

Mémoire présenté le :

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

Par : Youssef JABIR

Titre : Mise à jour d'un outil de tarification santé

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

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

Membres présents du jury de l'Institut  
des Actuaires


Signature

**ACTELIOR**  
Entreprise : **Actélior**  
Nom : 7 bis rue des Aulnes  
69410 Charnay-agne au Mont d'O  
Signature : Tel. : 04 78 66 30 00  
mail : actelior@actelior.com

Membres présents du jury de  
l'ISFA

Directeur de mémoire en entreprise :

Nom : Elodie Paget

Signature : 

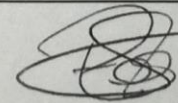
Invité :

Nom :

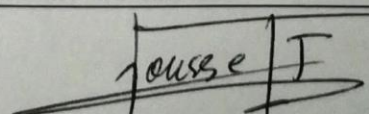
Signature :

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

Signature du responsable entreprise



Signature du candidat



Le 31/12/2021

## Remerciements

Je souhaiterais en premier lieu exprimer toute ma gratitude à David ECHEVIN pour m'avoir recruté dans son équipe. Je voudrais remercier également les personnes qui ont contribué tout au long de la construction de mon mémoire, en particulier à ma tutrice d'entreprise Elodie PAGET et ma collègue Julie VERNIN-BIANCALE qui m'ont fait confiance et m'ont aidé dans la consolidation du sujet.

Je tiens aussi à remercier toute l'équipe du cabinet Actélior pour son accueil et sa disponibilité depuis mon arrivée et pour son accompagnement dans la réalisation mon mémoire.

Je remercie également Monsieur Yahia SALHI, mon tuteur universitaire, qui m'a suivi dans le cadre de mon alternance et dans la réalisation de ce mémoire.

Finalement, je tiens à remercier ma famille, partie importante de mon projet, qui malgré la distance, m'a apporté tout son soutien durant les moments difficiles.

## Résumé

Le marché de la complémentaire santé a connu ces dernières années des grands bouleversements aussi bien dans la prise en charge que dans le coût des actes. Ces évolutions nécessitent une mise à jour régulière des bases de tarification de manière à adapter les tarifs aux nouvelles réglementations.

Actélior a mis en point un outil de tarification permettant de calculer des primes à travers l'approche dite « Coût moyen x Fréquence ». Pour produire des tarifs qui sont cohérents par rapport à la réalité et adaptés aux nouvelles réglementations, il a été nécessaire de mettre à jour régulièrement ces modèles mathématiques utilisés dans l'outil du cabinet.

Le processus de cette mise à jour est trop long vu que le nombre d'actes à modéliser dépasse 70, ce qui demande alors la recherche de plus de 140 modèles tout en prenant en compte les caractéristiques de chacun de ces actes. Une fois la modélisation des actes est terminée sur R, tous les résultats des modèles doivent être mises en place sur l'outil de tarification utilisé par l'entreprise et qui est programmé sur Excel.

Ce mémoire propose de résoudre cette problématique qui est la tarification des actes de santé en travers la mise à jour des modèles des coûts moyens et des fréquences en utilisant la méthode « Coût moyen x Fréquence » basée sur les modèles linéaires généralisés. L'avantage cette fois par rapport aux anciennes mises à jour est la création d'un algorithme permettant de faire tout le processus automatiquement à partir du traitement des données brutes jusqu'à l'exportation des coefficients qui seront prêts à utiliser directement sur fichier Excel de l'outil de tarification du cabinet.

### **Mots clés :**

Complément santé, coût moyen, fréquence, modèles linéaires généralisés, gamlss, outil de tarification.

# Abstract

The complementary health insurance market has undergone major changes in recent years, both in terms of coverage and in the cost of procedures. These changes require regular updates of the pricing bases in order to adapt the rates to the new regulations.

Actélior has developed a pricing tool that makes it possible to calculate premiums using the "Average Cost x Frequency" approach. In order to produce rates that are consistent with reality and adapted to new regulations, it has been necessary to regularly update the mathematical models used in the firm's tool.

The process of this update is too long since the number of procedures to be modeled exceeds 70, which then requires the search for more than 140 models while taking into account the characteristics of each of these procedures. Once the modeling of the acts is completed on R, all the results of the models must be implemented on the pricing tool used by the company and which is programmed on Excel.

This thesis proposes to solve this problem, which is the pricing of health care procedures, by updating the average cost and frequency models using the "Average Cost x Frequency" method based on generalized linear models. The advantage this time compared to the previous updates is the creation of an algorithm that allows the whole process to be done automatically from the processing of the raw data to the export of the coefficients that will be ready to use directly on the Excel file of the pricing tool of the firm.

## **Key words :**

Complément santé, coût moyen, fréquence, modèles linéaires généralisés, gamlss, outil de tarification.

# Sommaire

Remerciements .....	2
Résumé.....	3
Abstract.....	4
Sommaire .....	5
Liste des abréviations .....	9
Introduction .....	10
Partie I - L'assurance santé en France.....	11
1 La sécurité sociale.....	11
2 Le régime général.....	12
3 L'Assurance Maladie.....	14
5 La complémentaire santé .....	16
A. Les différents organismes complémentaires.....	16
B. Mécanisme de remboursement.....	16
C. Les réglementations importantes .....	18
i. L'Accord National Interprofessionnel.....	18
ii. Le contrat responsable.....	18
iii. Réforme « reste à charge zéro » ou « 100% santé » .....	22
D. Les facteurs influençant les dépenses .....	31
i. Niveau de garantie .....	31
ii. Type de contrat .....	31
iii. Bénéficiaires couverts.....	31
Partie II - Analyse et description du portefeuille.....	33
1 Informations sur le portefeuille.....	33
2 Etude des garanties et des codes actes .....	35
3 Etude des FR et BR moyens.....	37
4 Caractéristiques du portefeuille étudié .....	37
A. Données à disposition.....	37

B.	Etude du Turn-over .....	39
C.	Etude du portefeuille.....	41
i.	Répartition par année et par type ayant droit.....	41
ii.	Répartition de l'exposition par âge et par type d'assuré ...	42
iii.	Répartition des enfants par famille.....	43
iv.	Répartition par département.....	44
D.	Zoom sur les gammes étudiées.....	47
v.	Exposition en fonction des gammes étudiées.....	47
vi.	Démographie par gamme étudiée .....	48
vii.	Répartition de l'exposition par niveau de gamme et année	49
viii.	Démographie par niveau de gamme .....	50
E.	Conclusion .....	51
5	Etude de la base des effectifs .....	52
6	Etude de la base de prestations.....	55
7	Fusion des bases des prestations et des effectifs .....	57
8	Statistique descriptive de la base définitive.....	58
Partie III - Construction de l'outil de tarification .....		65
1	Introduction : Intelligence Artificielle (AI) .....	65
2	Apprentissage supervisé .....	66
A.	Classification :.....	68
B.	Régression :.....	68
3	Modèles de régression.....	69
A.	Régression non linéaire (NLR) :.....	70
B.	Régression linéaire (LR) :.....	71
i.	Régression linéaire simple.....	71
ii.	Régression linéaire multiple.....	72
iii.	Estimation des paramètres.....	72
4	Modèle Linéaire Généralisé (GLM) .....	77
A.	Introduction.....	77

i.	Distribution aléatoire .....	77
ii.	Fonction lien.....	78
B.	Test d'adéquation des lois .....	79
i.	Critères BIC et AIC.....	79
5	Application.....	81
A.	Statistiques sur les variables explicatives.....	81
B.	Hypothèses : fréquence $\perp$ coût moyen .....	82
C.	Intégration du PPAP .....	83
D.	Modélisation de la fréquence.....	84
E.	Modélisation du Coût Moyen .....	89
F.	Modèle final .....	95
G.	Impact de la crise sanitaire sur les tarifs.....	97
H.	Reste des libellés actes .....	100
I.	Intégration dans l'outil de tarification.....	100
Limites	.....	103
A.	Choix de la base de remboursement.....	103
B.	Prise en compte du plafond .....	103
C.	Prise en compte des variables explicatives .....	104
D.	Lien entre les actes.....	104
E.	Variable : taux de remboursement .....	104
F.	Niveau de gamme .....	105
G.	Actes avec forfait.....	105
Conclusion	.....	107
Table des illustrations	.....	108
Annexe	.....	110
Annexe 1	: Familles & libellés actes.....	110
Annexe 2	: Liste des lois continues positives de 'gamlss' .....	112
Annexe 3	: Liste des lois discrètes de 'gamlss'.....	114
Références	.....	117





## Liste des abréviations

- **FR** : Frais Réels
- **TM** : Ticket Modérateur
- **BR** : Base de Remboursement
- **RSS** : Remboursement de la Sécurité Sociale
- **BRySS** : Remboursement du Complément santé y compris le remboursement de la Sécurité Sociale.
- **FR** : Frais Réels
- **PASS/PMSS** : Plafond Annuel/Mensuel de la Sécurité Sociale
- **RO** : Remboursement Obligatoire
- **RC** : Remboursement du Complément santé
- **RAC** : Reste à Charge
- **AT/MP** : Accident du Travail - Maladies Professionnelles
- **CM** : Coût moyen
- **DPTAM** : Dispositif de Pratique Tarifaire Maitrisée.
- **OPTAM CO** : Option Pratique Tarifaire Maitrisées / Chirurgie Obstétrique
- **CAS** : Contrat d'Accès aux Soins.

# Introduction

Ces dernières années, le marché de l'assurance santé en France a connu de nombreuses réformes réglementaires telles que l'ANI (Accord National Interprofessionnel) en 2017 qui a généralisé la complémentaire santé dans le secteur privé et transféré une partie du marché de l'assurance santé individuelle vers l'assurance santé collective. La dernière réforme nommée « réforme 100% santé » se met en place progressivement depuis le 1er janvier 2019. Elle permet notamment aux assurés d'avoir accès à certains soins (en dentaire, optique et audiologie) sans reste à charge.

La première partie de ce mémoire dessine le cadre général de l'étude. Nous présenterons d'une part le contexte de l'assurance santé en France et d'autre part les nouvelles modalités de la réforme 100% santé à prendre en considération dans notre outil de tarification.

La deuxième partie sera consacrée à la présentation des traitements effectués sur les données utilisées, notamment la base des effectifs et la base des prestations, puis la préparation des tables de correspondances nécessaires pour le regroupement des actes de soins. Ensuite, nous concluons cette partie par la présentation du portefeuille utilisé dans la construction des modèles sous forme d'une étude statistique descriptive de ses différentes composantes.

Enfin, la dernière partie est dédiée au cadre théorique de la tarification santé et à l'explication du fonctionnement et de l'utilisation des modèles linéaires généralisés dans la tarification de produits d'assurance santé à travers la méthode « coût x fréquence ».

Nous allons par la suite mettre en pratique la méthode de régression linéaire généralisée (GLM) en appliquant sur une partie de la base (un acte de soins précis) issue des traitements présentés dans la deuxième partie pour objectif de trouver sa prime pure. Pour atteindre cet objectif, nous allons vérifier les hypothèses d'application, ensuite de modéliser le coût moyen et la fréquence d'utilisation de cet acte. Une fois la tarification est réalisée, nous allons effectuer un « Back-testing » afin de vérifier la précision du modèle trouvé.

# Partie I - L'assurance santé en France

Ce mémoire s'intéresse à la tarification des régimes complémentaires individuels.

Dans cette partie, nous présentons le fonctionnement de l'Assurance Santé en France, les différents régimes complémentaires ainsi que les mécanismes de remboursements. Nous nous attarderons notamment sur les évolutions réglementaires qui encouragent les services actuariels à mettre à jour leurs outils de tarification.

## 1 La sécurité sociale

La sécurité sociale est un service public de l'Etat créé en 1945. Elle permet de garantir pour chacun qu'en toutes circonstances, il disposera des moyens nécessaires pour assurer sa subsistance ainsi que celle de sa famille dans des conditions décentes.

Cet organisme a pour objectif d'assurer d'une manière obligatoire la protection des individus face aux conséquences financières de certains risques de la vie (décès, maladie, accident) ou certaines situations (charges familiales, vieillesse) en les accompagnant à chaque étape de leur existence. Elle repose sur le principe de la solidarité entre les individus, chacun participe à son financement selon ses moyens à travers une cotisation dite charge sociale.



▲ Figure 1 : Décomposition des régimes de la Sécurité Sociale.

La sécurité sociale comporte deux régimes principaux et des régimes spéciaux qui couvrent une ou plusieurs catégories socioprofessionnelles spécifiques et se caractérisent par des modalités de gestion et de prise en charge différentes :

- **Régime général** : couvre la majorité de la population (4 personnes sur 5) notamment les travailleurs salariés, les travailleurs indépendants (depuis le 1<sup>er</sup> janvier 2018) et les étudiants (depuis le 1<sup>er</sup> septembre 2019).
- **Le régime agricole** : couvre les exploitants et salariés exerçant des professions dans le domaine agricole.

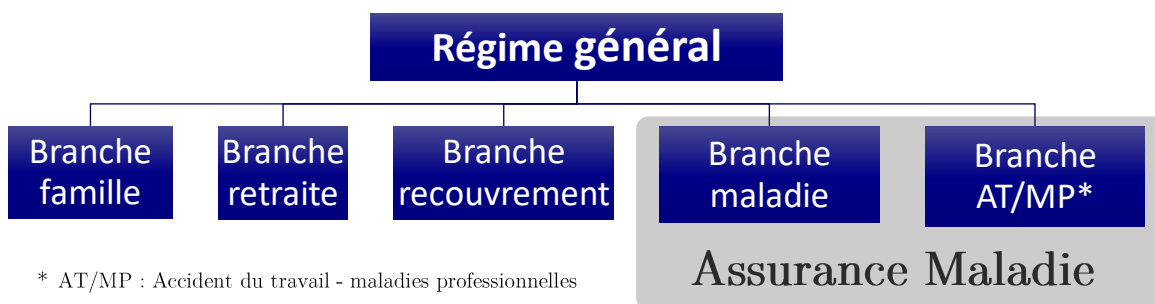
- **Régimes spéciaux** : on retrouve par exemple le régime d'Alsace-Moselle dont les habitants de cette région bénéficient d'un régime social spécifique qui se révèle plus avantageux. Nous retrouvons aussi d'autres régimes spéciaux comme le régime des agents de la SNCF, de la RTAP, etc.

Dans la suite de ce mémoire, nous nous intéresserons au régime général.

## 2 Le régime général

Le régime général de la sécurité sociale se compose de cinq branches qui ont à leur charge de gérer un ou plusieurs 'risques' ; ces risques sont définis comme des événements qui peuvent, au cours d'une vie, porter atteinte à la sécurité économique d'une personne dans un esprit de solidarité nationale.

### Présentation du régime général



▲ Figure 2 : Décomposition du régime général de la sécurité sociale.

- **La branche famille** : L'objectif est de réduire les inégalités de niveau de vie entre les familles en prenant en compte le nombre d'enfants.
- **La branche retraite** : Elle permet de verser les pensions aux retraités de l'industrie, des services et du commerce.
- **La branche de recouvrement** : Elle collecte les cotisations et contributions sociales pour les redistribuer au bénéfice des autres branches.

Dans la suite de ce mémoire, nous nous intéresserons à l'Assurance Maladie qui se compose de la « branche maladie » et de la « Branche Accident du travail – maladies professionnelles ».

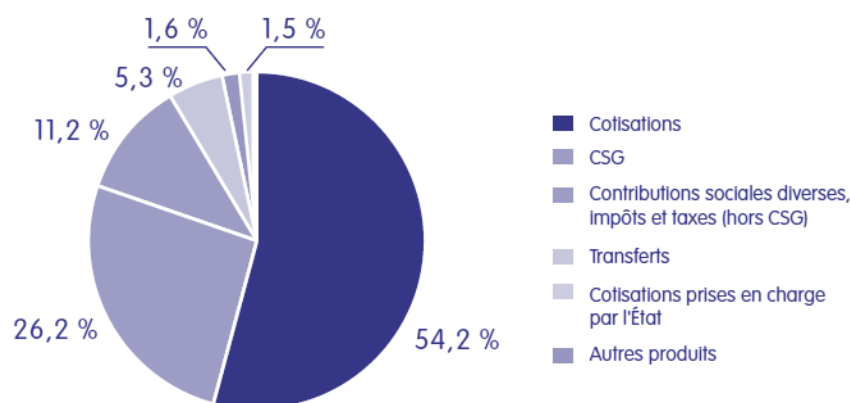
- **La branche maladie** : Elle assure la prise en charge des dépenses de santé des assurés et garantit l'accès aux soins. Elle favorise la prévention et contribue à

la régulation du système de santé français. Elle couvre les risques maladie, maternité, invalidité et décès.

- **La branche accident du travail - maladies professionnelles** : Elle gère les risques professionnels auxquels sont confrontés les travailleurs : accidents du travail, accidents de trajet et maladies professionnelles. À ce titre, elle indemnise les victimes et fixe la contribution respective des entreprises au financement du système. Elle met en œuvre une politique de prévention des risques professionnels.

### Financement du régime général

Le financement du régime général de sécurité sociale est assuré principalement par les cotisations et contributions assises sur les rémunérations qui représentent 80% des ressources totales. Les principaux financeurs sont les ménages à travers une part salariale et les entreprises à travers une part patronale. Il est financé également par les administrations publiques à travers les impôts, les taxes et d'autres contributions sociales diverses.



▲ Figure 3 : Répartition du financement du régime général.

- **Les cotisations** sont calculées à partir de taux fixés à l'échelon national et sont à la charge pour partie de l'employeur, et pour partie du salarié.
- **Les impôts et taxes affectés (ITAF)** sont des prélèvements obligatoires explicitement affectés au financement de la protection sociale, parmi lesquels la cotisation sociale généralisée (CSG) qui représente à elle seule plus de la moitié des ITAF.
- **La contribution sociale généralisée (CSG) et la contribution pour le remboursement de la dette sociale (CRDS)** sont prélevées sur les revenus

d'activité, les revenus de remplacement, les revenus du patrimoine, les produits de placement et des jeux.

### 3 L'Assurance Maladie

L'Assurance Maladie est la branche du régime général de la sécurité sociale qui couvre tout ou une partie des risques liés à la maladie, à la maternité et aux accidents de la vie privée et professionnelle, les maladies professionnelles et les invalidités. En santé, elle prend en charge cinq postes principaux :

- La pharmacie :

Les médicaments sont pris en charge sur prescription délivrée par un médecin, une sage-femme, un chirurgien-dentiste, un directeur ou directeur adjoint de laboratoire d'analyses. Il existe quatre taux de remboursement des médicaments :

- **100%** : pour les médicaments irremplaçables et coûteux.
- **65%** : pour les médicaments à service médical rendu majeur ou important.
- **30%** : pour les médicaments à service médical rendu modéré et certaines préparations magistrales.
- **15%** : pour les médicaments à service médical faible, les médicaments homéopathiques et les préparations magistrales homéopathiques (PMH).

- Les consultations :

Le remboursement de l'Assurance Maladie des consultations/visites d'un généraliste ou d'un spécialiste varie en fonction de la base de remboursement.

Ce remboursement est impacté aussi par le respect du parcours de soins : une consultation hors parcours de soins implique un remboursement minoré.

Si l'assuré a déclaré un médecin traitant, le taux de remboursement d'une consultation d'un généraliste est de 70% de la base de remboursement qui est de 25€. Hors parcours de soins, ce taux baissera à 30%.

Notons que différentes majorations (visite à domicile, consultation de nuit, jours fériés, ...) s'appliquent sur les tarifs de base. Le remboursement de ces majorations par la sécurité sociale peut être total, partiel ou nul.

- Le dentaire

Les principales familles d'actes remboursés par l'Assurance Maladie en dentaire sont les soins dentaires, les prothèses dentaires et les traitements d'orthodontie.

Les soins dentaires sont remboursés à 70% de la base de remboursement qui atteint 23€ généralement

Pour les actes prothétiques, les bases de remboursement ont beaucoup évolué ces dernières années du fait notamment de la réforme « 100% santé ». Des prix limite de vente sont notamment entrés en vigueur ces dernières années.

Les traitements d'orthodontie sont pris en charge par l'Assurance Maladie sous condition d'avoir un accord préalable de la caisse d'Assurance Maladie et que ces traitements commencent avant le 6<sup>ème</sup> anniversaire.

- L'optique

Ce poste concerne les lunettes de vue (montures et verres), les lentilles et les actes chirurgicaux des yeux.

Jusqu'à la fin de l'année 2019, la prise en charge par l'Assurance Maladie était modérée et le reste à charge conséquent pour les assurés. Depuis le 1<sup>er</sup> janvier 2020, les opticiens proposent un panier « 100% santé » composé d'une sélection de lunettes de vue de qualité qui sont pris en charge par l'Assurance Maladie et la complémentaire santé.

- L'hospitalisation

En cas d'hospitalisation, certains frais sont pris en charge par l'Assurance Maladie. Cependant, le forfait hospitalier, le ticket modérateur, les dépassements d'honoraires et les éventuelles dépenses supplémentaires pour une chambre particulière restent à la charge de l'assuré. Le reste à charge sera totalement ou partiellement pris en charge par la complémentaire santé.

## 5 La complémentaire santé

L'adhésion à une assurance complémentaire santé permet la prise en charge d'une partie ou de l'intégralité des dépenses de santé non remboursées par le régime obligatoire de la sécurité sociale.

Dans cette section, nous allons présenter les différents types d'organismes complémentaires, puis faire un point sur la réglementation et enfin expliquer le mécanisme de remboursement réparti entre l'assuré, l'Assurance Maladie et la complémentaire santé.

### A. Les différents organismes complémentaires

Nous distinguons 3 types d'organismes complémentaires : Cette complémentaire doit être souscrite auprès d'un des organismes suivants :

- **Les mutuelles régies par le code de la mutualité.** Ce sont des organismes à but non lucratif menant dans l'intérêt de leurs membres, moyennant le versement d'une cotisation, des actions de prévoyance, de solidarité et d'entraide.
- **Les institutions de prévoyance régies par le code de la sécurité sociale.** Ce sont des sociétés de droit privé à but non lucratif, qui gèrent des contrats collectifs d'assurance de personnes couvrant les risques de maladie, incapacité de travail et invalidité, dépendance et décès.
- **Les sociétés d'assurance régies par le code des assurances.** Ce sont des sociétés à but lucratif qui proposent à leurs assurés des contrats d'assurance dans divers secteurs tels que la santé, la prévoyance, l'assurance de biens ou encore la retraite.

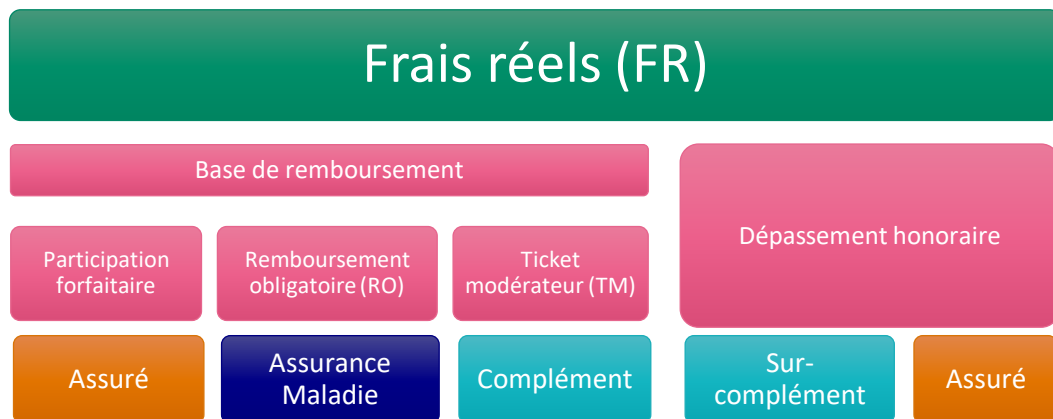
### B. Mécanisme de remboursement

La couverture du risque santé se décompose généralement en 4 parties :

- Couverture de base obligatoire financée par l'Assurance Maladie (obligatoire),
- Couverture complémentaire financée par l'organisme assureur (obligatoire ou facultative, collective ou individuelle).



- Couverture sur-complémentaire financée par l'organisme assureur également. Ce dernier garantit un remboursement supplémentaire de tout ou partie du reste à charge de l'assuré suite à l'intervention des deux remboursements cités ci-dessus :



▲ Figure 4 : Mécanisme de remboursement

- **Le reste à charge financé par l'assuré directement**

Lors d'un sinistre santé, l'Assurance Maladie ne couvre pas nécessairement la totalité des Frais Réels (FR) payés par l'assuré. Elle rembourse un pourcentage du tarif de convention appelé Base de Remboursement (BR). Ce taux varie en fonction des codes actes et médicaments, de la santé de l'assuré, du respect ou non du parcours de soins coordonnés.

Notons que l'assuré paie la participation forfaitaire (1 €) pour toutes les consultations ou actes réalisés par un médecin, examens radiologiques ou analyses, dès que le patient est majeur. L'assuré paie aussi un montant de franchise de 0,5 € par médicament ou par acte paramédical. Il paie également un montant de 2 € par transport sanitaire.

#### Exemple : Consultation médecin de secteur 2

Un assuré a fait une visite chez un médecin (spécialiste) et a payé 50€ : **Frais Réels (FR) = 50€**.

La base de remboursement (BR) est égale à 23€. Le taux de remboursement de la sécurité sociale est 70%. La participation forfaitaire dans ce cas est de 1€. On suppose dans cet exemple que le remboursement de la mutuelle est de 200% BR y compris le remboursement de la sécurité sociale. Alors :

- Participation forfaitaire (PF) = 1€ (à la charge de l'assuré).

- Remboursement obligatoire (RO) =  $70\% \times 23\text{€} - 1\text{€ (PF)} = 15,10\text{€}$
  - Remboursement Complémentaire santé (RC) : Minimum entre :
    - Taux RC x BR - RO - PF =  $200\% \times 23\text{€} - 1\text{€} - 15,1\text{€} = 29,9\text{€}$
    - FR - PF - RO =  $50\text{€} - 1\text{€} - 15,1\text{€} = 33,9\text{€}$
- ⇒ RC = 29,9€

Il est intéressant de noter ici que la complémentaire santé rembourse 29.9€ dont  $(100\%-70\%) * 23\text{€} = 6,9\text{€}$  de Ticket Modérateur (TM).

- Reste à charge = FR - RO - RC =  $50\text{€} - 15,1\text{€} - 29,9\text{€} = 5\text{€}$

Pour récapituler : consultation chez médecin de secteur 2 de FR =50€

- RO = 15,1€
- RC = 29,9€
- Reste à charge = 5€

### C. Les réglementations importantes

Les contrats de complémentaire santé sont soumis à certains textes réglementaires qui conditionnent la grille des garanties. Nous traiterons ci-dessous les trois principales dernières réformes à savoir l'ANI, le 100% santé et le contrat responsable.

#### i. L'Accord National Interprofessionnel

L'accord national interprofessionnel (ANI) du 11 janvier 2013 stipule que toute entreprise du secteur privé a l'obligation de souscrire à un contrat frais de santé pour l'ensemble de ses salariés (avant le 1<sup>er</sup> janvier 2016 pour le régime général et avant le 1<sup>er</sup> juillet 2016 pour le régime local).

Les principales contraintes sont la mise en place de garanties minimales (proches du contrat responsable sauf pour les prothèses dentaires, l'orthodontie et l'optique où les garanties sont supérieures) ainsi que la participation à hauteur de 50% minimum pour l'entreprise.

#### ii. Le contrat responsable

En 2018, 95% des contrats présents sur le marché sont des contrats dits « responsables ». En effet, afin de diminuer les dépenses de santé, la réforme de l'Assurance Maladie du 1<sup>er</sup> janvier 2006 a créé le contrat dit « responsable » dont les critères ont connu des modifications notamment en 2015.

Ci-dessous, nous présenterons les critères du contrat responsable avant et après 2015, puis nous citerons les avantages fiscaux et sociaux de ce type de contrat.

#### L'ancien contrat responsable :

La notion de contrat « responsable » a été introduite par la loi du 13 août 2004 qui a pour objectif de responsabiliser à la fois les assurés, les praticiens et les assureurs. En effet, pour inciter les complémentaires santé à mettre en conformité leurs contrats, des avantages fiscaux et sociaux sont mis en place. Aussi, les contrats collectifs responsables sont exonérés de cotisations sociales pour la participation de l'employeur et sont déductibles de l'impôt sur le revenu pour la participation de l'employé dans la limite de certains plafonds.

Au cours de cette décennie, le taux de la Taxe Spéciale sur les Conventions d'Assurance des contrats responsables a régulièrement fluctué pour passer de 0% à 3.5% puis enfin 7%. La contribution CMU de 5.90% a également été transformée en taxe de 6.27% en 2015.

Ainsi en 2015, la Taxe de Solidarité Additionnelle correspondant à la somme de la Taxe Spéciale sur les Conventions d'Assurance et de la contribution CMU convertie en taxe, a été créée.

Pour être **responsable**, un contrat doit, avant avril 2015 :

1- Garantir au minimum :

- 30 % de la base de remboursement des consultations du médecin traitant dans le cadre du parcours de soins coordonnés.
- 30 % de la base de remboursement des médicaments remboursables à 65 %.
- 35 % de la base de remboursement des examens de biologie médicale prescrits par le médecin traitant.
- Le ticket modérateur d'un minimum de deux prestations de prévention définies par décret et considérées comme prioritaires au regard des objectifs de santé publique.

2- Ne pas rembourser :

- Les dépassements et majorations liés au non-respect du parcours de soins coordonnés.
- La participation forfaitaire de 1 € applicable aux consultations et certains examens médicaux.
- Les franchises applicables sur les médicaments, les actes paramédicaux et les frais de transport.

## Le nouveau contrat responsable à partir d'avril 2015

Un nouveau contrat « responsable » a été défini par le décret n° 2014-1374 du 18 novembre 2014. Ainsi, pour pouvoir bénéficier des aides fiscales et sociales liées aux contrats responsables, un contrat doit, à partir du 1<sup>er</sup> avril 2015 :

### 1- Garantir au minimum :

- La prise en charge du ticket modérateur pour l'ensemble des dépenses de santé sauf pour les frais de cures thermales, l'homéopathie et les médicaments dont le service médical a été classé faible ou modéré.
- La prise en charge du forfait journalier hospitalier sans limitation de durée.

### 2- Garantir au maximum :

- 125% (100% à partir de 2017) des dépassements d'honoraires des médecins qui n'adhèrent pas au DPTAM<sup>1</sup>, dispositif ayant pour objet la maîtrise des dépassements d'honoraires des professionnels de santé conventionnés. La garantie doit également être inférieure à celle des dépassements d'honoraires de médecins qui adhèrent au dispositif d'au moins 20%.

### 3- Respecter des plafonds et des planchers en optique

- Les planchers et plafonds sont définis selon la correction optique. Le renouvellement de l'équipement ne peut se faire que tous les 2 ans (sauf pour les enfants ou si évolution de la vue pour les adultes).

Les planchers et plafonds du contrat responsable en optique sont définis ci-dessous :

Complexité	Type d'équipement	Plancher	Plafond
Simple	Equipement à verres simple foyer dont la sphère est comprise entre -6,00 et + 6,00 dioptries et dont le cylindre est inférieur ou égal à + 4,00 dioptries	50 €	470 €
	Un verre de type "a" et un verre de type "c"	125 €	610 €
Complexe	Equipement à verres simple foyer dont la sphère est hors zone de -6,00 à + 6,00 dioptries ou dont le cylindre est supérieur à + 4,00 dioptries et à verres multifocaux ou progressifs	200 €	750 €
	Un verre de type "a" et un verre de type "f"	125 €	660 €
	Un verre de type "c" et un verre de type "f"	200 €	800 €
Très Complexe	Equipement pour adulte à verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est hors zone de -8,00 à + 8,00 dioptries ou à verres multifocaux ou progressifs sphériques dont la sphère est hors zone de -4,00 à + 4,00 dioptries	200 €	850 €

Quel que soit l'équipement, la monture est prise en charge, au maximum, à hauteur de 150€ par le régime complémentaire.

▲ Figure 5 : Plancher et plafond en fonction de type d'équipement

<sup>1</sup> DPTAM : Dispositif de Pratique Tarifaire Maîtrisée.

Par exemple, pour deux verres de type « a », la complémentaire santé devra garantir un remboursement de l'équipement compris entre 50€ et 470€ ou à hauteur du ticket modérateur. De plus, le remboursement de la monture est plafonné à 150 €.

Un observatoire des prix et de la prise en charge en optique médicale est chargé d'analyser les prix de vente, les caractéristiques, la qualité des équipements d'optique ainsi que les niveaux de couverture proposés par les complémentaires santé.

### **Les avantages fiscaux et sociaux du nouveau contrat responsable**

Les contrats dits « responsables » présentent plusieurs avantages fiscaux et sociaux :

- 1- Taxe de solidarité additionnelle abaissée de 20,27% à 13,27%.
- 2- S'ils sont souscrits dans le cadre d'un contrat collectif et obligatoire, il y a une exonération des charges sociales pour les contributions versées par l'employeur (dans la limite d'un certain plafond).
- 3- S'ils sont souscrits dans le cadre d'un contrat collectif et obligatoire, il y a une déduction de la part salariale dans le calcul de l'impôt sur le revenu.

En 2018, 95% des complémentaires santé sont en conformité avec le nouveau contrat responsable.

### **Les évolutions qui accompagnent la création des contrats responsables**

Le contrat responsable façonne donc le marché de la santé. Tout assuré choisissant une complémentaire santé responsable est certain d'avoir un minimum de soins garantis :

- En optique, les plafonds instaurés par le contrat responsable ont permis de faire diminuer le remboursement des complémentaires santé sur ces postes et ainsi de diminuer les prix de vente. En effet, avant 2015, il était courant de voir des montures à plus de 300 € prises en charge intégralement par la mutuelle. Les montures sont maintenant remboursées à hauteur de 150 € maximum, les praticiens ont donc été obligés de s'adapter et de diminuer leurs tarifs.
- Le nombre de médecins adhérents au DPTAM a augmenté du fait du remboursement supérieur par les complémentaires santé. Ceci permet le contrôle des tarifs. Les médecins autorisés à pratiquer des honoraires libres signent alors un contrat avec l'Assurance Maladie (durée d'un an, renouvelable tacitement) dans lequel ils s'engagent à respecter le taux de dépassement moyen fixé ainsi que le pourcentage minimum d'activité à tarif opposable (sans dépassement) définis au

regard de leur pratique tarifaire observée. Ils sont 49,6% à avoir adhéré au DPTAM en 2018.

### iii. Réforme « reste à charge zéro » ou « 100% santé »

La réforme 100% santé, dont les différents décrets ont été signés entre 2018 et 2019, vient apporter un complément aux contrats responsables. Cette réforme, votée dans le cadre du PLFSS<sup>2</sup> 2019 bouleverse le secteur de la santé.

L'objectif de cette réforme est d'améliorer l'accès aux soins de qualité et de renforcer la prévention. Le reste à charge diminue sur de nombreux produits en optique, dentaire et audiologie (les 3 postes présentant actuellement le plus de reste à charge).

Pour la première fois, les professionnels (opticiens, audioprothésistes et dentistes) sont directement soumis à des lois de santé publique. Cette réforme se met progressivement en place entre le 1<sup>er</sup> janvier 2019 et le 1<sup>er</sup> janvier 2023.

#### Principes

La réforme « 100% santé » prévoit la création de différents paniers de soins en optique, audiologie et dentaire dont un panier de soins « 100% santé » pour chacun de ces postes. Ce panier permet une prise en charge intégrale du soin ou de l'équipement par la Sécurité sociale et la mutuelle pour tous les Français bénéficiant d'une complémentaire santé responsable ou de la CMU-C.

