Table des matières

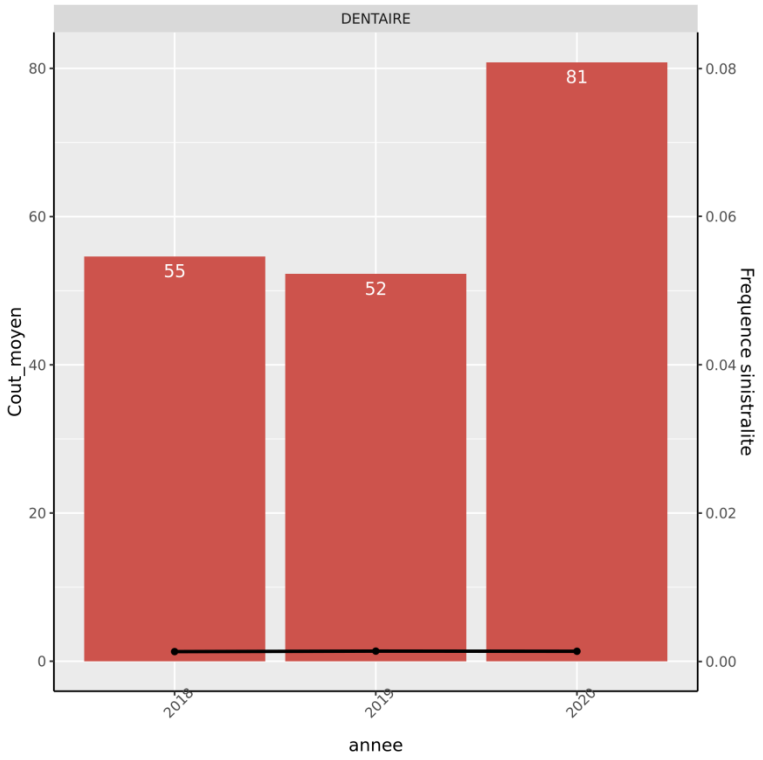
Synthèse

A travers ce mémoire, nous retraçons les différents travaux ayant mené à la mise en place d'un nouvel outil de tarification santé (utilisation des GLM) pour le service assurance de personnes de la Direction des Partenariats. Ces différents travaux prennent en compte les différentes évolutions réglementaires qui ont eu lieu dans le secteur de l'assurance santé. Il s'agit entre autres du 100% santé et de la résiliation infra annuelle.

L'optique, le poste dentaire et les aides auditives sont les principaux soins concernés par la réforme du 100% santé. Si cette réforme a eu des impacts notables pour le dentaire et les aides auditives avec une hausse des taux de recours à ces soins et aux équipements associés, l'impact est moins perceptible pour l'optique.

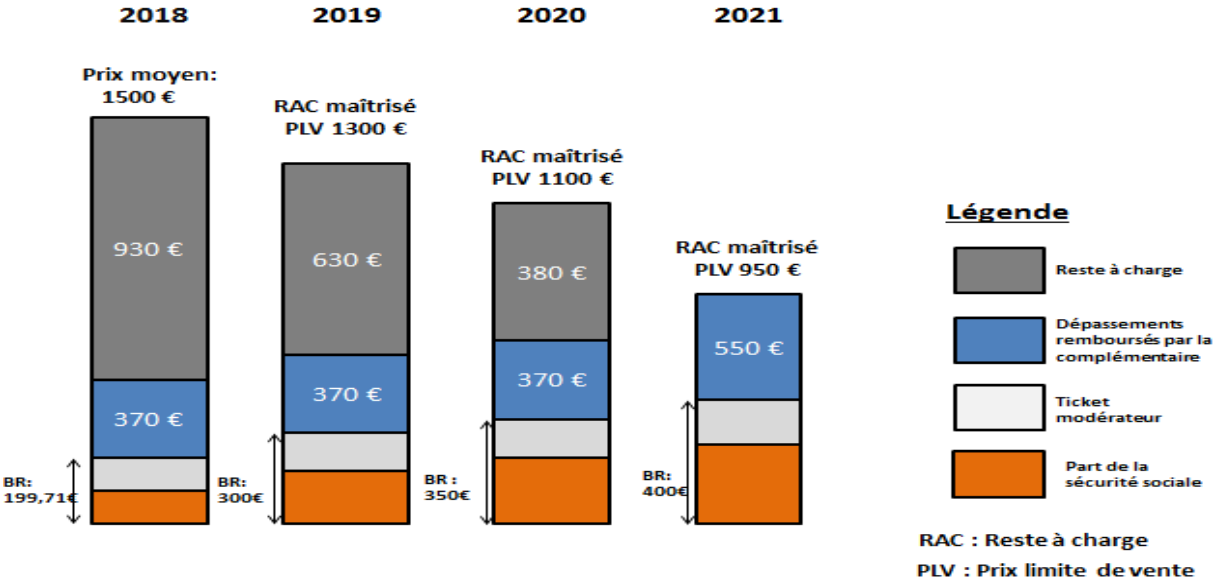
La première partie de nos travaux a consisté en la mise en place d'une base de données dite « as if » qui va intégrer les impacts liés notamment au 100% santé. Cette base retrace la sinistralité entre 2018 et 2020 d'un partenaire.

Concernant le dentaire ou la mise en place effective a été en 2020 on note un net impact sur le coût moyen. Les études menées sur notre portefeuille indiquent un impact plus important en 2021. On a donc appliqué une majoration des coûts moyens sur les trois années d'étude.

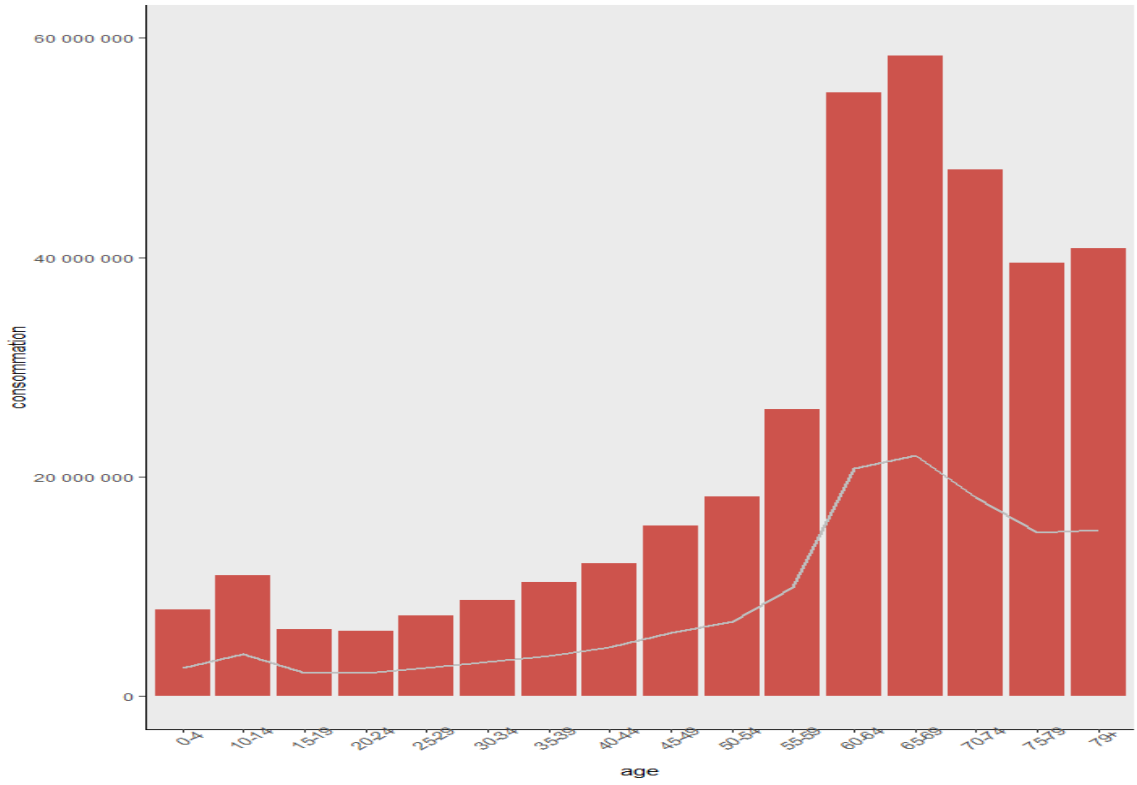


ADOVOEKPE Pierre-Marie
ISUP 2021

Des majorations ont également été opérées pour le coût moyen et la fréquence des sinistres ayant rapport aux aides auditives. On note ainsi une augmentation progressive des dépenses des complémentaires depuis la mise en place du 100% santé.



La base « as if » ainsi constituée, des retraitements ont été effectuées sur des variables comme les différentes dates, l'âge et le régime d'affiliation des assurés. De nouvelles variables telles que l'exposition, les niveaux de garantie homogénéisés, la fréquence et le coût moyen ont également été ajoutées. L'âge de l'assuré, le régime d'affiliation, le niveau de garanties et la zone sont les variables retenues pour la tarification. Ces variables ont en effet une grande influence sur la consommation des assurés. On remarque ainsi à titre illustratif une consommation qui augmente avec l'âge des assurés :



Pour réaliser la tarification, nous avons décidé de réaliser des tarifs par poste de garanties la consommation pouvant être différentes suivant le poste. Rappelons que nous disposons de sept postes de garanties que sont l'hospitalisation, les soins courants, la pharmacie, l'optique, le dentaire, l'appareillage et le poste prévention et bien-être.

L'utilisation des GLM pour la tarification suppose une indépendance entre le coût moyen et la fréquence. Les coefficients de spearman, de Kendall et de Pearson confirment cette hypothèse quel que soit le poste de garanties.

	Hospitalisation	Soins courants	Pharmacie	Optique	Dentaire	Appareillages	Prévention et bien-être
Spearman	0,159	-0,002	0,022	-0,041	0,033	0,075	-0,069
Kendall	0,112	-0,001	0,033	-0,03	0,022	0,053	-0,044
Pearson	0,069	-0,015	-0,005	-0,024	0,029	0,018	-0,053

La modélisation de la fréquence de certains postes a nécessité l'utilisation des Zero inflated GLM. Il s'agit de l'hospitalisation, l'optique, du dentaire et du poste prévention et bien-être. En effet ces postes présentent une part importante d'assuré n'ayant pas eu de sinistres :

	Hospitalisation	Soins courants	Pharmacie	Optique	Dentaire	Appareillages	Prévention et bien-être
Pourcentage de zéros	70%	17%	9%	70%	71%	55%	85%

Pour la modélisation du coût moyen les lois gamma, log normale et inverse gaussienne ont été testées afin de déterminer la plus adéquate selon le poste de garanties. Concernant la fréquence la loi binomiale négative et poisson ont été testées. Le choix du meilleur modèle se fait grâce à une analyse graphique, des tests d'adéquation, l'analyse de l'AIC et de la déviance (on choisit le modèle avec le plus petit AIC et la plus petite déviance).

Modèles sélectionnés pour le coût moyen :

HOSPITALISATION	Inverse gaussienne
SOINS COURANTS	Gamma
PHARMACIE	Gamma
OPTIQUE	Gamma
DENTAIRE	Log normale
APPAREILLAGES	Log normale
PREVENTION ET BIEN-ETRE	Gamma

La loi binomiale négative a été sélectionnée pour la modélisation de la fréquence quel que soit le poste de garanties. Les données présentaient en effet une sur dispersion ce qui explique un rejet systématique de la loi poisson.

Une fois les modèles choisis et la tarification réalisée, nous étudions la cohérence des primes issues de nos GLM. Le tarif obtenu augmente bien avec l'âge et le niveau de garanties. En ce qui concerne la zone, deux zoniers avaient été utilisés pour la tarification. Un de ces zoniers présentait des incohérences. En effet la zone 3 de ce zonier était plus chère que sa zone 2. La zone 1 devrait être la plus chère de toutes les zones suivie de la zone 2, de la zone 3 et de la zone 4. Le second zonier ne présentait pas d'incohérence et a été conservé pour la tarification.

Afin d'obtenir des tarifs pouvant être commercialisés quelques ajustements ont été faits sur les tarifs issus des GLM. Pour de réaliser nos GLM, nous avons créé des classes d'âge de 5 ans. Afin d'obtenir un tarif par âge, quelques ajustements ont donc dû être opérés. Dans un premier temps une prime unique du nouveau produit issu du GLM et une prime unique du produit de référence (prime déjà commercialisée) ont été calculées. La méthodologie de calcul est la suivante :

$$= \frac{prime_{0-4} * effectif_{0-4} + prime_{5-9} * effectif_{5-9} + \cdots + prime_{79-85} * effectif_{79-85}}{effectif_{0-85}}$$

Dans un second temps, on applique un coefficient à la prime unique du produit issu du GLM. Ce coefficient représente l'écart entre la prime pure du produit de référence à un âge i et la prime pure unique de ce même produit. Soit α_i ce coefficient. Si i est égale à 60 ans par exemple

$$\alpha_{60} = \frac{Prime\ pure\ produit\ de\ référence_{60\ ans}}{prime\ pure\ unique\ produit\ de\ réfrence}$$

La prime du nouveau produit pour un assuré de 60 ans est donc :

$$Prime\ pure = \alpha_{60} * prime\ pure\ unique\ du\ nouveau\ produit$$

On considère ainsi qu'à un âge i l'écart entre la prime pure du produit de référence à cet âge et la prime pure unique de ce même produit serait le même que l'écart la prime pure au même âge du nouveau produit et sa prime pure unique. Le coefficient α_i étant croissant avec l'âge, cela nous assure une prime pure croissante avec l'âge ainsi qu'un effet âge (écart entre les tarifs d'âges qui se suivent) similaire à celui du produit de référence.

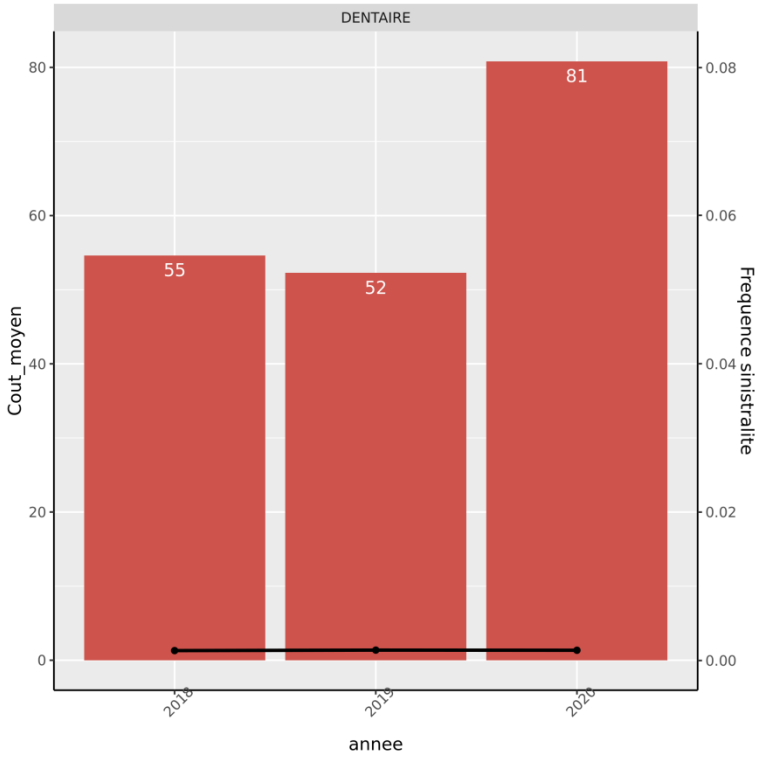
Summary

In this report, we retrace the various tasks that led to the implementation of a new health pricing tool (using GLM) for the personal insurance department of the Partnerships Department. This work takes into account the various regulatory changes that have taken place in the health insurance sector. These include 100% health insurance and intra-annual termination.

Optical, dental and hearing aids are the main treatments concerned by the 100% health reform. While this reform has had a significant impact on dental care and hearing aids, with an increase in the rate of use of these services and associated equipment, the impact is less noticeable for optical care.

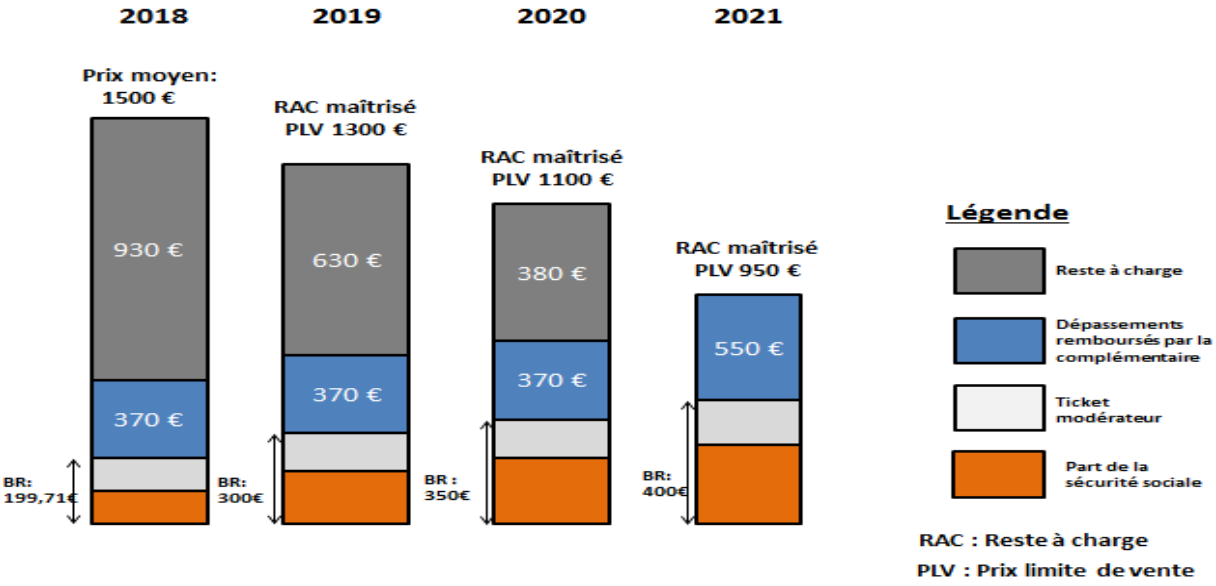
The first part of our work consisted in setting up an "as if" database that will integrate the impacts related to 100% health care in particular. This database tracks the claims experience of a partner between 2018 and 2020.

In the case of dental care, where the effective implementation was in 2020, there is a clear impact on the average cost. Studies conducted on our portfolio indicate a greater impact in 2021. We have therefore applied an increase in average costs over the three years of the study.

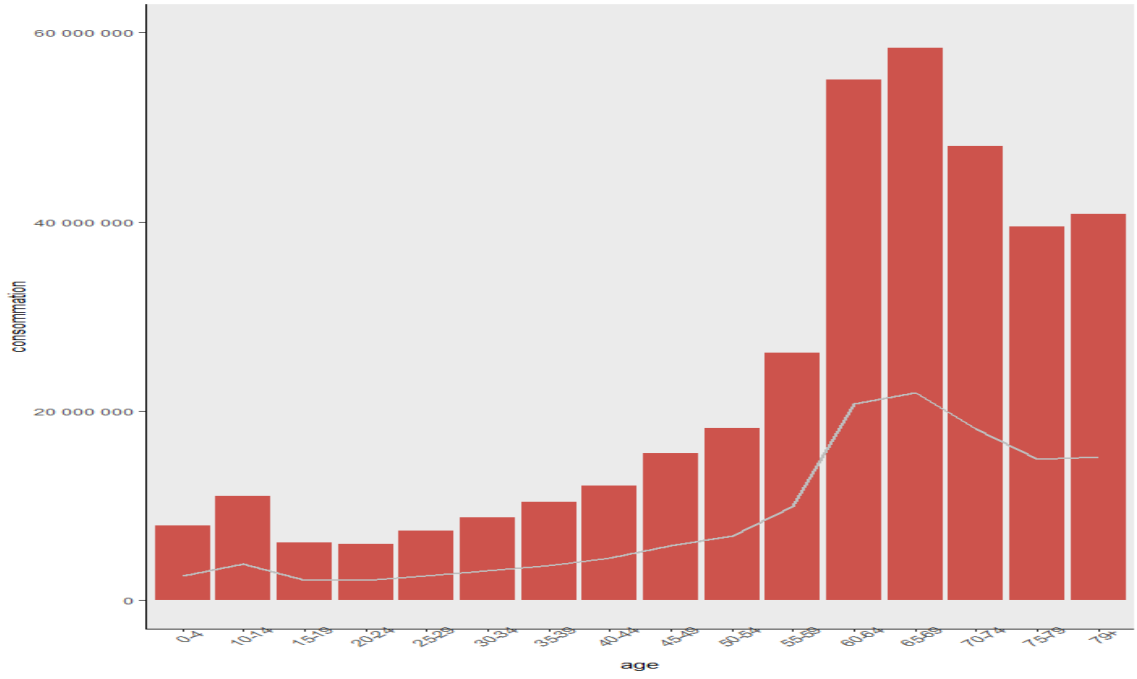


Increases have also been made in the average cost and frequency of claims related to hearing aids. There has been a gradual increase in the expenses of complementary health insurance since the

introduction of 100% health care.



The "as if" base was thus constituted, and adjustments were made to variables such as the various dates, ages and affiliation schemes of the insured. New variables such as exposure, homogenized coverage levels, frequency and average cost were also added. The age of the insured, the affiliation plan, the level of coverage and the zone are the variables retained for the pricing. These variables have a major influence on the consumption of the insured. For example, consumption increases with the age of the insured:



In order to carry out the pricing, we have decided to carry out tariffs by guarantee item, as consumption can be different according to the item. Let's remember that we have seven benefit items, which are hospitalization, routine care, pharmacy, optical, dental, appliances and the prevention and well-being item.

The use of GLMs for pricing assumes independence between average cost and frequency. Spearman's, Kendall's and Pearson's coefficients confirm this assumption regardless of the benefit item.

	Hospitalisation	Soins courants	Pharmacie	Optique	Dentaire	Appareillages	Prévention et bien-être
Spearman	0,159	-0,002	0,022	-0,041	0,033	0,075	-0,069
Kendall	0,112	-0,001	0,033	-0,03	0,022	0,053	-0,044
Pearson	0,069	-0,015	-0,005	-0,024	0,029	0,018	-0,053

The modeling of the frequency of certain items required the use of Zero inflated GLM. These are hospitalization, optical, dental and prevention and well-being. Indeed, these items present a significant proportion of insureds who have not had any claims:

	Hospitalisation	Soins courants	Pharmacie	Optique	Dentaire	Appareillages	Prévention et bien-être
Pourcentage de zéros	70%	17%	9%	70%	71%	55%	85%

For the modeling of the average cost, the gamma, log normal and inverse Gaussian distributions were tested in order to determine the most appropriate one according to the warranty item. For the frequency, the negative binomial and poisson distributions were tested. The choice of the best model is made through graphical analysis, goodness-of-fit tests, AIC and deviance analysis (the model with the smallest AIC and the smallest deviance is chosen).

Models selected for average cost:

HOSPITALISATION	Inverse gaussienne
SOINS COURANTS	Gamma
PHARMACIE	Gamma
OPTIQUE	Gamma
DENTAIRE	Log normale
APPAREILLAGES	Log normale
PREVENTION ET BIEN-ETRE	Gamma

The negative binomial distribution was selected for frequency modeling regardless of the guarantee item. The data presented an over dispersion which explains the systematic rejection of the poisson distribution.

Once the models have been chosen and the pricing performed, we study the consistency of the premiums obtained from our GLM. The rate obtained increases well with age and the level of coverage. As far as the zone is concerned, two zoners were used for the pricing. One of these zones was inconsistent. Zone 3 of this zone was more expensive than zone 2. Zone 1 should be the most expensive of all zones, followed by zone 2, zone 3 and zone 4. The second zoner did not have any inconsistencies and was retained for pricing.

