IRM | institut du risk management BY Institut.

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le 31/03/2022

Par : Faïçal CHARKI

Titre : Impact du 100% santé : Mesure de l'évolution du recours aux soins et de la sinistralité à l'aide des modèles classiques et exploration des méthodes machine learning

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

Membres présents du jury de l'Institut du Risk Management :

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise : AESIO_____

Nom : _____

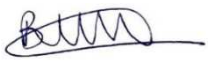
Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : ___Corine BENOIT

_____

Signature :



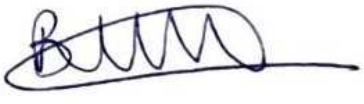
Invité :

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(s) du candidat(s)

Résumé

Les complémentaires santé comme les autres assureurs évoluent dans un marché très encadré mais en constante transformation. Des transformations qui peuvent être engendrées par l'évolution de la société avec l'émergence de nouveaux besoins par exemple, ou par les réformes réglementaires successives que connaît ce secteur. Selon leurs finalités, certaines de ces réformes peuvent influer le comportement de l'assuré et modifier de ce fait les informations jusqu'alors acquises concernant la sinistralité d'un portefeuille.

Dès lors, il devient nécessaire d'assurer une veille afin d'une part, se mettre en conformité et d'autre part, identifier les impacts de ces réformes sur les équilibres techniques. Le travail effectué le long de ce mémoire s'inscrit dans ce cadre et vise à mesurer le coût lié à la réforme du 100% santé qui aspire à favoriser l'accès aux soins et peut potentiellement conduire à une dégradation de la sinistralité.

Dans un premier temps, différents tests statistiques seront réalisés pour détecter toute évolution dans la consommation des soins dans un contexte de crise sanitaire. Par la suite, différents modèles seront élaborés pour permettre de mesurer la charge des sinistres de l'exercice 2020. Deux types de modèles seront proposés : des modèles appartenant à la famille GLM souvent utilisés dans le traitement des problématiques liées à la tarification, et des modèles de type machine Learning qui offrent plus de flexibilité et qui sont de plus en plus déployés.

Le mémoire s'achèvera avec une analyse critique des différents résultats obtenus

Abstract

Supplementary health insurance companies, like other insurers, operate in a highly regulated market that is constantly changing. These changes can be caused by the evolution of society with the emergence of new needs, for example, or by the successive regulatory reforms that this sector has undergone. Depending on their purpose, some of these reforms can influence the behaviour of the insured and thus modify the information previously acquired concerning the loss experience of a portfolio.

It is therefore necessary to monitor the situation in order to comply with the regulations and to identify the impact of these reforms on technical balances. The work carried out in this thesis is part of this framework and aims to measure the cost of the 100% health reform, which aims to promote access to care and could potentially lead to a deterioration in the claims experience.

Initially, various statistical tests will be carried out to detect any changes in the consumption of care in a health crisis context. Subsequently, different models will be developed to measure the burden of claims for the year 2020. Two types of models will be proposed: models belonging to the GLM family often used in the treatment of problems related to pricing and models of the machine learning type which offer more flexibility and which are increasingly deployed.

The thesis will end with a critical analysis of the different results obtained.

Remerciements

Je souhaite témoigner ma gratitude à l'ensemble des personnes qui m'ont accompagné tout au long du parcours et ont contribué à la réalisation de ce mémoire, et plus particulièrement :

- Madame Corine BENOIT, directeur de mon mémoire, pour sa relecture attentive et ses conseils avisés ;
- Monsieur Thomas SENNE, qui m'a fait bénéficier de sa très grande expérience du risque santé et sa connaissance du portefeuille AESIO ;
- Monsieur Laurent MASSOUTIER, pour ses précieux conseils et encouragements.

Je souhaite également formuler mes plus sincères remerciements au corps professoral du Centre d'Études Actuarielles pour le savoir transmis et les conseils prodigués tout au lent de la formation.

Je remercie, enfin, ma famille et mes proches pour leur soutien. Je souhaite remercier plus particulièrement ma femme pour la patience dont elle a fait preuve durant tout le temps consacré à la réalisation de ce mémoire.

Table des matières

Introduction

Suivre et analyser la sinistralité représente un enjeu majeur pour tout assureur. Cela permet en effet, d'adapter son tarif au risque couvert et de veiller à la suffisance de ce dernier pour faire face à la charge de sinistre attendue. Cette règle est d'autant plus vraie pour les organismes complémentaires couvrant le risque santé car ils évoluent dans un marché de plus en plus réglementé, très concurrentiel avec une importante pression tarifaire et où nous assistons à une érosion des marges et à des équilibres techniques fragiles.

L'un des objectifs recherchés dans le cadre du suivi et de l'analyse de la sinistralité d'un portefeuille, est de détecter, voire anticiper, toute dérive. Une dérive qui peut être la conséquence d'une inflation naturelle des prestations, d'une antisélection ou d'une évolution du cadre réglementaire affectant le comportement des assurés. Tous ces facteurs de dégradation de la sinistralité peuvent conduire à une augmentation des tarifs. Mais dans certains cas, ces analyses aboutissent aussi à un maintien ou une baisse des tarifs afin d'être plus concurrentiel sans pour autant dégrader les marges. Enfin, ce travail permet aussi de vérifier la pertinence des différentes segmentations en place et la nécessité ou pas de les faire évoluer.

L'une des récentes évolutions réglementaires affectant le marché de l'assurance santé, est la réforme 100% santé. Elle vise à supprimer le reste à charge dans le domaine de l'optique, des aides auditives et du dentaire de manière progressive à compter de 2019, en imposant aux professionnels de santé de proposer des paniers 100% santé et aux complémentaires santé de proposer des garanties avec un reste à charge à zéro sur ces paniers.

Le reste à charge, qui correspond à la participation des ménages aux dépenses de soins, a longtemps été considéré comme un levier de régulation des dépenses de santé par le système d'assurance maladie française. Néanmoins, plusieurs études et rapports démontrent que cette dépense constitue un frein et favorise le renoncement aux soins surtout si leurs restes à charge sont élevés. Le but principal de cette réforme est donc de favoriser l'accès aux soins dans les trois domaines concernés. Et cela passera forcément par un changement des habitudes de consommation avec un recours plus important aux soins en question. Il est donc nécessaire pour les complémentaires santé de mesurer l'impact de cette réforme afin de veiller aux équilibres techniques.

L'objectif de ce mémoire sera d'analyser le comportement des assurés suite à la mise en place de cette réforme, de modéliser la nouvelle charge de sinistres et de mesurer le recours aux paniers dits 100% santé. Il comportera cinq parties :

- Une première partie comportant un descriptif de la consommation de soins et de biens médicaux en France et ses sources de financements, un rappel des principes généraux de la réforme 100% santé ainsi qu'une présentation de la mutuelle AESIO ;
- Une seconde partie décrira les données qui seront utilisées dans le cadre de ce mémoire ;
- La troisième partie sera dédiée à la mesure d'impact en s'appuyant sur des modèles paramétriques ;
- La quatrième partie sera consacrée à la mesure d'impact à l'aide des méthodes de type Machine Learning ;
- La dernière partie, sous forme de conclusion, apportera une vision critique des différents analyses et résultats.

La consommation de soins et de biens médicaux en France et son financement

Partout dans le monde, se soigner présente un coût qui varie selon la nature des soins et peut devenir conséquent pour certains. Des soins qui sont par ailleurs, dans la majorité des cas, indispensables. Dès lors, il devient évident que l'état de santé d'un individu ainsi que son niveau d'accès aux soins dépendront, entre autres, des moyens de financement dont nous disposons pour couvrir les dépenses engagées.

En France, il existe trois sources de financement qui sont analysées chaque année dans plusieurs rapports dont le rapport « dépenses de santé » élaboré par la DREES[1] et consacré à l'analyse de la consommation de soins et de biens médicaux (CSBM) ainsi qu'à ses différentes sources de financement. Ces trois sources sont :

- Le financement par la sécurité sociale qui constitue la première brique du financement ;
- Le financement par la complémentaire santé qui intervient sur la part non prise en charge par la sécurité sociale ou le régime général ;
- Le financement personnel ou le reste à charge (RAC) qui correspond à la part non couverte ni par sécurité sociale ni par la complémentaire et qui reste donc à la charge de chaque individu.

Le schéma ci-dessous présente l'articulation de ces trois sources

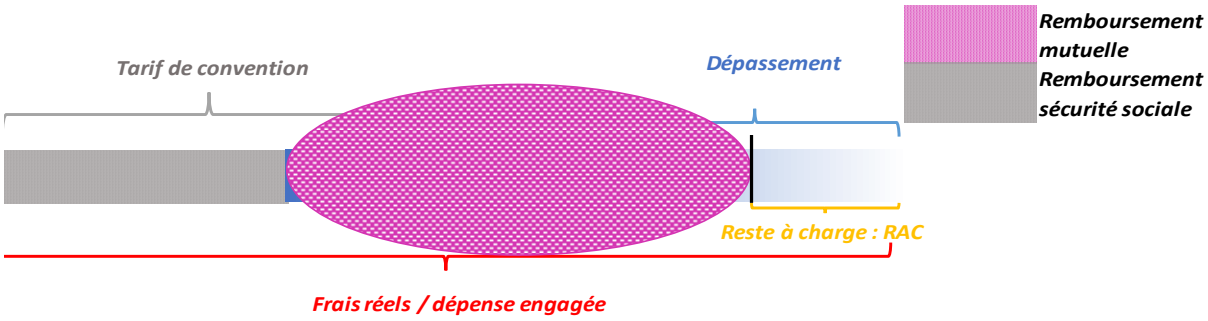


Figure 1 : Articulation entre les trois sources de financement de la CSBM

Le tarif de convention (TC) détermine pour chaque acte médical le montant sur lequel s'applique le taux de remboursement de la Sécurité Sociale. Il est régi par des conventions entre les professionnels de santé et les organismes de Sécurité Sociale. Le dépassement, lui, représente l'écart entre le TC et le tarif pratiqué par le professionnel de santé.

Le rapport « dépenses de santé 2019 » produit en 2020, rappelle qu'en France (pays avec le reste le plus faible), les dépenses de santé représentent 11.3% du PIB, ce qui classe notre pays 2ème au niveau européen derrière l'Allemagne. Il fait état aussi d'une accélération de la CSBM qui progresse de +2.1% pour atteindre 208 milliards d'euros. Cette progression, plus forte que les années précédentes (+1.6% en 2017 et 1.6% en 2018), est due principalement aux soins hospitaliers qui redeviennent le 1er facteur de croissance de cette consommation devant les soins de ville. En effet, ces derniers progressent de +2.4% pour une contribution à la CBSM de 1.1% contre une contribution de 0.7% pour les soins de ville. L'augmentation des prix de +1.3% en 2019 (contre +0.3% en 2018), constitue la principale explication de l'évolution des soins hospitaliers.

[1] Direction de la recherche, des études, de l'évaluation et des statistiques, appartenant l'administration centrale des ministères sanitaires et sociaux. Elle a été mise en place en 1998

Le tableau ci-dessous extrait du rapport cité précédemment, récapitule l'évolution en % de la CSBM en valeur avec ses principales composantes et en volume :

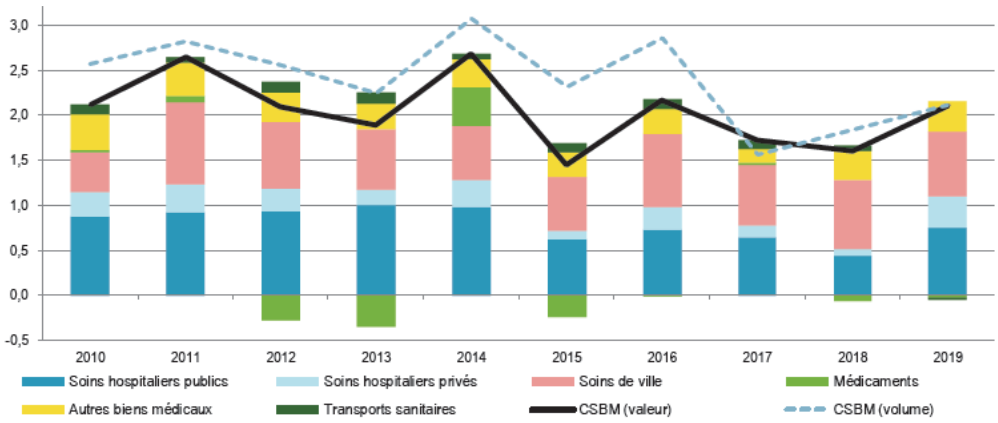


Figure 2: Évolution en % de la CSBM en valeur avec ses principales composantes et en volume

Enfin, la répartition de la CSBM 2019 par source de financement, montre que le premier financeur est la Sécurité Sociale avec 78% des parts, suivi par les complémentaires qui représentent 13.42%. Le RAC pèse, comme évoqué plus haut, 6.9% de ces dépenses. Le tableau suivant synthétise les montants des financements par source sur les années 2019 et 2018.

		2019		2018	
		Montant	%	Montant	%
Public	Sécurité sociale	162 708	78,21%	**158 981**	**78,03%**
	État, collectivités locales et CMU-Corganismes de base	3 104	1,49%	**2 991**	**1,47%**
	Sous total	**165 812**	**79,70%**	**161 972**	**79,50%**
Complémentaires	Mutuelles	13 557	6,52%	13 606	6,68%
	Sociétés d'assurances	8 920	4,29%	8 333	4,09%
	Institutions de prévoyance	5 445	2,62%	5 373	2,64%
	Sous total	**27 922**	**13,42%**	**27 312**	**13,40%**
RAC	Ménages	14 301	6,87%	14 463	7,10%
	Sous total	**14301**	**6,87%**	**14463**	**7,10%**
Total		**208 035**	**100%**	**203 747**	**100%**

Tableau 1: Montant de la CSBM par financeur en millions d'euros

Sécurité sociale

Naît en 1945 grâce à l'ordonnance du 04/10/1945, elle est définie selon l'exposé des motifs de ladite ordonnance, comme « la garantie donnée à chacun qu'en toutes circonstances il disposera des moyens nécessaires pour assurer sa subsistance et celle de sa famille dans des conditions décentes ». Elle regroupe cinq branches :

- L'assurance maladie qui représentait en 2019, 52.1 % des dépenses de la sécurité sociale ;
- La réparation des accidents du travail et des maladies professionnelles (2.9% des dépenses) ;
- Les retraites (33% des dépenses) ;
- Les prestations familiales (12% des dépenses) ;
- La branche consacrée à l'autonomie (depuis 2020).

Et se compose de quatre régimes de base :

- Le régime général qui représente 80 % des charges de l'ensemble des régimes de base et qui couvre 59.2 millions de bénéficiaires composés des salariés, des inactifs, des indépendants et des professions libérales (pour le risque maladie). Ces deux dernières catégories ont été rattachées à ce régime début 2018 ;
- Le régime des salariés et des exploitants agricoles, géré par la caisse centrale de la mutualité sociale agricole (CCMSA) ;
- Les régimes spéciaux, destinés à certaines professions comme les fonctionnaires, les salariés la SNCF, de la RATP ou encore ceux des industries électriques et gazières…

Le financement des différents régimes et branches et donc de la sécurité sociale dans son ensemble, est assuré par les charges sociales dues par les employeurs et leurs salariés, la CSG (contribution sociale généralisée) et enfin par d'autres contributions et taxes.

La branche maladie du régime général (41,68% des charges de la sécurité sociale) est gérée par la CNAM (Caisse nationale d'assurance maladie) qui supervise le réseau des caisses primaires d'assurance maladie (CPAM) .Elle couvre plusieurs dépenses comme les soins de ville (honoraires des professionnels de santé libéraux, les indemnités journalières, les dépenses ambulatoires de médicaments et de dispositifs médicaux, ainsi que les transports sanitaires), les dépenses liées à des établissements de santé, les dépenses Médico-sociales, selon la répartition suivante :

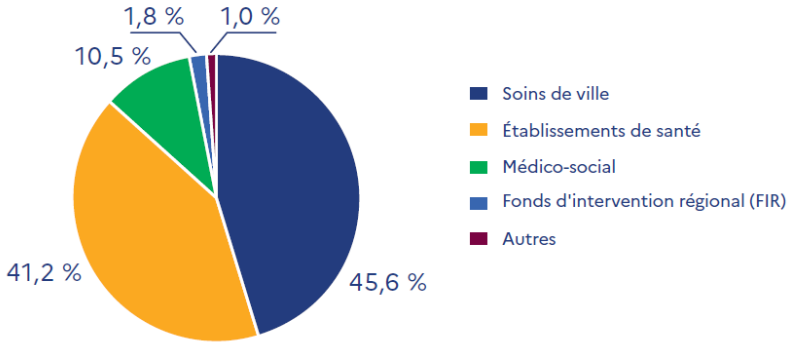


Figure 3: Répartition des dépenses de santé financées par l'Assurance maladie

Les établissements de santé ainsi que les soins de ville constituent, dans l'ordre, les plus importantes dépenses avec respectivement 45.6% et 41.2% des parts.

Comme évoqué dans le chapitre précédent, à travers cette branche, la sécurité sociale est le 1er financeur de la CBSM (78% des parts en 2019). Nous constatons toutefois des disparités assez fortes selon les postes de soins comme le montre le graphique suivant :

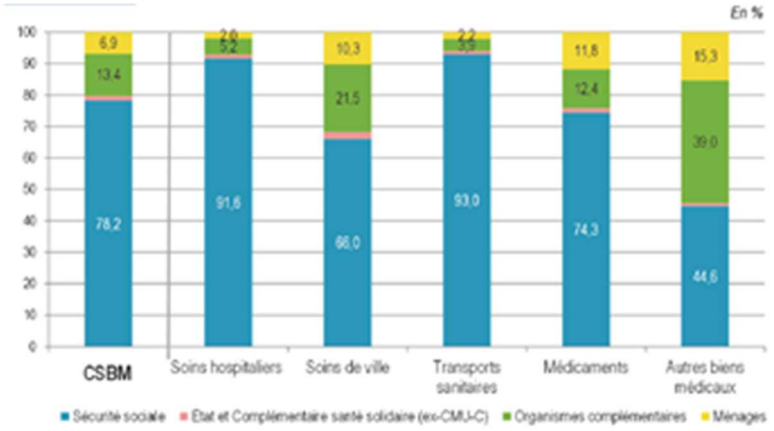


Figure 4 : Structure du financement des grands postes de la CSBM en 2019

En effet, la sécurité sociale finance 91,6% du coût des soins hospitaliers alors qu'elle n'intervient qu'à hauteur de 44,6% pour les autres biens médicaux. Toutefois, elle reste le 1er financeur quel que soit le type de la CSBM.

Il est à noter aussi que cette branche connait un solde négatif assez important depuis quelques années, avec une amélioration entre 2013 et 2018 et puis une dégradation en 2019 pour atteindre -1,5 Milliard d'euros et ce malgré les quelques 215 Milliards d'euros de produits nets destinés à financer cette branche. Le tableau suivant retrace l'évolution de son solde :

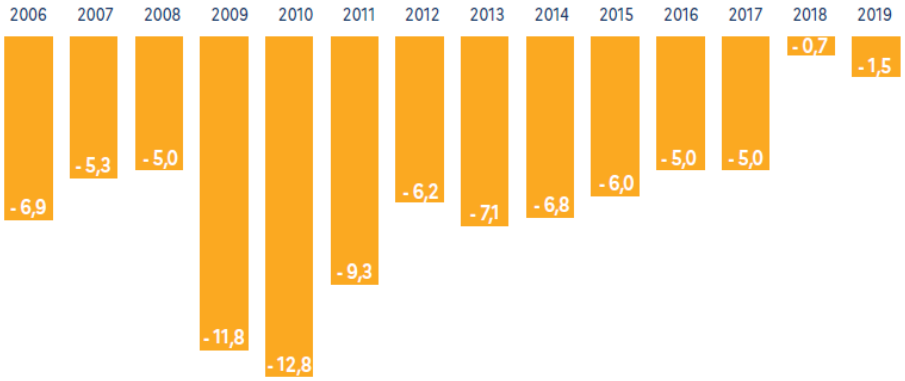


Figure 5: solde de la branche maladie en Milliard d'euros

Organismes complémentaires

Loin derrière la sécurité sociale, les organismes complémentaires (OC) constituent la deuxième source de financement de la CSBM. Ils sont composés de trois types de structures :

- Les mutuelles et groupements mutualistes : elles relèvent du Code de la Mutualité et représentent 86% du nombre total des OC. Ceux sont des sociétés du droit privé à but non lucratif. Soumises à l'origine à la charte de la mutualité de 1898, elles continuent de fonctionner aujourd'hui en vertu du principe de l'autogestion. Elles interviennent principalement dans le domaine de la complémentaire santé mais peuvent aussi mener des actions de prévoyance, de solidarité et d'entraide ;
- Les institutions de prévoyance : elles relèvent du Code de la Sécurité sociale et représentent 4% du nombre total des OC. Ces organismes de droit privé à but non lucratif, Couvrent les risques suivants : retraite, décès, incapacité, invalidité, en plus de la complémentaire santé ;

- Les sociétés d'assurance à but lucratif : elles relèvent du Code des Assurances et représentent 10% du nombre total des OC. Elles n'assurent pas de mission sociale contrairement aux précédents et sont exclues des comptes de la protection sociale.

Les trois types d'OC ont un fonctionnement assurantiel basé sur la mutualisation, à savoir qu'ils assurent leurs adhérents contre l'avènement d'un risque (risque financier lié aux problèmes de santé) en contre partie du paiement d'une cotisation. Les cotisations serviront principalement à couvrir les prestations versées aux adhérents au titre du risque assuré. La part restante permet de financer les frais de gestion et les bénéfices. A noter que ces sociétés évoluent dans un environnement très réglementé et en constante évolution mais aussi très concurrentiel ce qui a conduit à une érosion des marges. Parmi ces dernières évolutions nous pouvons citer :

- L'accord national interprofessionnel du 11 janvier 2013 : a conduit à la généralisation de la couverture complémentaire d'entreprise à l'ensemble des salariés du privé au 1er janvier 2016 quel que soit la taille de l'entreprise. L'accord prévoit que l'employeur finance le contrat collectif à hauteur de 50% minimum tout en lui garantissant la liberté de choix de son assureur. Le contrat en question doit lui inclure un niveau de garantie minimum ;
- La réforme des contrats solidaires et responsables entrée en vigueur en avril 2015 : d'après le décret n° 2014-1374 du 18 novembre 2014, impose une prise en charge complète de la participation de l'adhérent à ce type de contrat sur les tarifs de convention, sauf pour certains actes. Le forfait hospitalier doit donner lieu lui aussi à une couverture intégrale. La réforme encadre aussi la prise en charge des dépassements d'honoraires et des dépenses d'optique.
- Les réformes successives de la fiscalité avec la dernière en date applicable depuis 2016 et présente dans la loi de financement de la Sécurité sociale pour 2015. Elle a conduit à la fusion de la taxe de solidarité additionnelle (6.27%) et de taxe spéciale sur les conventions d'assurances (7%) et à la création d'une taxe unique de 13.27% pour les contrats responsables et de 2.27% pour les contrats non responsables ;
- L'entrée en vigueur au 01 janvier 2016 de la directive européenne solvabilité 2 : i est une réforme réglementaire qui vise le secteur de l'assurance et qui est venue renforcer les règles prudentielles. Les compagnies d'assurance et de réassurance sont dans l'obligation de mieux adapter les fonds propres aux risques qu'elles encourent dans leur activité, avec - dans certains cas - la nécessité de renforcer ces fonds ;
- La réforme du 100% santé avec la mise en place du RAC 0 garantit aux assurés l'absence de reste à charge après intervention des complémentaires santé dans les domaines de l'optique, des aides auditives et des prothèses dentaires. Les impacts de cette réforme (recours à ce dispositif, évolutions des prestations sur ce périmètre…) feront l'objet d'une analyse approfondie dans ce mémoire ;
- La loi n° 2019-733 du 14 juillet 2019 relative au droit de résiliation sans frais de contrats de complémentaire santé introduit pour les assurés la possibilité de résilier, après un an de souscription, leur contrat de complémentaire santé, à tout moment, sans frais ni pénalité. Auparavant, cette résiliation ne pouvait avoir lieu qu'une fois par an à la date d'anniversaire du contrat ;
- Les différents dispositifs d'aides de l'état pour bénéficier d'une couverture complémentaire : la Couverture maladie universelle complémentaire (CMU-C), l'aide à la complémentaire santé (ACS) et la complémentaire santé solidaire (CSS) née le 1er novembre 2019 de la fusion des deux premiers dispositifs .

Dans ce contexte, les OC ont versé des prestations santé à hauteur de 27,9 milliards d'euros en 2019 soit 13,42% de la CSBM. Une part qui reste stable dans le temps avec des prestations versées qui croissent à une vitesse similaire à celle de la CSBM, mais qui diffère sensiblement d'un type de soin à l'autre, comme le montre le graphique suivant :

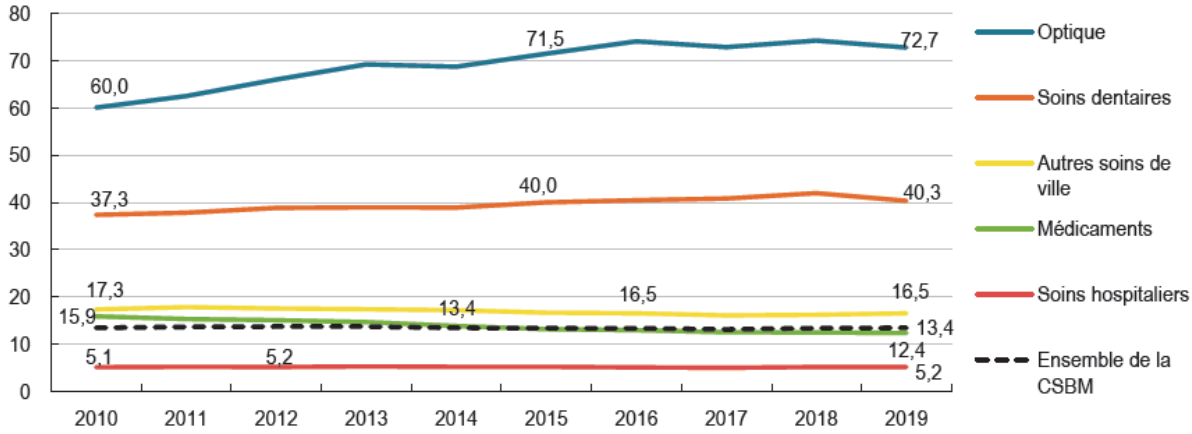


Figure 6: Part en % versée par les organismes complémentaires pour les principaux postes

Nous constatons effectivement qu'en 2019, les OC ont financé :

- 72,7% de la consommation en optique. Un pourcentage qui est en progression continue sur les dix dernières années ;
- 40,3% des soins dentaires avec une quasi-stabilité sur les cinq dernières années ;
- 5,2% des soins hospitaliers et sans évolution notable depuis dix ans.

En résumé, ces organismes assument une part importante des dépenses optiques et dentaires (part qui est amenée à croître sous l'effet de la mise en place du RAC 0) et une part faible à très faible pour les autres postes.

Enfin, l'activité des d'OC sur le marché de la complémentaire santé a connu quelques évolutions lors des dix dernières avec :

- Une progression de la part des contrats collectifs qui passe de 41% en 2011 à 48% en 2019 comme le montre le graphique suivant :

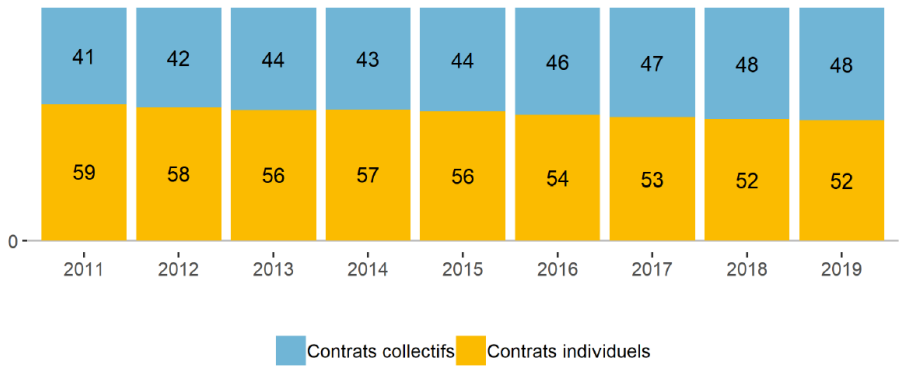


Figure 7: parts des contrats collectifs et contrats individuels en % des cotisations collectées

- Des sociétés d'assurance qui enregistrent au niveau des cotisations collectées, la croissance la plus forte depuis 2015 (5.8% en 2019 contre 2% pour les mutuelles), ce qui leur permet de gagner des parts de marché au détriment des mutuelles. Le graphique et le tableau ci-dessous illustrent bien ces deux aspects :

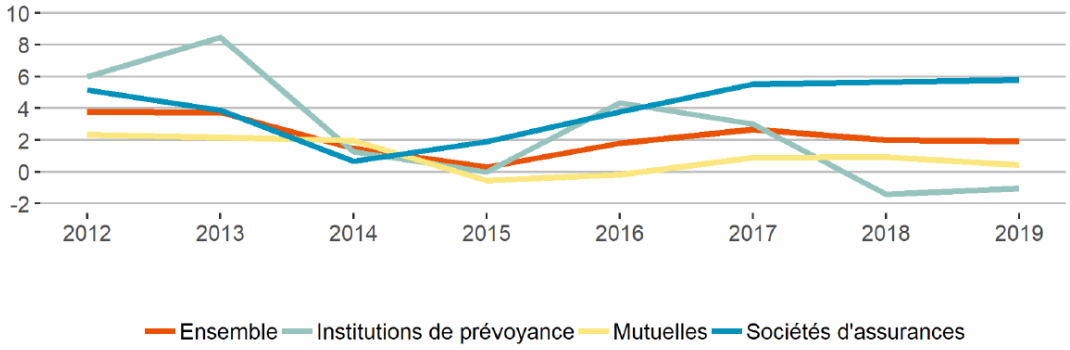


Figure 8: % d'évolution des cotisations collectées entre 2012 et 2019 par type d'OC

	2001	2006	2012	2013	2014	2015	2016	2017	2018	2019
Mutuelles	60	59	55	54	53	53	52	51	51	50
Sociétés d'assurances	19	24	27	28	28	29	30	31	32	33
Institutions de prévoyance	21	17	18	18	19	18	18	18	17	17

Tableau 2: parts en % des cotisations collectées par type d'OC

Reste à charge et le 100% santé

En France, il existe deux définitions du reste à charge :

- La première adoptée par la Cour des comptes, l'IRDES[2] et l'Assurance maladie elle-même considèrent que le reste à charge (dit public) correspond à la part des dépenses de soins non prise en charge par l'assurance maladie obligatoire dans la limite des tarifs de responsabilité qu'elle a fixés. Ce reste à charge a atteint 42.2 milliards d'euros en 2019, soit 20.3% de la CSBM. A titre de comparaison, ce reste à charge était de 42.9 milliards d'euros en 2016, soit 21,6 % de la CSBM ;
- La deuxième définition qui est utilisée par la DREES (et sera utilisée dans la suite de ce mémoire), considère que le reste à charge (dit cette fois-ci individuel) est la part des dépenses de santé assumée directement par l'assuré après intervention de l'assurance maladie obligatoire, de l'état et des organismes complémentaires. En 2019, cette part représentait 6.9% de la CSBM soit 14.3 milliards d'euros. A titre de comparaison, ce reste à charge représentait en 2016, 8.3 % de la CSBM soit 16.5 milliards d'euros.

Ces restes à charges sont le résultat de mécanismes propres au fonctionnement de la sécurité sociale. Des mécanismes qui existent depuis la création de cet organe et qui se sont multipliés à travers le temps. Nous pouvons citer par exemple : le ticket modérateur qui existe depuis la création de la sécurité sociale (art. 24, ordonnance du 19 octobre 1945) et qui correspond à la part de soins non pris en charge par l'assurance maladie (30 % pour les honoraires en médecine ambulatoire, 20 % pour les frais d'hospitalisation, 15 %, 35 % ou 65 % pour les médicaments selon le service médical rendu …), le forfait hospitalier qui représente la participation financière du patient aux frais d'hébergement et d'entretien entraînés par son hospitalisation, le déremboursement de certains médicaments qui ne

[2] Institut de recherche et documentation en économie de la santé, créé le 30 janvier 1985

seront plus pris en charge par la Sécurité Sociale, les dépassements d'honoraires dont la prise en charge par les OC est encadrée pour les contrats responsables voire limitée dans certains cas.

Longtemps considéré comme instrument de régulation des dépenses de santé, il commence à interpeler de plus en plus les classes politiques car malgré un niveau bas comparé à d'autres pays de l'OCDE (28% de la CSBM en Espagne en 2015 contre 7% en France par exemple), il peut varier fortement d'un type de soin à l'autre comme le montre le graphique suivant :

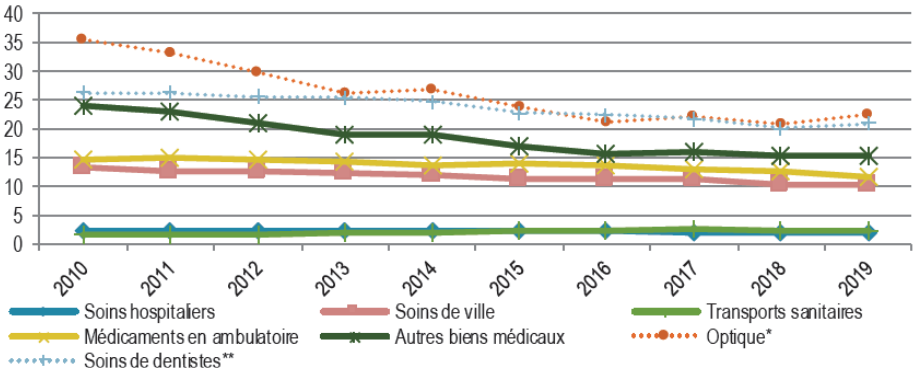


Figure 9 : Reste à charge en % de la CSBM par poste de soins entre 2010 et 2019

Contrairement aux transports sanitaires ou soins hospitaliers, le reste à charge représente une part significative des soins dentaires et optiques (soins généralement très onéreux) qui dépasse les 20% de la CSBM. En 2020, cette part est en baisse pour presque toutes les catégories de soins sauf pour l'optique et le dentaire. Cela a pour effet, une baisse de la part du reste à charge dans le financement de la CSBM qui passe de 7.1% en 2018 à 6.9% en 2019. Les facteurs de cette variation sont résumés dans le graphique suivant :

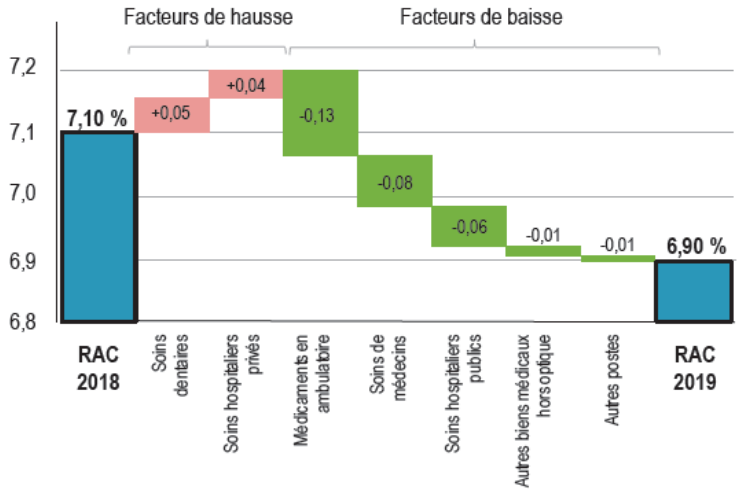


Figure 10 : Décomposition de la baisse du reste à charge en 2019 par poste

Nous pouvons naturellement nous 'interroger sur l'impact du reste à charge sur l'accès aux soins, surtout pour les populations précaires disposant d'un faible niveau de revenu et ne pouvant pas financer une complémentaire santé avec des niveaux élevés de couverture. Pour remédier à ce problème, plusieurs dispositifs ont été mis en place comme par exemple : le système des ALD qui limite le reste à charge des personnes atteintes de certaines pathologies reconnues comme affections de

longue durée ou le 100% santé qui garantit une prise en charge totale des dépenses engagées dans le cadre de certains soins.

Présenté comme un dispositif garantissant des soins pour tous et 100% pris en charge, il s'applique à certaines catégories des soins optiques, dentaires et audiologie. Son principal objectif est de réduire le taux de renoncement aux soins pour des raisons financière (il est de 10.1% pour les équipements d'optique et de 17% pour les soins dentaires concernés), tout en proposant des prestations de soins répondant aux besoins de santé nécessaires avec une garantie de qualité et dont peut bénéficier tout adhérent à une complémentaire santé. Il prévoit :

- En optique à partir de 2020 pour les assurés disposant d'une prescription médicale :
  - L'accès à des montures sans reste à charge avec une valeur inférieure ou égale à 30€ ;
  - L'accès à des verres dits de classe A qui sont 100% pris en charge et qui répondent à tous les besoins de corrections visuelles et garantissent un niveau de qualité élevé tant sur le plan esthétique que sur le plan des performances techniques. Les autres verres classés dans la catégorie B sont à tarif libre.
- En soins dentaires :
  - Un panier 100% santé incluant des prothèses dentaires avec zéro reste à charge ;
  - Un panier aux tarifs maitrisés avec des prothèses dont les prix sont plafonnés. Les assurés optant pour ce panier pourront avoir un reste à charge modéré selon le niveau de prise en charge inclus dans leurs contrats de complémentaire santé ;
  - Un panier à tarif libre.

  Les prothèses composant ces paniers seront disponibles de manière progressive à différente dates :
  - Au 1 Janvier 2020 : 8 prothèses dites "fixes" (couronnes et bridges) dans le panier 100% santé 6 dans le panier aux tarifs maitrisés ;
  - Au 1 Janvier 2021 : 50 autres prothèses dites "fixes" (couronnes et bridges) et amovibles (dentiers) entreront dans le panier 100% santé et 4 dans le panier aux tarifs maitrisés ;
  - Au 1 Janvier 2022 : 57 prothèses dans le panier aux tarifs maitrisés.
- En aides auditives :
  - Des aides auditives de classe 1 qui composent le panier 100% santé ;
  - Des aides auditives à tarifs libres.

  La prise en charge de ces aides de classe 1 augmentera progressivement à compter de 01 Janvier 2019 avec des tarifs plafonnés jusqu'à 2021 avec une prise en charge totale.

AESIO Mutuelle

Chiffres et dates clés

AESIO mutuelle est le fruit de la fusion de trois entités mutualistes : APRÉVA Mutuelle, ADREA Mutuelle et EOVI-MCD Mutuelle qui formaient jusqu'au 31/12/2020 le Groupe AÉSIO, union mutualiste de groupe créé en 2016. Cette fusion a été précédée par plusieurs étapes durant la période allant de 2016 à 2020, principalement marquée par le transfert du grand collectif en interlocution commerciale AESIO avec une affection à l'une des trois mutuelles en fonction des modalités de gestion et du département du siège social de l'entreprise couverte. Cinq dates clés résument ces étapes :

- 5 juillet 2016 : création de l'Union Mutualiste de Groupe AÉSIO, visant à une collaboration entre ADREA Mutuelle, APRÉVA Mutuelle et EOVI-MCD Mutuelle, pour proposer des offres auprès des grandes entreprises et des branches professionnelles ;
- Juin 2018 : Validation par les adhérents du projet de fusion ;
- Juin 2020 : vote de la fusion du groupe par les Assemblées générales des 3 mutuelles ainsi que celle de la Mutuelle des Anciens de Natixis (MAN) ;
- 31 décembre 2020 : les trois entités ADREA Mutuelle, APRÉVA Mutuelle et EOVI-MCD Mutuelle au sein de l'UMG ont fusionné pour donner naissance à une entité unique AÉSIO Mutuelle ;
- Janvier 2021 : fusion opérationnelle et politique effective de l'entité AÉSIO Mutuelle.