Les professionnels de santé (opticiens, audioprothésistes et dentistes) doivent obligatoirement proposer un devis normalisé avec un produit du panier « 100% santé », libre à l'assuré de le choisir ou non.

Les paniers « 100% santé » ont été définis par les professionnels de santé concernés. Ils mettent en place des dispositifs pour proposer un large choix de produits de qualité et répondre aux attentes de chacun. Les soins et équipements de ces paniers vont se voir attribuer un prix limite de vente que le praticien ne pourra pas dépasser. Leurs bases de remboursement évolueront également, elles seront souvent revues à la hausse.

Si l'assuré ne souhaite pas acheter un équipement ou soin de ce type, il peut se tourner vers les paniers de soins à tarifs libres ou négociés. Ces équipements ou soins vont, pour certains, voir leur base de remboursement évoluer. De plus, des plafonds et/ou planchers vont parfois être mis en place pour la complémentaire santé dans le cadre des contrats responsables.

En parallèle de la création de ces paniers de soins, certaines nomenclatures sont modifiées et des actions de prévention vont être mises en place. De plus, des enquêtes de satisfaction vont être créées et permettront pour certains postes

---

<sup>2</sup> PLFSS : Plan de Financement de la Sécurité sociale

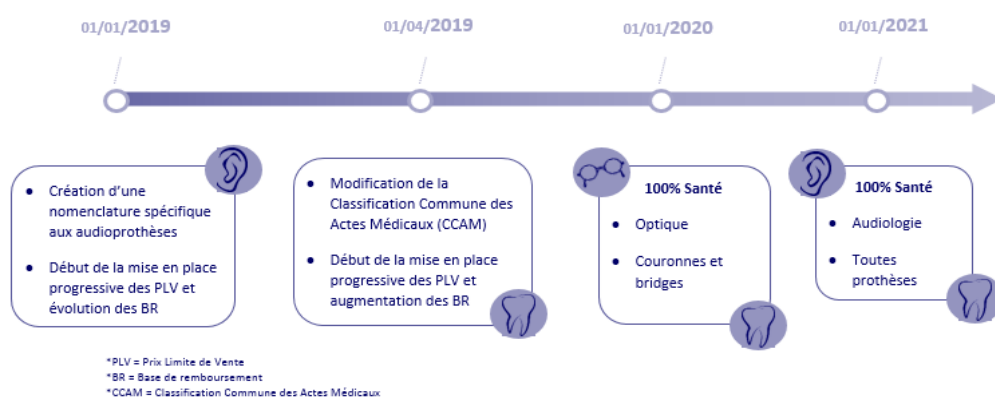
d'éventuellement ajuster les bases de remboursement et prix limites de vente en 2023.

### Impact pour les contrats non responsables

Les réévaluations des bases de remboursement et la mise en place de prix limites de vente vont impacter tous les contrats et donc par ricochet les contrats non responsables. De plus, la modification des nomenclatures oblige les complémentaires santés à mettre à jour leur système de gestion.

### Calendrier de mise en œuvre

La réforme « 100% santé » se met en place par étape, depuis le 1<sup>er</sup> janvier 2019 en audiologie, le 1<sup>er</sup> avril 2019 en dentaire et le 1<sup>er</sup> janvier 2020 pour l'optique. Ci-dessous la chronologie de la mise en œuvre de la réforme par l'intermédiaire d'une frise :



▲ Figure 6 : Calendrier de mise en application de la réforme 100% santé

### A - Optique

La réforme « 100% santé » en optique concerne uniquement les montures et les verres de lunettes de vue. Elle est entrée en vigueur intégralement au 1<sup>er</sup> janvier 2020 et est régie par trois textes réglementaires :

- l'avis du 21 juin 2018,
- l'arrêté du 3 décembre 2018
- le décret du 11 janvier 2019.

### Les paniers de soins

Deux paniers de soins distincts sont créés :

- Le panier « 100% santé » également appelé « panier classe A » présente des verres pour toutes les corrections avec des prix limites de vente ainsi que des montures d'un prix maximum de 30 euros. Ces équipements répondent à un

cahier des charges strict et sont pris en charge intégralement par la Sécurité sociale et la complémentaire santé. Les verres de classe A doivent par exemple présenter un antireflet (milieu de gamme) et un amincissement suffisant pour avoir une esthétique correcte.

- Le panier « classe B », aux tarifs libres, voit ses bases de remboursement revues à la baisse : 5 centimes d'euros par verre et par monture quelle que soit la correction. Les équipements du panier B doivent répondre à un cahier des charges plus strict que celui du panier A.

Il est possible pour l'assuré de faire un panachage : choisir des verres de « classe A » et une monture de « classe B » par exemple.

### Une nouvelle nomenclature

En parallèle de la création de ces deux paniers de soins, toutes les nomenclatures évoluent, il n'y a plus de distinction adulte / enfant. Des intervalles de dioptries plus précis sont créés et pour une même correction, deux codes actes correspondent : un code du panier A avec un prix limite de vente et un code du panier B avec une base de remboursement égale à 5 centimes.

### Evolution des contrats responsables

Plusieurs évolutions du contrat responsable accompagnent également la création de ces deux paniers :

- Les verres sont toujours classés en 3 types : « simple », « complexe » et « très complexe » mais leur définition a quelque peu évolué (évolution des seuils de dioptries).
- Les plafonds de remboursement des contrats responsables sont revus à la baisse : moins 50 euros. C'est en fait le plafond des montures qui est modifié et passe de 150 euros à 100 euros. Ceci n'impacte que les verres et montures du « panier B » car les plafonds ne seront jamais dépassés en « classe A » du fait des prix limites de vente.
- Les planchers de remboursement des contrats responsables ne concernent pas les équipements de « classe A » puisque la prise en charge est intégrale.

Ci-dessous, les nouveaux plafonds et planchers des contrats responsables en optique :



Complexité	Ancienne description de l'équipement	Nouvelle description de l'équipement	Ancien = Nouveau plancher	Ancien plafond	Nouveau plafond
Simple	Equipement à verres simple foyer dont la sphère est comprise entre -6,00 et + 6,00 dioptries et dont le cylindre est inférieur ou égal à + 4,00 dioptries	Equipement à verres unifocaux sphériques dont la sphère est comprise entre -6,00 et + 6,00 dioptries ; Equipement à verres unifocaux sphéro-cylindriques dont la sphère est comprise entre -6,00 et 0 dioptries et dont le cylindre est inférieur ou égal à + 4,00 dioptries ; Equipement à verre unifocaux sphéro-cylindriques dont la sphère est positive et dont la somme S (sphère + cylindre) est inférieure ou égale à 6,00 dioptries	50 €	470 €	420 €
	Un verre de type "a" et un verre de type "c"		125 €	610 €	560 €
Complexe	Equipement à verres simple foyer dont la sphère est hors zone de -6,00 à + 6,00 dioptries ou dont le cylindre est supérieur à + 4,00 dioptries et à verres multifocaux ou progressifs	Equipement à verres unifocaux sphériques dont la sphère est hors zone de -6,00 à + 6,00 dioptries ; Equipement à verre unifocaux sphéro-cylindriques dont la sphère est comprise entre -6,00 et 0 dioptries et dont le cylindre est supérieur à + 4,00 dioptries ; Equipement à verre unifocaux sphéro-cylindriques dont la sphère est inférieure à -6,00 dioptries et dont le cylindre est supérieur ou égal à 0,25 dioptrie ; Equipement à verre unifocaux sphéro-cylindriques dont la sphère est positive et dont la somme S est supérieure à 6,00 dioptries ; Equipement à verres multifocaux ou progressifs sphériques dont la sphère est comprise entre -4,00 et + 4,00 dioptries ; Equipement à verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est comprise entre -8,00 et 0,00 dioptries et dont le cylindre est inférieur ou égal à + 4,00 dioptries ; Equipement à verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est positive et dont la somme S est inférieure ou égale à 8,00 dioptries	200 €	750 €	700 €
	Un verre de type "a" et un verre de type "f"		125 €	660 €	610 €
	Un verre de type "c" et un verre de type "f"		200 €	800 €	750 €
Très Complexe	Equipement pour adulte à verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est hors zone de -8,00 à + 8,00 dioptries ou à verres multifocaux ou progressifs sphériques dont la sphère est hors zone de -4,00 à + 4,00 dioptries	Equipement à verres multifocaux ou progressifs sphériques dont la sphère est hors zone de -4,00 à + 4,00 dioptries ; Equipement à verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est comprise entre -8,00 et 0 dioptries et dont le cylindre est supérieur à + 4,00 dioptries ; Equipement à verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est inférieure à -8,00 dioptries et dont le cylindre est supérieur ou égal à 0,25 dioptrie ; Equipement à verres multifocaux ou progressifs sphéro-cylindriques dont la sphère est positive et dont la somme S est supérieure à 8,00 dioptries	200 €	850 €	800 €

Monture prise en charge au maximum à hauteur de 100€ par équipement (régime obligatoire + régime complémentaire).

▲ Figure 7 : Nouveaux plafonds et planchers contrat responsable optique

Il est important de préciser qu'un verre pourrait être considéré comme « simple » en 2018 et « complexe » en 2020, par exemple.

### Evolution des règles de renouvellement du contrat responsable

Pour que le contrat soit responsable, la prise en charge d'un équipement ne peut se faire, au-delà du ticket modérateur, qu'une fois tous les 2 ans pour un même bénéficiaire. Il existe cependant des conditions de dérogation au renouvellement tous les deux ans. Elles ont évolué en 2020 :

Jusqu'en 2019	A partir de 2020
Le renouvellement annuel est possible en cas de changement de vue.	Le renouvellement annuel est possible en cas de dégradation des performances oculaires d'un delta supérieur ou égale à 0,5 dioptrie (ceci correspond à une faible évolution de la vue).
Le renouvellement annuel est possible pour tous les enfants de moins de 18 ans.	Le renouvellement annuel est autorisé pour les enfants de moins de 16 ans (6 mois dans certains cas pour les enfants de moins de 6 ans), D'autres cas particuliers peuvent être remboursés par la complémentaire santé annuellement, ils ne sont pas listés ici, ils concernent notamment certaines maladies liées à l'œil.

▲ Figure 8 : Conditions de renouvellement : optique

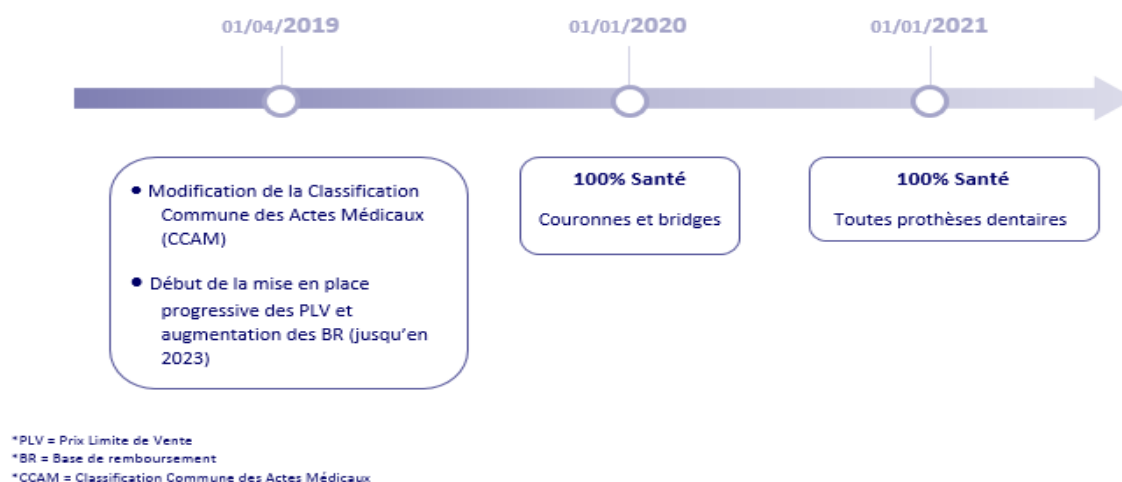
### Principales règles de vente

Les opticien-lunetiers sont également confrontés à des règles de vente strictes :

- Tous les verres de « classe A » et « classe B » doivent être référencés. Le praticien peut vendre un verre de « classe B » uniquement s'il propose également un devis avec un verre de « classe A » de correction identique. Il en est de même pour les montures, tout devis doit comprendre au moins une offre composée d'un équipement de « classe A ».
- Le praticien doit proposer au moins 35 montures de « classe A » pour adultes et 20 pour enfants. Plus précisément, 18 modèles adultes et 10 modèles enfants déclinés au maximum en deux couleurs.
- Pour les verres de « classe A », un remplacement total ou partiel de la monture en cas de casse pendant une période de 2 ans doit être garanti. Les verres progressifs sont également garantis dans les 3 premiers mois suivant l'achat (période d'adaptation).
- Des prestations d'appairage et de filtres doivent également être proposées par le praticien.

## B - Dentaire

En dentaire, la réforme « 100% santé » est régie par l'arrêté du 20 août 2018. Elle concerne les prothèses dentaires et les soins conservateurs et se met en place selon le calendrier suivant :



▲ Figure 9 : Mise en place progressive du dispositif « 100% santé » en dentaire du 01/04/2019 au 01/01/2023

### Les paniers de soins pour les prothèses

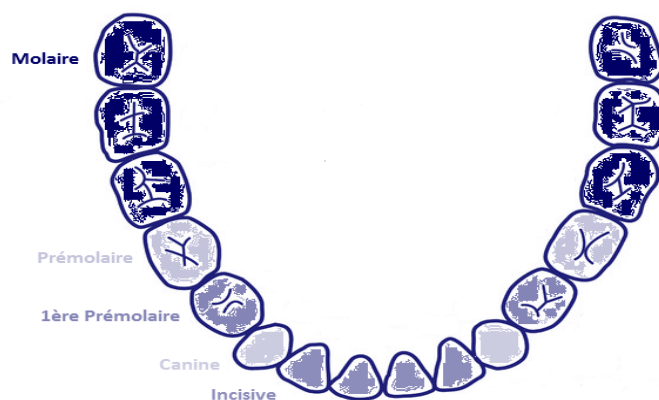
Depuis le 1<sup>er</sup> avril 2019, les prothèses dentaires sont référencées et classées selon trois paniers correspondant à trois niveaux qualitatifs :

- Le **panier de soins « 100% santé »** aussi appelé « Panier 1 » est pris en charge intégralement par l'association du Régime Obligatoire et du Régime Complémentaire depuis 2020 pour les couronnes et bridges et à partir de 2021 pour les autres prothèses dentaires. Des honoraires limites de facturation vont être associés à chacun des actes de ce panier progressivement entre 2019 et 2023. Pour plus de simplicité, à l'image des postes optique et audiologie, il y aura des prix limites de vente.
- Le **panier de soins à tarifs négociés**, aussi appelé « Panier 2 » présente également des prix limites de vente progressifs entre 2019 et 2023. Il n'y a aucune obligation de reste à charge nul sur ce panier pour les contrats responsables.

- Le **panier à Tarifs Libres**, aussi appelé « Panier 3 », n'aura pas de prix limite de vente. Il verra tout de même, comme les deux autres paniers, ses bases de remboursement évoluer progressivement entre 2019 et 2023.

	Couronnes Métalliques	Couronnes en Zircon	Couronnes Céramiques	Bridges Céramiques
RAC* avant le 100% santé	0 €	115 €	195 €	650 €
RAC* = 0 depuis 01/01/2020 pour :	Toutes les dents	<b>Incisive, Canine et Prémolaire</b> (et 1ère Prémolaire)	<b>Incisive, Canine et 1ère Prémolaire</b>	Incisive et <b>Canine</b>
Prix plafond depuis 01/01/2020	290 €	440 €	500 €	1 465 €

\*RAC : Reste à charge



▲ Figure 10 : Prise en charge à 100% des couronnes et bridges du panier 1 depuis 2020 (source : Ameli)

### La nouvelle nomenclature

En parallèle de la création de ces 3 paniers, la nomenclature a évolué le 1<sup>er</sup> avril 2019 : 6 codes actes prothétiques ont été supprimés et remplacés par 14 nouveaux codes. Cela permet une description plus précise de l'acte selon le matériau, la visibilité de la dent ou non et le panier.

Ainsi, 180 codes actes sont répartis en 3 paniers comme suit :

- 32% dans le panier « 100% santé »
- 37% dans le panier 2
- 31% dans le panier 3.

### Evolutions pour les soins conservateurs

La réforme « 100% santé » s'intéresse également à la revalorisation des soins conservateurs (par exemple, le traitement d'une carie ou un détartrage). Depuis avril 2019, une évolution progressive des bases de remboursement des soins conservateurs a été mise en place : 110 soins conservateurs (sur 218) voient leurs bases de remboursement évoluer, parmi lesquels, 83 à la hausse.

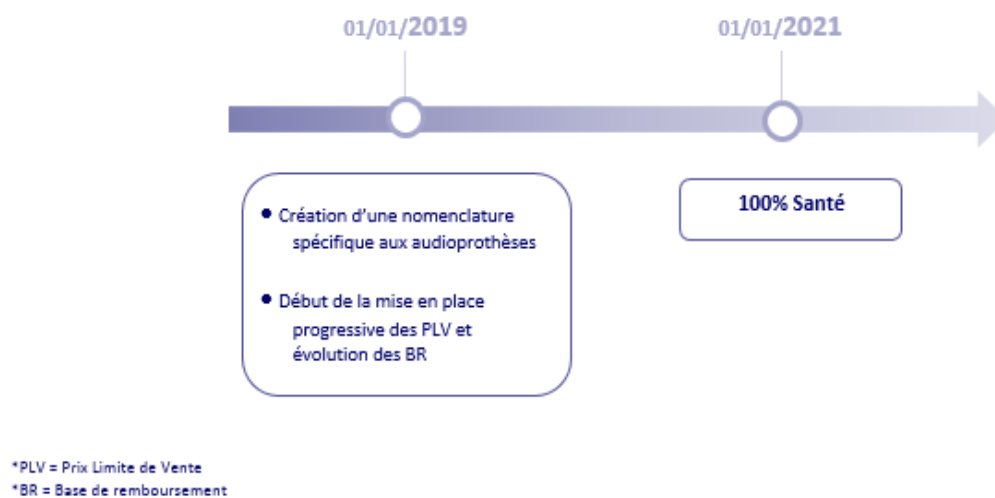
De nouveaux codes actes ont été créés et remboursés par la Sécurité sociale.

### **C - Audiologie**

En audiologie, ce sont uniquement les prothèses auditives et les piles qui sont concernées par la réforme « 100% santé ». Elle se met en place progressivement entre le 1<sup>er</sup> janvier 2019 et le 1<sup>er</sup> janvier 2022. Celle-ci est régie par :

- l'arrêté du 14 novembre 2018
- l'avis du 28 Novembre 2018
- le Décret du 11 janvier 2019.

Les trois types de prothèses auditives existants sont concernés par la réforme : « contour d'oreille », « contour à écouteur déporté » et « intra auriculaire ».



▲ Figure 11 : Mise en place progressive du dispositif « 100% santé » en audiologie du 01/01/2019 au 01/01/2021

### Les paniers de soins pour les prothèses

Depuis le 1<sup>er</sup> janvier 2019, les prothèses auditives sont référencées et classées selon deux paniers :

- Le **panier de soins « 100% santé »** aussi appelé « Panier 1 » est pris en charge intégralement par l'association du Régime Obligatoire et du Régime

Complémentaire responsable depuis le 1<sup>er</sup> janvier 2021. Des prix limites de vente sont également mis en place en 2019 avec des montants dégressifs de 2019 à 2021.

Pour être référencées dans ce panier, les prothèses auditives doivent respecter 12 critères de « base » et 3 critères parmi les 8 proposés dans la liste A (sauf exceptions).

- **Le panier à Tarifs Libres**, aussi appelé « Panier 2 », n'aura pas de prix limite de vente, un dispositif de contrôle des tarifs est tout de même mis en place par la profession. De plus, pour les contrats responsables, le remboursement total du régime obligatoire et du régime complémentaire est plafonné à 1700€ par oreille depuis le 1<sup>er</sup> janvier 2021. Pour être référencées dans ce panier, les prothèses auditives doivent respecter 12 critères de « base », 6 critères parmi les 8 de la liste A et un critère parmi les quatre de la liste B (sauf exceptions).

Les bases de remboursement des deux paniers de soins évoluent de la même façon.

#### **La nouvelle nomenclature**

La création de ces deux paniers de soins s'accompagne d'une nouvelle nomenclature pour les prothèses auditives et pour les piles.

Les prothèses auditives doivent être garanties par le praticien pendant 4 ans et pourront être remboursées au-delà du ticket modérateur que si le délai de renouvellement est supérieur à 4 ans.

Concernant leur base de remboursement, elle se décompose de la manière suivante :

- pour les moins de 20 ans est identique pour les deux paniers. Le prix limite de vente, qui s'applique uniquement au panier 1, est identique à la base de remboursement dès 2019. Le ticket modérateur devant être pris en charge par les contrats responsables, le reste à charge nul pour le panier 1 est ainsi garanti dès 2019. Pour le panier 2, la garantie maximale que pourra proposer la complémentaire, à partir de 2021, sera de 1 700€ (y compris Sécurité sociale).
- pour les plus de 20 ans, la base de remboursement (identique pour les deux paniers) est progressivement revalorisée. Les prix limites de vente du panier 1, quant à eux, diminuent progressivement. Comme pour les moins de 20 ans, à partir de 2021, le remboursement du Régime Obligatoire et du Régime complémentaire ne pourra pas dépasser 1 700€.

Quant à la base de remboursement des piles, elle n'est pas modifiée et la réforme n'impose pas un reste à charge nécessairement nul.

#### D. Les facteurs influençant les dépenses

Les dépenses d'un régime complémentaire santé sont principalement influencées par les facteurs suivants :

##### i. Niveau de garantie

- **'Entrée/bas de gamme'** : contrats basiques avec des remboursements limités au ticket modérateur (TM) (différence entre la base de remboursement (BR) et le remboursement du régime obligatoire (RO)). Les particularités de ce type de contrats sont qu'ils ne couvrent aucun dépassement d'honoraire et de faibles niveaux en optique et dentaire.
- **'Milieu de gamme'** : contrats comprenant des dépassements d'honoraires ainsi que des forfaits en optique, dentaire et audiologie plus élevée que ceux proposés précédemment.
- **'Haut de gamme'** : contrats de très bonne couverture qui remboursent la quasi-totalité des dépassements d'honoraires et des prestations peu ou pas remboursées par la Sécurité sociale (ostéopathes, diététiciens, etc.). L'assuré n'a presque aucun reste à charge.

##### ii. Type de contrat

Nous distinguons deux types de régimes complémentaires : les régimes individuels et les régimes collectifs. Pour ces derniers, l'adhésion peut être facultative ou obligatoire.

Ces deux types de contrats se distinguent notamment par la qualité du souscripteur : il s'agit d'une personne privée pour le contrat individuel alors qu'il s'agit d'une personne morale ou d'une association pour le contrat collectif.

##### iii. Bénéficiaires couverts

Dans un contrat collectif, les bénéficiaires sont l'assuré qui souscrit au contrat et dans certains cas ses ayants droits (conjoint, enfant, ascendant à charge...) qui peuvent aussi, à titre obligatoire ou non, être couverts par la complémentaire. Dans ce sens, il existe plusieurs types de cotisation, les plus courants étant :

- **Isolé / Famille** : si l'assuré principal s'assure seul, il bénéficie d'un tarif 'Isolé'. Dans le cas contraire, il bénéficie d'un tarif 'Famille'.
- **Adulte / Enfant** : L'assuré et son conjoint ont un tarif 'Adulte' et les enfants ont un tarif 'Enfant'.

- **Isolé / Duo / Famille** : La cotisation 'Isolé' correspond au salarié seul, la cotisation 'Duo' au salarié avec un enfant ou un conjoint et enfin la cotisation 'Famille' au salarié avec au minimum deux ayants droit.
- **Unique** : le tarif ne dépend pas de la composition familiale du salarié.



## Partie II - Analyse et description du portefeuille

L'objectif de cette partie est de présenter le portefeuille en détail et de montrer les traitements qui ont été appliqués sur les bases brutes afin d'avoir une table nettoyée et finale de la forme ci-dessous et qui sera utilisée dans la partie III de modélisation (GLM).

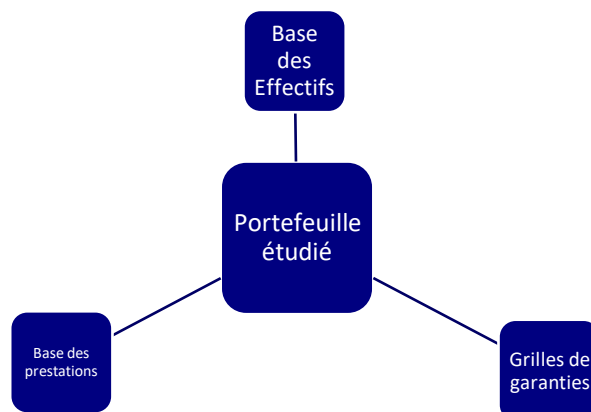
Acte	Remboursement	Age	Exposition	Nb_Actes	CM
Actes d'échographie - OPTAM	180%	23	1,000	0	0,00 €
Contraception	34 €	29	0,121	0	0,00 €
Consultations de généralistes - NON OPTAM	200%	6	0,997	2	23,40 €
Inlay onlay	160%	45	0,997	1	78,65 €
Lentilles acceptées	160%	30	0,997	0	0,00 €
Honoraires médicaux - OPTAM	180%	44	0,997	0	0,00 €
Cure thermique	100%+50€	12	0,997	3	121,54 €
Soins dentaires	150%	27	0,997	5	40,78 €
Actes d'échographie - NON OPTAM	150%	19	0,997	0	0,00 €

▲ Figure 12 : Forme de la base finale construite à la fin de partie II

### 1 Informations sur le portefeuille

Notre étude s'est portée sur les données d'un portefeuille de **plus de 140 000 personnes**, qui intègre l'ensemble des sinistres survenus entre le 1<sup>er</sup> janvier 2018 et le 31 décembre 2020. Nous avons à notre disposition :

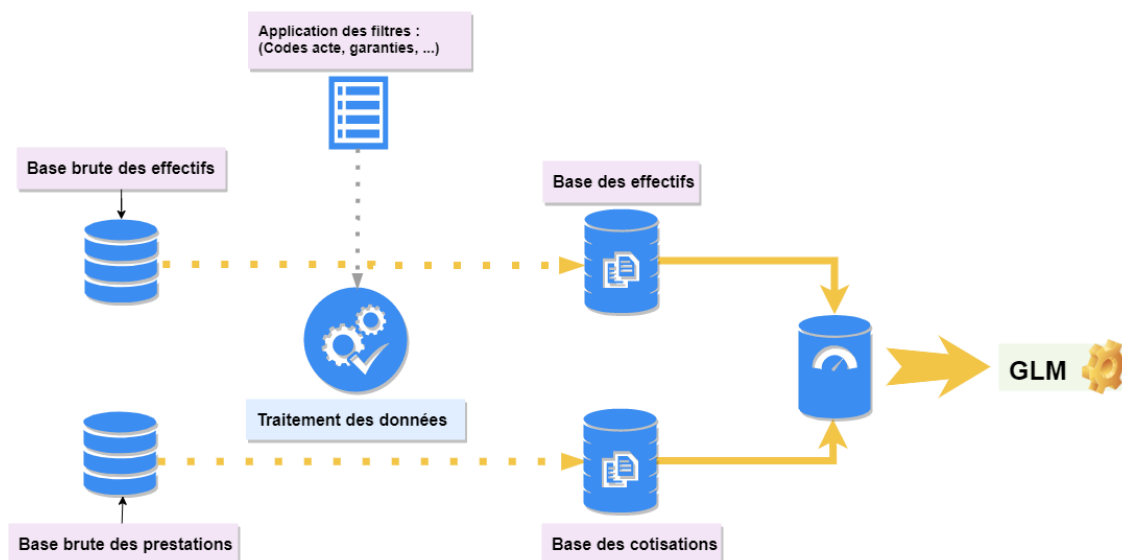
- Une base d'effectifs regroupant **133 690 adhérents** (dont **126 301 bénéficiaires**) répartis sur 3 exercices : **2018, 2019 et 2020**.
- Une base de prestations contient les lignes des sinistres survenus et payés entre le 1<sup>er</sup> janvier 2018 et le 31 décembre 2020, soit un volume total de **83 623 058 €**.
- Une base de grilles de garanties proposées à ces adhérents, contenant **13 tables de garanties différentes**.



▲ Figure 13 : Éléments du portefeuille étudié

En amont de la réalisation de notre outil de tarification, il a été nécessaire de retraiter les bases de données simultanément avec un traitement des garanties et des codes actes. Nos deux bases (de prestations et de cotisations) ont ensuite pu être fusionnées afin d'être étudiées plus aisément. Cette étape, cruciale, a été laborieuse mais nécessaire pour répondre à l'objectif de générer un modèle de tarification fiable et adapté à notre population assurée.

La fusion des bases des effectifs et des prestations est suivie par une analyse statistique univariée et multivariée afin de bien montrer les caractéristiques de notre portefeuille, ainsi que pour détecter les valeurs aberrantes qui seront retraitées ultérieurement. Le schéma ci-dessous illustre la démarche suivie :



▲ Figure 14 : Schéma de traitement des bases brutes avant la modélisation par GLM

La base des données disponible comporte l'information des soins pour 1130 codes actes (avec l'information CAS / NON-CAS). L'objectif de la sous-section suivante est de regrouper ces codes actes en une liste courte.

## 2 Etude des garanties et des codes actes

Le résultat de l'étude dans cette sous-section est de générer une table de correspondance de la forme suivante :

	Grille 1	.	.	Grille j	.	.	Grille 13
<b>Libellé brochure 1</b>	100%	150%	130%	100%	140%	155%	200%
.							
.							
<b>Libellé brochure i</b>	20 €	50 €	45 €	20 €	45 €	60 €	70 €
.							
.							
<b>Libellé brochure 73</b>	1% PMSS	1,4% PMSS	1,2% PMSS	1% PMSS	1,3% PMSS	1,5% PMSS	2% PMSS

▲ Figure 15 : Table de correspondance entre les grilles de garanties et les libellés actes

La base de données ne contient pas directement une information sur le niveau de remboursement du code acte.

Sur la base des prestations et sur la base des effectifs, nous avons une colonne qui s'appelle nom d'option, cette colonne nous permet de récupérer le nom de la grille des garanties à partir de laquelle nous pouvons récupérer le niveau de remboursement. Puisque nous avons 13 tables des garanties, cette colonne par conséquent contient 13 valeurs possibles noté dans la suite : Grille\_1, Grille\_2, ..., Grille\_13.

Sur les deux bases, nous avons également une colonne qui contient les codes actes. Nous avons au total 1130 codes actes différents. Un travail intermédiaire a été fait sur cette colonne afin regrouper ces codes actes. Ces codes actes ont été regroupé sous forme de 73 libellés brochures sous 7 familles d'actes différents :

La famille des actes « **Hospitalisation** » contient les actes d'anesthésie, de chirurgie, d'échographie, d'imagerie, d'obstétrique (avec la séparation OPTAM/NON OPTAM), forfait hospitalier, frais de séjour, chambre particulière, frais d'accompagnement et frais de transport.

La famille des actes « **Soins Courants** » contient les actes techniques médicaux, actes de radiologie, analyses médicaux, auxiliaires médicaux, consultations et visites

de généralistes/ spécialistes et de psychiatrie avec la séparation OPTAM / NON OPTAM.

La famille « **Pharmacie** » contient les trois types : Pharmacie 15%, 30% et 65%, vaccin anti-grippe et le forfait des médicaments non remboursés.

La famille « **Appareillage** » contient les prothèses auditives, l'appareillage et les accessoires.

La famille « **Optique** » contient les composantes d'un équipement : Monture, verre Simple, complexe et très complexe. Il contient aussi les lentilles acceptées/refusées et chirurgie des yeux.

La famille « **Dentaire** » contient les soins dentaires, Inlay-Onlay, Inlay core. Elle contient également les prothèses dentaires et l'orthodontie acceptées/ refusées par le régime obligatoire, les implants dentaires et la parodontologie.

La famille « **Autres** » contient le reste des actes, notamment les actes de médecine douce, cure thermale, soins à l'étranger, consultations sport, arrêt de tabac, contraception, fécondation in vitro, ostéodensitométrie, prestations appairage, prime de naissance et les vaccins non remboursés par la sécurité sociale (voir récapitulatif dans Annexe 1).



▲ Figure 16 : Regroupement : Familles d'actes

Après le regroupement des codes actes sous formes de 73 libellés brochure, nous pouvons désormais construire la table de correspondance illustrée par la figure 15.

Pour la construction d'un outil de tarification, Ce regroupement nous permet de :

- Limiter le nombre de modèles GLM que nous allons chercher
- Adapter l'outil de tarification avec les grilles de garanties de nos clients

Nous allons utiliser des grilles de garanties pour remplacer la colonne des codes actes par le niveau de remboursement (% ou €).

### 3 Etude des FR et BR moyens

L'outil de tarification à construire sur Excel sera alimenté à travers les grilles de garanties envoyées par le client. Le type de remboursement est différent d'un acte à l'autre. Par exemple, le forfait hospitalier est remboursé en fonction des frais réels, les consultations de généralistes sont remboursées en fonction de la **BR** (base de remboursement de la sécurité sociale) ou **BRySS** (base de remboursement y compris le remboursement de la sécurité sociale) ou **TM** (Ticket modérateur), les prothèses dentaires sont remboursées en fonction d'un taux d'une base (**BR**, **BRySS** ou **TM**) + un montant en € ou un % du **PMSS**. D'autre part, les détails des actes consommés du même libellé ont des BR différentes.

Pour cela, à travers les données des prestations, nous avons calculé une base de remboursement moyenne et une dépense réelle moyenne par libellé brochure.

## 4 Caractéristiques du portefeuille étudié

### A. Données à disposition

L'objectif de cette partie est de connaître et analyser le portefeuille utilisé pour la modélisation dans ce mémoire.

Pour cette étude, nous avons utilisé les données d'une mutuelle (fusion de plusieurs mutuelles) qui couvre la moitié sud de la France. Cette mutuelle propose des offres de couverture en santé pour les particuliers (individuels) et également pour les entreprises (contrats collectifs à adhésion obligatoire et des contrats surcomplémentaires à adhésion facultative).

Dans ce mémoire, nous n'avons modélisé que le portefeuille des particuliers (contrat individuels). Les produits sont regroupés en 4 gammes :

Une gamme qui cible les fonctionnaires territoriaux : 2 produits,

Une gamme qui cible les entrepreneurs : 3 produits,

Une gamme qui cible les fonctionnaires hospitaliers : 2 produits,

Une gamme individuelle standard qui regroupe 6 produits différents.

Tous les produits étudiés sont donc des produits à adhésion individuelle et par conséquent facultative.

Par souci de confidentialité, les noms des produits sont masqués. Dans la suite, les produits seront appelés **FT** (fonctionnaires territoriaux), **ENT** (entrepreneurs), **FH** (Fonctionnaires hospitaliers) et **GS** (gamme standard).

Les données sources se composent d'une base des effectifs et d'une base des prestations survenues et réglées du 01/01/2018 et 31/12/2020.

