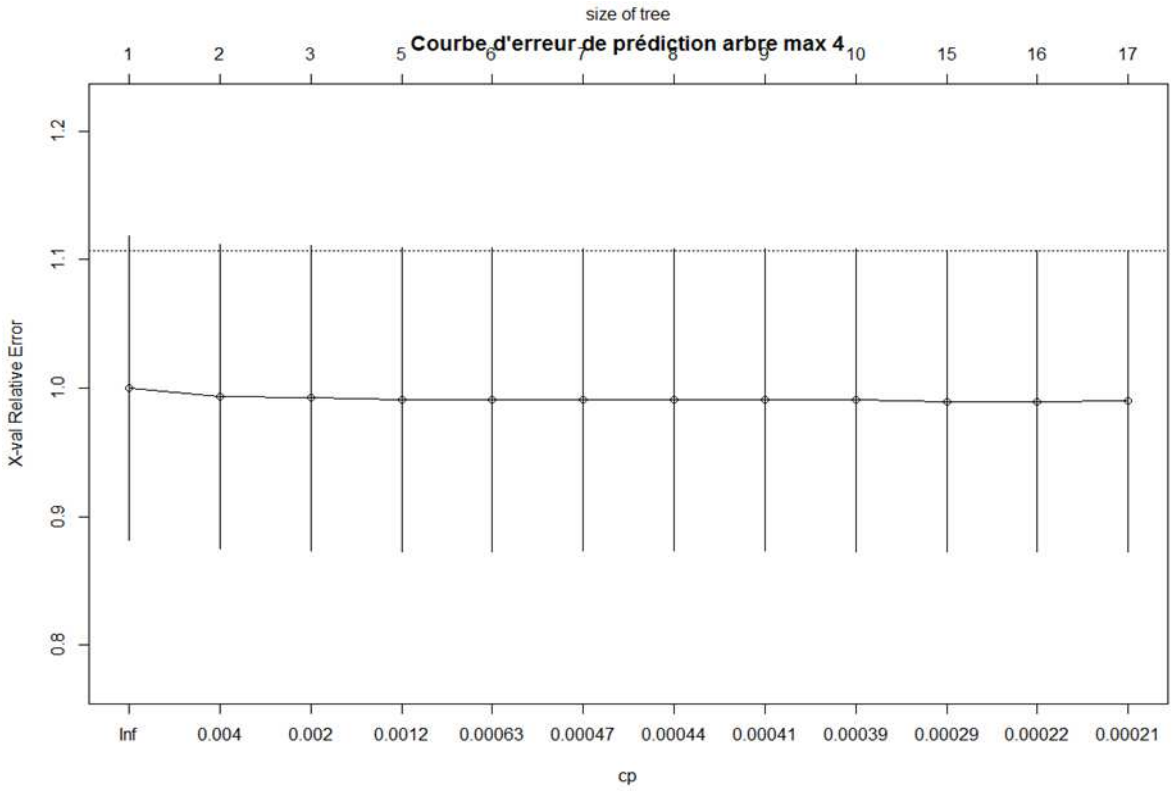


Figure 22. Courbe d'erreur de prédiction de l'arbre complet numéro 4

Le coefficient de pénalisation optimal est le coefficient pour lequel l'erreur est la plus basse. Il n'est pas si aisé de repérer l'abscisse correspondante mais nous remarquons une légère baisse pour la valeur 0,00022 du cp. C'est donc ce coefficient de pénalisation que nous utilisons alors pour l'élagage de l'arbre complet. L'erreur reste pratiquement constante en 1, ce qui est trop élevé pour prétendre à une modélisation acceptable. Néanmoins, nous sommes rassurés quant à notre calibrage car cette erreur est plus importante lorsque nous restons avec les valeurs par défaut du package rpart. Voici la courbe des erreurs de prédictions sans notre calibrage (avec minbucket=3 et minsplit=20).

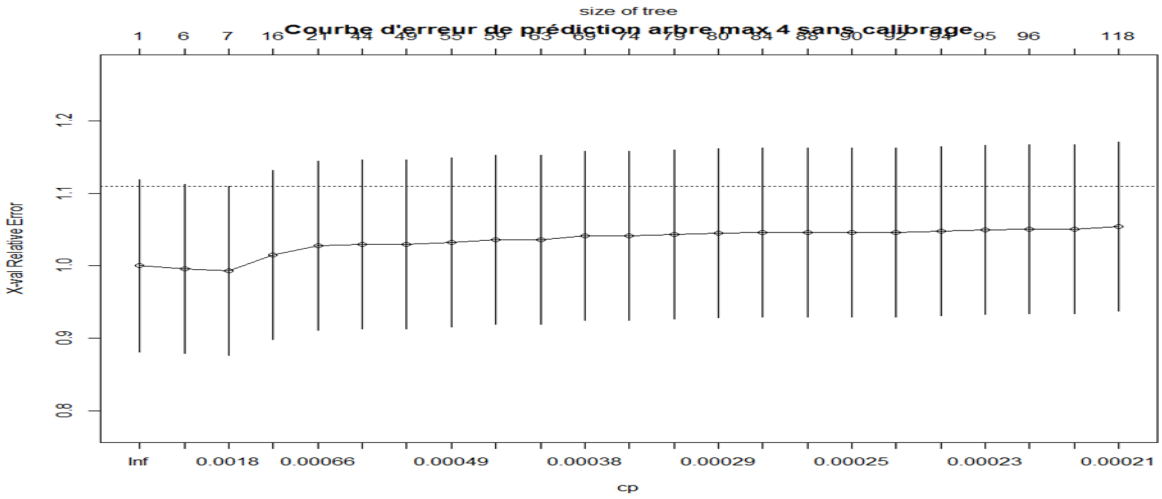


Figure 23. Courbe d'erreur de prédiction de l'arbre complet numéro 4, avec calibrage par défaut

Cette fois l'erreur minimale est aussi aux alentours de 1 mais on observe une augmentation plus significative de l'erreur ensuite. Une mauvaise appréciation dans le choix de notre coefficient optimal peut donc conduire à une plus grande erreur.

L’arbre élagué que nous obtenons après détermination du coefficient optimal est représenté par la figure ci-dessous :

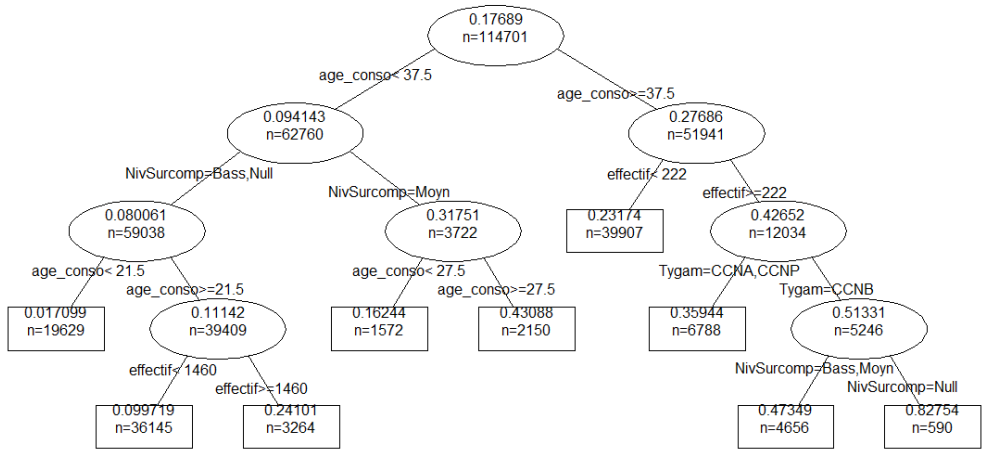


Figure 24. Arbre élagué numéro 4

Les variables discriminantes selon l’algorithme CART sont donc :

- *l’âge de l’assuré*
- *le niveau de la surcomplémentaire*
- *l’effectif*
- *le secteur d’activité*