Fort de ces 2.9 M de personnes protégées et ces 1.9 Milliards d'euros de chiffre d'affaires, AESIO occupe aujourd'hui la cinquième place des acteurs mutualistes en santé, prévoyance et assurance individuelle et collective en France. Elle compte aussi 3566 collaborateurs et 281 agences héritées des mutuelles fondatrices qui lui permettent de couvrir la majorité du territoire français comme le montre la carte ci-dessous :

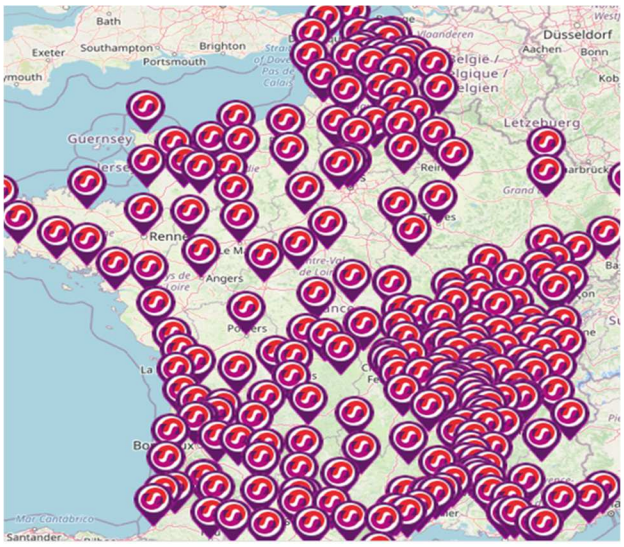


Figure 11: Implantation géographique des agences AESIO mutuelle

Un autre héritage concernant les activités historiques de ces mutuelles, fait qu'à AESIO mutuelle opère sur quatre branches : la branche 21 (Nuptialité-natalité), la branche 1 (Accidents), la branche 2 (Maladie) et enfin la branche 20 (Vie-Décès) grâce aux agréments administratifs dont elle dispose, conformément à l'article R 321-1 du Code des Assurances.

Parallèlement à cette fusion, la mutuelle a également entamé un rapprochement avec le groupe MACIF, ce qui a permis la création d'Aéma groupe et de l'UMG AÉSIO MACIF. Les liens entre les différentes entités sont résumés dans le schéma ci-dessous :

Figure 12: Les différentes entités d'Aéma groupe

Les principales composantes de ce groupe sont :

- La SGAM Aéma groupe, tête de groupe garante du pilotage prudentiel et économique ;
- La SAM Macif, pilote des métiers de l'IARD, assurance vie et finance-épargne ;
- L'UMG AÉSIO MACIF pole spécialisé dans les métiers de la santé et de la prévoyance individuelle et collective.

En termes de fonds propres AESIO dispose de 1.5 Milliards euros pour couvrir ses besoins de solvabilité. Valorisée à l'aide de la formule standard de la réglementation prudentielle Solvabilité II, l'entreprise affiche en 2019 et 2020, les ratios de couverture suivants :

Evolutions 2019/2020 (en K€)	Groupe AÉSIO 31/12/2020 (1)	AESIO Mutuelle 31/12/2020 (2)	Groupe AÉSIO 31/12/2019 (3)
Capital de solvabilité requis (SCR)	632 305	630 842	588 841
Fonds propres Solvabilité 2	1 643 423	1 644 022	1 582 973
Ratio de couverture du SCR	**260 %**	**261 %**	**269 %**
Ratio de couverture du MCR	**1022 %**	**1042 %**	**1051 %**

Tableau 3 : Fonds propres et ratio de couverture en 2019 et 2020

Le ratio de couverture du SCR qui est de 260% en 2020 est en légère baisse comparé à 2019, ceci s'explique par une hausse de 7 % du capital de solvabilité requis. Cette hausse est due à :

- La hausse du SCR Défaut des contreparties faisant suite au remboursement du fonds de solidarité au profit du Groupe AÉSIO et de son affectation en avoirs en banque ;
- La hausse du SCR souscription santé due à la prise en compte du SCR rachat.

Le profil de risque lui reste classique et se résume comme suit :

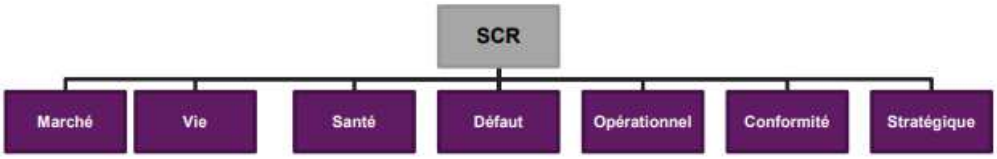


Figure 13 : Profil de risque du groupe AESIO

Le risque vie est lié à des garanties indemnitaires ou viagères d'allocation de frais d'obsèques. Le groupe est faiblement exposé à ce risque.

Présentation du portefeuille

AESIO intervient principalement sur le risque santé aussi bien sur le segment individuel que collectif. Les risques liés à la vie représentent une activité complémentaire mais marginale et ne pèse que 1.4% du chiffre d'affaires net de réassurance. Le tableau suivant récapitule le nombre de personnes protégées ces trois dernières années :

Nombre de personnes protégées	*PP Assurées au 31/12/20*	*PP Assurées au 31/12/19*	*PP Assurées au 31/12/18*
Effectif individuel santé (y.c. TNS)	1 521 391	1 601 783	1 666 319
Effectif Collectif santé	1 400 380	1 364 481	1 363 352
Total santé	2 921 771	2 966 264	3 029 671

Tableau 4 : Nombre de personnes protégées par segment

La répartition segment suit la tendance du marché avec une augmentation du segment collectif au détriment du segment individuel :

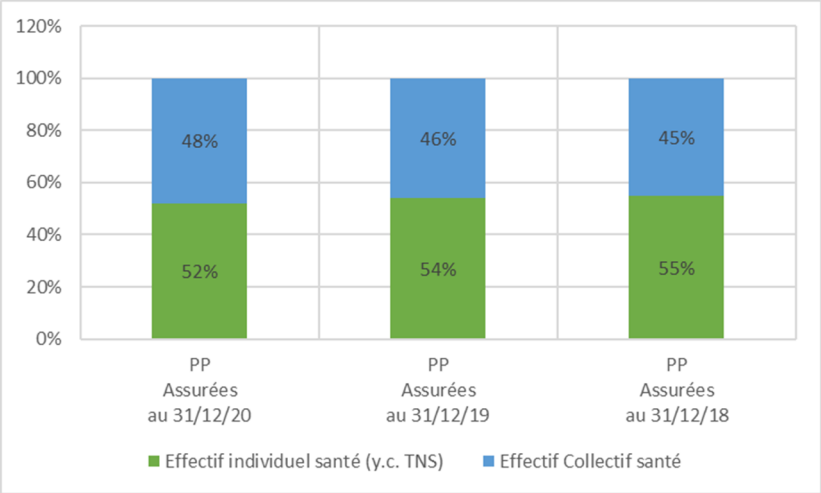


Figure 14: Répartition en % du nombre de personnes protégées

L'analyse des cotisations par nature du risque permet de confirmer les éléments précédents comme le montre le tableau suivant (dont les éléments proviennent du SFCR 2020) :

	Cotisations acquises nettes de réassurance		*Charge des prestations nettes de réassurance*		*Marge Brute*	
En K€	*31/12/2020*	*31/12/2019*	*31/12/2020*	*31/12/2019*	*31/12/2020*	*31/12/2019*
Vie	24 786	22 660	18 875	18 565	5 911	4 095
Non vie	1 714 924	1 760 030	1 388 026	1 462 100	326 898	297 930
Total	1 739 710	1 782 690	1 406 901	1 480 665	332 809	302 025

Tableau 5 : Cotisations, charge de prestations et marge brute par nature du risque

Les cotisations liées au risque non-vie qui correspond au risque santé, constituent la quasi-totalité du chiffre d'affaires. Un chiffre d'affaires en légère baisse en 2020 comparé à 2019 à l'image du nombre de personnes protégées. La marge brute, elle, est en amélioration. Une marge qui provient essentiellement du segment individuel malgré son recul dans le portefeuille.

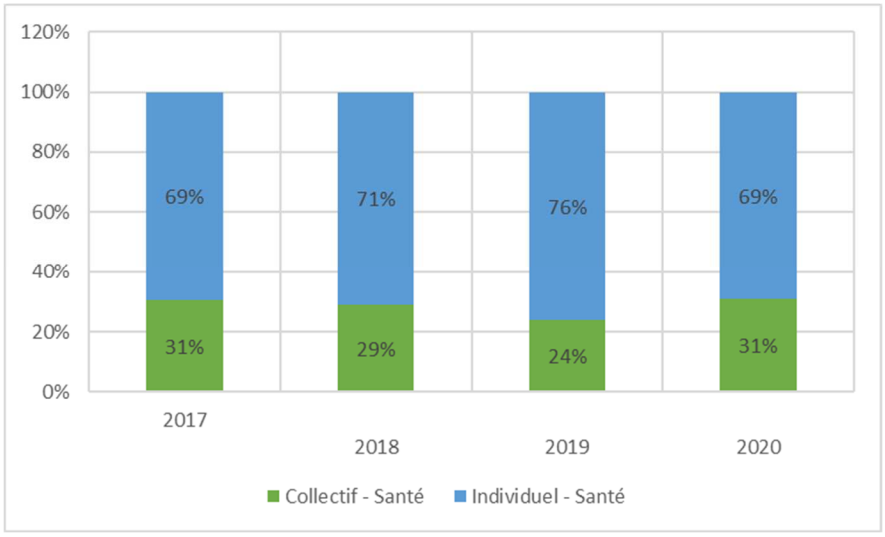


Figure 15 : Répartition en % de la marge brute par segment entre 2017 et 2020

1 Données

L'implantation géographique du réseau commercial d'AESIO Mutuelle lui permet de couvrir une grande partie du territoire français et de commercialiser ses produits composés principalement de garanties « frais de soins », auprès d'une large population. Un volume important des données relatives aux contrats souscrits est collecté et stocké, dans le respect de la réglementation, dans différents décisionnels. Les collaborateurs de la mutuelle disposent donc, dans le cadre de leurs missions, d'une considérable quantité de données. Toutefois, aux vues des capacités informatiques disponibles et des temps de traitements nécessaires, il s'avère très compliqué d'exploiter l'ensemble des données du portefeuille pour réaliser des études. De ce fait, il a été décidé de restreindre l'analyse à la gamme FLEX'ADREA qui est l'une des gammes phares commercialisée par la mutuelle.

1.1 Base de données

Toutes les données utilisées dans le cadre ce mémoire sont issues du datamart actuariat. Ce dernier est alimenté mensuellement avec les données de production provenant des outils de gestion et les données GRC qui est l'outil de la relation client. Plusieurs étapes sont réalisées avant la mise à disposition de ces flux. En effet, une première étape consiste a réalisé plusieurs transformations afin de disposer de données normées. Par la suite, des contrôles de premier niveau sont effectués pour vérifier la qualité des données constituées, suivi de contrôles de second niveau pour s'assurer de la cohérence des données. Les deux premières étapes sont réalisées par les équipes informatique alors que la dernière étape est de la responsabilité de la cellule data au sein de la direction actuariat. C'est une fois tous ces traitements effectués avec succès que les équipes actuariat peuvent exploiter la base en question.

1.1.1 Constitution de la base

Pour les besoins de ce mémoire, plusieurs extractions ont été effectuées depuis cet entrepôt. Cette étape nous a permis de disposer d'un ensemble de données disponibles nativement.

En tout, trois types de données ont été extraites (sous format csv) :

- Les données des effectifs, ce qui nous permet de disposer de plusieurs informations sur l'assuré : numéro d'adhérent, date de naissance, date d'adhésion, date de radiation, code postal…
- Les données des cotisations, qui permettent de récupérer le montant de la prime ainsi que le produit et l'exercice auquel la prime est rattachée. Il s'agit des données agrégées au niveau de l'exercice et du produit ;
- Les données des prestations, qui contiennent le montant de la prestation, l'acte de soin, la date de soin, la personne ayant eu recours à ce soin, la garantie couvrant ce soin.

Les données utilisées dans la suite, ont été extraites au 30/08/2021. Les prestations prises en compte sont les prestations payées au plus tard à cette date. L'analyse des coefficients de développement des triangles de paiements indique une quasi-stabilité à partir du milieu dès le 6ème mois de la deuxième année comme le montre le tableau suivant (les colonnes correspondent au mois de développement au-delà du 31/12/N) :

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2016	100,90%	100,37%	100,25%	100,13%	100,10%	100,08%	100,09%	100,06%	100,07%	100,11%	100,04%	100,03%	100,03%	100,01%	100,01%
2017	100,53%	100,38%	100,16%	100,13%	100,14%	100,13%	100,26%	100,23%	100,06%	100,03%	100,03%	100,01%	100,03%	100,01%	100,01%
2018	100,57%	100,34%	100,17%	100,17%	100,05%	100,06%	100,04%	100,05%	100,04%	100,03%	100,03%	100,02%	100,01%	100,01%	100,01%
2019	100,58%	100,27%	100,13%	100,08%	100,15%	100,15%	100,05%	100,08%	100,08%	100,03%	100,12%	100,03%	100,05%	100,02%	100,00%
2020	100,64%	100,35%	100,19%	100,15%	100,13%	100,06%	100,07%	100,00%							

Tableau 6 : Coefficients de développement des paiements

Les données extraites ont été par la suite importées dans le logiciel SAS. Plusieurs retraitements ont été effectués à l'aide de cet outil, afin de les enrichir et de mieux les structurer. Cette étape a pour principal but de rendre les données facilement exploitables et de gagner en lisibilité.

Ce travail permettra, par exemple, de disposer d'une vision des effectifs par exercice en lecture directe, d'associer un âge à une tranche d'âge, de disposer de l'âge de l'adhérent à la fin de chaque exercice, d'associer un produit à un niveau de gamme, de calculer le nombre moyen de personnes protégées, de formater certaines variables, de supprimer les mouvements d'encaissement ou de décaissement ayant donné lieu à des écritures d'annulation…

Un dictionnaire des données contenant une description succincte de chaque variable est proposé en annexe1.

1.1.2 Qualité des données

Une fois la base constituée, il était nécessaire de vérifier la qualité des données disponibles. En effet, l'utilisation des données de mauvaise qualité peut altérer la pertinence des études et conduire à des conclusions erronées.

A ce titre, les données effectifs et prestations, ont été soumises à deux types de contrôle :

- Des contrôles de premier niveau qui visent principalement à s'assurer que les données respectent le format attendu et qu'elles ne présentent ni doublons ni de valeurs manquantes ;
- Des contrôles de cohérence dont le but est de vérifier que la base ne contient pas de données aberrantes (par exemple : que les adhérents n'ont pas une date de radiation antérieure à la date d'adhésion).

Concernant les données de cotisations, seul un rapprochement avec les différents reportings existants a été effectué pour s'assurer de leur cohérence. En effet, l'extraction de ces données a été réalisée de manière agrégée et le niveau de granularité ne permet pas de réaliser des contrôles plus fins. Des contrôles qui ne sont pas par ailleurs nécessaires et dont l'absence n'impactera pas la qualité de l'étude et des conclusions.

1.1.2.1 Contrôles de premier niveau

La majorité des contrôles réalisés à cette étape se sont avérés satisfaisants. En effet, les données respectent le format attendu et aucun doublon n'a été identifié. Le contrôle des données manquantes nous a permis de détecter les anomalies suivantes :

❖ ***Effectifs :***

Le tableau si dessous résume le résultat du contrôle :

Obs.	Variables	Nombre d'observations	NB valeurs manquantes	% valeurs manquantes
1	ANNEE_MOIS_RAD_GAR	1097367	487983	44.47%
2	ANNEE_RAD_GAR	1097367	487983	44.47%
3	DATE_RAD_GAR	1097367	487983	44.47%
4	MOIS_RAD_GAR	1097367	487983	44.47%
5	ANNEE_ANC	1097367	144285	13.15%
6	ANNEE_MOIS_ANC	1097367	144285	13.15%
7	DATE_ANCIENNETE	1097367	144285	13.15%
8	MOIS_ANC	1097367	144285	13.15%
9	AGE_ARD	1097367	174	0.02%
10	AGE_EX	1097367	174	0.02%
11	ANNEE_MOIS_NAISS	1097367	174	0.02%
12	ANNEE_NAISS	1097367	174	0.02%
13	DATE_NAISS	1097367	174	0.02%
14	MOIS_NAISS	1097367	174	0.02%
15	ANNEE	1097367	0	0.00%
16	ANNEE_ADH_GAR	1097367	0	0.00%
17	ANNEE_MOIS_ADH_GAR	1097367	0	0.00%
18	CODE_DEPT	1097367	0	0.00%
19	CODE_GARANTIE	1097367	0	0.00%
20	DATE_ADH_GAR	1097367	0	0.00%
21	DT_DEB_EX	1097367	0	0.00%
22	DT_FIN_EX	1097367	0	0.00%
23	MOIS_ADH_GAR	1097367	0	0.00%
24	NB_MOIS_ANC	1097367	0	0.00%
25	NB_PP	1097367	0	0.00%
26	NB_PP_MOY	1097367	0	0.00%
27	NIVEAU_GAM	1097367	0	0.00%
28	NIVEAU_GAR	1097367	0	0.00%
29	NUM_PP	1097367	0	0.00%
30	PRODUIT	1097367	0	0.00%
31	SEXE	1097367	0	0.00%
32	TRANCHE_AGE_1	1097367	0	0.00%
33	TRANCHE_AGE_2	1097367	0	0.00%
34	TYPE_PP	1097367	0	0.00%
35	TYPE_PRODUIT	1097367	0	0.00%

Tableau 7 : Données manquantes par variable de la table effectifs

Les données manquantes concernent :

- Des variables calculées à partir de la date d'ancienneté : pour pallier cette anomalie, cette date sera recalculée à partir de la plus ancienne date d'adhésion disponible dans l'historique. La date d'adhésion étant toujours renseignée comme le montre le tableau ci-dessus, ce correctif permet de disposer de la vraie date d'ancienneté ;
- Des variables calculées à partir de la date de radiation : ce champ n'est effectivement renseigné que pour les adhérents ayant résiliés leurs contrats. L'ensemble des cas rencontrés correspondent à des adhérents avec des contrats encore en cours. Ces cas ne constituent donc pas une anomalie ;
- Des variables calculées à partir de la date de naissance :l'unique impact que peut avoir l'absence de la date de naissance concerne le calcul d'âge. Dans le peu de cas rencontrés, cet âge sera remplacé par l'âge moyen des adhérents ayant les mêmes critères (même produit, même catégorie de bénéficiaire, même sexe…) .

❖ ***Prestations :***

Comme le montre le tableau suivant les prestations ne présentent aucune donnée manquante :

bs.	Variables	Nombre d'observations	NB valeurs manquantes	% valeurs manquantes
1	ACTE_RC	12738109	0	0.00%
2	ACTE_RO	12738109	0	0.00%
3	ANNEE	12738109	0	0.00%
4	CODE_GARANTIE	12738109	0	0.00%
5	DT_DEB_SOIN	12738109	0	0.00%
6	DT_FIN_SOIN	12738109	0	0.00%
7	DT_PAIEMENT	12738109	0	0.00%
8	DT_TRAITEMENT	12738109	0	0.00%
9	LIB_ACTE_RC	12738109	0	0.00%
10	MT_AUTRE_MUT	12738109	0	0.00%
11	MT_DEPA	12738109	0	0.00%
12	MT_DEPE_ASSURE	12738109	0	0.00%
13	MT_RC	12738109	0	0.00%
14	MT_RO	12738109	0	0.00%
15	MT_TICKET_MOD	12738109	0	0.00%
16	NUM_PP	12738109	0	0.00%
17	PRODUIT	12738109	0	0.00%
18	RISQ_RC	12738109	0	0.00%
19	TX_RC	12738109	0	0.00%
20	TX_RO	12738109	0	0.00%

Tableau 8 : Données manquantes par variable de la table prestations

Enfin, un dernier traitement est appliqué pour ne retenir que les prestations rattachées à des adhérents présents dans l'effectif à la date des soins.

1.1.2.2 Contrôles de cohérence

Pour s'assurer de la cohérence des données dont nous disposons, plusieurs contrôles et traitements vont être effectués.

Effectifs :

- Le 1er contrôle concerne la table des effectifs et permet de vérifier que les dates des adhésions et des radiations sont cohérentes entre elles. Il se déroule comme suit :
  - Recherche des données pour lesquelles la condition n'est pas vérifiée. Cette recherche nous permet d'identifier 6146 cas qui présente cette anomalie ;
  - Analyse des données identifiées. Après vérification, il s'avère qu'il s'agit des erreurs de saisie qui sont rectifiées par la suite comme le montre l'exemple suivant :

Obs.	NUM_PP	TYPE_PP	CODE_GARANTIE	PRODUIT	SEXE	CODE_DEPT	DATE_ADH_GAR	DATE_RAD_GAR
1	20492	Chef de famille (assuré principal)	YF33	YF33	Homme	39	01/06/2013	31/07/2013
2	20492	Chef de famille (assuré principal)	YF23	YF23	Homme	39	01/06/2013	31/05/2013

Tableau 9 : Exemple d'anomalies

La 1ère ligne correspond à une adhésion au produit YF23 le 01/06/2013 pour une radiation le 31/05/2013 (ligne détectée en anomalie) et la 2ème ligne au correctif avec une adhésion le 01/06/2013 au produit YF33 pour lune radiation le 31/07/2013.

Tenant compte des résultats de ce contrôle, toutes les données identifiées vont être supprimées de la base sans risque de perte d'informations.

- Le 2ème contrôle concerne la variable âge et vise à détecter des valeurs aberrantes. Pour rappel, une valeur aberrante est liée soit à une erreur de mesure (dans notre cas erreur de saisie ou

problème informatique), soit à une loi de probabilité à queue lourde. Cette étude consistera à détecter uniquement les erreurs de mesure ;

La première approche pour identifier ces valeurs appelées aussi Oultliers, est l'utilisation de la règle 1,5 x écart interquartile.

Pour rappel :

- Soit X une variable aléatoire. X est $\mathrm{p}-$ quantile de X si $\mathrm{P}[\mathrm{X} \leq \mathrm{x}]=\mathrm{p}$.
- Q_1 correspend à un $\mathrm{p}=25\% \rightarrow \mathrm{P}[\mathrm{X} \leq \mathrm{Q}_1]=25\%$.
- Q_3 correspend à un $\mathrm{p}=75\% \rightarrow \mathrm{P}[\mathrm{X} \leq \mathrm{Q}_3]=75\%$.
- Ecart interquartile $=\mathrm{Q}_3-Q_1$.

La règle se présente sous le format suivant :

Soit :

- ✓ n le nombre d'observations ;
- ✓ $(X_1, X_2, \ldots, X_n)$ n réalisations de la variable aléatoire X ;
- ✓ α l'écart interquartile ;
- ✓ $A_1 = Q_1 - 1.5 * \alpha$ Et $A_2 = Q_3 + 1.5 * \alpha$;
- ✓ A l'ensemble de valeurs aberrantes.

A est définie donc comme suit :

$$A = \{X_i \quad \backslash 1 \leq \mathrm{i} \leq \mathrm{n} \text{ et } X_i \notin [A_1, A_2]\}.$$

Le tableau ci-dessous résume les valeurs de : Q1, Q3, P1 (quantile à 1%), P99 (quantile à 99%), A1, A2 et de la fréquence par catégorie de bénéficiaire :

Obs.	ANNEE	TYPE_PP	Fréquence	Quantile à 1%	Quantile à 99%	1er quartile	3ème quartile	A1	A2
1	2020	Autre bénéficiaire (exemple : ascendant)	53	7	62	20	51	-26,5	97,5
2	2020	Chef de famille (assuré principal)	116061	21	89	36	67	-10,5	113,5
3	2020	Conjoint	21187	25	86	49	70	17,5	101,5
4	2020	Enfant	37690	0	26	5	17	-13	35

Tableau 10 : Quantiles par catégorie de bénéficiaire

A l'analyse de ce tableau, nous pouvons conclure rapidement que cette méthode n'est pas efficace. En effet, l'âge d'un chef de famille sera considéré comme valeur aberrante s'il est hors intervalle [-10.5 ; 113.5]. Ce problème est dû à une importante dispersion de la variable âge.

Pour remédier à ce problème nous allons nous appuyer - dans notre analyse- sur les quantiles à 1% et 99% (notés respectivement P1 et P99) tout en déroulant les étapes suivantes :

a. Sélection des adhérents dont l'âge est inférieur à P1 ou supérieur P99 ;
b. Calcul des P1, P5 (quantile à 5%), P10 (quantile à 10%), Q1, Q3, P95 (quantile à 95%) et P99 pour cette sélection ;
c. Analyser les différentes valeurs obtenues :

ANNEE	TYPE_PP	Fréquence	Quantile à 1%	Quantile à 5%	Quantile à 10%	1er quartile	3ème quartile	Quantile à 95%	Quantile à 99%
2020	Chef de famille (assuré principal)	1869	4	12	18	20	93	98	101
2020	Conjoint	382	19	21	22	23	89	94	97
2020	Enfant	324	27	27	27	27	29	30	37

Tableau 11 : Quantiles par catégorie de bénéficiaire

Cette sélection comporte :

- ✓ 10% des adhérents avec un statut « chef de famille », soit 187 individus, ont un âge inférieur à 18 ans et un peu plus de 90 individus ont un âge inférieur à 12 ans ;

- ✓ 5% des adhérents ayant un statut « enfant », soit 16 individus, ont un âge supérieur à 30 ans et 3 ont un âge supérieur 37. Ces cas ne constituent pas pour autant des valeurs aberrantes. En effet, dans des situations très rares, un enfant peut être rattaché à la mutuelle de ses parents de fait de son état de santé par exemple.

Pour les cas concernant les « chef de famille », l'hypothèse que le type de bénéficiaire a été correctement saisi sera faite et l'âge sera remplacé par l'âge moyen constaté chez les individus disposant des mêmes profils (même département, même garantie...).

Prestations :

Les données de cotisations ont été extraites à titre d'information et ne vont pas servir dans la modélisation. De ce fait, ces éléments ne vont pas être concernés par les contrôles de cohérence.

Concernant les données prestations, seules celles en lien avec des actes dentaires ou optiques vont être contrôlées. En effet, l'objet de ce mémoire concerne uniquement ces deux types de prestations :

- Le 1er contrôle vise à s'assurer que la somme des montants des différents remboursements dont peut bénéficier l'adhérent ne dépasse pas le montant de la dépense.
  Soit :
  - ✓ MT_d Montant de la dépense ;
  - ✓ MT_{RO} Montant remboursé par le régime général, MT_{RC} Montant remboursé par régime complémentaire et $MT_{R.Autre}$ Montant par d'autre mutuelle.

  La règle serait donc : $MT_d \geq MT_{RO} + MT_{RC} + MT_{R.Autre}$.

  Aucun cas n'est identifié lors de ce contrôle.

- La suite sera consacrée à la recherche de valeurs aberrantes. Les variables concernées par ce contrôle sont :
  - ✓ MT_d , MT_{RO}, MT_{RC}, MT_{Autre} ;
  - ✓ MT_{TM} Montant du ticket modérateur ;
  - ✓ MT_{Dp} Montant du dépassement.

A titre d'illustration, le paragraphe suivant présentera la méthode utilisée ainsi que les résultats pour le MT_{RC} concernant les soins dentaires.

MT_{RC} Pour les soins dentaires :

Ce paragraphe s'intéresse aux montants remboursés par la mutuelle dans le cadre des soins dentaires. Tenant compte que ce montant dépend fortement du type d'acte avec des variations importantes d'un acte à l'autre, l'analyse s'effectuera par type d'acte.

La codification de ces actes est la suivante :

Code acte	Libellé acte
ADDC	Actes en D et DC
AUT4	Autres - Dentaire
CDEN	Consultations dentaires
IMPL	Implants dentaires
TOAC	Orthodontie acceptée
TORE	Orthodontie refusée
PARO	Parodontologie
SP21	Prothèses dentaires : autres dents ou non précisé RAC modéré
SP20	Prothèses dentaires : autres dents ou non précisé RAC0

Code acte	Libellé acte
SPR2	Prothèses dentaires acceptées : autres dents ou non précisé
SPR1	Prothèses dentaires acceptées : dents visibles
SPRN	Prothèses dentaires non prises en charge
SP11	Prothèses dentaires : dents visibles RAC modéré
SP10	Prothèses dentaires : dents visibles RAC0
RADD	Radiologie dentaire
SCON	Soins conservateurs (SC)
SCHN	Soins conservateurs hors nomenclature
ET04	Soins etrangers - Dentaire

Tableau 12 ; Référentiel des codes actes des soins dentaires

Pour détecter d'éventuelles valeurs aberrantes, nous analyserons, de manière visuelle la dispersion de notre variable par acte de soin. Le graphique suivant regroupe l'ensemble des boites à moustache utilisées à cet effet :

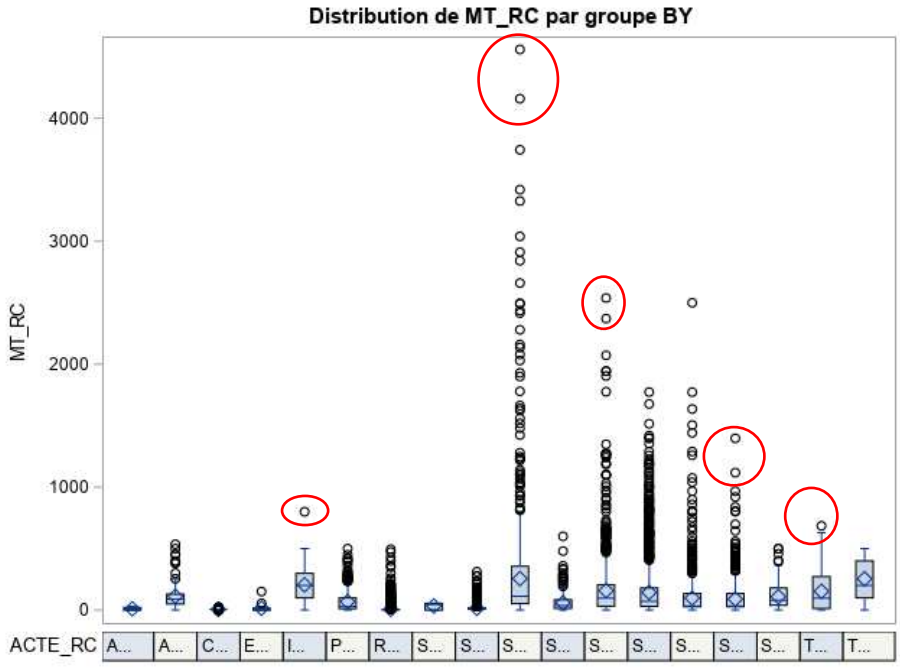


Figure 16 : Boîtes à moustaches de la variable MT_RC par code acte

Les valeurs entourées peuvent être assimilées à des valeurs aberrantes. Pour confirmer ce constat, nous allons aussi analyser les quantiles. Le tableau ci-dessous résume leurs valeurs par acte de soin.

Obs.	Quantile	ADDC	AUT4	CDEN	ET04	IMPL	PARO	RADD	SCHN	SCON	SP10	SP11	SP20	SP21	SPR1	SPR2	SPRN	TOAC	TORE
1	0% Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	5%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.45	0.00	6.45	0.00	0.00	1.00	0.00	0.00	59.00
4	10%	0.00	29.50	2.30	0.00	0.00	0.00	1.20	0.00	0.00	12.90	3.00	6.45	3.00	3.00	3.00	0.00	0.00	90.00
5	25% Q1	0.00	50.00	6.90	0.00	100.00	12.00	2.39	0.00	8.68	53.00	12.90	32.25	30.00	27.00	27.00	40.00	12.89	100.00
6	50% Médiane	6.07	87.71	6.90	1.85	200.00	28.92	4.79	45.38	13.02	112.00	36.00	96.75	70.95	32.25	32.25	80.00	96.75	200.00
7	75% Q3	21.25	131.00	6.90	16.56	300.00	100.00	6.00	55.00	18.28	356.00	85.00	206.00	180.00	135.00	135.00	180.00	272.25	400.00
8	90%	30.36	243.38	6.90	28.88	400.00	200.00	9.58	55.00	27.22	416.00	180.00	340.00	408.50	215.00	215.00	200.00	387.00	500.00
9	95%	30.36	300.00	6.90	54.94	500.00	250.00	13.18	55.00	34.70	832.00	180.00	580.50	559.00	322.50	322.50	300.00	483.75	500.00
10	99%	30.36	453.00	9.00	151.04	500.00	320.00	80.00	55.00	58.80	2080.00	360.00	1269.35	838.50	645.00	559.00	499.93	580.50	500.00
11	100Max 100%	30.36	535.62	24.58	151.04	800.00	500.00	495.00	55.00	312.61	4560.00	600.00	2538.70	1773.75	2500.00	1397.50	500.00	685.75	500.00

Tableau 13: Quantiles par codes actes

Pour presque tous les actes, le quantile à 99% possède une valeur très élevée comparée à la médiane. Par exemple, la médiane de l'acte codifié SP10 (Prothèses pour dents visibles) est de 112 alors que son quantile à 99% est de 2080€ ce qui signifie que 1% de ces actes, soit 135, ont bénéficié d'un remboursement annuel de plus de 2080€. Vu l'importance des quantités et des montants, il parait peu probable qu'il s'agisse d'erreur de saisie (cas à identifier), sachant qu'aucun de ces enregistrements n'a donné lieu à une annulation ou une correction. Nous pouvons conclure à l'absence de données aberrantes dans notre jeu de données. Conclusion qui peut être élargie à l'ensemble des variables pour les deux segments (optique/ dentaire) et ce après avoir obtenu les mêmes résultats en appliquant la même méthode sur l'ensemble du périmètre.

1.2 Présentation de la gamme

FLEX'ADREA représente la principale gamme commercialisée par l'ex mutuelle ADREA. Destinée au marché individuel, elle sera mise sur le marché en 2013. A fin 2020, elle couvrait quelques 149 mille personnes pour un chiffre d'affaires hors taxes de 83,6 M€.

L'une des caractéristiques principales de cette gamme est son aspect modulable, ce qui offre aux adhérents la possibilité de choisir leurs niveaux de garanties. En effet, chaque assuré qui souhaite souscrire à cette gamme se voit proposer six formules. Le client devra procéder à deux choix :

- Le choix de la formule qui définit le niveau de couverture pour les postes hospitalisation / Soins
- Le choix du niveau de couverture pour les postes optique, dentaire et prothèses et appareillages. Il existe là aussi six niveaux, mais en fonction de la formule souhaitée, seuls certains niveaux pourront être sélectionnés. En effet, l'écart entre le 1er et le 2ème choix ne pourra pas dépasser une unité. Le schéma ci-dessous récapitule l'ensemble des 16 combinaisons possibles.

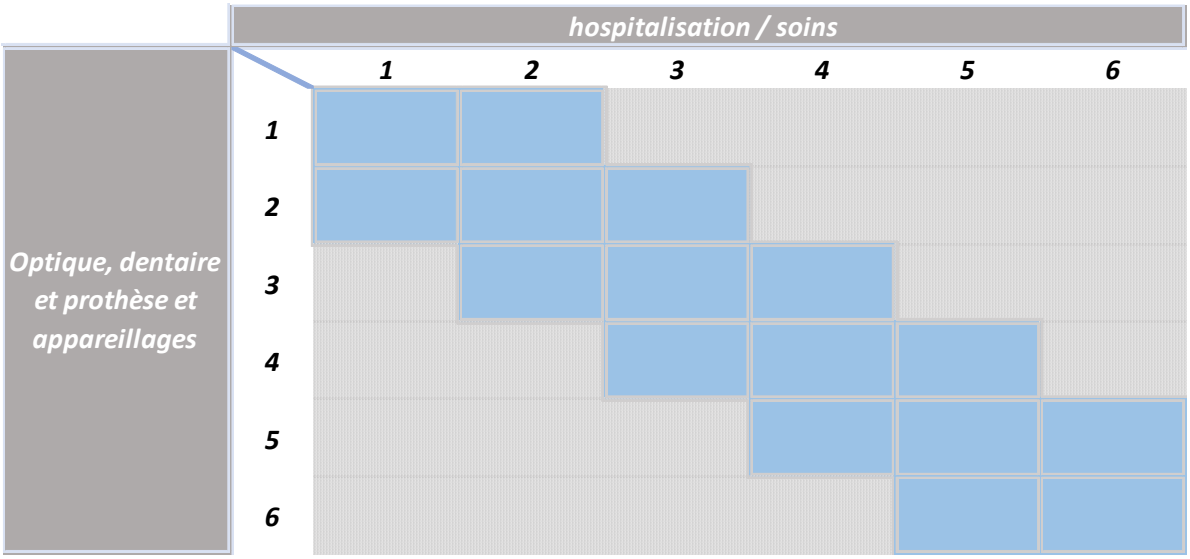


Figure 17: Les différentes combinaisons de la gamme FLEX ADREA

Ainsi, un adhérent qui souscrit à la formule 4, bénéficiera du 4ème niveau de couverture sur les postes hospitalisation / soins et pourra choisir entre les niveaux 3, 4 et 5 pour les postes optique, dentaire et prothèses et appareillages. En plus, il disposera des garanties assistance, pharmacie, tiers-payant et capital maladies redoutées dont les niveaux sont standards et communs à toutes les formules quels que soient les niveaux choisis.

Il est clair que le montant de la cotisation ainsi que le montant des remboursements pour certains actes, dépendront du niveau souscrit. Toutefois, les cotisations, elles, varieront aussi en fonction d'autres paramètres (qui sont listés ci-après) :

- L'âge de la personne protégée ;
- Le régime de rattachement. En effet, cette gamme contient cinq produits destinés aux personnes appartenant aux régimes Alsace Moselle. La codification de ces produits est la suivante : F21AM, F23AM, F34AM, F45AM, F66AM ;
- La zone d'habitation si l'assuré est rattaché au régime général.

Dans la suite de ce mémoire, et par souci de simplification, le niveau de couverture global de chaque assuré sera noté Fxy et appelé produit. Par exemple, si le choix se porte sur la formule 3 avec un niveau 2 pour les postes optique, dentaire et prothèses et appareillages, son niveau de couverture global sera noté F32.

Par ailleurs, le produit constituant un niveau fin d'analyse, un niveau intermédiaire a été mis en place afin de faciliter les analyses. Les différents produits ont été répartis en trois groupes :

- Entrée de gamme : composé des produits F11, F12, F21, F22 et F23 ;
- Milieu de gamme : composé des produits F32, F33, F34, F43, F44 et F45 ;
- Haut de gamme : composé des produits F54, F55, F56, F65, F66.

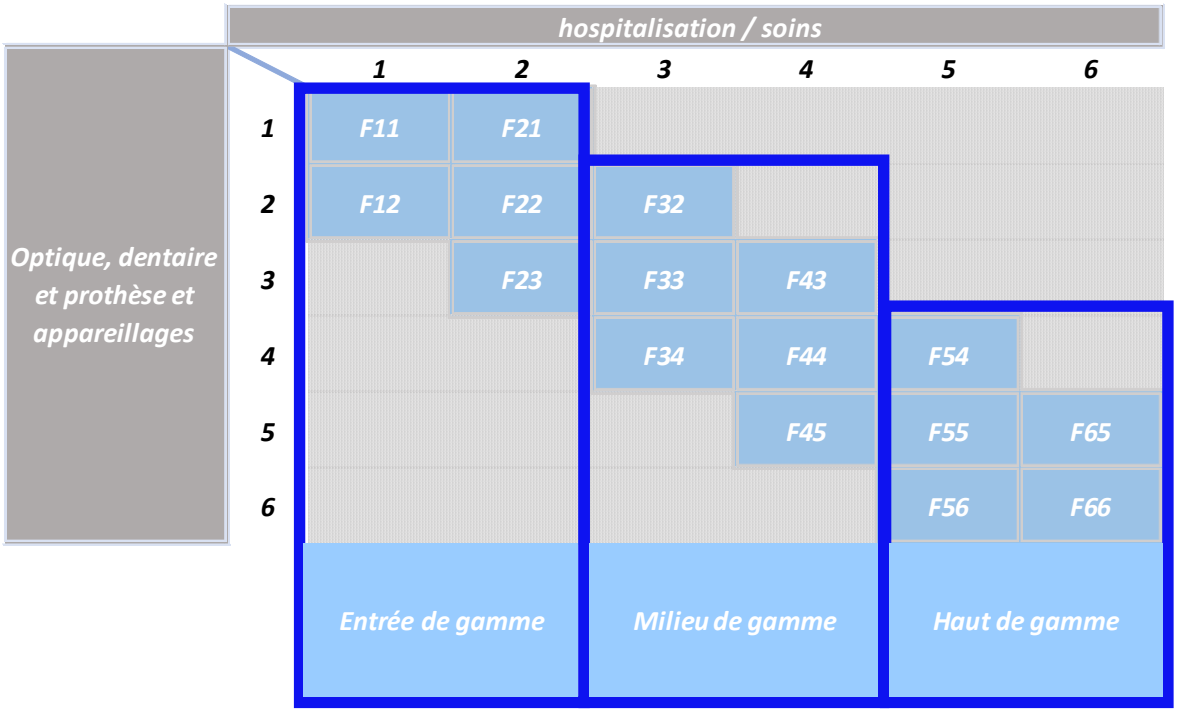


Figure 18 : Produits FLEX ADREA par niveau de gamme

1.2.1 Effectifs et cotisations

Depuis sa commercialisation, la gamme a connu une croissance continue marquée par une forte progression du nombre moyen des personnes protégées les premières années (par exemple, entre 2016 et 2017 sa progression était de +15%), puis un ralentissement les trois dernières années (entre 2019 et 2020 cette progression était de +2%). Le nombre moyen des personnes protégées est un compteur ou chaque individu est pondéré par son temps de présence sur l'exercice (un individu présent toute l'année sera comptabilisé pour 1 alors qu'un individu présent 6 mois sera comptabilisé pour 0.5).