La base des effectifs contient les informations sur les assurés, notamment l'âge, la qualité de l'ayant droit et les dates de début et de fin du contrat ainsi que la date d'adhésion à la mutuelle. La base des prestations contient des informations sur les remboursements, notamment le remboursement mutuelle, remboursement de la sécurité sociale et le montant de la dépense.

La mutuelle étudiée couvre plus de 130 000 adhérents particuliers, soit une exposition totale brute (avant traitement) de 316 762 sur les trois ans.

La base des effectifs se présente sous forme de 3 fichiers csv, un fichier par année civile de 2018 à 2020. Cette base contient les informations suivantes :

- **Ref\_personne** : code unique de l'adhérent
- **CodeFamille** : code unique de la famille
- **QualitéAy** : qualité de l'ayant droit (conjoint, enfant ...)
- **Type Contrat** : type du contrat (toujours individuel pour cette étude)
- **Naissance Ayant Droit** : date de naissance de l'ayant droit
- **Date Début Option** : date de début du contrat
- **Date Fin Option** : date de fin du contrat
- **Département** : numéro du département
- **Nom Option** : nom du produit / contrat
- **Année** : année d'abonnement
- **Année Cotisation HT** : montant de cotisation HT annuelle
- **Adhésion à la mutuelle** : date d'adhésion à la mutuelle

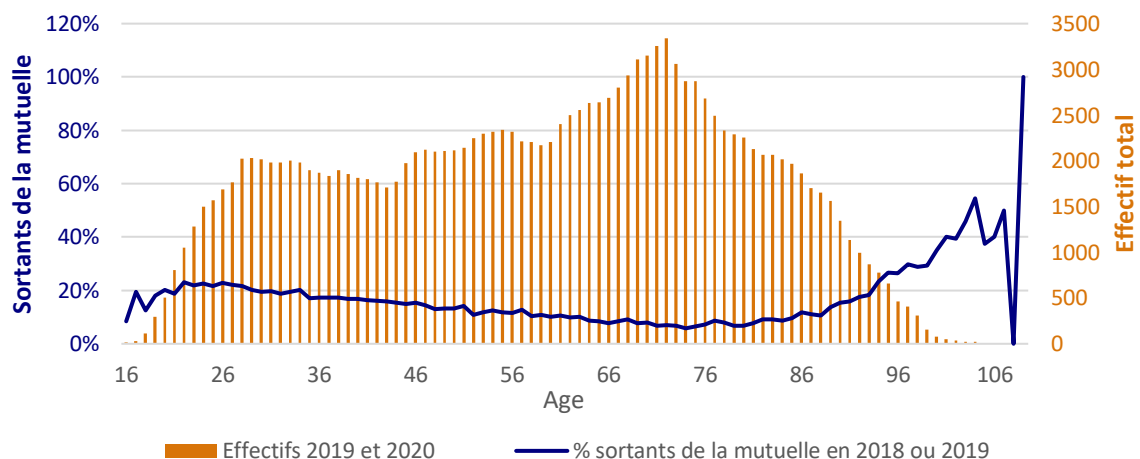
Cette dernière information est très intéressante puisqu'elle nous permet notamment de regarder l'ancienneté du portefeuille.

Malheureusement, nous ne disposons pas d'information sur le sexe et ne pouvons donc pas procéder à des analyses différenciées en fonction du genre.

## B. Etude du Turn-over

Le turn-over est un indicateur du renouvellement de l'effectif de notre portefeuille d'une année sur l'autre.

L'effectif dans les graphiques suivants indique le nombre d'adhérents présents au moins une journée dans le portefeuille sur la période d'étude (01 janvier - 31 décembre) des années 2018-2019 / 2019-2020. Le nombre de sorties est le nombre d'adhérents sortis de la mutuelle pendant la période d'étude.



▲ Figure 17 : Répartition des assurés et sortants de la mutuelle par âge 2018-2019

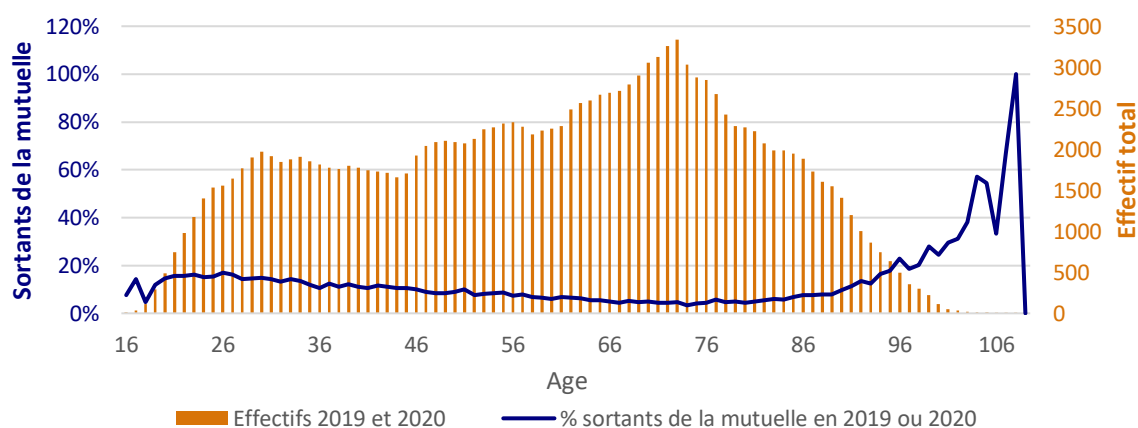
Dans le graphique ci-dessus, nous retrouvons la répartition des assurés par âge présentée dans la partie précédente (toutes gammes confondues) et avons ajouté le nombre de sorties en bleu en pourcentage des personnes présentes.

Nous constatons que les taux de turn-over sont les plus importants aux extrémités :

- En début d'activité professionnelle, cela s'explique par deux effets : d'une part des résiliations pour adhésions à un contrat collectif obligatoire et d'autre part une génération plus habituée à optimiser sa couverture et qui n'hésite pas à changer de mutuelle.
- En fin de vie provenant naturellement de décès.

Parmi plus de 77 700 adhérents en 2019, au cours de l'année 2019, 14% sont des sortis de la mutuelle et 11% qui sont entrés, soit au global une légère perte d'effectifs (3%).

Le graphique ci-dessous représente les assurés et les sortants par âge sur les deux années 2019 et 2020.



▲ Figure 18 : Répartition des assurés et sortants de la mutuelle par âge 2019-2020

Les constats sont les mêmes que pour le graphique précédent.

Le nombre d'adhérents en 2020 est estimé à 76 590. Nous constatons environ 12% de sorties entre 2019 et 2020 contre 11% d'entrées. Soit une perte de moins de 1% d'effectifs.

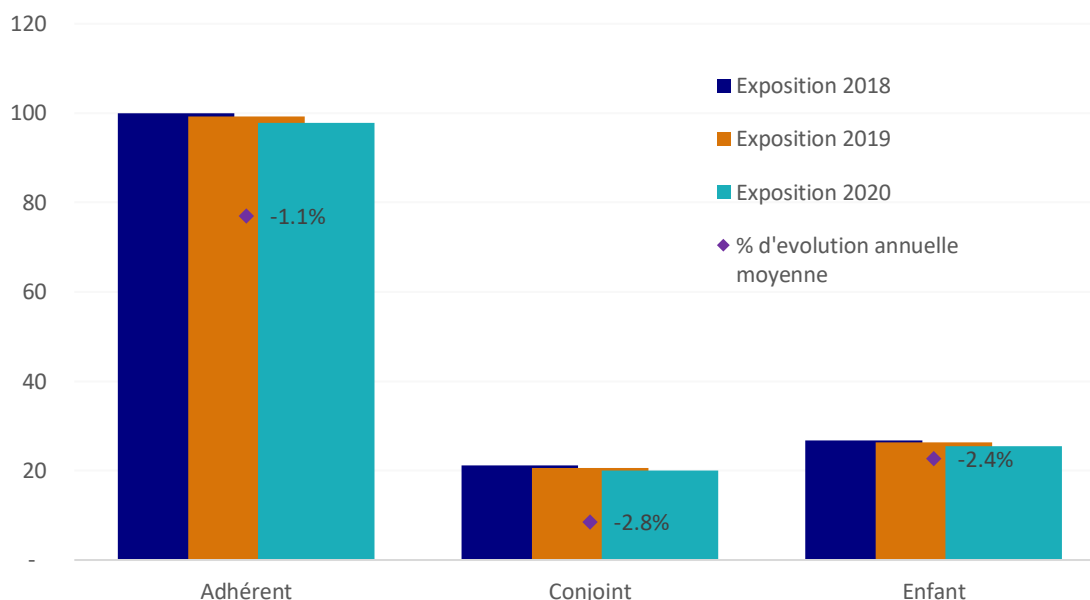
Nous constatons un renouvellement de la population (plus de 10% d'entrées et de sorties avec un écart léger entre les deux) ce qui pourrait faire craindre une modification de la population assurée. La partie suivante permet d'analyser la répartition de la population en fonction de différentes variables qui seront utilisées dans la modélisation. Le but est de vérifier la stabilité du portefeuille sur les 3 ans.



## C. Etude du portefeuille

### i. Répartition par année et par type ayant droit

Le graphique ci-dessous représente la somme d'exposition sur une base 100 par type d'ayant droit et par année. Il nous permet d'avoir une visibilité sur la répartition des ayants droit sur les 3 années d'étude. Le taux à droite correspond à l'évolution annuel moyen des expositions entre 2018 et 2020.



▲ Figure 19 : Répartition par année et par type ayant droit

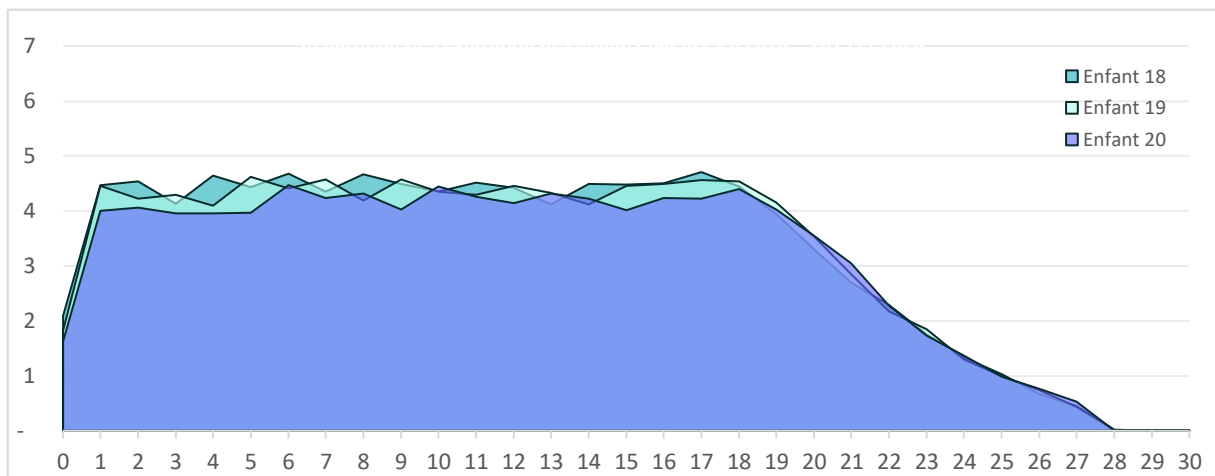
En termes d'exposition, la mutuelle couvre principalement les responsables, qui représentent 68% du portefeuille sur les 3 ans avec une baisse moyenne légère de 1% pendant cette période. Les conjoints et les enfants représentent 21% et 25% respectivement du portefeuille étudié avec une baisse moyenne d'un peu moins de 3%.

Le portefeuille reste plutôt stable sur la période : nous n'observons pas de grande évolution pour les adhérents principaux, ceux qui représentent la majorité du portefeuille étudié.

En revanche, nous observons des évolutions plus importantes pour les conjoints et les enfants. Nous vérifierons de manière plus fine dans la partie suivante la stabilité du portefeuille.

## ii. Répartition de l'exposition par âge et par type d'assuré

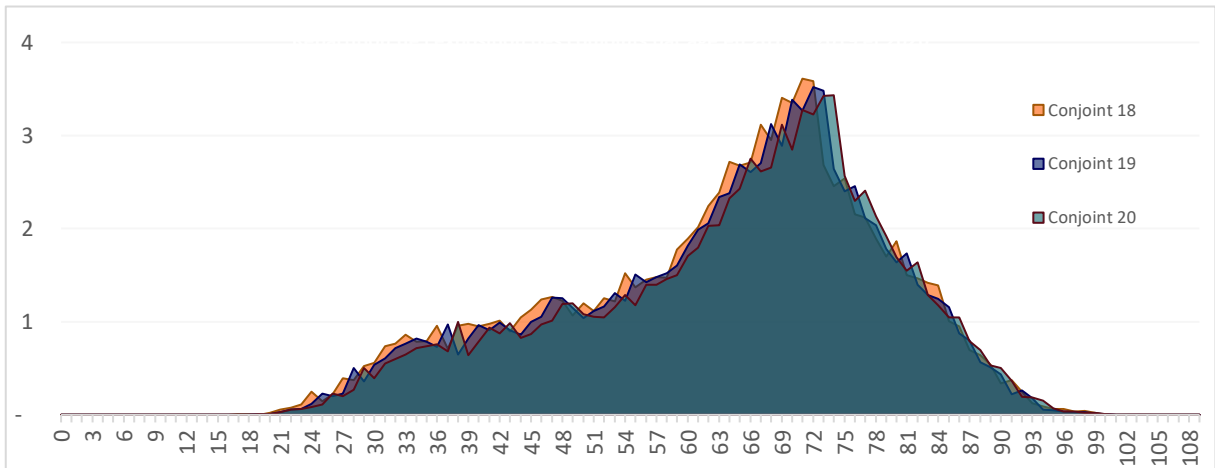
Pour les enfants, l'exposition totale sur une base de 100 est distribuée uniformément sur les âges de 0 à 20 ans comme nous montre le graphique suivant :



▲ Figure 20 : Répartition de l'exposition des enfants par âge en 2018 -- 2019 et 2020

L'exposition est répartie sur les âges de la même manière sur les 3 années. Nous validons l'hypothèse de stabilité de la répartition de l'exposition d'une année sur l'autre pour les enfants.

Pour les conjoints, la répartition des âges est similaire d'une année sur l'autre comme nous montre le graphique suivant (L'échelle à gauche représente l'effectif par âge sur une base de 100) :

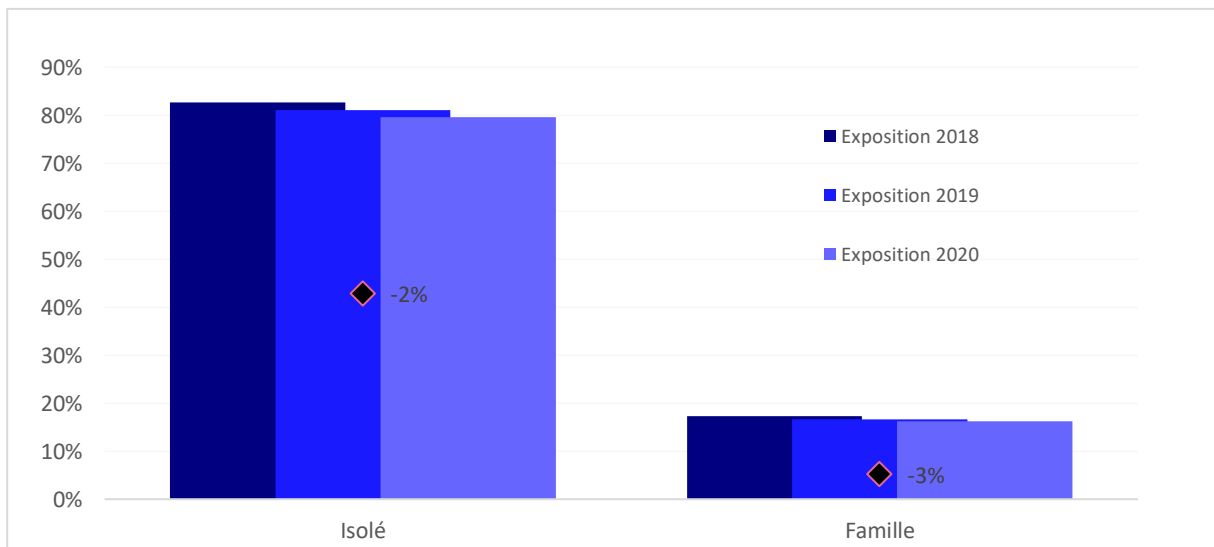


▲ Figure 21 : Répartition de l'exposition des conjoints par âge en 2018 -- 2019 et 2020

➔ Le nombre de conjoints d'un âge donné en 2018 est très proche du nombre pour ce même âge en 2019 et 2020. Le portefeuille reste stable en fonction du type d'ayant droit sur la période étudiée.

### iii. Répartition des enfants par famille

Le graphique suivant montre la répartition des responsables selon le critère isolé / famille à travers les variables : Type d'ayant droit, code famille et l'information sur l'année. Le taux à droite représente le taux d'évolution annuel moyen des expositions entre 2018 et 2020.

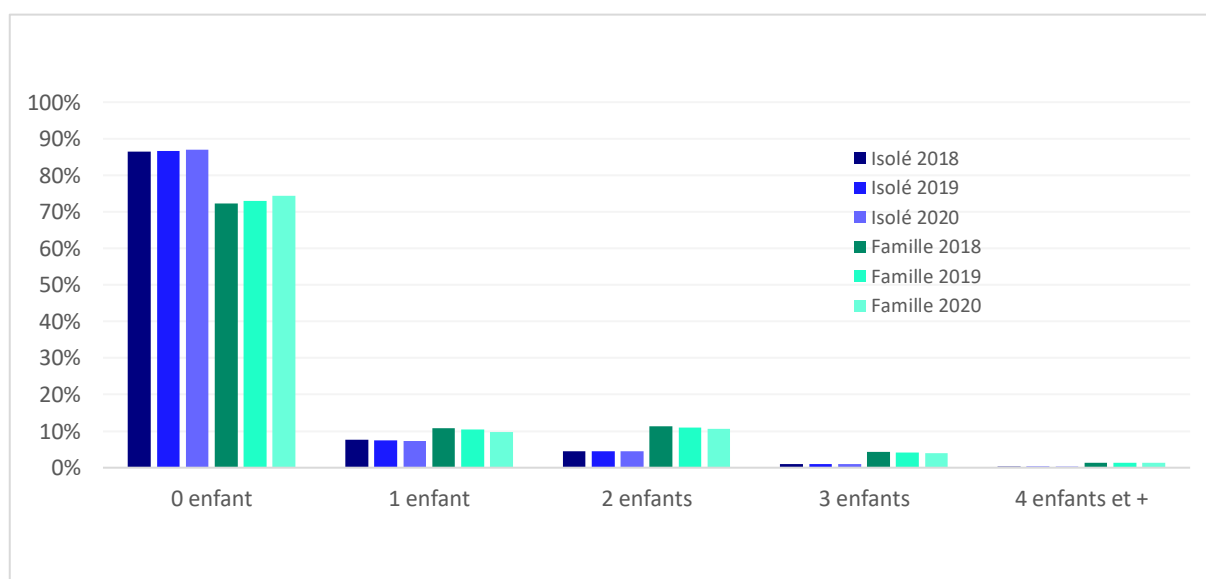


▲ Figure 22 : Répartition des assurés par : Famille / Seul

En termes d'exposition, les assurés 'isolés' (adhérent sans conjoint, avec ou sans enfants) représentent 83% du portefeuille étudié contre 17% des assurés en famille (adhérent avec conjoint, avec ou sans enfants).

Nous constatons une baisse annuelle moyenne de 2% pour les assurés isolés et 3% pour les assurés en famille. Nous considérons que le portefeuille est relativement stable par rapport au critère 'Seul / Famille' sans forte évolution.

La figure suivante présente la répartition du nombre d'enfants par année et par le critère solo / famille :



▲ Figure 23 : Répartition des enfants par type de famille

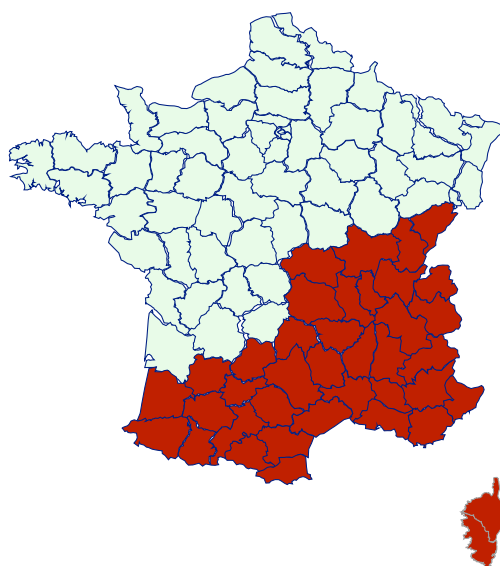
La répartition des enfants reste relativement stable aussi bien en fonction de la structure familiale qu'en fonction de l'année d'observation.

Au final, moins de 30% des responsables ont un enfant comme bénéficiaire.

#### iv. Répartition par département

La base des effectifs contient la variable « département » qui donne l'information sur le lieu de résidence de l'assuré. Le nombre de lignes dont le département est non renseigné ne représente que 0.03% de l'ensemble de la base.

En analysant la distribution des expositions sur toute la France, nous constatons que la mutuelle étudiée cible le Grand Sud français. Cette zone est formée les régions : *Auvergne-Rhône-Alpes*, *Occitanie*, *Provence-Alpes-Côte-d'Azur* et une partie de *Bourgogne-Franche-Comté* (zones en rouge dans la figure ci-dessous) :



▲ Figure 24 : Régions assurées par la mutuelle (rouge)

La répartition par départements ne présente pas de dégradé de couleurs pour des raisons de confidentialité.

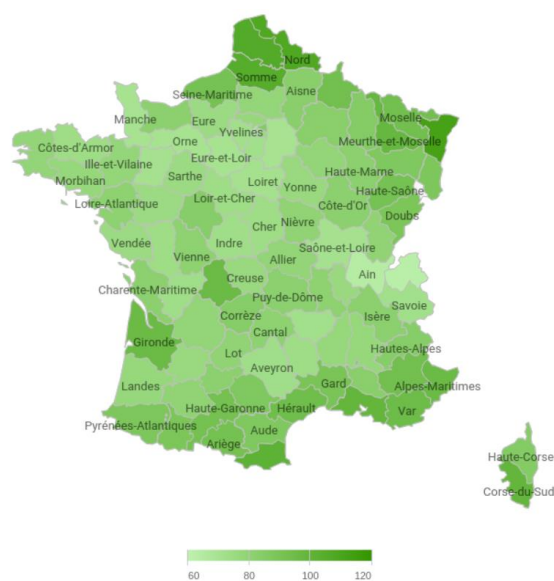
Les données à notre disposition ne couvrent pas toute la France et ne permettent pas de modéliser directement le coût de sinistre sur l'ensemble du territoire français.

En effet, les dépenses de santé sont impactées par la zone géographique de l'assuré. Ainsi, un assuré vivant à Nice aura globalement des dépenses de santé plus élevées par rapport à une autre personne vivant dans une zone rurale (toute chose étant égale par ailleurs).

La carte de l'INSEE suivante en présente les remboursements par personne en médecine générale en 2014 (en €)<sup>3</sup> pour un âge moyen de 42 ans pour les femmes et 40 ans pour les hommes.

---

<sup>3</sup> Source : <https://www.data.gouv.fr/en/reuses/quelle-est-linfluence-des-modes-de-vie-sur-les-depenses-de-sante/>



▲ Figure 25 : Remboursement par personne en médecine générale (€) - 2014 INSEE

Nous constatons des niveaux de remboursements différents en fonction du département.

Le remboursement moyen est élevé (vert foncé) en Lorraine, Alsace, Provence-Alpes-Côte d'Azur, Languedoc-Roussillon, autour de la zone Poitou-Charentes et en Corse-du-Sud.

La corrélation entre la zone géographique et la consommation en santé s'explique par le fait que les dépassements d'honoraires sont assez fréquents dans les grandes villes impliquant un coût total moyen par sinistre plus fort.

La fréquence de soins varie également entre les villes et les petits villages par le fait que l'accès au soin est plus difficile dans les villages, ceci implique également une fréquence plus faible dans ce cas. La distribution des médecins et des hôpitaux sur le pays explique une fréquence plus faible dans les petites villes que dans les grandes.

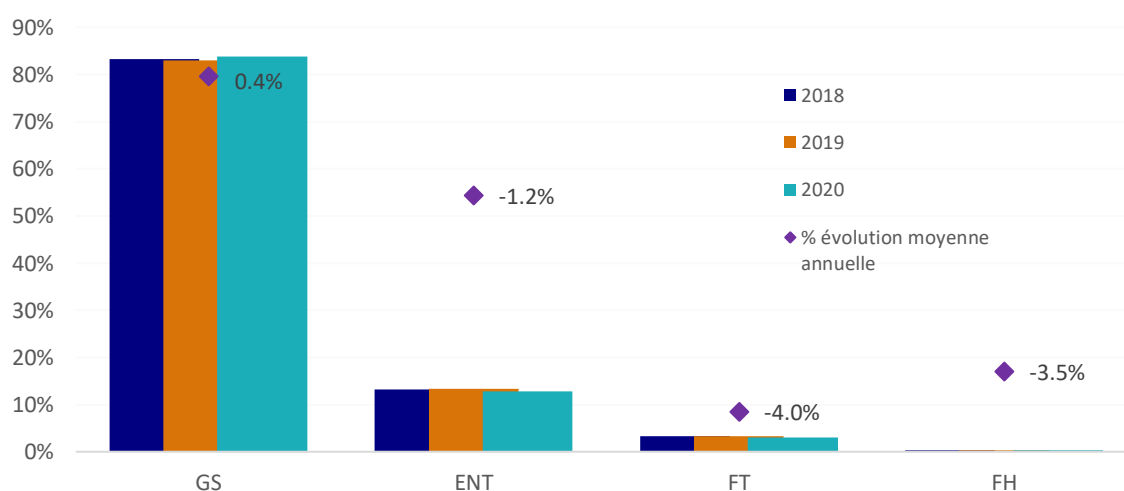
L'application directe des tarifs trouvés par la modélisation sur tous les assurés en France impliquera sûrement des résultats biaisés, d'où la nécessité d'appliquer les résultats d'un zonier sur les tarifs purs issus de la modélisation afin de prendre en compte la corrélation entre la zone géographique et la consommation, et afin de tarifier le reste des départements non couverts par la mutuelle.

En pratique cela n'est pas gênant car Actélior dispose de son propre zonier : les résultats de la présente étude pourront donc être étendu à la France entière par l'application de ce zonier.

## D. Zoom sur les gammes étudiées

### v. Exposition en fonction des gammes étudiées

Le graphique ci-dessous illustre la répartition des effectifs du portefeuille sur les gammes étudiées. Le taux à droite représente le taux d'évolution annuel moyen des expositions entre 2018 et 2020.



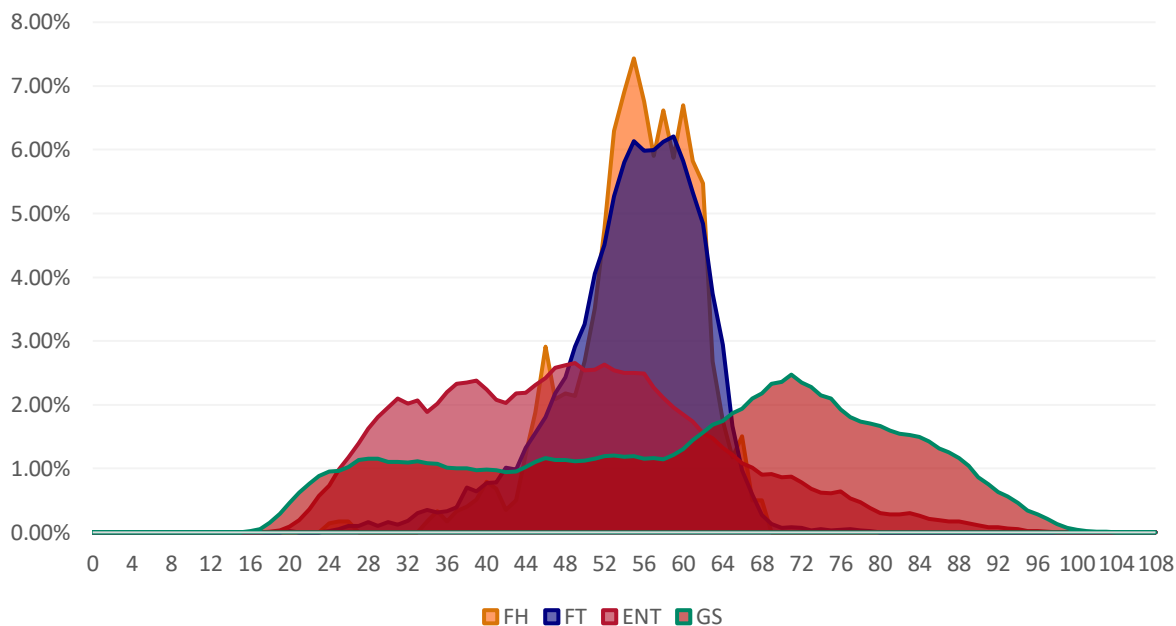
▲ Figure 26 : Répartition par année et par gamme

Le portefeuille est composé principalement des adhérents des contrats de la gamme standard. Les produits de cette gamme seule représentent 83% du portefeuille en termes d'exposition. Pour cette gamme, nous constatons une augmentation annuelle moyenne légère estimée à 0.4%. La gamme des entrepreneurs, représente 13% de l'exposition totale du portefeuille. L'effectif de cette gamme baisse de 1.2% en moyenne par an.

→ Le portefeuille est majoritairement représenté par la gamme individuelle standard (GS). La répartition du portefeuille par gammes étudiées est globalement stable dans le temps.

## vi. Démographie par gamme étudiée

Le graphique ci-dessous montre la répartition de l'exposition des adhérents sur les 3 exercices étudiés, par âge, pour chacune des gammes étudiées.



▲ Figure 27 : Démographie par gamme étudiée

Pour la gamme GS (gamme standard), la plus souscrite, nous constatons une distribution uniforme des âges entre 22 ans et 60 ans, puis un pic à 72 ans. La densité pour la gamme GS montre que nous disposons de suffisamment de données pour tous les âges entre 18 et 94 ans.

Pour les deux gammes FH et FT, tous les âges sont concentrés dans la tranche : [45 , 66]. La somme de ces deux gammes ne représente que 17% de l'exposition totale du portefeuille.

La gamme ENT qui cible les entrepreneurs est la gamme étudiée dont la population est la plus jeune. Les âges sont concentrés entre 24 ans et 70 ans environ. Ceci s'explique par le fait que les entrepreneurs sont, par définition, des personnes actives.

➔ Avec 83% du portefeuille, la gamme GS assure une large distribution de la population assurée selon l'âge.

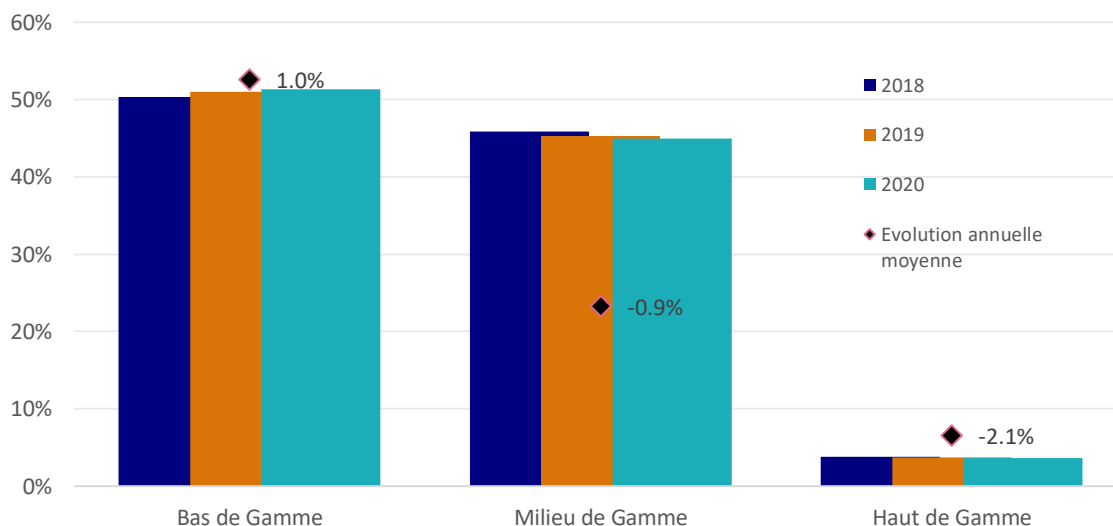


## vii. Répartition de l'exposition par niveau de gamme et année

La mutuelle propose systématiquement plusieurs niveaux de garanties responsables. On en distingue 3 principaux :

- **Bas de gamme** : concerne les formules de type « ticket modérateur » ou avec un léger dépassement d'honoraires (remboursement complémentaire inférieur à 150%).
- **Milieu de gamme** : concerne les formules intermédiaires, dont les remboursements sont compris entre 150% et 250% de la base de remboursement y compris sécurité sociale. Ces garanties prennent en charge des dépassements d'honoraires mais il peut rester du frais à la charge de l'assuré.
- **Haut de gamme** : concerne les formules qui offrent des remboursements entre 250% et 500% de la base de remboursement y compris le remboursement de la sécurité sociale. Ces produits présentent peu de reste à charge.

La figure suivante représente la répartition des responsables par niveau de gamme et par année de soins :



▲ Figure 28 : Exposition en fonction du niveau de gamme / année

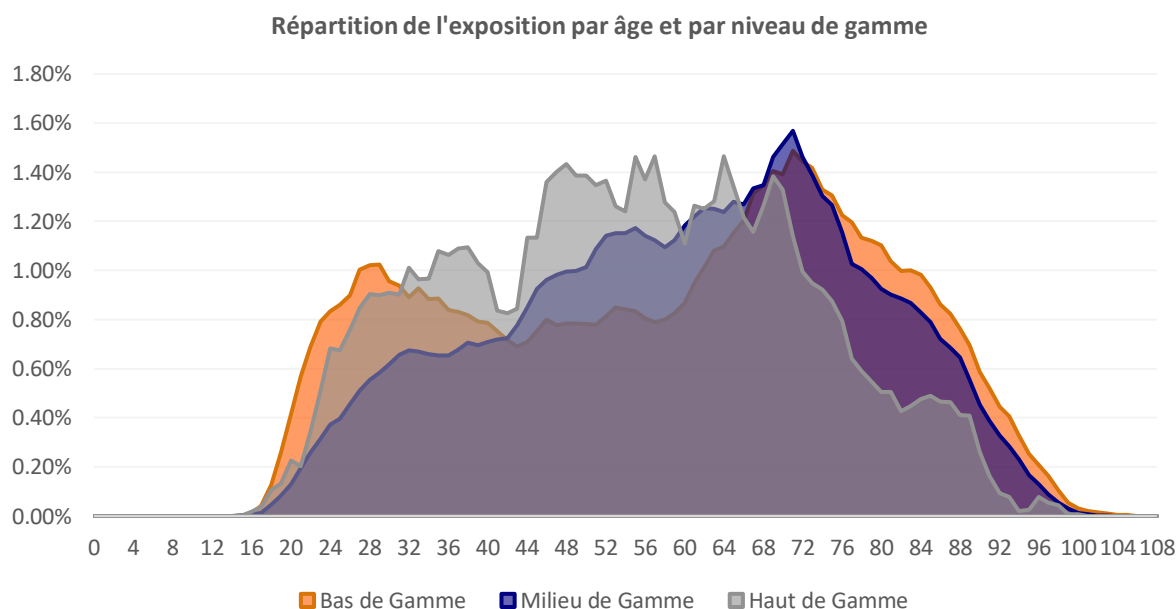
Les produits 'bas de gamme' représente la moitié du portefeuille étudié sur les 3 exercices. Ce niveau augmente en moyenne de 1% chaque année.