In order to obtain marketable rates, some adjustments were made to the GLM rates. To achieve our GLMs, we created 5-year age groups. In order to obtain a rate by age, some adjustments had to be made. First, a single premium for the new GLM product and a single premium for the reference product (premium already marketed) were calculated. The calculation methodology is as follows:

$$= \frac{premium_{0-4} * size_{0-4} + premium_{5-9} * size_{5-9} + \cdots + premium_{79-85} * size_{79-85}}{size_{0-85}}$$

In a second step, a coefficient is applied to the single premium of the product from the GLM. This coefficient represents the difference between the pure premium of the reference product at an age i and the single pure premium of this same product. Let α_i be this coefficient. If i is equal to 60 years for example

$$\alpha_{60} = \frac{Pure\ premium\ of\ the\ reference\ product_{60\ ans}}{single\ pure\ \ premium\ of\ the\ reference\ product}$$

The premium for the new product for a 60 year old insured is therefore:

Pure premium=α_60*single pure premium of the new product

It is thus considered that at an age i the difference between the pure premium of the reference product at this age and the single pure premium of this same product would be the same as the difference between the pure premium at the same age of the new product and its single pure premium. Since the coefficient α_i is increasing with age, this assures us of an increasing pure premium with age as well as an age effect (difference between rates at successive ages) similar to that of the reference product.

REMERCIEMENTS

Je tiens à remercier Madame Aicha SOUKI, manager du service Assurance de Personnes, Monsieur François HATAB, mon tuteur en entreprise, qui ont suivi et dirigé mes travaux d'étude pendant ce mémoire. Je tiens également à remercier mes collègues de la branche santé individuelle, Monsieur Arnaud MEBALE et Madame Karishma SWEENARAIN, qui ont été très disponible et m'ont aidé lors de la réalisation de mes différents travaux.

Je remercie également Monsieur Chérif AMADOU SOW du service modèle et innovation pour ses différents conseils pour la réalisation de ce mémoire.

J'exprime également ma gratitude à Monsieur Olivier LOPEZ, mon tuteur pédagogique.

INTRODUCTION

L'assurance santé en France est un secteur en constante évolution. Il a connu de nombreuses réformes tels que récemment le 100% santé ou la résiliation infra-annuelle. Il est donc important pour les assureurs de s'adapter à ces réformes, notamment en termes de tarifs, afin de s'assurer de leur solvabilité et faire face à un marché très concurrentiel.

L'objectif de ce mémoire est donc la mise en place d'un outil de tarification santé pour le service assurance de personnes de la Direction des Partenariats de GENERALI. Il remplacera l'ancien outil qui présente quelques défauts.

Ce mémoire comportera quatre parties. La première présentera le marché de l'assurance santé en France et déroulera les différentes réformes récemment mises en place. La deuxième mettra en exergue le processus de création de la base de données utilisée pour notre tarification, les différents retraitements faits sur cette base pour tenir comptes des nouvelles réformes et l'influence des variables tarifaires sur le comportement des assurés à travers des statistiques descriptives. La troisième partie quant à elle détaillera les différents modèles utilisés pour notre tarification et montrera la cohérence des tarifs issus de ces modèles. Enfin, la dernière partie permettra de juger si les tarifs issus de nos modèles sont compatibles avec le business des Partenariats.

Partie 1 – La Complémentaire santé en France

1.1. L'assurance Santé en France

En France, le remboursement des frais de santé est pris en charge par différents organismes. La plus grosse part (78%) est prise en charge par la Sécurité Sociale (Régime Obligatoire). En second niveau interviennent des complémentaires santé pour 13% de la consommation de soins et de biens médicaux. Enfin, une part de 5% reste à charge des ménages.

a. La Sécurité Sociale

L'Assurance Maladie de la Sécurité Sociale est le régime de base obligatoire. Chaque individu dépend d'un régime différent selon sa situation professionnelle :

- **Le Régime générale (RG)** prend en charge la majorité de la population, les travailleurs salariés ainsi que les travailleurs indépendants (**travailleurs non-salariés**) depuis le 1er janvier 2018 ainsi que toute personne bénéficiant de droits au titre de la résidence.
- **Le Régime agricole (AGRI)** prend en charge les exploitants et salariés agricoles
- **Le Régime Alsace Moselle (ALS) :** Le régime de maladie en Alsace-Moselle est un régime bien particulier de la Sécurité Sociale. L'Alsace et la Moselle gèrent leur assurance maladie de manière autonome. Celle-ci intervient en complément du régime général des salariés. Toute personne sous ce régime doit se soumettre à des cotisations plus élevées (1,80% du salaire brut) du bénéficiaire que dans les autres départements. Le montant des remboursements est quant à lui plus intéressant. Les complémentaires interviennent donc en troisième niveau et remboursent comparativement moins qu'avec les autres régimes.
- **Les Autres régimes** : il s'agit de nombreux régimes spéciaux, comme celui des marins, des mines, de la SNCF, de la RATP, du secteur de l'énergie, de l'Assemblée nationale, du Sénat, des clercs et employés de notaires.

L'Assurance Maladie rembourse des actes tels que les consultations, les médicaments, les soins et prothèses dentaires, l'optique, l'audition, l'hospitalisation, la chirurgie. Ces actes sont référencés par un code acte selon la classification commune des actes médicaux et appartiennent à des postes de garanties bien précis.

b. La complémentaire santé

En cas de dépenses de santé, la Sécurité Sociale ne rembourse pas la totalité de la dépense. La complémentaire santé complète tout ou partie des remboursements. Elle peut aussi prendre en charge des prestations qui ne sont pas du tout remboursées par l'Assurance maladie (par exemple l'ostéopathie ou certains vaccins) et proposer des services associés (assistance, prévention, etc.).

La complémentaire santé peut être collective, c'est-à-dire proposée par une entreprise à ses salariés, ou individuelle auquel cas l'assuré souscrit directement son contrat auprès d'organismes tels que les sociétés d'assurance, les institutions de prévoyance ou les mutuelles.

Ce mémoire portera exclusivement sur l'assurance santé individuelle, la Direction des Partenariats ne proposant pas de contrats collectifs.

c. Remboursements des frais de santé

L'Assurance Maladie calcule les remboursements à partir d'une base de remboursement exprimée en euros. Elle rembourse un pourcentage de cette base (taux de remboursement) et déduit une franchise non remboursable pour certaines dépenses telles que les consultations, les boîtes de médicaments. Le pourcentage restant est appelé ticket modérateur et est à la charge de l'assuré. C'est là qu'intervient la complémentaire santé : elle rembourse ce ticket modérateur.

Les tarifs ou honoraires de certains professionnels de santé sont parfois supérieurs à la base de remboursement de la sécurité sociale. On parle alors de dépassements d'honoraires. La complémentaire santé peut prendre en charge partiellement ou totalement les dépassements d'honoraires en plus du ticket modérateur. Cela dépend des garanties prévues au contrat.

Les garanties d'un contrat peuvent être libellées :

- En pourcentage de la base de remboursement (BR) – Exemple : 125%BR

- En montant en euros ou forfait– Exemple : 100€

- En pourcentage de frais réels (FR)– Exemple : 100%FR

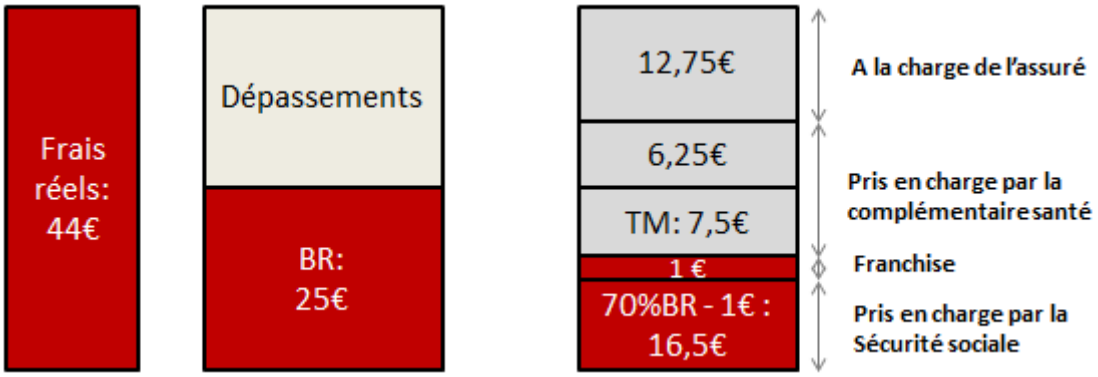
QUELQUES EXEMPLES

- **CONSULTATION CHEZ UN MEDECIN GENERALISTE**

Prenons l'exemple d'une consultation chez un médecin généraliste qui coûte 44€ à l'assuré.

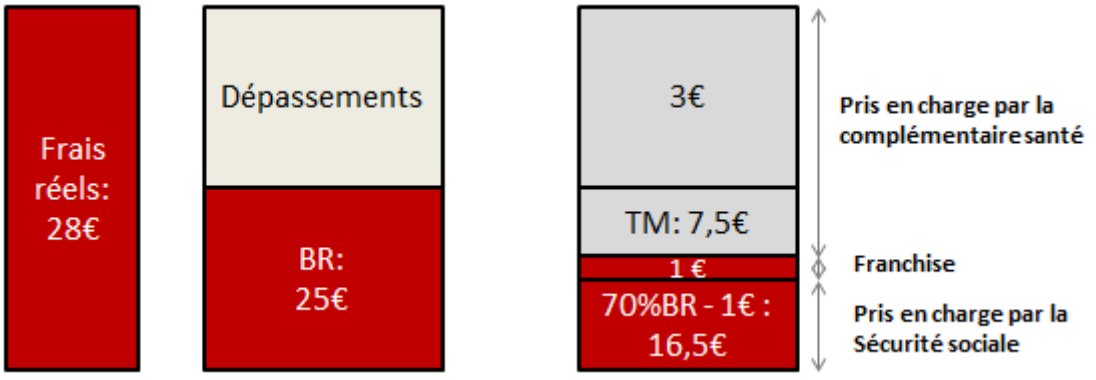
Sachant que la base de remboursement de cet acte est de 25€ et que la sécurité sociale prend en charge 70% de cette base de remboursement, si le niveau de garantie de la complémentaire santé est de 125% BR on a :

- La sécurité sociale prend en charge 70%*25€ = 17,5€ - 1€ de franchise à la charge de l'assuré.
- La complémentaire santé prend en charge le minimum entre 125%BR-70%BR=13,75€ et 44€-70%BR=26,5€.
- Il reste donc à la charge de l'assuré 44€ - 16,5€ -13,75€ = 13,75€



Suivant ce même exemple, si la consultation avait coûté 28€ :

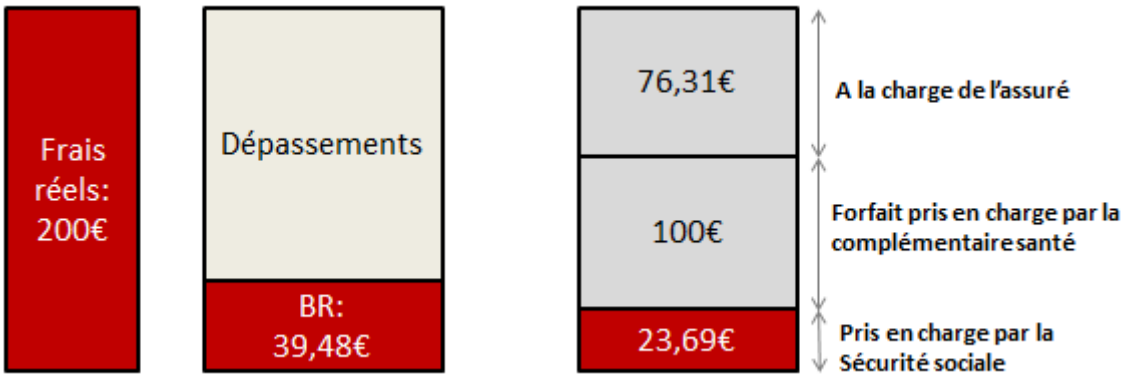
- La sécurité sociale prend en charge 70%*25€ = 17,5€ auxquels on retranche 1€ de franchise à la charge de l'assuré.
- La complémentaire santé prend en charge le minimum entre 125%BR-70%BR=13,75€ et 28€-70%BR=10,5€.
- Il reste à la charge de l'assuré 1€ de franchise



- **LENTILLES**

Les lentilles de contact sont remboursées à 60 % sur la base d'un forfait annuel, de date à date, par œil appareillé, fixé à 39,48 €.
Prenons l'exemple d'une lentille qui coûte 200€ à un assuré. En supposant que la complémentaire santé rembourse les lentilles par un forfait de 100€ par an :

- La sécurité sociale prend en charge 60%*39,48€ = 23,69€
- La complémentaire santé rembourse les 100€ de forfait
- Il reste à la charge de l'assuré 76,31€



d. Les différents acteurs et les modes de distribution

LES DIFFERENTS ACTEURS

Trois acteurs se partagent le marché de la complémentaire santé en France. Il s'agit des mutuelles, des institutions de prévoyance et des sociétés d'assurance.

- **Les mutuelles**

Elles sont régies par le code de la mutualité et sont des sociétés de personnes à but non lucratif appartenant à leurs assurés. Elles représentent le principal acteur du marché avec 73% des

organismes et 51% du chiffre d'affaires total. L'assurance santé individuelle est leur principal domaine d'activité (70% de leurs bénéficiaires et de leur chiffre d'affaires).

- **Les institutions de prévoyance**

Elles sont également à but non lucratif, soumises au code de la sécurité sociale, paritaires (co-dirigées par des syndicats de salariés et de patrons). Elles représentent 5% des organismes et 18% du chiffre d'affaires total. L'assurance santé collective (contrat souscrit par une entreprise pour ses salariés) est leur principal champ d'activité soit 88% de leurs bénéficiaires. La majorité des contrats santé individuels qu'ils ont en portefeuille est liée à d'anciens salariés des entreprises couvertes en santé collective.

- **Les sociétés d'assurance**

Elles sont régies par le code des assurances et sont à but lucratif. Elles représentent 22% des organismes du marché et 31% du chiffre d'affaires total. Leur activité est équilibrée entre le collectif (49% de leur chiffre d'affaires) et l'individuel (51% de leur chiffre d'affaires). Il est cependant à noter qu'elles gagnent du terrain en collectif avec un hausse du chiffre d'affaires des contrats collectifs de 22,3% entre 2015 et 2017.

LES MODES DE DISTRIBUTION

En France, les principaux canaux de distribution de l'assurance sont : les agents généraux, les courtiers, les banques, les réseaux salariés. Nous allons nous attarder sur les courtiers (à peu près 25 000 en France) qui représentent les principaux partenaires de la Direction des Partenariats.

Le courtier a un rôle d'intermédiaire entre les compagnies d'assurance et le client. Le courtier est mandaté par le client à la différence de l'agent général qui lui est mandaté par la compagnie : il a un rôle de conseil et doit lui trouver la meilleure solution du marché. Il négocie également avec les compagnies les meilleurs produits à proposer à ses clients.

On distingue deux types de courtiers : les courtiers grossistes et les courtiers directs. Les courtiers grossistes (avec lesquels nous travaillons le plus), participent à la conception des produits d'assurance, négocient les garanties et les tarifs auprès d'organismes assureurs qui portent le risque. Ils sont rémunérés par des commissions d'acquisition. Ils confient eux-mêmes la distribution des contrats à un réseau de courtiers partenaires qu'ils commissionnent également. Les courtiers directs remplissent les mêmes fonctions que les courtiers grossistes mais, contrairement à ces derniers, ils sont en contact sans intermédiaire avec les assurés.

Les ventes peuvent se faire :

- En face à face au domicile du client ou en boutique
- A distance via du démarchage téléphonique ou de la vente en ligne (sites internet dédiés, comparateurs)

1.2. Nouvelles règlementations et réformes en santé

a. Le 100% santé

La réforme 100% Santé, qui vise à lutter contre le renoncement aux soins, propose un ensemble de prestations d'équipements identifiés dans un panier spécifique pour trois postes de soins : audiologie, optique et dentaire. Ces paniers intègrent un large choix d'équipements de qualité qui seront pris en charge intégralement, sans frais supplémentaire à la charge de l'assuré.

Le 100 % Santé se déploie depuis 2019. En 2021, tout est en place pour les aides auditives et l'optique. Le dispositif pour le dentaire inclura encore plus de soins en 2022. Toute personne bénéficiant d'un contrat « responsable » par sa complémentaire santé peut bénéficier de l'offre 100 % Santé.

Qu'est-ce qu'un contrat responsable et solidaire ?

Une complémentaire santé responsable encourage le respect du parcours de soins coordonnés (recours à un médecin traitant qui peut orienter vers un spécialiste par la suite).

Comme mentionné sur le site Ameli, les complémentaires santé « responsables » qui couvrent 98 % des bénéficiaires des contrats complémentaires santé souscrits, remboursent au minimum :

* 30 % du tarif des consultations du médecin traitant (ou du médecin vers lequel il vous a orienté) dans le cadre du parcours de soins coordonnés,

* 30 % du tarif des médicaments remboursables à 65 % par l'assurance maladie obligatoire,

* 35 % du tarif des examens de biologie médicale prescrits par le médecin traitant (ou le médecin vers lequel il vous a orienté),

* le ticket modérateur d'au moins deux prestations de prévention fixées par la réglementation.

En revanche, elles ne remboursent pas :

* les dépassements et majorations liés au non-respect du parcours de soins ;

* la participation forfaitaire de 1 € applicable aux consultations et certains examens médicaux ;

* les franchises applicables sur les médicaments (exemple : 0,50 € par boîte de médicament), les actes paramédicaux et les frais de transport.

Une complémentaire santé est dite solidaire si l'assuré n'est pas soumis à un questionnaire médical à la souscription.

DEVIS PROPOSES PAR LES PRATICIENS

L'audioprothésiste ou l'opticien consultés proposent un devis : au moins un des équipements proposés doit être un équipement dit 100 % Santé.

Le dentiste consulté propose un devis avec un plan de traitement précisant les soins à réaliser. Dans le devis du dentiste doit figurer le panier 100 % Santé entièrement remboursé, si les soins à réaliser existent dans l'offre 100 % Santé.

Chacun reste libre de choisir son soin ou son équipement.

MISE EN PLACE DU 100% SANTE SUR LES POSTES CONCERNES

1. LES AIDES AUDITIVES

Depuis 2019, le reste à charge des assurés diminuent progressivement

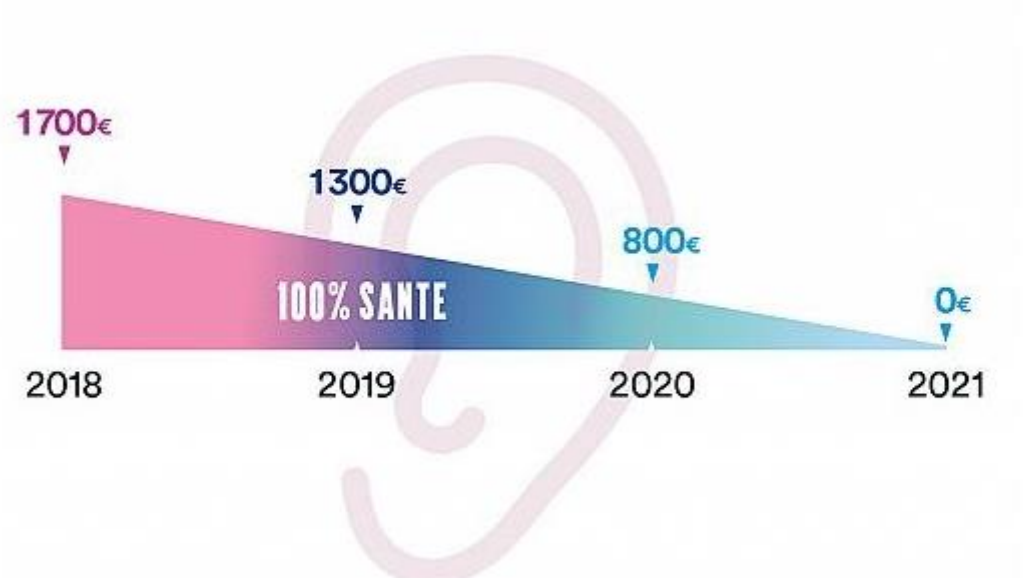


Figure 1 Evolution du reste à charge pour les aides auditives

Les aides auditives sont classées selon deux paniers : le panier 100% santé ou classe 1 et le panier hors 100% santé ou classe 2 qui n'est pas intégralement remboursé.

Concernant le panier 100% santé, la réforme semble pour l'instant portée ses fruits comme l'atteste ces chiffres publiés par les réseaux de soins et les données de l'assurance maladie. En effet :

- On note une hausse de 55% des remboursements de l'assurance maladie obligatoire sur les deux premiers mois de l'année par rapport à la même période l'année passée
- Le recours aux équipements 100% santé atteint les 30%, bien au-delà des attentes des pouvoirs publics

2. LE DENTAIRE

Trois paniers existent concernant l'offre dentaire :

- Un panier 100% santé
- Un panier avec reste à charge maîtrisé : où les prix sont plafonnés permettant ainsi au patient d'avoir un reste à charge limité qui peut être prise en charge par la complémentaire santé selon le niveau de garantie de l'assuré
- Un panier à tarif libre : le remboursement est totalement lié au niveau de garantie de l'assuré

L'offre 100% santé dentaire s'applique depuis le 1er janvier 2020 et permet d'avoir accès à un large choix de couronnes, de bridges, et de dentiers de qualité, totalement remboursés. Depuis le 1er janvier 2021, cette offre intègre des dentiers en résine et des prothèses mobiles.

On observe un fort recours au 100% santé. Ainsi le réseau de soins carte blanche partenaires remonte que les équipements 100% santé représentent 53% des prothèses vendues contre 32% en 2019. Une tendance baissière est observée sur les 2 autres paniers (-5% sur le panier modéré et -16% sur le panier à tarif libre par rapport à 2019).

Le recours au 100% santé dépend également de la position de la dent. Ainsi pour les dents de devant, on atteint un recours aux prothèses à 87% tandis qu'il ne dépasse pas les 16% pour les molaires invisibles.

3. L'OPTIQUE

Pour l'optique, les équipements (montures et verres) sont classés suivant deux paniers :

- Le panier 100% santé dit classe A , où les équipements sont pris en charge à 100%

- Le panier hors 100% santé ou classe B, où les remboursements dépendent du niveau de garantie

La prise en charge des montures est plafonnée à 100€. Un plafond existe également sur les verres en fonction de leur complexité.

Déployé depuis le 1er janvier 2020, le 100% santé en optique présente le taux de recours le plus faible des trois postes de santé concerné. Ainsi le ministre de la Santé Olivier Véran a annoncé un taux de diffusion des équipements du panier 100% santé d'environ 15,5 % (le rassemblement des opticiens de France annonce quant à lui un taux de recours de 17,6%).

b. Le COVID et la taxe de solidarité

LE COVID

La crise sanitaire liée au covid et les différents confinements qu'elle a engendrés, ont eu un impact sur la consommation en santé des Français :

- Pendant le premier confinement, on observe de ce fait une baisse de 60 à 70% des dépenses de santé par rapport à 2019 sur la même période. Cette baisse concerne globalement des postes comme l'optique et le dentaire et les soins jugés reportables.
- Après ce premier confinement, à partir de juin 2020, la reprise des dépenses de santé s'accélère et les soins reportés commencent à être pris en charge : on note une hausse de 6 à 10% des dépenses par rapport à l'été 2019.

Sur l'année 2020 les effets à la hausse et à la baisse devraient se compenser. Dans la suite de ce mémoire, nous allons nous poser la question de conserver ou non cette année atypique dans nos données. On analysera ainsi les impacts du covid sur le portefeuille utilisé pour cette étude.

LA TAXE DE SOLIDARITE

Cette taxe a été mise en place par le gouvernement pour les assureurs complémentaires santé pour compenser une partie du déficit de la sécurité sociale liée aux conséquences de la crise sanitaire. La sécurité sociale a en effet pris en charge à 100% certains actes tels que la téléconsultation et les tests de dépistage. La taxe de solidarité vise également à compenser les « profits » des complémentaires liés à la crise.

Cette taxe s'élève à 2,6% pour l'exercice 2020 et 1,3% en 2021 et sera assise sur les primes perçues par les complémentaires sur l'année 2020.

c. La résiliation infra annuelle

La résiliation infra annuelle est la possibilité pour l'assuré de pouvoir résilier son contrat sans condition après une année de souscription. Cette nouvelle réforme a pour objectif de rendre le marché des complémentaires plus ouvert et concurrentiel.

En effet, une plus grande volatilité des portefeuilles des assureurs est à prévoir obligeant ainsi ces derniers à innover et investir pour améliorer leur service et fidéliser leur client.