Nous avions d’autres variables discriminantes obtenues avec les neuf autres arbres élagués, mais nous avons sélectionné uniquement l’arbre 4, dont l’erreur de prédiction moyenne était la plus faible. L’âge de l’assuré et le niveau de surcomplémentaire nous semble être des facteurs discriminants tout à fait logiques. L’effectif est un facteur moins interprétable. En effet, il est difficile de déterminer les raisons entrainant des différences de fréquence de consommation entre entreprises de tailles différentes. Le secteur d’entreprise est un peu plus interprétable car c’est une variable corrélée au niveau de base. En effet, toutes les entreprises d’une même CCN ont approximativement le même niveau de couverture pour leur base.
L’ensemble de l’arbre propose une classification cohérente. Nous observons une plus forte fréquence de consommation pour les assurés ayant une surcomplémentaire élevée et d’âges supérieurs. Au niveau de la séparation par secteur d’activité, nous remarquons que le taux de recours est plus important chez les travailleurs du secteur boulanger. Or, cette CCN propose une base de niveau moyen en acte « Prothèses dentaires » tandis que les deux autres CCN proposent des niveaux bas. La séparation est donc là aussi pleinement justifiée et cohérente.
Nous sommes satisfaits de la classification pour sa simplicité de lecture et pour la cohérence des valeurs obtenues.

2. Modélisation des frais réels

Comme pour la fréquence, nous gardons toutes les variables explicatives lors de la construction de l’arbre. Cette fois ci, l’arbre élagué retenu est l’arbre 5.

La courbe des erreurs obtenue est la suivante :

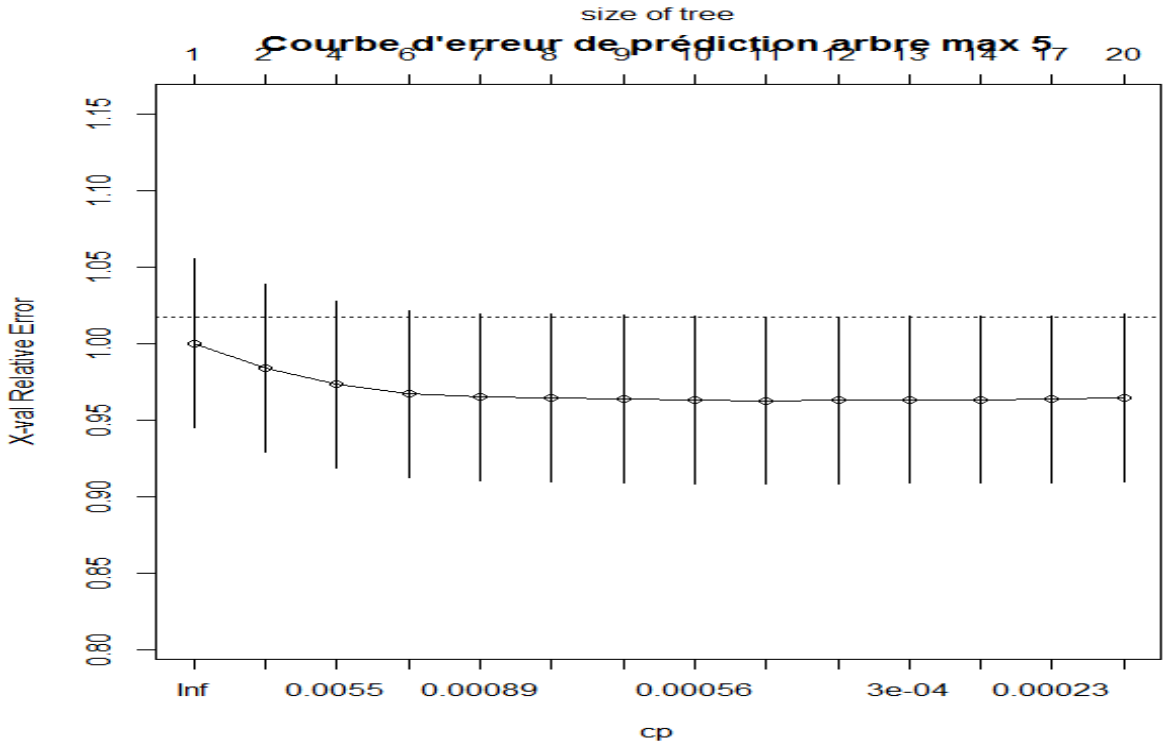


Figure 25. Courbe d'erreur de prédiction de l'arbre complet numéro 5

Cette courbe d'erreur de prédiction est plus intéressante que celle utilisée lors de la modélisation de la fréquence. L'erreur minimale est plus faible et avoisine les 0,95. Ce qui reste tout de même une valeur trop élevée. L'arbre final résultant de l'élagage est tel que le montre la figure suivante :

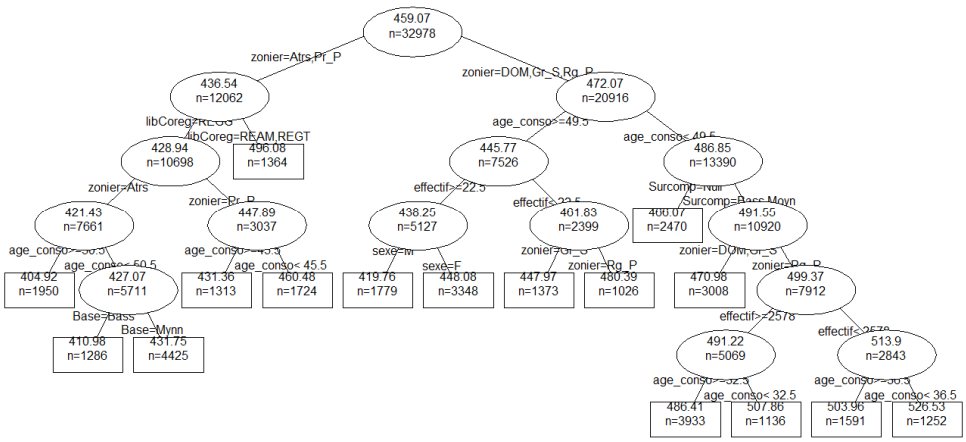


Figure 26. Arbre élagué numéro 5

Afin de n'afficher que les variables discriminantes de façon plus lisible, il est possible d'utiliser la combinaison de fonctions plot - text de R, qui donne alors pour cet arbre la figure qui suit :

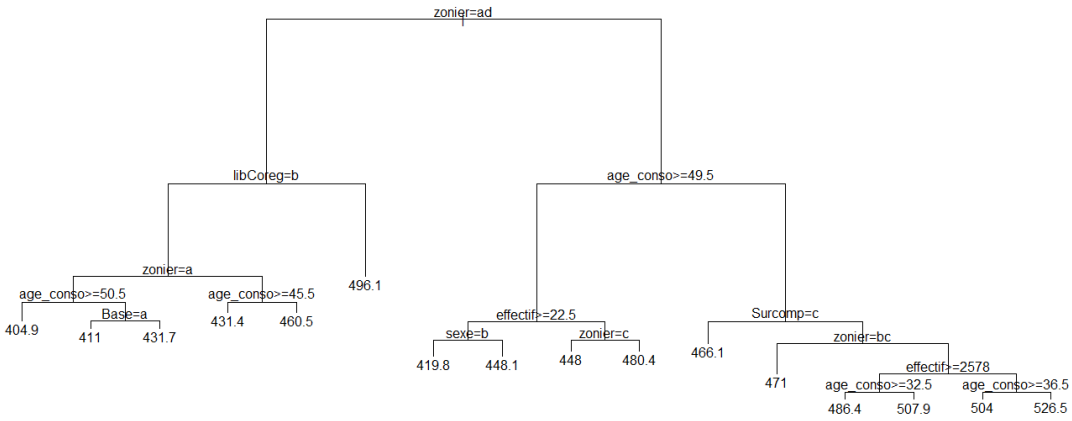


Figure 27. Arbre élagué numéro 5 plus lisible

Les variables explicatives retenues par l'algorithme CART sont les suivantes :