Les deux niveaux milieu et haut de gamme baissent en moyenne de 1% et 2.1% respectivement, chaque année.

➔ Les variations d'une année sur l'autre sont minimales. Le portefeuille est donc stable au regard des niveaux de gamme.

### viii. Démographie par niveau de gamme

Le graphique suivant montre la répartition de l'exposition par âge et par niveau de gamme.



▲ Figure 29 : Répartition de l'exposition par âge et par niveau de gamme

La répartition des effectifs entre les niveaux de gamme est assez différente. En effet, pour les bas de gamme, représentant la moitié de notre portefeuille, nous constatons deux pics : entre 20 et 35 ans puis entre 65 et 80 ans. Après 80 ans, le portefeuille décroît ce qui est cohérent avec la répartition de la population française.

Cela pourrait s'expliquer par une hypothèse que les personnes les plus intéressées par des « petites couvertures » sont des personnes avec des revenus plus faibles : les jeunes actifs et les seniors.

Pour les deux autres regroupements, les effectifs sont plus importants entre 40 et 75 ans. Ce sont des âges où les besoins en optique et en dentaire sont importants et donc la souscription de garanties de meilleure qualité est nécessaire.

➔ Au global, nous constatons une large distribution par âge de 20 à 90 ans.

## E. Conclusion

Le portefeuille étudié concerne des personnes assurées sur 3 ans (2018, 2019 et 2020). Ces assurés correspondent à différents types d'ayant droit (adhérent principal, conjoint et enfant), ont choisi une assurance complémentaire parmi plusieurs produits d'assurance regroupés en 4 gammes. Ces produits sont de différents niveaux de gamme (bas, milieu et haut de gamme).

L'étude du turn-over nous a permis d'analyser les sortants et les entrants en passant d'une année à l'autre. Nous constatons que les taux de sorties et les taux d'entrées sont cohérents sur les trois exercices. En revanche, le turn-over est indicateur sur le renouvellement de la population et non pas forcément de la stabilité du portefeuille dans ses caractéristiques permettant la réalisation d'une tarification à travers des modèles GLM.

La proportion des adhérents couvrant leur conjoint, ainsi que le nombre d'enfants couverts par adhérent restent stables sur la période. De plus, la distribution par âge, tant pour les enfants que pour les conjoints, reste également stable. On en conclut que la population des bénéficiaires reste globalement stable.

Les deux variables *ref\_personne* et *CodeFamille* nous ont permis d'avoir la répartition des expositions selon le critère « isolé / famille ». Nous constatons une légère baisse de l'exposition annuelle (entre 2% et 3%), mais la répartition selon le nombre d'enfants pour les familles et chez les personnes isolées reste stable sur les 3 années étudiées.

Les produits de la mutuelle sont regroupés en 4 gammes (**FT** : fonctionnaires territoriaux, **ENT** : entrepreneurs, **FH** : Fonctionnaires hospitaliers et **GS** : gamme standard). La gamme **GS** qui représente 83% du portefeuille en termes d'exposition connaît une très légère baisse annuelle moyenne (0.4%). L'exposition de cette gamme est répartie sur tous les âges d'une manière équilibrée. Nous concluons que nous avons suffisamment d'assurés à chaque âge pour les gammes étudiées.

Les produits du portefeuille sont classés également selon leur niveau de gamme (base, milieu ou haut de gamme). La somme l'exposition des deux niveaux 'bas' et 'milieu' représente plus de 95% de l'exposition totale du portefeuille. L'étude montre une évolution des effectifs de moins de 1% sur chacun de ces niveaux, ce qui confirme la stabilité du portefeuille d'une année sur l'autre en fonction du niveau de gamme. Notons bien que pour les 3 niveaux, nous constatons une large distribution de l'exposition par âge de

20 à 90 ans.

Comme conclusion, les caractéristiques du portefeuille utilisé pour la modélisation du coût des sinistres en santé sont globalement stables sur la période 2018-2020. Cette stabilité légitime l'utilisation des données pour la construction de l'outil de tarification. Compte tenu des effectifs réduits pour les produits de type haut de gamme, l'outil de tarification sera précis surtout pour des garanties de type « bas de gamme » et milieu de gamme.

Nous avons identifié dans cette étude une réserve concernant la répartition des assurés par département. Ce point sera corrigé dans l'outil final par l'intermédiaire du zonier disponible au sein d'Actélior.

## 5 Etude de la base des effectifs

La base brute des effectifs comprend les bénéficiaires présents dans le portefeuille d'assurés au moins une journée entre le 1<sup>er</sup> janvier 2018 et le 31 décembre 2020. Cette base contient **329 456** lignes avec **10** variables en colonnes réparties sur trois fichiers csv (un fichier pour chaque année). La figure ci-dessous représente un échantillon des données brutes des cotisations.

Numéro_adh	Qualité	Date_naissance	Date_début	Date_fin	Code_tarif	Libellé_tarif	Département	Code_option	Nom_option
4430	Assuré	09/04/1945	01/09/2016	31/12/2018	6G45V	A133	10	A133	Option_1
3493	Enfant	01/09/2010	01/06/2019	31/12/2018	341DE	A133	13	A133	Option_3
4953	Enfant	13/02/2009	01/01/2020	31/12/2020	34DER	B654	50	B654	Option_12
2394	Assuré	16/10/1988	01/01/2018	31/12/2018	22RR5	H899	52	H899	Option_12
1222	Assuré	11/06/1990	01/01/2017	31/12/2017	89KKI	R566	43	R566	Option_22

▲ Figure 30 : Échantillons de la base des effectifs

Afin d'être utilisée de manière fiable, nous avons dû procéder à la réalisation de plusieurs retraitements que nous énumérons ci-dessous :

- Correction des valeurs manquantes :

Lorsque la date de fin de couverture (colonne **Date\_fin**) n'est pas précisée, nous considérons que le bénéficiaire est en portefeuille jusqu'au 31/12 de l'année étudiée.

- Correction des bornes des colonnes Date\_début et Date\_fin :

Nous recherchons l'exposition sur chaque exercice étudié. Ainsi, si la **Date\_Début** est antérieure au 01/01 de l'exercice étudié alors nous retenons le 1<sup>er</sup> janvier. De même pour la **Date\_fin**, si elle est postérieure au 31/12 de l'exercice étudié alors nous retenons le 31 décembre. Nous obtenons donc les formules suivantes :

$$\begin{cases} Date\_début = \max (Date\_début, 01/01/N) \\ Date\_fin = \min (Date\_fin, 31/12/N) \end{cases}$$

avec N : l'exercice étudié.

- Calcul de l'exposition :

L'**Exposition** correspond à la période de couverture ou période d'exposition au risque pendant une année donnée. Par exemple, un bénéficiaire présent 6 mois en 2018 aura une exposition égale à 0,5 pour l'année 2018. Elle est calculée par la formule suivante :

$$Exposition = \frac{Date_{fin_N} - Date_{début_N} + 1}{Nbjours_N}$$

Ceci est équivalent à calculer directement l'exposition, sans recalculer les dates, par la formule suivante :

$$Exposition = \frac{\min(Date_{fin_N} ; 31-12-N) - \max(Date_{début_N} ; 01-01-N) + 1}{Nbjours_N}$$

- Calcul de la variable âge :

Nous ne pouvons pas utiliser immédiatement la variable « **Date de naissance** » comme variable explicative, pour cela, nous avons calculé la variable « **Age** » qui est plus parlante et qui sera intégrée ultérieurement dans nos modèles GLM de tarification. Cette variable est calculée par différence de millésime :

$$Age = N - année(Date\_naissance)$$

Notons bien que l'âge est calculé par rapport à l'exercice étudié et non pas par rapport à la date d'aujourd'hui.

Nous avons fusionné les trois bases résultantes après ces traitements. Nous avons remplacé les colonnes **Date\_début** et **Date\_fin** par la colonne **Exposition** et nous avons remplacé la colonne **Date\_naissance** par la colonne **Age**.

- Nom des garanties :

Ensuite, nous avons remplacé les colonnes **Code\_tarif**, **Libellé\_tarif**, **Code\_option** et **Nom\_option** par une seule colonne : **Nom\_option** provenant de la table de correspondance des codes garanties. Cette colonne contient 13 valeurs : GRILLE01, GRILLE02, ..., GRILLE13.

- Suppression des lignes répétées (doublons)

Après la fusion des trois bases, nous avons détecté des lignes qui correspondent au même adhérent sur la même année avec la même information sur toutes les colonnes y compris les dates de début et dates de fin. Nous avons supprimé ces doublons afin de ne pas avoir des lignes avec une exposition supérieure à 1 après l'agrégation par bénéficiaire.

Après l'analyse des variables et l'application de ces traitements sur les trois fichiers, nous avons fusionné toutes ces bases d'effectifs en une seule base en agrégeant la variable « **Exposition** » par somme.

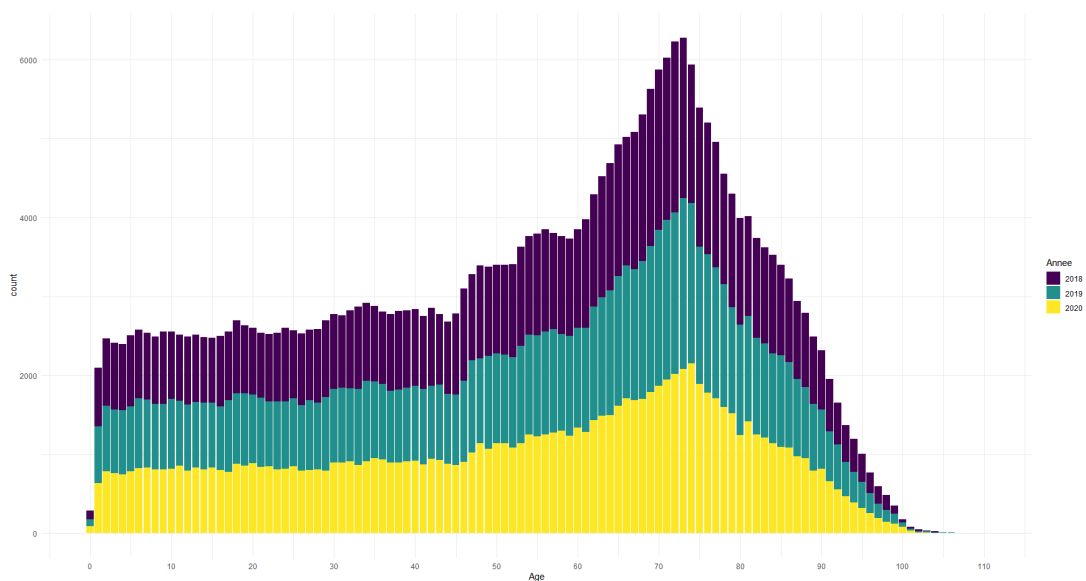
- Agrégation de la base des effectifs

Par chaque **Numero\_adh**, **Age**, **Année** et **Nom\_option** nous avons calculé l'exposition totale. Si l'exposition est égale à 1, cela signifie que le bénéficiaire est assuré pour une durée totale d'un an pour l'année indiquée sur la colonne **Année**.

La base finale contient **136 690** assurés (soit **126 301** consommateurs<sup>4</sup>) avec une exposition totale égale à **316 762** sur les trois ans, répartie de la manière suivante :

---

<sup>4</sup> Qui ont au moins un sinistre sur les trois années étudiées.



▲ Figure 31 : Répartition de l'exposition par âge et par Année

## 6 Etude de la base de prestations

La base des prestations recense tous les sinistres survenus pendant les années 2018, 2019 et 2020 pour les bénéficiaires présents dans notre base d'effectifs.

La base des prestations est répartie sur 61 fichiers csv. Le premier travail qui a été fait est d'importer ces fichiers sous R sous une forme unique et homogène (même choix des variables et mêmes noms des colonnes), puis de les fusionner afin d'avoir une seule base à traiter.

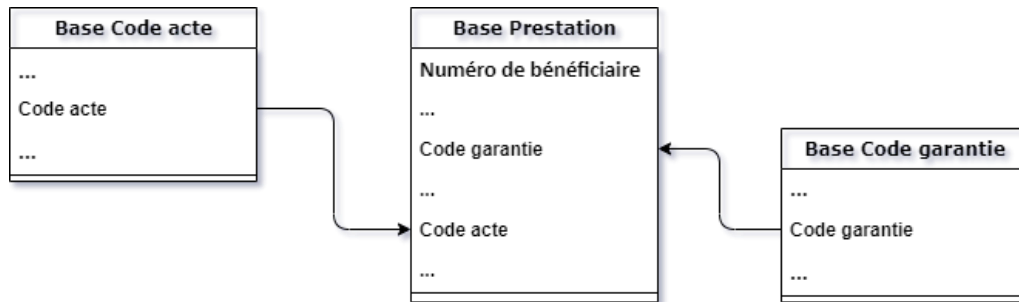
Cette base contient 14 variables en colonnes et plus de 18 millions de lignes de remboursement. Le tableau suivant présente l'ensemble des variables disponibles dans notre base :

Nom de variable	
Numero_adh	Code_option
Code_tarif	Nom_option
Libelle_tarif	Depenses
Annee	RC
CAS	RO
Code_acte	Nb_actes
Lib_acte	CAS / Hors CAS

▲ Figure 32 : Variables de la base des prestations (base brute)

Ensuite, nous avons regroupé les codes actes sur la colonne **Code\_acte** et la colonne **CAS/Hors CAS** en utilisant les tables de correspondance construites dans la partie « 3 Etude des codes actes ».

A l'aide des colonnes **RC**, **RO**, **Dépenses** et **Nb\_actes**, nous avons calculé les frais réels moyens, remboursement moyen du complément santé et nous avons estimé la base de remboursement moyenne par code acte et par garantie. Ce traitement est résumé dans la figure ci-dessous :



▲ Figure 33 : Filtres appliqués sur la base des prestations

### Agrégation des montants de remboursement

Avant d'agréger les montants dans la base des prestations, nous avons supprimé les lignes que nous n'allons pas utiliser dans la suite de la modélisation. Nous avons décidé de garder les variables : **Numero\_adh**, **Nom\_option**, **Annee**, **Code\_acte**, **Dépenses**, **RC**, **RO**, **Nb\_actes**.

Pour chaque numéro de bénéficiaire, nom d'option, code acte et année de survenance nous avons calculé la somme des remboursements de l'assureur (**RC**), somme de remboursement de sécurité sociale (**RO**), somme des dépenses (**Dépenses**) ainsi que le nombre de soins réalisés (**Nb\_actes**). Il nous reste ainsi plus qu'une ligne par bénéficiaire, par code acte, nom d'option et par année de survenance.

La base finale contient **1 967 028** lignes de prestations payées pour un volume de remboursement du complément santé égal à **85 119 324 €**.



La base des prestations issue de ces traitements est de la forme suivante :

Numéro_adh	Nom_option	Annee	Acte	Depenses	RC	RO	Nb_actes
4430	Option_1	2018	Actes d'échographie - OPTAM	48,80 €	23,75 €	36,83 €	1
3493	Option_3	2019	Contraception	39,10 €	31,36 €	0,13 €	2
4953	Option_12	2018	Consultations de généralistes - NON OPTAM	97,40 €	25,39 €	15,78 €	5
2394	Option_12	2019	Inlay onlay	419,29 €	43,63 €	47,19 €	1
1222	Option_22	2018	Lentilles acceptées	223,63 €	123,20 €	22,34 €	2

▲ Figure 34 : Échantillon de la base finale des prestations

## 7 Fusion des bases des prestations et des effectifs

La création de l'outil de tarification nécessite la modélisation de chaque libellé acte, ce qui revient à générer une base de données pour chaque libellé brochure.

Pour cela, nous avons filtré à chaque fois sur un libellé brochure dans la base des prestations, puis nous avons fusionné cette base avec la base entière des effectifs. Le résultat est un ensemble de 75 bases contenant le même nombre de données puisque les deux bases contiennent une ligne par adhérent, par année et par nom d'option. Devant chaque numéro d'adhérent, nous avons un montant en € si cet adhérent est bénéficiaire d'au moins un soin pendant l'année N de la ligne, sinon le code R va mettre NA comme valeur par défaut sur cette colonne. Le process est répété pour les autres colonnes de la base des prestations (**RO**, **Dépenses** et **Nb\_actes**).

Dans un second temps, nous remplaçons les valeurs manquantes (**NA**) par des 0, puis utilisons la table de correspondance entre les libellés brochures, les noms des options et les niveaux de remboursement pour remplacer les noms des options par leurs niveaux de remboursement. Finalement, nous avons supprimé la colonne **Numéro\_adh** et nous avons divisé la colonne **RC** par **Nb\_acte** pour avoir le coût moyen, nous avons renommé cette colonne par **CM**. Notons que le coût total des sinistres diminue de 0,92%, ceci est expliqué par une application de l'arrondi lors la fusion des données afin d'avoir le nombre total d'actes par coût moyen.

La base finale après tous les traitements, les agrégations et la fusion des données dont nous disposons est sous la forme suivante :

Niv_remb	Age	Exposition	Nb_acte	CM
160	10	1,000	2	27,40 €
160	15	0,121	0	0,00 €
180	26	0,997	0	0,00 €
160	25	0,997	1	14,00 €
160	11	0,997	0	0,00 €

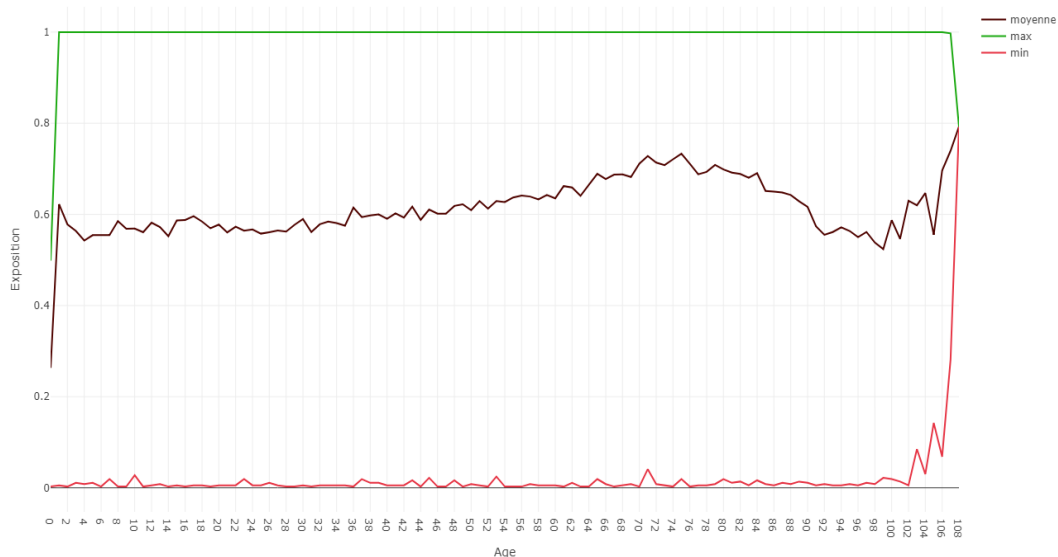
▲ Figure 35 : Exemple de la base finale après les traitements

## 8 Statistique descriptive de la base définitive

Dans cette partie, nous allons présenter les chiffres de la base finale utilisée par notre modèle GLM afin de vérifier si les données sont en cohérence avec les prix de revient.

### Suffisance d'exposition

Le premier contrôle réalisé consiste en la vérification du nombre de données pour chaque variable. Nous avons agrégé par moyenne, par max et par min les expositions de la base par âge pour vérifier que nous avons suffisamment d'exposition sur les âges. Nous avons le résultat suivant :

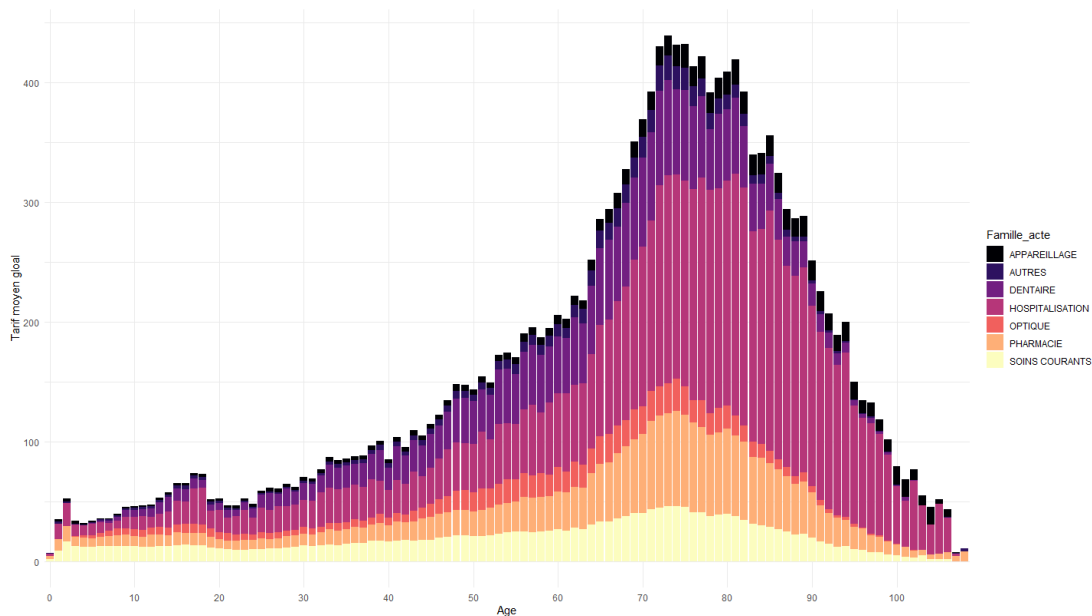


▲ Figure 36 : Exposition moyenne, min et max en fonction des âges

L'exposition minimale est très proche de 0 et l'exposition maximale est à 1 sur tous les âges. L'exposition moyenne est autour de 0,6 ce qui permet d'avoir un tarif annuel homogène. Pour l'âge de 0 (les enfants qui ont moins d'un an), l'exposition moyenne est autour de 0,2. Les tarifs pour cet âge sont à prendre par prudence. Notons que la somme du coût total des bénéficiaires qui ont moins d'un an est de 52 681,2 €, soit 0,06% du coût total des prestations payées ce qui est négligeable.

### Composition du Coût moyen mensuel

Pour avoir une vision sur la composition du tarif global de la population modélisée, nous avons calculé, pour chaque âge, le coût moyen qui est le coût du sinistre divisé par le nombre des adhérents assurés en prenant l'exposition moyenne. Le graphique suivant donne l'information sur le tarif mensuel par âge avec l'information sur le tarif moyen de chaque famille d'acte :

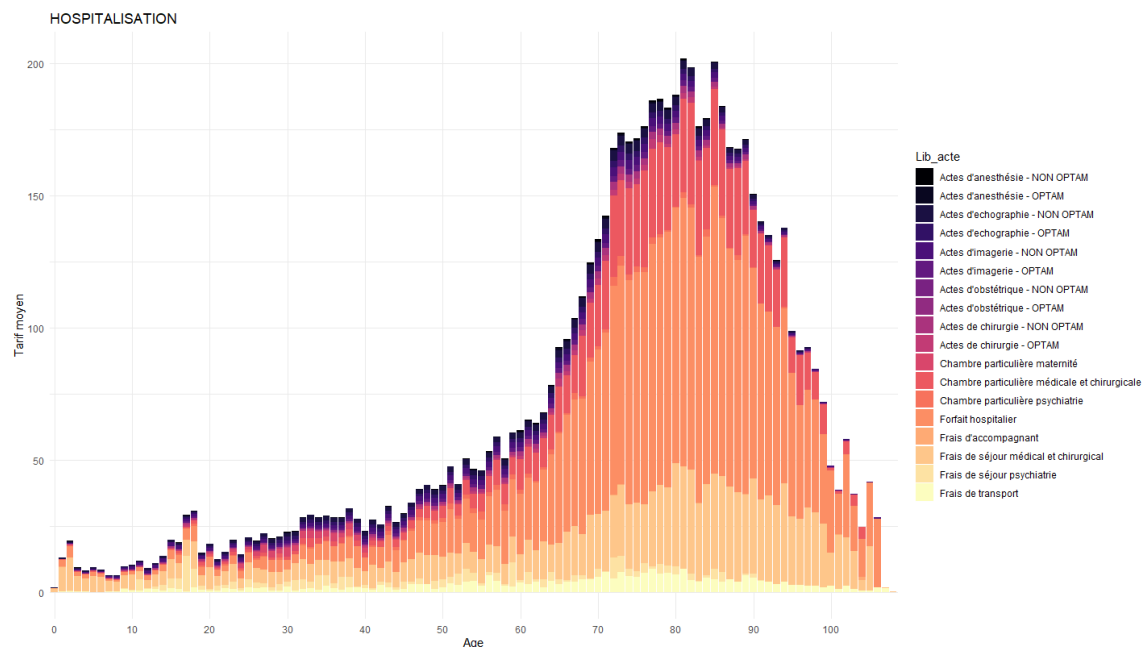


▲ Figure 37 : Coût moyen mensuel par âge

Le tarif mensuel est expliqué principalement le coût de l'hospitalisation puis par le coût des postes dentaire et pharmacie, surtout pour les plus âgés. Cette répartition est cohérente avec la typicité des contrats modélisés (individuels). En effet, une modélisation de contrats collectifs aurait donné des poids plus importants pour l'optique et le dentaire. Dans la suite, nous détaillons chaque poste pour trouver les libellés brochures les plus onéreux et pour expliquer également les hausses entre les âges 70 et 90 ans et les âges 10 et 20 ans également.

## Composition du coût moyen pour le poste : Hospitalisation

La figure suivante illustre la composition du coût moyen des sinistres sur le poste 'Hospitalisation' sur notre base de données :



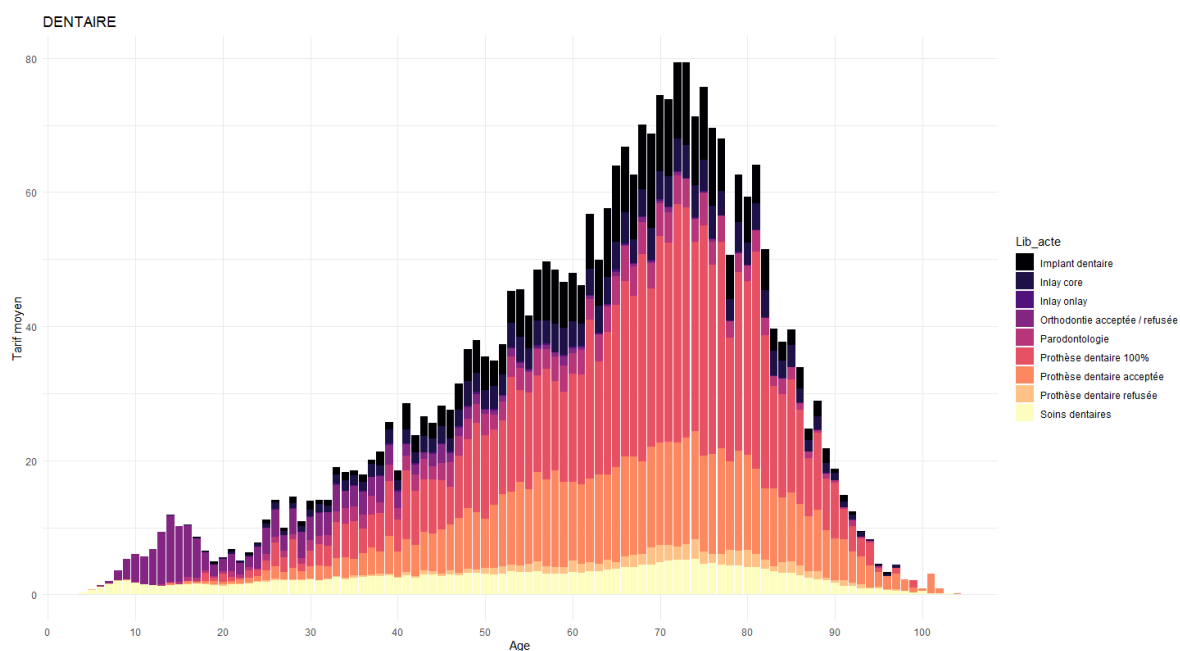
▲ Figure 38 : Coût moyen mensuel par âge : Hospitalisation

Pour l'hospitalisation, le forfait hospitalier et la chambre particulière augmentent fortement à partir de l'âge de 45 ans, ceci est expliqué par l'intégration de la base des prestations de 2020, cette année a été particulière au niveau des remboursements sur ses libellés brochures (sans 2020 le tarif pour 90 ans passe de 88€ à 60€)

Notons que dans la partie application (dernière partie), nous allons faire deux modèles, un sera avec l'année 2020 et l'autre est hors 2020 afin d'analyser l'impact de la Covid-19 sur les tarifs.

## Composition du coût moyen pour le poste : Dentaire

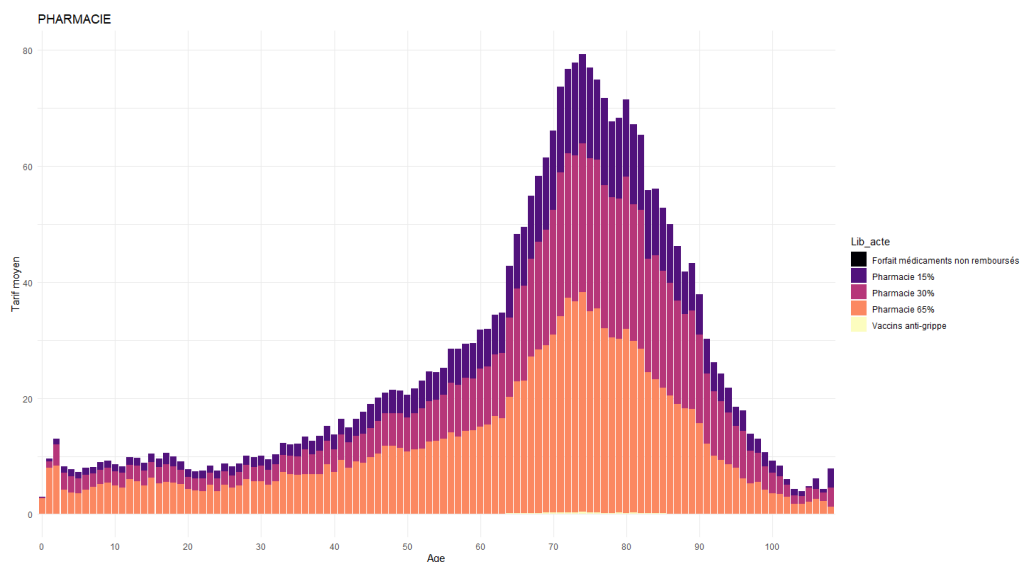
Sur le poste 'Dentaire', nous constatons (figure ci-dessous) que le coût moyen de sinistre est expliqué par le coût moyen des prothèses dentaires pour les plus âgés ce qui explique la hausse entre 70 et 80 ans et par l'orthodontie pour les enfants ce entre les âges 10 et 20 ans.



▲ Figure 39 : Coût moyen mensuel par âge : Dentaire

### Composition du coût moyen pour le poste : Pharmacie

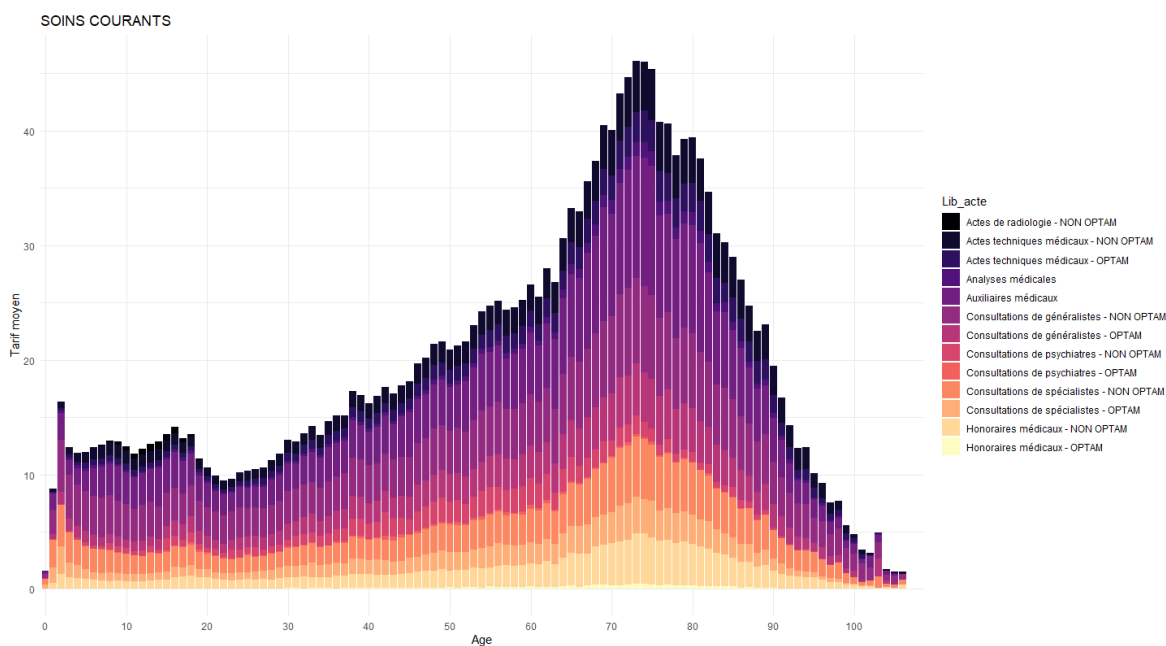
Le coût du poste 'Pharmacie' est expliqué globalement par les libellés brochures : 'Pharmacie 30%' et 'Pharmacie 65%'.