Sur le portefeuille santé, les effets de la réforme sont déjà palpables avec une réduction de l'écart entre la souscription d'un contrat et sa prise d'effet.

Avant la RIA (2018 ,2019) la majorité des contrats prenait effet soit très rapidement soit à peu près une année après la souscription (le temps que le précédent contrat de l'assuré qui dure 1 an prenne fin). Avec la RIA (2020) le pic est observé beaucoup plus tôt (aux alentours d'un mois d'écart).

On observe globalement que l'écart entre la date d'enregistrement et la prise d'effet du contrat est divisé par deux depuis l'entrée en vigueur de cette réforme.

D'un point de vue des résiliations aucun effet n'est pour le moment visible et le chiffres d'affaires ne se retrouve pour le moment pas impacté.

Partie 2 – Présentation des données

L'objectif de ce mémoire va être de réaliser un nouvel outil de tarification santé en se servant de GLM et en prenant en compte des impacts des nouvelles réformes de ce secteur de l'assurance. Pour ce faire nous nous servirons d'une base de données retraçant la sinistralité d'un partenaire de la Direction des Partenariats entre 2018U et 2020. Dans cette partie nous détaillons les différents retraitements que nous avons appliqués à cette base.

2.1. Présentation des produits GENERALI utilisés

Le portefeuille santé individuel de la Direction des Partenariats est composé de nombreux produits mis en place en collaboration avec des courtiers.

Dans le cadre de ce mémoire, la base de données utilisée sera constituée uniquement de produits issus d'un partenaire avec lequel nous avons mis en place une dizaine de prouits. On dispose d'énormément de données avec ces produits et ils représentent assez bien le portefeuille. De plus c'est l'un des seuls partenaires avec lequel la jointure entre les bases sinistre contrat et bénéficiaire a pu être réalisée. En effet, il était impossible d'obtenir pour les autres partenaires une clé de jointure entre les différentes bases.

Le produit le plus représenté dans la base va nous servir de produit de référence pour uniformiser les niveaux de garanties entre les différents produits. Ce point sera détaillé plus loin dans la partie consacrée à la construction de la base.

Structure tarifaire des produits

Quatre principales variables définissent le tarif des produits santé :

- **La formule :** elle correspond au niveau de garanties qui permet de déterminer le montant pris en charge par l'assureur. Plus la formule est élevée, plus le niveau de garanties est élevé et plus l'assuré a tendance à consommer.
- **L'âge :** les grilles tarifaires sont segmentées par âge et le tarif croît avec l'âge. En effet plus on vieillit plus on a tendance à avoir recours à des soins. La surreprésentation d'actes concernant des bénéficiaires entre 60 et 70 ans dans la base en témoigne.
- **La zone :** le tarif dépend de l'endroit où vit l'assuré, sa consommation médicale n'est pas la même d'une zone à une autre. En effet certaines zones peuvent présenter par exemple de fortes densités de praticiens ou des dépassements d'honoraires plus fréquents. On dispose de zoniers qui peuvent varier d'un partenaire à l'autre pour des raisons commerciales.
- **Le régime d'affiliation :** le régime général, agricole, TNS… Le tarif sorti correspond au régime général (concerne 4/5 des assurés) et des réductions sont appliquées pour déterminer les tarifs des autres régimes.

Des réductions sont également effectuées pour les couples et les familles.

Pour réaliser notre tarification, nous allons utiliser ces quatre variables comme variables explicatives. Concernant le régime d'affiliation, la variable comprendra les modalités régime générale et « hors régime générale » car les autres régimes sont sous représentés dans la base. De plus, des études réalisées sur notre portefeuille montrent que les réductions réalisées jusque-là sont cohérentes. Nous les conserverons donc pour déterminer le tarif des autres régimes.

2.2. Construction de la base de données

Pour réaliser notre tarification, nous utiliserons des bases de données fournies mensuellement par notre partenaire entre janvier 2018 et Décembre 2020. Il s'agira des bases contrats, des bases sinistres et des bases bénéficiaires.

- La base sinistre permet de retracer les différents paiements de sinistres effectués aux bénéficiaires. Elle renseigne sur l'acte, le poste de garanties concerné etc.

- La base contrat est l'image à l'instant t du portefeuille. Elle donne des informations sur les nombre de contrats souscrits et leur état (en cours, résilié, etc.) . La zone géographique, le niveau de garanties ainsi que les dates d'effet et de fin d'effet des contrats y également sont mentionnés.

- La base bénéficiaire qui donne des informations sur les bénéficiaires tels que : l'âge, la période de couverture et le rang. Le rang 10 correspond au souscripteur, le 20 correspond à son conjoint et les rangs 30,31,32… correspondent aux enfants.

La base finale nécessaire à l'étude a été fournie par un service transverse. Des variables comme le numéro de contrat et le rang du bénéficiaire ont été utilisé pour réaliser la jointure entre les trois différentes bases.

La construction de la base va suivre deux grandes étapes :

a. Retraitement de la base

La base fournie présentait quelques problèmes de qualité de données :

- **L'âge**

En parcourant la base de données on a pu observer des anomalies concernant l'âge des bénéficiaires. Des parents présentaient des âges anormalement bas tandis que dans certains cas les

enfants présentaient des âges anormalement élevés. De plus, il pouvait arriver qu'un même bénéficiaire dispose de plusieurs âges.

Pour résoudre ce problème nous avons dû refaire une jointure entre la base finale et la base bénéficiaire où sont stockés les âges.

Lorsque le problème de multiplicité d'âge pour un même bénéficiaire persistait nous avons conservé l'âge qui paraissait le plus cohérent compte tenu du rang de l'assuré.

- **Les dates**

Il s'agit ici des dates d'effet et de fin d'effet des contrats. Comme pour les âges, une jointure avec la base bénéficiaires a été faite pour régler des problèmes d'absence ou de multiplicité de dates pour un même bénéficiaire. Quand le problème de multiplicité persistait, on gardait la date la plus ancienne. Ce travail était primordial pour assurer la justesse de l'exposition.

Des variables ont également dû être ajoutées à la base pour réaliser la tarification :

- **L'exposition**

Elle correspond à la durée d'effet du contrat sur la période d'observation. Nous l'avons calculé de deux manières :

- En calculant l'écart entre la date d'effet et de fin d'effet du contrat lorsque nous avions ces deux variables renseignées.
- En additionnant les expositions par mois. En effet nous disposons de données mensuelles. Chaque bénéficiaire apparaît au moins une fois par mois (juste une fois quand il n'a pas de sinistre). Nous avons donc calculé les expositions par mois selon l'état du contrat (résilié, en cours etc .) :
  - Quand le contrat était en cours et que le mois de prise d'effet était différent du mois d'observation l'exposition du mois était égale à $\mathbf{30,5/365.}$
    Exemple : Date d'effet : 01/01/2018 , Mois d'observation : 01/2019.
  - Quand le contrat était en cours et que le mois de prise d'effet était le même que le mois d'observation l'exposition du mois était égale à $(\mathbf{30,5} - \boldsymbol{jour\ de\ prise\ d'effet})/\mathbf{365.}$
    Exemple : Date d'effet : 10/01/2018 , Mois d'observation : 01/2018.

  - Quand le contrat était résilié et que le mois de résiliation ne correspondait pas au mois d'observation, l'exposition sur le mois était de **0**.
    Exemple : Date de résiliation : 10/01/2018 , Mois d'observation : 01/2019.

- Quand le contrat était résilié et que le mois de résiliation correspondait au mois d'observation, l'exposition sur le mois était de $(\boldsymbol{jour\ de\ résiliation})/\mathbf{365}$.
  Exemple : Date de résiliation : 10/01/2018 , Mois d'observation : 01/2018.

- **Le niveau de garanties uniformisé**

Les différents actes de santé sont répartis par poste de garanties. Dans le cadre de ce mémoire nous cherchons à réaliser une tarification par poste, les manières de rembourser et la fréquence de sinistres pouvant être différentes d'un poste à l'autre. Nous recensons 7 principaux postes :

- **L'hospitalisation** : prend en compte des actes comme les frais de séjour, d'anesthésie, de chirurgie.

- **Les soins courants** : il s'agit de consultations de toute sorte, d'actes paramédicaux, d'imagerie médicale.

- **La pharmacie** : correspond aux frais liés aux médicaments, aux honoraires de dispensation.

- **Le dentaire :** est lié au remboursement de soins dentaires, prothèses dentaires, orthodontie.

- **L'optique :** frais concernant les montures, les verres, les lentilles de contact etc.

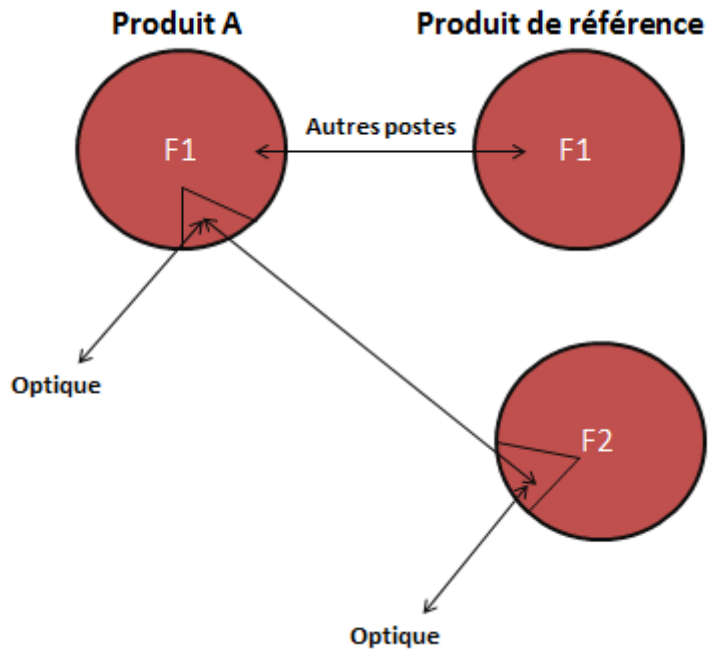
- **L'appareillage** : remboursement d'actes tels que : prothèses auditives, petit appareillage, accessoires, prothèses mammaires.

- **Prévention et bien-être :** regroupe des actes tels que la médecine douce, les cures thermales.

Les formules ou niveaux de garanties ne sont pas identiques d'un produit à l'autre. Ainsi la formule 1 d'un produit A peut avoir un niveau de remboursement proche de la formule 3 d'un produit B. Etant une variable tarifaire il est important d'uniformiser les formules de tous les produits en les regroupant par niveaux de garanties similaires. Nous avons réalisé ce travail en allant à la maille du poste de garantie au lieu de comparer les formules globalement entre elles.

Nous avons comparé tous les produits de la base de données à un produit de référence. Il s'agit du produit le plus représenté dans la base de données. Il possède 6 formules. Notre nouvelle variable disposera donc de 6 niveaux de garanties.

Attribution des niveaux



Dans cet exemple, pour un produit A de notre base, si le poste de garanties 'Optique' de sa formule 1 a un niveau de remboursement équivalent ou proche du poste de garanties 'Optique' de la formule 2 du produit de référence, On attribue à ce poste le 'NIVEAU 2'.

Concernant les autres postes, ils sont proches en termes de garanties de leur poste correspondant en formule 1 du produit de référence. On leur attribue donc le 'NIVEAU 1'.

- **Le nombre d'actes, fréquence et coût moyen**

Chaque ligne de la base sinistre correspond au remboursement d'un acte. Il peut arriver d'avoir des remboursements négatifs qui correspondent à des reprises de l'assureur après des remboursements qui ne devaient pas être effectués. Pour chaque remboursement négatif, il doit exister un remboursement positif.

Pour définir un acte, on va donc d'abord mettre en place une nouvelle base en sommant les remboursements suivant les variables : code_ acte, date de survenance du sinistre, zone etc. Une ligne de la nouvelle base sera considérée comme un acte si ladite somme de remboursements est supérieure à 0.

Calcul de la fréquence par bénéficiaire : $nombre\ d'actes\ du\ bénéficiaire/exposition$

Calcul du coût moyen par bénéficiaire : $coût/nombre\ d'actes\ du\ bénéficiaire$

La tarification se faisant par poste de garanties, nous divisons notre base de données suivant ces 7 postes. Le nombre d'actes, le coût, la fréquence et le coût moyen ont donc été calculé par poste de garanties.

b. Base de données « as if »

Afin que la tarification soit la plus juste possible, il est nécessaire de créer et d'effectuer notre étude une base dite « as if ». En effet depuis 2019 le 100% santé s'est mis progressivement en place. La manière de rembourser en 2021, pour les postes concernés, est différente des années d'observations retenues pour notre base.

DENTAIRE

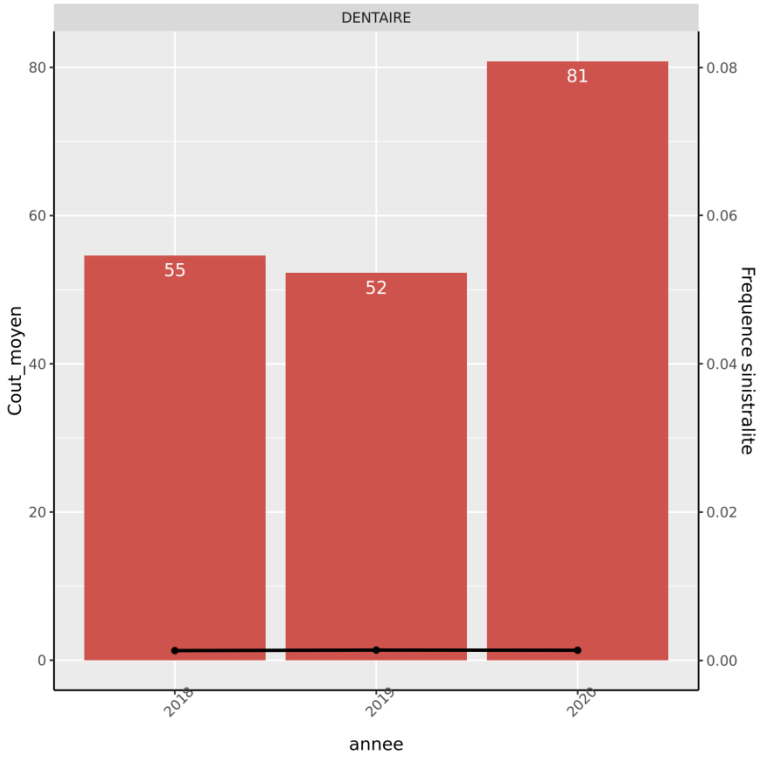


Figure 2 Evolution de la fréquence et du coût moyen pour le poste dentaire

Ainsi en 2020, année de mise en place du 100% en dentaire, le coût moyen est beaucoup plus élevé. Pour prendre en compte cet effet 100% santé, nous avons multiplié par 3 les coûts des années 2018 et 2019 et par 2 ceux de 2020. Ces chiffres ont été choisis compte tenu des coûts moyens observés sur notre portefeuille en 2021.

PROTHESES AUDITIVES

Concernant les prothèses, la réforme se met en place progressivement depuis 2019 avec une augmentation des bases de remboursement entraînant une baisse du reste à charge des assurés et une hausse de la prise en charge des complémentaires dans le même temps. Cette tendance s'observe dans notre base d'étude :

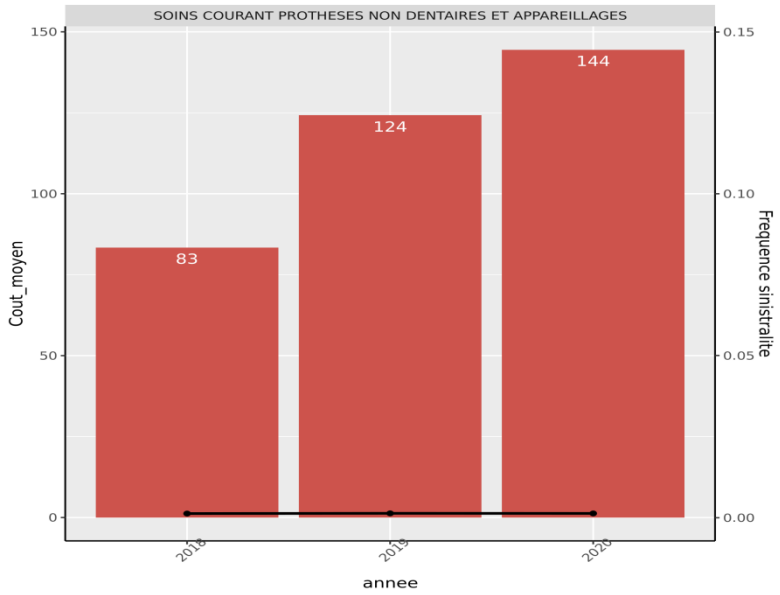


Figure 3 Evolution de la fréquence et du coût moyen pour les prothèses auditives

En 2021, le reste à charge des assurés sera de 0 sur le panier 100% santé des prothèses auditives.

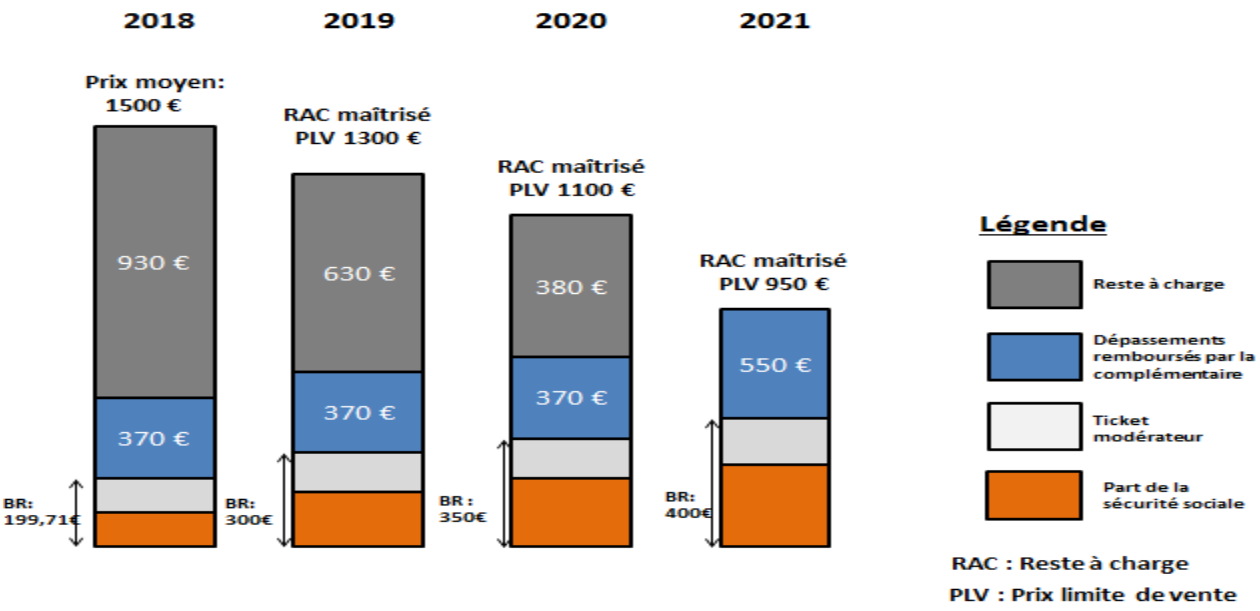


Figure 4 Evolution du reste à charge en audio prthèse

En partant de ce graphe on se rend compte que la part de la complémentaire santé était en moyenne de 450€ en 2018, 490€ en 2019, 510€ en 2020. En 2021 elle est de 710€. De plus les premiers chiffres remontés par les réseaux de soins en 2021 donnent un taux de recours à l'offre 100% santé d'à peu près un tiers.

Au vu de ces chiffres on applique aux coûts relatifs aux audioprothèses une majoration de 50% en 2018, 45% en 2019 et 40% en 2020. Ces majorations ont été effectuées à chaque fois sur un tiers des actes d'audioprothèses de l'année considérée. Sur le reste des actes d'audioprothèses, l'évolution de la base de remboursement a été appliquée aux coûts.

c. Effets covid

Nous avons mesuré les effets du covid sur la base de données afin de décider de conserver ou non l'année 2020 pour réaliser la tarification.

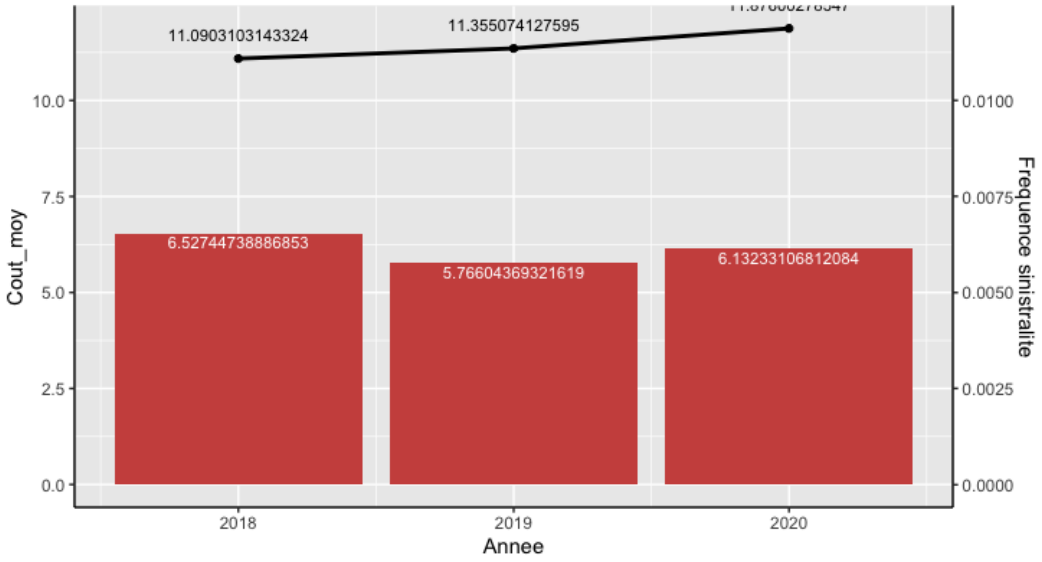


Figure 5 Evolution par année de la fréquence et du coût moyen

L'évolution du coût moyen et de la fréquence par année ne montre pas de dérive de consommation en 2020.

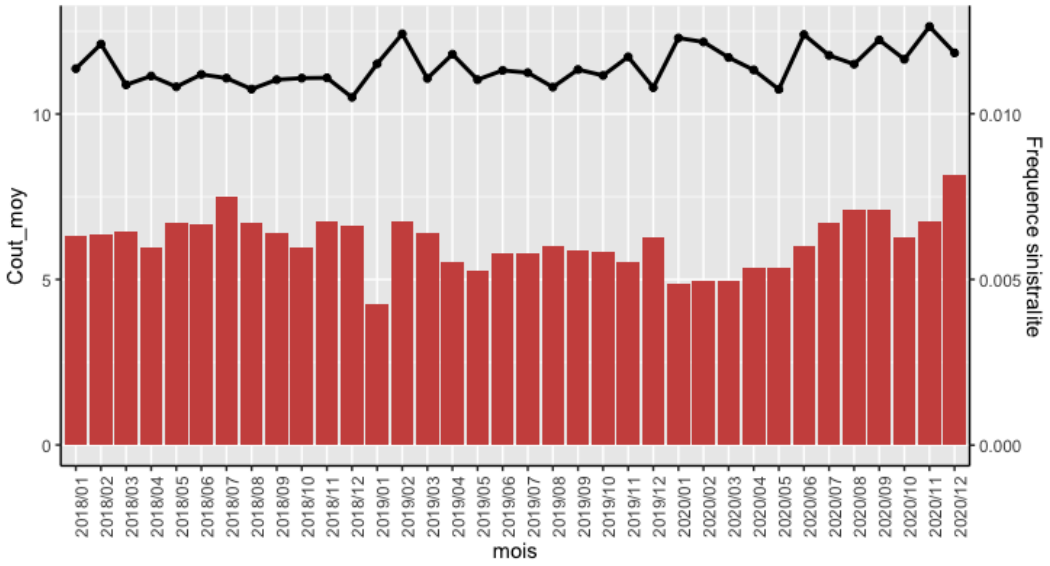


Figure 6 Evolution par mois de la fréquence et du coût moyen

L'observation par mois de l'évolution de la fréquence et du coût moyen montre une chute de la fréquence en début d'année liée au confinement. Cette chute est finalement rattrapée sur le reste de l'année.