- *le zonier*
- *l'âge*
- *le régime de SS*
- *l'effectif*
- *le niveau de base*
- *le niveau de surcomplémentaire*
- *le secteur d'activité*

Par rapport à l'arbre estimant la fréquence de consommation, nous ajoutons ici deux variables supplémentaires. Le régime de SS est assez intuitif car nous l'avons vu, le régime d'Alsace Moselle par exemple, rembourse plus que le régime général. Il est donc normal que les frais réels engagés soient plus élevés chez les assurés bénéficiant du régime d'Alsace Moselle. Idem en ce qui concerne le zonier. D'une part le zonier est corrélé au régime dans la mesure où tous les départements d'Alsace Moselle sont soumis au régime d'Alsace Moselle. D'autre part, et surement le plus important, il apparait évident que les tarifs proposés par les praticiens des grandes agglomérations telle que Paris ou Lyon seront plus élevés. Les montants de frais réels obtenus nous semblent, pour la majorité, cohérents. Seules certaines séparations apparaissent surprenantes lorsque la variable discriminante est l'âge de l'assuré. En effet, dans certains car les frais réels augmentent avec l'âge et inversement dans d'autres cas. Une étude statistique sur la consommation en prothèses dentaires en France et en fonction de l'âge pourrait être menée afin de mieux juger la fiabilité de ces divisions binaires.

d) Validation par backtesting

Afin de valider nos résultats, nous appliquons la même méthode par backtesting en utilisant les deux mêmes contrats que précédemment. Nous ne détaillerons donc pas de nouveau les caractéristiques propres aux deux entreprises et leur démographie. Les tarifs Homme pour la garantie « Prothèses dentaires » sont identiques aux tarifs Femme car le sexe n'est pas une variable retenue ni dans la modélisation de la fréquence ni dans celle des frais réels.

Prothèses dentaires

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la base	Frais réels	Fréquence	Reste à rembourser	Prime pure
Prothèses dentaires	450% de la Base de Remboursement	75,25	3,80	408,50	Moyenne	466,07	0,099719	390,82	38,97

Tableau 39. Prime pure adulte de la garantie « Prothèses dentaires » sur la base

Nous regardons si cette prime pure permet de recouvrir l'ensemble des prestations versées pour ce contrat.

Résultats de validation

Année	Nombre de bénéficiaires	Pourcentage d'hommes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Prime Individuelle	Somme des primes	Somme des prestations	P/C
2011	10,00	0,70	32,00	38,97	38,97	38,97	389,72		
Année	Nombre de bénéficiaires	Pourcentage d'hommes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Prime Individuelle	Somme des primes	Somme des prestations	P/C
2012	15,00	0,60	32,00	38,97	38,97	38,97	584,58	382,25	153%
Année	Nombre de bénéficiaires	Pourcentage d'hommes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Prime Individuelle	Somme des primes	Somme des prestations	P/C
2013	15,00	0,60	30,00	38,97	38,97	38,97	584,58	659,50	89%
Sur les 3 années									
	Somme des primes	Somme des prestations	P/C						
	1558,89	1041,75	150%						

Tableau 40. Ratios P/C pour la base

Notre prime pure est un peu sur évaluée et nécessitera surement des abattements si elle est retenue par l'assureur. Néanmoins, cette sur évaluation est impactée par le fait que l'année 2011 s'est déroulée sans aucune consommation. Nous pouvons donc penser que la fréquence de consommation est la composante que nous devrions revoir à la baisse.

Prothèses dentaires

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la base	Frais réels	Fréquence	Reste à rembourser (Montage)	Reste à rembourser	Prime pure
Prothèses dentaires	75% de la Base de Remboursement	75,25	4,55	489,125	Basse	486,41	0,47349	411,16	20,34	9,63

Tableau 41. Prime pure adulte de la garantie « Prothèses dentaires » sur l'option 1

Nous calculons comme précédemment le rapport P/C, et nous obtenons le tableau ci-après :

Résultats de validation

Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2011	2761,07	3229,76	40,00	9,63	9,63	57696,42	61710,93	93%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2012	2743,69	3246,77	40,00	9,63	9,63	57692,78	56485,63	102%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2013	2679,81	3208,59	40,20	6,39	13,13	56709,94	50796,20	112%
Sur les 3 années	Somme des primes	Somme des prestations	P/C					
	172099,14	168992,76	102%					

Tableau 42. Ratios P/C pour l'option 1

Seule l'année 2011 est en déficit. Les deux autres années, notre prime pure obtenue permet de recouvrir les engagements de l'assureur. Et ce de manière raisonnable, puisque sur les trois années, nous ne faisons que 2 % de marge supérieure. Les résultats nous semblent donc tout à fait corrects.

Prothèses dentaires

Garantie	Type et Niveau de remboursement	Montant de remboursement SS	Taux proposé en complément SS	Montant proposé en complément SS	Niveau de la surcomplémentaire	Frais réels	Fréquence	Reste à rembourser (Montage)	Reste à rembourser	Prime pure
Prothèses dentaires	125% de la Base de Remboursement	75,25	5,05	542,875	Moyenne	486,41	0,47349	411,16	20,34	9,63

Tableau 43. Prime pure adulte de la garantie « Prothèses dentaires » sur l'option 2

L'algorithme CART nous donne exactement le même tarif que pour l'option 1. Cela est étonnant et ne peut arriver en réalité. Dans un tel cas, il est évident que les assurés prendraient constamment l'option 2 qui les couvre plus et qui est au même prix que l'option 1. Nous pouvons déjà supposer que l'estimation du tarif n'est pas fiable.

Nous le vérifions grâce à notre tableau de validation qui suit :

Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2011	2761,07	5315,24	40,00	9,63	9,63	77781,28	272144,28	29%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2012	2743,69	5343,23	40,00	9,63	9,63	77883,41	220988,65	35%
Année	Nombre d'hommes	Nombre de femmes	Age	Prime Pure individuelle Homme	Prime Pure individuelle Femme	Somme des primes	Somme des prestations	P/C
2013	2679,81	5280,41	40,20	9,63	9,63	76663,17	191956,92	40%
Sur les 3 années	Somme des primes	Somme des prestations	P/C					
	232327,87	685089,85	33,9%					

Tableau 44. Ratios P/C pour l'option 1

Si nous gardons le tarif obtenu par l'arbre élagué numéro 5 pour l'option 2, nous sommes alors en total déficit. Chaque année, l'assureur verse beaucoup plus qu'il ne reçoit en primes. Notre modélisation est donc fausse, comme nous l'avions présagé plus haut. Les frais réels estimés étant les mêmes pour une option basse qu'une option moyenne, nous en déduisons que la modélisation des frais réels n'est pas adéquate. Ceux-ci ont été sous évalués par l'algorithme CART.

e) Bilan des résultats

L'algorithme CART nous a permis d'identifier les variables explicatives discriminantes de manière claire et lisible. Nous avons remarqué qu'il y avait des interactions entre variables. En effet, nous ne retrouvions pas nécessairement les mêmes variables séparatice en fonction du nœud. Cela se traduit donc par le fait qu'une modalité de variable est plus discriminante en fonction de la modalité d'une autre variable. Ceci est bien un effet d'interaction, bien apparent grâce à la classification de l'arbre élagué.
Au niveau des frais réels, nous trouvons des tarifs qui nous paraissent juste pour une base de niveau moyen et une surcomplémentaire de niveau bas. Néanmoins, le tarif final obtenu pour la base semble élevé par rapport aux prestations réellement versées par l'assureur. La fréquence obtenue étant relativement faible, nous supposons donc plutôt une sur estimation des frais réels engagés par les assurés.
Au niveau de la fréquence de consommation, celle-ci semble bien ajustée pour une surcomplémentaire de niveau bas, puisque les tarifs finaux obtenus sont très satisfaisants. Pour la base que nous avons tarifée, il est difficile de se prononcer, dans la mesure où le tarif final est important et peut être dû aussi bien aux frais réels qu'à la fréquence. La modélisation reste cependant logique puisque la fréquence est plus basse que pour les assurés ayant une surcomplémentaire.
En ce qui concerne la surcomplémentaire de niveau moyen, la modélisation n'est pas acceptable car elle mettrait en danger l'assureur vis-à-vis de ses engagements. L'arbre élagué retenu indique une fréquence de consommation similaire entre une surcomplémentaire basse et une moyenne. Cela n'est pas forcément surprenant car comme nous l'avons expliqué plus haut, il s'agit souvent d'un réel besoin qui n'est pas dû à un manque de responsabilisation de l'assuré. Cependant, il est très surprenant que l'arbre n'ait pas

différencié les deux niveaux de surcomplémentaire au niveau des frais réels. Nous pouvons donc supposer que les frais réels sont la composante pour laquelle la modélisation n'est pas adaptée.