▲ Figure 40 : Coût moyen mensuel par âge : Pharmacie

## Composition du Coût moyen pour le poste : Soins courants

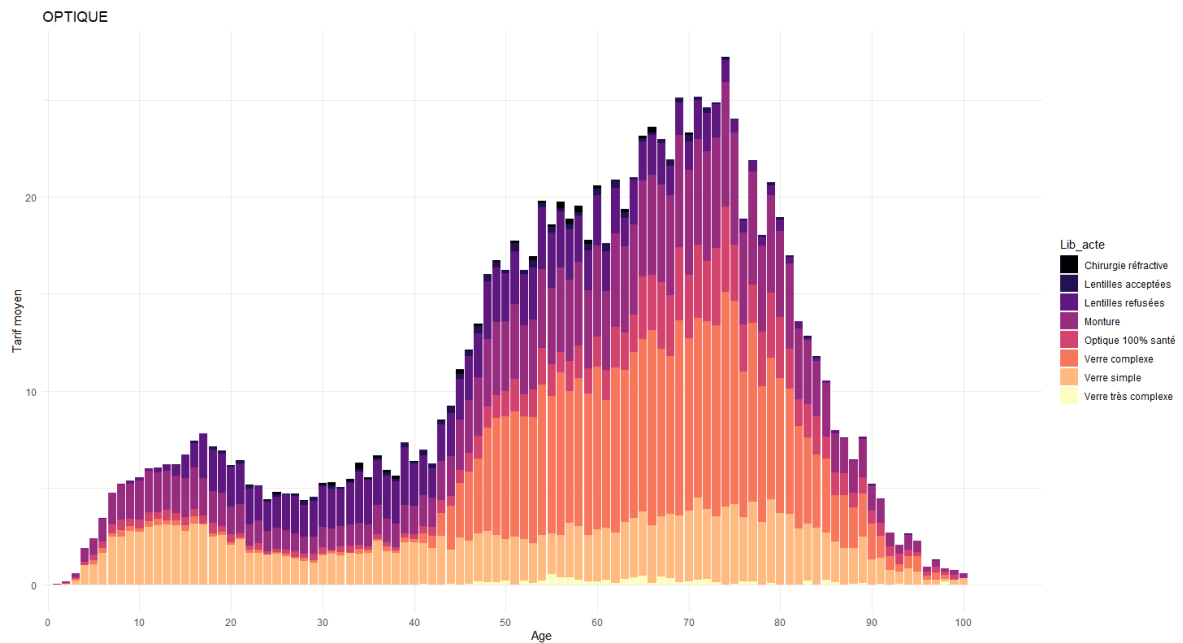
Le cout de sinistre est globalement homogène sur tous les libellés du poste. Ce sont les auxiliaires qui coûtent cher surtout pour les assurés de la tranche d'âge [70, 80].



▲ Figure 41 : Tarif mensuel moyen par âge : Soins courants

## Composition du coût moyen pour le poste : Optique

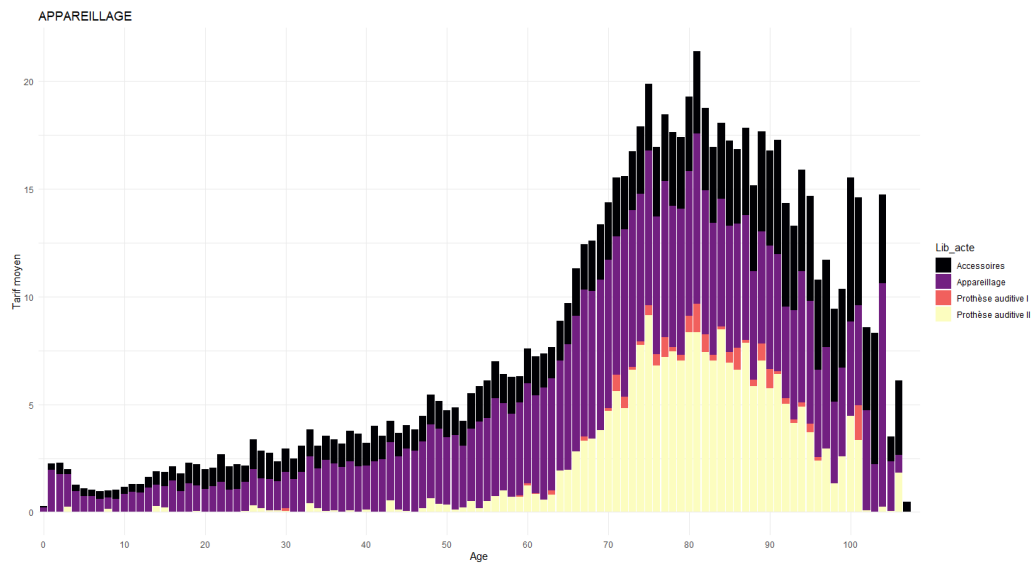
Pour le 'Optique', le coût des verres simples est globalement homogène sur tous les âges. Les verres complexes sont les plus coûteux pour les âges entre 45 et 90 ans et sont négligeables sur le reste des âges. Les lentilles refusées coutent plus chères également pour la tranche d'âge [45,75].



▲ Figure 42 : Coût moyen mensuel par âge : Optique

### Composition du coût moyen pour le poste : Appareillage

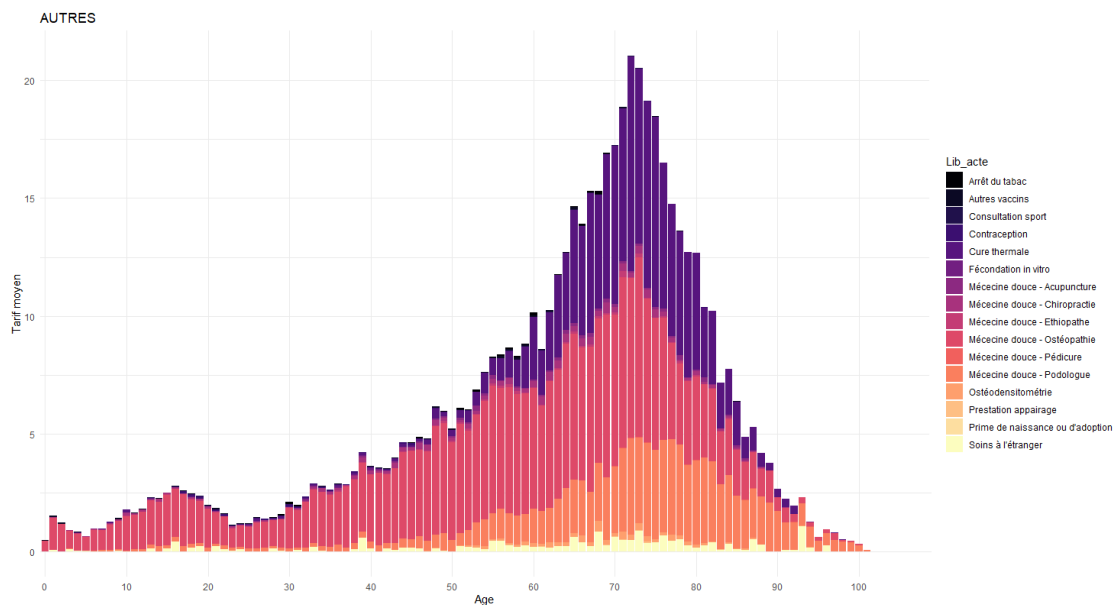
Nous notons la forte consommation de prothèses auditives à partir de 70 ans ce qui est cohérent avec ce que nous pressentions.



▲ Figure 43 : Coût moyen mensuel par âge : Appareillage

## Composition du coût moyen pour le poste : Autres

Ce poste est composé principalement des actes de prévention, de la médecine douce et de la cure thermal. Les actes de médecine douce sont les actes les plus onéreux. L'ostéopathie et la cure thermique commencent à impacter fortement le coût moyen du poste à partir de l'âge 55 ans et cela jusqu'à l'âge de 80 ans.



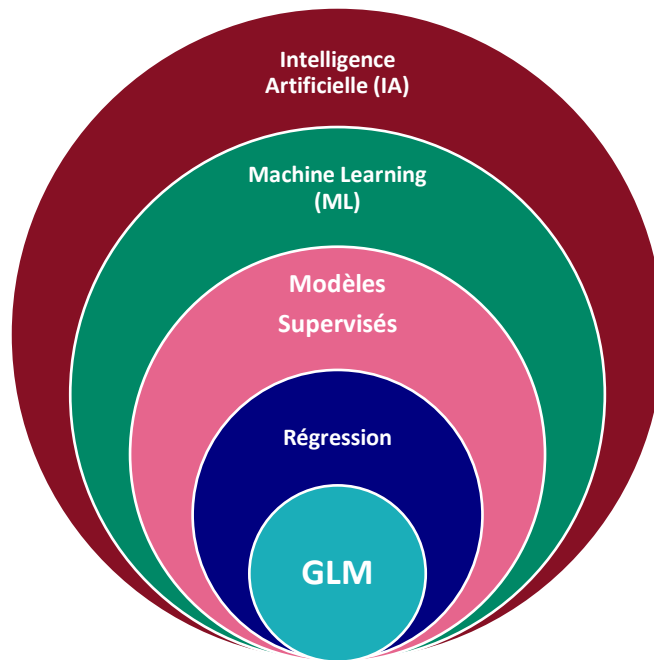
▲ Figure 44 : Coût moyen mensuel par âge : Autres



## Partie III - Construction de l'outil de tarification

La construction de notre outil de tarification est fondée sur des modèles linéaires généralisés (GLM).

Un modèle linéaire généralisé est un modèle supervisé de l'apprentissage statistique (Machine Learning).

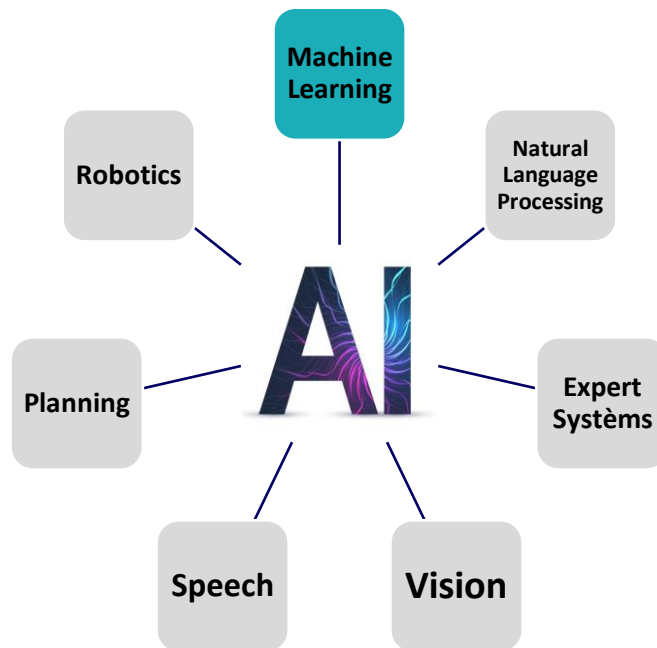


▲ Figure 45 : Situation de GLM par rapport à l'IA

### 1 Introduction : Intelligence Artificielle (AI)

L'Intelligence Artificielle (AI) est une science comme les mathématiques ou la biologie. Elle étudie les moyens de créer des programmes et des machines intelligentes capables de résoudre des problèmes de manière créative, ce qui a toujours été considéré comme une prérogative humaine.

Le Machine Learning (ML) ou l'apprentissage statistique est un sous-domaine de l'Intelligence Artificielle qui offre aux systèmes la possibilité d'apprendre et de s'améliorer automatiquement à partir de l'expérience sans être explicitement programmés.



▲ Figure 46 : Situation du Machine Learning par rapport à l'IA

Un algorithme d'apprentissage statistique peut modéliser un problème de différentes manières en fonction de son interaction avec l'expérience ou l'environnement ou les caractéristiques des données d'entrée.

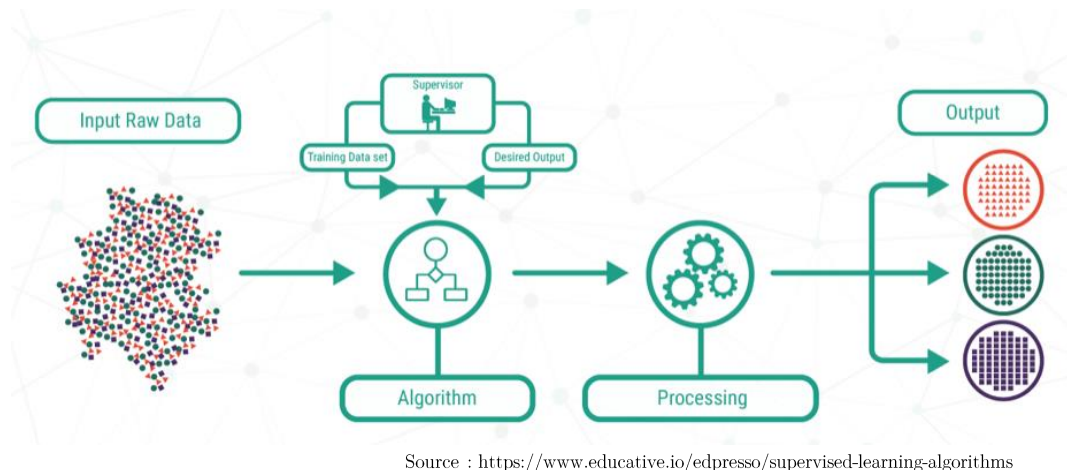
Il est courant dans les manuels d'apprentissage statistique et d'intelligence artificielle de considérer d'abord les styles d'apprentissage qu'un algorithme peut adopter. Cette manière d'organiser ces algorithmes est utile car elle nous oblige à réfléchir aux rôles des données d'entrée, au processus de préparation du modèle et à sélectionner le plus approprié pour notre problème afin d'obtenir les meilleurs résultats. On distingue quatre types d'apprentissage :

- Apprentissage supervisé
- Apprentissage semi-supervisé
- Apprentissage non supervisé
- Apprentissage par renforcement

## 2 Apprentissage supervisé

L'apprentissage supervisé est une méthode de Machine Learning. Elle se produit lorsque le modèle est formé sur un ensemble de données étiquetées. L'ensemble des données étiquetées est celui qui a à la fois des paramètres d'entrée et de sortie.

Dans ce type d'apprentissage, les ensembles de données de formation et de validation sont étiquetées comme indiqué dans la figure ci-dessous.



Source : <https://www.educative.io/edpresso/supervised-learning-algorithms>

▲ Figure 47 : Apprentissage supervisé

Ici, les données d'entrée sont appelées des données d'apprentissage ou 'training data'. Ex : historique des cours boursiers, historique des prestations payées par un assureur, etc.

Un modèle est préparé à travers un processus de formation ou de 'training' dans lequel il est nécessaire de faire des prédictions. Il est corrigé lorsque ces prédictions sont fausses ou biaisées. Le processus d'apprentissage se poursuit jusqu'à ce que le modèle atteigne le niveau de précision souhaité sur les données d'entraînement.

La modélisation linéaire généralisée (GLM) et le réseau neuronal sont deux exemples de ce style d'apprentissage statistique.

#### Entraînement du modèle :

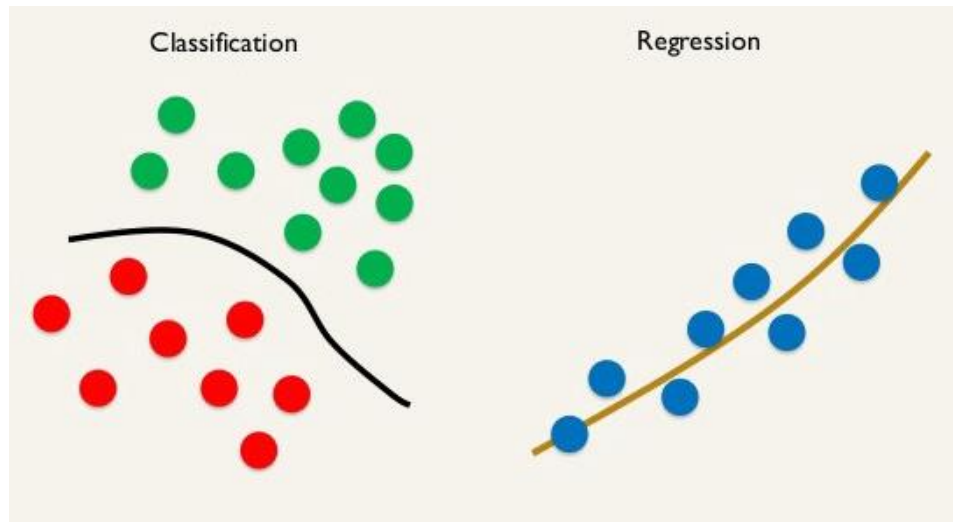
Lors de l'entraînement du modèle, les données sont généralement divisées dans un rapport de 60-40, c'est-à-dire 60% sous forme de données d'entraînement et 40% sur des données de test. Dans les données d'entraînement, nous alimentons les entrées et les sorties pour 60% des données.

Le modèle apprend uniquement à partir des données d'entraînement. Nous utilisons différents algorithmes d'apprentissage automatique afin de construire notre modèle. Le modèle construira sa propre logique en apprenant des données.

Une fois que le modèle est prêt, il doit être testé. Au moment du test, l'entrée est alimentée à partir des 40% de données restantes que le modèle n'a jamais vu auparavant, le modèle prédira une valeur que nous comparerons avec la sortie réelle et nous calculerons la prédiction.

## Types d'apprentissage supervisé :

Nous distinguons deux types d'apprentissage supervisés :



Source de la figure : <https://www.codeingschool.com/2019/06/regression-classification-supervised-machine-learning.html>

▲ Figure 48 : Types d'apprentissage supervisé

### A. Classification :

Il s'agit d'une tâche d'apprentissage supervisé où la sortie a des étiquettes définies (valeur discrète). Par exemple, dans A de la figure ci-dessus. Le but est de prédire des valeurs discrètes appartenant à une classe particulière  $\{0,1, \dots, n\}$  et de les évaluer sur la base de la prédiction.

Il peut s'agir d'une classification binaire ou multilinéaire. Dans la classification binaire, le modèle prédit 0 ou 1, oui ou non. En cas de classification multilinéaire, le modèle prédit plusieurs classes (ordonnées ou non ordonnées).

### B. Régression :

Il s'agit d'une tâche d'apprentissage supervisé où la sortie a une valeur continue. Par exemple, dans B de la figure ci-dessus, la variable RC n'a pas de valeurs discrètes mais est continue. Le but est de prédire une valeur aussi proche de la valeur de sortie réelle que notre modèle le peut, puis l'évaluation est effectuée en calculant la valeur d'erreur. Plus l'erreur est petite, plus la prédiction de notre modèle de régression est grande.

Le modèle linéaire généralisé est un modèle qui appartient à la famille des régressions linéaires. Nous allons présenter d'abord les méthodes de régression, la différence entre la régression linéaire et la régression non linéaire.

TYPE_BENEF	AGE	SEXE	GARANTIE	EXPOSITION	FREQUENCE
Assuré	42	F	220	1,00	33
Assuré	34	H	220	1,00	33
Conjoint	55	F	220	1,00	32
Assuré	57	F	220	1,00	32
Assuré	59	F	220	1,00	32
Enfant	6	H	220	1,00	32
Conjoint	57	F	250	1,00	32
Assuré	56	H	150	1,00	32
Enfant	25	F	220	1,00	31
Conjoint	46	F	220	0,87	31
Assuré	53	F	220	1,00	31
Assuré	54	F	220	1,00	31
Conjoint	55	F	220	1,00	31
Assuré	56	F	220	1,00	31
Conjoint	56	F	220	1,00	31
Assuré	45	H	220	1,00	31
Assuré	57	H	220	1,00	31
Conjoint	35	F	150	1,00	31
Assuré	60	H	150	1,00	31
Enfant	3	F	220	1,00	30
Assuré	41	F	220	1,00	30
Conjoint	42	F	220	1,00	30
Conjoint	48	F	220	1,00	30
Assuré	53	F	220	1,00	30
Conjoint	59	F	220	1,00	30
Assuré	54	H	220	1,00	30
Assuré	61	H	150	1,00	30

A - Classification

TYPE_BENEF	AGE	SEXE	GARANTIE	RC
Enfant	14	H	220	37,88 €
Enfant	6	H	220	37,74 €
Enfant	10	H	220	37,74 €
Assuré	46	H	220	37,74 €
Enfant	9	F	220	37,70 €
Enfant	8	F	220	37,50 €
Enfant	17	F	220	37,50 €
Enfant	19	F	220	37,50 €
Assuré	47	F	220	37,50 €
Conjoint	52	F	220	37,50 €
Conjoint	55	F	220	37,50 €
Assuré	59	F	220	37,50 €
Enfant	11	H	220	37,50 €
Assuré	30	H	220	37,50 €
Assuré	46	H	220	37,50 €
Assuré	63	H	220	37,50 €
Enfant	7	H	150	37,26 €
Assuré	56	F	220	37,23 €
Enfant	18	H	220	37,10 €
Assuré	62	F	220	37,06 €
Enfant	12	H	220	37,06 €
Enfant	12	H	220	37,06 €
Assuré	39	H	220	37,06 €
Assuré	60	H	220	37,06 €
Assuré	49	F	220	36,92 €
Conjoint	45	F	220	36,82 €
Conjoint	55	H	220	36,78 €

B - Régression

▲ Figure 49 : Types d'apprentissage supervisé

### 3 Modèles de régression

Le but de la régression est de construire un modèle pour prédire avec précision les cas inconnus.

La régression est généralement le processus de prédiction d'une variable continue comme le prix d'une mutuelle de santé, le prix d'une action boursière, etc. en utilisant des données historiques.

Dans la régression non linéaire, nous distinguons deux types de régressions :

- Régression simple
- Régression multiple

Ces régressions peuvent être linéaires ou non linéaires. La linéarité de la régression est fondée sur la nature de la relation entre les variables indépendantes et dépendantes.

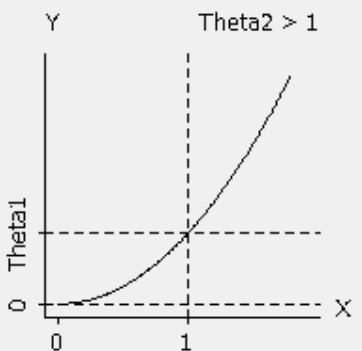
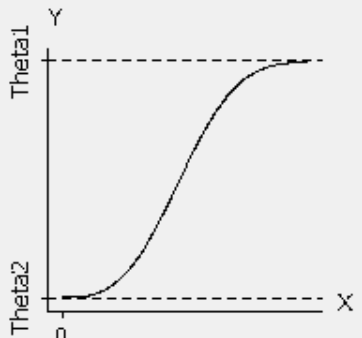
## A. Régression non linéaire (NLR) :

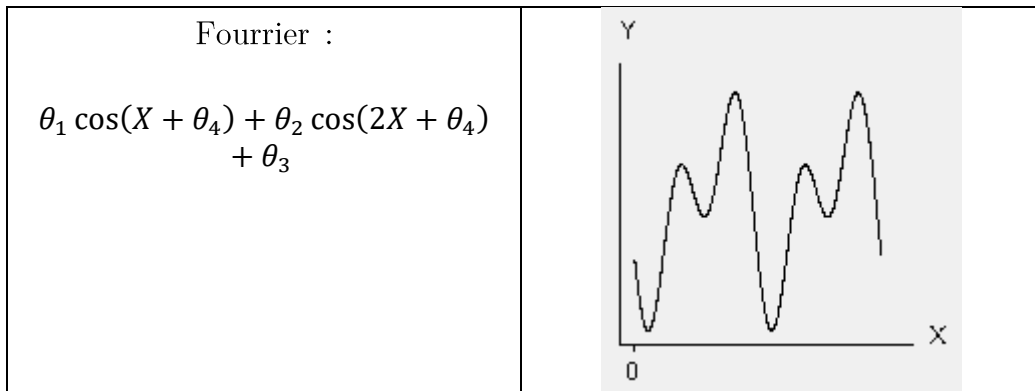
La régression non linéaire est toute relation entre une variable indépendante X et une variable dépendante Y qui se traduit par des données modélisées par une fonction non linéaire.

Essentiellement, toute relation qui n'est pas linéaire est qualifiée de non linéaire et est généralement représentée par le polynôme à k degrés (puissance maximale de X).

La flexibilité de ce type de modèle ouvre la porte à un grand nombre de formes possibles. Par conséquent, la régression non linéaire peut s'adapter à une grande variété de courbes.

Ci-dessous, nous présentons un ensemble d'exemples qui illustre la diversité des modèles de régression non linéaire.

Equations non linéaire	Graphique
<p>Puissance :</p> $\theta_1 \cdot X^{\theta_2}$	 <p>Y</p> <p>Theta2 &gt; 1</p> <p>Theta1</p> <p>0</p> <p>X</p> <p>0 1</p>
<p>Croissance Weibull :</p> $\theta_1 + (\theta_2 - \theta_1) \cdot \exp(-\theta_3 \cdot X^{\theta_4})$	 <p>Y</p> <p>Theta1</p> <p>Theta2</p> <p>0</p> <p>X</p>



▲ Figure 50 : Exemple de la régression non linéaire

Il est important de noter que  $R^2$  n'est pas valide pour les modèles non linéaires et que les logiciels statistiques ne peuvent pas calculer les p-values pour les variables indépendantes.

### B. Régression linéaire (LR) :

Un modèle est dit linéaire si ses paramètres et ses variables sont linéaires, ainsi que sa forme fonctionnelle. Un paramètre est linéaire s'il n'est pas multiplié avec un autre paramètre et s'il n'est pas élevé à une puissance supérieure à 1.

Ci-dessus, nous distinguons les deux types de régression linéaire : simple et multiple :

#### i. Régression linéaire simple

La régression linéaire simple est utile pour trouver la relation entre deux variables continues. La première variable est la variable indépendante (prédicteur), et l'autre est la variable dépendante (réponse).

L'objectif de la régression linéaire simple est de rechercher une relation statistique à condition qu'elle soit indéterministe. Cette relation illustre la ligne optimale qui passe par le maximum des couples  $(X, Y)$ .

L'équation de cette ligne est sous la forme :

$$Y = \beta_1 \cdot X + \beta_0 .$$

Avec :

- $Y$  : la variable expliquée.
- $X$  : la variable explicative.
- $\beta_1$  : le paramètre qui correspond à la variable  $X$ .
- $\beta_0$  : le paramètre d'interception.

Les paramètres  $\beta_1$  et  $\beta_0$  doivent être choisis de manière à minimiser l'erreur.

## ii. Régression linéaire multiple

La régression linéaire multiple est une technique statistique qui utilise plusieurs variables explicatives pour prédire le résultat d'une variable de réponse. Son but principal est de modéliser la relation linéaire, les variables explicatives (indépendantes) et la variable de réponse (dépendante).

Un modèle linéaire multiple repose sur les hypothèses suivantes :

### a) Hypothèses stochastiques :

- $H_1$  : Les  $X_i$  sont déterminées sans erreurs  $\forall i = 1, \dots, n$
- $H_2$  :  $\mathbb{E}(\epsilon_i) = 0, \forall i$  : Le modèle est bien spécifié en moyenne
- $H_3$  :  $\text{Var}(\epsilon_i) = \sigma^2, \forall i$  : Homoscédasticité des erreurs (variance constante)
- $H_4$  :  $\text{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$  : Pas d'autocorrélation des erreurs
- $H_5$  :  $\text{Cov}(X_i, \epsilon_j) = 0, \forall i \neq j$  : Les erreurs sont linéairement indépendantes des variables exogènes
- $H_6$  :  $\epsilon \sim \mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$  : Les erreurs suivent une loi normale multidimensionnelle

Remarque : L'hypothèse  $H_6$  implique  $H_2, H_3, H_4$ .

### b) Hypothèses structurelles :

- $H_7$  : Absence de colinéarité entre les variables explicatives, i.e. :  $X^t X$  est régulière,  $\det(X^t X) \neq 0$  et  $(X^t X)^{-1}$  existe

Remarque : c'est équivalent à  $\text{rang}(X) = \text{rang}(X^t X) = p + 1$

- $H_8$  :  $\frac{1}{n} \cdot X^t X$  tend vers une matrice finie non singulière  $Q$  lorsque  $n \rightarrow +\infty$ .
- $H_9$  :  $n > p + 1$  Le nombre d'observations est supérieur au nombre de variables +1 (la constante).

## iii. Estimation des paramètres

L'estimation des paramètres d'un modèle linéaire multiple sous la forme :

$$y_i = \beta_0 + \beta_1 \cdot x_{i,1} + \dots + \beta_p \cdot x_{i,p} + \epsilon_i$$

est fondée sur la méthode des moindres carrés ordinaires (MCO). Ces paramètres vont nous permettre d'estimer la valeur de  $\hat{y}_i$  en fonction des autres variables à travers l'équation :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i,1} + \dots + \hat{\beta}_p \cdot x_{i,p}$$

Où les  $\hat{\epsilon}_i = y_i - \hat{y}_i$  sont les résidus estimés entre la valeur observée et la valeur estimée.



La méthode des moindres carrés consiste à rechercher les valeurs des paramètres qui minimisent la somme des carrés des résidus.

$$\min \sum_{i=1}^n \hat{\epsilon}_i = \min_{\hat{\alpha}_0, \dots, \hat{\alpha}_p} \sum_{i=1}^n (y_i - \hat{\beta}_1 \cdot x_{i,1} - \dots - \hat{\beta}_p \cdot x_{i,p})^2$$

Ce qui revient à chercher les solutions de  $\frac{\partial(\sum \hat{\epsilon}_i^2)}{\partial \hat{\beta}_j} = 0$

Cela revient à résoudre un système de  $p+1$  équations normales. En passant l'opérateur de dérivation dans la somme, on a :

$$\sum_{i=1}^n x_{i,j} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_{i,1} - \dots - \hat{\beta}_p \cdot x_{i,p}) = 0$$

L'écriture de cette équation par la forme matricielle nous donne :

$$X^t(Y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta} = (X^tX)^{-1}X^tY$$

### c) Propriétés de l'estimation par MCO

- L'estimateur MCO est sans biais ( $\mathbb{E}(\hat{\beta}) = \beta$ ) sous les hypothèses  $H_1, H_2$  et  $H_5$ .

$$\begin{aligned} \text{En effet : } \mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^tX)^{-1}X^tY] \\ &= \mathbb{E}[\beta + (X^tX)^{-1}X^tY \cdot \epsilon] \\ &= \beta + (X^tX)^{-1}X^t \cdot \mathbb{E}[\epsilon] \quad (\text{sous } H_1 \text{ et } H_5) \\ &= \beta \end{aligned}$$

- L'estimateur MCO est le meilleur estimateur linéaire sans biais, sous les hypothèses  $H_1$  à  $H_5$ .
- L'estimateur MCO est distribué selon une loi normale  $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^tX)^{-1})$  sous les hypothèses  $H_1, H_2$  et  $H_6$ .
- L'estimateur MCO converge en probabilité vers  $\beta$  ( $\hat{\beta} \xrightarrow{\mathbb{P}} \beta$ ) sous les hypothèses  $H_6$  et  $H_8$ .

En réécrivant  $\hat{\beta} = \beta + \left(\frac{X^t X}{n}\right)^{-1} \frac{X^t \epsilon}{n}$

En passant la limite en probabilité nous avons :

$$\lim_{\mathbb{P}} \hat{\beta} = \beta + \lim_{\mathbb{P}} \left[ \left(\frac{X^t X}{n}\right)^{-1} \cdot \frac{X^t \epsilon}{n} \right]$$

D'après  $H_8$  :  $\frac{X^t X}{n} \xrightarrow{n \rightarrow +\infty} Q$  définie positive et nous avons :

$$\lim_{\mathbb{P}} \hat{\beta} = \beta + Q^{-1} \lim_{\mathbb{P}} \left(\frac{X^t X}{n}\right)$$

Or  $\mathbb{E}[x_i \epsilon_i] = 0$  d'après l'hypothèse  $H_6$ , nous avons alors :

$$\lim_{\mathbb{P}} \left(\frac{X^t X}{n}\right) = 0$$

D'où la convergence en probabilité de  $\hat{\beta}$  vers  $\beta$  ( $\hat{\beta} \xrightarrow{\mathbb{P}} \beta$ ).

- L'estimateur MCO suit asymptotiquement une loi normale :

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2 Q^{-1}}{n}\right)$$

- sous les hypothèses  $H_1$  à  $H_5$  et  $H_8$ .

#### d) Calcul de la matrice de variance-covariance des coefficients

Le calcul de la matrice de variance-covariance nous permet de faire des tests d'hypothèses et de calculer aussi l'intervalle de confiance des coefficients calculés, nous avons la variance de l'estimateur :  $\text{Var}(\hat{\beta}) = \sigma^2 (X^t X)^{-1}$ .

En effet, nous avons :  $\hat{a} = a + (X^t X)^{-1} X^t \epsilon$

On obtient :

$$\begin{aligned} \text{Var}[\hat{\beta}] &= \text{Var}[(X^t X)^{-1} X^t \epsilon] \\ &= (X^t X)^{-1} X^t \cdot \text{Var}[\epsilon] \cdot X (X^t X)^{-1} \\ &= (X^t X)^{-1} X^t \cdot \sigma^2 \mathbb{I} \cdot X (X^t X)^{-1} \quad \text{sous } H_3 \text{ et } H_4 \\ &= \sigma^2 \cdot (X^t X)^{-1} X^t \cdot X (X^t X)^{-1} \\ &= \sigma^2 \cdot (X^t X)^{-1} \end{aligned}$$

e) Estimation de la variance du résidu

Pour estimer la variance  $\sigma^2 = \text{Var}(\epsilon)$ , nous allons utiliser l'estimateur sans biais construit à partir de la variance des résidus observés :

$$s^2 \equiv \hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^N \hat{\epsilon}_i^2$$

Où  $n - p - 1$  : correspond au nombre de degrés de liberté du modèle.

Avec  $n$  : nombre d'observations

$p$  : nombre de coefficients à estimer

Les résidus  $\hat{\epsilon}$  correspondent aux résidus observés :  $\hat{\epsilon} = Y - \hat{Y}$

La matrice de variance-covariance des résidus est estimée par :

$$\widehat{\text{Var}}[\hat{\beta}] = \hat{\Sigma}_{\beta} = \hat{\sigma}^2 (\mathbf{X}^t \mathbf{X})^{-1}$$

La variance estimée  $\hat{\sigma}_{\hat{\beta}_j}^2$  du paramètre  $\hat{\beta}_j$  est lue sur la diagonale principale de cette matrice.

f) Estimation des coefficients

Le calcul de l'espérance de l'estimateur et l'estimation de sa variance nous permet de produire une estimation par intervalle et de réaliser des tests d'hypothèses.

En effet, à partir de l'hypothèse :  $\epsilon_i = \hat{\beta}_j - \beta_j \sim \mathcal{N}(0, \sigma^2)$ , nous avons ces deux résultats :

- $\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim \mathcal{N}(0, 1)$
- $(n - p - 1) \frac{\hat{\sigma}_{\hat{\beta}_j}^2}{\sigma_{\hat{\beta}_j}^2} \sim \chi_{n-p-1}^2$

Le rapport d'une loi normale centrée réduite et la racine d'une loi de  $\chi^2$  normalisée par ses degrés de liberté aboutit à une loi Student, nous avons :

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathbf{T}(n - p - 1)$$

### g) Evaluation globale de la régression

L'évaluation globale de la pertinence s'appuie sur l'équation d'analyse de variance :

$$\mathbf{SCT = SCE + SCR}$$

Où :

- **SCT** : Somme des carrées totaux, traduit la variabilité totale de l'endogène, où :

$$\mathbf{SCT = \sum_i (y_i - \bar{y})^2}$$

- **SCE** : Somme des carrées expliqués, traduit la variabilité expliquée par le modèle, où :

$$\mathbf{SCE = \sum_i (\hat{y}_i - \bar{y})^2}$$

- **SCR** : Somme des carrés résiduels, traduit la variabilité non-expliquée par le modèle, où :

$$\mathbf{SCR = \sum_i (\hat{y}_i - \bar{y})^2}$$

$R^2$  est un indicateur spécifique compris entre 0 et 1, permettant de traduire la variance expliquée par le modèle, il s'agit du **coefficient de détermination**. Sa formule est la suivante :

$$\mathbf{R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}}$$

Plus  $R^2$  est proche de 1, plus le modèle est intéressant. En revanche, il ne permet pas de savoir si le modèle est statistiquement pertinent pour expliquer les valeurs de  $Y$ .

La modélisation des données d'assurance nécessite l'utilisation de la régression linéaire généralisée qui est un cas général de la régression linéaire multiple, dans laquelle les coefficients  $\beta$  sont estimés par les méthodes qui suivent la même logique de la régression linéaire multiple, pour laquelle les coefficients  $\beta$  sont estimés par la maximisation de la vraisemblance cette fois. Dans la sous-section suivante nous allons définir un GLM et son application sur notre base de données.