Sur la base de ces données nous avons décidé de conserver l'année 2020 dans notre base.

2.3. Statistiques descriptives du portefeuille

Cette partie sur les statistiques descriptives permet d'avoir une meilleure connaissance de notre base d'étude et d'observer le caractère discriminant des variables tarifaires par rapport aux assurés.

Consommation et sinistralité par bloc de garanties

La tarification se faisant par bloc de garanties, nous analysons dans un premier temps la consommation et la sinistralité par bloc dans notre base d'étude.

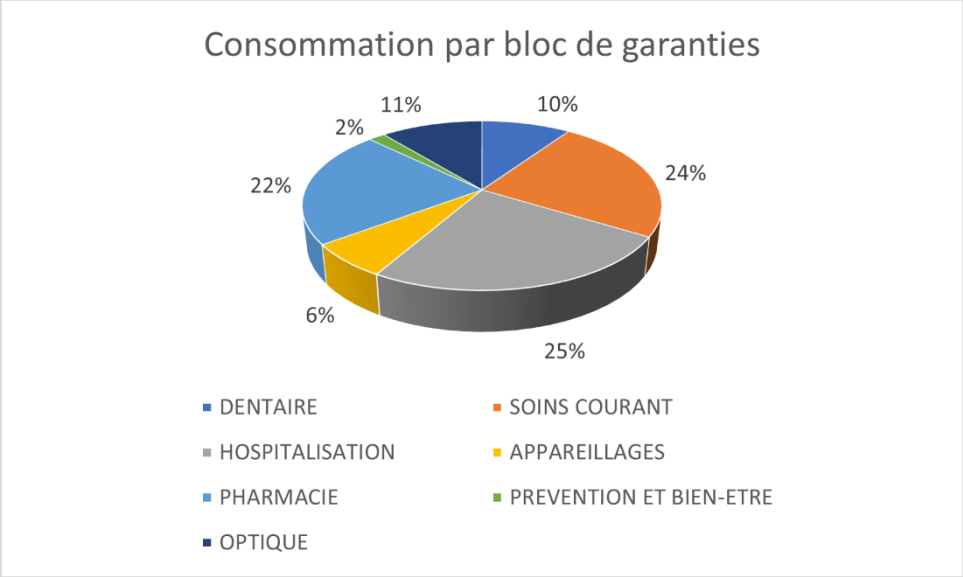


Figure 7 Consommation par bloc de garanties

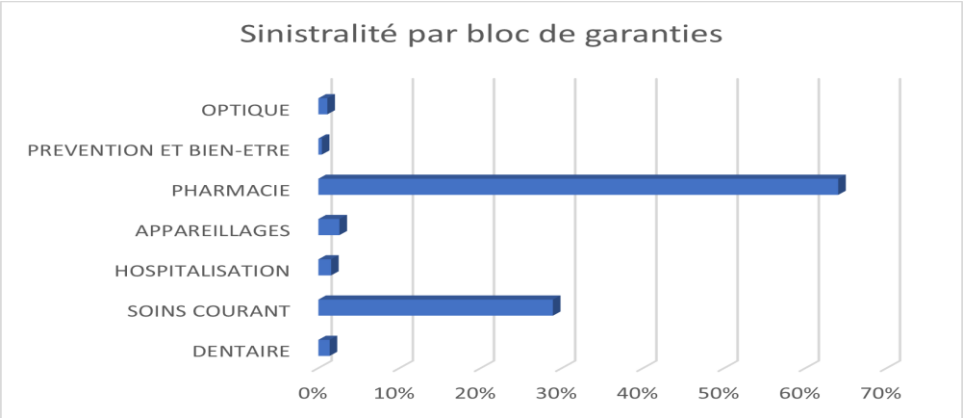


Figure 8 Sinistralité parr bloc de garanties

On note une forte représentation de la pharmacie et des soins courants. Les autres blocs sont assez peu représentés, mais nous disposons d'effectifs suffisant sur ces blocs pour réaliser nos modèles.

La pyramide des âges

La base compte à peu près 270 000 assurés sur 3 années d'observations (2018 à 2020). On remarque une prépondérance des assurés de plus de 60 ans, ce qui est tout à fait normal étant donné que les personnes âgées nécessitent plus de soins de santé que les plus jeunes. Les actifs ont également plus tendance à être couvert par l'assurance santé collective de leurs entreprises.

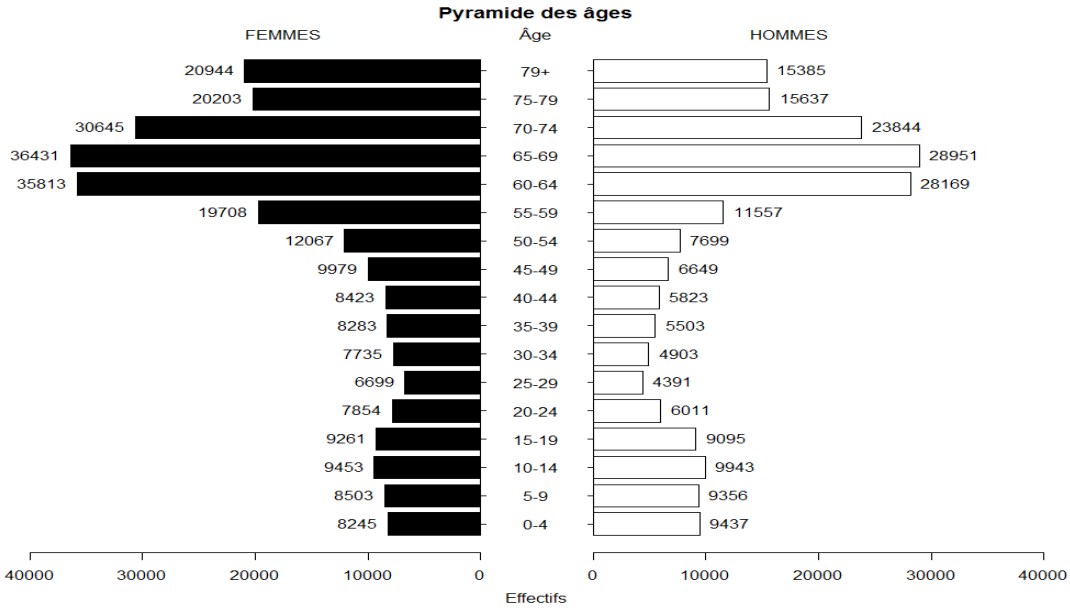


Figure 9 Pyramide des âges

Consommation et fréquence par âge

La pyramide des âges laisse entrevoir une influence de l'âge sur la consommation et la fréquence.

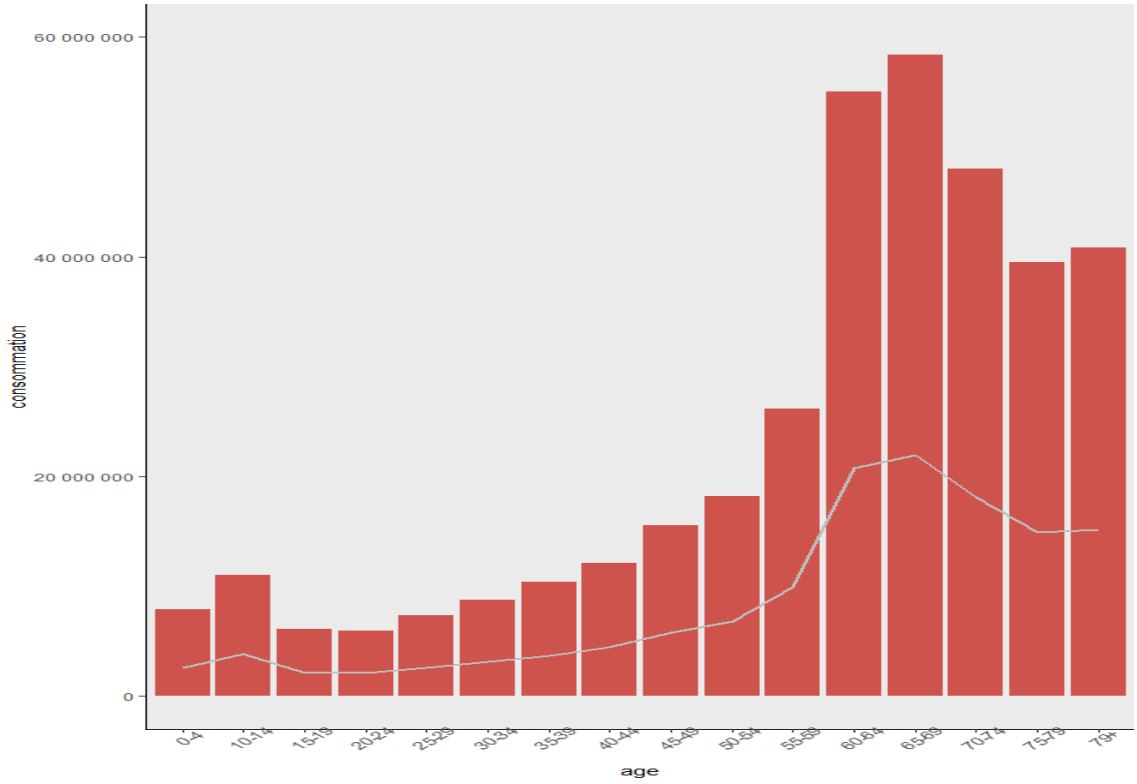


Figure 10 Consommation par âge

Frais réels **Dépenses de la complémentaire**

Le graphe ci-dessus montre globalement une augmentation de la consommation avec l'âge. Le constat est le même que ce soit pour la consommation en frais réels ou les dépenses de la complémentaire santé.

En terme de fréquence la même tendance est observée comme le montre le graphe ci-dessous.

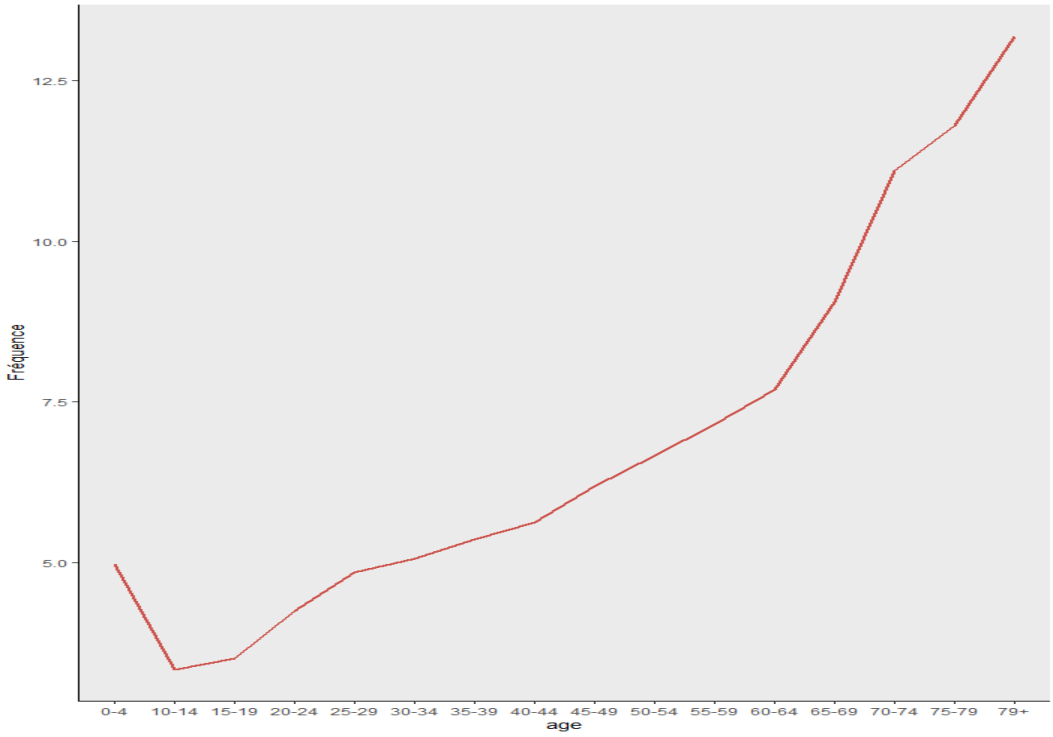


Figure 11 Fréquence par âge

Coût moyen et fréquence par niveau de garanties

Le coût moyen augmente avec le niveau de garanties, comme l'illustre le graphique ci-dessous.

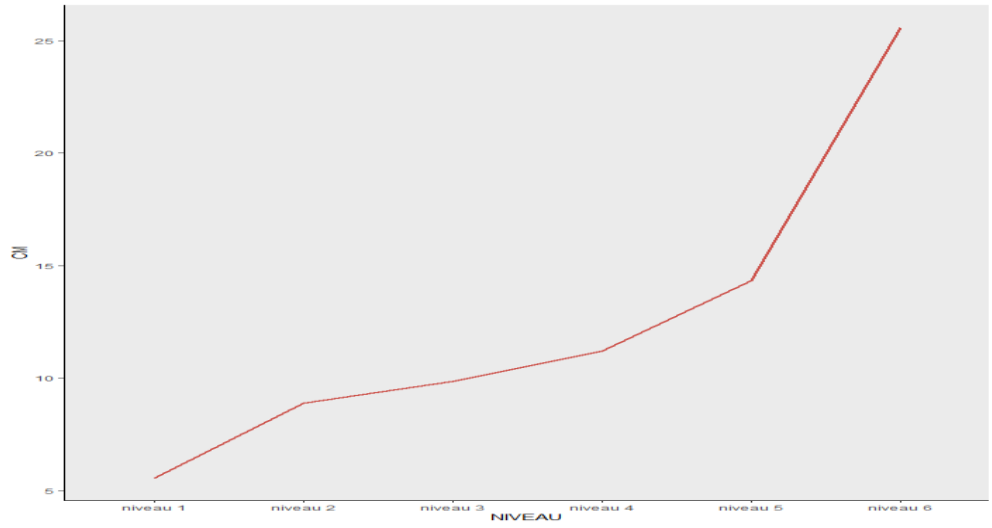


Figure 12 Coût moyen par niveau de garanties

La tendance contraire est observée pour la fréquence :

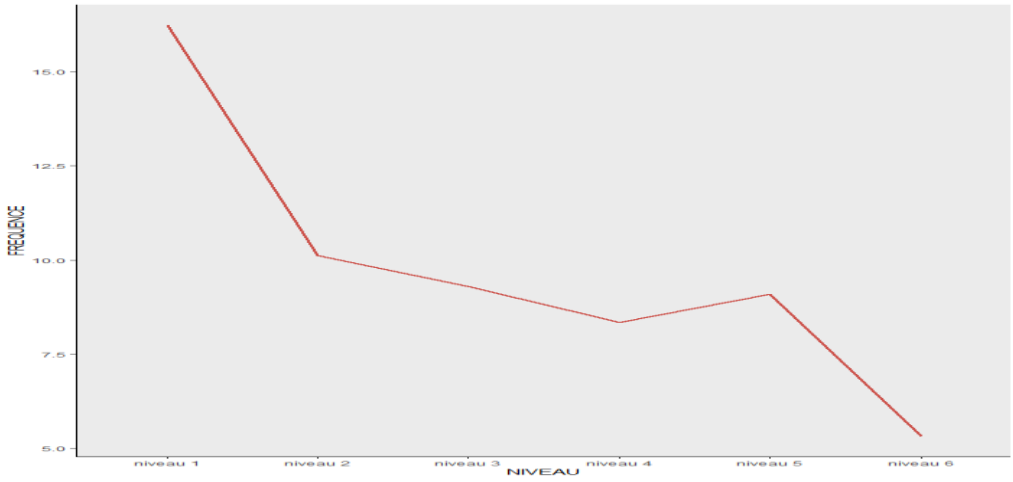


Figure 13 Fréquence par niveau de garanties

Les assurés souscrivent le plus souvent aux niveaux de garanties les plus bas ce qui explique cette tendance observée pour la fréquence.

Les assurés souscrivant aux niveaux les plus élevés visent des postes de garanties comme l'optique ou le dentaire qui sont assez chers et qui ne représentent pas énormément de sinistres.

Tout comme l'âge, le niveau de garanties influent également sur la consommation des assurés en santé.

Consommation par régime

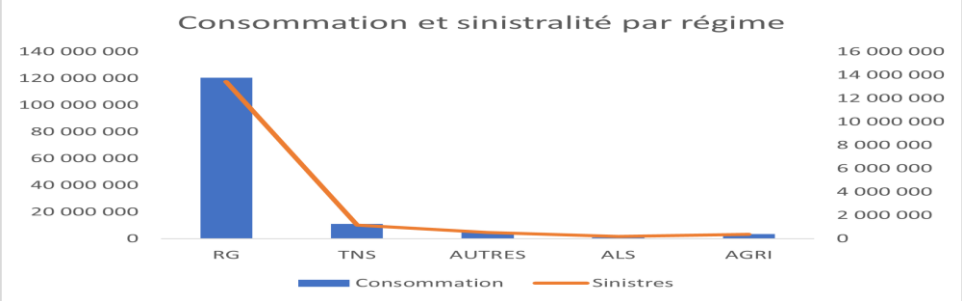


Figure 14 Consommation et sinistralité par régime

Comme on peut le voir sur le graphe ci-dessus, le régime général est sur-représenté dans la base. Pour une meilleure modélisation, on regroupe les autres régimes, qui sont très peu représentés, dans une modalité que l'on nommera **'Hors régime général'**.

Coût moyen et fréquence par zone

Afin de réaliser notre tarification, nous disposons de deux zoniers : le zonier A et le zonier B.

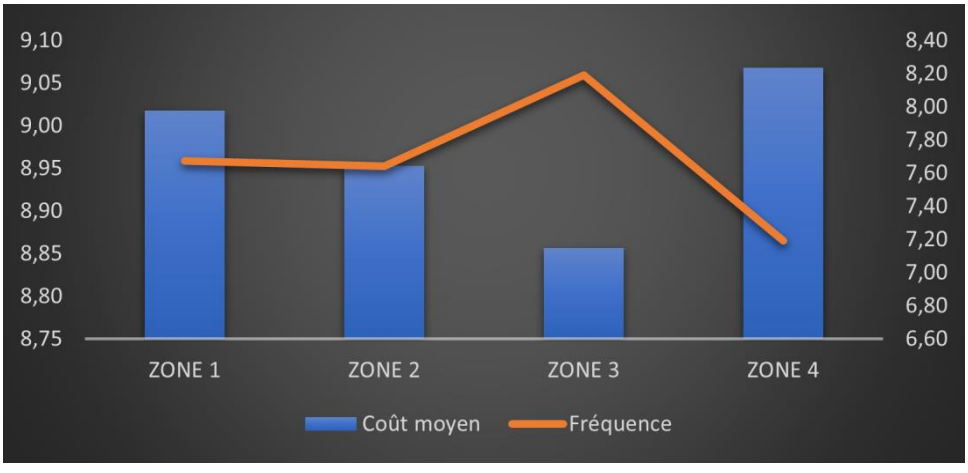


Figure 15 Coût moyen et fréquence zonier A

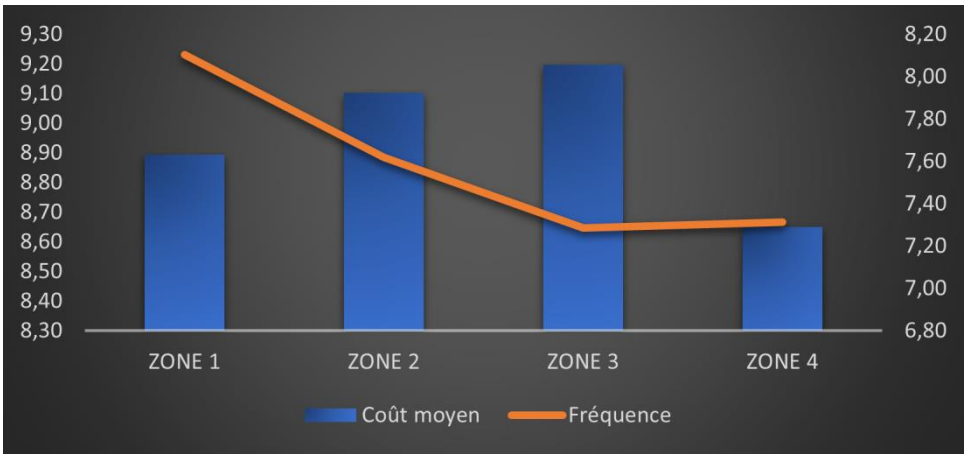


Figure 16 Coût moyen et fréquence zonier B

On n'observe pas une tendance claire pour la fréquence et le coût moyen avec le zonier A. Le zonier B montre quant à lui une baisse de la fréquence de la zone 1 à la zone 4. Ce constat est plutôt logique. En effet la zone 1 est constituée de départements à forte densité de populations et de praticiens. Le même constat aurait dû être fait avec le zonier A.

2.4. Zoniers

Pour tarifer, on dispose généralement de zoniers qui peuvent être différents selon le partenaire. Ces zoniers peuvent revêtir un aspect plus commercial afin de prendre en compte les demandes du partenaire.

Comme précisé plus haut, nous avons utilisé deux zoniers pour réaliser notre tarification :

- **Le zonier A** dont on se sert généralement pour réaliser nos tarifs. Ce zonier est commun à tous les produits de notre base d'étude et revêt un aspect commercial. Il regroupe les départements en 4 zones, la zone 1 étant la plus chère et la zone 4 la moins chère.
- **Le zonier B** qui est technique et ne dispose pas de l'aspect commercial du zonier A. Il regroupe également les départements en 4 zones, la zone 1 étant la plus chère et la zone 4 la moins chère.

Ces deux zoniers présentent beaucoup de disparités avec notamment 55 départements qui ne sont pas classés de la même façon.

Partie 3 – Tarification et application de GLM

3.1. La modélisation « coût x fréquence »

La tarification en assurance IARD se fait généralement suivant un modèle « coût x fréquence » dans lequel on modélise l'impact des variables explicatives sur le risque à travers des GLM.

L'utilisation de ce type de modèle suppose une indépendance entre le coût et le nombre de sinistres. En effet :

Soient **S** la charge totale de sinistres, $\mathbf{C_i}$ le coût unitaire d'un sinistre et **N** le nombre de sinistres. On a :

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

La formules des espérances conditionnelles totales nous permet d'écrire :

$$E(S) = \sum_{j=0}^{+\infty} E(S|N = j) * P(N = j)$$

$$= \sum_{j=0}^{+\infty} E\left(\sum_{i=1}^{N} Ci \,\middle|\, N = j\right) * P(N = j)$$

En supposant que les C_i sont indépendants de N et qu'ils sont indépendants et identiquement distribués :

$$E(S) = \sum_{j=0}^{+\infty} E(\sum_{i=1}^{j} Ci) * P(N = j)$$

$$= \sum_{j=0}^{+\infty} \sum_{i=1}^{j} E(C) * P(N = j)$$

$$= \sum_{j=0}^{+\infty} jE(C) * P(N = j)$$

$$= E(C) \sum_{j=0}^{+\infty} jP(N = j)$$

$$= E(C) * E(N)$$

Ce résultat n'est obtenu que sous la condition de l'indépendance entre le coût et le nombre de sinistres. Nous allons vérifier cette hypothèse sur tous les postes de santé afin de vérifier si la modélisation « coût x fréquence » est adaptée.

3.2. Etude de la dépendance entre coût et fréquence

Pour mesurer la dépendance entre ces deux variables nous allons utiliser **le coefficient de corrélation de Pearson, de Kendall et de Spearman**.

- **Le coefficient de corrélation de Pearson** mesure le degré de relation linéaire entre deux variables quantitatives et est compris entre -1 et 1. Si l'augmentation d'une variable entraîne la diminution de l'autre le coefficient est négatif. En revanche, si les deux variables tendent à évoluer dans le même sens, le coefficient est positif. Un coefficient proche en valeur absolue de 1 implique une forte dépendance linéaire. Lorsque deux variables sont indépendantes le coefficient est nul, la réciproque est cependant fausse .

  Formule de Pearson

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=0}^{n}(y_i - \overline{y})^2}}$$

- **Le tau de Kendall et le rho de Spearman** mesurent la corrélation de rang entre deux variables. Le calcul se fait donc sur les rangs et non sur les valeurs. Contrairement au coefficient de Pearson, ils permettent de capter une relation non linéaire. Les valeurs de ces coefficients sont comprises entre -1 et 1 et s'interprètent de la même manière que le coefficient de Pearson.