f) Discussion des résultats

Nous n'avons modélisé la fréquence et les frais réels que pour la garantie « Prothèses dentaires ». Nous n'avons donc pas assez de résultats pour pouvoir réellement émettre des hypothèses par rapport à cette méthode de modélisation. Néanmoins nous soulevons deux points qui peuvent être améliorés. Le premier point concerne l'exposition de l'assuré. Celle-ci peut être prise en compte de plusieurs manières possibles, que nous avons déjà évoquées lors de la modélisation par MLG. Il est donc tout aussi important d'y réfléchir dans le cadre d'un algorithme CART. D'autant plus que le package rpart permet de modifier la fonction d'hétérogénéité. L'effet de la durée de couverture peut ainsi être implémenté aussi bien dans l'algorithme que par calcul direct dans la base de données. Dans le premier choix, il faudrait alors définir la relation entre hétérogénéité et exposition. Le second point auquel nous avons pensé est le choix des nœuds. L'algorithme CART ne donne pas d'intervalles de confiance. Néanmoins, l'actuaire peut considérer qu'une division binaire ne lui semble pas adéquate et garder la valeur de la variable réponse proposé par le nœud juste inférieur. En faisant ainsi pour les frais réels engagés pour ceux ayant une option 2, nous nous ramenons à un ratio de 70%. Ce ratio reste déficitaire mais permet de se rapprocher beaucoup plus de la consommation réelle. Et la division binaire négligée est celle faite sur la variable effectif. Séparation qui peut sembler étonnante car il est difficile d'expliquer comment la taille d'une entreprise influe sur les frais réels. Certes, une entreprise de grande taille est souvent une compagnie importante proposant de bons contrats en santé collective. Mais cela est déjà intégré par les variables sur les niveaux de couverture. Nous pouvons penser que dans notre étude, quelques très grosses compagnies font partie d'une même CCN et influence notre modélisation. Une autre piste d'améliorations peut être d'intégrer de nouvelles variables explicatives. En effet, au lieu de segmenter régionalement, nous pourrions affiner la segmentation en fonction de la densité de médecins généralistes, du climat (pluviométrie, ensoleillement), du taux de chômage, du nombre de licenciés sportifs ; toutes ces variables pouvant avoir un effet conséquent sur la consommation en frais de santé. L'algorithme CART serait alors un outil de premier choix car il permet de travailler sur un très grand nombre de variables. Les travaux de YOHANNES.Y & HODDINOTT.J [2006] ou encore de FRIEDMAN.J & BREIMAN.L & OLSEN.R & STONE.C [1984] présentent d'autres exemples de mise en application de ce type d'algorithme, et permettent d'illustrer l'intérêt majeur de tels arbres de décision.

C. Comparaison des deux méthodes

Nous avons dans ce mémoire utilisé deux méthodes de modélisation totalement différentes. Afin de les comparer au mieux, nous allons donc séparer chaque aspect que nous jugeons importants lors d'une tarification.

Traitement des données

Le traitement des données nécessite approximativement le même travail que l'on opte pour une modélisation par MLG ou par CART. Les deux méthodes nécessitent un important volume de données, et sont sensibles aux données de l'étude. Il faudra donc veiller à retraiter efficacement les données. Néanmoins, l'algorithme CART est plus souple car il est possible de garder dans l'étude les individus pour lesquels des valeurs manquent. Alors que nous devons les supprimer lors d'un MLG, ces individus seront

tout de même classés - par l'intermédiaire d'une variable de substitution - lors d'une modélisation par CART.

Paramétrage

Réaliser une tarification par MLG implique une analyse statistique du portefeuille très approfondie. Celle-ci est indispensable à la réussite de la modélisation car elle nous permet d'orienter notre choix lors du paramétrage. Ce paramétrage doit être déterminé avec justesse car les résultats en dépendent totalement. L'actuaire doit ainsi choisir la distribution de la variable réponse, et la fonction de relation - dite fonction lien - entre cette variable et les variables explicatives. Le choix de la distribution se fait généralement en observant les statistiques descriptives effectuées auparavant et selon l'historique de consommation du groupe assureur. La fonction lien va plus souvent être choisie selon un souci d'interprétabilité et de facilité de calcul de la prime. Nous ne retrouvons pas tout ce travail lors d'une tarification par CART. C'est une méthode non paramétrique, et donc beaucoup moins contraignante. Nous devons néanmoins calibrer l'algorithme en lui indiquant par exemple le nombre d'individus minimum par classe ou encore le critère d'arrêt. Quant à l'exposition, l'algorithme CART et le MLG peuvent tous deux en prendre compte assez aisément. Le lecteur pourra trouver des informations pertinentes sur le paramétrage des MLG dans le mémoire LAGADEC.F [2009], et aussi sur le paramétrage d'un algorithme CART grâce au travail de VESIN.A [2006].

Effet d'interaction

L'interaction se traduit par le fait qu'une variable influe sur une autre. Puisque l'objectif d'une tarification est de déterminer l'impact de chaque variable sur la consommation, il est important de supprimer cet effet. Ou du moins de spécifier au modèle qu'une interaction existe. C'est ce qui se fait dans le cadre d'une tarification par MLG. L'utilisateur doit préciser les interactions entre variables. Cela exige donc un travail de recherche de toutes les interactions possibles. Travail qui peut donc être très long lorsque le nombre de variables est important. Dans notre étude, nous avons vu que plusieurs interactions étaient à prendre en compte. Certains croisements sont moins faciles à interpréter et à expliquer. Avec l'algorithme CART, aucune interaction n'est à rechercher car cela se fait automatiquement par l'algorithme. Les divisions successives sont représentatives des interactions possibles entre variables. L'algorithme CART est donc un outil de modélisation particulièrement puissant qui permet un gain de temps considérable, notamment lors d'une tarification avec un nombre de variables explicatives conséquent.

Lisibilité et clarté

Les deux méthodes donnent des résultats clairs et simples aussi bien à comprendre qu'à expliquer à un commercial ou à n'importe quel interlocuteur. Dans le cas du MLG, la notion de majoration/minoration est plus forte car chaque modalité est associée à un coefficient. Un MLG permet donc clairement d'identifier l'effet de chacune des variables sur la consommation. L'algorithme CART est différent dans le sens où il n'indique pas l'influence exacte de chaque variable. Il indique uniquement la variable et sa valeur qui cause responsable d'une différenciation sur la consommation. Dans cette méthode, nous perdons donc une information assez importante.

Bien que les deux méthodes soit parfaitement lisibles, la lecture d'un arbre issus de l'algorithme CART est beaucoup plus simple et appréciable. D'une part, nous remarquons de suite quelles sont les variables explicatives réellement importantes. D'autre part, la valeur de la variable à expliquer est directement affichée sur l'arbre.