## 4 Modèle Linéaire Généralisé (GLM)

### A. Introduction

Un GLM est défini en spécifiant deux composantes ; la variable réponse (Y) doit être modélisée par une **distribution aléatoire** et **une fonction lien** qui décrit comment la moyenne de la variable réponse est liée à la combinaison linéaire des variables explicatives (indépendantes).

#### i. Distribution aléatoire

La prime pure ‘homogène’ payée par un adhérent pour un libellé brochure par exemple est l’estimation de la consommation de ce libellé par toute la population divisée par le nombre des adhérents. Pour un adhérent avec des caractéristiques spécifiques (âge, niveau de remboursement, zone géographique, etc.), la prime pure est l’estimation de la consommation des adhérents qui ont les mêmes caractéristiques. Dans ce cas la prime pure est égale à l’estimation coût moyen d’une seule consommation multipliée par l’estimation du nombre moyen de consommation par année.

$$\mathbf{Prime\ pure = E(Fréquene) \cdot E(Coût\ Moyen)}$$

*En effet, nous avons la charge totale est C est de :*

$$C = \sum_{i=1}^N X_i$$

*Avec N est le nombre de sinistre par année (aléatoire)*

*Et  $(X_i)_{i \in \mathbb{N}^*}$  est la suite des prestations indépendants et identiquement distribués indexés par  $i \in \mathbb{N}$ .*

*Sous l’hypothèse d’indépendance entre les  $X_i$  et N, nous avons :*

$$\begin{aligned} \mathbb{N}[C] &= \mathbb{E} \left[ \sum_{i=1}^N X_i \right] = \mathbb{E} \left[ \mathbb{E} \left( \sum_{i=1}^N X_i | N \right) \right] = \mathbb{E} \left[ \sum_{i=1}^N \mathbb{E}(X_i | N) \right] \\ &= \mathbb{E}[N \cdot \mathbb{E}(X_i | N)] = \mathbb{E}[N] \cdot \mathbb{E}[X_i | N] = \mathbb{E}[N] \cdot \mathbb{E}[X_i] \end{aligned}$$

*De la même manière nous trouverons la variance de C est :*

$$\mathbb{V}[C] = \mathbb{E}[N] \cdot \mathbb{V}[X_i] + \mathbb{V}[N] \cdot \mathbb{E}[X_i]^2$$

Pour calculer l'espérance du coût moyen, nous allons d'abord choisir la bonne loi qui modélise les coûts moyens. Dans la partie pratique, nous avons choisi la loi la plus adéquate. Pour cela nous avons utilisé la bibliothèque « **gamlss** » qui nous propose une liste vaste de lois positives qui sont listées dans Annexe 3 et 4.

Pour trouver la loi optimale pour nos données, nous avons utilisé la fonction `FitDist` de la même bibliothèque. Cette fonction utilise les critères AIC et BIC et ordonne les meilleures lois présentatrices des données dans un ordre croissant.

## ii. Fonction lien

Lors de l'application de la modélisation linéaire, nous pouvons exprimer l'effet des variables indépendantes sur la variable réponse via un prédicteur linéaire :

$$\boldsymbol{\eta} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \dots + \boldsymbol{\beta}_p x_p = \mathbf{X}^t \boldsymbol{\beta}$$

Dans les GLM, la fonction lien  $g$  permet de décrire comment la moyenne de la variable expliquée  $\mathbb{E}(Y) = \mu$  est liée aux variables indépendantes par :

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{\theta}$$

A vrai dire, toute fonction continue, différentiable et monotone pourra être une fonction lien. Mais, dans la pratique, il existe des choix pratiques et communs pour chaque loi classique choisie en GLM.

La fonction lien standard relativement à sa loi est appelée la fonction lien canonique. Elle est le choix naturel dont les résultats sont les plus faciles à interpréter.

Le tableau ci-dessous représente les fonctions lien standards pour les lois usuelles :

Lois	Fonction lien	Fonction variance
Normale	$\eta = \mu$	1
Poisson	$\eta = \log(\mu)$	$\mu$
Binomiale	$\eta = \log\left[\frac{\mu}{1-\mu}\right]$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	$\mu^2$
Normale-Inverse	$\eta = \mu^{-2}$	$\mu^3$

Le reste des fonctions lien des lois de la bibliothèque sont indiqués dans Annexe 3 et 4.

## B. Test d'adéquation des lois

Le test d'ajustement d'un modèle statistique est une procédure suivie pour décrire dans quelle mesure le modèle s'adapte à un ensemble d'observations. Ce test résume généralement l'écart entre les valeurs observées et les valeurs attendues (théoriques) dans le modèle en question.

Le test d'ajustement peut être utilisé pour prendre une décision de choix entre deux hypothèses statistiques (test de normalité des résidus), pour vérifier si deux échantillons sont tirés de la même distribution (test de Kolmogorov-Smirnov) ou si des fréquences de survenance suivent une distribution spécifiée (test de  $\chi^2$ ).

### i. Critères BIC et AIC

AIC et BIC sont deux critères de sélection d'un modèle parmi un ensemble de modèles. Ils s'appliquent sur les modèles estimés par la méthode de maximum de vraisemblance.

Lors de l'ajustement du modèle, il est possible d'augmenter la probabilité en ajoutant des paramètres, cela peut entraîner un surapprentissage. Les critères AIC et BIC tentent tous les deux de résoudre ce problème en introduisant un terme de pénalité pour le nombre de paramètres dans le modèle.

Le critère BIC est défini par :

$$BIC = k \ln(n) - 2 \ln(\hat{L})$$

Où :

- $\hat{L}$  : la vraisemblance maximisée
- $n$  : la taille de l'échantillon ou le nombre d'observation
- $k$  : le nombre de paramètres estimés par le modèle

Le critère BIC s'exprime également en fonction des autres test connus :

$$BIC = n \ln(SCR/n) + k \ln(n) = \chi^2 + k \ln(n)$$

Où :

- $SCR$  : Somme résiduelle des carrés vu dans (partie modèle linéaire multiple).
- $\chi^2$  : la déviance du modèle

Lors de la sélection de plusieurs modèles, celui avec le BIC le plus bas est préféré. Le BIC est croissant en fonction de la variance de l'erreur  $\sigma_e^2$  et en fonction de  $k$ . Autrement dit, plus on augmente le nombre de variable explicatives, plus on augmente la valeur du BIC.

Un BIC inférieur implique soit moins de variables explicatives, soit un meilleur ajustement, soit les deux.

Le critère AIC est défini par :

$$AIC = 2k - 2 \ln (\hat{L})$$

Ce critère ne permet pas de faire des conclusions sur la qualité d'un modèle tout seul, mais par rapport aux autres modèles. Il est similaire au critère BIC, mais ici la pénalité est différente pour le nombre de paramètres. Avec AIC, la pénalité est  $2k$ , alors qu'avec BIC, la pénalité est  $\ln(n) k$ .



## 5 Application

Dans cette partie, nous allons montrer les étapes suivies pour modéliser un des libellés brochure, nous avons choisi l'acte de 'Chirurgie -NON OPTAM'. Pour le reste, un script R nous permet d'appliquer les mêmes étapes pour tous les libellés-brochure restants d'une manière automatique, notamment le choix de la meilleure loi et le choix du modèle optimal également.

Nous rappelons que la prime brute cherchée est le produit du coût moyen et de la fréquence moyenne de survenance. Dans la suite, nous allons modéliser la fréquence, puis le coût moyen du libellé-brochure choisi.

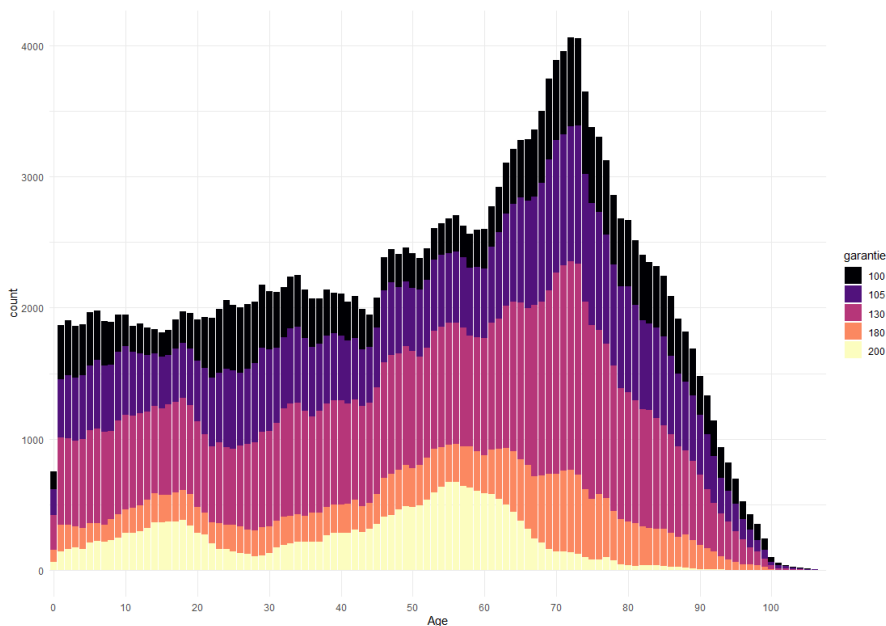
### A. Statistiques sur les variables explicatives

La base finale pour l'acte de 'Chirurgie -NON OPTAM' contient au total 222 831 lignes avec les variables : **garantie**, **Age**, **Exposition**, **CM** et **Nb\_actes** en colonnes.

La variable garantie est numérique, elle prend 5 valeurs différentes :

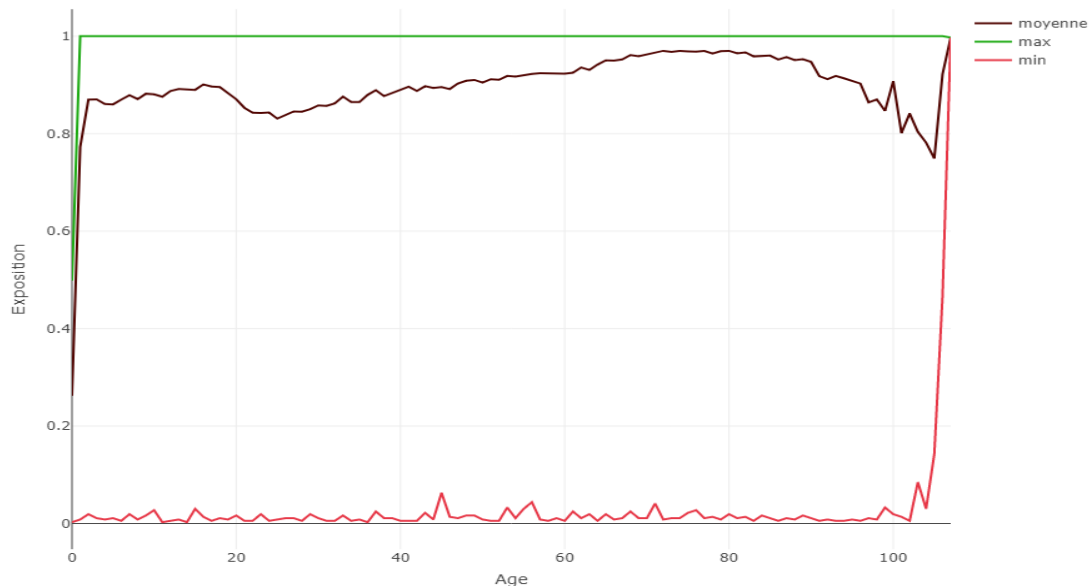
Valeur	100%	105%	130%	180%	200%
Nombre d'occurrences	36 230	55 953	80 435	26 375	23 838

En fonction de la variable 'Age' et de la variable 'garantie', les lignes d'adhérents dans la base sont distribuées de la manière suivante :



▲ Figure 51 : Nombre de d'adhérents en fonction de l'âge et le niveau de garantie

La forme de l'exposition pour l'acte de 'Chirurgie -NON OPTAM' ressemble à celle de la base globale de tous les codes actes ce qui est normal, puisque cet acte est présent sur toutes les tables de garanties. Elle dépasse globalement 0,8 en moyenne comme nous montre la figure suivante :



▲ Figure 52 : Etude de l'exposition pour 'Chirurgie -NON OPTAM'

D'après ces statistiques, les résultats pour les plus âgés et les nouveaux nés seront très volatiles à cause du nombre faible de données.

La précision du modèle sera donc satisfaisante pour les âges compris entre 10 ans et 85 ans et faible pour les âges extrêmes.

### B. Hypothèses : fréquence $\perp$ coût moyen

La décomposition de la prime pure sous forme de produit de l'espérance du coût moyen multiplié par l'espérance de la fréquence est basée sur l'hypothèse de l'indépendance entre les deux.

Pour justifier cette indépendance, nous avons travaillé sur la base du libellé brochure étudié contenant **222 831** lignes. Nous rappelons que l'objectif est de tester l'indépendance entre les deux variables : **CM** et **Nb\_actes**.

Pour ce faire, nous devons construire la table de contingence afin de faire le test de khi-deux. Nous avons découpé les deux variables en intervalles que nous bornons pour ne pas surcharger le tableau et garder les classes d'effectifs non nuls. L'idée du test de khi-deux est de calculer la mesure dite 'test de Pearson' entre toutes les

modalités et de comparer la somme trouvée avec la valeur théorique (en tenant compte le degré de liberté  $k = (n - 1)(p - 1)$ , avec  $n$  : le nombre de lignes et  $k$  est le nombre de colonnes.

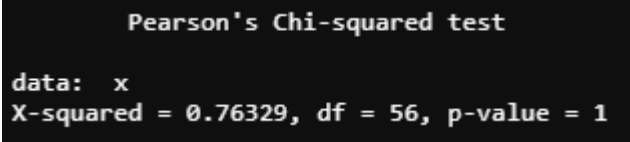
La variable **Nb\_actes** prend des valeurs comprises entre 0 et 82, nous avons décidé de couper cette variable sous forme des classes suivantes :

[0,10[, [10,20[, [20,30[, [30,40[, [40,50[, [50,60[, [60,70[, [70,80[et [80,90[.

La variable **CM** prend des valeurs comprises entre 0 et 1512. Nous avons décidé de couper cette variable sous forme des classes suivantes :

[0,200[, [200,400[, [400,600[, [600,800[, [800,1000[, [1000,1200[, [1200,1400[et [1400,1600[.

Nous avons comme résultat :



```

Pearson's Chi-squared test
data: x
X-squared = 0.76329, df = 56, p-value = 1
    
```

▲ Figure 53 : Résultat du test de Khi-deux

Le test de Khi-deux rejette hypothèses  $H_0$  : Indépendance. Cependant, le test de Khi-deux traite mal les grands effectifs. Pour cela, nous avons décidé d'appliquer le test de V-Cramer pour vérifier l'indépendance entre les deux variables. Ce test est défini par :

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, p-1)}}$$

Nous obtenons un test de V-Cramer égal à **0,19%** qui est très proche de 0, ce qui confirme l'absence d'association entre la variable **Nb\_actes** et **CM**.

### C. Intégration du PPAP

Les sinistres survenus à l'année N, réglés dans les années postérieures doivent impacter le compte de résultat de l'année N. La réglementation impose l'intégration de cette charge sous forme de provision PPAP (provision pour prestation à payer). Ce taux s'écrit :

$$\% PPAP = \frac{\text{Charge totale} - \text{Prestations payées}}{\text{Prestations payées}}$$

D'après les études que nous avons fait sur cette base de données au 31/12/2020, nous avons les taux suivants :

- PPAP pour l'année 2018 = 0% (les sinistres sont tous payés°)
- PPAP pour l'année 2019 = 0,3%
- PPAP pour l'année 2020 = 9,87%

Pour l'implémentation de ces taux, nous partons de la formule précédente. Nous avons :

$$\% PPAP = \frac{Charge\ totale - Prestations\ payées}{Prestations\ payées}$$

Implique que :

$$Charge\ totale = (1 + \% PPAP) \cdot Prestations\ payées$$

D'après l'hypothèse de l'indépendance, nous avons :

$$Charge\ totale = (1 + \% PPAP) \cdot \sum \mathbb{E}(fréquence) \cdot \mathbb{E}(Coût\ moyen)$$

Alors nous pouvons écrire :

$$Charge\ totale = \sum \mathbb{E}[(1 + \% PPAP) \cdot Fréquence] \cdot \mathbb{E}[Coût\ Moyen]$$

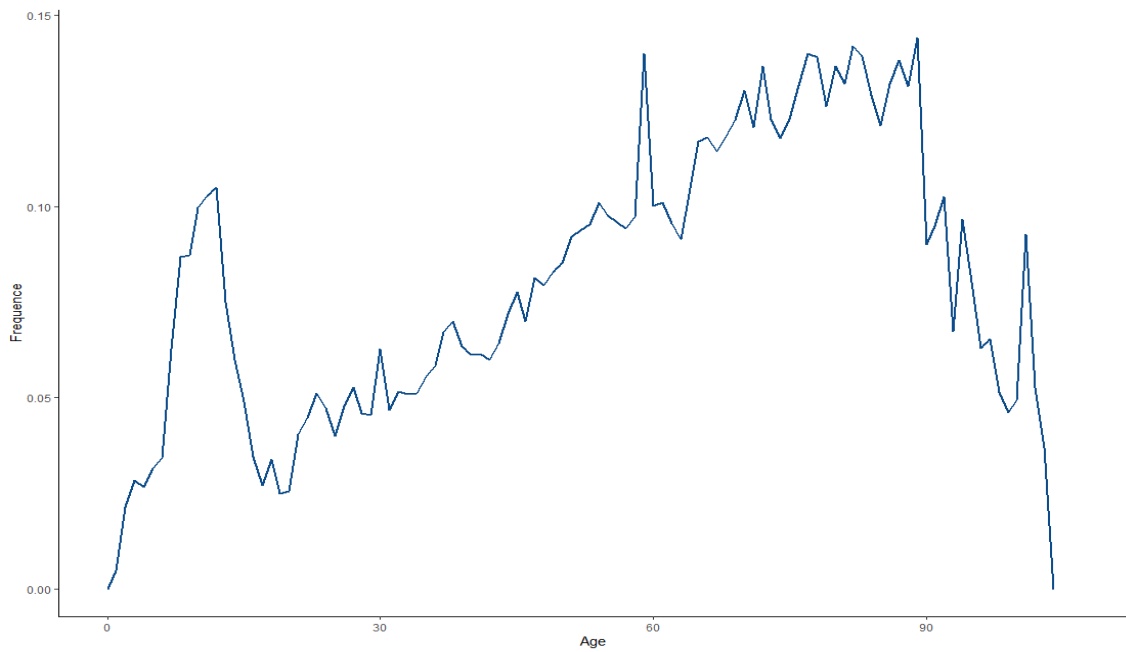
Dans la sous-section qui suit, les taux PPAP sont déjà intégré, en fonction de l'année, dans le calcul du nombre d'actes.

Dans la suite nous allons modéliser le coût moyen et la fréquence pour les sinistres qui sont souvenus en 2018 et 2019 afin de simuler des tarifs qui ne sont pas impactés par l'effet de la crise sanitaire Covid-19, puis à la fin de la partie **F**, nous allons intégrer l'année 2020 également pour mesurer l'impact entre les deux tarifs.

## D. Modélisation de la fréquence

Nous avons tout d'abord étudié la répartition des effectifs par âge. La base à utiliser pour modéliser le libellé brochure 'Chirurgie -NON OPTAM' contient une exposition totale de **203 071**.

La répartition du nombre moyen d'actes par âge est illustrée à travers la figure ci-dessous :

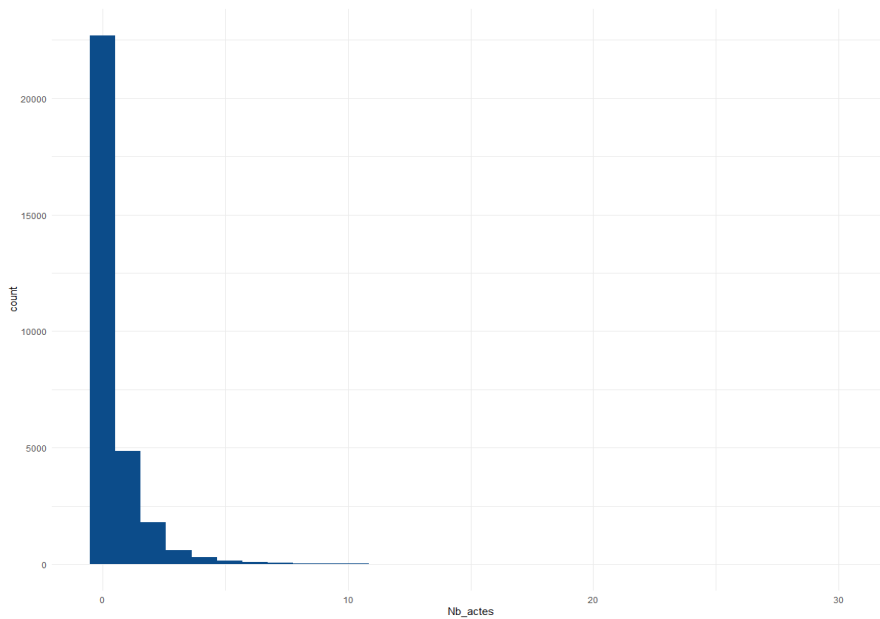


▲ Figure 54 : Répartition du nombre moyen d'actes (fréquence) en fonction de l'âge

Pour modéliser la fréquence, nous avons coupé la base en deux :

- 60% de la base sera utilisée dans la modélisation comme base d'apprentissage,
- 40% sera utilisée comme base de validation du modèle trouvé.

La densité de la variable 'Nb\_actes' est de la forme suivante :



▲ Figure 55 : Densité de la variable 'Nombre d'actes'

Sur la base, nous constatons que 26,5% seulement des adhérents ont reçu une prestation de la mutuelle pour ce libellé brochure.

La loi qui modélise la fréquence doit donc être une loi discrète avec un poids très important sur 0 et une densité décroissante sur le reste des possibilités. Cette loi appartient sûrement à la classe des lois ‘zero-inflated’ ou à la classe des lois ‘zero-ajusted’.

Nous avons utilisé la fonction **FitDist** de la bibliothèque « **gamlss** » fondée sur l’algorithme : **Espérance-Maximisation** (abrégé : **EM**). Cette fonction permet de tester toutes les lois discrètes possibles à modéliser par la bibliothèque ‘gamlss’ (voir annexe 4).

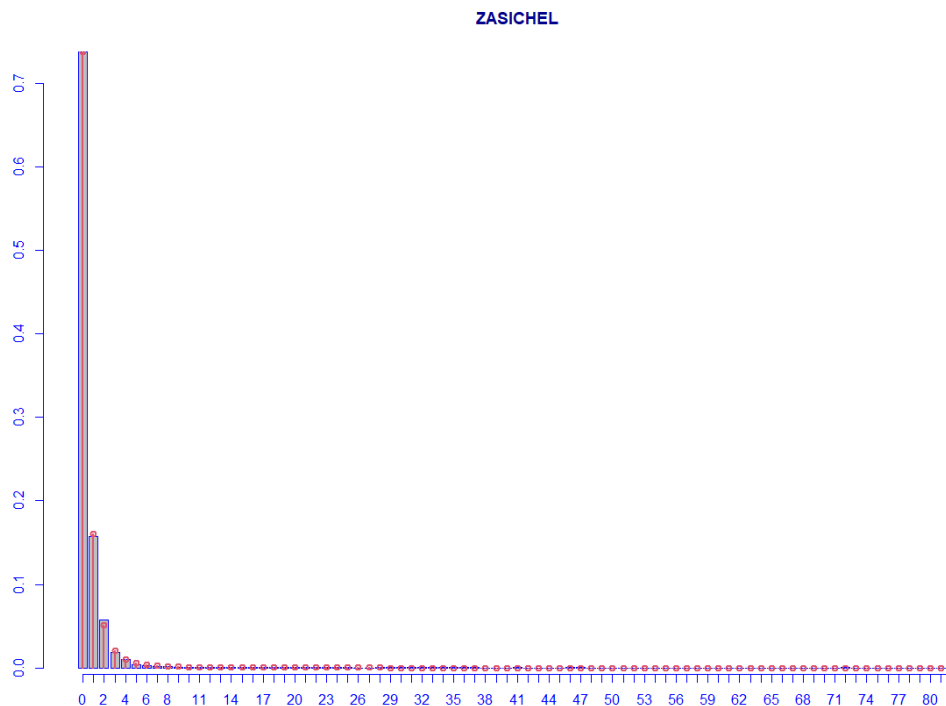
Nous obtenons une table des lois listées par ordre décroissante selon la pertinence du critère AIC de chacune de ces lois :

Loi	AIC
ZASICHEL	32000,9319
ZISICHEL	32000,9319
BNB	32029,9798
...	...
YULE	32772,8943
PO	33566,8534

▲ Figure 56 : récapitulatif des résultats des lois discrètes

La loi **ZASICHEL** (Zero Ajusted Sichel) est la meilleure loi représentative des données de la base d’apprentissage avec un critère AIC égal à **32 000,9319** comme nous montre la figure ci-dessus. La loi **PO** (Poisson) est la loi discrète la moins représentative des données selon le **critère AIC** qui est égal à **33 566,8534**.

La fonction `histDist` de la bibliothèque `gamlss` nous permet de comparer la densité réelle (à partir des données) qui est représentée par des barres avec la densité optimale de la loi choisie.



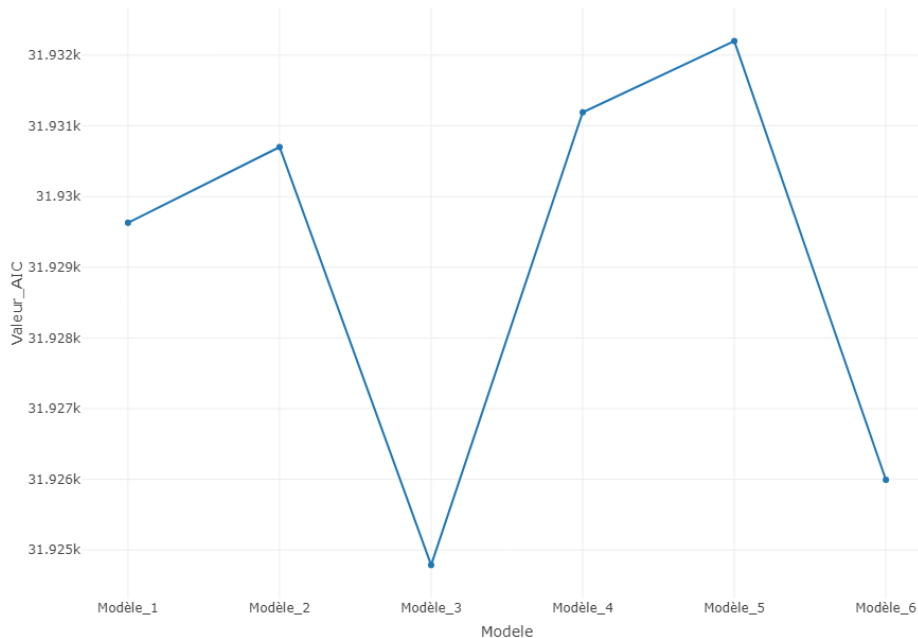
▲ Figure 57 : Comparaison de la densité réelle avec la densité de la loi ZISICHEL

Ensuite, nous avons cherché la combinaison linéaire optimale des variables explicatives qui expliquent la fréquence. Nous avons essayé le modèle qui contient l'âge en polynôme de degré 3, le niveau de garantie et l'exposition. Cette dernière est intégrée en `offset` de la commande R car elle n'est pas considérée comme une variable explicative.

Avec la commande `gamlss`, nous avons modélisé 6 modèles différents (avec la même distribution optimale trouvée), ces modèles sont les suivants :

$$\begin{aligned}
 \mathbf{Nb}_{actes} &= \mathit{Age} + \mathit{garantie} + \mathit{offset}(\log(\mathit{Exposition})) \\
 \mathbf{Nb}_{actes} &= \mathit{Age} + \mathit{Age}^2 + \mathit{garantie} + \mathit{offset}(\log(\mathit{Exposition})) \\
 \mathbf{Nb}_{actes} &= \mathit{Age} + \mathit{Age}^2 + \mathit{Age}^3 + \mathit{garantie} + \mathit{Age} * \mathit{garantie} + \mathit{offset}(\log(\mathit{Exposition})) \\
 \mathbf{Nb}_{actes} &= \mathit{Age} + \mathit{Age}^2 + \mathit{Age}^3 + \mathit{garantie} + \mathit{offset}(\log(\mathit{Exposition})) \\
 \mathbf{Nb}_{actes} &= \mathit{Age} + \mathit{garantie} + \mathit{Age} * \mathit{garantie} + \mathit{offset}(\log(\mathit{Exposition})) \\
 \mathbf{Nb}_{actes} &= \mathit{Age} + \mathit{Age}^2 + \mathit{garantie} + \mathit{Age} * \mathit{garantie} + \mathit{offset}(\log(\mathit{Exposition}))
 \end{aligned}$$

En comparant le critère AIC des six modèles, nous obtenons le graphique suivant qui montre que le **modèle 3** a la valeur minimale avec une valeur AIC = 31,915k. Il s'agit donc du meilleur modèle pour nos données :



▲ Figure 58 : Valeur de l'AIC par Modèle\_FREQ

Nous rappelons que le modèle 6 est celui qui contient comme variables explicatives : **Age**, **Age<sup>2</sup>**, **Age<sup>3</sup>**, **garantie** et **Age \* garantie** pour prendre en considération la corrélation entre l'âge et le niveau de garantie.

Ci-dessous, nous avons un résumé sur le modèle, notamment la loi choisie, la formule du modèle, la fonction lien utilisée par défaut et un récapitulatif des coefficients trouvés pour chaque variable ainsi que le coefficient de significativité des variables.

```

*****
Family: c("ZASICHEL", "zero adjusted Sichel")
Call: gamlss(formula = Nb_actes ~ I(Age) + I(Age^2) + I(Age^3) + garantie + Age * garantie + offset(Exposition), family = names(LoiFreq$fits[indice_loi_FREQ]),
  data = base_appren_freq)
Fitting method: RS()
-----
Mu link function: log
Mu Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.002e+00  4.900e-01  -4.086  4.41e-05 ***
I(Age)       -4.199e-02  2.242e-02  -1.873  0.06104 .
I(Age^2)     1.871e-03  4.798e-04  3.900  9.65e-05 ***
I(Age^3)     -1.295e-05  3.128e-06  -4.140  3.49e-05 ***
garantie     5.424e-03  3.085e-03  1.758  0.07869 .
garantie:Age -1.342e-04  5.137e-05  -2.613  0.00899 **
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '.' 0.1 ' ' 1
-----

```

▲ Figure 59 : Résumé des résultats de la modélisation de la fréquence



Dans ce cas, toutes les variables sont significatives pour ce modèle au seuil de 10%.

Nous rappelons que la fonction lien de la loi ZASICHEL est  $g(x) = \log(x)$ .

Le meilleur modèle pour la fréquence du libellé brochure : ‘Chirurgie-NON OPTAM’ est alors sous la forme :

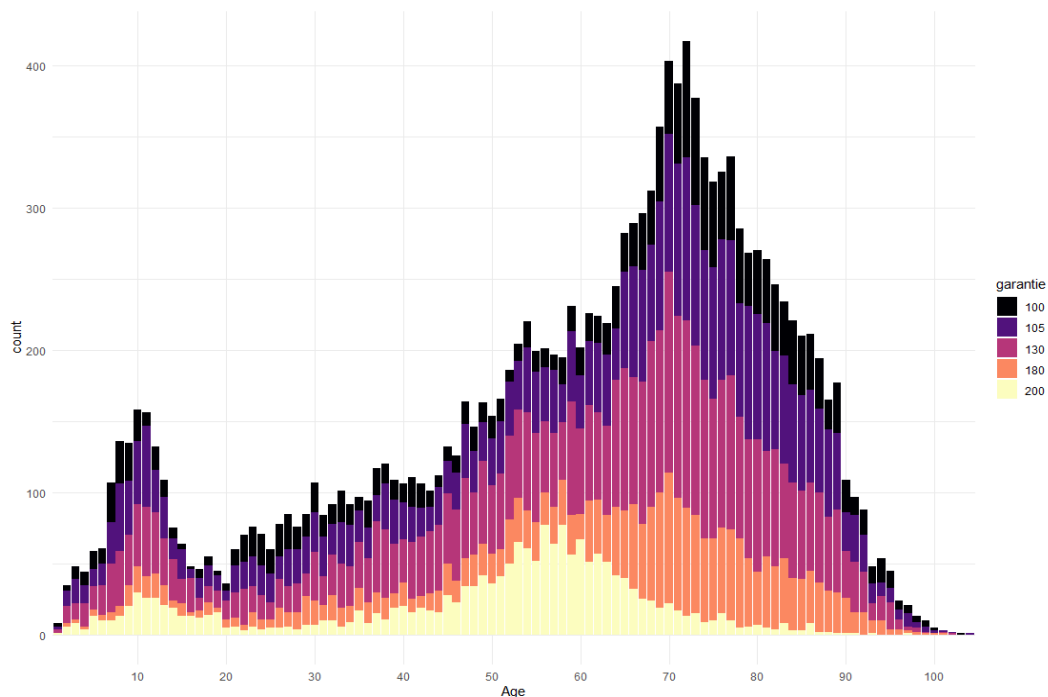
$$\mathbb{E}(\text{Fréquence}) = g^{-1} \left( \sum_i \beta_i \cdot X_i \right) = \exp \left( \sum_i \beta_i \cdot X_i \right)$$

$$\mathbb{E}(\text{Fréquence}) = \exp \left( -2,002 - 4,199e - 02 * \text{Age} + 1,871e - 03 * \text{Age}^2 - 1,295e - 05 * \text{Age}^3 + 5,424e - 03 * \text{garantie} - 1,342e - 04 * \text{Age} * \text{garantie} \right)$$

### E. Modélisation du Coût Moyen

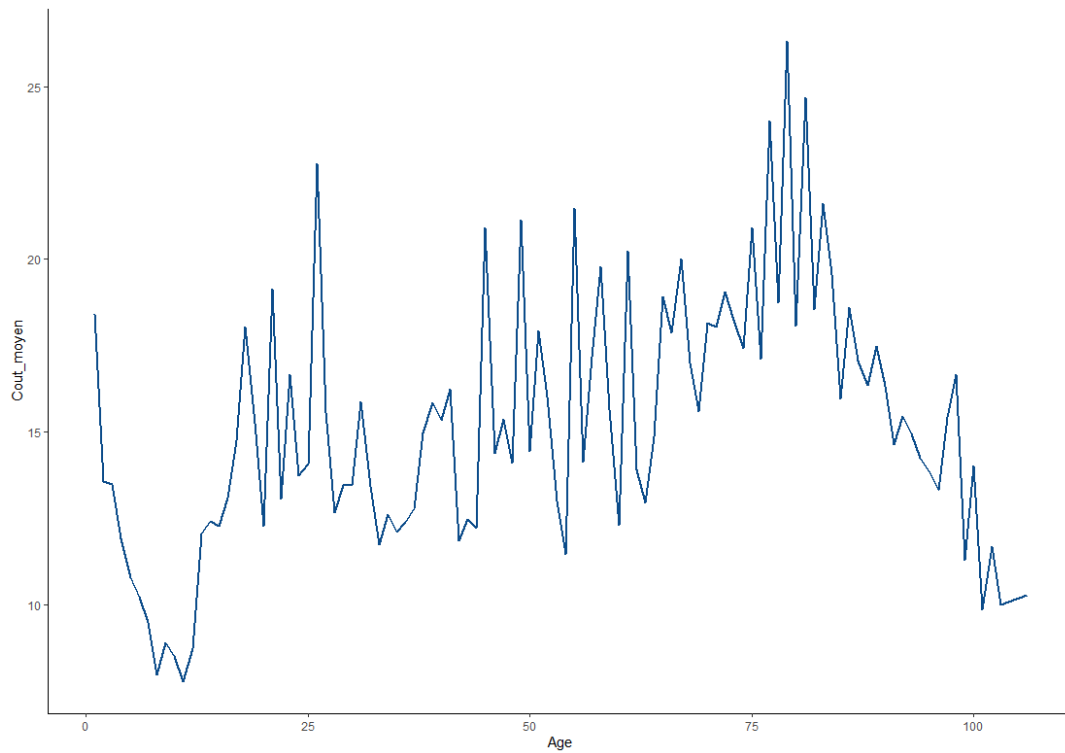
La modélisation du coût moyen nécessite la suppression de toutes les lignes avec un  $CM = 0$ . Pour cela, nous avons pris la même base de départ, puis nous avons filtré les lignes avec  $CM > 0$ . Cette base contient 15 267 lignes.