  Récapitulatif des différents indicateurs par poste de garanties

  Afin d'appliquer ces différents coefficients, nous avons conservé les assurés ayant eu au moins un sinistre pour le poste concerné. En effet la relation « sinistre=0 alors coût=0 » biaiserait les coefficients.

	Hospitalisation	Soins courants	Pharmacie	Optique	Dentaire	Appareillages	Prévention et bien-être
Spearman	0,159	-0,002	0,022	-0,041	0,033	0,075	-0,069
Kendall	0,112	-0,001	0,033	-0,03	0,022	0,053	-0,044
Pearson	0,069	-0,015	-0,005	-0,024	0,029	0,018	-0,053

Tableau 1 Coefficients de corrélation par poste de garanties

Pour chacun des postes, les trois indicateurs sont très proches de 0. On peut donc supposer une indépendance entre la fréquence et le coût moyen pour chaque poste. La modélisation « coût x fréquence » semble donc adaptée.

3.3. Les GLM et Zero-inflated GLM

a. Les Modèles Linéaires Généralisés (GLM)

Rappel sur les modèles linéaires gaussiens

Le principe de ces modèles est d'établir une relation linéaire entre une variable aléatoire Y à expliquer et des variables explicatives X_1,….,X_n. Ainsi le modèle répond à l'équation :

$$Y = X\theta + \epsilon$$

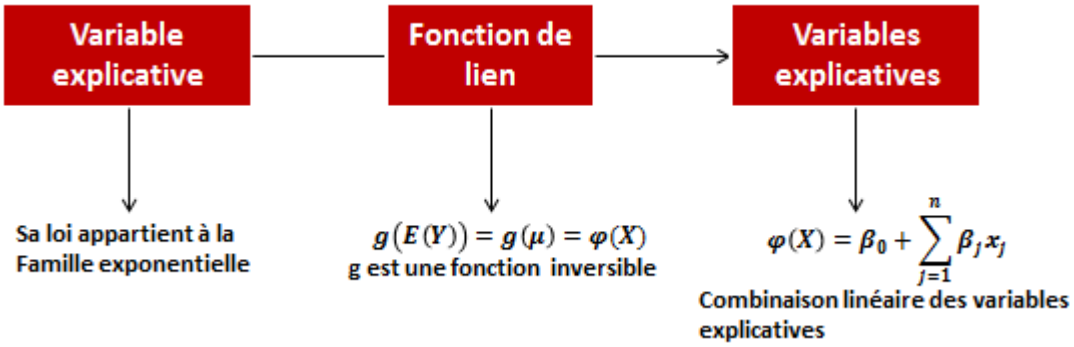
Avec :

- X le vecteur de variables explicatives déterministes (1, X_1,….,X_n)
- θ le vecteur (θ_0, θ_1,…., θ_n) de paramètres à estimer
- ϵ la variable aléatoire représentant les résidus c'est-à-dire l'écart entre la valeur prédite et celle observée. Les résidus ($\epsilon_1, \dots, \epsilon_n$) sont indépendants et identiquement distribués. Ils suivent une loi normale d'espérance 0 et de variance σ^2(homoscédasticité).

Y doit donc suivre une loi normale ce qui est un inconvénient dans le secteur assurantiel où l'on ne trouve pas souvent ce genre de variable. Les modèles linéaires généralisés permettent de contourner ce problème.

Les GLM

Les modèles linéaires généralisés modélisent une fonction (fonction de lien) de l'espérance de la variable Y à expliquer contrairement aux modèles linéaires gaussiens qui modélise directement la variable à expliquer. Les modèles linéaires généralisés suivent donc le schéma suivant :



Le choix de la loi de Y est une étape importante de la modélisation. Cette loi doit appartenir à une famille exponentielle de densité :

$$f_{\theta,\varphi}(y) = c_{\varphi}(y)\exp\left(\frac{y\theta - a(\theta)}{\varphi}\right)$$

Où :

- θ est le paramètre canonique et φ le paramètre de de dispersion
- $a(\theta)$ est de classe C^2 et convexe
- $c_{\varphi}(y)$ ne dépend pas de θ

Les lois gamma, normale, binomiale négative, poisson, inverse gaussienne sont des exemples de loi appartenant à la famille exponentielle :

- Les lois de Poisson (en absence de sur dispersion) et Binomiale négative sont généralement utilisées pour modéliser la fréquence.
- Les lois gamma, log normale et inverse gaussienne sont quant à elles généralement utilisées pour modéliser le coût moyen.

Le choix entre les différentes lois se fait graphiquement et grâce à des indicateurs comme l'AIC.

Le choix de la fonction de lien est libre. Pour notre modélisation nous avons choisi la fonction logarithme népérien $g(x) = \ln(x)$ pour profiter de son caractère multiplicatif et éviter un tarif négatif grâce à l'exponentielle. En effet :

$$\ln\left(E(y_i)\right) = \alpha_0 + \sum_{j=1}^{k} \alpha_j \, x_{i,j}$$

$$E(y_i) = \exp\left(\alpha_0 + \sum_{j=1}^{k} \alpha_j \, x_{i,j}\right)$$

$$E(y_i) = \exp(\alpha_0) * \prod_{j=1}^{k} \exp(\alpha_j x_{i,j})$$

b. Estimation des coefficients par la méthode du maximum de vraisemblance

L'estimation des coefficients α_j se fait par la méthode se fait par la méthode du maximum de vraisemblance. La vraisemblance s'écrit comme le produit de fonction de densités :

$$L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Où les x_i sont n-échantillons indépendants et θ le paramètre à observer. Lorsqu'on ne connaît pas θ on peut l'estimer en cherchant ses valeurs qui maximisent $L(x_1, x_2, \dots, x_n; \theta)$. On obtient l'estimateur du maximum de vraisemblance noté $\hat{\theta}$.

Afin de rechercher cet estimateur, on passe souvent par la log vraisemblance qui est le logarithme de la fonction de vraisemblance. La fonction logarithme étant strictement croissante, la log vraisemblance et la vraisemblance atteignent leur maximum au même point. De plus, le logarithme permet de passer d'un produit de densités à une somme de densités. Cette somme est beaucoup plus simple à dériver qu'un produit.

La log vraisemblance s'écrit donc :

$$\ln(L(x_1, x_2, \dots, x_n; \theta)) = \ln\left(\prod_{i=1}^{n} f(x_i; \theta)\right)$$

$$= \sum_{i=1}^{n} \ln(f(x_i; \theta))$$

Déterminer le maximum de cette fonction consiste donc à trouver les valeurs de θ pour lesquelles sa dérivée est nulle et sa dérivée seconde négative. Ces deux conditions s'exprimes ainsi :

$$\frac{\partial \ln(L(x_1, x_2, \dots, x_n; \theta))}{\theta} = 0$$

$$\frac{\partial^2 \ln(L(x_1, x_2, \dots, x_n; \theta))}{\theta} < 0$$

c. Validation du modèle

Etude de la statistique et des résidus de Pearson

Pour juger de la bonne qualité de l'ajustement du modèle on peut se servir du test de Pearson de statistique :

$$X_p^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i)^2}{V(\widehat{Y}_i)}$$

Cette statistique est approximativement distribuée suivant une loi du khi 2 à n-p degrés de liberté.

Les résidus de Pearson quant à eux suivent approximativement une loi normale centrée réduite. Les individus ayant en valeur absolue un résidu ≥ 2 sont à surveiller. Ils peuvent réduire la qualité du modèle.

Etude de la déviance et de la déviance résiduelle

Un autre moyen de vérifier le bon ajustement du modèle est l'étude de la déviance. La déviance permet de comparer la vraisemblance du modèle estimé à celle d'un modèle parfait en termes d'ajustement aux données : le modèle saturé. Ce modèle à autant de paramètres que de variables explicatives, la même fonction de lien et la même distribution que le modèle estimé.

Soit L_{sat} la vraisemblance du modèle saturé et L_n celle du modèle estimé. La déviance s'écrit :

$$\delta = 2 * (L_{sat} - L_n)$$

$$\delta = 2 * \sum_{i=1}^{n} \frac{Y_i\left(\widehat{\theta_{sat,i}} - \widehat{\theta_{n,i}}\right) + a\left(\widehat{\theta_{sat,i}}\right) - a\left(\widehat{\theta_{n,i}}\right)}{\varphi}$$

Les résidus de déviance s'écrivent :

$$\gamma_i = sign(y_i - \mu_i)\sqrt{\gamma_i^2}$$

Où :

$$\gamma_i^2 = 2\,\frac{Y_i\left(\widehat{\theta_{sat,i}} - \widehat{\theta_{n,i}}\right) + a\left(\widehat{\theta_{sat,i}}\right) - a\left(\widehat{\theta_{n,i}}\right)}{\varphi}$$

Si le modèle est bien ajusté δ converge en loi vers une loi du khi 2 à n-p degrés de liberté, où p est le nombre de paramètres à estimer.

Les résidus de déviance≥ 2 peuvent indiquer un défaut d'ajustement.

La cross validation

La validation croisée permet de tester la fiabilité du modèle et de vérifier qu'il ne fait pas de sur apprentissage. Elle permet également d'utiliser l'intégralité des données pour l'entraînement et la validation du modèle.

La validation croisée k-folds consiste à scinder le jeu de données en k parties. L'une après l'autre chaque partie servira de jeu de test pendant que les autres parties seront utilisées pour l'entraînement.

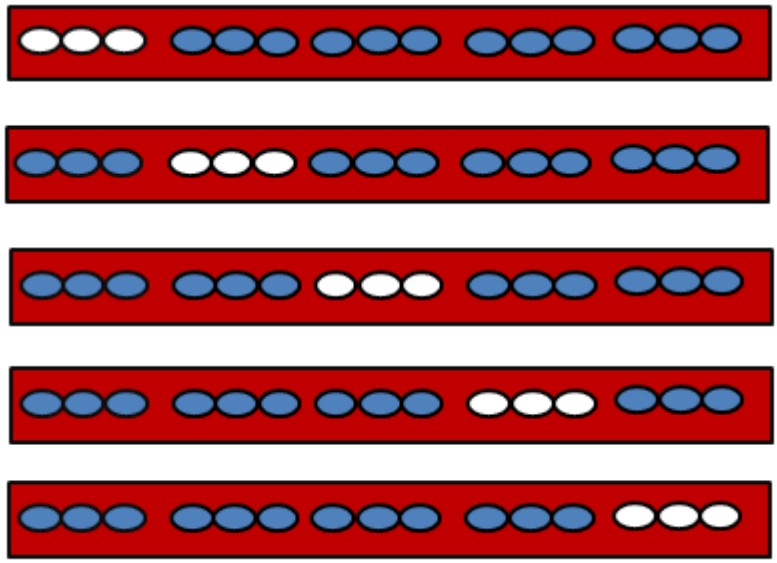


Figure 17 Illustration de la validation croisée

Cette image illustre une cross validation à 5 parties avec en blanc les jeux de test et en bleu ceux d'entraînements.

Une fois que tous les ensembles sont testés, on peut obtenir une estimation unique en moyennant les résultats obtenus sur les k-folds. La cross validation nous permettra d'obtenir des indicateurs comme le RMSE (Root Mean Square Error) qui mesure l'amplitude des écarts entre les valeurs prédites et celles observées. Soient ε_i ces écarts et n le nombre d'observations :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2}$$

d. Les Zero-Inflated GLM

Les Zero-Inflated GLM pour modéliser une variable de comptage présentant une proportion importante de zéros. Le phénomène 'nombre de sinistres = 0' peut s'expliquer par l'absence du phénomène ou un nombre d'apparition nul sur la période d'observation. Pour le poste Prévention et bien-être par exemple un 'nombre de sinistres = 0' pour un jeune assuré est tout à fait normal. Pour un assuré âgé ce phénomène peut s'expliquer par le fait que cet assuré n'a pas eu l'opportunité d'avoir recours aux soins sur la période d'observation. La régression Zero-Inflated permet de prendre en compte ce double sens du phénomène 'nombre de sinistres = 0'.

Dans notre étude cette situation se présente pour 4 postes : l'hospitalisation, l'optique, le dentaire et la prévention et le bien-être.

	Hospitalisation	Soins courants	Pharmacie	Optique	Dentaire	Appareillages	Prévention et bien-être
Pourcentage de zéros	70%	17%	9%	70%	71%	55%	85%

Tableau 2 Part des assurés n'ayant pas eu de sinistres

Ces quatre postes présentent en effet plus de 70% d'assurés n'ayant pas eu de sinistres sur la période d'observation.

Explication du modèle

La régression Zero-Inflated combine une régression logistique et une régression Poisson ou binomiale négative. La régression logistique permet de déterminer en fonction des variables explicatives la survenance ou non de 'nombre de sinistres = 0'. La régression Poisson ou binomiale négative permet le comptage du nombre de sinistres y compris possiblement la valeur 0.

Soient Y le nombre de sinistres, X et Z les variables explicatives respectivement du modèle Poisson et du modèle logistique. Y a pour distribution et espérance :

$$P(Y|X,Z) = \begin{cases} \pi + (1-\pi)e^{-\mu} & si\ y = 0 \\ (1-\pi)\dfrac{\mu^y e^{-\mu}}{y!} & si\ y > 0 \end{cases}$$

$$E(Y|X,Z) = (1-\pi)\mu$$

Avec π la probabilité qu'on ait 0 et μ le paramètre de la loi Poisson. X et Y peuvent être identiques, disjoints ou avoir des variables en commun.

La régression logistique permet de modéliser le paramètre π avec une fonction de lien logit :

$$\ln\left(\frac{\pi}{1-\pi}\right) = Z\beta$$

$$\hat{\pi} = \frac{\exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i z_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{n}\beta_i z_i\right)}$$

Où :

- Z est le vecteur de variables explicatives déterministes (1, Z_1 ,….,Z_n)
- β est le vecteur (β_0, β_1 ,…., β_n) de paramètres à estimer

La régression de Poisson permet de modéliser le paramètre μ avec une fonction de lien log :

$$\ln(\mu) = X\theta$$

$$\hat{\mu} = \exp\left(\theta_0 + \sum_{i=1}^{n}\theta_i x_i\right.$$

Avec :

- X le vecteur de variables explicatives déterministes (1, X_1 ,….,X_n)
- θ le vecteur (θ_0, θ_1 ,…., θ_n) de paramètres à estimer

Ainsi la prédiction de l'espérance du nombre de sinistres de l'individu i est :

$$E(\hat{y_i}|X,Z) = (1-\hat{\pi})\hat{\mu}$$

Le test de Vuong

Ce test nous permet de définir le meilleur modèle entre deux modèles (GLM et Zero-Inflated dans notre cas). Ces modèles peuvent être imbriqués ou non. La statistique de Vuong teste l'hypothèse H_0 selon laquelle les deux modèles sont proches de façon équivalente de la vraie densité d'une variable Y conditionnellement à des variables explicatives (X, Z) contre une hypothèse H_1 selon laquelle l'un des modèles est le plus proche. Le modèle 1 est par exemple préféré au modèle 2 à un niveau de significativité α si :

$$Z = \frac{LR_N(\beta_{ML,1}, \beta_{ML,2)}}{\sqrt{N} w_N}$$

est supérieure au quantile d'ordre $(1-\alpha)$ d'une loi normale centrée réduite.

Le numérateur :

$$LR_N(\beta_{ML,1}, \beta_{ML,2)} = L_N^1 - L_N^2 - \frac{K_1 - K_2}{2}\log(N))$$

Représente la différence entre le maximum de vraisemblance des deux modèles.

3.4. Résultats et test d'adéquation

a. Etude de la dépendance entre les variables explicatives

La réalisation des GLM nécessite l'absence de corrélation entre les variables explicatives. En effet une forte corrélation entre certaines variables peut réduire la précision des estimations. On peut être emmené à conclure que certaines variables n'ont pas d'effet sur le modèle alors qu'elles en ont une fois qu'elles sont prises individuellement.

Pour mesurer la dépendance entre nos variables, nous utiliserons le V de Cramer qui permet de traiter la présence de variables qualitatives comme c'est le cas dans notre étude. Il se base sur le test du khi 2 d'indépendance entre les variables. Il est défini par :

$$V = \sqrt{\frac{X^2}{N inf(r-1, s-1)}}$$

Avec :

- N l'effectif total
- r le nombre de lignes de la table de contingence
- s le nombre de colonnes de la table de contingence
- $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{i,j} - \frac{n_{i.}n_{.j}}{N}\right)^2}{\frac{n_{i.}n_{.j}}{N}} \cdot \frac{n_{i.}n_{.j}}{N}$ correspond à l'effectif théorique sous l'hypothèse d'indépendance.

V est compris entre 0 et 1. Plus il est proche de 1, plus les variables sont dépendantes entre elles. Plus il est proche de 0, plus la corrélation entre les variables étudiées est faible.

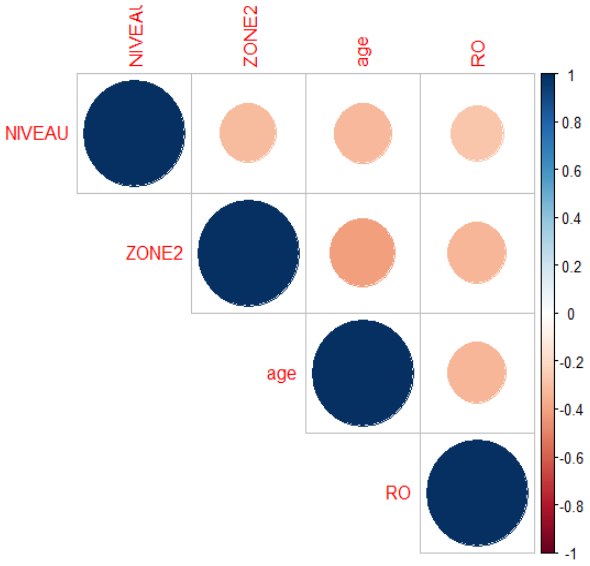


Figure 18 V de Cramer Hospitalisation

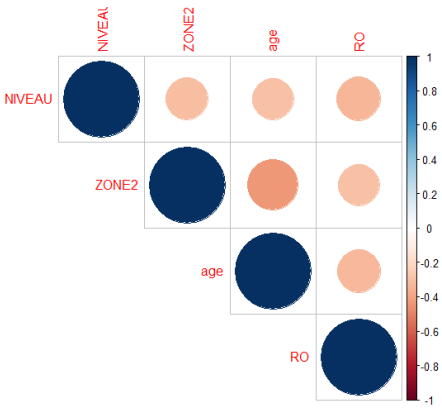


Figure 19 V de Cramer Pharmacie

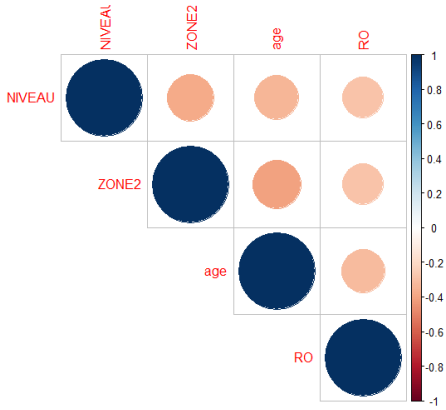


Figure 20 V de Cramer Soins courants

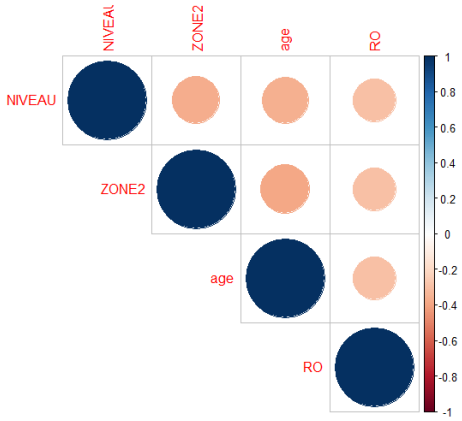


Figure 21 V de Cramer Dentaire

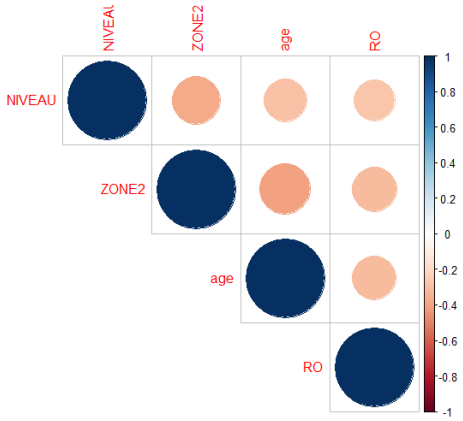


Figure 22 V de Cramer Optique

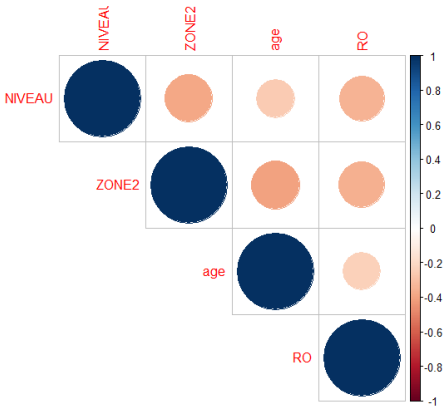


Figure 23 V de Cramer Prévention et bien-être

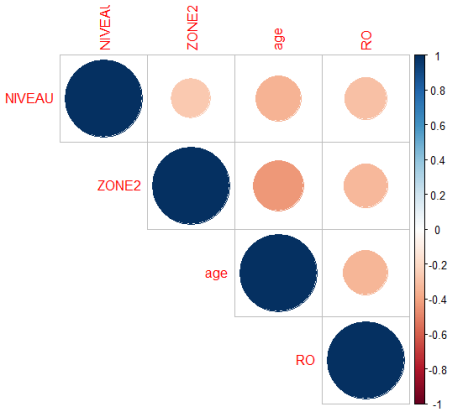


Figure 24 V de Cramer Appareillages

Quel que soit le poste on remarque une faible corrélation entre les variables tarifaires. On peut donc toutes les conserver pour la modélisation.

b. Choix des modèles de coût moyen

Pour modéliser le coût moyen nous allons tester les lois gamma, inverse gaussienne et log normale.

Critères de sélection

Le premier critère de sélection sera graphique. Il va consister à tracer la distribution empirique et la comparer aux différentes distributions théoriques afin de déterminer quelle loi est la mieux adaptée à nos données.

Le deuxième critère de sélection concernera les tests d'adéquation qui permettent de vérifier l'hypothèse nulle :

$$H_0 : F = F_n$$

Où F est la fonction de répartition théorique de la loi testée et F_n la fonction de répartition empirique de nos données. Deux tests d'adéquation sont généralement utilisés : le test Cramer-Von Mises ou celui de Kolmogorov-Smirmov. Nous utiliserons le premier cité qui est moins sensible aux valeurs extrêmes.

Le test d'adéquation conduit généralement à un rejet de l'hypothèse nulle à cause d'un trop grand nombre de données. Cela nous conduit à utiliser un troisième critère de sélection qui sera basé sur l'AIC et la déviance du modèle.

L'AIC utilise le maximum de vraisemblance tout en pénalisant les modèles ayant trop de variables ce qui limite les effets du sur apprentissage. Il s'écrit comme suit :

$$AIC = -2\ln\big(L(\theta)\big) + 2k$$

Avec L la vraisemblance du modèle et k le nombre de paramètres.

On choisit la loi qui minimise L'AIC et la déviance.

Application aux différents postes

- **L'HOSPITALISATION**

Dans un premier temps nous avons réalisé une analyse graphique en comparant la distribution empirique aux différentes distributions théoriques des lois gamma, inverse gaussienne et log normale.