Segmentation

Nous l'avons en partie cerné dans le paragraphe précédent, la notion de segmentation est beaucoup plus forte dans une modélisation par MLG. Chaque modalité de variable ayant un coefficient qui lui est affecté, nous aurons donc un tarif final différent pour chaque individu ayant des caractéristiques

différentes. Cette segmentation est moins présente avec l'algorithme CART car la séparation se fait à chaque nœud selon un seuil. Pour l'âge par exemple, le but n'est plus d'estimer la consommation pour chaque tranche d'âge mais d'estimer la consommation en-dessous et au-dessus d'un certain âge. C'est donc pour cette raison qu'intervient le critère d'hétérogénéité dans l'algorithme CART. Les groupes formés étant plus larges, il est important de voir le degré d'homogénéité au sein des classes. Il en ressort que la méthode par MLG est plus adaptée lorsque l'étude porte sur l'estimation de l'effet d'une variable en particulier.

Calculs supplémentaires

Alors que l'algorithme CART nous indique directement la valeur estimée, le MLG ne nous donne que l'estimation des coefficients. C'est donc à nous de calculer ensuite la valeur finale des frais réels et de la fréquence, en fonction des caractéristiques de l'individu. Cela peut être un travail plus ou moins long en fonction du nombre de variables explicatives. Dans notre étude, le temps de calcul était raisonnable. Mais nous aurions pu utiliser beaucoup plus de variables et le calcul aurait alors été particulièrement long.

Résultats

La tarification de la garantie « Prothèses dentaires » a mis en relief des résultats sensiblement différents, notamment pour la base et la surcomplémentaire de niveau moyen. Les tarifs trouvés par MLG pour la base ont été rejetés car ils étaient bien trop élevés pour être retenus. Au contraire, les tarifs sortis par l'algorithme CART ont été jugés beaucoup plus acceptables car ils collaient plus à la consommation réelle. En ce qui concerne la surcomplémentaire de niveau moyen, c'est l'inverse qu'il s'est produit. Les tarifs issus de l'algorithme CART étant beaucoup trop bas pour être acceptés tandis que ceux obtenus par le MLG étaient particulièrement proche de la consommation réelle. Pour la surcomplémentaire de niveau bas, les deux méthodes nous ont donnés satisfaction, avec un léger avantage pour l'algorithme CART car celui-ci recouvre l'ensemble des prestations de manière plus juste (2% d'écart contre 6% d'écart).

Les problèmes principaux rencontrés avec le MLG ont porté sur la fréquence de consommation, qui a impacté à la hausse les tarifs finaux. Le problème de modélisation de la fréquence semble moins présent avec l'algorithme CART. En effet, alors que la fréquence de consommation - pour les assurés souscrivant uniquement à la base - est égale à 0,34 avec le MLG, elle est de 0,099719 avec l'algorithme CART. Selon l'algorithme CART, ces assurés ont donc recours aux actes de prothèses dentaires 3,5 fois moins que l'estimation faite par MLG. Cependant, la modélisation des frais réels n'a pas été aussi satisfaisante avec l'algorithme CART qu'elle ne l'a été avec le MLG. Nous avons obtenu des résultats incohérents au niveau de la surcomplémentaire de niveau moyen pour l'acte « Prothèses dentaires ».

Il est difficile ici de déterminer quelle a été la meilleure modélisation. Une tarification juste de la base est indispensable car elle correspond au contrat auquel les assurés souscrivent le plus. Il est donc important pour l'assureur de fixer un tarif qui lui permette de bien couvrir ses engagements. Nous pourrions donc dire que l'algorithme CART est la méthode à retenir pour modéliser les tarifs de la base. Au niveau des options, le grand écart entre la consommation réelle et l'estimation faite par l'algorithme CART nous conduise à retenir le MLG. Au final, le MLG serait surement retenu par l'assureur car le MLG fournit des tarifs plus sécurisants et le tarif de la base pourra être abattu ensuite en fonction du marché.

Pilotage du risque

Comme nous l'avons vu lors de la validation par backtesting, le MLG propose une solution très intéressante pour l'actuaire qui peut estimer le risque par l'intermédiaire d'intervalle de confiance. Ceci confère au MLG un très gros avantage par rapport à l'ensemble des méthodes de tarification dont l'algorithme CART. Ainsi, bien que la tarification de la base par MLG ait été rejetée au départ, nous avons pu produire des tarifs beaucoup plus justes en fonctionnant avec les bornes inférieures des coefficients. Le ratio P/C a alors été de 151% contre 150% pour l'algorithme CART. Un des objectifs pour l'actuaire est de constamment estimer le risque pris lors de ses choix. Le MLG permet donc de répondre à cet objectif.

Conclusion

Il n'est pas approprié de définir une méthode comme étant meilleure qu'une autre car plusieurs paramètres sont à prendre en considération. L'un des plus importants est de savoir quel est l'objectif réel de l'actuaire lorsqu'il effectue une tarification. Le but peut être d'obtenir un modèle avec un fort pouvoir explicatif ou le but peut être encore d'obtenir des résultats parfaits sans pouvoir les interpréter. Compte tenu de notre travail, il nous apparait que l'algorithme CART est préférable pour sa lisibilité, sa facilité de mise en place et sa capacité de détection des interactions. Le MLG, quant à lui, est plus intéressant pour son pouvoir explicatif de chaque variable et son pilotage du risque. L'intérêt de ce mémoire n'est pas simplement d'effectuer une tarification, mais aussi d'estimer l'impact du montage base - surcomplémentaire. Le MLG nous fournit des coefficients de majoration/minoration pour chaque modalité issue du croisement des deux variables (plus l'âge) et nous permet ainsi de répondre à notre problématique : comment l'assuré se comporte en fonction du niveau de sa base et du niveau de sa surcomplémentaire ? La modélisation par MLG est donc celle que nous devons retenir dans notre étude. Néanmoins, construire un algorithme CART et en analyser les résultats avant d'effectuer un MLG peut être une méthode efficace d'éliminer les variables inutiles et d'identifier les interactions à spécifier.

D. Hypothèse d'indépendance

Lorsque nous avons présenté la méthode Fréquence - coûts, nous avons émis l'hypothèse d'une indépendance entre la fréquence de consommation et les montants de sinistres par actes. Il s'agit là d'une hypothèse très forte qui est à la base de ce modèle. Est-elle vraiment justifiée ? De manière générale, l'hypothèse d'indépendance est acceptable pour les risques de haute fréquence tels que les risques « bris de glace » en assurance auto, qui concernent des risques de faibles coûts. Tandis que les risques très rares tels que les risques de catastrophes, générant des montants de remboursements très élevés, ne vérifient pas cette hypothèse. La santé est portée plus particulièrement sur les risques de haute fréquence et nous pouvons donc adopter cette hypothèse. De plus, la segmentation permet d'accentuer cette indépendance. Les prix ont tendance à s'uniformiser entre praticiens d'un même département et n'influent donc plus sur la fréquence. Ceci va être matérialisé dans nos modèles par une segmentation régionale. Pour certains actes, fréquence et coûts varient avec le niveau de couverture. La segmentation par niveau de couverture va permettre d'éliminer ce facteur commun. Ces deux exemples illustrent en partie l'importance d'une bonne segmentation. Afin de s'assurer de la validité de cette hypothèse sur notre base de données, nous avons étudié l'indépendance de la fréquence et des coûts engagés en obtenant les coefficients de Spearman pour chacune des garanties. Pour chacune d'entre elles -lorsque qu'aucune segmentation n'est faite- avec un seuil d'acceptation de 5 pourcent, l'indépendance est rejetée par le test. La p - value est, pour la garantie « prothèses dentaires » inférieur à 0,0001. Pourtant, en travaillant par région, le test d'indépendance affiche une p-value bien supérieur à 0,005 pour la plupart des garanties étudiées. La p-value, pour la garantie « prothèses dentaires » au sein de la région parisienne, est alors égale à 0,2529, laissant ainsi supposer l'indépendance entre la fréquence et les coûts engagés. Néanmoins, ces résultats ne se retrouvent pas pour la garantie « Chambre particulière », ce qui n'est pas surprenant au vu de ce qui a été énoncé précédemment. En effet, la garantie « chambre particulière », contrairement à des garanties comme « consultations généralistes » peut être assimilée à des actes plus rares.