Ces lignes des bénéficiaires des prestations sont distribuées sous la forme :



▲ Figure 60 : Nombre de bénéficiaires en fonction de l'âge et du niveau de garantie

Le coût moyen est réparti en fonction de l'âge selon la figure suivante :

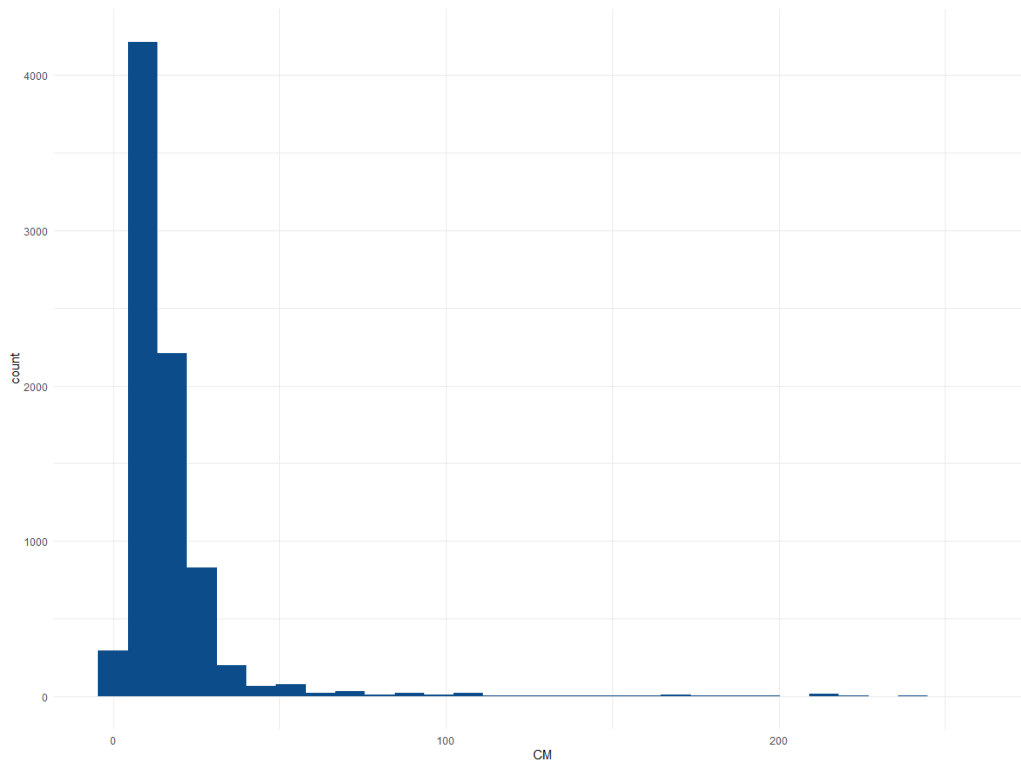


▲ Figure 61 : Coût moyen en fonction de l'âge

Les assurés qui ont un âge entre 75 et 80 ans sont les assurés dont leur coût moyen par acte est le plus élevé par rapport aux autres âges.

Pour modéliser le coût moyen, nous avons coupé aléatoirement la base en deux (base d'apprentissage à 60% et base de validation du modèle à 40%).

La densité du coût moyen est de la forme suivante :



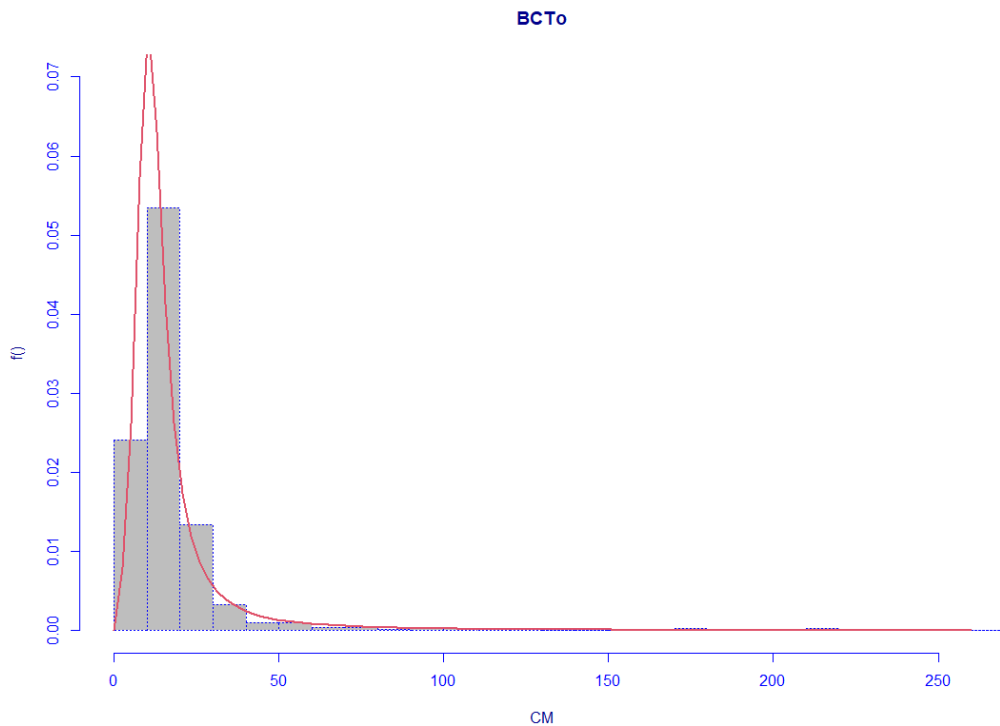
▲ Figure 62 : Densité de la variable\_CM

La loi qui modélise le coût moyen doit être une loi continue positive avec un poids très important sur les valeurs qui sont proches de 0. En utilisant la fonction `FitDist`, nous obtenons les résultats suivants :

Loi	AIC
BCTo	40786,8992
GB2	40797,2159
BCT	40798,3724
...	...
WEI3	44047,0930
EXP	44075,7596

▲ Figure 63 : récapitulatif des résultats des lois continues

La loi **BCT-o** (Box-Cox t.-orig.) est la meilleure distribution représentative du coût moyen dans la base d'apprentissage comme nous montre la figure ci-dessous :



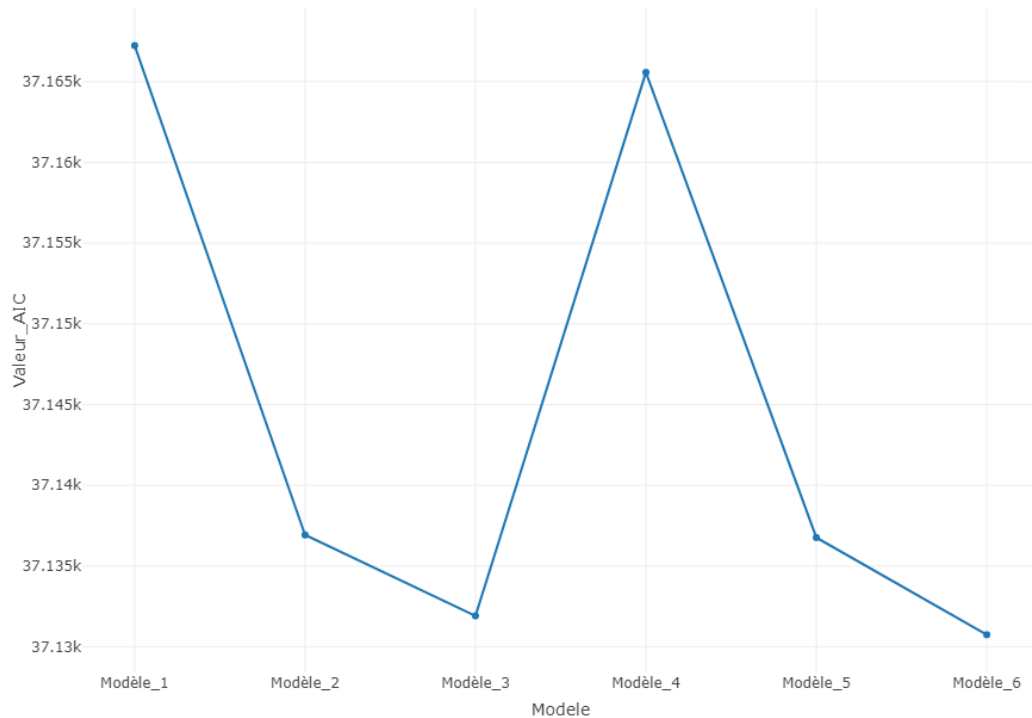
▲ Figure 64 : Comparaison de la densité réelle avec la densité de la loi BCTo

En comparant la densité en rouge et la densité réelle de la figure 51 nous constatons que les deux sont fortement croissantes à partir de 0 jusqu'à ~ 20€ puis diminuent fortement jusqu'à 40€. A partir de 40€, les deux fonctions sont faiblement décroissantes. Cette comparaison nous confirme le choix de cette distribution comme loi optimale en représentation des données.

Ensuite, nous avons cherché la combinaison linéaire optimale des variables explicatives qui expliquent le coût moyen. Comme dans le modèle de la fréquence, nous avons ajouté l'âge comme variable explicative en polynôme de degré 3 ainsi que le niveau de garantie et la corrélation entre les deux. Les modèles que nous testés sont :

$$\begin{aligned}
 & \mathbf{CM} = \mathit{Age} + \mathit{garantie} \\
 & \mathbf{CM} = \mathit{Age} + \mathit{Age}^2 + \mathit{garantie} \\
 & \mathbf{CM} = \mathit{Age} + \mathit{Age}^2 + \mathit{Age}^3 + \mathit{garantie} \\
 & \mathbf{CM} = \mathit{Age} + \mathit{garantie} + \mathit{Age} * \mathit{garantie} \\
 & \mathbf{CM} = \mathit{Age} + \mathit{Age}^2 + \mathit{garantie} + \mathit{Age} * \mathit{garantie} \\
 & \mathbf{CM} = \mathit{Age} + \mathit{Age}^2 + \mathit{Age}^3 + \mathit{garantie} + \mathit{Age} * \mathit{garantie}
 \end{aligned}$$

En comparant le critère AIC des six modèles, nous obtenons le graphique suivant qui montre que le **modèle 6** a la valeur minimale avec une valeur AIC = **37,131k**. Il s'agit donc du meilleur modèle pour nos données :



▲ Figure 65 : Valeur de l'AIC par Modèle\_CM

Nous rappelons que le modèle 6 est celui qui contient comme variables explicatives : **Age**, **Age<sup>2</sup>**, **garantie** et la variable **Age \* garantie** pour prendre en considération la corrélation entre l'âge et le niveau de garantie.

Ci-dessous, nous avons un résumé des résultats sur le modèle, notamment la loi choisie, la formule du modèle, la fonction lien utilisée par défaut et un récapitulatif des coefficients trouvés pour chaque variable ainsi que le coefficient de significativité des variables.

```

*****
Family: c("BCto", "Box-Cox-t-orig.")
Call:  gamlss(formula = CM ~ I(Age) + I(Age^2) + I(Age^3) + garantie + Age * garantie, family = names(LoiCM$fits[indice_loi_CM]), data = base_appren_CM)
Fitting method: RS()

-----
Mu link function: log
Mu Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.267e+00  7.809e-02  29.048 < 2e-16 ***
I(Age)       1.135e-02  3.311e-03   3.233  0.00123 **
I(Age^2)     -1.667e-04  7.639e-05  -2.183  0.02912 *
I(Age^3)     6.935e-07  5.063e-07   1.370  0.17084
garantie     -6.990e-04  5.017e-04  -1.393  0.16360
garantie:Age  2.735e-05  8.728e-06   3.134  0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
-----

```

▲ Figure 66 : Résumé des résultats de la modélisation du coût moyen

La variable garantie n'est pas significative, en revanche, le produit entre Age et garantie est significatif au seuil de 5%. En conséquence, et pour nous permettre de construire un outil de tarification, nous avons tout de même conservé la variable garantie dans notre modèle final.

La fonction lien de la moyenne associée à la loi 'Box-Cox t-orig.' est 'log'. Le meilleur modèle pour la moyenne de la fréquence du libellé brochure : 'Chirurgie-NON OPTAM' est alors sous la forme :

$$\mathbb{E}(\text{Coût moyen}) = g^{-1} \left( \sum_i \beta_i \cdot X_i \right) = \exp \left( \sum_i \beta_i \cdot X_i \right)$$

$$\mathbb{E}(\text{Coût moyen}) = \exp (2,26 - 1,135e - 02 * Age - 1,667e - 04 * Age^2 - 6,935e - 07 * Age^3 - 6,990e - 04 * garantie + 2,735e - 05 * Age * garantie)$$

## F. Modèle final

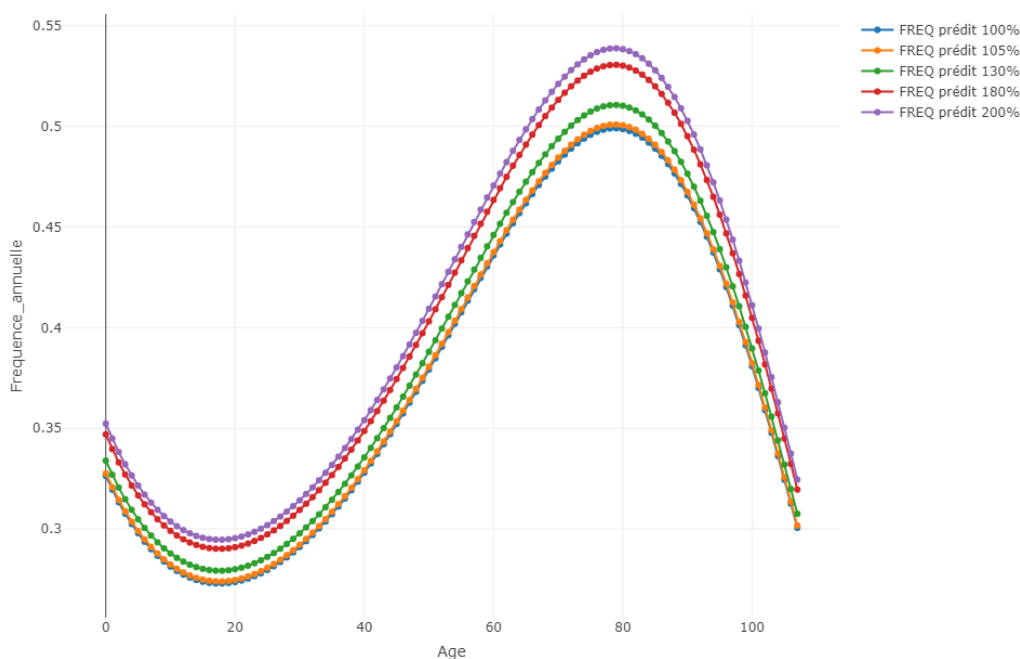
Le modèle final qui nous permet de calculer le tarif annuel du libellé brochure est le produit des deux modèles trouvés :

$$\text{Tarif}_{\text{Chirurgie-NON OPTAM}} = \mathbb{E}(\text{Coût moyen}) \cdot \mathbb{E}(\text{Fréquence})$$

La charge totale de l'échantillon de validation est de 128 400,4 € et la somme des tarifs prédite pour cette base (de validation) est de 131 057 €, soit une erreur de :

$$\frac{\text{Prédiction}}{\text{Réel}} - 1 = \mathbf{2,083\%}$$

La figure suivante illustre les fréquences en fonction de l'âge :

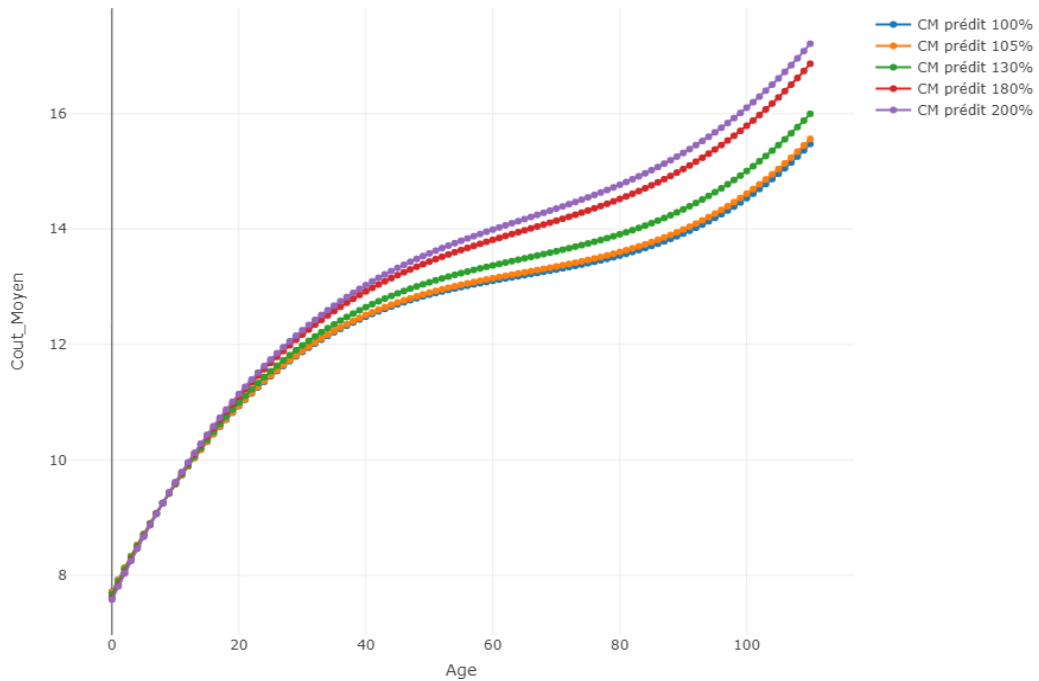


▲ Figure 67 : Fréquence annuelle en fonction de l'âge et du niveau de remboursement (Hors Covid-19)

Le graphique ci-dessus montre que les courbes sont cohérentes entre elles avec des fréquences plus faibles pour des assurés qui ont optés pour de plus faibles garanties.

De plus, après un recours important lors des premières années de vie, ce libellé brochure est moins utilisé pendant l'enfance. La fréquence se remet à augmenter vers la majorité jusqu'à 80 ans où nous notons ensuite une baisse du recours à ce libellé brochure.

Pour le coût moyen nous obtenons le graphique suivant :

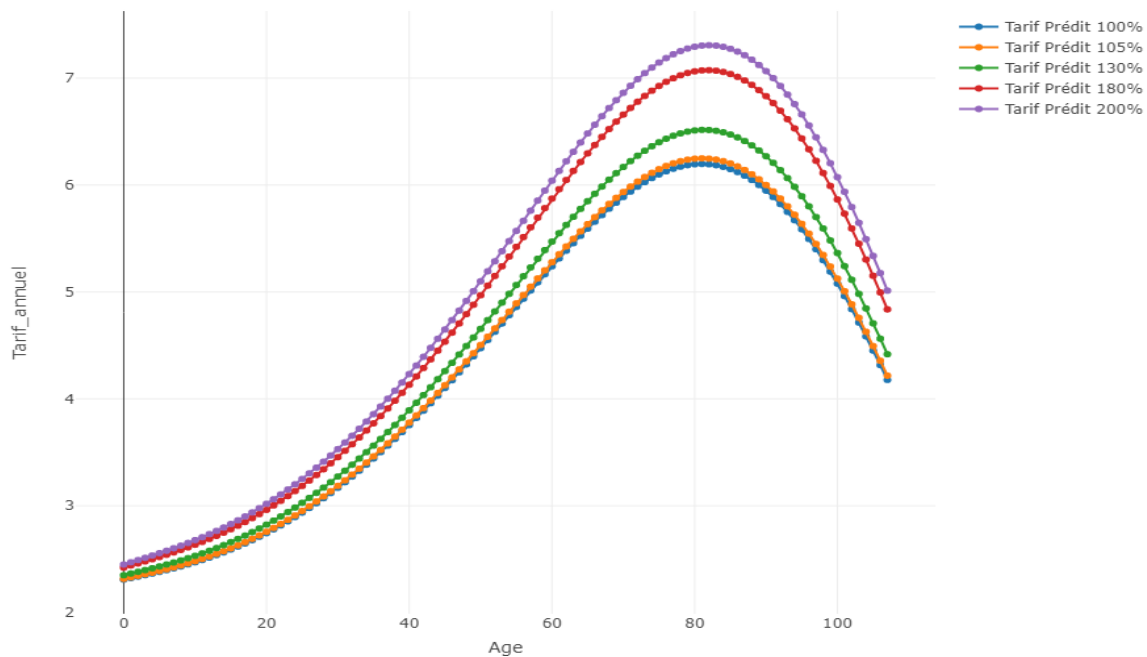


▲ Figure 68 : Coût moyen annuel en fonction de l'âge et du niveau de remboursement (Hors Covid-19)

Le coût moyen annuel est croissant en fonction de l'âge. En revanche il est stable en fonction du niveau de garantie pour les premiers âges jusqu'à 20 ans, puis croissant.

Le produit du coût moyen et de la fréquence nous donne les tarifs ci-dessous :





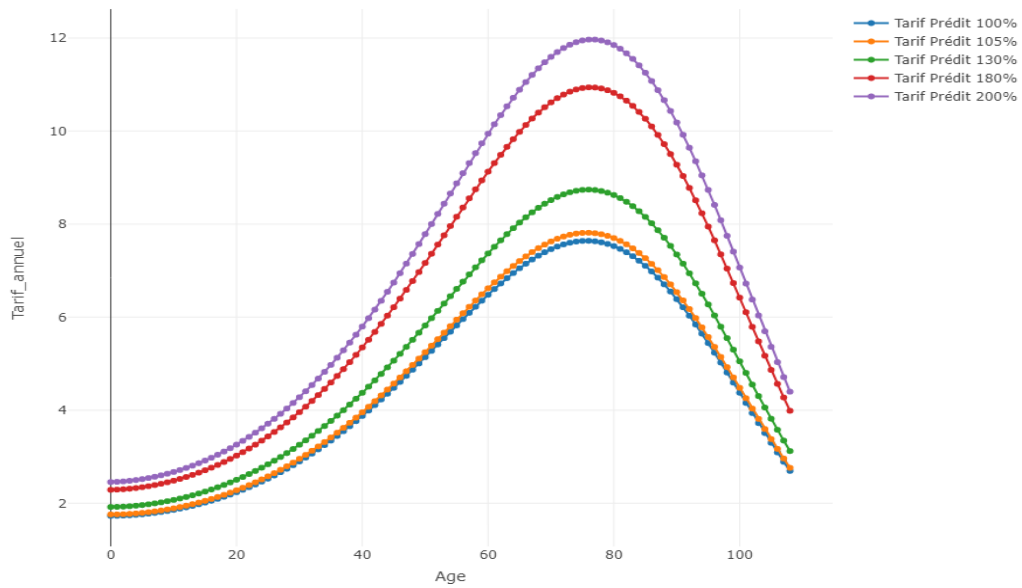
▲ Figure 69 : Tarif annuel en fonction de l'âge et du niveau de garantie (Hors Covid-19)

Les tarifs annuels sont strictement croissants en fonction de l'âge jusqu'à 80 ans, puis strictement décroissants. Cette monotonie locale par intervalle d'âge est normale car elle prend la forme de la fréquence annuelle. L'impact du coût moyen alors est d'agrandir l'écart entre les tarifs en fonction du niveau de garantie pour les âges qui sont autour de 80 ans.

### G. Impact de la crise sanitaire sur les tarifs

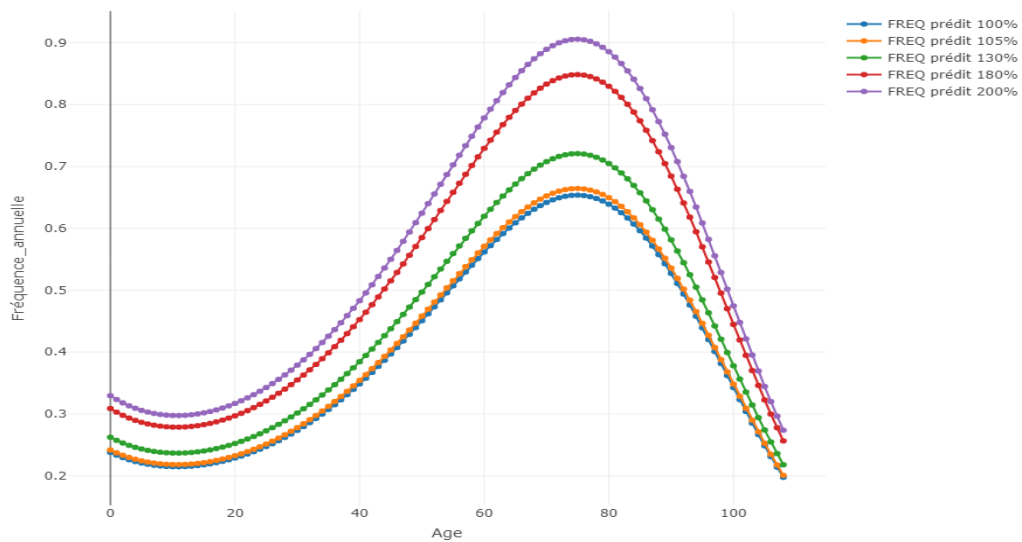
Dans cette section nous avons fait la même modélisation en intégrant l'année 2020 pour voir l'impact de la crise sanitaire qui a commencé en mars 2020.

La modélisation nous donne les tarifs annuels suivants :



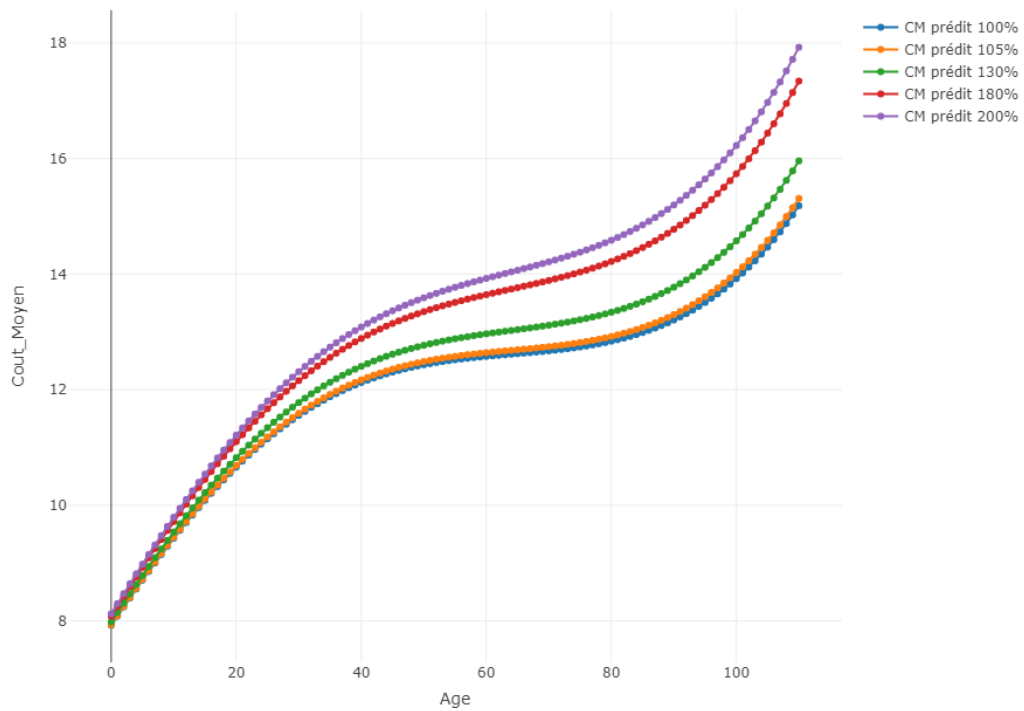
▲ Figure 70 : Tarif annuel en fonction de l'âge et du niveau de garantie (Covid-19)

En intégrant l'année 2020, nous constatons que les tarifs augmentent de 70% en moyenne. La fréquence prend toujours la forme de monotonie locale, mais augmente de 66% en moyenne comme nous montre la figure suivante :



▲ Figure 71 : Fréquence annuelle en fonction de l'âge et du niveau de remboursement (Covid-19)

L'écart entre les nouveaux tarifs (Figure 57) est quasiment le même pour les âges entre 0 et 60 ans en Covid et Hors Covid (Figures 56 et 57), ceci est expliqué par le fait que les coûts moyens par rapport aux deux approches (Covid : Figure 59 et Hors Covid : Figure 55) n'ont pas trop évolués au global pour cette tranche d'âge. A partir de 60 ans, plus nous avons écart au niveau des coûts moyens, plus nous constatons un écart entre les tarifs trouvés.



▲ Figure 72 : Coût moyen annuel en fonction de l'âge et du niveau de remboursement (Covid-19)

## H. Reste des libellés actes

La variable 'garantie' est construite à partir de la table de correspondance des grilles de garanties. Elle prend des valeurs de formes différentes selon le type du libellé brochure. Par exemple pour les postes :

- Actes d'anesthésie, actes d'échographie et consultations de généralistes ou spécialistes : c'est % de la BRySS, de la BR, du TM ou des FR.
- Chambre particulière de maternité, médicale et chirurgicale ou de psychiatrie : en €, en %PMSS, en €/jour ou en %PMSS/jour.

Pour chaque libellé brochure, nous avons cherché un type de remboursement unique que nous avons par la suite modéliser à travers la variable garantie. Par exemple, pour la chambre particulière, nous allons mettre toutes les valeurs de la variable garantie en €.

## I. Intégration dans l'outil de tarification

La tarification santé au sein du cabinet se fait sur Excel. Pour mettre en place les modèles GLM trouvés, nous avons utilisé deux feuilles : une feuille de saisie et une feuille qui contient les modèles ainsi que les calculs intermédiaires pour avoir un tarif brut.

Sur l'outil de tarification sur Excel nous faisons plusieurs tarifications sur le même fichier, pour cela nous sommes obligés d'optimiser tous les calculs, tout en gardant une procédure de calcul simple et lisible. Pour cela, nous avons résumé le calcul de ces deux formules :

$$\mathbb{E}(\text{Coût moyen}) = \exp(2,26 - 1,135e - 02 * \text{Age} - 1,667e - 04 * \text{Age}^2 - 6,935e - 07 * \text{Age}^3 - 6,990e - 04 * \text{garantie} + 2,735e - 05 * \text{Age} * \text{garantie})$$

$$\mathbb{E}(\text{Fréquence}) = \exp(-2,002 - 4,199e - 02 * \text{Age} + 1,871e - 03 * \text{Age}^2 - 1,295e - 05 * \text{Age}^3 + 5,424e - 03 * \text{garantie} - 1,342e - 04 * \text{Age} * \text{garantie})$$

En effet, pour un âge X et un niveau de garantie Y, nous avons :

$$\text{Tarif}_{X,Y} = \mathbb{E}(\text{Coût moyen}_{X,Y}) \cdot \mathbb{E}(\text{Fréquence}_{X,Y})$$

$$\text{Tarif}_{X,Y} = \exp(2,26 - 1,135e - 02 * \text{Age} - 1,667e - 04 * \text{Age}^2 - 6,935e - 07 * \text{Age}^3 - 6,990e - 04 * \text{garantie} + 2,735e - 05 * \text{Age} * \text{garantie}) \cdot \exp(-2,002 - 4,199e - 02 * \text{Age} + 1,871e - 03 * \text{Age}^2 - 1,295e - 05 * \text{Age}^3 + 5,424e - 03 * \text{garantie} - 1,342e - 04 * \text{Age} * \text{garantie})$$

$$\begin{aligned} \text{Tarif}_{X,Y} = & \exp(-6,990e-04 * \text{garantie}) \cdot \exp(2,735e-05 * \text{Age} * \text{garantie}) \\ & \cdot \exp(2,26 - 1,135e-02 * \text{Age} - 1,667e-04 * \text{Age}^2 - 6,935e-07 \\ & * \text{Age}^3) \cdot \exp(5,424e-03 * \text{garantie}) \cdot \exp(-1,342e-04 * \text{Age} \\ & * \text{garantie}) \cdot \exp(-2,002 - 4,199e-02 * \text{Age} + 1,871e-03 * \text{Age}^2 \\ & - 1,295e-05 * \text{Age}^3) \end{aligned}$$

$$\begin{aligned} \text{Tarif}_{X,Y} = & \exp [(-6,990e-04 + 2,735e-05 * \text{Age}) * \text{garantie}] \\ & \cdot \exp[2,26 - 1,135e-02 * \text{Age} - 1,667e-04 * \text{Age}^2 - 6,935e-07 \\ & * \text{Age}^3] \cdot \exp [(5,424e-03 - 1,342e-04 * \text{Age}) * \text{garantie}] \\ & \cdot \exp [-2,002 - 4,199e-02 * \text{Age} + 1,871e-03 * \text{Age}^2 - 1,295e-05 \\ & * \text{Age}^3] \end{aligned}$$

$$\text{Tarif}_{X,Y} = f_{\text{CM}}(\text{garantie}) \cdot A \cdot f_{\text{FREQ}}(\text{garantie}) \cdot B$$

$$\text{Tarif}_{X,Y} = f_{\text{CM}}(\text{garantie}) \cdot \mathbb{E}(\text{Coût moyen}_{X,0}) \cdot f_{\text{FREQ}}(\text{garantie}) \cdot \mathbb{E}(\text{Fréquence}_{X,0})$$

$$\text{Tarif}_{X,Y} = f_{\text{CM}}(\text{garantie}) \cdot f_{\text{FREQ}}(\text{garantie}) \cdot \text{Tarif}_{X,0}$$

Les deux fonctions  $f_{\text{CM}}(\text{garantie})$  et  $f_{\text{FREQ}}(\text{garantie})$  ne dépendent que du niveau de remboursement et seront donc intégrées facilement sur Excel. Cette fonction simple va récupérer le niveau de remboursement ajouté sur la feuille 'Saisie' :

	Max de jours ou séances		TR mutuelle en % BR		Remboursement forfaitaire		Tarif brut (30 ans)
	Nombre max	Assiette	%	Assiette	Montant ou %	Assiette	
<b>HOSPITALISATION</b>							
Actes d'anesthésie - OPTAM			250	BRySS			0,05 €
Actes d'anesthésie - NON OPTAM			230	BRySS			0,15 €
Actes de chirurgie - OPTAM			250	BRySS			0,05 €
Actes de chirurgie - NON OPTAM			130	BRySS			0,06 €
Actes d'échographie - OPTAM			250	BRySS			0,37 €
Actes d'échographie - NON OPTAM			230	BRySS			0,55 €
Actes d'imagerie - OPTAM			250	BRySS			0,96 €
Actes d'imagerie - NON OPTAM			230	BRySS			1,18 €
Actes d'obstétrique - OPTAM			250	BRySS			0,00 €
Actes d'obstétrique - NON OPTAM			230	BRySS			0,00 €
Forfait hospitalier							0,00 €
Frais de séjour médical et chirurgical							0,00 €

▲ Figure 73 : Feuille de 'saisie' sur Excel

La fonction  $\text{Tarif}_{X,0}$  est la prédiction du tarif pour un âge X et pour un niveau de garantie 0 (valeur par défaut pour faciliter les calculs).