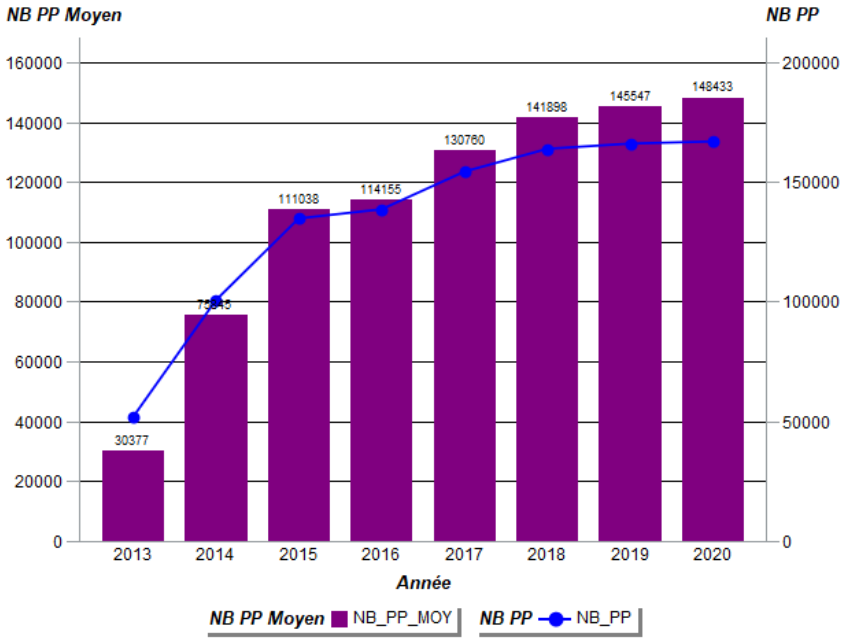


Figure 19 : Evolution du nombre de personnes protégées et du nombre moyen par année

L'analyse de l'évolution du portefeuille par tranche d'âge permet de constater une augmentation continue de la part des personnes dont l'âge dépasse soixante ans. Cette tendance a eu un impact sur la moyenne d'âge qui passe de 37.6 en 2015 à 42.5 en 2016. Les deux graphiques qui suivent, confirment ces deux tendances :

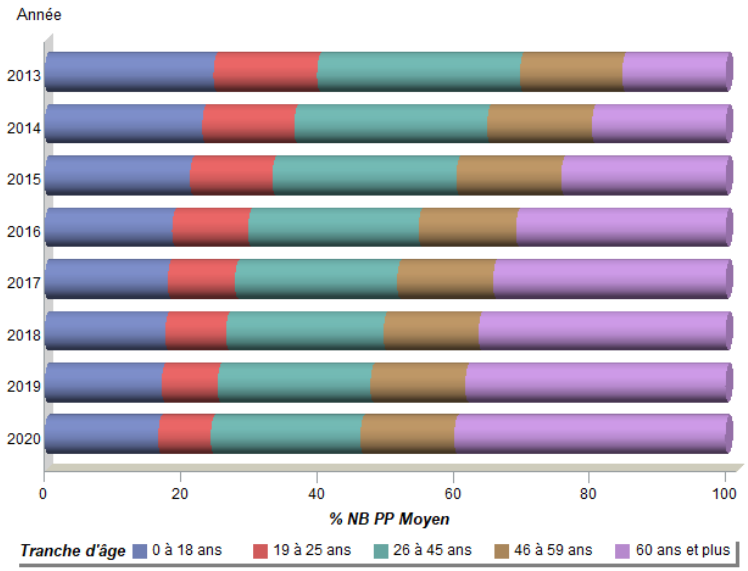


Figure 20 : Répartition en % du nombre moyen des personnes protégées par tranche d'âge

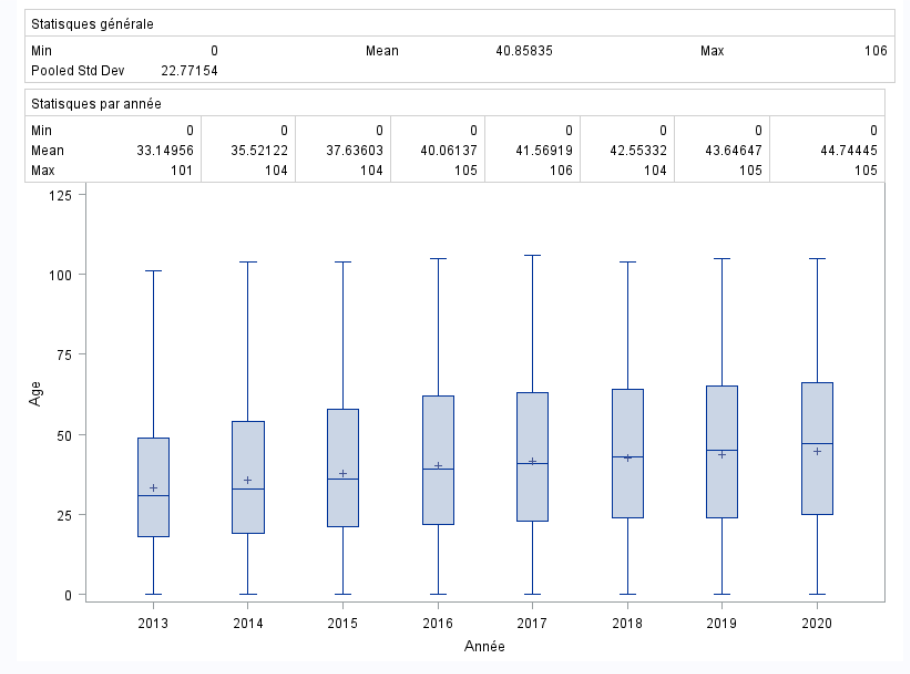


Figure 21 : boîtes à moustache de la variable âge par année

Cette évolution peut s'expliquer par l'entrée en vigueur en 2016, de l'accord national interprofessionnel (transposé dans la loi n 2013-504 du 14 juin 2013 relative à la sécurisation de l'emploi) qui instaure la généralisation de la complémentaire santé dans les entreprises du secteur privé. Grâce à cette réforme, les actifs bénéficient automatiquement d'une complémentaire souscrite par leur employeur et donc ont moins recours à des contrats individuels. Elle est aussi la conséquence d'un rééquilibrage tarifaire sur le segment des séniors.

La répartition des effectifs par niveau de gamme montre que l'offre « entrée de gamme » constitue la majorité du portefeuille à plus de 80%. Une part qui augmente légèrement d'une année sur l'autre.

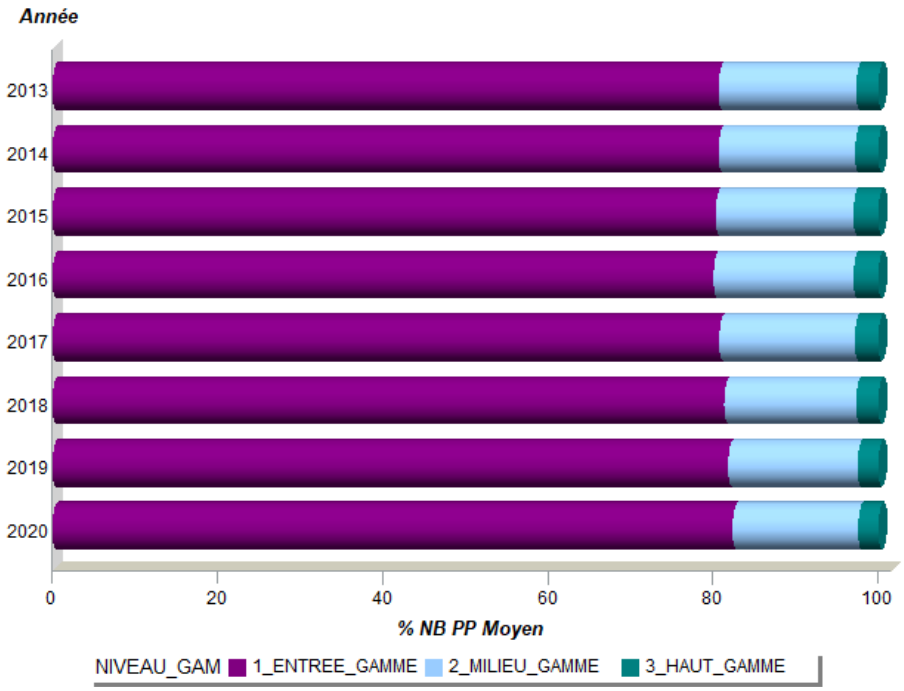


Figure 22 : Répartition en % du nombre moyen des personnes protégées par niveau de gamme

L'analyse des cotisations, permet de confirmer les constats précédents :

- Une augmentation continue des cotisations, année après année, qui passe de 56,72 M€ en 2016 à 91,03 M€ en 2020, soit une progression de + 60%. Cette évolution est due à la progression du nombre des adhérents, l'augmentation de la moyenne d'âge et aux indexations tarifaires annuelles visant à garantir l'équilibre technique et à faire face à l'inflation des prestations ;
- La part la plus importante des cotisations est issue des contrats dits « entrée de gamme ».

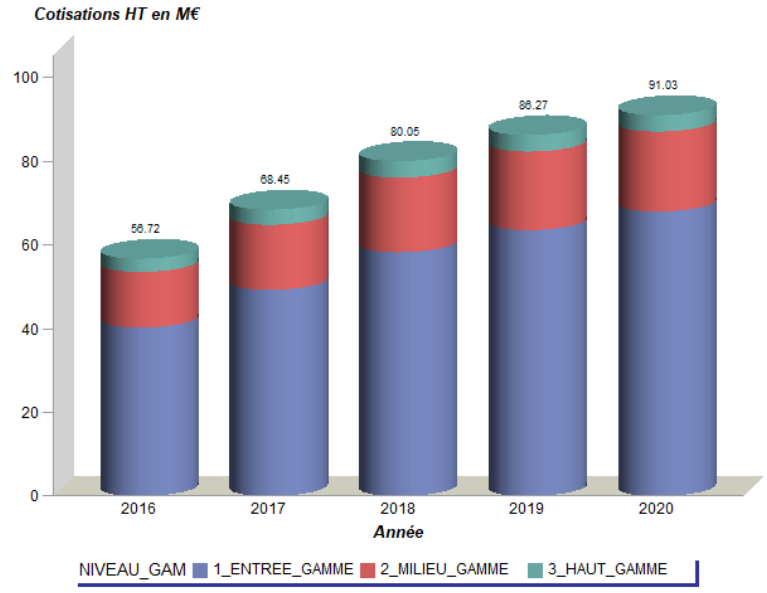


Figure 23 : Répartition en % du chiffre d'affaires HT par niveau de gamme entre 2016 et 2020

La prime moyenne par personne protégée, suit aussi la même tendance. Tendance qui s'explique comme pour le chiffre d'affaires par une moyenne d'âge du portefeuille qui augmente annuellement et des indexations tarifaires qui interviennent chaque année.

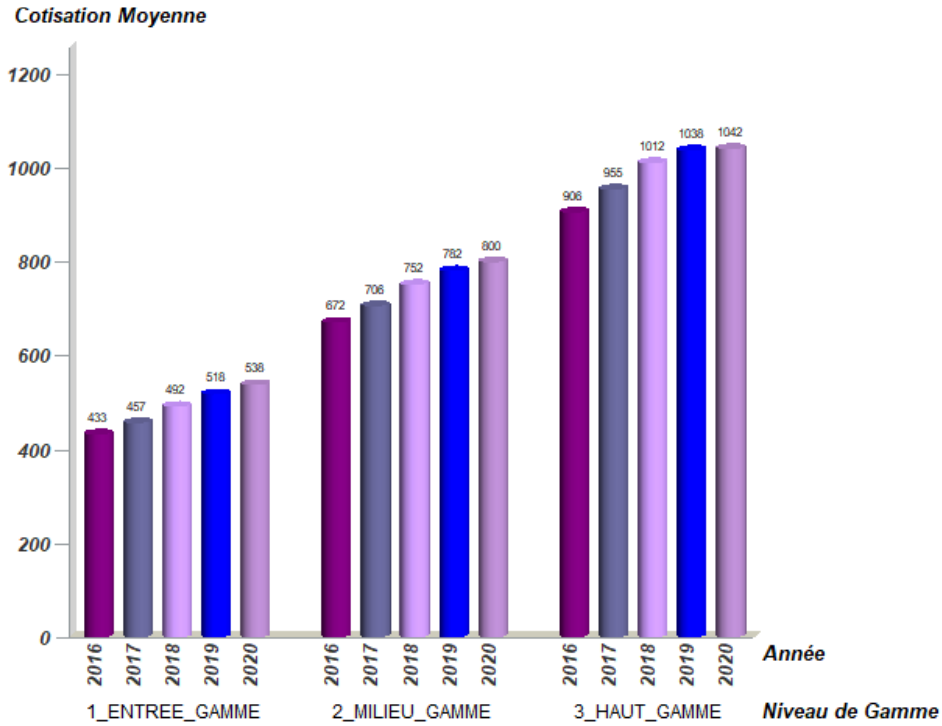


Figure 24 : Prime moyenne par année et par niveau de gamme

Entre 2016 et 2020, les évolutions enregistrées sont les suivantes : +5,1% en 2017, +7% en 2018, +4,6% en 2019 et enfin +3,2% en 2020. Ce qui nous conduit à une augmentation globale de +21,4% en 2020 comparée à 2016.

Ces évolutions varient toutefois d'un niveau de gamme à l'autre. Par exemple, en 2019, l'augmentation des primes moyennes sur le haut de gamme était de +2,9% alors que sur la même période, les primes des produits « entrée de gamme » ont augmenté de + 5,33%. A noter aussi, que l'entrée de gamme connait l'augmentation la plus forte chaque année.

2017			2018			2019			2020		
Entrée de Gamme	Milieu de Gamme	Haut de Gamme	Entrée de Gamme	Milieu de Gamme	Haut de Gamme	Entrée de Gamme	Milieu de Gamme	Haut de Gamme	Entrée de Gamme	Milieu de Gamme	Haut de Gamme
5,60%	5,23%	5,38%	7,74%	6,76%	6,14%	5,33%	4,00%	2,88%	3,95%	2,44%	0,08%

Tableau 14: Evolution en % de la prime moyenne par niveau de gamme

1.2.2 Sinistralité

À la vue de la répartition d'effectifs par niveau de gamme, il est évident que l'entrée de gamme représentera une part importante des prestations et que ces mêmes prestations connaitront des augmentations annuelles, comme le montre le graphique ci-dessous :

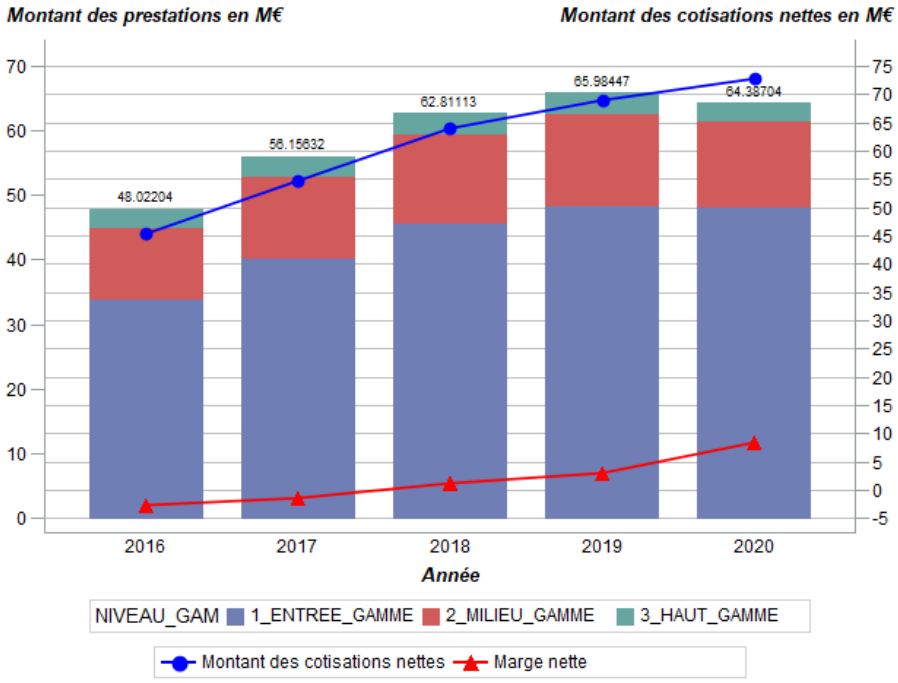


Figure 25 : Prestations par niveau de gamme, chiffre d'affaires net et marge par année

Ce graphique met en évidence aussi une amélioration continue de la marge nette qui passe de -2 M€ en 2016 à + 3 M€ en 2019, soit en progression de plus de 5 millions d'euros en quatre ans.

L'année 2020 sera marquée par une forte progression de la marge nette. Celle-ci est en effet de 8,4 M€ soit une évolution de +173% comparée à 2019. Ceci s'explique principalement par la baisse des prestations sous l'effet du premier confinement imposé en 2020 entre le 17 mars et le 11 mai, afin de limiter la propagation de la Covid 19.

Le montant des cotisations nettes ainsi que la marge nette sont calculés comme suit :

- ✓ $Montant\ Cotisation_{Net} = (1 - i) * Montant\ cotisation_{Hors\ taxe}$ / Où i représente le taux de frais ;
- ✓ $Marge\ _{Nette} = Montant\ cotisation_{Net} - Montant\ des\ prestations$.

Cette marge provient principalement des produits « entrée et milieu de gamme » alors que les produits « hauts de gamme » réalisent très peu de marge, voire sont déficitaires dans certains cas. En effet, et comme l'illustre le graphique ci-dessous, le produit réalisant la marge la plus importante est le F23 suivi du F12, alors que les produits F54 et F56 enregistrent des pertes.

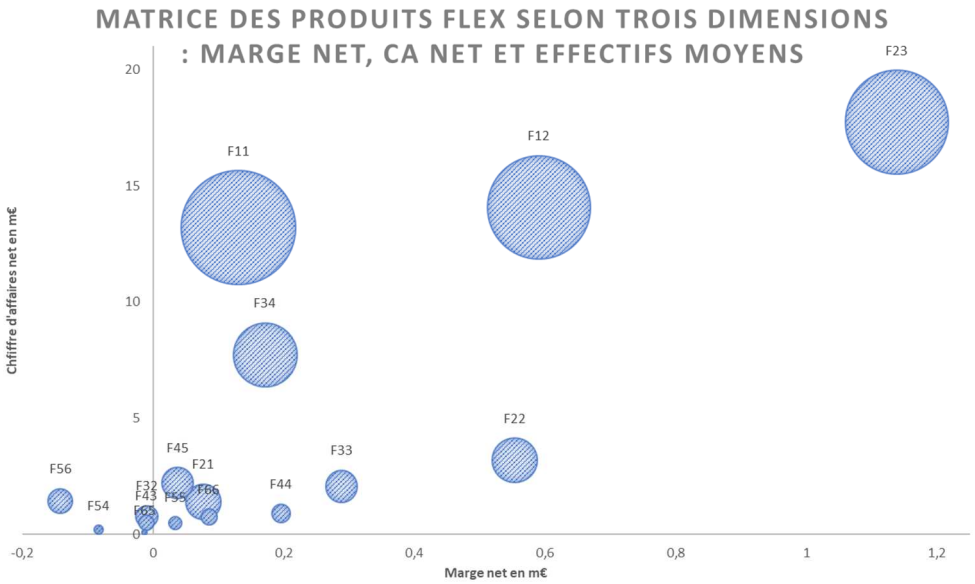


Figure 26 : Matrice des produits FLEX selon trois dimensions

Par ailleurs, l'analyse des prestations mois par mois permet de mieux mettre en évidence l'effet du 1er confinement sur la consommation des soins :

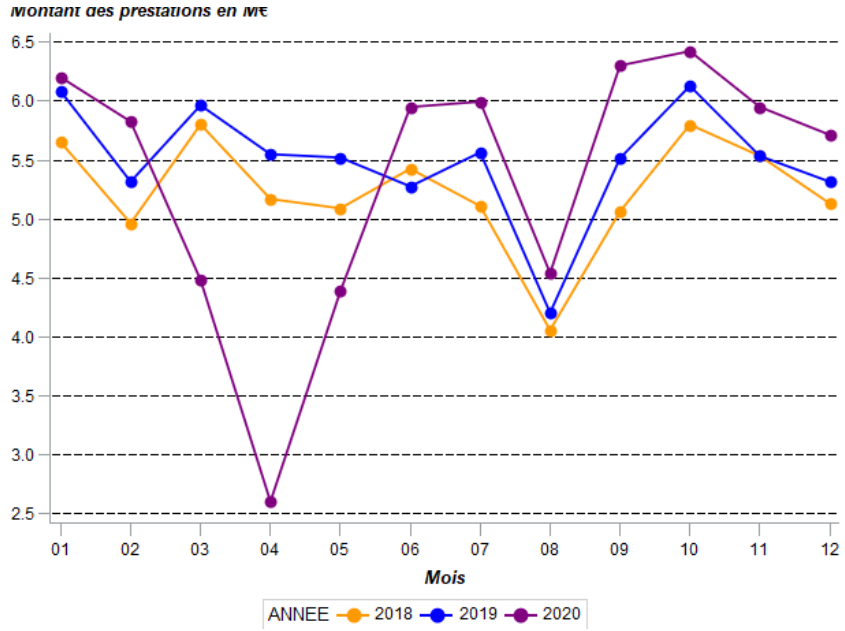


Figure 27 : montant des prestations par mois et par année

En effet, ce graphique laisse apparaître une chute des prestations sur les mois de Mars, avril et Mai pour l'année 2020.

Enfin, la répartition des prestations par poste de soins est assez stable d'une année à l'autre.

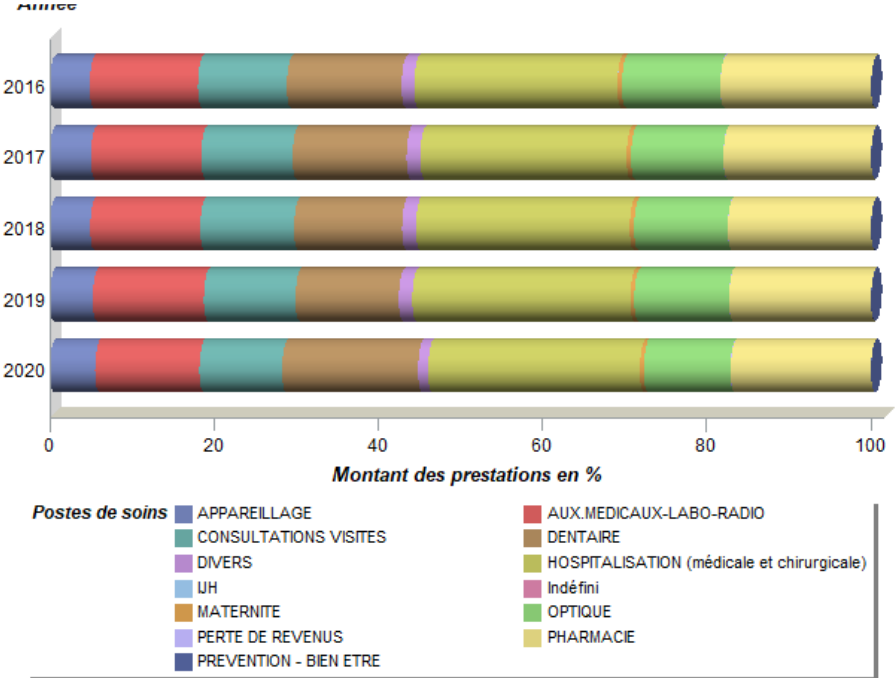


Figure 28 : répartition en % des prestations par poste de soins entre 2016 et 2020

Il est à noter que :

- L'hospitalisation constitue le poste le plus important en termes de dépenses. A titre d'exemple, les remboursements liés à des actes d'hospitalisation, représentaient 25% en 2017 (et 27% en 2019) du total des prestations ;
- Une augmentation de la part des actes dentaires qui passe de 12,4% en 2019 à 16,5% en 2020.

1.2.3 Zoom sur le dentaire et l'optique

Malgré une baisse générale des prestations en 2020 par rapport à 2019 sous l'effet du confinement, les remboursements liés aux soins dentaires sont eux en nette progression (+29%) :

	2019	2020	Evolution en %
APPAREILLAGE	3 341 351 €	3 431 898 €	3%
AUX.MEDICAUX-LABO-RADIO	8 605 820 €	7 856 341 €	-9%
CONSULTATIONS VISITES	7 095 460 €	6 222 546 €	-12%
DENTAIRE	7 925 104 €	10 226 696 €	29%
DIVERS	1 035 581 €	746 711 €	-28%
HOSPITALISATION (médicale et chirur	17 060 849 €	16 069 833 €	-6%
MATERNITE	311 196 €	284 563 €	-9%
OPTIQUE	7 293 385 €	6 501 658 €	-11%
PHARMACIE	11 058 894 €	10 638 930 €	-4%
PREVENTION - BIEN ETRE	17 105 €	11 987 €	-30%
Total général	63 744 746 €	61 991 167 €	-3%

Figure 29 : montant des prestations et évolutions en % par poste de soins

Cette évolution est liée certainement à l'entrée en vigueur du 100% santé associée à l'augmentation des effectifs qui est de 2% entre 2019 et 2020. Toutefois, selon le niveau de gamme et le produit, ces tendances sont accentuées ou inversées, comme le montrent les deux graphiques suivants :

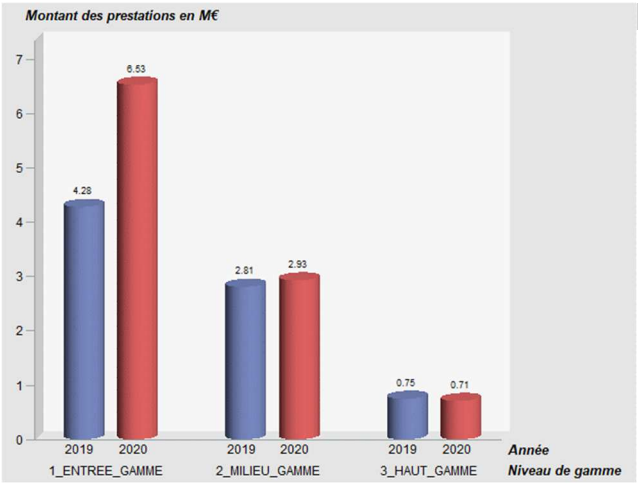


Figure 30 : montant des prestations des soins dentaires par année et par niveau de gamme

	2019	2020	Evolution en %
F11	0,45	1,22	171%
F12	1,23	1,94	58%
F21	0,04	0,09	135%
F22	0,23	0,34	48%
F23	2,38	2,96	24%
F32	0,05	0,06	23%
F33	0,21	0,26	20%
F34	1,75	1,81	4%
F43	0,04	0,04	3%
F44	0,11	0,12	5%
F45	0,65	0,62	-4%
F54	0,02	0,02	-1%
F55	0,08	0,07	-8%
F56	0,55	0,53	-5%
F65	0,01	0,01	-5%
F66	0,12	0,14	14%
Total général	7,93	10,23	29%

Tableau 15 : Montant des prestations des soins dentaires par année et par produit

Les montants des prestations dentaires connaissent :

- Une forte augmentation pour l'entrée de gamme. Le produit F11 enregistre l'évolution la plus importante (+171%) toute gamme confondue, suivi du F21 (+135%) ;
- Une légère augmentation pour le milieu de gamme. Augmentation constatée pour tous les produits de ce niveau sauf pour le F45 qui enregistre une légère baisse ;
- Un léger recul pour le haut de gamme tout produit confondu sauf pour le F66 qui est en augmentation de +14% .

L'analyse du périmètre du 100% santé qui correspond aux prothèses dentaires montre une augmentation du poids des prothèses dentaires dans les prestations réglées comme le décrit le graphique suivant :

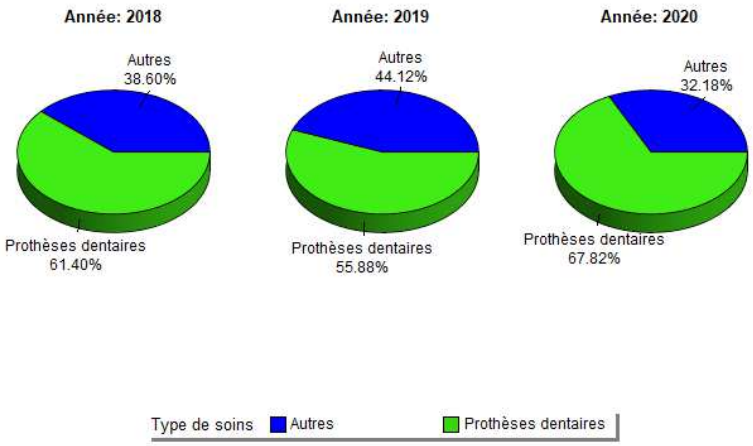


Figure 31:Répartition du montant des prestations dentaires par type de soins

Nous constatons que le poids des prothèses dentaires passe de 56% en 2019 à 68% en 2020. Une augmentation = accompagnée par une hausse du poids du panier 100% santé qui représente 57% en 2020 contre 27% en 2019 :

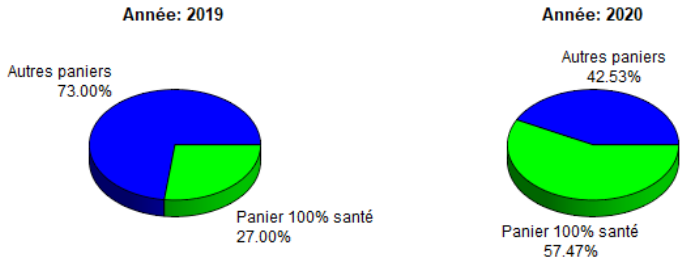


Figure 32: Répartition du montant des prestations prothèses dentaires par type de panier

Ces variations impactent aussi bien le montant des prestations que le montant du reste à charge, comme le montre le tableau suivant :

	Vairation 2018-2019				Variation 2019-2020			
	GAMME			Total	GAMME			Total
	1_ENTREE	2_MILIEU	3_HAUT		1_ENTREE	2_MILIEU	3_HAUT	
Montant moyen des prestations par acte	4%	-9%	-12%	-8%	74%	23%	-6%	27%
Montant moyen du reste à charge par acte	-1%	-11%	-13%	-8%	-35%	-21%	-19%	-26%

Tableau 16 : Variation des indicateurs moyens des prestations de type prothèse dentaire

Le montant moyen du reste à charge a connu deux années successives de baisse. La baisse est toutefois plus marquée sur 2020. Le montant des prestations connait lui une forte hausse en 2020 après avoir enregistré une baisse modérée en 2019. Cette hausse est plus marquée sur l'« entrée de gamme » suivi du « milieu de gamme » alors que le « haut de gamme » enregistre une baisse.

En optique, le constat est un peu plus contrasté. En effet, le montant des prestations enregistre une baisse de 11% par rapport à 2019. Une tendance baissière constatée sur tous les produits sauf les deux produits F11 et F21 nous constatons des augmentations de respectivement +9% et +8% comme indiqué dans le tableau ci-dessous :

	2019	2020	Evolution en %
F11	0,77	0,83	9%
F12	1,43	1,31	-8%
F21	0,07	0,07	8%
F22	0,25	0,23	-9%
F23	2,48	2,13	-14%
F32	0,05	0,04	-6%
F33	0,21	0,19	-10%
F34	1,16	0,92	-20%
F43	0,04	0,04	-12%
F44	0,09	0,08	-18%
F45	0,33	0,29	-14%
F54	0,02	0,01	-42%
F55	0,06	0,05	-14%
F56	0,23	0,22	-5%
F65	0,01	0,01	-13%
F66	0,10	0,08	-20%
Total général	7,29	6,50	-11%

Tableau 17 : Montant des prestations optique par année et par produit

D’ailleurs, le recours au panier 100% santé en optique reste marginal comparé aux prothèses dentaires, comme le montre le graphique suivant :

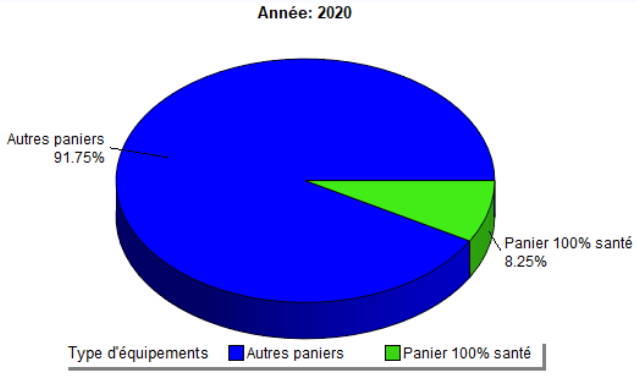


Figure 33: Répartition du montant des prestations des équipements optiques par type de panier

2 Impact du 100% santé sur la charge des sinistres

Comme précisé précédemment, l'objectif de la mise en place de la réforme 100% santé est de favoriser le recours aux soins. L'atteinte de cet objectif aura comme effet une évolution de la charge des sinistres supportée par les organismes complémentaires. Cette partie sera donc consacrée à identifier des tendances pouvant avoir un impact sur cette dernière après l'entrée en vigueur de la réforme. Son périmètre étant limité à certains soins, l'étude sera restreinte à ce même périmètre. De plus, les actes de soins « aides auditives » seront eux exclus car le calendrier de mise en place court jusqu'à fin de 2021 alors que la base de données recense uniquement les soins réalisés au plus tard le 31/12/2020 et vus au 31/08/2021.

2.1 Fréquence de consommation

Diminuer le renoncement aux soins conduira automatiquement à une augmentation du nombre des consommateurs. Une première approche pour évaluer l'impact de la réforme est d'analyser les fréquences de consommation par année, par niveau de couverture et par type de soins (dentaire ou optique).

Pour rappel, Le constat ayant conduit à l'introduction de cette réforme est que le reste à charge représente un frein à l'accès aux soins, ce qui peut conduire à déduire que plus élevé sera le niveau de couverture, plus faible sera l'impact de la réforme. La gamme étudiée propose six niveaux de couverture en optique et en dentaire.

Plusieurs tests statistiques permettent de comparer des proportions. Un de ces tests consiste à comparer les proportions observées sur deux échantillons.

Construction du test :

Soit A et B deux échantillons distincts de tailles respectives n_A et n_B et définissant des variables indépendants. Sera noté P_A la proportion des consommateurs observé sur le premier échantillon et P_B sur le deuxième. Les hypothèses du test sont les suivants :

- L'hypothèse nulle $\mathcal{H}_0 : P_A = P_B$;
- L'hypothèse alternative $\mathcal{H}_1 : P_A \neq P_B$;
- La statistique du test est la suivante : $Z = \frac{P_A - P_B}{\sqrt{\frac{p*q}{n_A} + \frac{p*q}{n_B}}}$ $o\grave{u}\ p = 1 - q$.

  Avec p la proportion théorique des consommateurs. Elle est inconnue mais peut être approchée par la moyenne pondérée des deux proportions observées :

$$p = \frac{n_A * P_A + n_B * P_B}{n_A + n_B}.$$

Sous $\mathcal{H}_0$ et si les échantillons sont de tailles suffisamment grandes, la variable Z suit approximativement la loi normale centrée réduite. Au risque α l'intervalle de confiance est le suivant :

$$IC_\alpha = [-z; +z] \; tel \; que \; P(|Z| < z) = 1 - \alpha \; .$$

Autrement dit, z est le quantile au niveau $1 - \frac{\alpha}{2}$ d'une loi normale centrée réduite. Les intervalles de confiance au niveau 5% par exemple, sont les suivants :

$$Cas \; billatérale : \; IC_{5\%} = [-1.96; 1.96]; cas \; unilatérale \; \text{à} \; droite : \; IC_{5\%} =]-\infty; 1.64].$$

La règle de décision est la suivante :

- $Si : Z \in IC_\alpha \Rightarrow on \; ne \; rejette \; pas \; \mathcal{H}_0$;
- $Sinon, rejet \; de \; \mathcal{H}_0 \; et \; existence \; d'une \; différence \; significative \; entre \; P_A \; et \; P_B$.

Application numérique et résultats :

Le tableau qui suit résume les nombres et les proportions des assurés par année (2018, 2019 et 2020) et par niveau de gamme selon qu'ils aient consommé des prestations de type « prothèses dentaires » dans l'année ou pas :

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Nombre de personnes protégées		Nombre de personnes protégées		Nombre de personnes	
Année	Consommateur	Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
2018	NON	96 535	95%	53 068	89%	6 800	85%
	OUI	5 049	5%	6 834	11%	1 231	15%
	Total année CONSO	101 584	100%	59 902	100%	8 031	100%
2019	NON	99 225	95%	53 515	89%	6 812	86%
	OUI	5 193	5%	6 901	11%	1 075	14%
	Total année CONSO	104 418	100%	60 416	100%	7 887	100%
2020	NON	100 277	94%	52 939	89%	6 888	88%
	OUI	6 043	6%	6 515	11%	944	12%
	Total année	106 320	100%	59 454	100%	7 832	100%

Tableau 18 : Nombre et fréquence des assurés. Soins : prothèses dentaires ; période : année

L'objectif est de comparer la proportion des consommateurs de l'année N (P_A) à celle de l'année N-1 (P_B). Autrement dit, comparer la proportion 2019 à celle de 2018 et la proportion 2020 à celle de 2019. Les valeurs de la statique Z obtenues par niveau de couverture sont les suivantes :

	Entrée de gamme	Milieu de gamme	haut de gamme
Z 2018 --2019	0,00	0,11	-3,03
Z 2019 --2020	7,26	-2,58	-2,96

Tableau 19 : Statistiques Z par niveau de gamme. Soins : p.dentaires ; période : année

D'après ces statistiques, l'année 2020 a connu une hausse de la fréquence de consommation pour le segment « entrée de gamme » alors que les deux autres segments enregistrent des baisses pouvant être associées au contexte sanitaire. D'autre part, la comparaison de 2019 à 2018 montre une baisse pour le « haut de gamme ». Une explication plus détaillée de ces statistiques est proposée en annexe 2 pour l'ensemble des tests effectués.

Ces premiers résultats laissent suggérer qu'il y aurait éventuellement une augmentation de la consommation des prothèses dentaires entre 2019 et 2020 sur l'entrée de gamme. Sachant que l'année 2020 a connu plusieurs confinements, dont un strict à partir de mars, qui ont eu pour effet une baisse générale des prestations. Il serait intéressant de comparer ces mêmes proportions sur les deux premiers mois de l'année. Par soucis de simplification, il sera considéré que les résiliations se font au 31/12 de chaque année et que les nouvelles souscriptions se font au 01/01 de chaque année. Les statistiques obtenues sont les suivantes :

	Entrée de gamme	***Milieu de gamme***	***Haut de gamme***
Z 2018 --2019	*0,00*	*-0,31*	*-1,12*
Z 2019 --2020	*4,75*	*1,65*	*-1,25*

Tableau 20 : Statistiques Z par niveau de gamme. Soins : p. dentaires ; période : deux mois

Cette fois-ci le test détecte une augmentation en 2020 pour les segments « entrée de gamme » et « milieu gamme » alors qu'une différence n'est identifiée pour le « haut de gamme ». Les statistiques de comparaison de l'année 2019 à l'année 2018 ne permettent pas de conclure à des différences.