Au final, nous pouvons supposer l'indépendance entre nos deux composantes fréquence et coûts, tout en gardant en mémoire qu'un test d'indépendance ne suffit pas à valider une hypothèse et que les résultats pourront en être impactés.

E. Tarification d’un contrat

L’étude ayant été réalisée dans l’équipe santé collective, nous devons donc proposer une méthode de tarification d’un contrat collectif. Une solution triviale serait une tarification tête par tête. Cela reviendrait donc à réaliser la même méthode utilisée pour le backtesting. Nous récupérons toutes les informations nécessaires sur l’entreprise ainsi que l’âge moyen de l’ensemble des salariés. Enfin, une prime pure est établie pour chaque assuré en fonction de ses critères propres telle que sa situation familiale et son sexe. Or, d’une part l’assureur n’a pas toujours accès à toutes les informations sur les assurés (nombre d’enfants à charge, marié ou divorcé et autres caractéristiques). D’autre part, les règlementations interdisent la différenciation non objective. C’est-à-dire une différenciation faite sur les caractères propres à l’individu. L’organisme assureur ne peut proposer des tarifs différents que pour des catégories objectives, en particulier en fonction du statut professionnel. Dans la majorité des cas, l’entreprise pourra donc opter pour une ou plusieurs structures de cotisations telle qu’une structure uniforme ou encore une structure familiale. Chaque assuré pourra alors choisir la structure de cotisation souhaitée. Ces cotisations sont alors établies soit par rapport à la démographie réelle de l’entreprise si l’assureur dispose de l’information, soit par rapport à une démographie théorique. Dans ce type de démographie, l’assureur estime la répartition de la population grâce à son expérience.

L’assureur doit établir un unique tarif pour les adultes. Néanmoins, en pratique, la différence de consommation entre les hommes et les femmes est prise en compte. Si cette différente n’est pas explicite dans la modélisation choisie, elle pourra être implémentée ensuite dans le calcul de la cotisation finale, en pondérant par le rapport de consommation historique observé dans le portefeuille du groupe.

VI. Conclusion générale

Nous avons dans ce mémoire fait une revue des différentes méthodes de tarification les plus usuelles ainsi que les méthodes en plein développement. Notre travail nous as permis d'identifier toute les étapes et leurs importances pour à une tarification acceptable. De plus, nous avons mis en relief les différents objectifs d'une tarification, qui s'ajoutent à celui de couvrir tous les engagements futurs. Ces objectifs dictent souvent le travail de l'actuaire et déterminent le choix de l'outil de modélisation.

Dans le cadre de l'ANI, de nombreuses modifications sur les garanties et donc sur les contrats proposés sont à prévoir. La notion de « contrats responsables » conduit les assureurs à proposer des contrats selon un montage de type Base plus Surcomplémentaire, avec un niveau de base assez faible. De plus, le secteur de la santé collective, contrairement à la santé individuelle, est souvent en déficit et nécessite une meilleure appréciation des risques réels couverts par l'assureur. Or, la généralisation de la couverture d'entreprise issue de l'ANI entrainera une augmentation des contrats collectifs au détriment des contrats individuels. Aussi, notre mémoire s'est penché sur deux méthodes pouvant répondre à ces enjeux futurs. Une méthode Fréquence - Coût par MLG, permettant d'estimer au mieux l'effet induit par le nouveau type de contrat. Et une méthode par apprentissages statistiques, l'algorithme CART, qui assure une lecture simple par arbre de la consommation d'un individu.

Dans notre processus de modélisation, nous avons préféré modéliser la consommation réelle de l'individu (frais réels) plutôt que le montant engagé par l'assureur. Les raisons sont multiples. D'une part, il est possible de retrouver le montant du remboursement par simples calculs en fonction du type et du niveau de couverture proposé par l'assureur. Nous pouvons ainsi passer aisément des frais réels aux remboursements effectués par l'assureur. D'autre part, cette approche nous permet d'actualiser plus facilement les bases tarifaires dans le cadre d'évolutions législatives ou lors d'un changement de type de garantie souhaité par l'assuré. Cette approche est d'ailleurs justifiée par l'environnement actuel de l'ANI exigeant la mise en place de contrats responsables avec des types de couverture différents de ceux proposés au sein d'AG2R. Enfin, pour un assureur désirant identifier l'impact réel d'une surcomplémentaire sur la consommation, il est intéressant de voir ce que l'assuré engage réellement comme montants.

La mise en place de ces deux méthodes nous as conduit à revoir les aspects théoriques sous - jacents et à réfléchir sur les points susceptibles d'être améliorés. Nous avons notamment jugé la prise en compte de l'exposition comme un point essentiel de recherche future par l'actuaire, afin de modéliser au mieux la fréquence de consommation. Les deux méthodes ont donné des résultats très différents.
La méthode par MLG nous a satisfaits car elle a permis d'affecter un coefficient de majoration/minoration pour chaque type de montage Base - Surcomplémentaire. De plus, cette méthode nous a montrés tout son intérêt lorsque nous avons utilisé les intervalles de confiance obtenus, permettant un pilotage du risque efficace. Grâce à cette méthode, nous avons pu proposer des tarifs cohérents et utilisables par l'assureur pour la majorité des garanties que nous avons étudiées.
La méthode par algorithme CART a été appliquée uniquement sur la garantie « Prothèses dentaires ». Elle nous a tout de même montrés son potentiel de détection des variables discriminantes. Sa clarté et sa facilité de mise en place en font une méthode particulièrement intéressante qui nécessite d'être mieux développée dans le domaine de l'assurance santé.
Néanmoins, notre modélisation - par MLG et par CART - a rencontré certains problèmes.

La modélisation de la fréquence de consommation n'a pas été aussi satisfaisante que celle des frais réels. Cette difficulté de modélisation peut être due à plusieurs sources, telles qu'un mauvais paramétrage ou une mauvaise prise en compte de l'exposition. Il se peut aussi que la fréquence de consommation soit une

variable plus difficile à modéliser car elle est avant tout dépendante du comportement de l'assuré et de ses besoins. Elle est donc propre à chaque assuré et nous pourrions ainsi avancer qu'il y a autant de fréquences possibles que d'assurés. À l'inverse, les frais réels dépendent avant tout des tarifs demandés par les praticiens. Or, ceux-ci ne peuvent pas non plus fixer n'importe quels tarifs et doivent, afin de rester compétitifs vis-à-vis de la concurrence, respecter des tranches de prix. Il est donc bien plus aisé de modéliser une telle variable. C'est pourquoi il est nécessaire d'avoir une base de données importante afin de modéliser au mieux la fréquence de consommation d'un individu.

Nous avons aussi identifié deux autres points de discussion, causés par le fait que nous travaillions garantie par garantie. Tout d'abord, toutes les garanties ne sont pas toujours concernées par les surcomplémentaires. Prenons le cas de deux assurés A et B. Le premier, étant employé dans une entreprise proposant une surcomplémentaire en optique. Le second, travaillant dans une entreprise ne proposant aucune surcomplémentaire. De plus, les deux entreprises offrent la même base, et l'assuré A souscrit à la surcomplémentaire. Or, nous ne savons pas pour quelles garanties en particulier l'assuré A a souscrit à la surcomplémentaire. Est-il forcément justifié d'avancer l'idée que l'assuré A désire se couvrir plus en optique. Il se peut qu'il ait souscrit à la surcomplémentaire plutôt pour une autre garantie. Notre théorie, basée sur le fait que la souscription à une surcomplémentaire est indicatrice du comportement de l'assuré, peut dont être totalement fausse pour certaines garanties. Finalement, la seule information réelle que nous ayons, est que cet assuré a la possibilité d'engager plus de frais réels pour l'optique s'il le désire.
Ensuite, dans une approche garantie par garantie, nous négligeons les dépendances qui existent entre certains actes. Le parcours de soins induit nécessairement ce type de dépendance. Ainsi, un assuré ira la plupart du temps en pharmacie après avoir d'abord obtenu une ordonnance de la part d'un médecin généraliste. Il ira chez l'opticien après avoir été au préalable chez un ophtalmologue. La fréquence de consommation de certains actes clés tels que les « consultations généralistes » influent donc fortement le taux de recours à d'autres actes. Une solution possible pour cette problématique serait alors d'adopter une vision client, comme illustrée dans le mémoire PAGLIA.A [2010]. L'objectif réel d'une tarification étant d'établir la prime pure de l'assuré pour un contrat, il est étonnant que les actuaires ne modélisent pas directement le montant global payé par l'assuré, en agrégeant donc les garanties couvertes. Une telle approche pourrait d'ailleurs être plus intéressante dans le nouveau cadre imposé par l'ANI car les contrats de base proposés par les assureurs seront approximativement tous de même niveau.