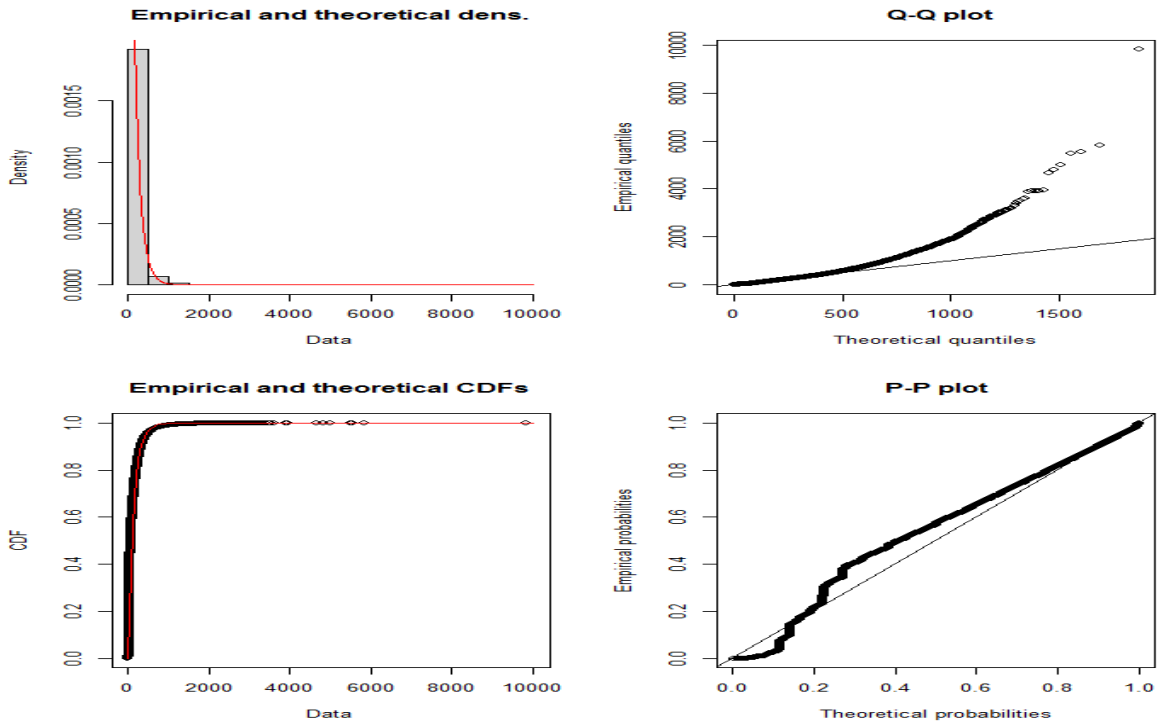


Figure 25 Distribution de la loi gamma Hospitalisation

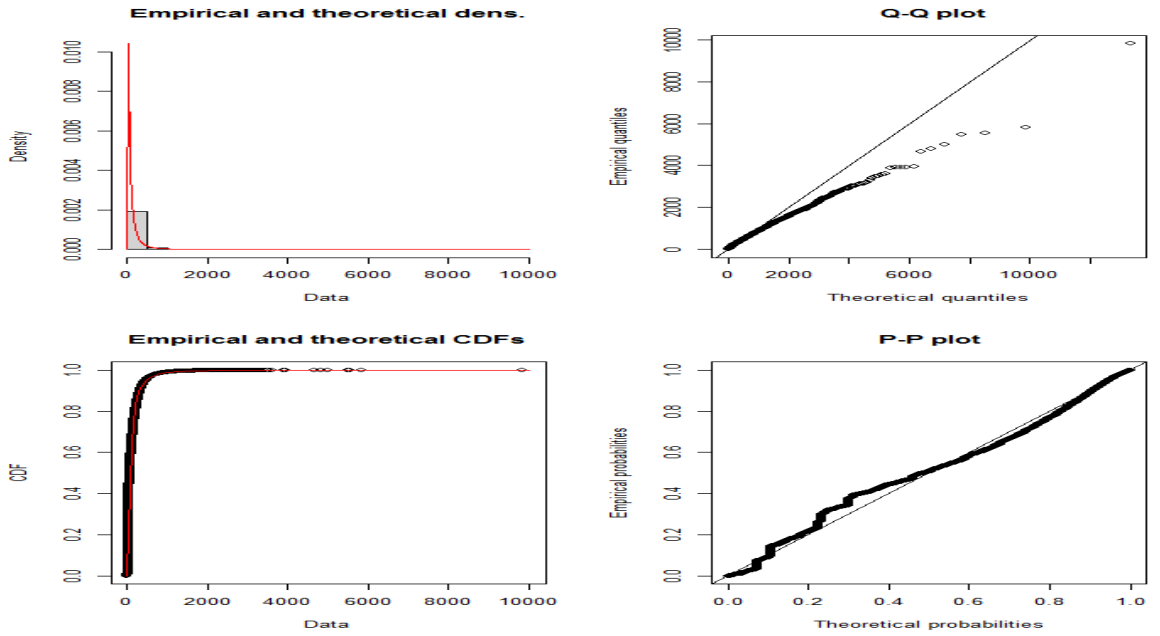


Figure 26 Distribution de la loi log normale Hospitalisation

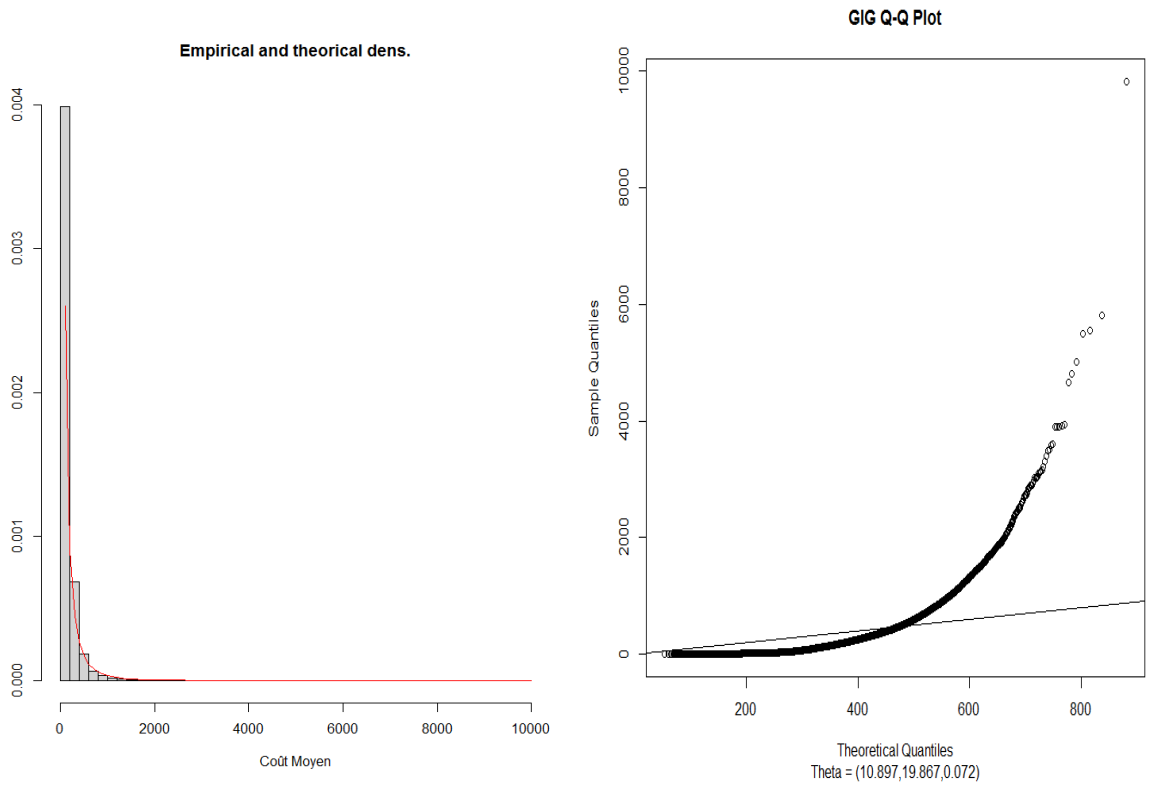


Figure 27 Distribution de la loi inverse gaussienne Hospitalisation

En observant ces trois graphes, les lois gamma et inverse gaussienne semblent le mieux correspondre à nos données. La loi log normale ne semble pas du tout adaptée lorsque l'on compare les distributions empirique et théorique.

A partir du test de CRAMER -VON MISES, de l'AIC et de la déviance, nous choisissons le meilleur modèle entre les GLM gamma et inverse gaussienne :

	Inverse gaussienne	Gamma
stat de Cramer-Von Mises	76,248	145,14
AIC	799 594	814 483
Deviance	91 574	99 299

Dû au grand nombre de données la p valeur du test de CRAMER-VON MISES n'est pas utilisable. La comparaison se fera à travers la statistique de test.

La loi inverse gaussienne minimise les trois critères. On la sélectionne donc pour la modélisation.

➢ SOINS COURANTS

Le modèle utilisant la loi inverse gaussienne ne convergeant pas elle a été exclue pour la modélisation de ce poste. L'analyse graphique concerne donc les lois gamma et log normale.

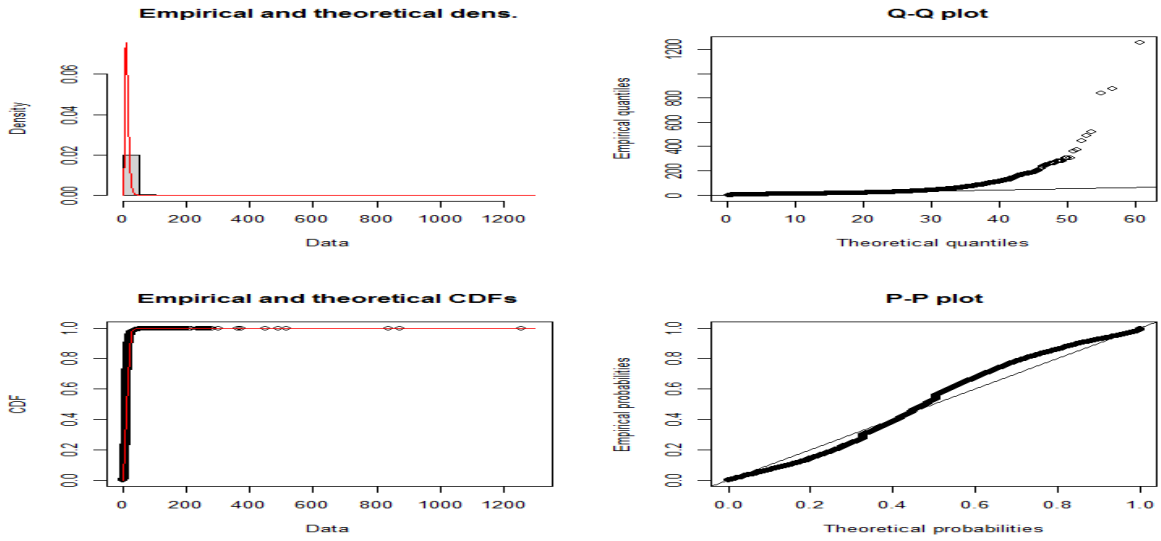


Figure 28 Distribution de la loi gamma Soins courants

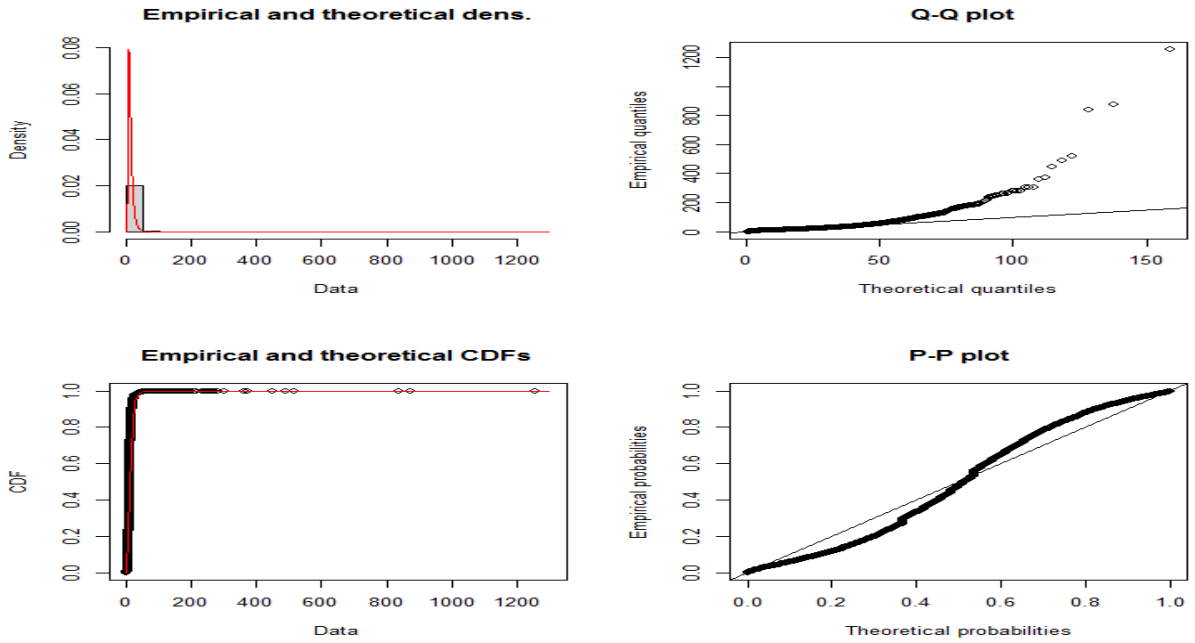


Figure 29 Distribution de la loi log normale soins courants

Il est difficile de choisir graphiquement entre les deux lois. On s'en remet donc aux autres critères de sélections :

	log normale	Gamma
stat de Cramer-Von Mises	780	480
AIC	1 230 955	1 187 702
Deviance	70 118	53 043

La loi gamma minimise les trois critères. On la conserve donc pour la modélisation de ce poste.

- **PHARMACIE**

Le modèle utilisant la loi inverse gaussienne ne convergeant pas elle a été exclue pour la modélisation de ce poste. L'analyse graphique concerne donc les lois gamma et log normale.

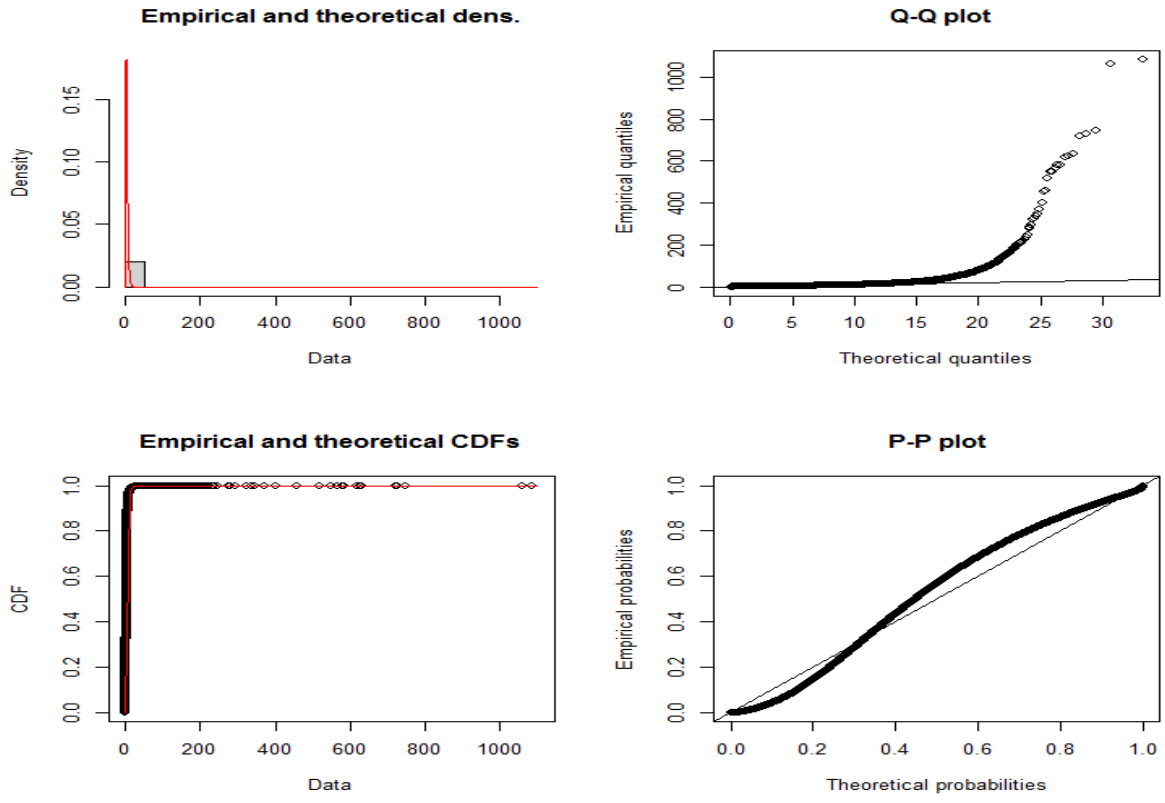


Figure 30 Distribution de la loi gamma Pharmacie

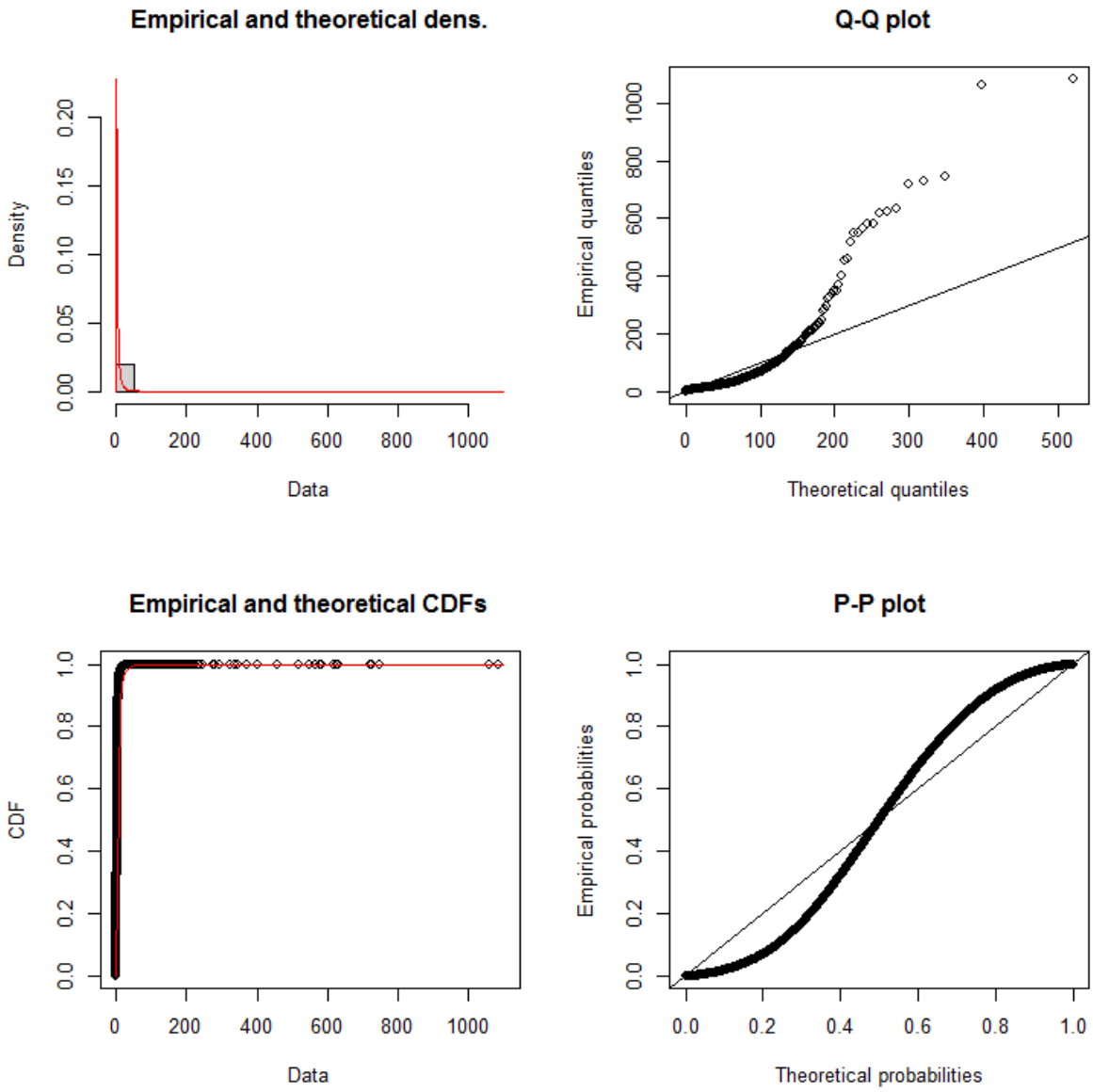


Figure 31 Distribution de la loi log normale Pharmacie

Les différents graphiques ne nous permettent pas de départager les deux lois. On utilise alors les autres critères pour notre choix final :

	log normale	Gamma
stat de Cramer-Von Mises	1737	632
AIC	1 070 646	940 789
Deviance	143 217	129 564

La loi gamma minimise les trois critères. On la conserve donc pour la modélisation de ce poste.

➢ OPTIQUE

Le modèle utilisant la loi inverse gaussienne ne convergeant pas elle a été exclue pour la modélisation de ce poste. L'analyse graphique concerne donc les lois gamma et log normale.

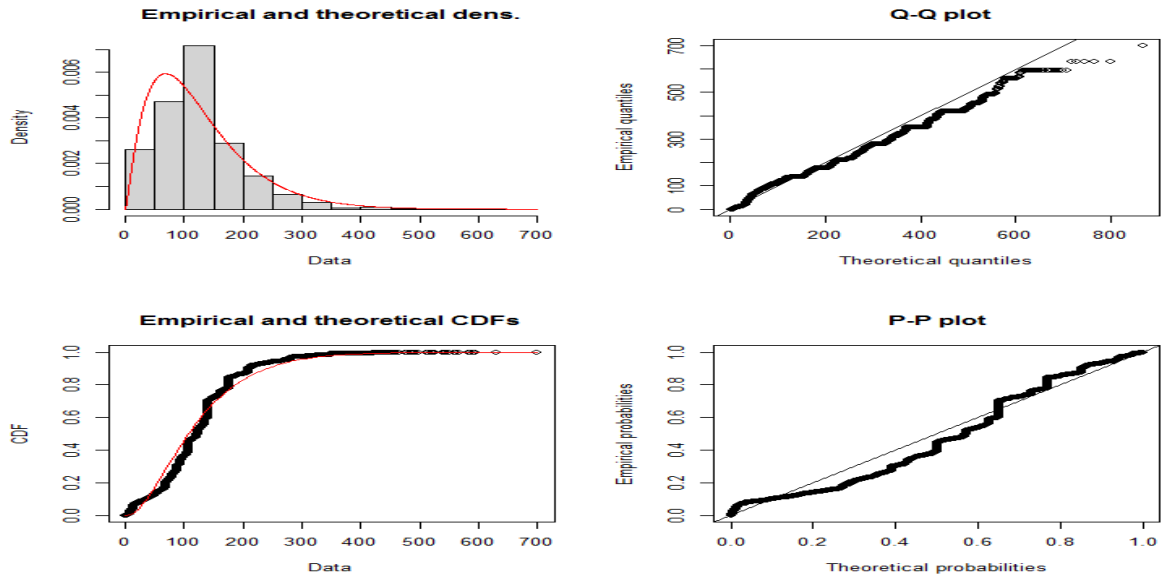


Figure 32 Distribution de la loi gamma Optique

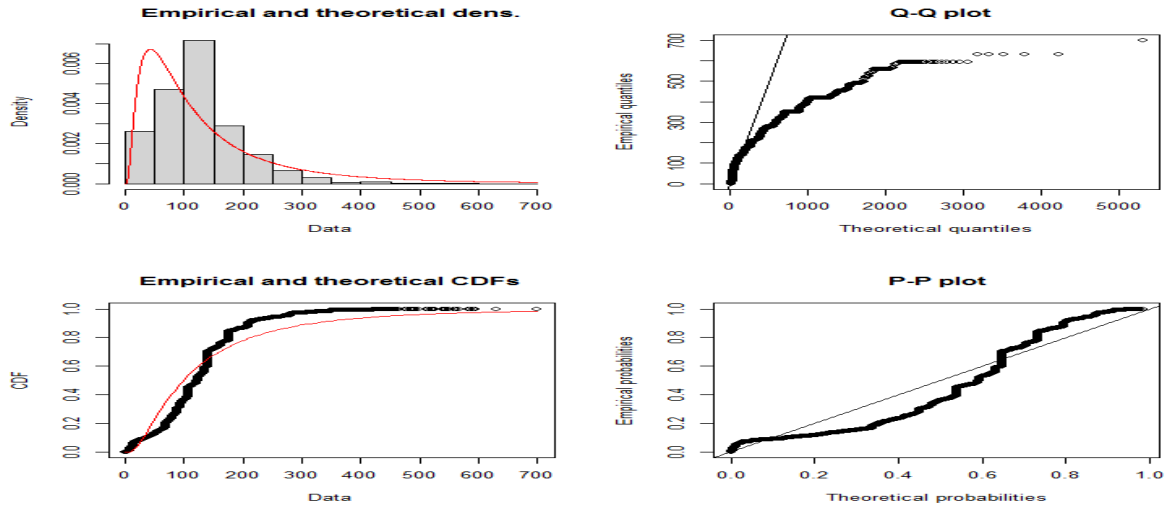


Figure 33 Distribution de la loi log normale Optique

Graphiquement la loi gamma semble mieux adaptée. L e qq plot et le pp plot montrent également un meilleur ajustement pour cette loi. Nous tentons de confirmer ce constat avec les autres critères.

	log normale	Gamma
stat de Cramer-Von Mises	713	281
AIC	816 088	785 122
Deviance	34 211	28 763

La loi gamma minimise les trois critères. On la conserve donc pour la modélisation.