Une des explications possibles de l'absence de différence pour le « haut de gamme » est que les assurés avec ces garanties ont moins recours au panier 100% santé comme indiqué dans le tableau suivant :

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Nombre de personnes protégées		Nombre de personnes protégées		Nombre de personnes protégées	
		Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
Année	Panier 100% santé						
2020	NON	259	20.42 %	423	26.70 %	96	38.86 %
	OUI	1 009	79.57 %	1 161	73.29 %	151	61.13 %
	Total année	1 268	100.00 %	1 584	100.00 %	247	100.00 %

Tableau 21 : Consommateurs par type de panier. Soins : p. dentaires ; période : deux mois

Nous pouvons constater que le panier 100% santé est plus utilisé par les adhérents « entrée de gamme » que par les adhérents « haut de gamme ». A noter que cette répartition reste identique en prenant les prestations sur toute l'année 2020. L'analyse du nombre d'actes et du montant des prestations confirme cette tendance. Les tableaux contenant ces indicateurs sont disponibles en annexe 2.

Nous pouvons donc envisager que l'impact de la réforme est très important sur les prestations des contrats « Entrée de gamme », relativement important sur les contrats « milieu de gamme » et faible sur les contrats « haut de gamme ».

Concernant les équipements optiques, les tableaux de garanties stipulent que - sauf exception - seul un équipement est autorisé tous les deux ans. La consommation d'un individu dépend donc de celle de l'année précédente. Par exemple, et sauf exception, une personne qui a consommé en 2018, ne

pourra pas consommer en 2019. Ceci peut éventuellement influencer les résultats du test. Les deux tableaux suivants récapitulent les résultats du test appliqués aux deux premiers mois de l'année :

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Nombre de personnes protégées		Nombre de personnes protégées		Nombre de personnes protégées	
Année	Consommateur	Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
2018	NON	87 305	97%	45 804	95%	5 912	95%
	OUI	2 440	3%	2 316	5%	320	5%
	Total année CONSO	89 745	100%	48 120	100%	6 232	100%
2019	NON	89 050	97%	45 670	95%	5 859	95%
	OUI	2 723	3%	2 656	5%	336	5%
	Total année CONSO	91 773	100%	48 326	100%	6 195	100%
2020	NON	90 245	97%	46 035	94%	5 949	95%
	OUI	3 094	3%	2 711	6%	336	5%
	Total année	93 339	100%	48 746	100%	6 285	100%

Tableau 23 : Nombre et fréquence des assurés. Soins : é. optiques ; période : deux mois

	Entrée de gamme	Milieu de gamme	Haut de gamme
Z 2018 --2019	3,21	4,78	0,72
Z 2019 --2020	4,32	0,48	-0,20

Tableau 22 : Statistiques Z par niveau de gamme. Soins : é. optiques ; période : deux mois

Il est à noter que, d'après ces statistiques, le segment « entrée de gamme » a connu deux années consécutives d'augmentation de la fréquence de consommation sur les deux premiers mois de l'année. Le « milieu de gamme » a connu une hausse en 2019. Pour le reste, les résultats ne détectent pas de changement de tendance.

Le tableau suivant qui répartit les assurés selon leurs dates d'adhésion (un nouvel adhérent correspond à un assuré ayant souscrit sa garantie en cours d'année) et selon qu'ils aient consommé dans l'année ou pas, fournit une explication de ces résultats :

			GAMME					
			1_ENTREE		2_MILIEU		3_HAUT	
			Nombre de personnes protégées		Nombre de personnes protégées		Nombre de personnes protégées	
ANNEE	Nouveau adhérent	Consommateur	Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
2018	NON	NON	62 821	70%	36 220	75%	4 608	74%
		OUI	2 102	2%	1 986	4%	285	5%
	OUI	NON	24 484	27%	9 584	20%	1 304	21%
		OUI	338	0%	330	1%	35	1%
		Total	89 745	100%	48 120	100%	6 232	100%
	Nouveau adhérent	Consommateur						
2019	NON	NON	65 889	72%	37 098	77%	4 683	76%
		OUI	2 416	3%	2 360	5%	297	5%
	OUI	NON	23 161	25%	8 572	18%	1 176	19%
		OUI	307	0%	296	1%	39	1%
		Total	91 773	100%	48 326	100%	6 195	100%
	Nouveau adhérent	Consommateur						
2020	NON	NON	69 442	74%	38 919	80%	4 834	77%
		OUI	2 774	3%	2 488	5%	304	5%
	OUI	NON	20 803	22%	7 116	15%	1 115	18%
		OUI	320	0%	223	0%	32	1%
		Total	93 339	100%	48 746	100%	6 285	100%

Tableau 24 : Nombre et fréquence des assurés et des nouveaux adhérents. Soins : é. optiques

L'augmentation constatée entre 2018 et 2019 ainsi qu'entre 2019 et 2020 pour « l'entrée de gamme » est due à l'augmentation de la consommation des anciens adhérents. La stagnation constatée en 2020 est due à une baisse de la consommation des nouveaux adhérents alors qu'elle était en progression en 2019.

Ces différents résultats ne permettent pas de conclure à un effet 100% santé. Le tableau qui suit permet de constater que les paniers « 100% santé » sont moins utilisés que pour les actes dentaires :

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Nombre de personnes protégées		Nombre de personnes protégées		Nombre de personnes protégées	
Année	Panier 100% santé	Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
2020	NON	2 575	74.01 %	2 605	92.53 %	334	99.40 %
	OUI	904	25.98 %	210	7.46 %	2	0.59 %
	Total année	3 479	100.00 %	2 815	100.00 %	336	100.00 %

Tableau 25 : Consommateurs par type de panier. Soins : é. optiques ; période : deux mois

Il montre un recours moins important au panier 100% santé comparé aux actes dentaires. En effet, seuls 26% des consommateurs « entrée de gamme » ont eu recours à ce panier en 2020 alors que pour les actes dentaires ce taux est de 80%. Ce taux chute à 0.6% pour les consommateurs « haut de gamme » alors qu'ils sont 61% à y avoir eu recours pour les actes dentaires. Nous retrouvons ces mêmes proportions en analysant les indicateurs : nombre d'actes et montant des prestations (analyse qui est proposée en annexe 2).

Les tests effectués ont permis de mettre en évidence une tendance haussière de la fréquence de consommation sur les deux premiers mois de l'année. Cette tendance concerne :

- Les soins de type prothèses dentaires pour les assurés avec des garanties « entrée de gamme » ou « milieu de gamme » ;
- Les soins de type équipements optiques pour les assurés équipés de garanties « entrée de gamme». Il s'agit là toutefois, d'une tendance de fond car la fréquence 2019 est aussi satistiquement supérieure à celle de 2018.

Ces premiers résultats laissent suggérer que l'effet de la réforme serait plus perceptible concernant le recours aux soins de type « prothèses dentaires » des assurés appartenant aux segments « entrée de gamme » et « milieu de gamme ». L'analyse des autres variables permettra d'enrichir cette première analyse.

2.2 Nombre d'actes, montant des prestations et reste à charge

Une autre approche pour évaluer l'impact de cette réforme sur la sinistralité du portefeuille est l'analyse des variations annuelles des variables aléatoires : nombre d'actes, montant des prestations et montant du reste à charge. Comme pour la fréquence, l'objectif sera de comparer les observations de l'année N à celles de l'année N-1 au cours des trois dernières années.

Les assurés du portefeuille seront répartis en deux groupes et deux tests seront utilisés :

- Le test des rangs signés de Wilcoxon utilisé dans la comparaison des données du premier groupe ;
- Test de Wilcoxon - Mann – Whitney qui sera utilisé dans la comparaison des données du second groupe.

2.2.1 Test des rangs signés de Wilcoxon

Dans cette analyse la variable étudiée sera notée X, chaque année représentera un échantillon, X observé l'année N-1 une première mesure de la variable X et X observée l'année N une deuxième mesure de la variable X. Par ailleurs, les individus sont les mêmes dans les deux échantillons. L'avantage de ce test qui appartient à la famille des tests non paramétriques et à la catégorie des tests des signes, est qu'il dispense de l'hypothèse d'indépendance des échantillons ainsi que de toute hypothèse de distribution de notre variable. Les tests des signes utilisent dans leurs constructions les écarts entre deux mesures observées sur les mêmes individus. Celui de Wilcoxon utilise en plus l'importance relative de ces écarts, ce qui le rend plus riche et plus puissant. Ces tests peuvent s'écrire de la manière suivante :

- L'hypothèse nulle $\mathcal{H}_0 : F_1(X) = F_2(X + \theta), \theta = 0$;
- L'hypothèse alternative $\mathcal{H}_1 : F_1(X) = F_2(X + \theta), \theta \neq 0$.

Où θ représente un paramètre de translation qui traduit le décalage entre les fonctions de répartitions.

Construction du test :

Soit :

- n la taille de nos deux échantillons ;
- X_1 la variable de l'année N-1 et x_{i1} une observation de cette variable sur l'individu i ;
- X_2 la variable de l'année N et x_{i2} une observation de cette variable sur l'individu i ;
    1. $\forall\, 1 \leq i \leq n : |d_i| = |x_{i1} - x_{i2}|$;
    2. Définition des rangs r_i : la plus petite valeur de $|d_i|$ reçoit le 1 et la plus grande valeur reçoit le rang n ;
    3. $T^+ = \sum_{i:\, d_i>0} r_i$ et $T^- = \frac{n*(n+1)}{2} - T^+$;
    4. Pour n assez grand (n>15) et sous $\mathcal{H}_0$: $T^+ \sim N(\frac{1}{4} * n * (n+1); \frac{1}{24} * n * (n+1) * (2n+1))$;
    5. L'expression de notre statistique de test est la suivante :

$$Z = \frac{T^+ - \frac{1}{4} * n * (n+1)}{\sqrt{\frac{1}{24} * n * (n+1) * (2n+1)}} \sim N(0;1) ;$$

    6. L'intervalle de confiance bilatérale au niveau α est $IC_{5\%} = [-1.96; 1.96]$.

Dans la construction du test, les individus avec des écarts nuls sont supprimés ce qui a pour conséquence une modification de la taille des échantillons. De même un retraitement est effectué

pour les individus du même rangs (appelés ex-aequo) en utilisant un principe appelé principe du rang moyen qui vient modifier la variance de T^+ sans modifier la variable elle-même.

Pour les tests unilatéraux, l'hypothèse alternative s'écrit sous la forme :

- L'hypothèse alternative H_1 : $F_1(X) = F_2(X+\theta), \theta < 0\ si\ test$ à $gauche\ ou\ \theta > 0\ si\ test$ à $droite$;
- La statistique du test s'écrit :

$$Z = \frac{T^+ - \frac{1}{4} * n * (n+1) + 0{,}5}{\sqrt{\frac{1}{24} * n * (n+1) * (2n+1)}}\ si\ gauche\ ou\ Z = \frac{T^+ - \frac{1}{4} * n * (n+1) - 0{,}5}{\sqrt{\frac{1}{24} * n * (n+1) * (2n+1)}}\ sinon\ .$$

Le logiciel SAS, lui, utilise une autre approximation, très proche de la précédente, faisant appel à la loi de Student à n-1 degrés de liberté :

$$\frac{\sqrt{n-1} * S}{\sqrt{n * V - S^2}} \sim T(n-1)\ Avec :$$

$$S = T^+ - \frac{1}{4} * n * (n+1) \qquad et \qquad V = \frac{1}{24} * n * (n+1) * (2n+1) - \frac{1}{2} \sum_k t_k * (t_k + 1) * (t_k - 1).$$

S représente la statistique du test. Et t_k désigne le nombre d'ex-aequo dans le groupe k.

Application numérique et résultats :

Pour réaliser ce test, il était nécessaire de se mettre dans le cadre d'échantillons appariés. Les individus retenus sont les assurés présents sur deux années consécutives (N-1 et N) et qui ne changent pas du niveau de gamme d'une année à l'autre. Par ailleurs, ce test a nécessité la construction d'une nouvelle variable correspondant à la différence entre notre variable d'origine X mesurée l'année N et mesurée l'année N-1 :

$$Diff_X = X_N - X_{N-1}.$$

Appliqué à nos échantillons sur le périmètre des prothèses dentaires, le test produit les résultats suivants :

		Comparaison sur une année entière				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	1 064 513	Pr >= |S|	<,0001
	2_MILIEU	Rang signé	S	-2 385 690	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	-45 329	Pr >= |S|	0,0016
2019-2018	1_ENTREE	Rang signé	S	217 919	Pr >= |S|	0,2723
	2_MILIEU	Rang signé	S	-349 969	Pr >= |S|	0,2052
	3_HAUT	Rang signé	S	-32 286	Pr >= |S|	0,0669

Tableau 26 : Wilcoxon appariés pour le nombre d'actes. Soins : p. dentaires ; période : année

Le test met en évidence une différence du nombre d'actes entre l'année 2019 et l'année 2020 quel que soit le niveau de gamme. La statistique de test étant positive pour l'entrée de gamme, nous pouvons en déduire que le nombre d'actes en 2020 est supérieur à celui de 2019. Ce même test ne détecte aucune différence du nombre d'actes entre 2018 et 2019.

Seront présentés par la suite, pour l'ensemble des variables uniquement, les résultats concernant la comparaison sur les deux premiers mois de l'année. Les résultats des tests concernant la comparaison

des exercices complets ainsi qu'une interprétation détaillée de ces résultats est disponible en annexe 3.

Pour la même raison que pour la comparaison des fréquences de consommateurs, ce test a été effectué sur un périmètre plus restreint. Le nouveau périmètre correspond au nombre d'actes enregistré les deux premiers mois de l'année. Les résultats obtenus sont les suivants :

		Comparaison sur les 2 premiers mois de l'année				
		Test		Statistique		p-value
2020-2019	1_ENTREE	Rang signé	S	46 239	Pr >= |S|	0,0089
	2_MILIEU	Rang signé	S	-78 373	Pr >= |S|	0,0006
	3_HAUT	Rang signé	S	-779	Pr >= |S|	0,562
2019-2018	1_ENTREE	Rang signé	S	-5 180	Pr >= |S|	0,7081
	2_MILIEU	Rang signé	S	-37 359	Pr >= |S|	0,0765
	3_HAUT	Rang signé	S	-1 443	Pr >= |S|	0,2969

Tableau 37 : Wilcoxon appariés pour le nombre d'actes. P. dentaires ; période : 2 mois

Cette fois ci le test met en évidence une différence du nombre d'actes entre l'année 2019 et l'année 2020 uniquement pour « l'entrée et le milieu de gamme » (avec des p-values moins importantes) alors qu'aucune différence n'est identifiée pour le haut de gamme. D'après les signes des statistiques, la différence est à la hausse pour le segment « entrée de gamme » et la baisse pour le segment « milieu de gamme ». Ce même test ne détecte aucune différence des nombres d'actes entre 2018 et 2019. La différence sur l' « entrée de gamme » est bien visible en traçant la fonction de répartition par année pour le niveau « entrée de gamme » :

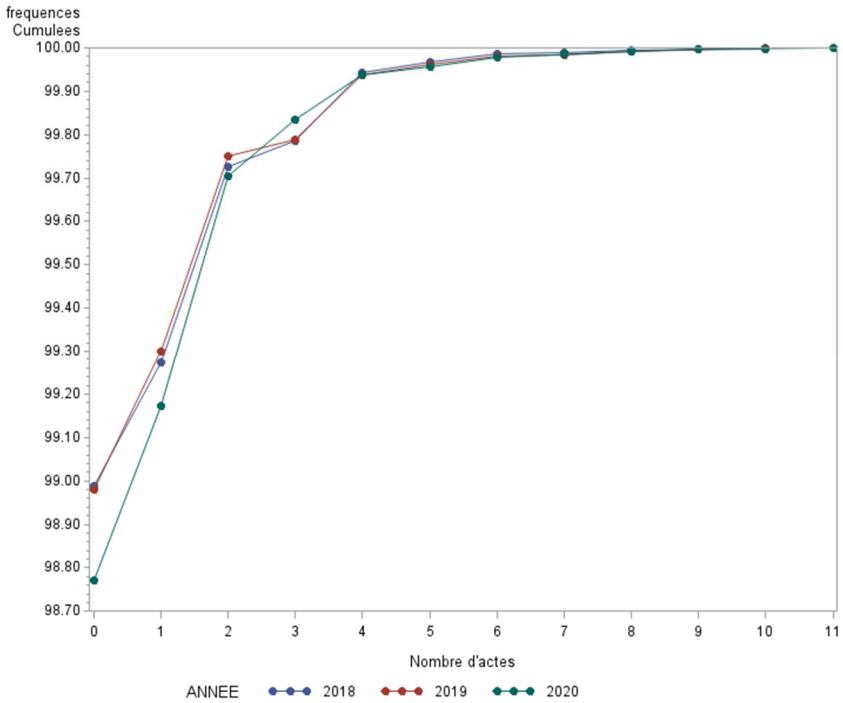


Figure 34: fonction de répartition par année pour les contrats "Entrée de gamme"

Le graphique montre que les courbes des fonctions de répartition des années 2018 et 2019 sont quasiment confondues et qu'elles sont en décalage comparées à celles de 2020. Le graphe montre aussi que pour la première partie des courbes :

$$F_{X_{2020}}(x) < F_{2019}(x) \Leftrightarrow P(X_{2020} \leq x) < P(X_{2019} \leq x) \Leftrightarrow 1 - P(X_{2020} \leq x) \geq 1 - P(X_{2019} \leq x)$$
$$\Leftrightarrow \quad P(X_{2020} > x) \geq P(X_{2019} > x)\,.$$

Cette relation laisse suggérer que la variable X_{2020} a tendance à prendre des valeurs plus importantes que la variable X_{2019}.

Les résultats de ce test appliqués aux équipements optiques sont les suivants :

		Comparaison sur les 2 premiers mois de l'année				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	934 894	Pr >= |S|	<,0001
	2_MILIEU	Rang signé	S	479 098	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	8 375	Pr >= |S|	0,0019
2019-2018	1_ENTREE	Rang signé	S	668 810	Pr >= |S|	<,0001
	2_MILIEU	Rang signé	S	725 936	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	10 926	Pr >= |S|	<,0001

Tableau 27 : Wilcoxon appariés pour le nombre d'actes. Soins : é. optiques

La comparaison 2019 à 2020 relève une différence significative pour tous les segments. Les trois enregistrent une augmentation. Le même test détecte toutefois la même tendance en comparant 2019 à 2018.

Le périmètre d'analyse du montant des prestations concernera :

- Uniquement les assurés ayant eu recours à une prestation sur deux années successives, ce qui évitera de détecter des différences dues à une variation de la fréquence de consommation d'une année à l'autre ;
- Le montant moyen de prestation par acte, ce qui évitera de détecter des différences dues à une variation du nombre d'actes d'une année à l'autre.

Les résultats du test sont les suivants :

- **Pour les prothèses dentaires**

		Comparaison sur les 2 premiers mois de l'année				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	278	Pr >= |S|	0,0023
	2_MILIEU	Rang signé	S	162,5	Pr >= |S|	0,2579
	3_HAUT	Rang signé	S	9,5	Pr >= |S|	0,3008
2019-2018	1_ENTREE	Rang signé	S	211,5	Pr >= |S|	0,0735
	2_MILIEU	Rang signé	S	36	Pr >= |S|	0,8119
	3_HAUT	Rang signé	S	-10	Pr >= |S|	0,5525

Tableau 28 : Wilcoxon appariés pour le montant moyen des prestations. Soins : p. dentaires

Les statistiques et les P-values indiquent une différence pour le segment « entrée de gamme ». Une différence qui se matérialise par une hausse du montant moyen des prestations.

- **Pour les équipements optiques**

		Comparaison sur les 2 premiers mois de l'année				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	-328	Pr >= |S|	0,0596
	2_MILIEU	Rang signé	S	-971,5	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	-58	Pr >= |S|	0,0012
2019-2018	1_ENTREE	Rang signé	S	86,5	Pr >= |S|	0,3502
	2_MILIEU	Rang signé	S	63	Pr >= |S|	0,3089
	3_HAUT	Rang signé	S	29,5	Pr >= |S|	0,0676

Tableau 29: Wilcoxon appariés pour le montant moyen des prestations des é. optiques

Le test détecte une baisse significative du montant moyen des prestations en 2020. Cette baisse concerne les assurés équipés de garanties « milieu de gamme » et « haut de gamme ».

Les conclusions à tirer de ce test appliqué au montant de prestations devront toutefois être nuancées car le montant de prestations est sujet à inflation, ce qui génère des variations des coûts d'une année à l'autre et ne permet pas d'isoler l'effet du 100% santé.

La dernière variable à analyser est le montant du reste à charge. En effet, la réforme vise principalement à baisser ce dernier. Nous pouvons donc nous attendre à des variations significatives de ce montant. Le périmètre du test sera identique à celui du montant des prestations, à savoir concernera uniquement les assurés ayant eu recours à une prestation sur deux années successives et la variable à analyser sera le montant moyen du reste à charge par acte. Les résultats sont les suivants :

- **Pour les prothèses dentaires**

		Comparaison sur les 2 premiers mois de l'année				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	-472	Pr >= |S|	<,0001
	2_MILIEU	Rang signé	S	-466	Pr >= |S|	0,0013
	3_HAUT	Rang signé	S	8,5	Pr >= |S|	0,3594
2019-2018	1_ENTREE	Rang signé	S	315	Pr >= |S|	0,0043
	2_MILIEU	Rang signé	S	258	Pr >= |S|	0,1321
	3_HAUT	Rang signé	S	1,5	Pr >= |S|	0,946

Tableau 30 : Wilcoxon appariés pour le montant moyen du reste à charge des p. dentaires

Le résultat du test montre une différence significative entre 2020 et 2019 pour les segments « entrée de gamme » et « milieu de gamme » avec une tendance à la baisse. Contrairement à 2019 ou ce montant a connu une hausse pour le segment « entrée de gamme » comparé à 2018.

- **Pour les équipements optiques**

		Comparaison sur les 2 premiers mois de l'année				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	312,5	Pr >= |S|	0,067
	2_MILIEU	Rang signé	S	709,5	Pr >= |S|	0,001
	3_HAUT	Rang signé	S	31	Pr >= |S|	0,1167
2019-2018	1_ENTREE	Rang signé	S	-83	Pr >= |S|	0,5356
	2_MILIEU	Rang signé	S	35,5	Pr >= |S|	0,6836
	3_HAUT	Rang signé	S	-1,5	Pr >= |S|	0,9632

Tableau 31 : Wilcoxon appariés pour le montant moyen du reste à charge des é. optiques

La seule différence identifiée par le test concerne le segment « milieu de gamme » qui enregistre en 2020 une hausse significative du montant moyen du reste à charge.

L'augmentation de la fréquence du recours aux soins de type prothèses dentaires des assurés « entrée de gamme » identifiée précédemment, est accompagnée par une hausse du nombre d'actes et du montant moyen des prestations ainsi qu'une baisse de montant moyen du reste à charge. Celle du segment « milieu de gamme » est accompagnée par une baisse du reste à charge. Ces différents effets sont à mettre en lien directement avec la réforme qui vise effectivement à favoriser le recours aux soins tout en limitant le reste à charge et en garantissant un niveau élevé de la qualité des soins. Les effets identifiés pour les soins de type « équipements optiques » montrent encore une fois que la réforme n'a pas influencé le comportement des assurés pour ce type de soins.

2.2.2 Test de Wilcoxon - Mann – Whitney

Comme le test précédent, il s'agit d'un test non paramétrique qui s'appuie, comme la majorité des tests basés sur les rangs, sur une statistique de rang linéaire qui s'écrit généralement sous la forme :

$$T = \sum_{i=1}^{n} c_i * f(r_i) \text{ où } f(r_i) \text{ } s'appelle \text{ } score \text{ } ou \text{ } code \text{ } et \text{ } f(.) \text{ } une \text{ } fonction \text{ } score \text{ .}$$

Dans le cas de ce test la fonction $f(.)$ correspond à la fonction identité et les codes correspondent directement aux rangs bruts.

Pour les besoins de ce test, le premier échantillon sera constitué des assurés ayant adhéré à leurs garanties au cours de l'année N-1 et le deuxième échantillon des assurés ayant adhéré à leurs garanties au cours de l'année N.

Les hypothèses du test sont identiques à celle du test des rangs signés de Wilcoxon à savoir :

- L'hypothèse nulle $\mathcal{H}_0 : F_1(X) = F_2(X + \theta), \theta = 0$;
- L'hypothèse alternative $\mathcal{H}_1 : F_1(X) = F_2(X + \theta), \theta \neq 0$ ($dans \text{ } le \text{ } cas \text{ } unilatéral : \theta < 0 \text{ } si \text{ } test$ à $gauche \text{ } ou \text{ } \theta > 0 \text{ } si \text{ } test$ à $droite$.

Construction du test :

Soit :

- $n_1 \text{ } et \text{ } n_2$ les tailles de nos échantillons ;
- $S_k = \sum_{i=1}^{n_k} r_{ik}$;

  $ou \text{ } k \in \{1; 2\} \text{ } et \text{ } r_{ik} \text{ } le \text{ } rang \text{ } de \text{ } la \text{ } valeur \text{ } x_{ik} \text{ } rattachée$ à $l'individu \text{ } i \text{ } du \text{ } groupe \text{ } k \text{ } et$

- $U_k = S_k - \frac{n_k * (n_k + 1)}{2}$;
- $La \text{ } statistique \text{ } du \text{ } test \text{ } est : U = \min(U_1; U_2)$;
- $Sous \text{ } \mathcal{H}_0 \text{ } l'espérance \text{ } et \text{ } la \text{ } variance \text{ } de \text{ } U \text{ } s'ecrivent$:
  - $E(U) = \frac{1}{2} n_1 \, n_2$;
  - $V(U) = \frac{1}{12}(n_1 + n_2 + 1) n_1 n_2$.

Pour des échantillons d'une taille suffisamment grande ($n_k > 8$), la distribution de la variable U converge vers une loi normale de moyenne $E(U)$ et de variance $V(U)$. La statistique centrée réduite est définie comme :

- $Z = \frac{U - E(U)}{\sqrt{V(U)}} \text{ } dans \text{ } le \text{ } cas \text{ } bilatéral$;
- $Z = \frac{U - E(U) \pm 0{,}5}{\sqrt{V(U)}} \text{ } dans \text{ } le \text{ } cas \text{ } unilatéral$.

Si les échantillons présentent des ex-aequo, une correction est apportée à la variance de U :

$$\tilde{V}(U) = V(U) * \left(1 - \frac{\sum_{g=1}^{G} t_g ({t^2}_g - 1)}{n^3 - n}\right),$$

$O\grave{u} \; n = n_1 + n_2 \text{ } et \text{ } G \text{ } le \text{ } nombre \text{ } de \text{ } valeurs \text{ } distinctes \text{ } et \text{ } t_g \text{ } et \text{ } le \text{ } nombre \text{ } d'observation \text{ } g$.

Application numérique et résultats :

Les variations éventuellement identifiées en comparant 2020 à 2019 pouvant être dues aux confinements, seules les données de la période allant de début janvier fin février seront analysées dans ce paragraphe. Les variables qui seront analysées sont : le nombre d'actes, le montant moyen des prestations par acte ainsi que le montant moyen du reste à charge par acte. Pour les raisons évoquées précédemment, le test concernant la première variable tiendra compte de l'ensemble des assurés appartenant à la sélection. Les deux dernières variables seront analysées sur un échantillon plus restreint qui comporte uniquement les nouveaux adhérents de chaque année ayant consommé dans l'année.

Le paragraphe suivant permettra de présenter les différentes statistiques et les p-values obtenues ainsi qu'une interprétation synthétique de ces dernières. Une analyse détaillée des résultats du test sur une année est disponible en annexe 4.

- **Prothèses dentaires :**
  - **Variable nombre d'actes**

		2020-2019	2019-2018
1_ENTREE	Statistique	967 664 071	1 112 304 110
	Z	-3,9459	-4,8609
	Unilatéral	<,0001	<,0001
	Bilatéral Pr > |Z|	<,0001	<,0001
2_MILIEU	Statistique	163 805 011	214 749 786
	Z	-2,3729	-3,0567
	Unilatéral	0,0088	0,0011
	Bilatéral Pr >	0,0176	0,0022
3_HAUT	Statistique	3 810 055	4 726 191
	Z	0,3444	0,243
	Unilatéral Pr	0,3653	0,404
	Bilatéral Pr > |Z|	0,7305	0,808

Tableau 32 : Wilcoxon pour le nombre d'actes des p. dentaires

Ces valeurs permettent de déduire qu'il existe une différence significative entre le nombre d'actes de l'année 2020 et celui de l'année 2019 pour les segments « entrée gamme » et « milieu gamme », avec un nombre d'actes en 2020 qui est statistiquement plus élevé que celui de l'année précédente. Le même constat est toutefois valable lors de la comparaison du nombre d'actes de l'année 2019 à celui de l'année 2018.

  - **Variables montant moyen des prestations et montant moyen du reste à charge**

		Montant moyen des prestations		montant moyen du reste à charge	
		2020-2019	2019-2018	2020-2019	2019-2018
1_ENTREE	Statistique	33 024	35 977	46 302	31 486
	Z	-3,2464	1,9047	5.6761	-1,3627
	Unilatéral Pr	0,0006	0,0284	<.0001	0,0865
	Bilatéral Pr > |Z|	0,0012	0,0568	<.0001	0,173
2_MILIEU	Statistique	63 864	82 028	77 777	75 771
	Z	-2,5522	0,6215	3.8923	-1,9171
	Unilatéral Pr	0,0054	0,2671	<.0001	0,0276
	Bilatéral Pr > |Z|	0,0107	0,5343	<.0001	0,0552
3_HAUT	Statistique	2 326	4 090	2 453	4 322
	Z	0,7951	0,2366	1.6486	0,7247
	Unilatéral Pr	0,2133	0,4065	0.0496	0,2343
	Bilatéral Pr > |Z|	0,4265	0,813	0.0992	0,4686

Tableau 33: Wilcoxon pour les montant moyen des prestations du reste à charge des p. dentaires

Ces résultats indiquent que les montants moyens des prestations des segments « entrée de gamme » et « milieu de gamme » sont plus faibles en 2019 par rapport à 2020 alors que les montants moyens du reste à charge sont plus élevés.

- **Equipements optiques :**
  - **Variable nombre d'actes**

		Nombre d'actes	
		2020-2019	2019-2018
1_ENTREE	Statistique	5 583 339 749	5 181 927 329
	Z	-5,3478	5,7986
	Unilatéral	<,0001	<,0001
	Bilatéral Pr > |Z|	<,0001	<,0001
2_MILIEU	Statistique	1 736 686 164	1 659 943 476
	Z	0,4898	-6,9612
	Unilatéral	0,3121	<,0001
	Bilatéral Pr > |Z|	0,6242	<,0001
3_HAUT	Statistique	27 846 147	27 575 365
	Z	0,1679	-1,5894
	Unilatéral Pr	0,4333	0,056
	Bilatéral Pr > |Z|	0,8667	0,112

Tableau 34 : Wilcoxon pour le nombre d'actes des é. optiques

D'après ces résultats, le nombre d'actes de l'année 2020 est généralement plus élevé que celui de l'année 2019 qui est lui plus faible que celui 2018 pour le segment « entrée de gamme ».

  - **Variables montant moyen des prestations et montant moyen du reste à charge**

		Montant moyen des prestations		montant moyen du reste à charge	
		2020-2019	2019-2018	2020-2019	2019-2018
1_ENTREE	Statistique	6 987 974	4 731 554	6 299 970	4 785 716
	Z	13,9718	-1,3617	1,4526	0,163
	Unilatéral Pr	<,0001	0,0866	0,0732	0,4353
	Bilatéral Pr > |Z|	<,0001	0,1733	0,1463	0,8706
2_MILIEU	Statistique	6 435 690	4 617 269	5 395 616	4 571 650
	Z	14,2269	1,1059	-6,7328	0,0567
	Unilatéral Pr	<,0001	0,1344	<,0001	0,4774
	Bilatéral Pr > |Z|	<,0001	0,2688	<,0001	0,9548
3_HAUT	Statistique	98 699	76 670	80 744	77 515
	Z	3,1923	0,1208	-4,8738	0,5432
	Unilatéral Pr	0,0007	0,4519	<,0001	0,2935
	Bilatéral Pr > |Z|	0,0014	0,9038	<,0001	0,587

Tableau 35 : Wilcoxon pour les montant moyen des prestations du reste à charge des é. optiques

Ces résultats indiquent que le montant moyen des prestations de l'année 2019 est plus élevé que celui de l'année 2020 quel que soit le segment, alors que le montant moyen du reste à charge de l'année 2019 est plus faible que celui de l'année 2020 pour les segments « milieu de gamme » et « haut de gamme ».

Comme pour la première population, les effets identifiés qui peuvent être associés à la réforme, concernent les soins de type prothèses dentaires et les assurés « entrée de gamme » et « milieu de gamme » avec principalement une augmentation du montant moyen des prestations et une baisse du montant moyen du reste à charge.

3 Modélisation de la charge de sinistre

Les différents tests effectués montrent une augmentation de la consommation accompagnée d'une baisse du reste à charge des soins concernés par la réforme. Une évolution davantage perceptible sur « l'entrée de gamme » et les prothèses dentaires alors qu'elle est plus modérée voire absente sur le milieu et le haut de gamme ainsi que sur les équipements optiques. De plus, l'année 2020 reste une année atypique qui ne permet pas de mesurer le vrai impact de la réforme, d'où la nécessité de procéder à une modélisation en tenant compte des modifications qu'a engendré la crise sanitaire

Notons :

- n_a *le nombre d'assurés*, X_i *la charge des sinistres imputable* à *l'assuré* i ;
- N_s *le nombre des sinistres du portefeuille* ;
- K_i *le nombre des sinistres pour l'individu* i. $Y_{i,j}$ *la valeur du sinsitre* j *pour l'individu* i .

Nous aurons donc :

$$X_i = \sum_{j=1}^{K_i} Y_{i,j} \ et \ N_s = \sum_{i=1}^{n_a} K_i \rightarrow \sum_{i=1}^{n_a} X_i = \sum_{i=1}^{n_a}\sum_{j=1}^{K_i} Y_{i,j} = \sum_{l=1}^{N_s} Y_l \ (\ en\ renumérotant\ les\ sinistres).$$

La quantité à estimer est donc l'espérance suivante :

$$E\left[\frac{\sum_{i=1}^{n_a} X_i}{n_a}\right] = E\left[\frac{1}{n_a}\sum_{l=1}^{N_s} Y_l\right] = E\left[E\left[\sum_{l=1}^{N_s}(Y_l|N_s)\right]\right].$$

Deux méthodes peuvent être utilisées pour modéliser la charge de prestations :

- La première, basée sur le modèle coût* fréquence, est souvent utilisée en IARD. Elle est simple à implémenter et permet d'approcher de manière satisfaisante la consommation de l'assuré. Elle consiste à modéliser séparément, d'une part le coût d'un sinistre (appelé coût moyen) et d'autre part la fréquence des sinistres. La charge des prestations qui représente la prime pure n'est alors que le produit de la fréquence probable des sinistres par le coût probable d'un sinistre. Elle s'appuie toutefois sur une hypothèse forte qui doit être vérifiée avant d'y avoir recours. En effet, il faut s'assurer de l'indépendance entre les deux variables fréquence et coût. Dans ce cadre l'écriture de la charge des sinistres est simplifiée et devient :
$$E\left[\frac{\sum_{i=1}^{n_a} X_i}{n_a}\right] = E\left[\frac{1}{n_a} N_s * E[Y|N_s]\right] = E\left[\frac{1}{n_a} N_s E[Y]\right] = E\left[\frac{N_s}{n_a}\right] * E[Y]\,.$$
- La deuxième consiste à modéliser la consommation globale. Elle peut être utilisée lorsque l'hypothèse d'indépendance n'est pas vérifiée. Elle semble par ailleurs mieux adaptée au risque santé ou nous pouvons éventuellement nous 'attendre à ce que les assurés disposant d'un niveau de couverture élevé aient recours plus souvent aux soins.

Pour arrêter le choix de la méthode de modélisation, il est donc nécessaire de vérifier l'hypothèse d'indépendance ; un travail qui sera réalisé dans le paragraphe suivant.

3.1 Liaison entre fréquence et coût moyen

Plusieurs tests statistiques permettent de tester l'indépendance entre deux variables au sens probabiliste du terme avec comme hypothèse nulle : nos deux variables sont indépendantes. Le choix du test dépend de la nature des variables et de leurs lois respectives. La liste suivante donne quelques exemples de ces derniers :

- Le test du chi-deux qui peut s'appliquer dans le cas de variables de tout type. Il est préférable tout de même de s'assurer que les nombres des modalités distinctes de ces dernières ne sont pas très importants.
- Le coefficient de corrélation linéaire (ou de Pearson) utilisé dans le cas de variables numériques. Il caractérise la linéarité de la dépendance de deux variables. La nullité de ce coefficient est équivalente à l'indépendance de nos variables. Cette équivalence est vraie uniquement dans le cas de variables de loi normale. Dans les autres cas, l'indépendance des variables implique la nullité de coefficient de corrélation linéaire mais la réciproque n'est pas vraie.
- Le test du rapport de vraisemblance adapté à tout type de variables sera développé dans la suite de ce paragraphe.

3.1.1 Le test du rapport de vraisemblance

Tout d'abord, précisons quelques définitions et notations :

- $Soit\ X\ et\ Y\ deux\ variables\ observées\ sur\ une\ population\ constituée\ de\ n\ individus$;
- $X_1, X_2, \dots, X_p\ les\ modalités\ de\ la\ variable\ X\ et\ Y_1, Y_2, \dots, Y_q\ \ celle\ de\ la\ variable\ \ Y$;
- $n_{ij}\ l'effectif du\ couple\ \left(X_i, Y_j\right)$;
- $et\ \ n_{i.} = \sum_{j=1}^{q} n_{ij}\ \ , n_{.j} = \sum_{i=1}^{p} n_{ij}\ \ les\ effectifs\ marjinaux$;
- $(x_1, y_1), \dots, \left(x_i, y_j\right), \dots, (x_n, y_n)\ n\ réalisations\ indépendantes\ du\ couple\ (X, Y)$;
- $Les\ probabilités\ p_{ij} = P\left[(X, Y) = \left(x_i, y_j\right)\right]\ \ définissent\ la\ loi\ du\ couple\ (X, Y)$;
- $L\left(p_{ij}\right) = \prod_{i=1}^{p} \prod_{j=1}^{q} (p_{ij})^{n_{ij}}$ la vraisemblance du modèle.

Notons que si :

$$\forall\ 1 \leq i \leq p\ et\ \forall\ 1 \leq j \leq q : \ p_{ij} \geq 0\ et\ \sum_{i=1}^{p} \sum_{j=1}^{q} p_{ij} = 1\ .$$

Une estimation des p_{ij} est la suivante : $\hat{p}_{ij} = \frac{n_{ij}}{n}$.

Construction du test :

Le test consiste à tester l'hypothèse :

$$\mathcal{H}_0 : X\ et\ Y\ sont\ inédpendantes\ .$$

Sous $\mathcal{H}_0$:

- $p_{ij} = \ p_{i.} * \ p_{.j}\ avec\ \ p_{i.} > 0\ ;\ \ p_{.j} > 0\ et\ \sum_{i=1}^{p} p_{i.} = \sum_{j=1}^{q} p_{.j}\ \ ;$
- Une estimation des p_{ij} peut s'écrirecomme : $\hat{p}^0{}_{ij} = \frac{n_{i.} * n_{.j}}{n}$.

Le rapport des maximas de vraisemblance s'écrit alors comme suit :

$$\lambda = \frac{L(\hat{p}^0{}_{ij})}{L(\hat{p}_{ij})}\ .$$

La statistique : $K = -2 \log(\lambda)$ suit asymptotiquement sous H_0 une loi de chi-deux à (p-1) * (q-1) degré de liberté. La région critique au niveau α est :

$$W = \{K > \chi^2_{1-\alpha}((p-1)(q-1))\}\ .$$

Une forte valeur de K conduit au rejet de H_0

Application numérique et résultats :

Les données prises en compte dans la réalisation de ce test se limitent à celles des assurés ayant eu recours au moins une fois aux soins concernés. En effet, les autres assurés qui représentaient en 2019, 92.4% du portefeuille pour les prothèses dentaires et 81.4% pour les équipements optiques, se trouvent automatiquement avec un nombre d'actes et un montant de prestations moyen nuls.