S'ajoute à cela une autre problématique : en gardant toujours l'exemple ci-dessus, nous observons qu'il est très difficile de comparer la consommation réelle que pourraient avoir les assurés A et B. En effet, dans notre modèle, nous considérons que l'assuré A désire se couvrir plus en optique que ce que lui permet le contrat de base. Nous supposons que sa consommation dépassera les montants maximum proposés par la base, et pourra atteindre les niveaux maximum de la surcomplémentaire. Cependant, en ce qui concerne l'assuré B, nous ne pouvons rien avancer car ce dernier n'a pas eu le choix. Il se peut que ce niveau de base soit trop faible pour lui. Il est donc possible que cet assuré consomme totalement les niveaux de garanties qui lui sont proposés, et dépasse même la consommation de l'assuré A. Dans notre étude, nous avons donc une information supplémentaire car nous savons que c'est l'assuré qui a choisi de se couvrir plus. Il serait intéressant d'établir une méthode qui modélise nos variables réponses conditionnellement à cette possibilité pour l'assuré de choisir ou non une surcomplémentaire.

Un autre aspect problématique a été identifié dans ce mémoire : la méthode fréquence - coût n'est valable que si l'hypothèse d'indépendance est vérifiée. Celle-ci n'a pas été vérifiée pour toutes nos garanties et peut expliquer les écarts que nous obtenons. Il est donc primordial pour l'actuaire de déterminer les actes qui pourront respecter cette hypothèse et ceux qui devront être placés dans la catégorie risques « extrêmes » et être donc tarifés autrement. Plusieurs méthodes, que nous ne détaillerons pas ici, existent pour cette classe de risques.

Enfin, nous avons présenté les avantages d'une modélisation directe des frais réels. Néanmoins, cette méthode implique un biais non négligeable par rapport aux remboursements finalement effectués par l'assureur. En effet, nous considérons toujours dans nos calculs de passage le remboursement maximal proposé par l'assureur. Or, ce remboursement n'est pas toujours accordé à l'assuré. Un assuré ne respectant pas le parcours de soins bénéficiera d'un remboursement diminué. Cette possibilité n'est donc pas prise en compte dans notre méthode et induit ainsi une marge d'erreur dans nos résultats finaux.

Pour conclure ce mémoire, il s'avère que la méthode par MLG reste encore la méthode la plus efficace dans un modèle Fréquence - Coût et garantie par garantie. Méthode plus pratique aussi dans le cadre des contrats sur mesure, où l'entreprise choisit spécifiquement les garanties et les niveaux de couverture qu'elle désire. Néanmoins, les méthodes par apprentissages statistiques, dont l'algorithme CART, peuvent devenir les nouvelles méthodes majeures si les chercheurs continuent de les adapter au monde de l'assurance. Avec l'ANI notamment et pour les contrats standard du groupe AG2R La Mondiale, une approche client combinée à notre algorithme CART, pourrait s'avérer plus utile afin d'estimer le coût de la base. L'estimation de l'impact des surcomplémentaires se ferait alors par MLG. Quant à la modélisation des frais réels, celle-ci représente une alternative intéressante mais des hypothèses supplémentaires doivent être introduites par la suite, afin de s'approcher au maximum de l'engagement futur réel de l'assureur.

VII. Bibliographie

Littéraire

FARAWAY J.J. [2006] “Extending the Linear Model with R : Generalized Linear, Mixed Effects and Non parametric Regression Models”, *Boca Raton, FL: Chapman & Hall/CRC*

FRIEDMAN J., BREIMAN L., OLSHEN R.A. & STONE C.J. [1984] “Classification and Regression Trees”, *Wadsworth Statistics/Probability, First Edition*

HASTIE T., TIBSHIRANI R. & FRIEDMAN J. [2009] “The Elements of Statistical Learning”, *Springer Series, Second Edition*

LAGADEC F. [2009] “Tarification d’un contrat de complémentaire santé par un modèle linéaire généralisé”, *Mémoire d’actuariat EURIA*

PAGLIA A. [2010] “Tarification des risques en assurance non-vie, une approche par modèle d’apprentissage statistique”, *Mémoire d’actuariat EURIA*

VESIN A. [2006] “Construction d’un arbre de décision”, *Inserm U578 – OUTCOMEREA*

YOHANNES Y., HODDINOTT J. [1999] “Classification and Regression Trees: an introduction”, *Washington, DC: International Food Policy Research Institute*

Internet

Argus de l’assurance, http://www.argusdelassurance.com/institutions/complementaire-sante-le-desequilibre-financier-des-couvertures-d-entreprise-etude.80716

Assurance Maladie, http://www.ameli.fr/

Institut National de la Statistique et des Études Économiques, http://www.insee.fr/fr/

Section actuariat, http://freakonometrics.hypotheses.org/category/actuarial-science

Sécurité Sociale, http://www.securite-sociale.fr/

VIII. Annexes

ANNEXE-I

	Classe d'âge								
Modalités	0 - 24	25 - 29	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 et plus

Tableau 45. Modalités de la variable « Classe d'âge »

	Sexe	
Modalités	Homme	Femme

Tableau 46. Modalités de la variable « Sexe »

	Classe d'effectif			
Modalités	0 à 4	5 à 19	20 à 100	100 et plus

Tableau 47. Modalités de la variable « Classe d'effectif »

	Régime		
Modalités	Général	Alsace Moselle	TNS

Tableau 48. Modalités de la variable « Régime »

	Base		
Modalités	Basse	Moyenne	Haute

Tableau 49. Modalités de la variable « Base »

	Surcomplémentaire			
Modalités	Nulle	Basse	Moyenne	Haute

Tableau 50. Modalités de la variable « Surcomplémentaire »

	Situation familiale	
Modalités	Seul	En couple

Tableau 51. Modalités de la variable « Situation familiale »

	Secteur d'activité		
Modalités	Boulangerie	Patisserie	AFFLEC

Tableau 52. Modalités de la variable « Secteur d'activité »

	Statut professionnel				
Modalités	EVIN	Ensemble du personnel	Cadre	Non cadre	TNS

Tableau 53. Modalités de la variable « Statut professionnel »

	Zonier			
Modalités	Autre	Région parisienne	Proche de Paris ou de Lyon	Grand sud

Tableau 54. Modalités de la variable « Zonier »

Les départements sont classés dans chaque zone tel quel :

- Région parisienne : 75,77,78,91,92,93,94,95
- Proche de Paris ou de Lyon : 50,14,76,60,02,10,52,55,70,89,45,28,41,36, 18,61,27,71,01,42,38,26,07
- Grand sud : 69,73,06,83,13,84,30,34,31,64,33
- Autre : le reste des départements

ANNEXE-II

Les coefficients obtenus pour le croisement âge * Base * Surcomplémentaire et sexe*âge sont regroupés dans les tableaux qui suivent :