Le calcul dans la page précédente est valable tant que nous avons la fonction lien log (ce qui donne la fonction exponentielle dans la prédiction). Afin de généraliser le calcul, nous avons gardé l'information de la loi du libellé brochure, ainsi que de sa fonction lien-:

	Informations sur GLM utilisé				Informations SS (estimées)			Calcul du Niveau de garantie						
	Coût Moyen		Fréquence		FR Moyen	BR Moyenne	% de SS	Max de jours ou séances		TR mutuelle en % BR		Remboursement forfaitaire		variable garantie
	Loi	Lien	Loi	Lien				Nombre max	Assiette	%	Assiette	Montant ou %	Assiette	
HOSPITALISATION														
• Actes d'anesthésie - OPTAM	IG	log	NBII	log	225,30 €	147,63 €	80%	0	0	250	BRySS	0	0	250
• Actes d'anesthésie - NON OPTAM	BCPEo	log	NBII	log	204,38 €	136,38 €	80%	0	0	230	BRySS	0	0	230
• Actes de chirurgie - OPTAM	GIG	identité	ZASICHEL	log	277,75 €	176,48 €	80%	0	0	250	BRySS	0	0	250
• Actes de chirurgie - NON OPTAM	BCTo	log	ZISICHEL	log	77,42 €	51,98 €	80%	0	0	130	BRySS	0	0	130
• Actes d'échographie - OPTAM	BCTo	log	ZASICHEL	log	56,83 €	46,03 €	80%	0	0	250	BRySS	0	0	250
• Actes d'échographie - NON OPTAM	BCT	identité	ZASICHEL	log	55,85 €	48,08 €	80%	0	0	230	BRySS	0	0	230
• Actes d'imagerie - OPTAM	GB2	log	ZASICHEL	log	32,29 €	29,32 €	80%	0	0	250	BRySS	0	0	250
• Actes d'imagerie - NON OPTAM	BCTo	log	ZIPIG	log	29,07 €	25,68 €	80%	0	0	230	BRySS	0	0	230

▲ Figure 74 : Feuille de 'GLM' : Paramétrage

Voici un exemple d'alimentation des modèles GLM dans l'outil de tarification :

Niveau garantie	Type tarif	Coeff Niveau Garantie		TARIF BRUT			COUT MOYEN			FREQUENCE		
		CM	FREQ	1	2	3	1	2	3	1	2	3
250	Taux (%)	0,0112	0,0088	0,01 €	0,01 €	0,02 €	9,77 €	9,74 €	9,71 €	0,000	0,000	0,000
230	Taux (%)	0,0224	0,0118	0,13 €	0,13 €	0,12 €	1,00 €	1,00 €	0,99 €	0,001	0,001	0,001
250	Taux (%)	0,0054	0,0012	0,06 €	0,06 €	0,06 €	23,25 €	23,46 €	23,68 €	0,021	0,020	0,020
130	Taux (%)	0,0001	0,0033	0,05 €	0,05 €	0,05 €	7,93 €	8,01 €	8,08 €	0,045	0,044	0,043
250	Euro (€)	0,0010	-0,0009	0,12 €	0,13 €	0,14 €	19,78 €	19,40 €	19,04 €	0,071	0,079	0,087
230	Euro (€)	2,2600	-0,0014	31,13 €	30,05 €	29,05 €	9,81 €	9,89 €	9,97 €	0,966	0,933	0,902
250	Taux (%)	0,0002	0,0005	0,48 €	0,49 €	0,50 €	7,43 €	7,47 €	7,51 €	0,643	0,657	0,672
230	Taux (%)	0,0001	0,0000	0,35 €	0,37 €	0,40 €	6,98 €	6,98 €	6,97 €	0,579	0,619	0,661

▲ Figure 75 : Feuille de 'GLM' : Modèles GLM

Les fonctions  $\mathbb{E}(\text{Coût moyen}_{X,0})$  et  $\mathbb{E}(\text{Fréquence}_{X,0})$  trouvées par R seront collées en dur sur les colonnes du coût moyen en vert et sur les colonnes de la fréquence en rouge.

Les fonctions  $f_{\text{CM}}(\text{garantie})$  et  $f_{\text{FREQ}}(\text{garantie})$  seront calculées directement sur les colonnes de tarif brut en bleu en utilisant les coefficients de niveau de garantie qui sont sur les colonnes Y et Z et les coefficients collés du coût moyen et de la fréquence à travers la formule :

$$\text{Tarif}_{X,Y} = f_{\text{CM}}(\text{garantie}) \cdot \mathbb{E}(\text{Coût moyen}_{X,0}) \cdot f_{\text{FREQ}}(\text{garantie}) \cdot \mathbb{E}(\text{Fréquence}_{X,0})$$

# Limites

## A. Choix de la base de remboursement

Les codes actes de la mutuelle étudiée ont été regroupés en libellés actes (soins dentaires, actes de chirurgie OPTAM, etc.) correspondant aux libellés affichés dans les tables de garanties transmises par les clients pour réaliser des tarifications.

Ce regroupement nous amène forcément à choisir une base de remboursement ‘représentative’ du libellé acte. Dans cette étude, l’idée a été de calculer la base de remboursement moyenne de tous ces codes actes.

La précision du calcul pourrait être améliorée en modélisant cette base de remboursement moyenne. L’idée de cette amélioration consiste à modéliser la base de remboursement (information disponible dans la base de prestations) en fonction des caractéristiques de l’individu (informations disponibles dans la base des effectifs).

En définitive, la tarification sera plus précisée puisque plus adaptée aux profils de risques.

## B. Prise en compte du plafond

Parmi les limites, la modélisation réalisée ne permet pas d’intégrer le plafond des remboursements pour un poste de soins (ex. : dentaire) ou sur un groupe de libellés actes (ex. : Prothèses dentaires remboursées par la sécurité sociale + Implants dentaires).

En effet, il existe des grilles de garanties dans lesquelles l’organisme complémentaire prend en charge le remboursement sur certains libellés actes à condition de ne pas dépasser un montant X€ par année. Par exemple, nous avons une table de garanties qui propose les remboursements suivants :

- Soins courants : **225% BRySS**.
- Prothèses dentaires hors 100% santé : **300% BRySS**.
- Actes non remboursés par la sécurité sociale (Implantologie Prothèses, orthodontie, parodontologie) : **250% BRySS**.

Avec condition :

- Plafond dentaire annuel (hors soins) incluant les prestations 100% santé<sup>5</sup> : **2 000 €** :

Une piste d’amélioration de l’outil de tarification pourrait être de modéliser ce plafond afin de l’intégrer correctement dans les tarifications.

---

<sup>5</sup> Le panier 100% santé et le ticket modérateur sont toujours pris en charge dans un contrat responsable.

## C. Prise en compte des variables explicatives

Le coût moyen et la fréquence d'un libellé acte sont expliqués par des variables explicatives telles que l'âge, le niveau de remboursement ou encore la qualité de l'ayant droit. Le tarif résultant de la modélisation ne présente qu'une estimation du coût réel avec une marge d'erreur. Cette erreur peut être minimisée par l'optimisation de la méthode mathématique modélisant le libellé acte comme trouver une loi plus adaptée à nos données ou encore changer la combinaison des variables explicatives utilisées.

Cependant, la précision du modèle pourrait être optimisée en améliorant la qualité des données utilisées, en ajoutant d'autres variables qui impactent le coût moyen du libellé actes ou sa fréquence. Par exemple, ajouter la variable sexe, statut de l'adhérent ou encore plus d'informations sur la localisation de l'assuré. Cette dernière variable permet de faire une étude complémentaire sur l'impact de la distance entre l'assuré et l'offre de soins sur la consommation.

## D. Lien entre les actes

Nous savons qu'il existe un lien entre les différents libellés actes. Par exemple, une consultation de médecin généraliste s'accompagne souvent d'une prescription de médicaments. La consommation « pharmacie » est donc liée à la consommation « médecin généraliste ».

La modélisation dans mon mémoire se fait pour chaque libellé acte séparément. Par conséquent, elle ne prend pas en compte de corrélation entre ces actes de soins. L'intégration de la corrélation entre les libellés actes serait une piste d'amélioration des résultats de notre outil de tarification.

## E. Variable : taux de remboursement

La variable « taux de remboursement » a été intégrée comme une variable numérique continue. Il existe des actes de soins pour lesquels cette variable ne prend qu'un nombre très limité de valeur. Par exemple, la variable « frais de transport » ne prend que la valeur '100% BRySS' dans notre base de données. Le modèle final pour ce libellé acte n'intégrera pas cette variable et l'outil final dans ce cas ne permettra pas de tarifier une garantie différente de 100% BRySS. Cette problématique n'est a priori pas trop gênante, le nombre limité de valeur de ces variables étant le reflet de l'uniformité des garanties pour ces actes sur le marché (pour l'exemple proposé ci-avant, la prise en charge du ticket modérateur est la norme).



## F. Niveau de gamme

L'étude réalisée dans (*Partie II-4-D : Zoom sur les gammes étudiées*) montre que la modélisation réalisée est limitée par les tableaux de garanties retenus. En fait, nous n'arriverons pas à modéliser des niveaux hauts de remboursement car plus de 97% des garanties modélisées sont des garanties bas de gamme et milieu de gamme.

L'ajout des nouvelles bases de données avec des garanties haut de gamme sera un point d'amélioration des modèles utilisés dans la tarification.

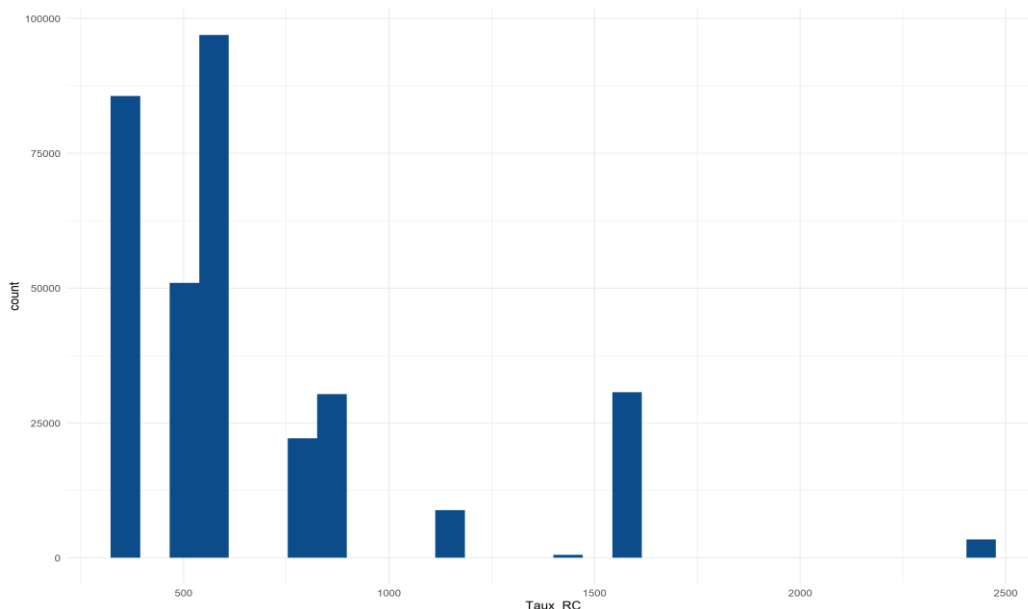
## G. Actes avec forfait

La mutuelle pourrait proposer, pour un libellé acte, un remboursement en % de la base de remboursement de la sécurité sociale, un forfait en € ou les deux combinés, notamment pour certains actes comme les prothèses dentaires.

La méthode de modélisation suivie dans ce mémoire consiste à expliquer le coût moyen et la fréquence en fonction des autres variables explicatives à travers une régression optimale.

Dans le cas où nous avons un manque de données pour un libellé acte ou le cas où nous avons un nombre très limité de montants de remboursement proposés par la mutuelle, comme par exemple, pour toute la population assurée, le montant de forfait proposé par la mutuelle ne prend que 9 valeurs possibles du coût moyen.

Le graphique suivant illustre la densité des montants de forfait proposés par la mutuelle en prothèses dentaires :



▲ Figure 76 : densité des forfaits pour prothèses dentaires remboursées par la sécurité sociale

Dans ce cas il est difficile de modéliser en utilisant une régression. En effet, l'approche GLM consiste à trouver une loi positive continue 'optimale' qui permet de présenter la densité réelle du coût moyen. L'algorithme de régression GLM diverge dans cet exemple car pour un nombre limité d'itérations  $n$ , il n'arrive pas à minimiser l'écart entre cette densité réelle et une des densités classiques listées dans l'annexe.

Pour ce libellé acte, l'ajout d'une nouvelle base de données avec des forfaits diversifiés permettra d'avoir une nouvelle densité réelle plus facile à modéliser par GLM. Sinon, le passage par une méthode de classification permettra probablement de résoudre ce problème.

## Conclusion

La construction de l'outil de tarification santé du cabinet est fondée sur des résultats de modèles linéaires généralisés. Mettre à jour cet outil nécessite de créer de nouveaux modèles à partir de données récentes.

L'objectif principal de ce mémoire a été de créer un algorithme sous R qui permette, à partir de l'implémentation d'une base de données brute, de réaliser des modèles linéaires généralisés de manière optimale. Cet algorithme, qui ne nécessite aucune intervention manuelle, permet également d'exporter les résultats de ces modèles sous un format facile à interpréter et à intégrer dans l'outil de tarification Excel.

Tout au long de ce mémoire, les avantages des modèles linéaires généralisés, qui permettent de fournir des tarifs en utilisant l'approche « Coût x Fréquence », ont été mis en évidence. Cette méthode a été préférée à la méthode de tarification déterministe puisqu'elle est plus adaptée à l'utilisation de variables explicatives continues comme l'âge ou le niveau de remboursement.

Plusieurs tests ont été réalisés afin de valider la cohérence des résultats issus de l'algorithme, notamment la tarification des échantillons choisis aléatoirement dans la base globale. Une phase de validation et de recettage commence maintenant pour adapter les résultats de l'outil de tarification en fonction de notre connaissance du risque et du marché de l'assurance santé.

Des études complémentaires, réalisées par le cabinet, permettent par exemple de confronter les tarifs obtenus dans l'outil « théorique » avec la réalité technique de d'autres portefeuilles.

D'autres études permettent d'intégrer des coefficients de zoniers ou encore de mutualisation en fonction de la typologie du contrat (individuel/collectif) et de la taille de l'entreprise.

Enfin, la tarification des actes liés au 100% santé est toujours en cours de développement. Nous observons les premiers résultats et attendons la fin de l'exercice 2021 pour avoir un peu plus de recul et mettre à jour les travaux d'analyse réalisés lors d'un précédent mémoire.

## Table des illustrations

▲ Figure 1 : Décomposition des régimes de la Sécurité Sociale.....	11
▲ Figure 2 : Décomposition du régime général de la sécurité sociale.....	12
▲ Figure 3 : Répartition du financement du régime général.....	13
▲ Figure 4 : Mécanisme de remboursement.....	17
▲ Figure 5 : Plancher et plafond en fonction de type d'équipement.....	20
▲ Figure 6 : Calendrier de mise en application de la réforme 100% santé.....	23
▲ Figure 7 : Nouveaux plafonds et planchers contrat responsable optique.....	25
▲ Figure 8 : Conditions de renouvellement : optique.....	26
▲ Figure 9 : Mise en place progressive du dispositif « 100% santé » en dentaire du 01/04/2019 au 01/01/2023.....	27
▲ Figure 10 : Prise en charge à 100% des couronnes et bridges du panier 1 depuis 2020 (source : Ameli).....	28
▲ Figure 11 : Mise en place progressive du dispositif « 100% santé » en audiologie du 01/01/2019 au 01/01/2021.....	29
▲ Figure 12 : Forme de la base finale construite à la fin de partie II.....	33
▲ Figure 13 : Éléments du portefeuille étudié.....	34
▲ Figure 14 : Schéma de traitement des bases brutes avant la modélisation par GLM .....	34
▲ Figure 15 : Table de correspondance entre les grilles de garanties et les libellés actes .....	35
▲ Figure 16 : Regroupement : Familles d'actes.....	36
▲ Figure 17 : Répartition des assurés et sortants de la mutuelle par âge 2018-2019... 39	39
▲ Figure 18 : Répartition des assurés et sortants de la mutuelle par âge 2019-2020... 40	40
▲ Figure 19 : Répartition par année et par type ayant droit.....	41
▲ Figure 20 : Répartition de l'exposition des enfants par âge en 2018 -- 2019 et 2020 42	42
▲ Figure 21 : Répartition de l'exposition des conjoints par âge en 2018 -- 2019 et 2020 .....	43
▲ Figure 22 : Répartition des assurés par : Famille / Seul.....	43
▲ Figure 23 : Répartition des enfants par type de famille.....	44
▲ Figure 24 : Régions assurées par la mutuelle (rouge).....	45
▲ Figure 25 : Remboursement par personne en médecine générale (€) - 2014 INSEE 46	46
▲ Figure 26 : Répartition par année et par gamme.....	47
▲ Figure 27 : Démographie par gamme étudiée.....	48
▲ Figure 28 : Exposition en fonction du niveau de gamme / année.....	49
▲ Figure 29 : Répartition de l'exposition par âge et par niveau de gamme.....	50
▲ Figure 30 : Échantillons de la base des effectifs.....	52
▲ Figure 31 : Répartition de l'exposition par âge et par Année.....	55
▲ Figure 32 : Variables de la base des prestations (base brute).....	55
▲ Figure 33 : Filtres appliqués sur la base des prestations.....	56
▲ Figure 34 : Échantillon de la base finale des prestations.....	57
▲ Figure 35 : Exemple de la base finale après les traitements.....	58
▲ Figure 36 : Exposition moyenne, min et max en fonction des âges.....	58
▲ Figure 37 : Coût moyen mensuel par âge.....	59

▲ Figure 38 : Coût moyen mensuel par âge : Hospitalisation.....	60
▲ Figure 39 : Coût moyen mensuel par âge : Dentaire.....	61
▲ Figure 40 : Coût moyen mensuel par âge : Pharmacie.....	61
▲ Figure 41 : Tarif mensuel moyen par âge : Soins courants.....	62
▲ Figure 42 : Coût moyen mensuel par âge : Optique.....	63
▲ Figure 43 : Coût moyen mensuel par âge : Appareillage.....	63
▲ Figure 44 : Coût moyen mensuel par âge : Autres.....	64
▲ Figure 45 : Situation de GLM par rapport à l'IA.....	65
▲ Figure 46 : Situation du Machine Learning par rapport à l'IA.....	66
▲ Figure 47 : Apprentissage supervisé.....	67
▲ Figure 48 : Types d'apprentissage supervisé.....	68
▲ Figure 49 : Types d'apprentissage supervisé.....	69
▲ Figure 50 : Exemple de la régression non linéaire.....	71
▲ Figure 51 : Nombre de d'adhérents en fonction de l'âge et le niveau de garantie....	81
▲ Figure 52 : Etude de l'exposition pour 'Chirurgie -NON OPTAM'.....	82
▲ Figure 53 : Résultat du test de Khi-deux.....	83
▲ Figure 54 : Répartition du nombre moyen d'actes (fréquence) en fonction de l'âge	85
▲ Figure 55 : Densité de la variable 'Nombre d'actes'.....	85
▲ Figure 56 : récapitulatif des résultats des lois discrètes.....	86
▲ Figure 57 : Comparaison de la densité réelle avec la densité de la loi ZISICHEL....	87
▲ Figure 58 : Valeur de l'AIC par Modèle_FREQ.....	88
▲ Figure 59 : Résumé des résultats de la modélisation de la fréquence.....	88
▲ Figure 60 : Nombre de bénéficiaires en fonction de l'âge et du niveau de garantie.	89
▲ Figure 61 : Coût moyen en fonction de l'âge.....	90
▲ Figure 62 : Densité de la variable_CM.....	91
▲ Figure 63 : récapitulatif des résultats des lois continues.....	91
▲ Figure 64 : Comparaison de la densité réelle avec la densité de la loi BCTo.....	92
▲ Figure 65 : Valeur de l'AIC par Modèle_CM.....	93
▲ Figure 66 : Résumé des résultats de la modélisation du coût moyen.....	94
▲ Figure 67 : Fréquence annuelle en fonction de l'âge et du niveau de remboursement (Hors Covid-19).....	95
▲ Figure 68 : Coût moyen annuel en fonction de l'âge et du niveau de remboursement (Hors Covid-19).....	96
▲ Figure 69 : Tarif annuel en fonction de l'âge et du niveau de garantie.....	97
▲ Figure 70 : Tarif annuel en fonction de l'âge et du niveau de garantie.....	98
▲ Figure 71 : Fréquence annuelle en fonction de l'âge et du niveau de remboursement (Covid-19).....	98
▲ Figure 72 : Coût moyen annuel en fonction de l'âge et du niveau de remboursement (Covid-19).....	99
▲ Figure 73 : Feuille de 'saisie' sur Excel.....	101
▲ Figure 74 : Feuille de 'GLM' : Paramétrage.....	102
▲ Figure 75 : Feuille de 'GLM' : Modèles GLM.....	102
▲ Figure 76 : densité des forfaits pour prothèses dentaires remboursées par la sécurité sociale.....	105

# Annexe

## Annexe 1 : Familles & libellés actes

### HOSPITALISATION

- Actes d'anesthésie - OPTAM
- Actes d'anesthésie - NON OPTAM
- Actes de chirurgie - OPTAM
- Actes de chirurgie - NON OPTAM
- Actes d'échographie - OPTAM
- Actes d'échographie - NON OPTAM
- Actes d'imagerie - OPTAM
- Actes d'imagerie - NON OPTAM
- Actes d'obstétrique - OPTAM
- Actes d'obstétrique - NON OPTAM
- Forfait hospitalier
- Frais de séjour médical et chirurgical
- Frais de séjour psychiatrie
- Chambre particulière maternité
- Chambre particulière médicale et chirurgicale
- Chambre particulière psychiatrie
- Frais d'accompagnant
- Frais de transport

### APPAREILLAGE

- Prothèse auditive I
- Prothèse auditive II
- Appareillage
- Accessoires

### PHARMACIE

- Pharmacie 15%
- Pharmacie 30%
- Pharmacie 65%
- Vaccins anti-grippe
- Forfait médicaments non remboursés

### SOINS COURANTS

- Actes techniques médicaux - OPTAM
- Actes techniques médicaux - NON OPTAM
- Actes de radiologie - OPTAM
- Actes de radiologie - NON OPTAM
- Analyses médicales
- Auxiliaires médicaux
- Consultations de généralistes - OPTAM
- Consultations de généralistes - NON OPTAM
- Consultations de spécialistes - OPTAM
- Consultations de spécialistes - NON OPTAM
- Consultations de psychiatres - OPTAM
- Consultations de psychiatres - NON OPTAM

### AUTRES

- Cure thermale
- Médecine douce - Acupuncture
- Médecine douce - Chiropractie
- Médecine douce - Diététique
- Médecine douce - Ethiope
- Médecine douce - Ostéopathie
- Médecine douce - Pédicure
- Médecine douce - Podologie
- Soins à l'étranger
- Consultation sport
- Arrêt du tabac
- Autres vaccins
- Contraception
- Fécondation in vitro
- Ostéodensitométrie
- Prestation appairage
- Prime de naissance ou d'adoption

## OPTIQUE

- Optique 100% santé
- Monture
- Verre simple
- Verre complexe
- Verre très complexe
- Lentilles acceptées
- Lentilles refusées
- Chirurgie réfractive

## DENTAIRE

- Soins dentaires
- Inlay onlay
- Inlay core
- Prothèse dentaire 100%
- Prothèse dentaire acceptée
- Prothèse dentaire refusée
- Orthodontie acceptée / refusée
- Implant dentaire
- Parodontologie

## Annexe 2 : Liste des lois continues positives de ‘gamlss’

Distribution	gamlss		Parameter link functions			
	name	Range $R_Y$	$\mu$	$\sigma$	$\nu$	$\tau$
beta	BE	(0, 1)	logit	logit	-	-
Box-Cox Cole-Green	BCCG	(0, $\infty$ )	ident.	log	ident.	-
Box-Cox Cole-Green orig.	BCCGo	(0, $\infty$ )	log	log	ident.	-
Box-Cox power exponential	BCPE	(0, $\infty$ )	ident.	log	ident.	log
Box-Cox power expon. orig.	BCPEo	(0, $\infty$ )	log	log	ident.	log
Box-Cox $t$	BCT	(0, $\infty$ )	ident.	log	ident.	log
Box-Cox $t$ orig.	BCTo	(0, $\infty$ )	log	log	ident.	log
exponential	EXP	(0, $\infty$ )	log	-	-	-
exponential Gaussian	exGAUS	( $-\infty$ , $\infty$ )	ident.	log	log	-
exponential gen. beta 2	EGB2()	( $-\infty$ , $\infty$ )	ident.	log	log	log
gamma	GA	(0, $\infty$ )	log	log	-	-
generalized beta type 1	GB1	(0, 1)	logit	logit	log	log
generalized beta type 2	GB2	(0, $\infty$ )	log	log	log	log
generalized gamma	GG	(0, $\infty$ )	log	log	ident.	-
generalized inv. Gaussian	GIG	(0, $\infty$ )	log	log	ident.	-
generalized $t$	GT	( $-\infty$ , $\infty$ )	ident.	log	log	log
Gumbel	GU	( $-\infty$ , $\infty$ )	ident.	log	-	-
inverse Gamma	IGAMMA	(0, $\infty$ )	log	log	-	-
inverse Gaussian	IG	(0, $\infty$ )	log	log	-	-
Johnson’s SU repar.	JSU	( $-\infty$ , $\infty$ )	ident.	log	ident.	log
Johnson’s SU original	JSUo	( $-\infty$ , $\infty$ )	ident.	log	ident.	log
logistic	LO	( $-\infty$ , $\infty$ )	ident.	log	-	-
logit normal	LOGITNO	(0, 1)	ident.	log	-	-
log normal	LOGNO	(0, $\infty$ )	ident.	log	-	-



Distribution	gamlss name	Range $R_Y$	Parameter link functions			
			$\mu$	$\sigma$	$\nu$	$\tau$
log normal 2	LOGNO2	$(0, \infty)$	log	log	-	-
log normal (Box-Cox)	LNO	$(0, \infty)$	ident.	log	fixed	-
NET	NET	$(-\infty, \infty)$	ident.	log	fixed	fixed
normal	NO, NO2	$(-\infty, \infty)$	ident.	log	-	-
normal family	NOF	$(-\infty, \infty)$	ident.	log	-	-
Pareto 2	PARETO2	$(0, \infty)$	log	log	-	-
Pareto 2 original	PARETO2o	$(0, \infty)$	log	log	-	-
Pareto 2 repar	GP	$(0, \infty)$	log	log	-	-
power exponential	PE	$(-\infty, \infty)$	ident.	log	log	-
reverse gen. extreme	RGE	$y > \mu - (\sigma/\nu)$	ident.	log	log	-
reverse Gumbel	RG	$(-\infty, \infty)$	ident.	log	-	-
sinh-arcsinh	SHASH	$(-\infty, \infty)$	ident.	log	log	log
sinh-arcsinh original	SHASHo	$(-\infty, \infty)$	ident.	log	ident.	log
sinh-arcsinh original 2	SHASHo2	$(-\infty, \infty)$	ident.	log	ident.	log
skew normal type 1	SN1	$(-\infty, \infty)$	ident.	log	ident.	-
skew normal type 2	SN2	$(-\infty, \infty)$	ident.	log	log	-
skew power exp. type 1	SEP1	$(-\infty, \infty)$	ident.	log	ident.	log
skew power exp. type 2	SEP2	$(-\infty, \infty)$	ident.	log	ident.	log
skew power exp. type 3	SEP3	$(-\infty, \infty)$	ident.	log	log	log
skew power exp. type 4	SEP4	$(-\infty, \infty)$	ident.	log	log	log
skew $t$ type 1	ST1	$(-\infty, \infty)$	ident.	log	ident.	log
skew $t$ type 2	ST2	$(-\infty, \infty)$	ident.	log	ident.	log
skew $t$ type 3	ST3	$(-\infty, \infty)$	ident.	log	log	log
skew $t$ type 3 repar	SST	$(-\infty, \infty)$	ident.	log	log	log-2
skew $t$ type 4	ST4	$(-\infty, \infty)$	ident.	log	log	log
skew $t$ type 5	ST5	$(-\infty, \infty)$	ident.	log	ident.	log
t Family	TF	$(-\infty, \infty)$	ident.	log	log	-
t Family repar	TF2	$(-\infty, \infty)$	ident.	log	log-2	-
Weibull	WEI	$(0, \infty)$	log	log	-	-
Weibull (PH)	WEI2	$(0, \infty)$	log	log	-	-
Weibull ( $\mu$ the mean)	WEI3	$(0, \infty)$	log	log	-	-

### Annexe 3 : Liste des lois discrètes de ‘gamlss’

Distribution	gamlss name	Range	Parameter range		
		$R_Y$	$\mu$	$\sigma$	$\nu$
beta binomial	BB	$\{0, 1, \dots, n\}$	logit	log	-
binomial	BI	$\{0, 1, \dots, n\}$	logit	-	-
Delaporte	DEL	$\{0, 1, 2, \dots\}$	log	log	logit
geometric	GEOM	$\{0, 1, 2, \dots\}$	log	-	-
logarithmic	LG	$\{1, 2, 3, \dots\}$	logit	-	-
negative binomial type I	NBI	$\{0, 1, 2, \dots\}$	log	log	-
negative binomial type II	NBII	$\{0, 1, 2, \dots\}$	log	log	-
Poisson	PO	$\{0, 1, 2, \dots\}$	log	-	-
Poisson inverse Gaussian	PIG	$\{0, 1, 2, \dots\}$	log	log	-
Sichel	SI	$\{0, 1, 2, \dots\}$	log	log	identity
Sichel ( $\mu$ the mean)	SICHEL	$\{0, 1, 2, \dots\}$	log	log	identity
Waring ( $\mu$ the mean)	WARING	$\{0, 1, 2, \dots\}$	log	log	-
Yule ( $\mu$ the mean)	YULE	$\{0, 1, 2, \dots\}$	log	-	-
zero altered beta binomial	ZABB	$\{0, 1, \dots, n\}$	logit	log	logit
zero altered binomial	ZABI	$\{0, 1, \dots, n\}$	logit	logit	-
zero altered logarithmic	ZALG	$\{0, 1, 2, \dots\}$	logit	logit	-
zero altered neg. binomial	ZANBI	$\{0, 1, 2, \dots\}$	log	log	logit
zero altered Poisson	ZAP	$\{0, 1, 2, \dots\}$	log	logit	-
zero inflated beta binomial	ZIBB	$\{0, 1, 2, \dots\}$	logit	log	logit
zero inflated binomial	ZIBI	$\{0, 1, \dots, n\}$	logit	logit	-
zero inflated neg. binomial	ZINBI	$\{0, 1, 2, \dots\}$	log	log	logit
zero inflated Poisson	ZIP	$\{0, 1, 2, \dots\}$	log	logit	-
zero inflated Poisson ( $\mu$ the mean)	ZIP2	$\{0, 1, 2, \dots\}$	log	logit	-
zero inflated Poisson inv. Gaussian	ZIPIG	$\{0, 1, 2, \dots\}$	log	log	logit

## Annexe 4 : Résultat des lois simulateurs de la fréquence

Loi_Fréquence	Coefficient
Loi	Coefficient
ZASICHEL	32000,9319
ZISICHEL	32000,9319
BNB	32029,9798
ZABNB	32031,9717
ZIBNB	32031,9798
WARING	32042,7565
SI	32047,8211
SICHEL	32047,8211
ZAPIG	32050,9187
ZIPIG	32050,9187
PIG	32050,9912
GPO	32064,0943
DEL	32070,3831
ZANBI	32070,8505
NBI	32083,3333
NBII	32083,3333
ZALG	32084,052
NBF	32085,3333
ZINBI	32085,3333
ZINBF	32087,3333
ZINBF	32087,3333
GEOM	32102,4739
GEOMo	32102,4739
DPO	32159,0704
ZIP2	32466,8125
ZIP	32466,8125
ZAP	32466,8125
ZAZIPF	32528,2743
YULE	32772,8943

Loi_CM	Coefficient
BCTo	40786,8992
GB2	40797,2159
BCT	40798,3724
BCPEo	40804,941
BCPE	40804,941
BCCG	40905,1802
BCCGo	40905,1802
GG	40939,824
IGAMMA	40971,2748
GIG	40973,2748
LOGNO	41373,482
LOGNO2	41373,482
IG	41681,9208
exGAUS	42917,6174
PARETO2	43256,7401
GP	43256,7401
PARETO2o	43256,7401
GA	43957,3649
WEI	44047,093

## Références

- [1] Arthur CHARPENTIER et Christophe DUTANG, L'actuariat avec R, 2012
- [2] Michel DENNUI et Arthur CHARPENTIER, Mathématiques de l'assurance non-vie, Tome II, Economica, 2005
- [3] Mikis D. STASINOPOULOS et Robert A. RIGBY et Gillian Z. HELLER et Vlasios VOUDOURIS et Fernanda DE BASTIANI, Flexible Regression and Smoothing Using GAMLSS in R, CRC PRESS, 2017
- [4] Xavier MILHAUD, Data Science, Machine Learning et Actuariat, Cours de Master ISFA, 2019
- [5] Arthur Charpentier, Actuariat de l'Assurance Non-Vie, Cours de Master ENSAE, 2017
- [6] Cyrille CHERY, Construction d'un outil de tarification de contrats complémentaires santé, Mémoire d'actuaire ISFA 2015
- [7] Julie VERNIN-BIANCALE, Création d'un outil permettant d'estimer les impacts de la réforme 100% santé à partir de l'historique d'un portefeuille, Mémoire d'actuaire ISFA 2020
- [8] Mathieu ADAM, Différentes méthodes de tarification de garanties complémentaires santé et intégration de ces modèles dans un outil de tarification de contrat, Mémoire d'actuaire ISUP 2013
- [9] Fatemeh ABDOLLAHI, Tarification d'une complémentaire santé à destination des séniors, modulaire par poste de garanties et l'impact sur la solvabilité, Mémoire d'actuaire ISUP, 2018
- [10] Florence MILLET, Analyse de la sinistralité et tarification d'un produit d'assurance complémentaire santé, Mémoire d'actuaire ISFA, 2014
- [11] Ngoc Trung Phuong NGUYEN, **Construction de bases de tarification pour des contrats complémentaires santé collectifs par le Modèle Linéaire Généralisé**, Mémoire d'actuaire ISFA, 2013