La réalisation du test a nécessité la construction d'un tableau de contingence croisant les deux variables qui se présente sous la forme suivante :

$$\begin{pmatrix} n_{11} & \cdots & n_{1q} \\ \vdots & \ddots & \vdots \\ n_{p1} & \cdots & n_{pq} \end{pmatrix}.$$

Il a donc été nécessaire de découper la variable coût moyen en p classes de valeurs et la variable fréquence en q classes de valeurs. Chaque classe correspond à un intervalle de valeurs de chaque variable et la valeur n_{ij} de notre tableau représente le nombre d'assurés ayant un coût moyen appartenant à la classe i et une fréquence appartenant à la classe j. Le nombre de classes a été choisi de telle sorte à disposer d'assez d'effectifs dans chaque case. La définition des classes, elle, s'est fait en fonction des quantiles.

Les résultats du test appliqués aux prestations prothèses dentaires et équipements optiques sont résumés dans le tableau suivant :

	Prothèses dentaires			*Equipements optiques*		
Statistique	*DDL*	*Valeur*	*Prob*	*DDL*	*Valeur*	*Prob*
Khi-2	18	11029,6863	<,0001	18	73652,2983	<,0001
Test du rapport de vraisemblance	18	11806,8845	<,0001	18	76248,3336	<,0001
Khi-2 de Mantel-Haenszel	1	3773,4562	<,0001	1	50797,2156	<,0001
Coefficient Phi		0,4211			0,7098	
Coefficient de contingence		0,3881			0,5788	
V de Cramer		0,2431			0,4098	

Tableau 36 : Test d'indépendance entre le coût moyen et la fréquence

Les p-values sont dans les deux cas (prothèses dentaires et équipements optiques) inférieures au seuil de 5%, ce qui conduit au rejet de l'hypothèse H_0. Ce test ne permet donc pas de conclure à l'indépendance entre le coût moyen et la fréquence. La méthode coût*fréquence parait donc inadaptée dans la modélisation de la charge des sinistres du périmètre de l'étude.

3.1.2 Mesure de corrélation

Ce paragraphe sera dédié à la mesure des liens de corrélation entre les variables. Cela permettra de compléter l'analyse de l'indépendance réalisée précédemment entre le coût moyen et la fréquence mais aussi de vérifier les relations existantes entre nos différents variables quantitatives. Pour ce faire, il est possible d'utiliser le coefficient de Pearson ou de Sperman par exemple. Le premier mesure l'intensité de la liaison linéaire entre deux variables, alors que le 2ème mesure l'intensité d'une relation monotone. Ils sont tous les deux adaptés à des variables continues. Il convient donc de vérifier la nature de la relation entre les variables avant d'opter pour l'un ou pour l'autre coefficient. Les variables qui seront étudiées sont : le coût moyen, la fréquence, l'âge ainsi que le nombre de mois d'ancienneté dans le portefeuille. Les différents graphiques ci-dessous, correspondent au nuage des points des individus représenté sur un plan à deux dimensions et ou chaque dimension correspond à une variable :

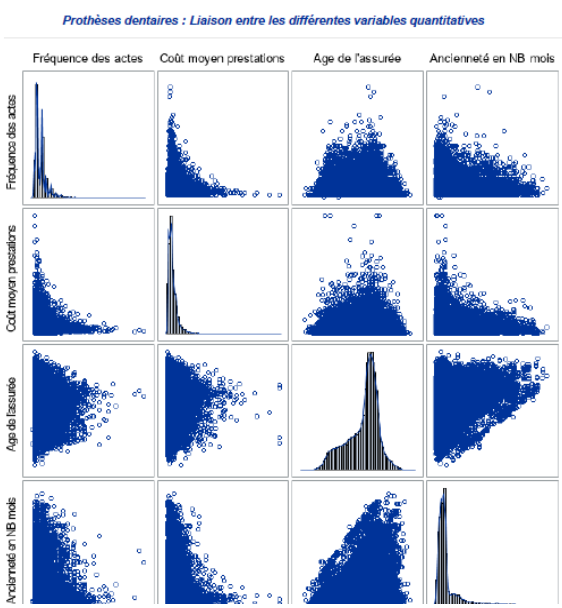


Figure 36 : Nuages des points pour les prothèses dentaires

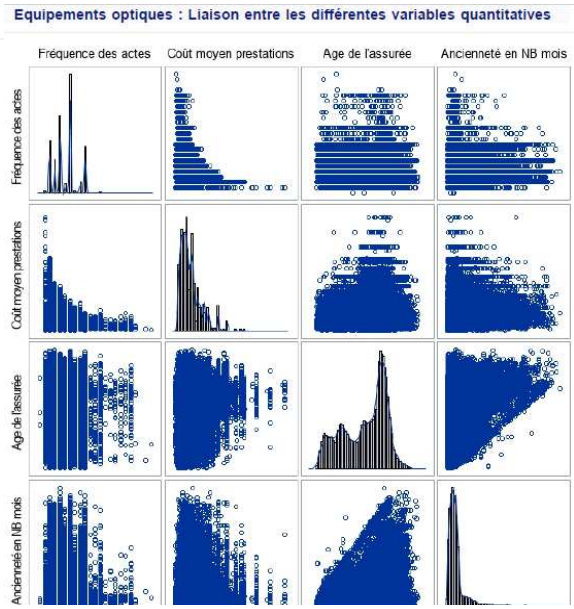


Figure 35 : Nuages des points pour les équipements optiques

En plus des nuages de points ces graphiques intègrent des histogrammes ainsi que les courbes de densité de Kernel qui représentent une méthode non paramétrique d'estimation de la densité de probabilité d'une variable aléatoire.

Aux vues des formes des différents nuages de points, il est difficile de détecter visuellement une relation linéaire entre les différentes variables. Le coefficient qui sera donc utilisé pour mesurer la corrélation sera celui de Sperman. Pour calculer sa valeur, il convient de classer chaque variable par ordre croissant et attribuer un rang à chaque valeur, la plus petite valeur prenant la valeur 1, aux ex-aequo sera attribué un rang moyen. Deux nouvelles variables sont créées : r_X représente les rangs attribués à la variable X et r_Y ceux attribués à la variable Y. Le coefficient de corrélation des rangs de Spearman est le coefficient de corrélation linéaire entre les variables r_X et r_Y.
Pour rappel, le coefficient de corrélation de Pearson pour deux variables X et Y s'écrit de la manière suivante :

$$\widehat{corr}(X,Y) = r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} = \frac{S_{XY}}{S_X S_Y}.$$

Où $x_i \; et \; y_i$ sont les valeurs prises par les variables X et Y pour l'individu i. $\bar{x} \; et \; \bar{y}$ leurs moyennes empiriques. ${S_X}^2 \; et \; {S_Y}^2$ leurs variances empiriques. S_{XY} covariance empirique et estimateur convergent de la covariance théorique : $\sigma_{XY} = [(X - E[X]) * (Y - E[Y])]$.

Le coefficient de corrélation de Spearman peut donc s'écrire comme suit :

$$\rho_S(X,Y) = r_{r_X r_Y}.$$

ρ_S : Mesure la "ressemblance" entre les deux classements. Il vérifie :

- $-1 \leq \rho_S \leq 1$;
- $\rho_S = 1$ Signifie que les deux classements sont identiques ;
- $\rho_S = -1$ Signifie que les deux classements sont inverses l'un de l'autre.

Les coefficients de corrélation entre les différentes variables sont résumés dans les tableaux suivants :

	Prothèses dentaires		Equipements optiques	
	NB_ACTE	MT_RC_MOYEN	NB_ACTE	MT_RC_MOYEN
NB_ACTE	1	0,23635	1	0,60976
		<,0001		<,0001
	62214	62212	146199	146196
MT_RC_MOYEN	0,23635	1	0,60976	1
	<,0001		<,0001	
	62212	62212	146196	146196
AGE_EX	0,08788	0,08088	0,04578	0,41819
	<,0001	<,0001	<,0001	<,0001
	62204	62202	146168	146165
NB_MOIS_ANC	0,02988	0,00332	0,10098	0,0678
	<,0001	0,4079	<,0001	<,0001
	62214	62212	146199	146196

Tableau 37 : Coefficients de corrélations de Spearman

Le coefficient de corrélation entre le coût et la fréquence est assez important aussi bien pour les équipements optiques que pour les prothèses dentaires.

3.2 Le modèle linéaire généralisé

Cette partie du mémoire sera consacrée à la modélisation de la charge des sinistres concernant les actes « prothèses dentaires » et « équipements optiques » à l'issue du déploiement la nouvelle réforme sur ce périmètre. Le modèle le plus souvent utilisé dans le monde de l'assurance pour ce type de problème est le modèle GLM. Il est adapté pour les variables continues et se base sur le même principe du modèle de régression linéaire multiple, à savoir définir la loi d'une variable à expliquer Y en fonction d'un vecteur de variables explicatives X. Le modèle en question suppose qu'il existe une relation linéaire entre la variable à expliquer et les variables explicatives. Cette relation peut s'écrire sous la forme suivante :

$$y = \beta_0 + \sum_{i=0}^{p} \beta_i * x_i + \varepsilon \,,$$

Où :

- Y est la variable à expliquer et qui prend ces valeurs dans $\mathbb{R}$;
-
- $x_1, x_2, \dots, x_p$ P variables explicatives ;
- $\beta_0, \beta_1, \dots, \beta_p$ P+1 paramètres à estimer ;
- ε est le terme d'erreur aléatoire du modèle.

Pour n observations, le modèle peut s'écrire : $y_i = \beta_0 + \sum_{j=0}^{p} \beta_j * x_{ij} + \varepsilon_i$.

Supposons que :

- ε_i est une variable aléatoire, non observée appelé terme d'erreur ;
- x_{ij} est observé et non aléatoire ;
- y_i est observé et aléatoire ;
- $E[\varepsilon_i] = 0 \iff E[y_i] = \beta_0 + \sum_{j=0}^{p} \beta_j * x_{ij}$. Les termes d'erreur sont centrés ;
- $V[\varepsilon_i] = \sigma^2 \iff V[y_i] = \sigma^2$. Il s'agit de l'hypothèse d'homoscédasticité (≃ homogénéité des variances). σ^2 est un paramètre du modèle qu'il faut estimer ;
- $\forall\, i \neq i' cov\,(\varepsilon_i\,, \varepsilon_{i'}) = 0 \iff cov\,(y_i\,, y_{i'}) = 0$. Les termes d'erreurs sont non corrélés.

Deux méthodes permettent l'estimation des paramètres $\beta_0, \beta_1, \dots, \beta_p$:

- La méthode des moindres carrés qui ne nécessite pas d'hypothèse supplémentaire sur la distribution de ε_i mais qui ne fournit pas d'estimateur pour σ^2;
- La méthode du maximum de vraisemblance qui est fondée sur la normalité de ε_i.

Une fois les paramètres estimés, des tests de nullité sont réalisés afin de mesurer la pertinence de ces derniers. La réalisation de ces tests nécessite l'introduction d'une hypothèse forte qui est la normalité de nos termes d'erreur et qui implique la normalité de la variable à expliquer , d'où tout l'intérêt du modèle linéaire généralisé qui dispense de cette hypothèse.

Le modèle linéaire généralisé s'appuie sur trois composantes :

- La variable aléatoire y dont la loi est supposée appartenir à la famille des lois exponentielles. Les lois telles que : la loi normale, la loi de poisson, la loi binomiale et la loi gamma et la loi gauss inverse appartiennent toutes à cette famille et leurs densités peuvent toutes s'écrire sous la forme suivante :

  $$f(y_i, \theta_i, \varphi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\right) \ ;$$

  Les fonctions a, b et c sont spécifiées en fonction du type de loi exponentielle. θ_i est appelé paramètre canonique et est supposé inconnu. Le paramètre φ est appelé paramètre de dispersion et est supposé connu.
  Par exemple si y_i suit une loi exponentielle de paramètre λ, les relations suivantes peuvent être déduites :

$$\theta_i = -\lambda = -\frac{1}{E[y_i]} \ ; \ a(\varphi) = 1 \ ; \ b(\theta_i) = 0 \ et \ c(y_i, \varphi) = \ln(\lambda) \Longrightarrow f(y_i, \theta_i, \varphi) = \lambda e^{-\lambda y_i} \ .$$

- $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ vecteur de variables déterministes appelées variables explicatives de y_i ;
- La fonction lien noté g qui décrit la relation entre la combinaison linéaire des variables ; explicatives et l'espérance mathématique de la variable à expliquer. Son choix est libre et peut être lié à la nature de la variable à expliquer, ce qui permet de modéliser l'espérance μ (cas de régression linéaire) ou modéliser une fonction monotone et différentiable g(μ). Dans ce dernier cas, le modèle s'écrit sous la forme :

  $$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \ .$$

  Les fonctions utilisées le plus souvent sont :

  - ❖ La fonction identité ce qui revient à modéliser :

    $$E[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \ ;$$

    Et ramène à un modèle additif

  - ❖ La fonction logarithme népérien qui se traduit par $g(\mu) = log(\mu)$. Elle est adaptée pour les modèles log-linéaires. Le modèle s'écrit donc sous la forme :

    $$\ln(E[y_i]) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \ ,$$
    $$\Leftrightarrow \ E[y_i] = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}} ;$$

    Le modèle devient donc un modèle multiplicatif

❖ La fonction logit avec $g(\mu) = log\left(\frac{1}{1-\mu}\right)$ qui modélise le rapport du log de la chance. Elle est adaptée pour les cas ou μ est comprise entre 0 et 1. Par exemple la probabilité de succès d'une loi binomiale.
La fonction lien qui associe la moyenne u_i au paramètre naturel est appelée fonction lien canonique. Nous obtenons dans ce cas :

$$g(u_i) = \theta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} .$$

Une fois la loi et la fonction de lien identifiées, il faut déterminer les paramètres β_i. La méthode la plus souvent utilisée est celle du maximum de vraisemblance dont le cadre est le suivant :

- Soit $y_1, y_2, \dots, y_n$ n observations de la variable y et $x_1, x_2, \dots, x_p$ p variables explicatives ;
- $f(x, \theta)$ la densité de la variable y.

La vraisemblance s'écrit alors de la manière suivante :

$$L(y_1, y_2, \dots, y_n, \theta) = \prod_{i=1}^{n} f(y_i, \theta) \Leftrightarrow \ln\big(L(y_1, y_2, \dots, y_n, \theta)\big) = \sum_{i=1}^{n} \ln\left(f(y_i, \theta)\right) .$$

La recherche porte sur le paramètre θ qui maximise le log de la vraisemblance. Cette valeur est appelée maximum du log-vraisemblance et vérifiée deux conditions :

$$\frac{\partial \ln\big(L(y_1, y_2, \dots, y_n, \theta)\big)}{\partial \theta} = 0 ,$$

Et

$$\frac{\partial^2 \ln\big(L(y_1, y_2, \dots, y_n, \theta)\big)}{\partial \theta^2} < 0 .$$

Notons l_i la contribution de la ième observation à la log-vraisemblance. Cette contribution est définie par :

$$\ell_i = \ell(y_i, \theta_i, \varphi) = \ln\left(f(y_i, \theta_i, \varphi)\right) ,$$

Ce qui signifie que :

$$\ell(y_i, \theta_i, \varphi) = \frac{y_i - b'(\theta_i)}{a(\varphi)} + c(y_i, \varphi),$$

Et que :

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\varphi)} \; ; \; \frac{\partial^2 \ell}{\partial {\theta_i}^2} = -\frac{b"(\theta_i)}{a(\varphi)} .$$

Comme il s'agit d'une loi issue de structures exponentielles qui vérifie les conditions de régularité, nous pouvons écrire :

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \; et - E(\frac{\partial^2 l}{\partial \theta^2}) = E\left(\frac{\partial l}{\partial \theta}\right)^2 ,$$

$$\Rightarrow E\left(\frac{\partial l}{\partial \theta_i}\right) = E\left(\frac{y_i - b'(\theta_i)}{a(\varphi)}\right) = \frac{1}{a(\varphi)}\big(E(y_i) - b'(\theta_i)\big) = 0 ,$$

$$\Rightarrow E(y_i) = b'^{(\theta_i)} = \mu_i .$$

De même

$$-E\left(-\frac{b\,(\theta_i)}{a(\varphi)}\right) = E\left(\frac{y_i - b'^{(\theta_i)}}{a(\varphi)}\right)^2,$$

$$\Rightarrow \frac{b"(\theta_i)}{a(\varphi)} = \frac{1}{a^2(\varphi)} E\big(y_i - E(y_i)\big)^2 = \frac{Var(y_i)}{a^2(\varphi)},$$

$$\Rightarrow b"(\theta_i) * a(\varphi) = Var(y_i)\,.$$

Soit maintenant le vecteur β composé des p paramètres du modèle. Estimer ce vecteur revient à calculer :

$$\frac{\partial \ell_i}{\partial \beta_j}.$$

Cette dérivée partielle peut s'écrire de la manière suivante :

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} * \frac{\partial \theta_i}{\partial \beta_j}.$$

Nous supposons que $g(\mu_i)$ est la fonction de lien canonique. Ceci se traduit par :$g(\mu_i) = \theta_i$

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} * \frac{\partial g(\mu_i)}{\partial \beta_j},$$

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \mu_i}{a(\varphi)} \; et \; \frac{\partial g(\mu_i)}{\partial \beta_j} = x_{ij}\,,$$

Ce qui donne finalement :

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\varphi)} * x_{ij}\,.$$

Ces équations sont généralement non linéaires et les estimateurs des paramètres n'ont pas d'expressions bien définies. Pour les estimer, il faut avoir recours à des logiciels qui utilisent des algorithmes itératifs pour la résolution d'équations non linéaires.

Par ailleurs, une fois le choix du modèle et l'estimation des paramètres réalisés, il est nécessaire d'effectuer d'autres actions afin de juger de la pertinence du modèle. Des actions comme :

- Le calcul des statistiques permettant d'apprécier l'adéquation du modèle aux données ;
- Les tests d'hypothèses concernant les coefficients du modèle ;
- La construction d'intervalles de confiance pour les coefficients du modèle ;
- L'estimation de la moyenne ;
- L'analyse des résidus.

3.2.1 Définition des paramètres du modèle

Le test d'indépendance entre le coût moyen et le nombre d'actes ainsi que l'analyse de corrélation ont permis de mettre en évidence un lien entre ces deux variables. L'existence de ce lien ne permet pas l'utilisation du modèle coût*fréquence. Il faudra donc faire appel à la méthode alternative qui consiste à modéliser la charge globale par individu. La réalisation de cette modélisation à l'aide du modèle GLM nécessite de faire une hypothèse concernant la distribution à modéliser. Parmi les différentes distributions évoquées dans la littérature, celles de Tweedie tiennent une place importante, surtout lorsqu'il s'agit d'études actuarielles. Il s'agit d'une sous-catégorie de la famille des lois exponentielles qui possède une densité de distribution qui s'exprime sous la forme suivante :

$$f(y,\mu,\phi) = a(y,\phi) * \exp\{\frac{1}{\phi}[y\theta(\mu) - \kappa(\theta(\mu)]\},$$

Où

$$\theta(\mu) = f(x) = \begin{cases} \frac{\mu^{1-\gamma}}{1-\gamma}, & \gamma \neq 1 \\ \log(\mu), & \gamma = 1 \end{cases} et\ \kappa\big(\theta(\mu)\big) = \begin{cases} \frac{\mu^{2-\gamma}}{2-\gamma}, & \gamma \neq 2 \\ \log(\mu), & \gamma = 2 \end{cases}.$$

La densité dépend de trois paramètres qui sont : la moyenne μ, la dispersion ϕ et γ paramètre de puissance Tweedie. La variance, elle, est une fonction de la moyenne, ce qui constitue un avantage. En effet, pour Y variable aléatoire de distribution de Tweedie, la variance est définie par :

$$Var(Y) = \phi\mu^{\gamma}.$$

D'ailleurs, plusieurs lois usuelles appartiennent à cette catégorie, parmi lesquelles nous pouvons citer :

- La loi normale qui correspond à un $\gamma = 0$;
- La loi de poisson qui correspond à un $\gamma = 1$;
- La loi Gamma qui correspond à un $\gamma = 2$.

Dans le cas où γ est compris entre 1 et 2 nous dirons que Y suit une loi de poisson composée avec des sauts de Gamma. La densité s'écrit alors de la manière suivante :

$$f(y,\mu,\phi) = a(y,\phi) * \exp\{\frac{1}{\phi}[y\frac{\mu^{1-\gamma}}{1-\gamma} - \frac{\mu^{2-\gamma}}{2-\gamma}]\},$$

Avec

$$a(y,\phi) = \frac{1}{y}\sum_{j=1}^{\infty} \frac{y^{-j\alpha}(\gamma-1)^{\alpha j}}{\phi^{j(1-\alpha)}(2-\gamma)^{j} j!\,\Gamma(-j\alpha)}\ avec\ \alpha = \frac{2-\gamma}{1-\gamma}.$$

L'expression de la densité de la famille de lois exponentielle s'obtient avec les paramètres suivants :

$$\theta_i = \frac{\mu^{1-\gamma}}{1-\gamma}\ et\ b(\theta_i) = \frac{1}{2-\gamma}[(1-\gamma)\theta_i]^{\frac{2-\gamma}{1-\gamma}}.$$

Cette catégorie présente aussi l'avantage de pouvoir intégrer dans la modélisation un volume de probabilités nulles, ce qui est souvent le cas des données de sinistralités issues des portefeuilles d'assurance et s'expliquant par le fait que tous les assurés ne sont pas amenés à connaitre des sinistres durant un exercice. Toutefois, et par prudence, nous n'allons pas nous appuyer sur cette caractéristique. En effet, comme le montre l'analyse de la fréquence de consommation réalisée

précédemment, la part des assurés sans sinistre est très importante ce qui peut conduire à surestimer la fréquence des valeurs proches de zéro et sous-estimer les queues de distribution. Pour pallier ce problème, nous allons procéder à deux modélisations :

- Modélisation de la variable $X_i * 1_{T_i=1}$ représentant la charge de sinistre annuelle de l'individu i pour les assurés ayant connu un sinistre au cours de l'exercice. Nous supposons que cette variable suit une distribution de Tweedie ;
- Modélisation de la variable T_i qui suit une loi de Bernoulli et prend la valeur 1 si l'assuré i a connu au moins un sinistre durant l'exercice et 0 sinon.

La loi de T_i sera modélisée à l'aide de la régression logistique qui constitue un cas particulier des modèles linéaires généralisées. Pour rappel :

$$T_i \sim B(p) \Longrightarrow P(T_i = y) = p^y(1-p)^{1-y} = \exp\left[y * log\left(\frac{p}{1-p}\right) + log(1-p)\right] et\ E[T_i] = p.$$

Les paramètres de la densité de la famille exponentielle sont les suivants :

$$\theta_i = log\left(\frac{p}{1-p}\right)\ ;\ b(\theta_i) = -log(1-p) = \log\left(1+e^{\theta_i}\right)\ ;\ a(\varphi) = 1.$$

La fonction de lien naturel est la fonction LOGIT définie par :

$$logit(p) = log\left(\frac{p}{1-p}\right).$$

Ce qui signifie que le modèle s'écrie sous la forme

$$log\left(\frac{E[T_i]}{1-E[T_i]}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},$$

$$\Rightarrow E[T_i] = \frac{1}{1+e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}}$$

Une fois l'estimation des paramètres composants les vecteurs β associés aux variables $X_i * 1_{T_i=1}$ et T_i réalisée, il suffira d'appliquer la formule d'espérance totale pour calculer l'espérance de X_i. Pour rappel, la formule des espérance totales est la suivante :

$$Pour\ X\ v.a\ discréte\ prenant\ ces\ valeurs\ dans\ l'ensemble\ \{1,2,\dots,n\}\ et\ Y\ v.a.r$$

$$E[Y] = \sum_{i=1}^{n} P(X=i).E[Y|X=i)],$$

Ce qui revient dans notre cas à :

$$E[X_i] = P(T_i = 1).E[X_i|T_i = 1)] + P(T_i = 0).E[X_i|T_i = 0)],$$

Sachant que

$$E[X_i|T_i = 0)] = 0,$$

Nous obtenons la formule suivante :

$$\Rightarrow E[X_i] = P(T_i = 1).E[X_i|T_i = 1)].$$

Remarque : Si l'objectif de la régression logistique est de prédire la classe de la variable T_i, il faut choisir un point de coupure c (qui peut être optimisé) et utiliser la règle suivante :

$$si\ \hat{p} < c\ \Rightarrow\ \hat{T}_i = 0\ .\ Sinon\ \hat{T}_i = 1\ avec\ \hat{p} = \hat{\mu}.$$

3.2.2 Choix des données à modéliser

Les différents tests de comparaison effectués montrent une évolution de la consommation en 2020. Une partie de cette évolution peut s'expliquer par la crise sanitaire qui a eu pour effet général une baisse de la sinistralité du fait des différents confinements. Toutefois, ces mêmes tests détectent sur certains segments et pour certains soins des tendances haussières. Par exemple, les prestations liées aux prothèses dentaires connaissent une nette augmentation en 2020. Les deux graphiques suivants permettent de visualiser cet effet :

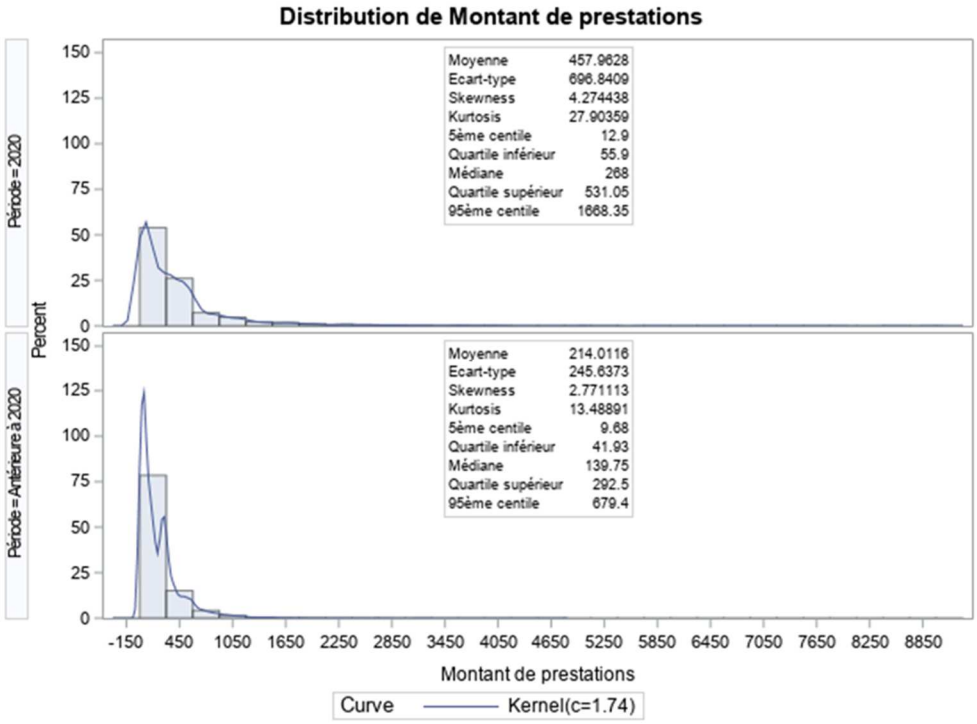


Figure 37: distribution des prestations des prothèses dentaires pour l'entrée de gamme

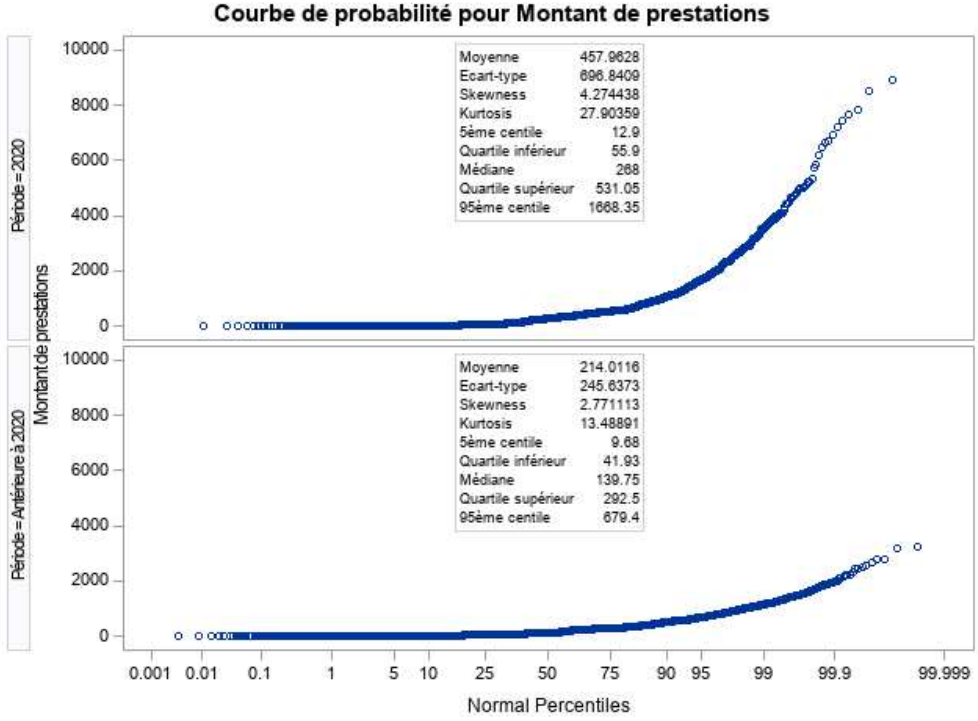


Figure 38: Courbe de probabilité des prestations des prothèses dentaires de l'entrée de gamme

Nous pouvons contraster une nette différence entre les deux distributions des variables : montant des prestations 2020 et montant des prestations des exercices antérieurs à 2020. Ces différences se traduisent par une hausse en 2020 des différents indicateurs : moyenne, écart-type, coefficient d'asymétrie (Skewness), coefficient d'aplatissement (kurtoris), quantiles… et d'un changement de l'allure de la densité de distribution estimé à l'aide de la méthode de noyau (une description de cette méthode ainsi que la définition des coefficients d'asymétrie et d'aplatissement sont disponibles en annexe 5). Dès lors, nous pouvons penser que l'utilisation des données historiques pour modéliser la charge des sinistres future conduira à une sous-estimation de cette dernière. Pour remédier à cela, nous avons décidé de modéliser la nouvelle charge à partir de 2020 et d'utiliser les données antérieures pour modéliser la charge des sinistres sur les années passées. Les données 2020 peuvent elles aussi conduire à une sous-estimation de la variable du fait des perturbations liées à la crise. Nous faisons l'hypothèse que cet effet a impacté uniquement la fréquence de consommation avec une diminution de la part des assurés ayant eu recours aux soins et que la sinistralité des assurés ayant consommé n'a pas été déformée.

Les données utilisées dans le cadre de la modélisation sont :

- Variable à expliquer : Montant annuel des prestations qui correspond à notre variable explicative
- Variables explicatives :
    - Zone géographique : 5 modalités ;
    - Tranche d'âge : 13 modalités ;
    - Gamme : 3 modalités ;
    - Ancienneté : Nombre d'années d'anciennetés ;
    - Sexe ;
    - Type de personnes protégées : 4 modalités.

  Ces données ont toutes été recodifiées selon les modalités présentées en annexe.

Enfin les données de modélisation seront réparties en deux échantillons :

- Le premier, appelé échantillon d'apprentissage, sera composé de 70% des individus qui seront sélectionnés de manière aléatoire. Il permettra d'ajuster le modèle, de calculer les différentes statistiques et d'estimer le vecteur β ;
- Le second, appelé échantillon de validation, sera composé des individus restants (soit 30%). Il permettra de comparer les estimations obtenues à l'aide du modèle de la variable Y à ces valeurs réelles.

3.2.3 Résultats et adéquations des modèles

Ce paragraphe sera consacré à la présentation des statistiques permettant d'ajuster ou de juger de la pertinence des modèles retenus ainsi que des résultats obtenus à l'aide de ses modèles.

En amont de la présentation de ces résultats, nous allons tout d'abord rappeler quelques notions qui seront utilisées par la suite dans l'interprétation des sorties du logiciel :

- AIC (critère d'information d'Akaike) : il s'agit d'un critère qui peut aider à apprécier la qualité du modèle lorsque l'estimation des paramètres du modèle se fait à l'aide de la méthode du maximum de vraisemblance. Il varie en fonction du nombre des variables explicatives retenues, se présente sous la forme d'une fonction de la log-vraisemblance et intègre une pénalité pour le nombre de coefficients β. Son expression est la suivante :

$$AIC = -2\ell(\hat{\beta}) + 2p \text{ où } p: nombre\ de\ paramètres\ \beta\ et\ \ell\ la \log vraisemblance.$$

Nous pouvons constater qu'il est composé de deux termes qui vont avoir des variations de sens opposés lors du rajout d'un paramètre et peut donc aussi bien augmenter que baisser. Il connaitra une baisse uniquement si la baisse du premier terme est suffisante pour compenser le fait que le terme 2p augmente à 2(p +1). La règle à retenir lors de l'utilisation de ce critère est que : plus sa valeur est petite, meilleure est l'adéquation du modèle.

- Le Test de l'hypothèse nulle globale $\beta = 0$, qui teste l'hypothèse $\mathcal{H}_0$: les paramètres sont nuls, contre l'hypothèse alternative : au moins un des paramètres est différent de zéro ;
- Les p-values correspondant aux tests des hypothèses $\mathcal{H}_0 : \beta_j = 0$ versus $\mathcal{H}_1 : \beta_j \neq 0$. Il renseigne sur l'effet de la variable x_j. Si cette valeur est inférieure au seuil de test (à savoir 0.05 dans notre cas), nous pourrons conclure que l'effet de la variable est statistiquement différent de zéro.
- Le test du rapport de vraisemblance : Il découle de la méthode d'estimation du maximum de vraisemblance et est donc généralement applicable lorsque nous estimons les paramètres avec cette méthode. Il consiste à ajuster deux modèles emboîtés :
    - Le premier modèle, dit modèle complet, contient tous les paramètres du vecteur β dont l'estimateur est noté $\hat{\beta}$;
    - Le deuxième modèle, appelé modèle réduit, correspond à l'hypothèse nulle $\mathcal{H}_0$, contient tous les paramètres avec les restrictions imposées sous $\mathcal{H}_0$. L'estimateur dans ce cas sera noté $\hat{\beta}_0$.

Le test est basé sur la statistique :

$$D = -2\{\ell(\hat{\beta}_0) - \ell(\hat{\beta})\}.$$

Autrement dit, c'est la différence entre -2 Log L pour le modèle réduit et -2 Log L pour le modèle complet. Cette différence D, lorsque l'hypothèse $\mathcal{H}_0$ est vraie, suit approximativement une loi khi-deux avec un nombre de degrés de liberté égal au nombre de paramètres testés (le nombre de restrictions sous $\mathcal{H}_0$). Nous pouvons donc calculer la p-value en utilisant la distribution du khi-deux.

- Offset et exposition : lorsque la variable à expliquer dans le cas d'un modèle linéaire généralisé dépend également linéairement d'une autre variable, cette dernière est appelée offset ou variable de décalage. Lorsque la variable à expliquer est la charge annuelle des sinistres, nous pouvons naturellement prendre comme variable offset l'exposition qui représente le temps de présence de l'individu dans le portefeuille. En effet, nous pouvons supposer que la charge annuelle est proportionnelle à l'exposition. L'exposition dans notre cas est égale au PP moyen. Les quelques lignes suivantes expliquent comment ces termes modifient le modèle. Reprenons le modèle avec une fonction lien log et dans lequel l'exposition sera notée B et considérée comme variable déterministe. Le modèle peut s'écrire sous la forme :

$$\begin{aligned}
&\ln\left(\frac{E[y_i]}{B}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\ , \\
&\Leftrightarrow \frac{E[y_i]}{B} = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}\ , \\
&\Leftrightarrow E[y_i] = B * e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}, \\
&\Leftrightarrow E[y_i] = \exp\left[\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \log(B)\right].
\end{aligned}$$

Dans nos différents modèles utilisant un offset, ce dernier sera égal à log(exposition)

- Approche Stepwise : Il s'agit d'une approche pas à pas qui vérifie que l'ajout d'une variable ne provoque pas la suppression d'une variable déjà introduite. La première itération contient le modèle avec l'intercepte uniquement, à la seconde itération la première variable citée dans le modèle est ajoutée et ainsi de suite jusqu'à parcourir dans l'ordre l'ensemble des variables.

3.2.3.1 Recours aux soins

Comme précisé précédemment, la modélisation choisie nécessite de modéliser la variable recours aux soins. Cette variable appelée T_i prend deux valeurs : 1 si l'assuré a eu recours aux soins lors de l'exercice en question et 0 sinon. Cette section sera consacrée à la présentation des différents résultats obtenus pour les prothèses dentaires et les équipements optiques. En plus d'avoir distingué les deux types de soins lors de la modélisation, nous avons aussi réalisé une étude par niveau de gamme. Ce choix a été motivé par la différence notable de consommation selon le niveau de gamme dont l'assuré bénéficie.

- **Prothèses dentaires**

Pour chaque niveau de gamme deux modèles ont été réalisés. Un modèle avec les données 2020 qui donne une estimation du recours aux soins en 2020, et un modèle avec les données antérieures à 2020 qui donne une estimation du recours aux soins pour les années antérieures. Les résultats concernant le deuxième modèle sont disponibles en annexe 7.

Entrée de gamme

Comme le montre le tableau ci-dessous, les p-values du test de l'hypothèse nulle globale nous conduit à rejeter l'hypothèse nulle ce qui implique qu'au moins une des variables explicatives a un effet statistiquement différent de zéro.

Test de l'hypothèse nulle globale : BETA=0			
Test	*khi-2*	*DDL*	*Pr > khi-2*
Rapport de vrais	3 562,5	28	<.0001
Score	2 912,1	28	<.0001
Wald	1 474,9	28	<.0001

Tableau 38 : Test d'hypothèse nulle globale. Soins : P. dentaires ; Segment : Entrée

La méthode Stepwise a permis par la suite de sélectionner les variables suivantes :

Récapitulatif sur la sélection Stepwise						
	Effet			*Nombre*	*Khi-2*	*Khi-2*	
Etape	*Saisi*	*Supprimé*	*DDL*	*dans*	*du score*	*de Wald*	*Pr > khi-2*
1	CD_TRANCHE_AGE		12	1	2 641,9		<.0001
2	NIVEAU		1	2	161,0		<.0001
3	ANNEE_ANC		8	3	75,2		<.0001
4	CD_SEXE		1	4	16,0		<.0001
5	CD_ZONE		6	5	20,8		0.0020

Tableau 39: Variables retenues. Soins : P. dentaires ; Segment : Entrée

Les p-values de l'ensemble des variables sont très significatives sauf pour la variable « zone géographique » que nous pouvons décrire comme moins significative que les autres. Ces statistiques nous conduisent à conclure que l'ensemble des variables explicatives ont des effets statistiquement différents de zéro sur le recours aux soins.