- **DENTAIRE**

Le modèle utilisant la loi inverse gaussienne ne convergeant pas elle a été exclue pour la modélisation de ce poste. L'analyse graphique concerne donc les lois gamma et log normale.

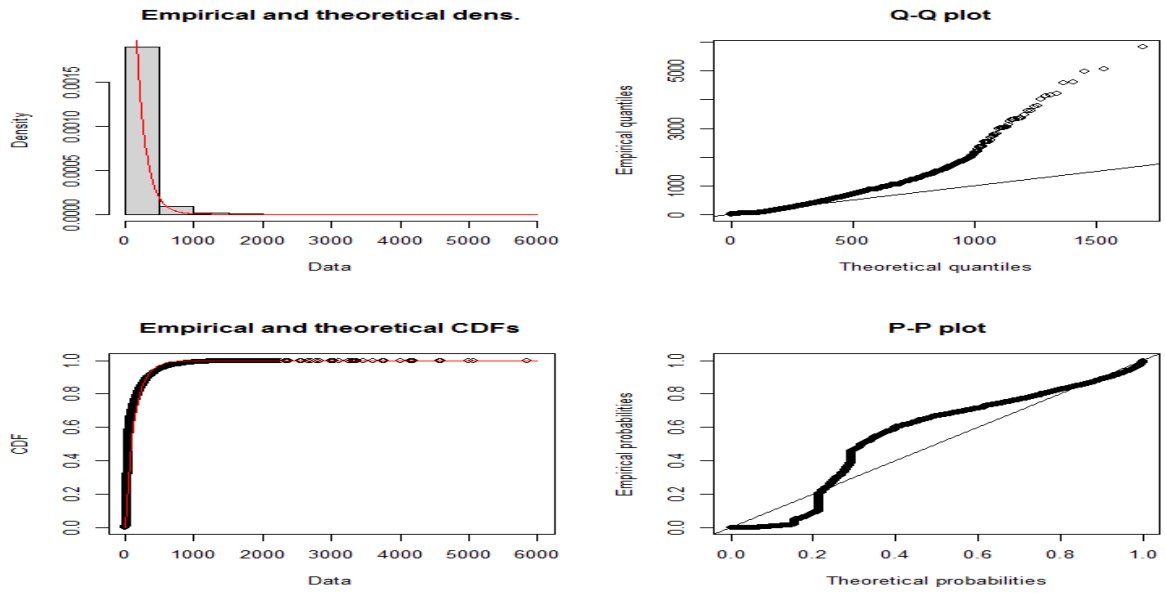


Figure 34 Distribution de la loi gamma Dentaire

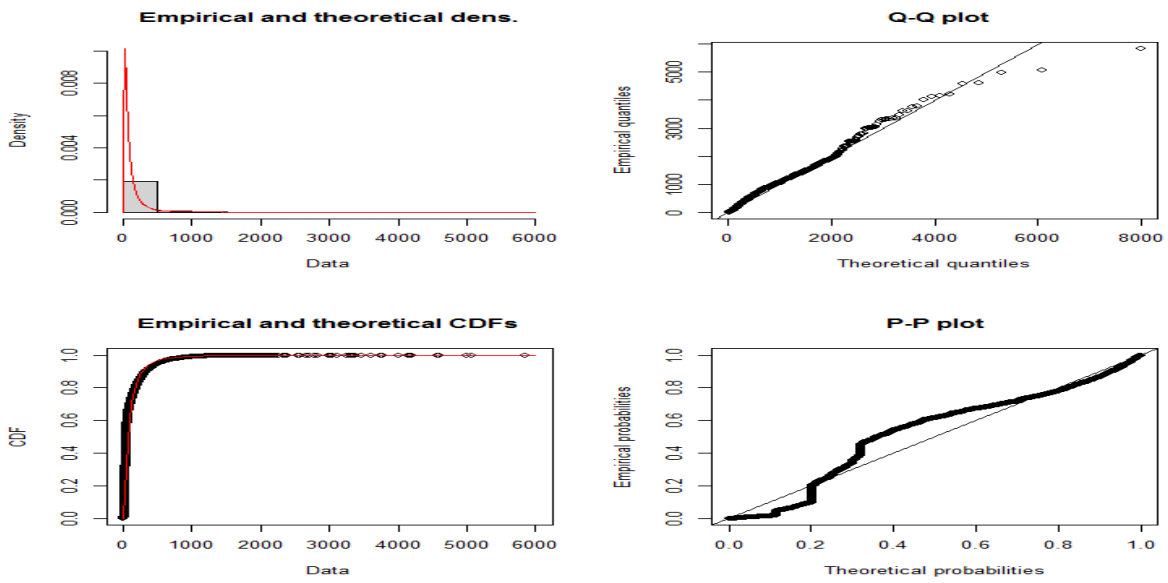


Figure 35 Distribution de la loi log normale Dentaire

Les pp plot et qq plot montrent un meilleur ajustement de la loi log normale. Les graphes comparant les densités empiriques et théoriques laissent cependant penser à une meilleur ad équation de la loi gamma. Nous nous servons des autres critères pour trancher :

	log normale	Gamma
stat de Cramer-Von Mises	373	808
AIC	753 871	771 322
Deviance	68 203	81 913

La loi log normale minimise les différents critères. Nous la conservons donc pour la modélisation de ce poste.

- **APPAREILLAGES**

Le modèle utilisant la loi inverse gaussienne ne convergeant pas elle a été exclue pour la modélisation de ce poste. L'analyse graphique concerne donc les lois gamma et log normale.

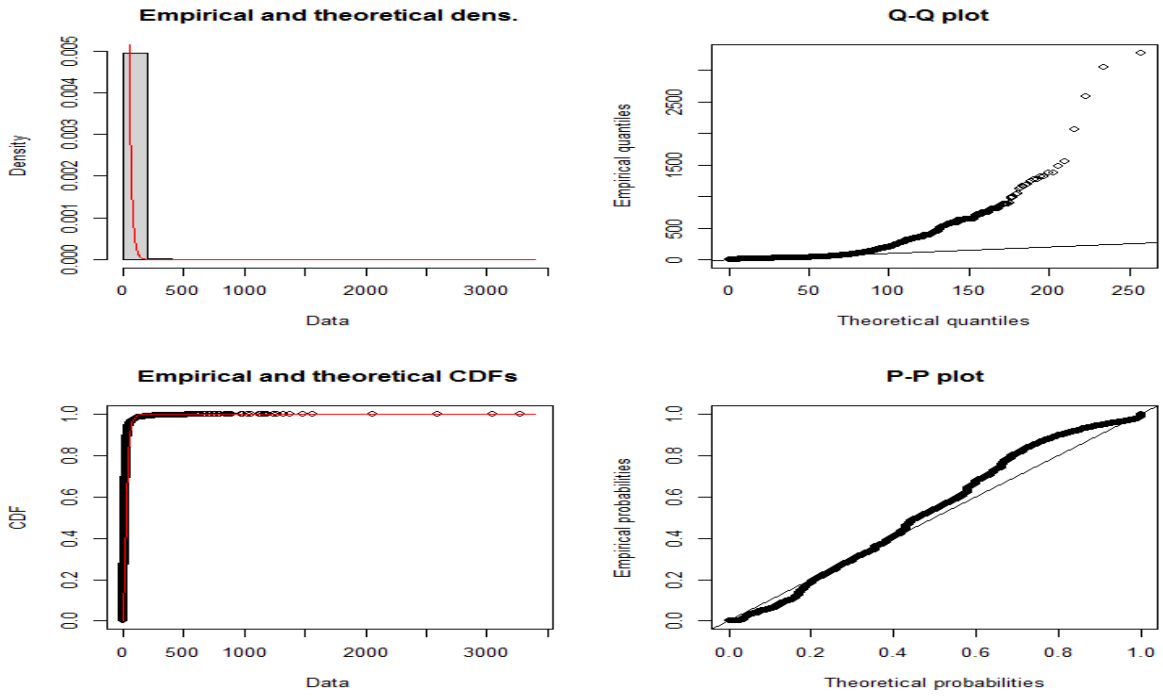


Figure 36 Distribution de la loi gamma Appareillages

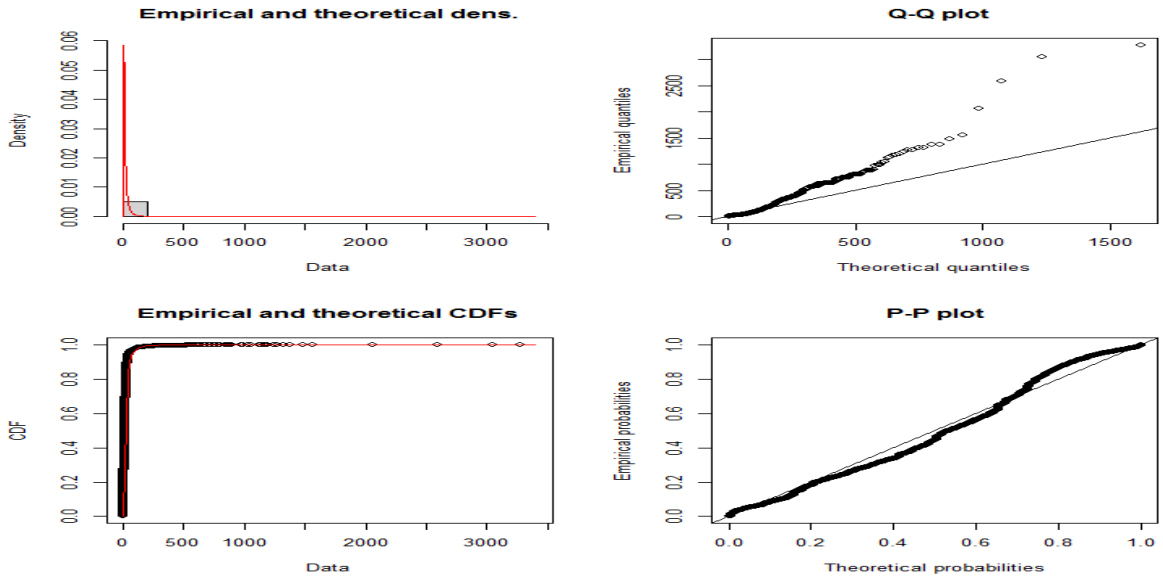


Figure 37 Distribution de la loi log normale Appareillages

L'analyse graphique ne permet pas de sélectionner une loi aux détriments de l'autre. Nous nous en remettons donc aux autres critères de sélection :

	log normale	Gamma
stat de Cramer-Von Mises	148	282
AIC	722 421	731 549
Deviance	100 794	102 569

La loi log normale minimisent les trois critères. Elle est donc conservée pour la modélisation de ce poste.

- **PREVENTION ET BIEN-ETRE**

Le modèle utilisant la loi inverse gaussienne ne convergeant pas elle a été exclue pour la modélisation de ce poste. L'analyse graphique concerne donc les lois gamma et log normale.

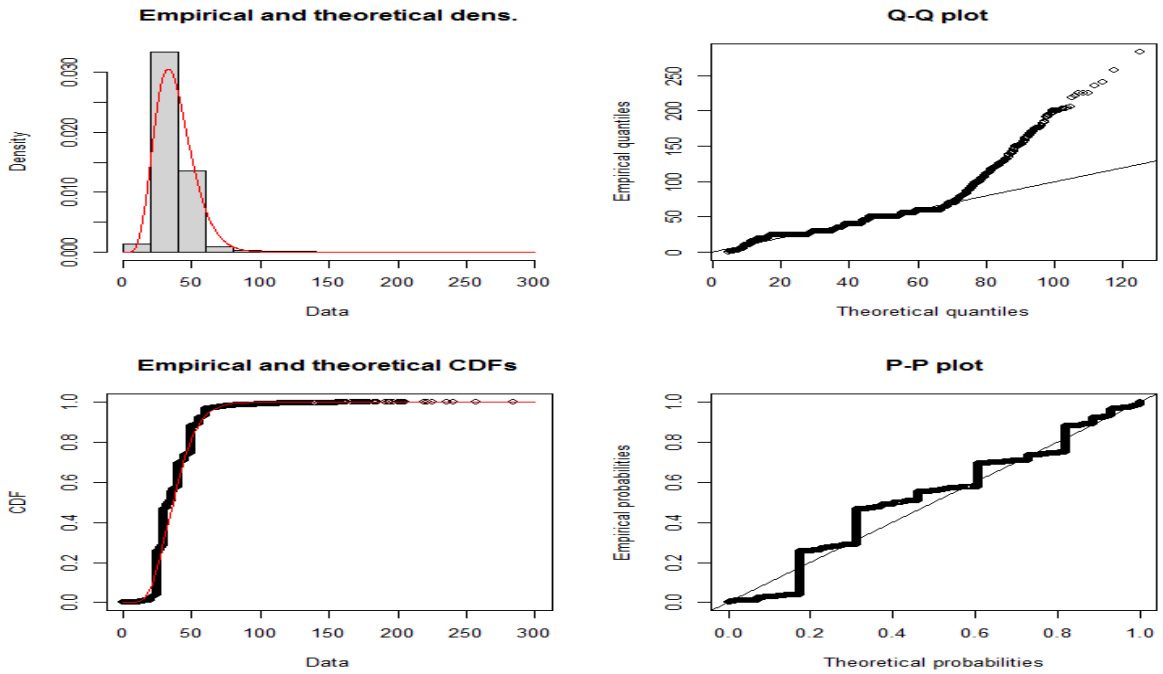


Figure 38 Distribution de la loi gamma Prévention et bien-être

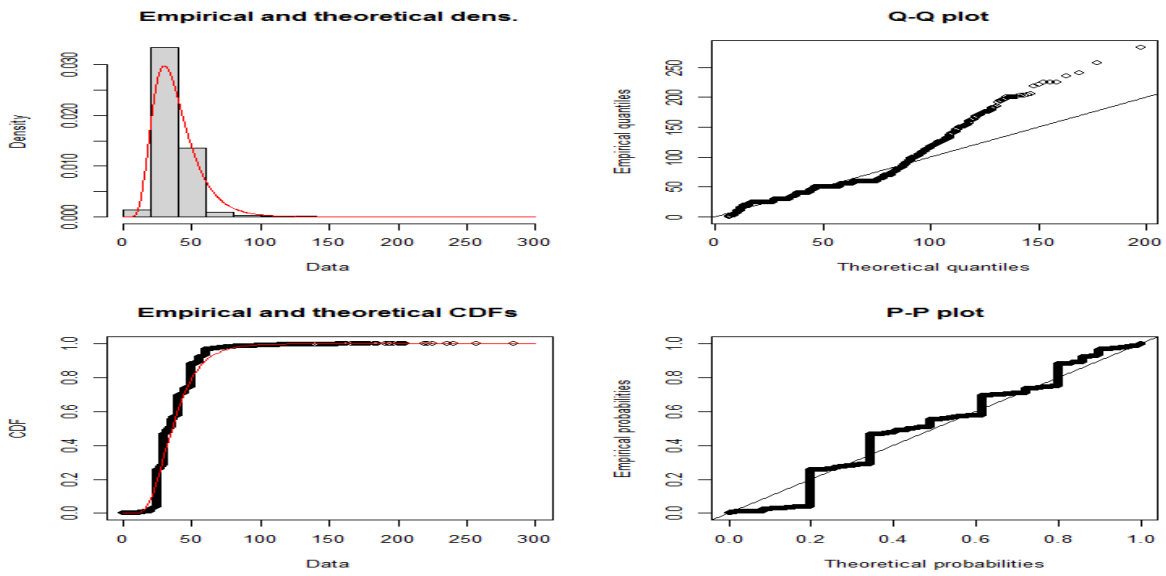


Figure 39 Distribution de la loi log normale Appareillages

L'analyse graphique ne permet pas de sélectionner une loi au détriment de l'autre. Nous nous en remettons donc aux autres critères de sélection :

	log normale	Gamma
stat de Cramer-Von Mises	110	101
AIC	269 315	262 775
Deviance	11 008	4 103

La loi gamma minimise les trois critères. Elles est conservée pour la modélisation.

Récapitulatif des lois par poste

HOSPITALISATION	Inverse gaussienne
SOINS COURANTS	Gamma
PHARMACIE	Gamma
OPTIQUE	Gamma
DENTAIRE	Log normale
APPAREILLAGES	Log normale
PREVENTION ET BIEN-ETRE	Gamma

Sélection des variables

Pour tester la significativité des variables explicatives, nous avons utilisé deux procédures automatiques que sont la fonction « Anova » du package « car » sur R et la fonction « step ».

L'Anova permet de tester un modèle contraint avec moins de variables avec un modèle non contraint. Sous H_0 les variables supplémentaires du modèle non contraint sont supposées nulles et sous H_1 au moins une de ces variables est considérée non nulle. L'Anova va ainsi un test du ratio de vraisemblance de statistique de test :

$$\alpha = -2\log\left(\frac{l(\widehat{\theta_0})}{l(\widehat{\theta})}\right)$$

Avec $l(\widehat{\theta_0})$ la vraisemblance des paramètres θ estimés sous H_0 et $l(\widehat{\theta})$ la vraisemblance des estimateurs des paramètres θ.

La fonction step se base sur l'Aic. Elle teste la suppression de variables qui pourraient détériorer l'ajustement du modèle et donc augmenter l'AIC. On retient à la fin le modèle avec le plus petit AIC .

Dans notre étude, les variables âge, niveau, régime d'affiliation et zone se sont avérées à chaque fois significatives avec l'Anova comme le montre les tableaux ci-dessous. La fonction step a confirmé à chaque fois les résultats issus de l'Anova.

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	838.67	16	< 2.2e-16	***
Zone	28.30	3	3.148e-06	***
Niveau	40.55	5	1.157e-07	***
Régime	22.63	1	1.959e-06	***

Anova Hospitalisation

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	919.57	16	< 2.2e-16	***
Zone	502.41	3	< 2.2e-16	***
Niveau	604.02	5	< 2.2e-16	***
Régime	5.48	1	0.01926	*

Anova Soins courants

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	419.83	16	< 2.2e-16	***
Zone	54.68	3	8.037e-12	***
Niveau	102.16	5	< 2.2e-16	***
Régime	10.67	1	0.4626	*

Anova Pharmacie

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	7551.4	16	< 2.2e-16	***
Zone	199.8	3	< 2.2e-16	***
Niveau	12779.2	5	< 2.2e-16	***
Régime	36.7	1	1.37e-09	***

Anova Optique

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	8733.1	16	< 2.2e-16	***
Zone	649.3	3	< 2.2e-16	***
Niveau	12466.9	5	< 2.2e-16	***
Régime	7.8	1	0.00512	**

Anova Dentaire

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	507.42	16	< 2.2e-16	***
Zone	41.38	3	5.442e-09	***
Niveau	28.91	5	2.415e-05	***
Régime	5.54	1	0.01864	*

Anova Appareillages

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	427.73	16	< 2.2e-16	***
Zone	127.18	3	< 2.2e-16	***
Niveau	1469.23	5	< 2.2e-16	***
Régime	15.30	1	9.193e-05	***

Anova Prévention et bien-être

c. Choix des modèles de fréquence

Généralement, deux modèles servent à la modélisation de la fréquence en assurance non-vie : le modèle Poisson et le modèle Binomiale-négatif.

L'utilisation du modèle de Poisson suppose une absence de sur dispersion dans les données. En effet selon la loi Poisson la variance des données est égale à leur moyenne. On parle de sur dispersion lorsque la variance est supérieure à la moyenne. Ce problème peut engendrer de faibles p valeurs et conduire à une mauvaise interprétation de la significativité des variables explicatives.

Un moyen de vérifier la sur dispersion est d'observer le ratio $déviance\ résiduelle/nombre\ de\ degrés\ de\ liberté$. Un ratio significativement supérieur à 1 indique une sur dispersion.

Un moyen de contourner ce problème est d'utiliser une « loi Binomiale négative ».

On essaie d'avoir une idée de la présence de sur dispersion des données en comparant la variance et la moyenne de nos données par poste.

	Moyenne	Variance
Hospitalisation	0,91	5,58
Soins courants	16,77	1948,12
Pharmacie	37,20	2191,12
Optique	0,64	1,86
Dentaire	0,81	3,47
Appareillages	1,51	18,56
Prévention et bien-être	0,25	0,66

La totalité des postes présentent une variance supérieure à la moyenne ce qui peut laisser penser à une sur dispersion et une mauvaise adéquation de la loi Poisson.

Résultats par poste

- **SOINS COURANTS, PHARMACIE ET APPAREILLAGES**

Pour la modélisation de ces postes on teste les modèles Poisson et Binomiale-négatif. Pour effectuer notre on observe l'AIC, la déviance ainsi que le rapport $déviance\ résiduelle/nombre\ de\ degrés\ de\ liberté$.

	Poisson	Binomiale négative
Déviance	6 126 219	235 358
AIC	6 803 658	1 445 148
Déviance/ddl	30,22	1,16

Soins courants

	Poisson	Binomiale négative
Déviance	6 685 175	242 026
AIC	7 589 729	1 792 752
Déviance/ddl	33,00	1,19

Pharmacie

	Poisson	Binomiale négative
Déviance	1 629 645	160 992
AIC	1 900 207	721 432
Déviance/ddl	8,05	0,80

Appareillages

Ces trois postes montrent une meilleure adéquation de la loi binomiale négative. Elle est donc conservée pour leur modélisation.

- **HOSPITALISATION, OPTIQUE, DENTAIRE ET PREVENTION ET BIEN-ETRE**

Pour ces quatre postes on observait une sur représentation de zéro pour le nombre de sinistres. De ce fait, pour modéliser la fréquence de ces postes, on compare le modèle zero-inflated binomial négatif et le modèle binomial-négatif usuel grâce au test de vuong.

	Vuong Z-test	H1	p-value
RAW	-13.479944	model2>model1	< 2.22e-16
AIC-CORRECTED	-12.867786	model2>model1	< 2.22e-16
BIC-CORRECTED	-9.740095	model2>model1	< 2.22e-16

Test de vuong Hospitalisation

	Vuong Z-test	H1	p-value
RAW	-28.73965	model2>model1	< 2.22e-16
AIC-CORRECTED	-28.46881	model2>model1	< 2.22e-16
BIC-CORRECTED	-27.08503	model2>model1	< 2.22e-16

Test de vuong Optique

Vuong Z-test	H1	p-value

RAW	-18.74815	model2>model1	< 2.22e-16
AIC-CORRECTED	-18.20406	model2>model1	< 2.22e-16
BIC-CORRECTED	-15.42417	model2>model1	< 2.22e-16

Test de vuong Dentaire

	Vuong Z-test	H1	p-value
RAW	-14.91504	model2>model1	< 2.22e-16
AIC-CORRECTED	-14.27691	model2>model1	< 2.22e-16
BIC-CORRECTED	-11.01611	model2>model1	< 2.22e-16

Test de vuong Prévention et bien-être

Dans chaque cas on rejette l'hypothèse H_0 selon laquelle les deux modèles sont équivalents et on suppose que le modèle 2, qui est le modèle zero-inflated, est mieux adapté que le modèle binomiale négatif classique. On conserve donc le modèle zero-inflated binomial négatif pour la modélisation de la fréquence de ces postes.

d. Analyse des résidus

Tests d'adéquation

Pour valider nos modèles on réalise de tests d'adéquation du khi 2 de Pearson et de déviance. Si nos modèles sont bien ajustés, les statistiques de ces tests les statistiques de ces tests convergent vers une loi du khi 2 à n-p degrés de libertés, où p représente le nombre de paramètres à estimer.