Frais réels

	Age	Base	Surcomplémentaire	Coefficient	Borne inf	Borne sup
age*Base*Surcomplémentaire	60 ans +	Moyenne	Nulle	1,53603	1,49429	1,57895
age*Base*Surcomplémentaire	60 ans +	Moyenne	Moyen	1,75686	1,70922	1,80583
age*Base*Surcomplémentaire	60 ans +	Moyenne	Haute	1,84963	1,76334	1,94014
age*Base*Surcomplémentaire	60 ans +	Moyenne	Basse	1,47531	1,41530	1,53787
age*Base*Surcomplémentaire	60 ans +	Haute	Nulle	1,70074	1,65785	1,74474
age*Base*Surcomplémentaire	60 ans +	Haute	Basse	1,71909	1,67456	1,76481
age*Base*Surcomplémentaire	55-59 ans	Moyenne	Nulle	1,56927	1,53022	1,60931
age*Base*Surcomplémentaire	55-59 ans	Moyenne	Moyen	1,71875	1,67787	1,76062
age*Base*Surcomplémentaire	55-59 ans	Moyenne	Haute	1,88724	1,82121	1,95567
age*Base*Surcomplémentaire	55-59 ans	Moyenne	Basse	1,52254	1,47754	1,56890
age*Base*Surcomplémentaire	55-59 ans	Haute	Nulle	1,61586	1,57750	1,65516
age*Base*Surcomplémentaire	55-59 ans	Haute	Basse	1,69494	1,65604	1,73477
age*Base*Surcomplémentaire	50-54 ans	Moyenne	Nulle	1,48859	1,45234	1,52574
age*Base*Surcomplémentaire	50-54 ans	Moyenne	Moyen	1,67165	1,63221	1,71204
age*Base*Surcomplémentaire	50-54 ans	Moyenne	Haute	1,75819	1,69430	1,82450
age*Base*Surcomplémentaire	50-54 ans	Moyenne	Basse	1,44926	1,40832	1,49139
age*Base*Surcomplémentaire	50-54 ans	Haute	Nulle	1,53590	1,50004	1,57262
age*Base*Surcomplémentaire	50-54 ans	Haute	Basse	1,59821	1,56155	1,63572
age*Base*Surcomplémentaire	45-49 ans	Moyenne	Nulle	1,39339	1,35944	1,42819
age*Base*Surcomplémentaire	45-49 ans	Moyenne	Moyen	1,51169	1,47493	1,54936
age*Base*Surcomplémentaire	45-49 ans	Moyenne	Haute	1,67337	1,61583	1,73297
age*Base*Surcomplémentaire	45-49 ans	Moyenne	Basse	1,33387	1,29607	1,37278
age*Base*Surcomplémentaire	45-49 ans	Haute	Nulle	1,46846	1,43425	1,50349
age*Base*Surcomplémentaire	45-49 ans	Haute	Basse	1,47834	1,44421	1,51328
age*Base*Surcomplémentaire	40-44 ans	Moyenne	Nulle	1,17836	1,14778	1,20975
age*Base*Surcomplémentaire	40-44 ans	Moyenne	Moyen	1,29441	1,25901	1,33080
age*Base*Surcomplémentaire	40-44 ans	Moyenne	Haute	1,45531	1,38129	1,53328
age*Base*Surcomplémentaire	40-44 ans	Moyenne	Basse	1,13230	1,09042	1,17578
age*Base*Surcomplémentaire	40-44 ans	Haute	Nulle	1,25917	1,22882	1,29026
age*Base*Surcomplémentaire	40-44 ans	Haute	Basse	1,31988	1,28754	1,35304
age*Base*Surcomplémentaire	35-39 ans	Moyenne	Nulle	1,00000	0,98101	1,03819
age*Base*Surcomplémentaire	35-39 ans	Moyenne	Moyen	1,13645	1,10095	1,17309
age*Base*Surcomplémentaire	35-39 ans	Moyenne	Haute	1,19566	1,11234	1,28523
age*Base*Surcomplémentaire	35-39 ans	Moyenne	Basse	1,00000	0,98350	1,07452
age*Base*Surcomplémentaire	35-39 ans	Haute	Nulle	1,00000	0,99992	1,05376
age*Base*Surcomplémentaire	35-39 ans	Haute	Basse	1,07386	1,04319	1,10542
age*Base*Surcomplémentaire	30-34 ans	Moyenne	Nulle	1,00000	0,97258	1,02885
age*Base*Surcomplémentaire	30-34 ans	Moyenne	Moyen	1,11074	1,07559	1,14705

age*Base*Surcomplémentaire	30-34 ans	Moyenne	Haute	1,22522	1,15230	1,30276
age*Base*Surcomplémentaire	30-34 ans	Moyenne	Basse	1,00000	0,92595	1,01175
age*Base*Surcomplémentaire	30-34 ans	Haute	Nulle	1,00000	0,96738	1,01929
age*Base*Surcomplémentaire	30-34 ans	Haute	Basse	1,06212	1,03116	1,09400
age*Base*Surcomplémentaire	25-29 ans	Moyenne	Nulle	0,96841	0,94181	0,99575
age*Base*Surcomplémentaire	25-29 ans	Moyenne	Moyen	1,07624	1,03987	1,11390
age*Base*Surcomplémentaire	25-29 ans	Moyenne	Haute	1,27952	1,20688	1,35654
age*Base*Surcomplémentaire	25-29 ans	Moyenne	Basse	0,94346	0,90255	0,98623
age*Base*Surcomplémentaire	25-29 ans	Haute	Nulle	1,00000	0,96286	1,01277
age*Base*Surcomplémentaire	25-29 ans	Haute	Basse	1,04130	1,01160	1,07186
age*Base*Surcomplémentaire	0-24 ans	Moyenne	Nulle	0,95655	0,92953	0,98435
age*Base*Surcomplémentaire	0-24 ans	Moyenne	Moyen	1,07079	1,02468	1,11898
age*Base*Surcomplémentaire	0-24 ans	Moyenne	Haute	1,00000	0,95342	1,17627
age*Base*Surcomplémentaire	0-24 ans	Moyenne	Basse	0,93252	0,88066	0,98744
age*Base*Surcomplémentaire	0-24 ans	Haute	Nulle	0,95511	0,93180	0,97900
age*Base*Surcomplémentaire	0-24 ans	Haute	Basse	1,00000	1,00000	1,00000

Tableau 55. Coefficients obtenus par MLG sur les frais réels pour le croisement âge*base*surcomplémentaire

Fréquence

	Sexe	**Age**	**Coefficient**	**Borne inf**	**Borne sup**
Sexe*âge	60 ans +	M	1,539499	1,413187	1,677101
Sexe*âge	60 ans +	F	1,835681	1,703436	1,978192
Sexe*âge	55-59 ans	M	1,434817	1,346215	1,529251
Sexe*âge	55-59 ans	F	1,653430	1,567800	1,743737
Sexe*âge	50-54 ans	M	1,364103	1,285061	1,448006
Sexe*âge	50-54 ans	F	1,754746	1,671090	1,842590
Sexe*âge	45-49 ans	M	1,313891	1,241218	1,390819
Sexe*âge	45-49 ans	F	1,722958	1,643036	1,806769
Sexe*âge	40-44 ans	M	0,922456	0,870449	0,977570
Sexe*âge	40-44 ans	F	1,304272	1,240387	1,371447
Sexe*âge	35-39 ans	M	0,754728	0,710065	0,802200
Sexe*âge	35-39 ans	F	1,083939	1,027388	1,143602
Sexe*âge	30-34 ans	M	0,804092	0,758376	0,852564
Sexe*âge	30-34 ans	F	1,140665	1,081311	1,203277
Sexe*âge	25-29 ans	M	0,736122	0,696389	0,778121
Sexe*âge	25-29 ans	F	1,150396	1,094061	1,209632
Sexe*âge	0-24 ans	M	0,634706	0,604588	0,666325
Sexe*âge	0-24 ans	F	1,000000	1,000000	1,000000

Tableau 56. Coefficients obtenus par MLG sur la fréquence pour le croisement âge*base*surcomplémentaire