L'analyse du critère AIC confirme aussi la pertinence de la sélection de l'ensemble des variables comme le montre le tableau ci-dessous :

		Statistique d'ajustement du modèle
2ème itération	AIC	29 673
	-2 Log L	29 667
4ème itération	AIC	29 431
	-2 Log L	29 415
6ème itération	AIC	29 421
	-2 Log L	29 401

Tableau 40: Critère AIC par itération. Soins : P. dentaires ; Segment : Entrée

Le modèle obtenu à la 6ème itération possède le critère AIC le plus petit, ce qui prouve que le meilleur ajustement est obtenu avec ce modèle. La statistique D (utilisée par le test du rapport de vraisemblance) entre le modèle complet et le modèle sous contrainte obtenu à la 4ème itération vaut 16 et possède une p-value de 0,00033 inférieur au seuil de test de 0,05, ce qui conduit à rejeter l'hypothèse nulle et à privilégier le modèle complet.
Le logiciel produit aussi des statistiques permettant de comparer l'effet de chaque niveau de variable comme le montre le tableau suivant :

	Estimation du rapport de cotes		
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_SEXE 1 vs 2	0.878	0.823	0.937
ANNEE_ANC 0 vs 8	0.686	0.466	1.010
ANNEE_ANC 1 vs 8	0.977	0.665	1.436
ANNEE_ANC 2 vs 8	0.973	0.661	1.433
ANNEE_ANC 3 vs 8	0.973	0.660	1.433
ANNEE_ANC 4 vs 8	1.020	0.691	1.504
ANNEE_ANC 5 vs 8	1.105	0.749	1.631
ANNEE_ANC 6 vs 8	0.994	0.672	1.470
ANNEE_ANC 7 vs 8	0.956	0.640	1.429
CD_TRANCHE_AGE 1 vs 13	<0.001	<0.001	>999.999
CD_TRANCHE_AGE 2 vs 13	0.034	0.021	0.054
CD_TRANCHE_AGE 3 vs 13	0.148	0.111	0.198
CD_TRANCHE_AGE 4 vs 13	0.225	0.177	0.287
CD_TRANCHE_AGE 5 vs 13	0.344	0.274	0.432
CD_TRANCHE_AGE 6 vs 13	0.477	0.381	0.597
CD_TRANCHE_AGE 7 vs 13	0.618	0.496	0.770
CD_TRANCHE_AGE 8 vs 13	0.749	0.605	0.927
CD_TRANCHE_AGE 9 vs 13	0.873	0.707	1.079
CD_TRANCHE_AGE 10 vs 13	1.002	0.833	1.205
CD_TRANCHE_AGE 11 vs 13	1.151	0.962	1.378
CD_TRANCHE_AGE 12 vs 13	1.302	1.071	1.584
NIVEAU 1 vs 2	0.671	0.628	0.717
CD_ZONE 1 vs 7	0.531	0.342	0.825
CD_ZONE 2 vs 7	0.651	0.428	0.992
CD_ZONE 3 vs 7	0.630	0.418	0.949
CD_ZONE 4 vs 7	0.650	0.430	0.982
CD_ZONE 5 vs 7	0.606	0.359	1.022
CD_ZONE 6 vs 7	0.732	0.484	1.108

Tableau 41: Rapport des côtes. Soins : P. dentaires ; Segment : Entrée

Les résultats de ce tableau peuvent être interprétés de la manière suivante :

- Un homme a moins de chance de recourir à des soins de type prothèses dentaires ;
- Les assurés avec une ancienneté inférieure à 4 ans ou supérieur à 5 ans auront moins recours aux soins concernés en comparaison avec une ancienneté de 8 ans ;
- Les tranches d'âge 10 et 11 auront plus de chance de recourir aux soins que la dernière tranche. Les autres en auront moins ;
- Le niveau de garantie 1 aura moins de chance d'avoir recours aux soins ;
- Les assurés de la zone 7 ont plus de chance de recourir aux soins que ceux des autres zones.

Une autre manière de juger de la qualité du modèle est de comparer les valeurs réelles aux valeurs prédites comme c'est le cas dans le tableau suivant :

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	74.8	D de Somers	0.497
Pourcentage discordant	25.1	Gamma	0.497
Pourcentage lié	0.1	Tau-a	0.053
Paires	296 278 548	c	0.748

Tableau 42 : % de concordance. Soins : P. dentaires ; Segment : Entrée

Nous pouvons constater que le pourcentage des concordants est de 74.8%. Le même modèle obtient le score de 94% sur l'échantillon de validation comme le montre tableau suivant :

		Consommateurs prédits NON Nombre d'assurés	
		Nombre	Fréquence
Consommateurs réels	Bien classé		
NON	OUI	30 150	100
	Total	30 150	94
OUI	Bien classé		
	NON	1 793	100
	Total	1 793	6

Tableau 43 : Echantillon de validation : Prédits / observés. Soins : P. dentaires ; Segment : Entrée

Malgré un bon score des biens classés, nous pouvons constater que tous les assurés ont été classés comme non-consommateurs. Ceci est dû au taux très élevé des assurés qui n'ont pas recours aux soins. Cette remarque restera valable pour l'ensemble des modèles. Son impact sur l'estimation de la charge annuelle restera toutefois limité car ce qui nous intéresse finalement c'est la probabilité de recours aux soins et non la classification en elle même

Milieu de gamme

Les variables retenues par la méthode Stepwise sont les suivantes :

Récapitulatif sur la sélection Stepwise						
Etape	Effet Saisi	Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
1	CD_TRANCHE_AGE		12	1	1 490,35		<.0001
2	NIVEAU		1	2	28,44		<.0001
3	ANNEE_ANC		8	3	43,40		<.0001
4	CD_SEXE		1	4	7,97		0.0048
5	CD_TYPE_PP		2	5	9,31		0.0095

Tableau 44: Variables retenues. Soins : P. dentaires ; Segment : Milieu

L'ensemble des variables explicatives retenues possèdent des p-values très significatives sauf pour le sexe et le type de personnes protégées ayant une p-value que nous pouvons décrire comme moins significative que les autres. Ces statistiques nous conduisent à conclure que ces variables explicatives ont des effets statistiquement différents de zéro avec un effet moins important du sexe et du type de personnes protégées.

A l'aide du tableau ci-dessous, nous pouvons comparer le modèle retenu (modèle sous contrainte) au modèle complet :

	Statistique d'ajustement du modèle			
	Modèle sous contrainte	Modèle complet		Test du rapport de la vraisemblance
AIC	26 854,8	26 854,0	D	2,779
-2 Log L	26 838,8	26 836,0	P-value	0,24

Tableau 45 : Mesure de la qualité du modèle. Soins : P. dentaires ; Segment : Milieu

Nous pouvons constater que, bien que le modèle complet possède un critère AIC plus faible que le modèle réduit (ou modèle sous contrainte), c'est bien le deuxième qui a été retenu. Le test du rapport de vraisemblance valide toutefois ce choix car la p-value associée à la statistique D vaut 0,24 bien au-dessus du seuil 0,05 ce qui ne permet pas de rejeter l'hypothèse nulle.
Les statistiques permettant l'analyse des effets sont résumées dans le tableau suivant :

	Estimation du rapport de cotes		
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_TYPE_PP 2 vs 4	0.926	0.852	1.006
CD_TYPE_PP 3 vs 4	0.509	0.296	0.876
CD_SEXE 1 vs 2	0.909	0.853	0.970
ANNEE_ANC 0 vs 8	1.291	0.892	1.868
ANNEE_ANC 1 vs 8	1.335	0.924	1.930
ANNEE_ANC 2 vs 8	1.190	0.822	1.721
ANNEE_ANC 3 vs 8	1.100	0.760	1.592
ANNEE_ANC 4 vs 8	1.065	0.736	1.541
ANNEE_ANC 5 vs 8	1.075	0.743	1.555
ANNEE_ANC 6 vs 8	0.986	0.680	1.431
ANNEE_ANC 7 vs 8	0.972	0.665	1.422
CD_TRANCHE_AGE 1 vs 13	0.005	<0.001	0.041
CD_TRANCHE_AGE 2 vs 13	0.044	0.020	0.094
CD_TRANCHE_AGE 3 vs 13	0.192	0.120	0.308
CD_TRANCHE_AGE 4 vs 13	0.275	0.186	0.407
CD_TRANCHE_AGE 5 vs 13	0.392	0.269	0.572
CD_TRANCHE_AGE 6 vs 13	0.493	0.340	0.715
CD_TRANCHE_AGE 7 vs 13	0.568	0.392	0.823
CD_TRANCHE_AGE 8 vs 13	0.670	0.465	0.964
CD_TRANCHE_AGE 9 vs 13	0.751	0.524	1.076
CD_TRANCHE_AGE 10 vs 13	0.910	0.646	1.283
CD_TRANCHE_AGE 11 vs 13	1.104	0.785	1.552
CD_TRANCHE_AGE 12 vs 13	1.289	0.899	1.848
NIVEAU 3 vs 4	0.823	0.767	0.883

Tableau 46 : Rapport des côtes. Soins : P. dentaires ; Segment : Milieu

Les résultats de ce tableau peuvent être interprétés de la manière suivante :

- Un homme a moins de chance de recourir à des soins de type prothèses dentaires ;
- Les assurés avec une ancienneté inférieure à 6 auront plus recours aux soins concernés en comparaison avec une ancienneté de 8 ans ;
- Les tranches d'âge 10 et 11 auront plus de chance de recourir aux soins que la dernière tranche. Les autres en auront moins ;
- Le niveau de garantie 3 aura moins de chance d'avoir recours aux soins ;
- Un assuré avec un statut « conjoint » ou « enfant » aura moins de chance d'avoir recours aux soins qu'un assuré avec un statut « autre ».

L'analyse de la concordance montre que ce modèle est moins performant que celui du segment précédent comme précisé dans les deux tableaux suivants :

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	67.1	D de Somers	0.348
Pourcentage discordant	32.4	Gamma	0.349
Pourcentage lié	0.5	Tau-a	0.068
Paires	168 807 028	c	0.674

Tableau 48 : % de concordance. Soins : P. dentaires ; Segment : Milieu

		Consommateurs prédits NON Nombre d'assurés	
		Nombre	Fréquence
Consommateurs réels	Bien classé		
NON	OUI	15 815	100
	Total	15 815	89
OUI	Bien classé		
	NON	1 951	100
	Total	1 951	11

Tableau 47 : Echantillon de validation : Prédits / observés. Soins : P. dentaires ; Segment : Milieu

Le score des biens classés est de 67% pour l'échantillon d'apprentissage et de 89% pour l'échantillon de validation.

Haut de gamme

Les variables retenues par la méthode Stepwise sont les suivantes :

Récapitula						
Etape	Effet Saisi	Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
1	CD_TRANCHE_AGE		12	1	326,7		<.0001
2	CD_TYPE_PP		2	2	12,2		0.0022
3	NIVEAU		1	3	7,4		0.0067

Tableau 49 : Variables retenues. Soins : P. dentaires ; Segment : Haut

Seules trois variables explicatives ont été retenues pour ce segment, à savoir la tranche d'âge, le type de personnes protégées et le niveau de couverture. Pour vérifier la pertinence de ce choix, il est nécessaire d'analyser les valeurs du critère AIC ainsi que les résultats du test du rapport de vraisemblance. Le tableau ci-dessous fournit ces informations :

	Statistique d'ajustement du modèle			Test du rapport de la vraisemblance
	Modèle sous contrainte	Modèle complet		
AIC	3 525,3	5 226,5	D	-1719,201
-2 Log L	3 493,3	5 212,5	P-value	1

Tableau 50 : Mesure de la qualité du modèle. Soins : P. dentaires ; Segment : Haut

D'une part, le modèle complet possède un critère AIC plus grand que le modèle réduit (ou modèle sous contrainte), d'autre part, la p-value associée à la statistique D vaut 1 ce qui ne permet pas de rejeter

l'hypothèse nulle. Une analyse des p-values du modèle complet aurait aussi d'ailleurs conduit au même résultat comme le montre le tableau suivant :

Analyse des valeurs estimées du Paramètre	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept	1	-6,6021	0,4963	176,9709	<,0001
CD_TYPE_PP	1	0,3067	0,0609	25,3288	<,0001
CD_SEXE	1	0,00643	0,0725	0,0079	0,9293
ANNEE_ANC	1	0,0249	0,0169	2,1563	0,142
CD_TRANCHE_AGE	1	0,2429	0,0139	306,7914	<,0001
NIVEAU	1	0,2786	0,0727	14,6891	0,0001
CD_ZONE	1	0,0302	0,0292	1,0688	0,3012

Tableau 51: P-values modèle complet. Soins : P. dentaires ; Segment : Haut

Toutes les variables explicatives non retenues par la méthode Stepwise possèdent des p-values supérieures au seuil de 0,05.

Les résultats de l'analyse des effets sont résumés dans le tableau suivant :

Estimation du rapport de Effet	Estimation du point	Intervalle de confiance de Wald	
CD_TYPE_PP 2 vs 4	0.608	0.453	0.815
CD_TYPE_PP 3 vs 4	0.648	0.231	1.818
CD_TRANCHE_AGE 1 vs 13	1.726	<0.001	>999.999
CD_TRANCHE_AGE 2 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 3 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 4 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 5 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 6 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 7 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 8 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 9 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 10 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 11 vs 13	>999.999	<0.001	>999.999
CD_TRANCHE_AGE 12 vs 13	>999.999	<0.001	>999.999
NIVEAU 5 vs 6	0.788	0.664	0.936

Tableau 52 : Rapport des côtes. Soins : P. dentaires ; Segment : Haut

Les résultats de ce tableau peuvent être interprétés de la manière suivante :

- Les assurés avec un statut « enfant » ou « conjoint » auront moins de chance d'avoir recours aux soins qu'un assuré avec un statut « autre » ;
- Les assurés avec une ancienneté inférieure à 6 auront plus recours aux soins concernés en comparaison avec une ancienneté de 8 ans ;
- Les tranches d'âge 10 et 11 auront plus de chance de recourir aux soins que la dernière tranche. Les autres en auront moins ;
- Le niveau de garantie 5 aura moins de chance d'avoir recours aux soins ;
- Les résultats obtenus pour la tranche d'âge ne doivent pas être interprétés car aucun assuré de la dernière tranche n'a eu recours aux soins. Pour remédier à cela, il faudra préciser à l'algorithme que la référence à utiliser pour comparer les effets des différents

niveaux de cette variable correspond à un autre niveau. Dans l'exemple suivant, nous précisons que le niveau de référence est la tranche 12 :

Estimation du rapport de Effet	*Estimation du point*	*Intervalle de confiance de Wald*	
CD_TYPE_PP 2 vs 4	0.608	0.453	0.815
CD_TYPE_PP 3 vs 4	0.648	0.231	1.818
CD_TRANCHE_AGE 1 vs 12	<0.001	<0.001	>999.999
CD_TRANCHE_AGE 2 vs 12	0.005	<0.001	0.046
CD_TRANCHE_AGE 3 vs 12	0.216	0.087	0.536
CD_TRANCHE_AGE 4 vs 12	0.441	0.215	0.902
CD_TRANCHE_AGE 5 vs 12	0.381	0.184	0.790
CD_TRANCHE_AGE 6 vs 12	0.513	0.249	1.059
CD_TRANCHE_AGE 7 vs 12	0.582	0.287	1.181
CD_TRANCHE_AGE 8 vs 12	0.569	0.282	1.148
CD_TRANCHE_AGE 9 vs 12	0.757	0.383	1.496
CD_TRANCHE_AGE 10 vs 12	0.754	0.399	1.423
CD_TRANCHE_AGE 11 vs 12	0.993	0.529	1.865
CD_TRANCHE_AGE 13 vs 12	<0.001	<0.001	>999.999
NIVEAU 5 vs 6	0.788	0.664	0.936

Tableau 53 : Rapport des côtes. Soins : P. dentaires ; Segment : Haut

Tous les assurés avec des tranches d'âge inférieures à la tranche 12 auront moins de chance d'avoir recours aux soins.

L'analyse de la qualité du modèle basée sur le score des biens classés montre encore que ce modèle est moins performant que le modèle du segment « entrée de gamme » comme le l'indique les deux tableaux suivants :

Association des probabilités prédites			
Pourcentage concordant	69.5	D de Somers	0.442
Pourcentage discordant	25.3	Gamma	0.466
Pourcentage lié	5.2	Tau-a	0.093
Paires	3 111 550	c	0.721

Tableau 55 : % concordance. Soins : P. dentaires ; Segment : Haut

		Consommateurs NON Nombre	
		Nombre	*Fréquence*
Consommateurs réels	*Bien classé*		
NON	*OUI*	2 093	100
	Total	2 093	88
OUI	*Bien classé*	291	100
	NON		
	Total	291	12

Tableau 54 : Echantillon de validation : Prédits / observés. Soins : P. dentaires ; Segment : Haut

- **Equipements optiques**

Dans la même logique que pour les prothèses dentaires, nous allons présenter dans cette section les différents résultats de la modélisation des recours aux équipements optiques. Seule la sélection des variables ainsi que les scores de concordance seront présentés dans ce qui suit. L'analyse de la qualité de l'ajustement des modèles ainsi que l'étude de l'effet des variables sont disponibles en annexe 6.

Entrée de gamme

Les variables retenues par la méthode Stepwise sont les suivantes :

Récapitulatif sur la sélection Stepwise						
Etape	Saisi	Effet Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
1	CD_TRANCHE_AGE		12	1	1 864		<.0001
2	NIVEAU		1	2	275		<.0001
3	CD_SEXE		1	3	251		<.0001
4	ANNEE_ANC		8	4	123		<.0001
5	CD_ZONE		6	5	20		0.0024

Tableau 56 : Variables retenues. Soins : E. optiques ; Segment : Entrée

Cinq variables explicatives ont été retenues. Ces variables ont donc des effets statistiquement différents de zéro quant au recours aux soins de type équipements optiques. La variable type de personnes protégées a elle été exclue.

Les scores de concordances sont présentés dans les deux tableaux suivants :

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	64.8	D de Somers	0.297
Pourcentage discordant	35.1	Gamma	0.298
Pourcentage lié	0.1	Tau-a	0.076
Paires	702826842	c	0.649

Tableau 58 : % concordance. Soins : E. optiques ; Segment : Entrée

		Consommateurs prédits NON Nombre d'assurés	
		Nombre	Fréquence
Consommateurs réels	Bien classé		
NON	OUI	27 266	100
	Total	27 266	85
OUI	Bien classé		
	NON	4 899	100
	Total	4 899	15

Tableau 57 : Echantillon de validation : Prédits / observés. Soins : E. optiques ; Segment : Entrée

Le score de concordance est de 65% pour l'échantillon d'apprentissage et de 85% pour l'échantillon de validation.

Milieu de gamme

Les variables retenues par la méthode Stepwise sont les suivantes :

Etape	Effet Saisi	Effet Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
1	CD_TRANCHE_AGE		12	1	365.3151		<.0001
2	CD_SEXE		1	2	96.7506		<.0001
3	CD_TYPE_PP		2	3	10.9080		0.0043
4	ANNEE_ANC		8	4	21.4019		0.0062
5	CD_ZONE		6	5	13.7459		0.0326

Tableau 59 : Variables retenues. Soins : E. optiques ; Segment : Milieu

Seules la variable niveau de garantie n'a pas été retenue. Les variables tranche d'âge, sexe de l'assuré, type de personnes protégées, nombre d'années d'ancienneté et zone géographique ont des effets statistiquement différents de zéro sur le recours à ces soins.
Les scores de concordance sont résumés dans les deux tableaux suivants :

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	57.3	D de Somers	0.149
Pourcentage discordant	42.4	Gamma	0.150
Pourcentage lié	0.3	Tau-a	0.053
Paires	303715276	c	0.575

Tableau 61 : % de concordance. Soins : E. optiques ; Segment : Milieu

		Consommateurs prédits NON Nombre d'assurés	
		Nombre	Fréquence
Consommateurs réels	Bien classé		
NON	OUI	13 920	100
	Total	13 920	78
OUI	Bien classé		
	NON	3 950	100
	Total	3 950	22

Tableau 60 : Echantillon de validation : prédits / observés. Soins : E.optiques ; Segment : Milieu

Les scores sont de 78% pour l'échantillon de validation et de 57% pour l'échantillon d'apprentissage.

Haut de gamme

Les variables retenues par la méthode Stepwise sont les suivantes :

Récapitulatif sur la sélection Stepwise						
Etape	Effet Saisi	Effet Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
1	CD_TRANCHE_AGE		12	1	55.3366		<.0001
2	CD_ZONE		6	2	18.6204		0.0049
3	CD_SEXE		1	3	8.3362		0.0039

Tableau 62 : Variables retenues. Soins : E. optiques ; Segment : Haut

Trois variables explicatives ont été retenues pour la construction du modèle de ce périmètre. Bien que toutes les variables en question possèdent des effets statistiquement différents de zéro, l'importance de ces effets diffère d'une variable à l'autre. La tranche d'âge aura un effet plus important que n'importe quelle autre variable alors que le sexe de l'assuré aura un effet moins important que les autres.

Les deux tableaux suivants résument les scores de concordances :

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	57.2	D de Somers	0.164
Pourcentage discordant	40.8	Gamma	0.168
Pourcentage lié	2.0	Tau-a	0.060
Paires	5533590	c	0.582

Tableau 64 : % de concordance. Soins : E. optiques ; Segment : Haut

		Consommateurs prédits NON Nombre d'assurés	
		Nombre	Fréquence
Consommateurs réels	Bien classé		
NON	OUI	1 767	100
	Total	1 767	76
OUI	Bien classé		
	NON	557	100
	Total	557	24

Tableau 63 : Echantillon de validation : Prédits / observés. Soins : E. optiques ; Segment : Haut

Les scores de concordances sont 76% pour l'échantillon de validation et de 57% pour l'échantillon d'apprentissage.

Nous constatons que globalement la modélisation du recours aux équipements optiques fournit des scores de concordances plus faibles que la modélisation du recours aux prothèses dentaires. Une manière de visualiser cette différence est de comparer les courbe ROC.

- **Comparaison des performances de classement des modèles dentaires et optiques**

La courbe ROC (Receiver Operating Characteristic) représente la sensibilité en fonction de 1 – spécificité. La sensibilité est la capacité du test à bien détecter - dans notre cas - un assuré ayant eu recours aux soins, la spécificité est la capacité du test à bien détecter les assurés n'ayant pas eu recours à ces soins. L'aire sous la courbe ROC (ou Area Under the Curve, AUC) peut être interprétée comme la probabilité que, parmi deux assurés choisis au hasard, un consommateur des soins et un non-consommateur, la valeur du marqueur soit plus élevée pour le premier. Par conséquent, une AUC de 0,5 (50%) indique que le marqueur est non-informatif. Une augmentation de l'AUC indique une amélioration des capacités discriminatoires, avec un maximum de 1,0 (100%). Concrètement, nous nous retrouvons dans la situation suivante :

		Valeur réelle		***Total***
		Consommateur	*Non consommateur*	
Valeur prédite	*Consommateur*	*TP*	*FP*	*P+N*
	Non consommateur	*FN*	*TN*	*N*
Total		*P*	*N*	*P+N*

Ou P et N sont respectivement le nombre des assurés consommateurs et des assurés non-consommateurs de soins. La courbe ROC utilise les informations suivantes :

- TP (resp. TN) : le nombre de vrais positifs (resp. négatifs) qui permet d'obtenir la sensibilité avec la formule suivante :

$$Sensibilité = \frac{TP}{P} ;$$

- FP (resp. FN) : le nombre de faux positifs (resp. négatifs) qui permet d'obtenir la spécificité avec la formule suivante :

$$Spécificité = 1 - \frac{FP}{N} .$$

Nous allons présenter dans ce qui suit les courbes obtenues par type de soins et par niveau de couverture selon différents scénarii intégrant le modèle complet et le modèle retenu.

Entrée de gamme :

Les deux graphiques suivants présentent les courbes ROC pour les prothèses dentaires et les équipements optiques :

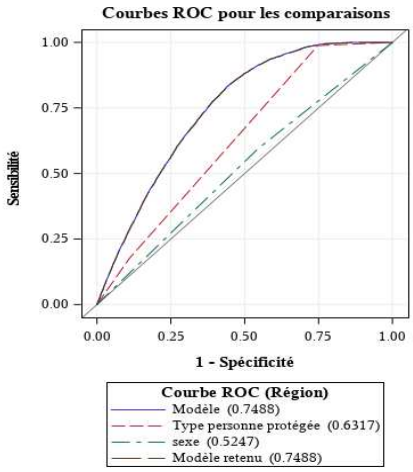


Figure 40 : Courbe ROC. Soins : Prothèses dentaires ; Segment : Entrée de gamme

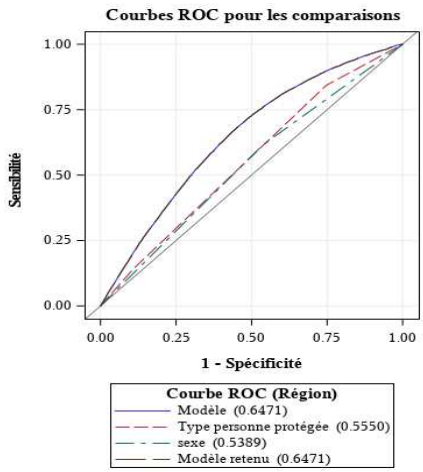


Figure 39 : Courbe ROC. Soins : Equipements optiques ; Segment : Entrée de gamme

Deux remarques s'imposent :

- D'une manière générale, l'air contenu sous la courbe est plus important pour les soins de type prothèses dentaires ce qui veut dire que le score de concordance est plus important pour ce modèle ;
- Les courbes du modèle (modèle complet) et du modèle retenu se superposent. La différence de score de concordance entre les deux modèles est négligeable

Milieu et haut de gamme :

Les courbes ROC pour ces deux segments sont les suivantes :

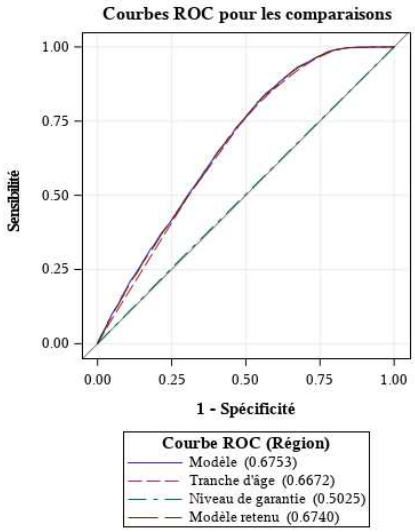


Figure 44 : Courbe ROC. Soins : Prothèses dentaires ; Segment : Milieu de gamme

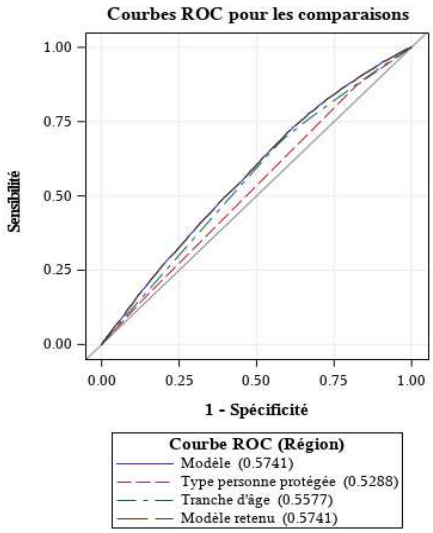


Figure 43 : Courbe ROC. Soins : Equipements optiques ; Segment : Milieu de gamme

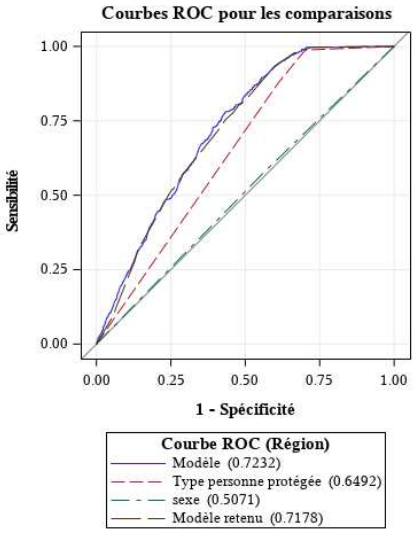


Figure 42 : Courbe ROC. Soins : Prothèses dentaires ; Segment : Haut de gamme

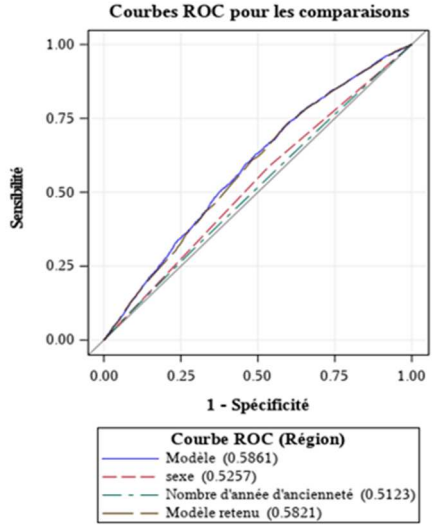


Figure 41 : Courbe ROC. Soins : Equipements optiques ; Segment : Haut de gamme

Nous constatons les mêmes effets que pour le segment « entrée de gamme ».

Une manière d'améliorer le modèle est d'intégrer des effets croisés des variables dans le modèle. Le logiciel SAS propose des commandes qui identifient la meilleure combinaison possible selon plusieurs méthodes et critères. Nous décidons de choisir comme méthode la méthode Stepwise et comme

critère de décision le critère AIC. Les deux graphiques suivants présentent les courbes ROC obtenues pour les segments « entrée de gamme » et « haut de gamme » en ce qui concerne le recours aux équipements optiques :

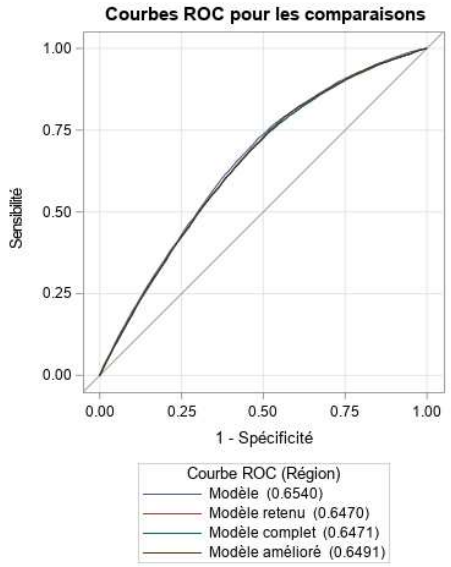


Figure 46 : Courbe ROC améliorée. Soins : Equipements optiques ; Segment : Entrée de gamme

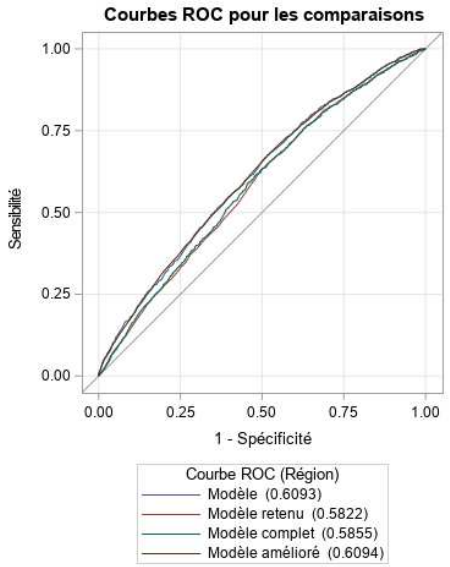


Figure 45: Courbe ROC améliorée. Soins : Equipements optiques ; Segment : Haut de gamme

Nous pouvons constater une légère amélioration des scores de concordance qui passent de 64.70% à 64.91% pour l'entrée de gamme et de 58.22% à 60.94% pour le haut de gamme. L'utilisation de ces modèles complexifie toutefois l'interprétation des résultats et des effets pour des gains qui restent assez faibles. A titre d'exemple le modèle proposé par l'algorithme pour le segment « Entrée de gamme » s'écrit sous la forme :

```
CD_SEXE*ANNEE_ANC     CD_TYPE_PP*CD_SEXE*ANNEE_ANC     CD_TRANCHE_AGE
CD_SEXE*ANNEE_ANC*CD_TRANCHE_AGE
CD_TYPE_PP*CD_SEXE*ANNEE_ANC*CD_TRANCHE_AGE   NIVEAU   CD_SEXE*NIVEAU
CD_SEXE*ANNEE_ANC*NIVEAU CD_SEXE*CD_TRANCHE_AGE*NIVEAU
```

Nous décidons donc de ne pas apporter de modifications aux précédents modèles.

3.2.3.2 Charge annuelle des prestations

La deuxième variable à modéliser pour aboutir au résultat final est la variable $X_i * 1_{T_i=1}$ qui correspond à la charge annuelle des assurés ayant eu recours aux soins définissant le périmètre de l'étude. Seront présentés dans cette section les différents résultats obtenus pour les prothèses dentaires et les équipements optiques. Les périmètres de modélisation définis pour la variable T_i seront repris pour la variable $X_i * 1_{T_i=1}$.

- **Prothèses dentaires**

Comme pour la variable recours aux soins, deux modèles ont été réalisés pour chaque niveau de gamme. Un modèle avec les données 2020 qui donne une estimation de la charge annuelle en 2020, et un modèle avec les données antérieures à 2020 qui donne une estimation de charge annuelle pour les années antérieures. Les résultats concernant le deuxième modèle sont aussi disponibles en annexe 7.

Entrée de gamme

Tout d'abord revenons sur la nécessité de modéliser uniquement la charge des sinistres des assurés ayant consommé et comparons graphiquement la qualité du modèle obtenu contre le modèle incluant avec tous les assurés :

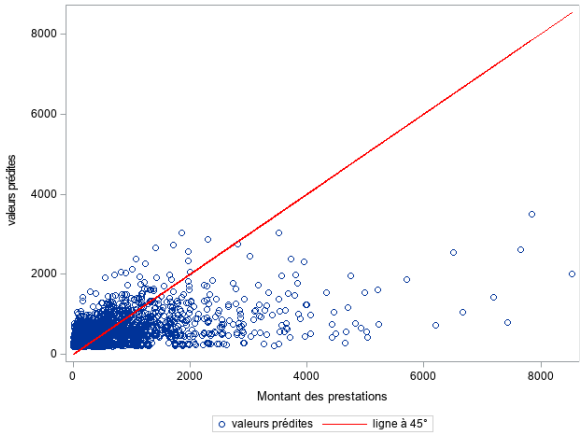


Figure 48: Valeurs prédites des assurés ayant consommés. Soins : P. dentaires

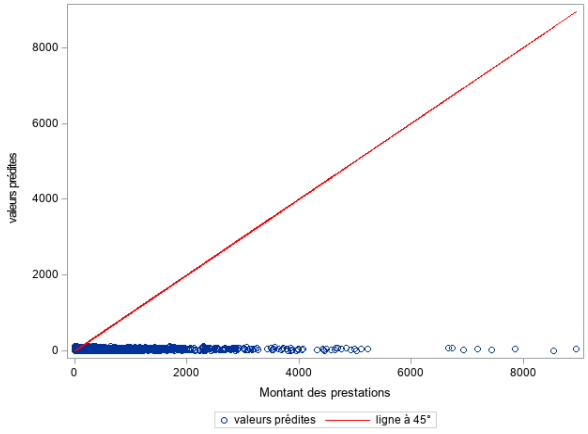


Figure 47 : Valeurs prédites tout assuré. Soins : P. dentaires

La diagonale en rouge représente le modèle parfait où la valeur prédite est égale à la valeur réelle. Nous constatons que le modèle obtenu en prenant tous assurés (à droite) a tendance à sous-estimer la charge, d'où la nécessité de réduire le périmètre de modélisation aux assurés ayant consommé uniquement. Ce comportement constaté pour les soins de type prothèses dentaires chez les assurés « entrée de gamme » est constaté sur l'ensemble des autres périmètres de modélisation.

Quant aux résultats de la sélection des variables avec la méthode Sptewise, ils sont résumés dans le tableau suivant :

Résultats estimés des paramètres					
Paramètre	DDL	Estimation	Erreur type	Khi-2	Pr > khi-2
Intercept	1	6.616504	0.122239	2929.7781	<.0001
ANNEE_ANC	1	0,087252	0.010456	69.6336	<.0001
CD_TRANCHE_AGE	1	0,031874	0.010893	8.5626	0.0034
NIVEAU	1	0.143603	0.050378	8.1254	0.0044
Dispersion	1	0.307719	0.029287	.	.
Puissance	1	2.324027	0.019326	.	.

Tableau 65 : Variables retenues. Soins : p. dentaires ; segment : Entrée

Parmi les cinq variables explicatives proposées, seules trois ont des effets statiquement différents de zéro sur le montant de la charge annuelle individuelle. Les variables en question sont : la tranche d'âge de l'assuré, le nombre d'années d'ancienneté et le niveau de garantie. A noter que la tranche d'âge de l'assuré a un pouvoir explicatif moins important que les deux autres. Enfin, le logiciel permet de calculer aussi le paramètre de la puissance de la loi de Tweedie, qui vaut dans notre cas 2,32. Ce modèle appliqué à l'échantillon de validation aboutit à une surestimation de notre variable de l'ordre de 13% comme le montre le tableau ci-dessous :

Montant des prestations réel	Montant des prestations prédit
Montant	Montant
866 259	981 406

Tableau 66 : Résultat échantillon de validation. Soins : p. dentaires ; segment : Entrée

Milieu de gamme

Les variables retenues sont :

Résultats estimés des paramètres					
Paramètre	DDL	Estimation	Erreur type	Khi-2	Pr > khi-2
Intercept	1	5.703779	0.157153	1317.2934	<.0001
CD_TYPE_PP	1	0.065371	0.022968	8.1008	0.0044
ANNEE_ANC	1	0,070075	0.007436	88.8056	<.0001
NIVEAU	1	0.230850	0.039500	34.1566	<.0001
Dispersion	1	0.174150	0.024157	.	.
Puissance	1	2.322526	0.025104	.	.

Tableau 67 : Variables retenues. Soins : p. dentaires ; segment : Milieu

L'algorithme identifie uniquement trois variables ayant des effets statistiquement différents de zéro sur la variable à expliquer. Les variables explicatives retenues sont : le type de personnes protégées, le nombre d'année d'ancienneté et le niveau de couverture. Le paramètre puissance de la loi de Tweedie vaut 2,32 proche de celui de segment « entrée de gamme ». Appliqué à l'échantillon de validation, le modèle estime la charge suivante :

Montant des prestations réel	Montant des prestations prédit
Montant	Montant
1 163 764	1 270 868

Tableau 68 : Résultat échantillon de validation. Soins : p. dentaires ; segment : Milieu

Le modèle obtenu surestime cette fois-ci la consommation de 9%

Haut de gamme

Les variables retenues sont :

Résultats estimés des paramètres					
Paramètre	DDL	Estimation	Erreur type	Khi-2	Pr > khi-2
Intercept	1	6.711305	0.497448	182.0196	<.0001
ANNEE_ANC	1	0,123423	0.017360	50.5485	<.0001
CD_TRANCHE_AGE	1	0,063157	0.020465	9.5238	0.0020
NIVEAU	1	0.157837	0.080245	3.8688	0.0492
CD_ZONE	1	0.071224	0.032650	4.7586	0.0292
Dispersion	1	0.545813	0.240299	.	.
Puissance	1	2.096652	0.074747	.	.

Tableau 69 : Variables retenues. Soins : p. dentaires ; segment : Haut

Les variables nombre d'années d'ancienneté, tranche d'âge, niveau de garantie et zone géographique sont les variables ayant des effets statiquement différents de zéro. Le paramètre puissance vaut lui 2,09 assez proche de 2, ce qui veut dire que la loi de notre variable à expliquer s'approche d'une loi Gamma. En appliquant le modèle obtenu aux données de l'échantillon de validation, nous obtenons les résultats suivants :

Montant des prestations réel	Montant des prestations prédit
Montant	Montant
211 049	252 823

Tableau 70 : Résultat échantillon de validation. Soins : p. dentaires ; segment : Haut

Le modèle surestime la variable de presque 20%. Cela peut s'expliquer par le volume de données assez faible pour cette catégorie d'assurés.

- **Equipements optiques**

Nous allons présenter dans cette section les résultats de la modélisation de la charge individuelle des sinistres pour les assurés ayant bénéficié d'un équipement optique lors de l'année 2020.

Entrée de gamme

Comme pour les prothèses dentaires, la comparaison du modèle intégrant tous les assurés contre le modèle se limitant aux assurés ayant eu recours aux soins, confirme la pertinence de notre choix. Le premier modèle a tendance à sous-estimer la charge des sinistres associée à ces soins comme le montre les deux graphiques suivants :

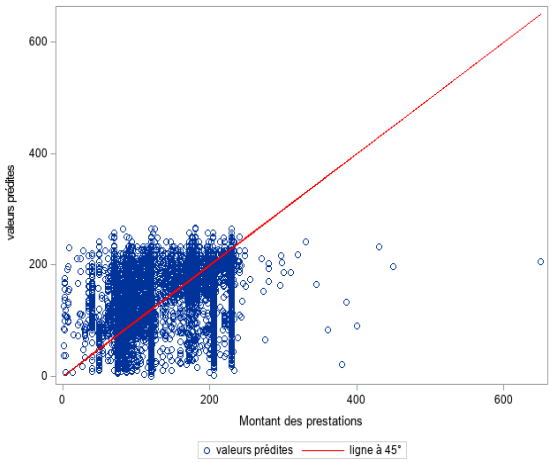


Figure 50: Valeurs prédites des assurés ayant consommés. Soins : E. optiques

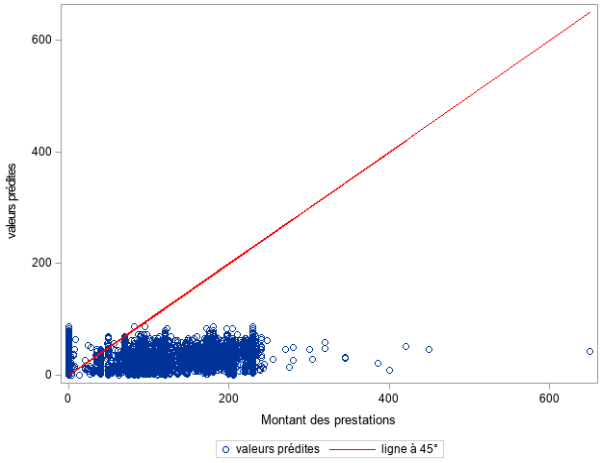


Figure 49 : Valeurs prédites tout assuré. Soins : E. optiques

Le tableau suivant récapitule les variables explicatives retenues sur ce périmètre :

Résultats estimés des paramètres					
Paramètre	*DDL*	*Estimation*	*Erreur type*	*Khi-2*	*Pr > khi-2*
Intercept	1	4.232596	0.015999	69990.5293	<.0001
ANNEE_ANC	1	0,033414	0.001800	344.7462	<.0001
CD_TRANCHE_AGE	1	0.074127	0.001149	4160.3827	<.0001
NIVEAU	1	0.205858	0.007724	710.2518	<.0001
Dispersion	1	13.228904	0.313616	.	.
Puissance	1	1.117457	0.004776	.	.