Le test sur la déviance démontre une bonne adéquation pour tous les postes modélisés avec une p valeur systématiquement égale à 1.

Le test de Pearson conduit au rejet de l'hypothèse de bonne adéquation des modèles pour la pharmacie et les modèles de fréquence des postes soins courants et appareillages. Des p valeurs à 0 ont été observées pour les modèles concernés.

Il est cependant compliqué de se fier à ces deux tests qui sont sensibles au nombre de données. Nous appuyons donc l'analyse en testant la normalité des résidus.

Normalité des résidus

Une distribution normale centrée et réduite des résidus peut être un indicateur de la bonne adéquation des différents modèles.

- **HOSPITALISATION**

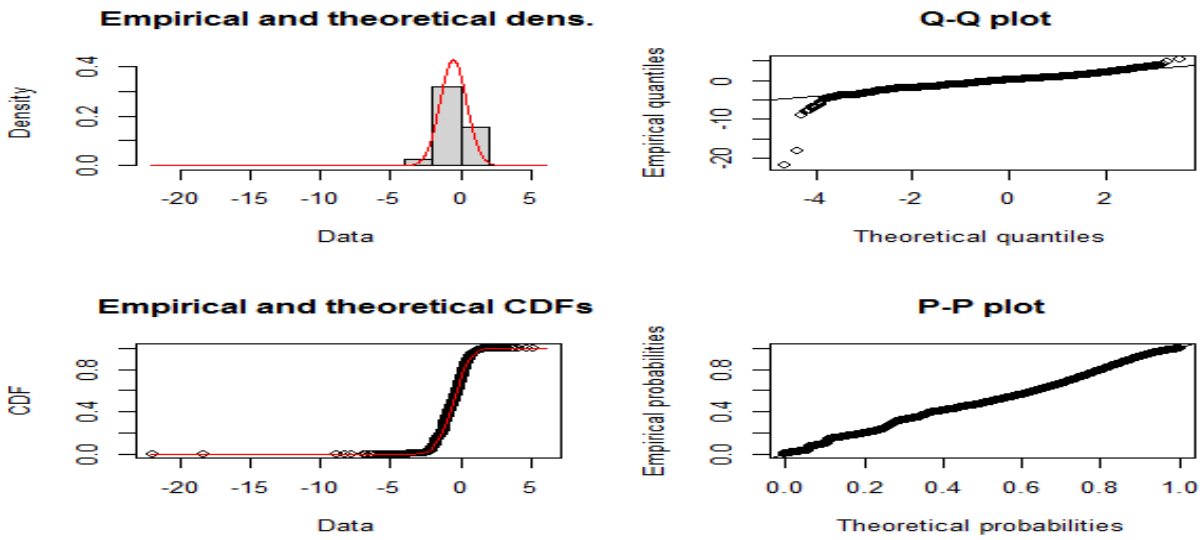


Figure 40 Distribution des résidus standardisés de la modélisation du coût moyen

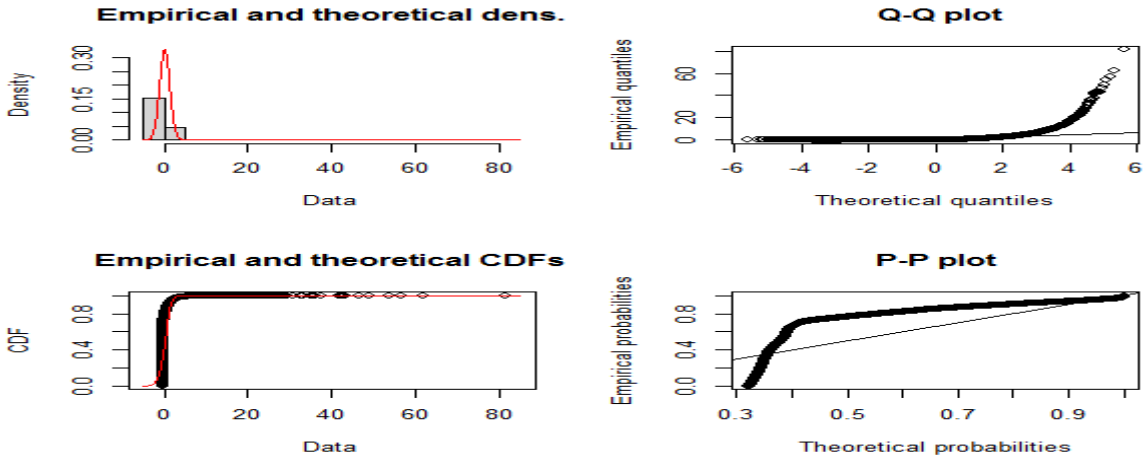


Figure 41 Distribution des résidus standardisés de la modélisation de la fréquence

Les graphes montrent une bonne adéquation à la loi normale surtout pour le coût moyen avec une moyenne proche de 0. Le même constat a été fait sur les autres postes.

3.5 Analyse des résultats

a. Calcul de la prime pure

Soit un assuré ayant 70 ans, affilié au régime générale, habitant en zone 2 et ayant souscrit à une formule de niveau 4.

Considérons que les coefficients issus de la modélisation du poste Hospitalisation sont les suivants :

	Coefficients coût moyen	coefficiecients fréquence	coefficients modèle logistique
Intercept	4,93	-1,36	-1,03
Âge 70 ans	-0,09	1,19	-8,58
Zone 3	0,04	0,05	-0,15
Niveau 4	0,08	0,06	-0,45

Les « coefficients modèle logistique » interviennent pour les postes dont la modélisation a nécessité un modèle Zero-Inflated.

$$coût\ moyen = \exp(4{,}83 - 0{,}09 + 0{,}04 + 0{,}08)$$

$$fréquence = \exp(-1{,}36 + 1{,}19 + 0{,}05 + 0{,}06)$$

$$P(N > 0) = 1 - \frac{\exp(-1{,}03 - 8{,}58 - 0{,}14 - 0{,}45)}{1 + \exp(-1{,}03 - 8{,}58 - 0{,}14 - 0{,}45)}$$

Avec N le nombre de sinistres.

$$Prime\ pure = coût\ moyen * fréquence * P(N > 0)$$

$$= 121{,}5$$

Pour les postes n'ayant pas nécessité de modèle Zero-Inflated :

$$Prime\ pure = coût\ moyen * fréquence$$

b. Cohérence de la prime modélisée

Pour cette partie une nouvelle tarification du produit de référence a été faite en utilisant les résultats du GLM. L'objectif est d'observer si les primes obtenues sont cohérentes en termes de tarifs par zone, par âge et par niveau.

Tarifs par zone

Pour la réalisation des différents modèles deux zoniers ont été utilisé. Tous les résultats sortis plus haut viennent de la modélisation avec le zonier B. Ce choix a été fait car la tarification avec le zonier A présente quelques incohérences.

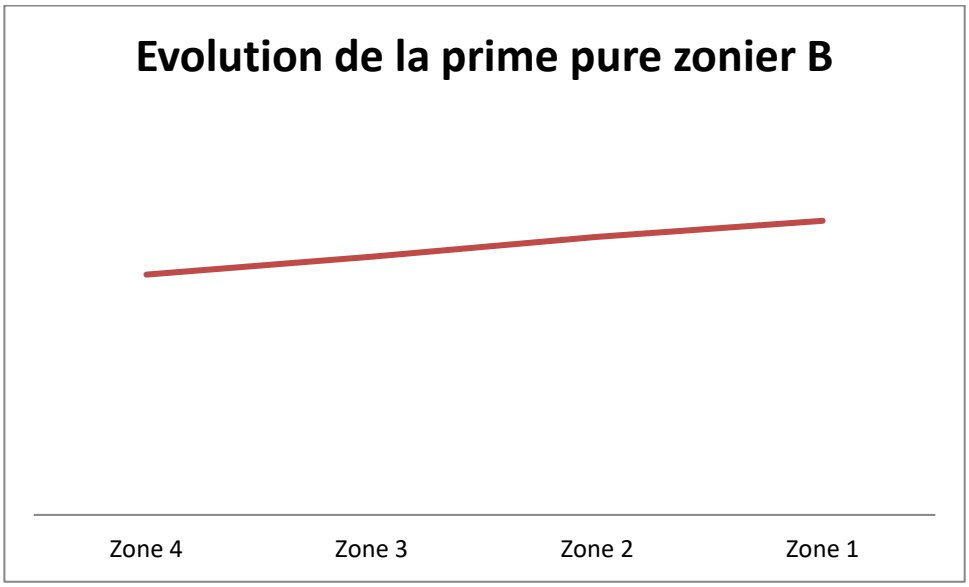


Figure 42 Evolution de la prime pure zonier B

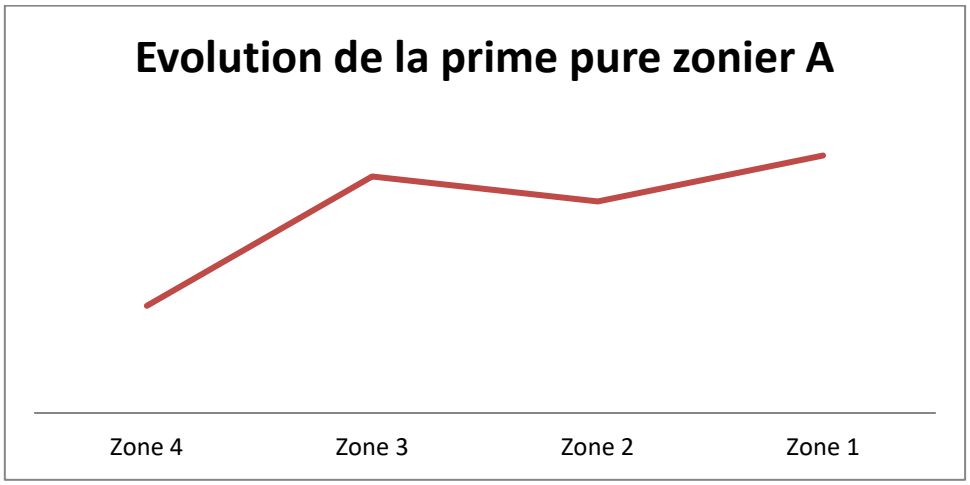


Figure 43 Evolution de la prime pure zonier A

On devrait observer un tarif qui croît entre la zone 4 et la zone 1. Ce n'est pas le cas avec le zonier A où le tarif en zone 3 est supérieur à celui en zone 2. On avait déjà remarqué cette anomalie en réalisant les statistiques descriptives. En effet le coût moyen en zone 3 était supérieur à celui en zone 2. De plus, la variable zone issue de ce zonier ne s'est pas avérée significative lors de la modélisation de certains postes comme l'Hospitalisation. Les méthodes de sélections automatiques que sont l'Anova et la fonction step suggéraient que cette variable n'avait pas un réel impact sur le modèle.

	LR Chisq	Df	Pr(>Chisq)	Significativité
Age	924.26	16	< 2.2e-16	***
Zone	2.09	3	0.5539471	
Niveau	24.72	5	0.0001581	***
Régime	29.55	1	5.442e-08	***

Anova zonier A

Le zonier B ne présentant aucune de ces incohérences, il a été conservé pour réaliser notre tarification.

Tarifs par âge

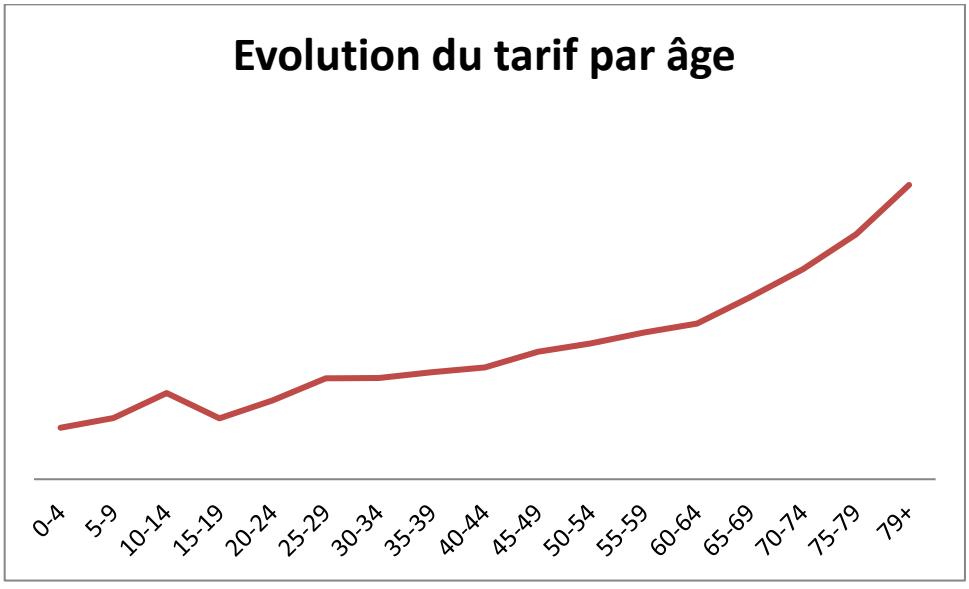


Figure 44 Evolution du tarif par âge

Le tarif devrait être croissant avec l'âge. On l'observe partout sauf lors du passage de 10-14 ans à 15-19 ans. Nous commercialisons des produits pour lesquels tous les assurés de -19 ans ont la même prime. Nous calculons donc une prime pour les -19 ans :

$$= \frac{prime_{0-4} * effectif_{0-4} + prime_{5-9} * effectif_{5-9} + \cdots + prime_{15-19} * effectif_{15-19}}{effectif_{0-19}}$$

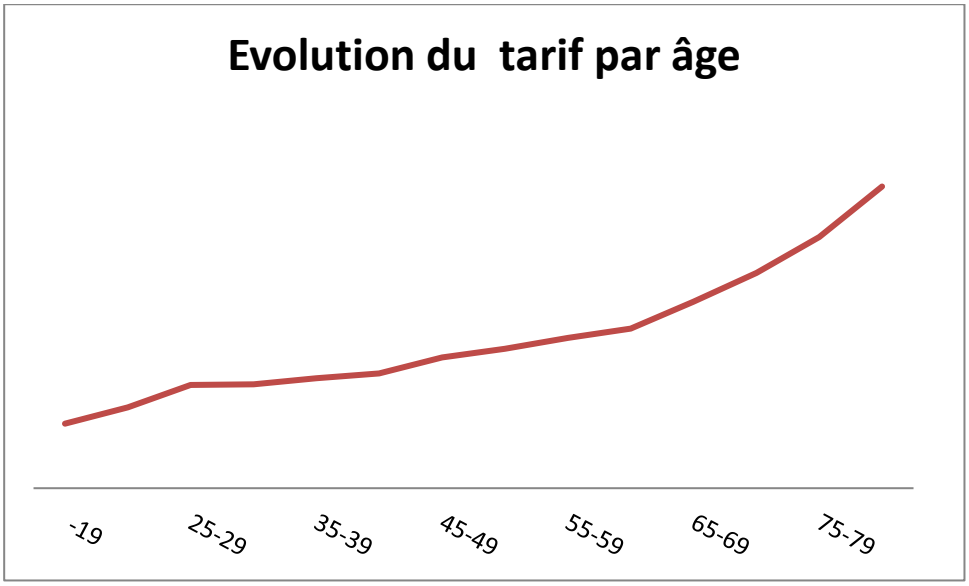


Figure 45 Evolution du tarif par âge retravaillé

Avec cette prime unique pour les moins de 19 ans, on observe bien un tarif qui croit avec l'âge.

Tarifs par niveau de garanties

Pour cette partie, notre tarif serait cohérent s'il augmentait avec le niveau de garanties. En effet plus la garantie est élevée, plus la complémentaire santé est amenée à dépenser pour couvrir l'assuré.

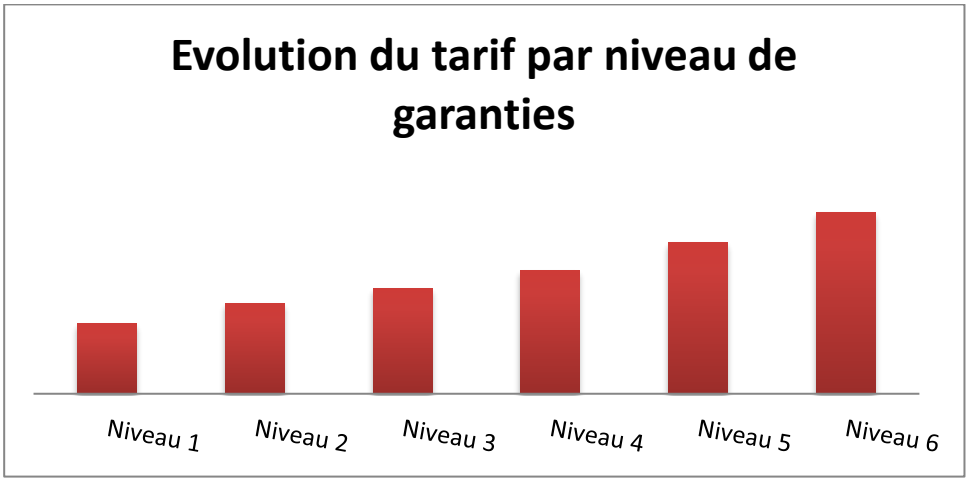


Figure 46 Evolution du tarif par niveau de garanties

Ce graphe montre bien un tarif croissant avec le niveau de garanties ce qui confirme la cohérence de la prime modélisée.

c. Ajustement

Afin de réaliser nos GLM, nous avons créé des classes d'âge de 5 ans. Afin d'obtenir un tarif par âge, quelques ajustements ont dû être opérés. Dans un premier temps une prime unique du nouveau produit issu du GLM et une prime unique du produit de référence (prime déjà commercialisée) ont été calculées. La méthodologie de calcul est la même que celle du calcul de la prime unique pour les moins de 19 ans. Dans un second temps, on applique un coefficient à la prime unique du produit issu du GLM. Ce coefficient représente l'écart entre la prime pure du produit de référence à un âge i et la prime pure unique de ce même produit. Soit α_i ce coefficient. Si i est égale à 60 ans par exemple

$$\alpha_{60} = \frac{Prime\ pure\ produit\ de\ référence_{60\ ans}}{prime\ pure\ unique\ produit\ de\ réfrence}$$

La prime du nouveau produit pour un assuré de 60 ans est donc :

$$Prime\ pure = \alpha_{60} * prime\ pure\ unique\ du\ nouveau\ produit$$

On considère ainsi qu'à un âge i l'écart entre la prime pure du produit de référence à cet âge et la prime pure unique de ce même produit serait le même que l'écart la prime pure au même âge du nouveau produit et sa prime pure unique. Le coefficient α_i étant croissant avec l'âge, cela nous assure une prime pure croissante avec l'âge ainsi qu'un effet âge (écart entre les tarifs d'âges qui se suivent) similaire à celui du produit de référence.

Partie 4 – Application au business plan des Partenariats et différentes limites de l'outil

Dans cette partie nous comparons les tarifs ajustés de notre produit de référence aux tarifs déjà commercialisés.

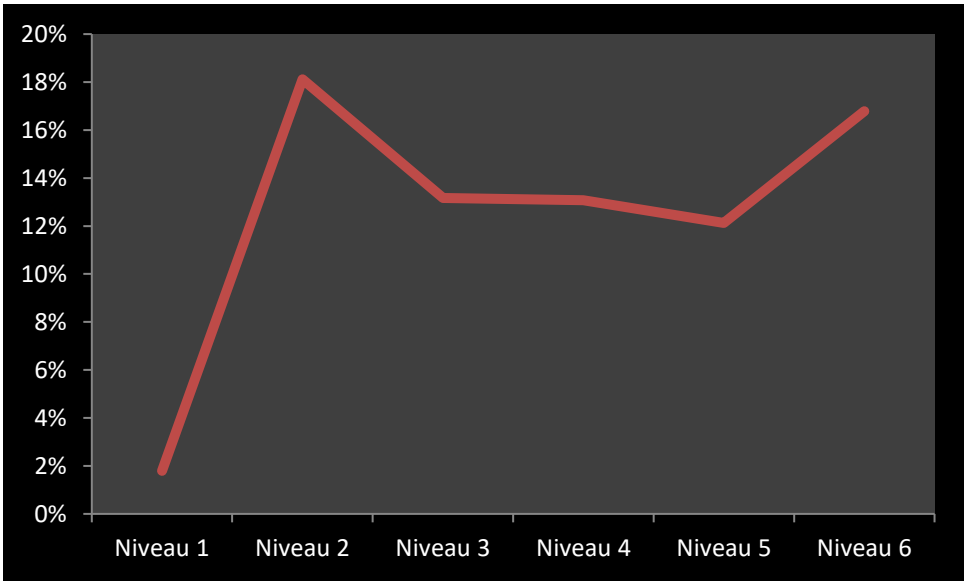


Figure 47 Comparaison du tarif modélisé et du tarif existant

On note ainsi que les primes pures ajustées issues de nos GLM sont supérieures à celles commercialisées. Si l'écart reste assez faible sur le premier niveau de garanties, il est assez notable sur les autres niveaux avec un écart qui varie entre 13% et 18%.

Limites de l'outil

Le principal problème de cet outil est qu'il nous limite dans la réalisation de tarifs sur mesure pour nos partenaires. En effet notre tarification se fait par poste, mais chaque poste présente des sous-postes. Le poste Soins courants a par exemple les consultations, les auxiliaires médicaux, les examens médicaux en laboratoires comme sous-poste. Chaque sous-poste présente des niveaux de remboursement. Pour attribuer un niveau de garantie global au poste (qui a été fait dans la partie avec la création de la variable niveau de garanties) on a donc comparé les niveaux de remboursement de ces sous-postes. On obtient ainsi des niveaux allant de 1 à 6. A chaque niveau correspond des remboursements assez proches. En prenant l'exemple d'une consultation chez le généraliste le niveau 1 correspond à des niveaux de remboursement proche de 100%BR. On peut ainsi classer un remboursement à 105%BR en niveau 1. Cela cause un problème dans la réalisation de tarifs sur mesure. En effet passer de 100%BR à 105%BR doit avoir un effet sur le tarif ce qui n'est pas le cas avec notre outil étant donné que le niveau ne change pas.

Il existe également de niveaux de remboursement que l'on peut appeler « entre deux ». A titre illustratif, si on considère que le niveau 1 correspond à des remboursements proches de 100%BR et le niveau 2 correspond à des remboursements proches de 150%BR, il est difficile d'attribuer un niveau à une garantie qui présente un remboursement à 125%BR.

L'idéal aurait été de réaliser un tarif par sous-poste. Cependant le manque de données pour certains sous-postes nous a conduits à renoncer à cette idée.

CONCLUSION

Dans un contexte où de nombreuses réformes se mettent en place en assurance santé, ce mémoire a permis de mettre en place une tarification par poste de garanties qui tient compte de ces évolutions à travers des majorations effectues sur la consommation et la sinistralité des postes comme le dentaire et l'appareillage. Cette tarification s'est faite grâce aux modèles linéaires généralisés. La sur représentation de « nombre de sinistres=0 » nous a cependant amené à faire appel aux Zero-Inflated GLM pour la modélisation de la fréquence de certains. Le test de Vuong nous a permis de montrer que ces modèles sont effectivement mieux adaptés que des GLM classiques dans cette situation. Les tarifs que nous avons pu tirer de nos modèles ce sont avérés être cohérents en termes d'évolution par zone, par âge ou par niveau de garanties.

Tables des illustrations

ANNEXE

BIBLIOGRAPHIE

1- **Tarification d'une complémentaire santé à destination des séniors , modulaire par poste de garanties et l'impact sur la solvabilité : Fatemeh ABDOLLAHI mémoire d'actuariat.**

2- **Impact de la réglementation 100% Santé sur la tarification d'un contrat d'assurance santé individuel : Wael Redouane AALILOU, Mémoire d'actuariat.**

3- **Elaboration d'un racier et tarification des produits en assurance santé animale : Sébastien LEFEVRE Mémoire d'actuariat.**

4- **Construction d'un zonier en Santé en utilisant les méthodes de lissage : Le krigeage. Chérif Amadou SOW Mémoire d'actuariat.**

5- **Maud THOMAS. Econométrie de l'assurance non-vie.**

6- **www.drees.sante.gouv.fr.**

7- **L'assurance maladie en ligne. www.ameli.fr**.

8- **http://eric.univlyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_ZIP_Regression_R_Python.pdf**