Tableau 71 : Variables retenues. Soins : é. Optiques ; segment : Entrée

Les variables sexe, type de personnes protégées et zone géographique ne possèdent pas, statistiquement parlant, de pouvoir explicatif sur la variable à expliquer. Le modèle appliqué à notre échantillon de validation donne le résultat suivant :

Montant des prestations réel	*Montant des prestations prédit*
Montant	*Montant*
748 478	747 511

Tableau 72 : Résultat échantillon de validation. Soins : é. optiques ; segment : Entrée

La charge estimée est légèrement inférieure à la charge réelle de 0,10%.

Milieu de gamme

Les variables explicatives sélectionnées par la méthode Stepwise sont les suivantes :

Résultats estimés des paramètres					
Paramètre	*DDL*	*Estimation*	*Erreur type*	*Khi-2*	*Pr > khi-2*
Intercept	1	4.430680	0.030820	20667.4514	<.0001
CD_TYPE_PP	1	0.013278	0.004351	9.3123	0.0023
ANNEE_ANC	1	0,025408	0.001574	260.5816	<.0001
CD_TRANCHE_AGE	1	0.061469	0.001139	2913.2771	<.0001
NIVEAU	1	0.194233	0.007218	724.0449	<.0001
Dispersion	1	14.653971	0.208139	.	.
Puissance	0	1.100000	0	.	.

Tableau 73 : Variables retenues. Soins : é. optiques ; segment : Milieu

Les variables : tranche d'âge, niveau de couverture, nombre d'années d'ancienneté et le type de personnes protégées ont des effets statistiquement différents de zéro sur la variable à expliquer. Le modèle produit le résultat suivant sur notre échantillon de validation :

Montant des prestations réel	*Montant des prestations prédit*
Montant	*Montant*
1 010 705	1 014 998

Tableau 74 : Résultat échantillon de validation. Soins : é. optiques ; segment : Milieu

Notre modèle surestime la charge annuelle de 0.4%

Haut de gamme

Les variables explicatives sélectionnées par la méthode Stepwise sont les suivantes :

Résultats estimés des paramètres					
Paramètre	*DDL*	*Estimation*	*Erreur type*	*Khi-2*	*Pr > khi-2*
Intercept	1	4.471007	0.124171	1296.4972	<.0001
ANNEE_ANC	1	0,039071	0.005102	58.6440	<.0001
CD_TRANCHE_AGE	1	0.070663	0.003383	436.3256	<.0001
NIVEAU	1	0.158942	0.021570	54.2992	<.0001
CD_ZONE	1	0.033814	0.008847	14.6081	0.0001
Dispersion	1	27.566835	1.062651	.	.
Puissance	0	1.100000	0	.	.

Tableau 75 : Variables retenues. Soins : é. optiques ; segment : Haut

Seule la variable sexe n'a pas été retenue. Cette variable ne participe donc pas à l'amélioration du pouvoir explicatif et ne possède pas d'effet significativement différent de zéro. La charge estimée sur l'échantillon de validation est la suivante :

Montant des prestations réel	*Montant des prestations prédit*
Montant	*Montant*
193 315	191 977

Tableau 76 : Résultat échantillon de validation. Soins : é. optiques ; segment : Haut

Il existe très peu d'écart entre la valeur réelle et la valeur prédite. Cet écart représente 0.7% de la valeur réelle.

Nous avons constaté que, d'une manière générale, les modèles construits sur le périmètre des équipements optiques fournissent une meilleure estimation de la charge annuelle. Une des explications possibles de ce constat est le volume des données de modélisation qui est plus conséquent en optique qu'en dentaire. En effet, presque 19% des assurés consomment des équipements optiques alors qu'ils sont que 7,7% à recourir aux prothèses dentaires.

3.2.3.3 Modèle complet

Maintenant que nous disposons de modèles pour les variables T_i et $X_i * 1_{T_i=1}$, nous pouvons réaliser une estimation de la charge annuelle attendue par assuré. Il suffira d'appliquer la formule d'espérance totale que nous rappelons ci-après :

$$E[X_i] = P(T_i = 1).E[X_i|T_i = 1)] .$$

En pratique :

- Les modèles associés à la variable T_i permettront de calculer une probabilité de survenance de l'évènement (recours aux soins) pour chaque assuré ;
- Les modèles associés à la variable $X_i * 1_{T_i=1}$ permettront d'estimer une charge de sinistre annuelle pour l'ensemble des assurés.

Les modèles réalisés avec les données antérieures à 2020 seront appliqués au portefeuille 2020 et les modèles ajustés avec les données 2020 seront appliqués au portefeuille 2019.

Pour prendre en compte l'effet des confinement trois scénarii seront analysés :

- Le premier correspond à une absence d'effet sur la consommation avec offset calculé à partir de l'exposition réelle ;
- Dans le deuxième scénario, nous allons considérer que le manque de consommation représente un mois de prestations. L'offset sera calculé sur la base de l'exposition réelle à laquelle nous avons retranché l'équivalent d'un mois;
- Dans le troisième scénario, nous allons considérer que le manque de consommation représente deux mois de prestations.

A noter que la plupart des analyses réalisées font état d'un effet rattrapage sur le quatrième trimestre 2020 avec la réalisation d'une partie des soins non réalisés lors du confinement.

- **Prothèses dentaires**

Nous allons présenter dans ce paragraphe les résultats obtenus concernant les soins prothèses dentaires déclinés par segment

Entrée de gamme

L'échantillon de validation de la régression logistique a été réutilisé pour mesurer l'impact de l'estimation :

Montant des prestations réel	*Montant prédit * probabilité de survenance*	*Ecart en %*
Montant	*Montant*	*pourcentage*
870 796	913 319	4,9%

Nous constatons que le modèle surestime la charge annuelle de presque 5%

Le modèle 2020 appliqué aux données du portefeuille 2019 selon les trois scénarii donne les résultats suivants :

	Montant des prestations réel 2019	*Montant prédit * probabilité de survenance*	*Montant des prestations réel 2020*
Scenario	*Montant*	*Montant*	*Montant*
1	1 006 640	2 794 606	2 767 469
2	1 006 640	2 991 032	2 767 469
3	1 006 640	3 435 857	2 767 469

Nous pouvons constater que la charge estimée est très élevée par rapport à la charge réelle. Toutefois, celle obtenue avec le scénario 1 et 2 reste dans le même ordre de grandeur que celle constatée en 2020.

Le modèle avec des données antérieures appliqué aux données du portefeuille 2020 donne, lui, le résultat suivant :

Montant des prestations réel 2019	*Montant prédit * probabilité de survenance*	*Montant des prestations réel 2020*
Montant	*Montant*	*Montant*
1 006 640	1 093 846	2 767 469

Nous remarquons que cette fois ci le modèle dont les données d'apprentissage sont constituées d'informations antérieures à 2020 prédit une charge de sinistre 2020 équivalente à celle constatée en 2019. Les estimations fournies semblent satisfaisantes et produisent des approximations correctes comparées aux données réelles.

Pour une analyse plus juste, il est préférable de s'intéresser à la charge moyenne par assuré, ce qui permet de supprimer les effets liés aux variations des effectifs. Le tableau suivant permet de comparer d'une part, les charges moyennes réelles et estimées 2020, et d'autre part les charges moyennes réelles 2020 et estimées 2019 :

	charge moyenne réelle 2019	*Charge moyenne estimée 2019*	*Charge moyenne réelle 2020*	*Réelle 2020 vs estimée 2019*	*2019 : réelle vs estimée*
Scenario	*Montant*	*Montant*	*Montant*	*Pourcentage*	*Pourcentage*
1	8,2	22,7	21,8	4%	178%
2	8,2	24,2	21,8	11%	197%
3	8,2	27,9	21,8	28%	241%

La charge moyenne estimée pour 2019 est légèrement au-dessus de la charge moyenne constatée sur 2020 pour le scénario 1. Cette même charge est deux fois plus élevée que la charge moyenne constatée en 2019.

Milieu de gamme :

Appliqué à l'échantillon de validation, le modèle surestime la charge annuelle de presque 5.6% comme le montre le tableau suivant :

Montant des prestations réel	*Montant prédit * probabilité de survenance*	*Ecart en %*
Montant	*Montant*	*pourcentage*
1 111 725	1 174 492	5,6%

Les estimations de la charge annuelle 2019 avec le modèle construit à partir des données 2020 ainsi que celle de 2020 obtenue à l'aide du modèle basé sur des données antérieures à 2020 sont résumées dans les deux tableaux suivants :

	Montant des prestations réel 2019	*Montant prédit * probabilité de survenance*	*Montant des prestations réel 2020*
Scenario	*Montant*	*Montant*	*Montant*
1	2 875 512	3 794 439	3 748 798
2	2 875 512	4 195 415	3 748 798
3	2 875 512	4 522 361	3 748 798

Montant des prestations réel 2019	*Montant prédit * probabilité de survenance*	*Montant des prestations réel 2020*
Montant	*Montant*	*Montant*
1 006 640	1 093 846	2 767 469

Les deux modèles fournissent des estimations dans les mêmes ordres de grandeurs que les réels 2019 et 2020.

Pour les mêmes raisons que pour le segment « entrée de gamme », il est nécessaire d'analyser les primes moyennes :

	charge moyenne réelle 2019	Charge moyenne estimée 2019	Charge moyenne réelle 2020	Réelle 2020 vs estimée 2019	2019 : réelle vs estimée
Scenario	Montant	Montant	Montant	Pourcentage	Pourcentage
1	119,1	157,2	156,4	0%	32%
2	119,1	173,8	156,4	11%	46%
3	119,1	187,3	156,4	20%	57%

Le scénario 1 permet d'estimer une charge moyenne 2019 équivalente à la charge constatée en 2020, mais en hausse de 32% par rapport à celle constatée en 2019. Le scénario 2, lui, estime une charge 2019 en hausse de 11% par rapport à celle de la charge réelle 2020 et de 46% par rapport à celle de 2019. Le scénario trois estime, lui, un impact beaucoup plus important car nous faisons l'hypothèse que la baisse des prestations enregistrée en 2020, du fait de la crise sanitaire, est équivalente à deux mois de prestations d'une année ordinaire.

Haut de gamme

Ci-dessous l'estimation de la charge des données de validation :

Montant des prestations réel	Montant prédit * probabilité de survenance	Ecart en %
Montant	Montant	pourcentage
211 701	227 873	7,6%

Nous pouvons constater que l'écart entre le réel et l'estimé est plus important que pour les deux précédents segments. Comme précisé précédemment, nous disposons d'un volume de données moins important ce qui peut altérer la qualité de l'ajustement du modèle et produire des estimations moins précises.

Le tableau suivant récapitule la charge estimée selon les différents scénarii :

	Montant des prestations réel 2019	Montant prédit * probabilité de survenance	Montant des prestations réel 2020
Scenario	Montant	Montant	Montant
1	749 604	761 814	713 393
2	749 604	906 230	713 393
3	749 604	943 857	713 393

Comme l'ont démontré les différents tests de comparaison, l'impact sur « le haut de gamme » de l'entrée en vigueur du 100% santé est faible, voire absent. Ceci peut expliquer l'écart très faible entre la charge estimée et la charge réelle constatée en 2019. A noter que sur ce segment les effectifs 2019 sont plus importants que les effectifs 2020, ce qui peut expliquer la baisse des prestations réglées. Le deuxième modèle (avec des données d'apprentissage antérieures à 2020), estime pour 2020 la charge suivante :

Montant des prestations réel 2019	Montant prédit * probabilité de survenance	Montant des prestations réel 2020
Montant	Montant	Montant
749 604	744 755	713 393

Elle est équivalente à celle constatée en 2019 avec une charge moyenne réelle de l'ordre de 195€ et une charge moyenne estimée de 196€ qui légèrement plus élevée (+ 0,4%).

En termes d'estimation de la charge moyenne 2019, nous obtenons les résultats suivants :

	charge moyenne réelle 2019	Charge moyenne estimée 2019	Charge moyenne réelle 2020	Réelle 2020 vs estimée 2019	2019 : réelle vs estimée
Scenario	Montant	Montant	Montant	Pourcentage	Pourcentage
1	195,3	198,5	187,8	6%	2%
2	195,3	236,1	187,8	26%	21%
3	195,3	245,9	187,8	31%	26%

Le scénario 1 estime une charge moyenne 2019 légèrement au-dessus de celle constatée. Les scénarii 2 et 3 prédisent des augmentations plus conséquentes.

- **<u>Equipements optiques</u>**

Dans ce paragraphe seront présentés les résultats obtenus concernant les soins de type équipements optiques déclinés par segment.

<u>Entrée de gamme</u>

Le modèle réalisé permet d'obtenir le résultat suivant sur notre échantillon de validation :

Montant des prestations réel	Montant prédit * probabilité de survenance	Ecart en %
Montant	Montant	pourcentage
738 653	705 650	4,5%

Nous constatons que l'écart entre la charge prédite et la charge réelle représente 4,5% de la charge réelle. Appliqué au portefeuille 2019, le modèle prédit les valeurs suivantes :

	Montant des prestations réel 2019	Montant prédit * probabilité de survenance	Montant des prestations réel 2020
Scenario	Montant	Montant	Montant
1	2 543 531	2 238 296	2 487 136
2	2 543 531	2 444 775	2 487 136
3	2 543 531	2 682 963	2 487 136

Ce qui est équivalent en charge moyenne aux variations suivantes :

	charge moyenne réelle 2019	Charge moyenne estimée 2019	Charge moyenne réelle 2020	Réelle 2020 vs estimée 2019	2019 : réelle vs estimée
Scenario	Montant	Montant	Montant	Pourcentage	Pourcentage
1	20,6	18,1	19,6	-7%	-12%
2	20,6	19,8	19,6	1%	-4%
3	20,6	21,8	19,6	11%	5%

Nous pouvons constater que, malgré l'hypothèse forte associée au scénario 3, la charge moyenne estimée dépasse à peine de 5% la charge réelle constatée sur 2019 ; le modèle construit à partir des données d'apprentissage antérieure à 2020, fournit pour la charge moyenne attendue en 2020 l'estimation suivante :

charge moyenne réelle 2019	Charge moyenne estimée 2020	Charge moyenne réelle 2020	2020 : réelle vs estimée	Réelle 2019 vs estimée 2020
Montant	Montant	Montant	Pourcentage	Pourcentage
20,6	20,5	19,6	5%	0%

Nous pouvons constater que la charge moyenne prédite pour 2020 est quasiment stable comparée à la charge réelle 2019 et plus élevée de presque 5% comparée à celle constatée en 2020. Ces résultats laissent penser que finalement le modèle détecte l'absence d'effet liée à l'évolution de la réglementation associée à un effet baissier qui peut être dû à la crise sanitaire.

Milieu de gamme

L'estimation de la charge associée à notre échantillon de validation est la suivante :

Montant des prestations réel	Montant prédit * probabilité de survenance	Ecart en %
Montant	Montant	pourcentage
1 018 472	980 179	-3,8%

La charge estimée est légèrement inférieure à la charge réelle constatée. L'écart en pourcentage par rapport à la charge réelle est de presque 4%.

Las autres quantités qui sont la charge moyenne 2019 estimée par scénario de 2019 et la charge moyenne 2020 estimée sont résumées dans les deux tableaux suivants :

	charge moyenne réelle 2019	Charge moyenne estimée 2019	Charge moyenne réelle 2020	Réelle 2020 vs estimée 2019	2019 : réelle vs estimée
Scenario	Montant	Montant	Montant	Pourcentage	Pourcentage
1	165,2	135,8	140,5	-3,3%	-17,8%
2	165,2	148,6	140,5	5,8%	-10,0%
3	165,2	164,9	140,5	17,4%	-0,1%

charge moyenne réelle 2019	Charge moyenne estimée 2020	Charge moyenne réelle 2020	2020 : réelle vs estimée	Réelle 2019 vs estimée 2020
Montant	Montant	Montant	Pourcentage	Pourcentage
165,2	158,8	140,5	13%	-4%

Encore une fois, il faut le scénario 3 pour atteindre un niveau de charge moyenne équivalent à celui constaté en 2019. L'estimation 2020 est-elle plus élevée de 17% que la charge moyenne réelle constatée. Les mêmes conclusions faites pour le segment « entrée de gamme » sont aussi valables pour ce segment.

Haut de gamme

L'estimation de la charge associée à notre échantillon de validation est la suivante :

Montant des prestations réel	Montant prédit * probabilité de survenance	Ecart en %
Montant	Montant	pourcentage
168 781	185 428	9,9%

La charge estimée est légèrement inférieure à la charge réelle constatée. L'écart en pourcentage par rapport à la charge réelle est de presque 10%.

Les autres quantités que sont la charge moyenne 2019 estimée par scénario et la charge moyenne 2020 estimée sont résumées dans les deux tableaux suivants :

	charge moyenne réelle 2019	Charge moyenne estimée 2019	Charge moyenne réelle 2020	Réelle 2020 vs estimée 2019	2019 : réelle vs estimée
Scenario	Montant	Montant	Montant	Pourcentage	Pourcentage
1	186,3	170,1	164,2	3,6%	-8,7%
2	186,3	178,2	164,2	8,5%	-4,3%
3	186,3	195,0	164,2	18,8%	4,7%

charge moyenne réelle 2019	Charge moyenne estimée 2020	Charge moyenne réelle 2020	2020 : réelle vs estimée	Réelle 2019 vs estimée 2020
Montant	Montant	Montant	Pourcentage	Pourcentage
186,3	183,8	164,2	12%	-1%

Au vu des estimations obtenues nous pouvons reprendre la même analyse que pour les deux précédents segments.

Chez AESIO mutuelle, l'équilibre technique s'apprécie au global. En effet, la tarification se fait tous postes confondus. Pour évaluer l'impact du 100% santé sur les équilibres techniques, il faudra donc estimer la sinistralité du portefeuille en question. Pour réaliser ce travail, nous proposons de retenir comme périmètre la sinistralité de l'année 2019 et de faire les hypothèses suivantes :

- La sinistralité des soins de type prothèses dentaires a évolué selon le scénario 2 ;
- La sinistralité des soins de type « équipements optiques » a évolué selon le scénario 3 ;
- La sinistralité des autres soins n'a pas connu de changement.

Le premier impact concernera les poids que représentent ces soins dans la sinistralité comme le montrent les deux graphiques suivants :

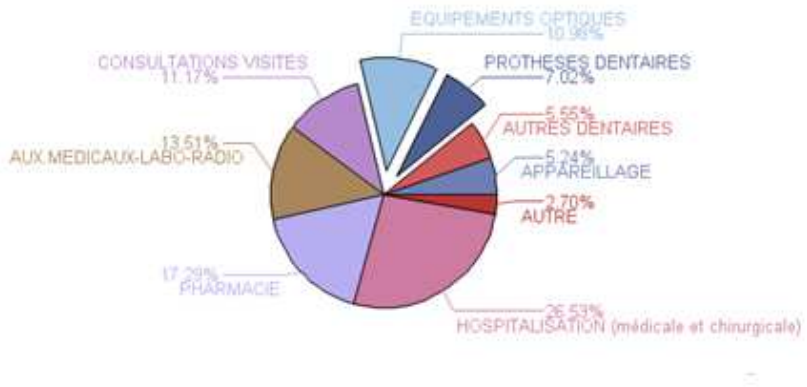


Figure 52 : Répartition de la sinistralité réelle

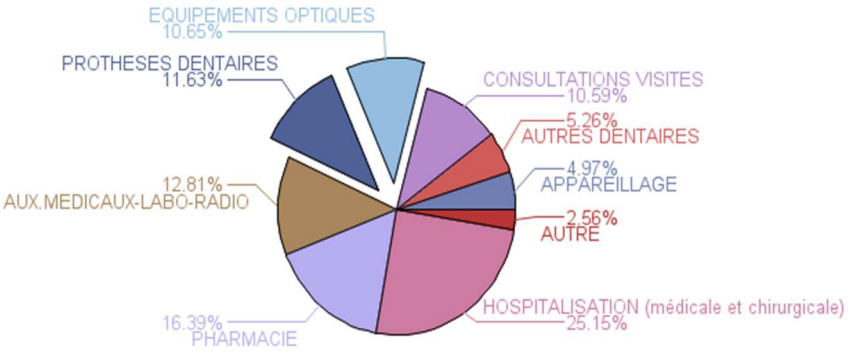


Figure 51 : Répartition de la sinistralité prédite

Nous pouvons constater que le poids des soins de type « prothèses dentaires » connait une forte augmentation et passe de 7% à 11.63%, alors que le poids des soins de type « équipements optiques » est stable. Cette stabilité est due au faible impact identifié de la mesure sur ces soins.

Le deuxième impact concerne l'évolution des montants des prestations :

	Sinistralité		
	Pédite Montant des prestations	Réelle Montant des prestations	Evolution
	Montant	Montant	Pourcentage
type de soins			
EQUIPEMENTS OPTIQUES	7 413 665,00	7 246 186,00	2,3%
PROTHESES DENTAIRES	8 092 677,00	4 631 756,00	74,7%
AUTRES SOINS	54 106 528,00	54 106 528,00	0,0%
Total	69 612 870,00	65 984 470,00	5,5%

Tableau 77: Evolution de la sinistralité

Comme précisé, l'impact concernant les « équipements optiques » est assez faible alors pour les « prothèses dentaires » l'impact est très fort, ce qui conduit à une dégradation de la sinistralité du fait de cette mesure de 5.5% selon l'estimation effectuée. Cette hausse ne pourra toutefois être associée à la réforme que si elle est accompagnée d'une baisse du reste à charge, d'où la nécessité de modéliser la variable « reste à charge » et d'analyser les résultats.

3.2.3.4 Variable reste à charge

La modélisation du reste à charge respectera la même logique que celle utilisée lors de la modélisation de la charge annuelle des prestations. Nous allons, entre autres, réutiliser la variable T_i représentant le recours aux soins, modéliser le reste à charge pour les assurés ayant eu recours aux soins puis s'appuyer sur la formule d'espérance totale pour obtenir le modèle de la variable « reste à charge ». Nous nous allons présenter dans ce paragraphe les résultats du scénario 2 par niveau de gamme et par type de soins.

Le tableau suivant contient les estimations obtenues sur le portefeuille 2019 pour les soins prothèses dentaires :

	Effectifs	Montant réel du reste à charge (B)	B/A	Montant reste à charge prédit* probabilité de survenance (C)	C/A
	Nombre	Montant	pourcentage	Montant	pourcentage
Entée de gamme	123 343	2 755 382	22	2 174 574	18
milieu de gamme	24 144	3 072 207	127	2 646 237	110
Haut de gamme	3 839	451 115	118	363 784	95

Les rapports B/A et C/A représentent respectivement le reste à charge moyen par assuré constaté sur l'année 2019 et le reste à charge moyen estimé sur la même année. Nous pouvons remarquer que ce modèle prédit une baisse du reste à charge, baisse qui est plus marquée sur le segment « entrée de gamme ».

Concernant les « équipements optiques », le modèle prédit une stagnation, voire une hausse de reste à charge comme le montre le tableau suivant :

	Montant réel des prestations (A)	Montant réel du reste à charge (B)	B/A	Montant reste à charge prédit* probabilité de survenance (B)	B/A
	Montant	Montant	pourcentage	Montant	pourcentage
Entée de gamme	126 829	2 654 648	21	2 723 125	21
milieu de gamme	23 964	2 281 208	95	2 495 868	104
Haut de gamme	3 799	234 135	62	292 770	77

Cette hausse est d'ailleurs constatée aussi sur les données réelles 2020 comme indiqué dans le tableau suivant :

Entée de gamme	milieu de gamme	Haut de gamme
23	104	68

4 Machine Learning et apprentissage supervisé

Les modèles linéaires généralisés représentent, aujourd'hui, l'outil principal utilisé dans le monde de l'actuariat pour répondre à des problématiques de tarification et de mesure de sinistralité en assurance non-vie. Toutefois, de plus en plus de publications proposent une autre approche pour résoudre ce genre de problèmes. Cette approche se base sur des méthodes de machine Learning appelées aussi apprentissage statistique (ou encore apprentissage automatique). Comme les GLM, ces méthodes permettent de s'affranchir de la plupart des hypothèses imposées par la statistique classique et qui sont dans bien souvent des cas impossibles à vérifier, mais à l'inverse des GLM, leur utilisation ne nécessite aucune hypothèse concernant la distribution des variables.

Ces méthodes sont constituées d'un ensemble d'algorithmes qui permettent de créer de manière automatique des modèles. Des modèles qui aideront à construire des classes homogènes dans la population étudiée ou à réaliser des prédictions en se basant sur des statistiques et sur des analyses prédictives. Les premiers algorithmes ont été créés à la fin des années 1950 et le plus connu d'entre eux n'est autre que le Perceptron[3]. Ces algorithmes dont les premiers remontent à la 2ème moitié du siècle précèdent vont connaitre un vrai développement et susciter de plus en plus d'intérêt dès la fin des année 90 et l'arrivée du Big DATA. Le big data a permis en effet, de disposer d'une masse de données importante et variée afin d'entrainer ces algorithmes et d'accroitre leurs capacités prédictives mais aussi de disposer de capacités de calcul plus conséquentes ce qui a permis de réduire les temps de calculs et d'améliorer l'attrait de ces méthodes.

Ces algorithmes peuvent être répartis en deux catégories :

- Apprentissage non supervisé : dans ce type de problème, le but est d'apprendre une caractéristique structurelle des variables observées X. Les algorithmes de classification (clustering) qui représentent une catégorie des algorithmes d'apprentissage, permettent de répondre à ce besoin en créant des groupes d'individus rassemblés sur la base de la proximité de leurs valeurs de X.
- Apprentissage supervisé : dans ce type de problème, on cherche à définir une règle de prédiction d'une variable à prédire Y en fonction de variables prédictives (ou explicatives) X. Pour établir cette règle, il est nécessaire de disposer d'un ensemble de données constitué des observations de X et de Y. Le choix de la règle se fera par la suite parmi une famille de règles possibles en veillant à optimiser un critère de qualité à définir.

Les méthodes qui seront développées par la suite appartiennent toutes à la catégorie des méthodes d'apprentissage supervisé. Leur utilisation se base sur la seule hypothèse suivante : les observations de la variable Y à prédire sont générées de manières indépendantes et identiques à l'aide d'un processus qui se base sur les valeurs des variables explicatives X. De telles méthodes nécessitent de préciser :

- Un ensemble $\mathcal{R}$ de règles candidates. La règle $R : \mathcal{X} \rightarrow \mathcal{Y}$ sera sélectionnée parmi cet ensemble et permettra de calculer la valeur prédite de Y : $Y_{pred} = R(X)$;
- Un critère de qualité f à optimiser.

Pour un critère de qualité f à minimiser et un ensemble de règles $\mathcal{R}$, nous pouvons définir alors :

[3] Le perceptron est un algorithme d'apprentissage supervisé de classifieurs binaires (c'est-à-dire séparant deux classes). Il a été inventé en 1957 par Frank Rosenblatt1 au laboratoire d'aéronautique de l'université Cornell. Il s'agit d'un neurone formel muni d'une règle d'apprentissage qui permet de déterminer automatiquement les poids synaptiques de manière à séparer un problème d'apprentissage supervisé. Si le problème est linéairement séparable, un théorème assure que la règle du perceptron permet de trouver une séparatrice entre les deux classes.

$$R = \underset{R \in \mathcal{R}}{\operatorname{argmin}} f(R(X), Y) .$$

Par exemple, L'erreur quadratique moyenne est souvent choisie comme critère à minimiser quel que soit le type de variable à prédire. Pour un jeu de données de taille N, elle est définie par :

$$f(Y_{pred}, Y_{obs}) = \frac{1}{N}\sum_{i=1}^{N}\left({y^{pred}}_i - {y^{obs}}_i\right)^2 \; ou \; f(Y_{pred}, Y_{obs}) = \frac{1}{N}\sum_{i=1}^{N}\left({y^{pred}}_i - \mathrm{R}({x^{obs}}_i)\right)^2 .$$

Le schéma suivant récapitule le cadre général associé à ces algorithmes :

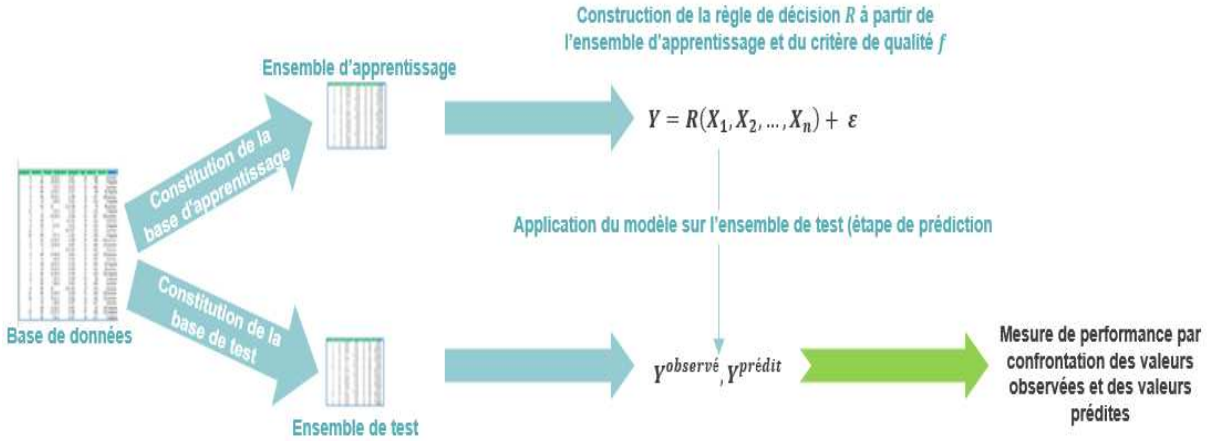


Figure 53: Schéma explicatif d'un algorithme d'apprentissage supervisé

Il est à noter que la qualité de prédiction se mesure à partir d'un ensemble appelé ensemble de test. Il est conseillé en effet, de séparer le jeu de données en deux ensembles : un ensemble d'apprentissage et un ensemble de test. L'ensemble d'apprentissage permettra de définir la meilleure règle de décision (celle qui minimise le critère de qualité), alors que l'ensemble de tests permettra de mesurer la performance du modèle. Cette séparation permet de contrer des phénomènes de sur-apprentissage qui peuvent être la conséquence d'une prédiction quasi-parfaite où les valeurs prédites sont quasiment identiques aux valeurs observées avec un critère de qualité presque nul.

Un autre concept très important dans l'exploitation de ces algorithmes est la validation croisée. Il s'agit d'une méthode composée de k itérations. L'ensemble d'apprentissage (noté D) est réparti en deux sous-ensemble A et V de la manière suivante :

$$\forall\, i \in \{1,2,\dots,k\}: A_i \cup V_i = D \; et \; A_i \cap V_i = \emptyset .$$

A chaque itération, le modèle sera entrainé sur l'ensemble A_i et l'erreur de prédiction sera calculée sur l'ensemble V_i. Cette méthode permet de calculer le risque (ou l'erreur) espéré estimé par la moyenne des erreurs de prédictions relevées à chaque itération.

Les sections qui suivent seront consacrées à la présentation de quelques-unes de ces méthodes qui seront utilisées pour prédire le recours aux soins, le recours aux paniers 100% santé et enfin la charge des sinistres. Les méthodes utilisées pour les deux premières variables seront les arbres de décision et les forêts aléatoires. La dernière variable sera modélisée à l'aide d'un arbre de régression. Le périmètre de l'étude sera réduit aux soins « prothèses dentaires ». Ce choix a été motivé par le fait que les modèles GLM ont identifié l'impact le plus conséquent sur ce périmètre.

Pour exploiter ces différents méthodes et algorithmes, le langage python sera utilisé à travers le navigateur de la distribution open source Anaconda. La principale bibliothèque qui sera utilisée est la bibliothèque Sklearn.

Plusieurs outils seront utilisés pour évaluer la pertinence des modèles obtenus et comparer les performances de ces derniers. Les différents outils utilisés sont les suivants :

- La validation croisée à K blocs (ou K-fold cross validation) : désigne une technique d'évaluation d'un algorithme de Machine Learning qui consiste à découper un jeu de données en K sous-ensemble (ou K folds). Un seul sous-ensemble sera sélectionné pour servir d'ensemble de test (validation set), les K-1 restants seront utilisés pour entrainer le modèle (training set). On répète l'opération sur toutes les combinaisons possibles. On obtient K mesure de performance dont la moyenne représente la performance de l'algorithme. Le schéma ci-dessous (source : site scikit learn) représente le déroulement des itérations (K=5) :

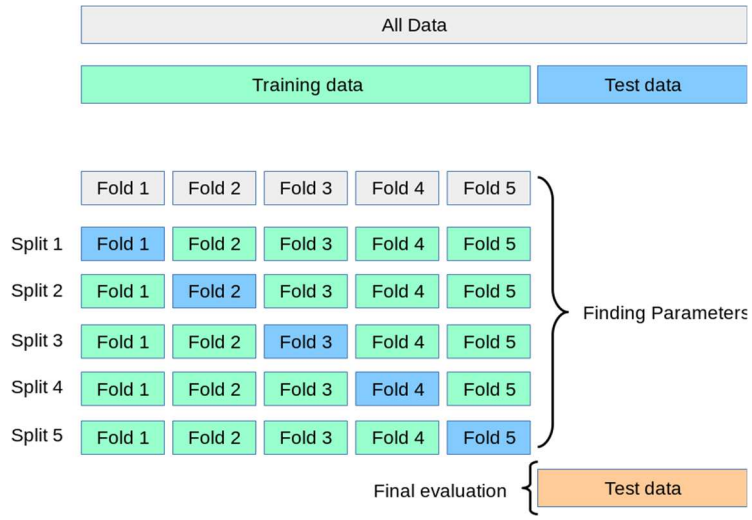


Figure 54: Schéma de validation croisée

- Matrice de confusion : Comme pour les modèles GLM, il s'agit d'une matrice qui mesure la qualité d'une classification. Chaque ligne correspond à une classe réelle et chaque colonne à une classe estimée. Pour une variable binaire, la matrice sera composée de deux lignes et deux colonnes comme le montre l'exemple ci-dessous :

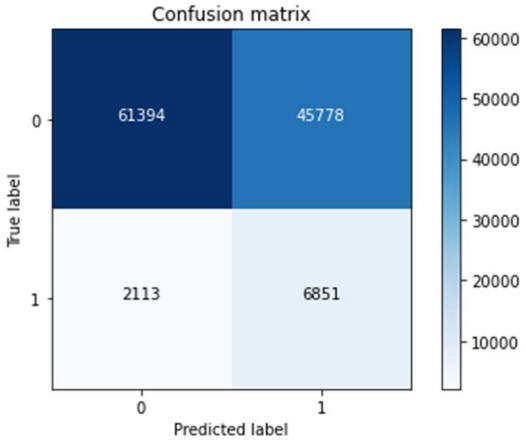


Figure 55: Exemple matrice de confusion

- Le rapport de précision qui permet de disposer de plusieurs indicateurs : le score de rappel (recall) correspond pour la classe d'étiquette « a » au nombre d'observations dans la classe estimée « a » rapporté au nombre d'observations de la classe réelle « a », le score de précision correspond pour la classe d'étiquette « a » au nombre d'observations dans la classe réelle « a » rapporté au nombre d'observations de la classe estimée « a » et enfin le F1-Score qui est une moyenne entre rappel et précision ;
- La courbe ROC précédemment définie

Remarque :

L'utilisation de ces modèles nécessite une transformation des variables catégorielles. Cette transformation consiste à créer une variable binaire par modalité et se traduit dans notre cas par la création de deux variables en remplacement du sexe, de quatre variables en remplacement du type de la personne protégée et d'une centaine de variables pour le code département. Ce travail a bien été réalisé mais ne sera pas présenté dans le cadre de ce mémoire. En effet, cette modification augmente sensiblement le nombre des variables, ce qui complexifie les modèles et leurs interprétations et rend nécessaire une étude approfondie pour aboutir à des modèles lisibles et exploitables ou des traitements automatiques à même d'exploiter les résultats.

4.1 Recours aux soins

Ce paragraphe sera consacré à la mesure de l'impact de la réforme sur le recours aux soins. Deux algorithmes seront utilisés pour prédire cette variable.

4.1.1 Arbre de décision

Un arbre de décision correspond à une méthode de classification destinée à prédire des variables binaires. Il présente l'avantage d'être relativement facile à utiliser et simple à interpréter quand il n'est pas très grand. Il se présente graphiquement sous la forme suivante :

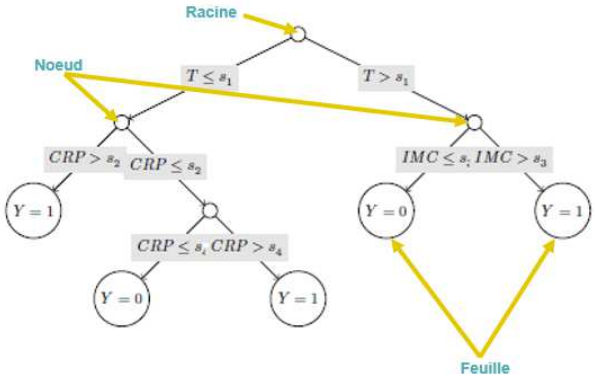


Figure 56 : Exemple arbre de décision

Une observation i parcours l'arbre de la racine jusqu'à une feuille dont l'étiquette va indiquer la décision prise. Le choix de l'embranchement se fait en fonction de la valeur de la variable explicative concerné à chaque étape. L'algorithme le plus souvent utilisé est l'algorithme CART (*Classification and Regression Tree) (Breiman et al., 1984)*.

La construction de l'arbre se base sur des indices appelés indices d'impuretés et des critères de séparations. Le principe lui est assez simple : chaque feuille est séparée en deux si elle est jugée pas assez pure, de façon à obtenir deux filles avec des indices d'impuretés optimaux. Une fois la construction de l'arbre terminée, chaque feuille portera l'étiquette de classe majoritaire qui la compose. Deux indices d'impuretés peuvent être utilisés :

- L'indice d'impureté de Gini : soit $(p_1, p_2, \dots, p_K)$ les proportions de chaque modalité de Y dans le S. Cet indice est défini alors de la manière suivante :

$$I_{Gini}(S) = \sum_{j=1}^{K} p_j(1 - p_j)\,.$$

  Cet indice sera nul si l'ensemble est pur car composé d'une seule modalité de Y dont la proportion sera égale à 1. Il sera maximal si toutes les modalités sont équi-représentées ;
- L'indice d'impureté de l'entropie qui s'exprime sous la forme suivante :

$$I_{\text{Entropie}}(S) = \sum_{j=1}^{K} p_j log(p_j)\,.$$

Le critère de séparation consiste à choisir un couple (variable, seuil de variable) qui sépare S en (S_1, S_2) de taille n_1 et n_2. L'algorithme parcours alors l'ensemble des variables possibles et l'ensemble des valeurs de ces variables afin de choisir le couple qui maximise la quantité suivante :

$$I(S) - (\frac{n_1}{n_1 + n_2} I(S_1) + \frac{n_2}{n_1 + n_2} I(S_2))\,.$$

Cette opération peut être répétée jusqu'à l'obtention d'un indice d'impureté nul sur toutes les feuilles avec le risque de se retrouver en situation de sur-apprentissage. Une façon d'éviter ce problème est

de fixer un seuil non nul pour l'indice d'impureté ou de construire l'arbre total puis supprimer les branches n'améliorant pas la quantité à maximiser.

Application à la variable recours aux soins

Comme pour les modèles GLM, les variables explicatives retenues sont : l'âge, le sexe, la zone géographique, le type de personnes protégées et l'exposition. L'âge observé à la fin de l'exercice remplacera la tranche d'âge alors que la zone géographique sera remplacée par le département. L'indice d'impureté retenu est celui de Gini. La construction de l'arbre se fera à l'aide de la fonction « DecisionTreeClassifier » présente dans la bibliothèque « sklearn ».

La performance du modèle construit sera mesurée à l'aide des outils présentés précédemment. Les résultats obtenus sont résumés ci-dessous :

- Score de la validation croisée : Decision tree cross-validation score : 85.63 ;
- Matrice de confusion :

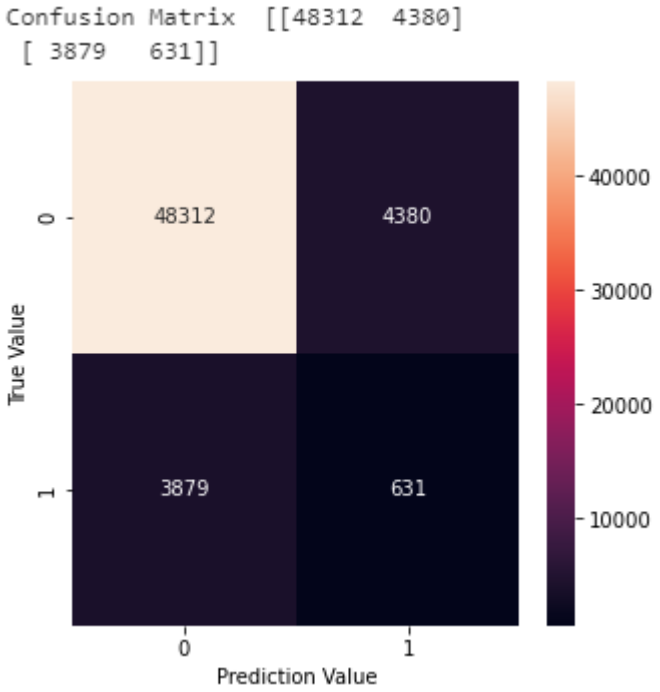


Figure 57: Matrice de confusion de l'arbre de décision recours aux soins

- Rapport de précision :

	precision	recall	f1-score	support
0	0.93	0.92	0.92	52692
1	0.13	0.14	0.13	4510
accuracy			0.86	57202
macro avg	0.53	0.53	0.53	57202
weighted avg	0.86	0.86	0.86	57202

Tableau 78 : rapport de précision de l'arbre de décision recours aux soins

Le score de la validation croisée de 86% laisse penser que le modèle obtenu est moins précis que certains modèles GLM développés dans la section précédente car plus faible. Mais l'analyse de la matrice de confusion et du rapport de précision permet de rejeter cette affirmation. En effet, contrairement au modèle GLM, l'arbre de décision obtenu permet de classer certains assurés du

groupe test dans la catégorie des consommateurs, là où les modèles GLM classaient l'ensemble des assurés dans la catégorie des non-consommateurs. Toutefois, la proportion des consommateurs classés par le modèle en tant que tels reste relativement faible (14%).

Une autre sortie permet de disposer du pouvoir discriminant des variables explicatives comme le montre le tableau suivant :

	Variable	Importance 2020
2	AGE_EX	0.59
1	CODE_DEPT	0.17
6	ANNEE_ANC	0.11
4	CD_SEXE	0.04
0	NIVEAU	0.03
3	NB_PP_MOY	0.03
5	CD_TYPE_PP	0.03

En adéquation avec la majorité des modèles GLM vus précédemment, l'âge représente la variable ayant le pouvoir discriminant le plus important. L'importance d'une variable étant sa contribution à la qualité globale du modèle.

Le modèle entrainé sur les données 2020, appliqué au portefeuille 2019, prédit une augmentation du nombre de consommateurs de 19%. L'analyse des variations par niveau de gamme montre que l'augmentation est plus marquée sur les niveaux de gamme les plus faibles, comme le décrit le tableau ci-dessous :

	1	***2***	***3***	***4***	***5***	***6***	***Total***
variation	*51%*	*22%*	*16%*	*5%*	*14%*	*3%*	*19%*

Il est aussi possible de disposer des règles de décision sous le format suivant :

```
|--- AGE_EX <= 38.79
|   |--- AGE_EX <= 24.49
|   |   |--- AGE_EX <= 20.16
|   |   |   |--- AGE_EX <= 16.93
|   |   |   |   |--- AGE_EX <= 12.33
|   |   |   |   |   |--- AGE_EX <= 9.77
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- AGE_EX >  9.77
|   |   |   |   |   |   |--- AGE_EX <= 9.78
|   |   |   |   |   |   |   |--- NIVEAU <= 3.50
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- NIVEAU >  3.50
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- AGE_EX >  9.78
|   |   |   |   |   |   |   |--- ANNEE_ANC <= 2.50
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- ANNEE_ANC >  2.50
|   |   |   |   |   |   |   |   |--- ANNEE_ANC <= 3.50
|   |   |   |   |   |   |   |   |   |--- CODE_DEPT <= 26.50
|   |   |   |   |   |   |   |   |   |   |--- CODE_DEPT <= 24.00
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- CODE_DEPT >  24.00
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 4
|   |   |   |   |   |   |   |   |   |--- CODE_DEPT >  26.50
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- ANNEE_ANC >  3.50
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |--- AGE_EX >  12.33
|   |   |   |   |   |--- AGE_EX <= 12.34
|   |   |   |   |   |   |--- ANNEE_ANC <= 4.50
```

L'arbre étant très développé, il est compliqué d'interpréter ce type de sortie.

L'arbre de décision appartient à la catégorie « apprentissage supervisé », il doit donc permettre de définir une règle de prédiction d'une variable à prédire Y en fonction de variables prédictives.

L'analyse des différents résultats a permis de constater que le modèle possède un pouvoir prédictif relativement important avec un score de 86% de « biens classés ». Cependant, ce pouvoir prédictif est très faible pour la classe minoritaire avec un score de 13% et l'interprétation des règles de prédiction est complexe du fait du nombre important des feuilles et des nœuds.

Un premier travail a été effectué pour optimiser le modèle afin d'améliorer sa qualité de prédiction mais son succès. Il est donc nécessaire de tester d'autres modèles

4.1.2 Forêt aléatoire

La forêt aléatoire ou le Random Forest est un algorithme basé sur l'agrégation d'un grand nombre de modèles. Il consiste à construire plusieurs arbres de décision (une forêt) dans le but d'améliorer l'ajustement et réduire l'erreur de prévision. Il s'appuie sur le principe de bagging tout en ajoutant de l'aléa et du hasard dans le choix des variables explicatives intervenant dans les modèles. Dans le cas d'un jeu de données composés n observations :

$$z = \{(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \dots, (\mathrm{x}_n, y_n)\}\,.$$

Avec $y_{i\,;1\leq i\leq n}$ n réalisations d'une variable Y appelée variable à expliquer supposée qualitative.

Et $\mathrm{x}_{i\,;1\leq i\leq n}$ n réalisations d'un vecteur $\mathrm{X} = (X^1, X^2, \dots, X^p)$ appelé vecteur des variables explicatives.

Si f(x) est un modèle fonction de x et $\{z_b\}_{b=1,B}$ avec B échantillons de z alors la prévision est définie par :

$$\hat{f}_B(.) = argmax_j card\{b | f_{z_b}\widehat{(.)} = j\}\,.$$

Autrement dit, le choix de la prévision correspond au vote majoritaire des différents modèles. A noter que l'estimation d'un arbre sur un échantillon se fait avec la randomisation des variables : la recherche de chaque division optimale est précédée d'un tirage aléatoire d'un sous-ensemble de m variables explicatives parmi les p disponibles.

<u>Application à la variables recours aux soins</u>

L'indice d'impureté ainsi que les variables explicatives sont identiques à ceux utilisés dans la précédente modélisation. Les résultats obtenus sont les suivants :

- Score de la validation croisée : Random Forest cross-validation score: 90.45
- Matrice de confusion :

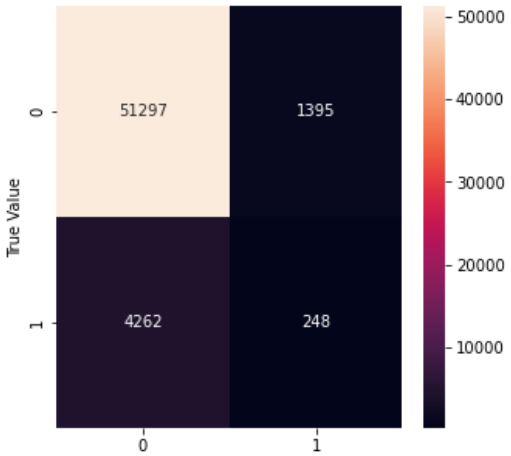


Figure 58 : Matrice de confusion forêt aléatoire recours aux soins

- Rapport de précision :

	precision	recall	f1-score	support
0	0.92	0.97	0.95	52692
1	0.15	0.05	0.08	4510
accuracy			0.90	57202
macro avg	0.54	0.51	0.51	57202
weighted avg	0.86	0.90	0.88	57202

Tableau 79 : rapport de précision forêt aléatoire recours aux soins

Dans 90% des cas, le modèle arrive à classer correctement les individus. Ce pourcentage est de 5% pour la catégorie des assurés ayant eu recours aux soins et 97% pour la deuxième catégorie.

Malgré un pouvoir prédictif plus important au global que le modèle précédent, les prédictions de la classe positive représentant les consommateurs restent peu satisfaisantes. Ce modèle est même moins performant concernant cette classe.

Les problèmes rencontrés dans la capacité des deux modèles à prédire de manière satisfaisante le recours aux soins sont dus au fait que les assurés composant cette classe sont très minoritaires dans l'échantillon. Les deux classes consommateurs/non consommateurs sont donc très déséquilibrées avec seulement 8% des assurés qui ont eu recours à des soins de type « prothèses dentaires ».

4.1.3 L'algorithme SMOT pour données déséquilibrées

Pour remédier au problème des données déséquilibrées, il est possible d'avoir recours à des méthodes permettant un rééquilibrage de ces dernières. Ces méthodes appelées « data-level solutions » font appel à deux principes :

- Le sous-échantillonnage (undersampling) : une partie des individus majoritaires est retirée, ce qui permet d'accorder plus d'importance aux individus minoritaires. Cette approche permet de diminuer la redondance des informations apportées par le grand nombre d'individus majoritaires ;
- Le sur-échantillonnage (oversampling) : cette approche a pour finalité l'augmentation du nombre d'individus minoritaires. Elle conduit à donner plus d'importance à ce groupe d'individus lors de la modélisation. Différentes solutions sont possibles, comme le "clonage" aléatoire ou le SMOTE.

Le sous-échantillonnage et le sur-échantillonnage peuvent être combinés pour corriger le déséquilibre plus efficacement.

SMOTE pour Synthetic Minority Oversampling Technique, est une méthode qui se base sur le principe suréchantillonnage des individus minoritaires. Afin d'augmenter cette classe d'individus, l'algorithme va créer de nouveaux individus synthétiques ressemblant à ceux de la classe minoritaire sans toutefois être identiques ; ce qui la distingue du simple clonage. La classe d'individus minoritaires est densifiée donc de façon plus homogène.

La création des individus synthétiques à l'aide de l'algorithme du SMOTE respecte les étapes décrites ci-dessous :

- Une première observation minoritaire (observation initiale) est sélectionnée de manière aléatoire ;
- Identification de k plus proches voisins parmi les observations minoritaires (k est un paramètre défini par l'utilisateur) ;
- Un des k plus proches voisins est sélectionné de manière aléatoire.

- Un coefficient $0 < \alpha < 1$ est généré de manière aléatoire ;
- Un individu est créé entre l'observation initiale et le voisin sélectionné. Cette nouvelle observation est positionnée à une distance α de l'observation initiale et classé dans la classe minoritaire.

Ces étapes seront répétées jusqu'à ce que l'algorithme atteigne le taux d'observations minoritaires α_{os} (os pour "oversampling") renseigné par l'utilisateur. Le nombre d'observations synthétiques générées est en effet déduit de ce taux. Le choix de paramètre k et α_{os} a un impact sur la qualité de la prédiction et le temps de calcul. Un k très grand peut par exemple conduire à créer des individus synthétiques non représentatifs des données réelles et risque de diminuer les performances du modèle.

Application à la variables recours aux soins

L'indice d'impureté ainsi que les variables explicatives sont identiques à ceux utilisés dans la précédente modélisation. Les données étant composées de données à la fois catégorielles et numériques, l'algorithme utilisé est SMOTE-NC (SMOTE-Nominal Continuous), il sera appliqué uniquement aux données d'apprentissage. Les données enrichies seront utilisées pour entrainer deux algorithmes : le premier correspond au DecisionTreeClassifier et le second est le Radom Forest.

Dans ce qui suit, seront présentés les résultats des deux algorithmes :

- Score de la validation croisée :

Decision tree	***Random Forest***
66.40	67.46

- Matrice de confusion :

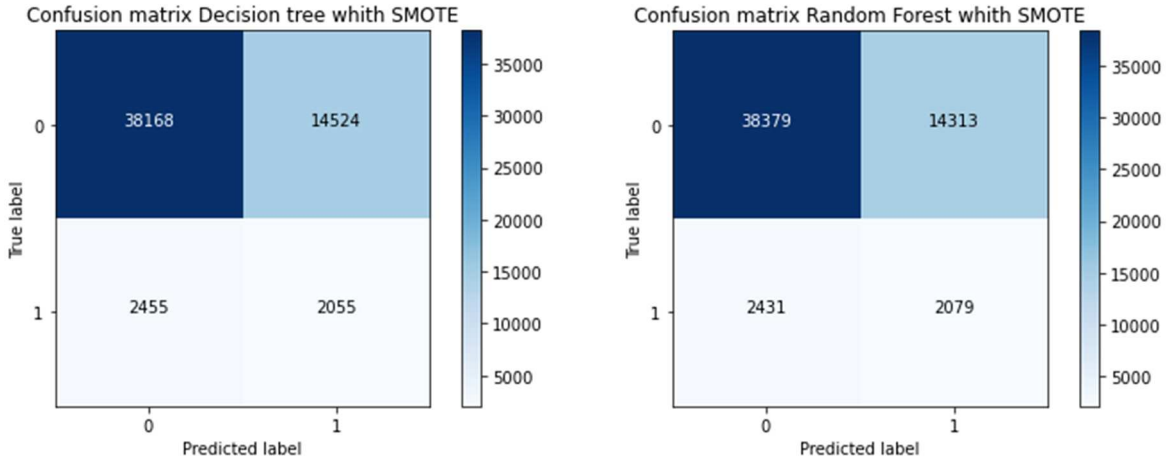


Figure 59: Matrices de confusion avec la méthode SMOTE

- Rapport de précision :

rapport de performance Decision tree whith SMOTE

	precision	recall	f1-score	support
0	0.94	0.72	0.82	52692
1	0.12	0.46	0.19	4510
accuracy			0.70	57202
macro avg	0.53	0.59	0.51	57202
weighted avg	0.88	0.70	0.77	57202

rapport de performance Random Forest whith SMOTE

	precision	recall	f1-score	support
0	0.94	0.73	0.82	52692
1	0.13	0.46	0.20	4510
accuracy			0.71	57202
macro avg	0.53	0.59	0.51	57202
weighted avg	0.88	0.71	0.77	57202

Tableau 80 :Rapports de précision avec la méthode

Malgré une dégradation du score global, une amélioration est constatée dans la prédiction de la classe « 1 » avec un score de rappel qui atteint 46% alors qu'il n'était que de 14% pour l'arbre de décision et de 5% pour la forêt aléatoire.

L'étape d'après consistera à optimiser les hyperparamètres à l'aide de l'algorithme GridSearchCV dont le principe est d'optimiser les paramètres en explorant toutes les combinaisons de valeurs possibles dans un ensemble donné. Les différents paramètres qui seront optimisés sont les suivants :

- max_depth : seuil de la profondeur maximale de l'arbre ;
- min_samples_split : nombre minimum d'échantillons requis pour diviser un nœud interne ;
- min_samples_leaf : nombre minimal d'échantillons dans un nœud feuille ;
- n_estimators : nombre d'arbres (propre au Random Forest).

Ce travail d'optimisation a permis de construire quatre modèles :

- Modèle 1 : arbre de décision obtenue grâce à l'optimisation du paramètre « max_depth ». Le score d'optimisation utilisé est la précision (accuracy) ;
- Modèle 2 : arbre de décision obtenue grâce à l'optimisation des paramètres « max_depth » et « min_samples_split ». Le score d'optimisation utilisé est le f1-score ;
- Modèle 3 : arbre de décision obtenue grâce à l'optimisation des paramètre « max_depth » et « min_samples_split ». Le score d'optimisation utilisé est le auc_roc qui correspond à l'air sous la courbe ROC ;
- Modèle 4 : Forêt aléatoire obtenue grâce à l'optimisation des paramètres « max_depth » et « min_samples_leaf » et « n_estimators ».

Les performances des quatre modèles sont résumées ci-dessous :

- Score de la validation croisée :

Modèle 1 :	39.55	***Modèle 3 :***	51.78
Modèle 2 :	48.19	***Modèle 4 :***	53.24

- Matrice de confusion :

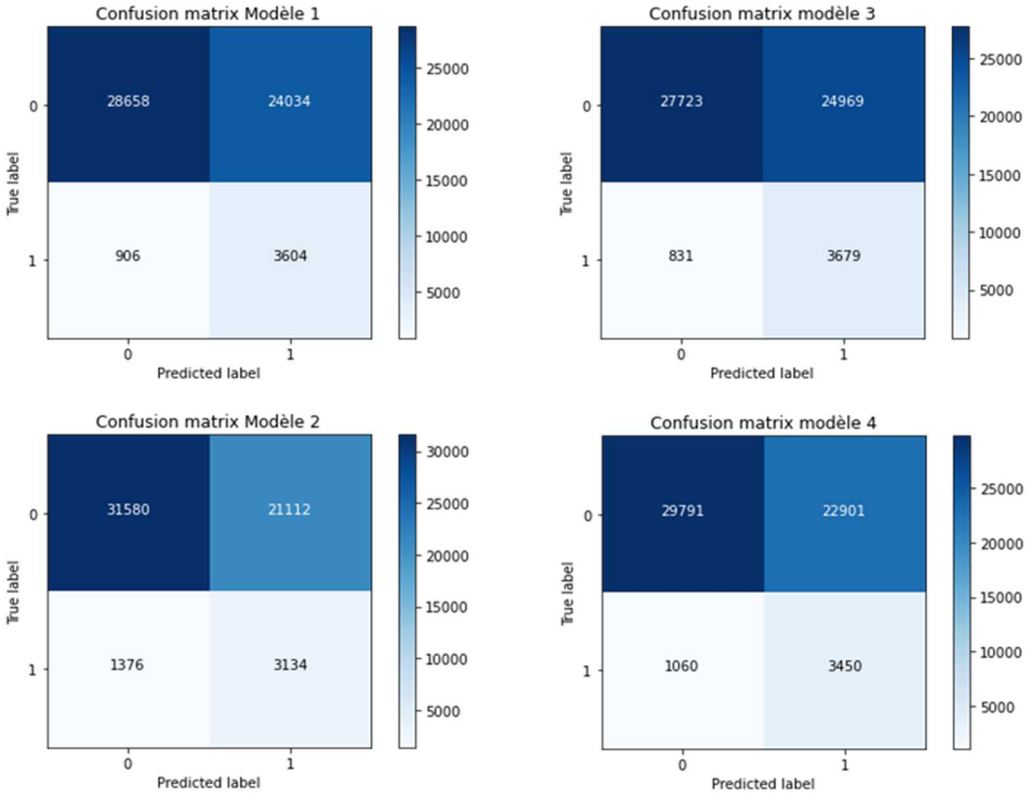


Figure 60 : Matrices de confusion des modèles 1, 2, 3 et 4

- Rapport de précision :

rapport de performance modèle 1

	precision	recall	f1-score	support
0	0.97	0.54	0.70	52692
1	0.13	0.80	0.22	4510
accuracy			0.56	57202
macro avg	0.55	0.67	0.46	57202
weighted avg	0.90	0.56	0.66	57202

rapport de performance modèle 3

	precision	recall	f1-score	support
0	0.97	0.53	0.68	52692
1	0.13	0.82	0.22	4510
accuracy			0.55	57202
macro avg	0.55	0.67	0.45	57202
weighted avg	0.90	0.55	0.65	57202

rapport de performance modèle 2

	precision	recall	f1-score	support
0	0.96	0.60	0.74	52692
1	0.13	0.69	0.22	4510
accuracy			0.61	57202
macro avg	0.54	0.65	0.48	57202
weighted avg	0.89	0.61	0.70	57202

rapport de performance modèle 4

	precision	recall	f1-score	support
0	0.95	0.67	0.79	52692
1	0.13	0.59	0.22	4510
accuracy			0.67	57202
macro avg	0.54	0.63	0.50	57202
weighted avg	0.89	0.67	0.74	57202

Tableau 81 : Rapports de précision des modèles 1, 2, 3 et 4

Le modèle 3 présente le meilleur score de rappel pour la classe 1 avec un score de 82%. La précision globale reste toutefois faible et ne dépasse pas 67%, score obtenu par le modèle 4. La comparaison des courbes ROC des diférents modèles montre aussi que le modèle 3 est le plus performant. Le graphique ci-dessous reproupe les courbes des quatre modèle ainsi que celles des modèles avec des paramètres non optimisés et celles des mdoèles construits sans la méthode SMOTE :

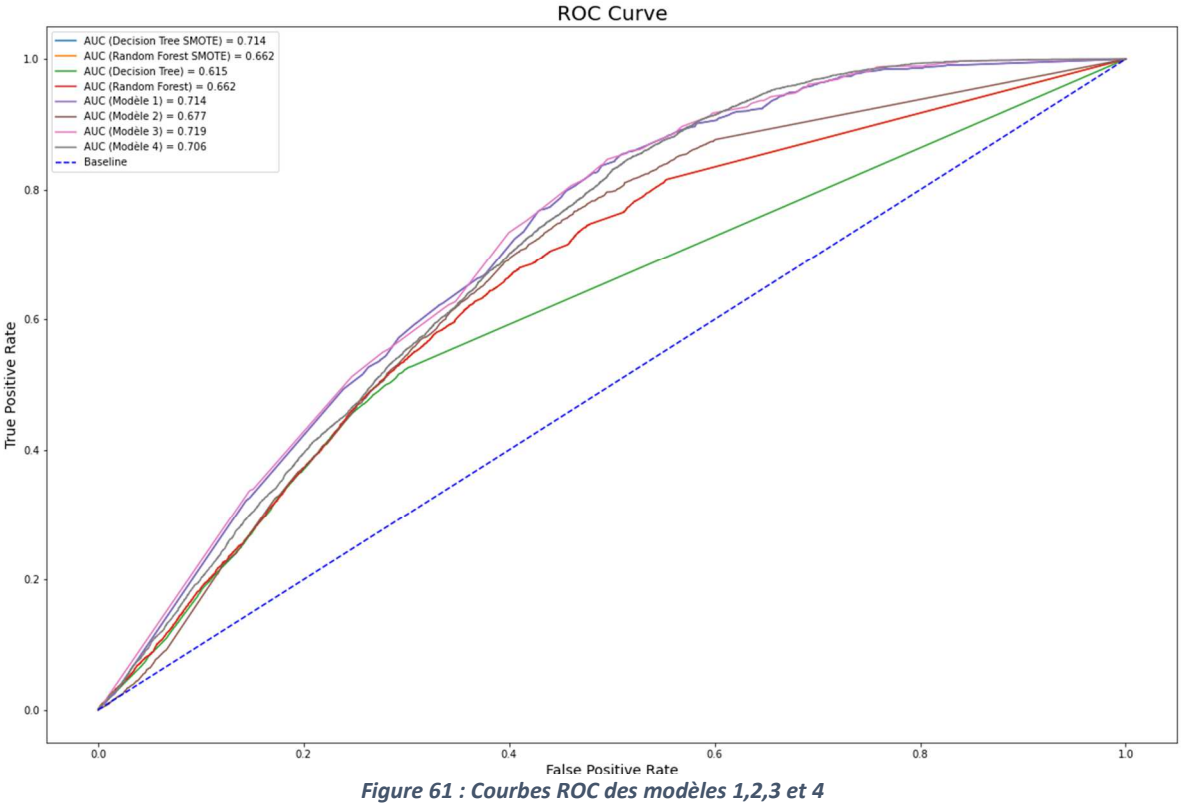


Figure 61 : Courbes ROC des modèles 1,2,3 et 4

La méthode SMOTE améliore sensiblement l'air sous la courbe pour l'algorithme « DecisionTreeClassifier ». L'optimisation des paramètres permet aussi d'obtenir une légère amélioration en fonction du choix du score d'optimisation. Concernant l'algorithme « RandomForestClassifier » seule l'optimisation des hyperparamètres permet d'améliorer l'air sous la courbe.

Les premiers modèles de machine Learning développés ont permis d'obtenir une précision globale équivalente à celle des modèles GLM avec toutefois une meilleure précision pour la classe minoritaire (classe des consommateurs). La méthode SMOTE a aidé à améliorer cette dernière, tout en dégradant la précision globale. Si la finalité est de définir des profils de consommateurs, le choix pourra se porter sur le modèle 3 par exemple, mais si la finalité est de prédire le nombre de consommateurs, l'arbre de décision sans méthode SMOTE sera un choix judicieux.

Si l'objectif est de disposer de règles permettant de prédire le comportant de l'assuré et son recours aux soins (variable à expliquer) en fonction de ces caractéristiques (variables explicatives), l'algorithme permet de disposer d'un l'arbre de décision qui se présente sous la forme:

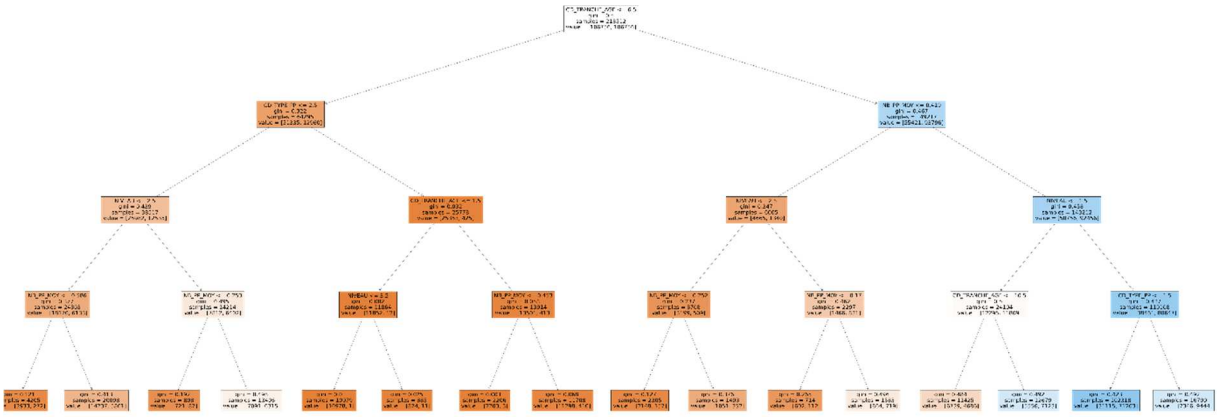


Figure 62 : Arbre de décision modèle 3

Cet arbre reste peu laisible. L'algrothime permet de disposer de ces mêmes règles sous un autre format :

Données d'entrainement 2020	***Données d'entrainement 2019***
|--- CD_TRANCHE_AGE <= 4.50 | |--- CD_TYPE_PP <= 2.50 | | |--- NIVEAU <= 2.50 | | | |--- CD_SEXE <= 1.50 | | | | |--- NB_PP_MOY <= 0.34 | | | | | |--- CODE_DEPT <= 51.50 | | | | | | |--- CODE_DEPT <= 26.00 | | | | | | | |--- weights: [205.00, 24.00] class: 0 | | | | | | |--- CODE_DEPT > 26.00 | | | | | | | |--- weights: [136.00, 0.00] class: 0 | | | | | |--- CODE_DEPT > 51.50 | | | | | | |--- CODE_DEPT <= 61.50 | | | | | | | |--- weights: [7.00, 11.00] class: 1 | | | | | | |--- CODE_DEPT > 61.50 | | | | | | | |--- weights: [146.00, 26.00] class: 0 | | | | |--- NB_PP_MOY > 0.34 | | | | | |--- NB_PP_MOY <= 1.00 | | | | | | |--- NB_PP_MOY <= 0.92 | | | | | | | |--- weights: [644.00, 441.00] class: 0 | | | | | | |--- NB_PP_MOY > 0.92 | | | | | | | |--- weights: [58.00, 181.00] class: 1 | | | | | |--- NB_PP_MOY > 1.00 | | | | | | |--- CD_TYPE_PP <= 1.50 | | | | | | | |--- weights: [2576.00, 1007.00] class: 0 | | | | | | |--- CD_TYPE_PP > 1.50 | | | | | | | |--- weights: [64.00, 106.00] class: 1 | | | |--- CD_SEXE > 1.50 | | | | |--- NB_PP_MOY <= 0.34 | | | | | |--- CODE_DEPT <= 72.50 | | | | | | |--- NIVEAU <= 1.50 | | | | | | | |--- weights: [381.00, 11.00] class: 0 | | | | | | |--- NIVEAU > 1.50 | | | | | | | |--- weights: [207.00, 0.00] class: 0 | | | | | |--- CODE_DEPT > 72.50 | | | | | | |--- NB_PP_MOY <= 0.31	|--- CD_TRANCHE_AGE <= 6.50 | |--- CD_TYPE_PP <= 2.50 | | |--- NIVEAU <= 2.50 | | | |--- NB_PP_MOY <= 0.59 | | | | |--- NB_PP_MOY <= 0.33 | | | | | |--- NB_PP_MOY <= 0.17 | | | | | | |--- weights: [1324.00, 13.00] class: 0 | | | | | |--- NB_PP_MOY > 0.17 | | | | | | |--- weights: [1236.00, 54.00] class: 0 | | | | |--- NB_PP_MOY > 0.33 | | | | | |--- NIVEAU <= 1.50 | | | | | | |--- weights: [860.00, 77.00] class: 0 | | | | | |--- NIVEAU > 1.50 | | | | | | |--- weights: [513.00, 128.00] class: 0 | | | |--- NB_PP_MOY > 0.59 | | | | |--- NIVEAU <= 1.50 | | | | | |--- CD_TRANCHE_AGE <= 4.50 | | | | | | |--- weights: [4527.00, 878.00] class: 0 | | | | | |--- CD_TRANCHE_AGE > 4.50 | | | | | | |--- weights: [3696.00, 1589.00] class: 0 | | | |--- NIVEAU > 1.50 | | | | |--- CODE_DEPT <= 34.50 | | | | | |--- weights: [2584.00, 1897.00] class: 0 | | | | |--- CODE_DEPT > 34.50 | | | | | |--- weights: [3430.00, 1497.00] class: 0 | |--- NIVEAU > 2.50 | | |--- NB_PP_MOY <= 0.25 | | | |--- CODE_DEPT <= 91.50 | | | | |--- NIVEAU <= 3.50 | | | | | |--- weights: [479.00, 31.00] class: 0 | | | | |--- NIVEAU > 3.50 | | | | | |--- weights: [236.00, 49.00] class: 0 | | | |--- CODE_DEPT > 91.50 | | | | |--- NB_PP_MOY <= 0.19 | | | | | |--- weights: [5.00, 0.00] class: 0

L'algorithme permet en effet de dectecter quelques évolutions des comportements. Par exemple, l'utilisation des données 2019 pour l'apprentissage, permet de disposer d'un modèle qui ne prédit aucun recours aux soins pour les assurés de type « assuré principal » ou « conjoint » dont l'âge est inférieur à 40 ans et dont le niveau de garantie ne dépasse pas le 2ème niveau, alors que le modèle entrainé avec des données 2020 prédit que des asurés dans cette catégorie auront recours aux soins.

4.2 Recours au panier 100% santé et montant des prestations :

Le but de ce paragraphe est d'analyser le profil des assurés ayant recours au panier 100% et de vérifier s'ils se distinguent des autres assurés ayant recours aux autres paniers.
Le périmètre de l'analyse sera restreint aux assurés ayant eu recours à des soins de type « prothèses dentaires » en 2020. Les assurés seront répartis en deux classes :

- Classe A : composée des assurés ayant consommé un panier 100% santé ;
- Classe B : composée des assurés ayant consommé un panier autre que le panier 100% santé.

Le tableau ci-dessous contient trois extraits de l'arbre de décision obtenue après optimisation des hyperparamètres :

Extrait 1	***Extrait 2***	***Extrait 3***
|--- CD_TRANCHE_AGE <= 11.50 | |--- NIVEAU <= 1.50 | | |--- CD_TRANCHE_AGE <= 8.50 | | | |--- CODE_DEPT <= 23.00 | | | | |--- CODE_DEPT <= 2.00 | | | | | |--- weights: [13.00, 32.00] class: 1 | | | | |--- CODE_DEPT > 2.00 | | | | | |--- ANNEE_ANC <= 2.50 | | | | | | |--- weights: [1.00, 24.00] class: 1 | | | | | |--- ANNEE_ANC > 2.50 | | | | | | |--- weights: [6.00, 24.00] class: 1 | | | |--- CODE_DEPT > 23.00 | | | | |--- CD_TRANCHE_AGE <= 6.50 | | | | | |--- CODE_DEPT <= 26.00 | | | | | | |--- ANNEE_ANC <= 2.50 | | | | | | | |--- weights: [23.00, 19.00] class: 0 | | | | | | |--- ANNEE_ANC > 2.50 | | | | | | | |--- weights: [14.00, 18.00] class: 1 | | | | | |--- CODE_DEPT > 26.00 | | | | | | |--- NB_PP_MOY <= 0.93 | | | | | | | |--- weights: [7.00, 24.00] class: 1 | | | | | | |--- NB_PP_MOY > 0.93 | | | | | | | |--- ANNEE_ANC <= 4.50 | | | | | | | | |--- CODE_DEPT <= 72.00 | | | | | | | | | |--- CODE_DEPT <= 42.00 | | | | | | | | | | |--- weights: [21.00, 33.00] class: 1 | | | | | | | | | |--- CODE_DEPT > 42.00 | | | | | | | | | | |--- weights: [4.00, 25.00] class: 1 | | | | | | | | |--- CODE_DEPT > 72.00 | | | | | | | | | |--- weights: [18.00, 20.00] class: 1 | | | | | | | |--- ANNEE_ANC > 4.50	|--- CD_TRANCHE_AGE <= 11.50 | |--- NIVEAU > 1.50 | | |--- CD_TRANCHE_AGE > 8.50 | | | | |--- NIVEAU > 4.50 | | | | | |--- CODE_DEPT > 67.50 | | | | | | |--- CODE_DEPT <= 74.50 | | | | | | | |--- ANNEE_ANC <= 3.50 | | | | | | | | |--- CODE_DEPT <= 69.50 | | | | | | | | | |--- ANNEE_ANC <= 1.50 | | | | | | | | | | |--- weights: [19.00, 9.00] class: 0 | | | | | | | | | |--- ANNEE_ANC > 1.50 | | | | | | | | | | |--- weights: [17.00, 8.00] class: 0 | | | | | | | | |--- CODE_DEPT > 69.50 | | | | | | | | | |--- CODE_DEPT <= 73.50 | | | | | | | | | | |--- weights: [14.00, 15.00] class: 1 | | | | | | | | | |--- CODE_DEPT > 73.50 | | | | | | | | | | |--- weights: [16.00, 10.00] class: 0 | | | | | | | |--- ANNEE_ANC > 3.50 | | | | | | | | |--- NIVEAU <= 5.50 | | | | | | | | | |--- weights: [15.00, 27.00] class: 1 | | | | | | | | |--- NIVEAU > 5.50 | | | | | | | | | |--- weights: [18.00, 15.00] class: 0 | | | | | |--- CODE_DEPT > 74.50 | | | | | | |--- weights: [30.00, 15.00] class: 0	|--- CD_TRANCHE_AGE > 11.50 | |--- CODE_DEPT <= 65.50 | | |--- CD_TRANCHE_AGE <= 12.50 | | | |--- CODE_DEPT <= 2.00 | | | | |--- ANNEE_ANC <= 2.50 | | | | | |--- weights: [8.00, 22.00] class: 1 | | | | |--- ANNEE_ANC > 2.50 | | | | | |--- weights: [10.00, 41.00] class: 1 | | | |--- CODE_DEPT > 2.00 | | | | |--- CODE_DEPT <= 33.50 | | | | | |--- ANNEE_ANC <= 3.50 | | | | | | |--- CODE_DEPT <= 8.50 | | | | | | | |--- weights: [5.00, 29.00] class: 1 | | | | | | |--- CODE_DEPT > 8.50 | | | | | | | |--- ANNEE_ANC <= 1.50 | | | | | | | | |--- weights: [2.00, 52.00] class: 1 | | | | | | | |--- ANNEE_ANC > 1.50 | | | | | | | | |--- weights: [1.00, 46.00] class: 1 | | | | | |--- ANNEE_ANC > 3.50 | | | | | | |--- NIVEAU <= 2.50 | | | | | | | |--- ANNEE_ANC <= 4.50 | | | | | | | | |--- weights: [3.00, 29.00] class: 1 | | | | | | | |--- ANNEE_ANC > 4.50 | | | | | | | | |--- ANNEE_ANC <= 5.50 | | | | | | | | | |--- weights: [15.00, 36.00] class: 1 | | | | | | | | |--- ANNEE_ANC > 5.50 | | | | | | | | | |--- weights: [5.00, 32.00] class: 1 | | | | | |--- NIVEAU > 2.50 | | | | | | |--- CODE_DEPT <= 22.00 | | | | | | | |--- weights: [4.00, 23.00] class: 1 | | | | | | |--- CODE_DEPT > 22.00 | | | | | | | |--- weights: [1.00, 31.00] class: 1 | | | | |--- CODE_DEPT > 33.50 | | | | | |--- ANNEE_ANC <= 2.50 | | | | | |--- CODE_DEPT <= 38.50

Les règles de prédiction pouvant être déduites de ces extraits sont les suivants :

- Extrait 1 : les assurés équipés du premier niveau de couverture et dont l'âge est inférieur à 75 ans auront recours aux paniers 100% santé, exception faite des assurés dont le code département est inférieur à 26 et avec une ancienneté de moins de deux ans ;
- Extrait 2 : les assurés ayant recours aux autres paniers sont des assurés avec un niveau de garantie supérieur au quatrième niveau dont l'âge dépasse 50 ans et dont le code département de leurs résidences est inférieur à 69 ;
- Extrait 3 : les assurés âgés de plus de 75 ans auront recours au panier 100% santé quelques soient leurs autres caractéristiques.

La modélisation du montant des prestations permet aussi d'identifier quelques évolutions. Le tableau suivant contient les règles de décision issues de deux modèles : un premier modèle obtenu en utilisant comme données d'apprentissage le portefeuille 2020 et le second obtenu à l'aide des données 2019 :

Données d'entrainement 2020	***Données d'entrainement 2019***
|--- AGE_EX <= 40.01 | |--- AGE_EX <= 25.78 | | |--- AGE_EX <= 20.16 | | | |--- AGE_EX <= 16.93 | | | | |--- value: [0.14] | | | |--- AGE_EX > 16.93 | | | | |--- value: [2.60] | | |--- AGE_EX > 20.16 | | | |--- CODE_DEPT <= 92.50 | | | | |--- value: [9.52] | | | |--- CODE_DEPT > 92.50 | | | | |--- value: [105.78] | |--- AGE_EX > 25.78 | | |--- NIVEAU <= 3.50 | | | |--- NIVEAU <= 2.50 | | | | |--- value: [13.53] | | | |--- NIVEAU > 2.50 | | | | |--- value: [33.47] | | |--- NIVEAU > 3.50 | | | |--- NIVEAU <= 5.50 | | | | |--- value: [59.49] | | | |--- NIVEAU > 5.50 | | | | |--- value: [108.25] |--- AGE_EX > 40.01 | |--- NIVEAU <= 2.50 | | |--- NIVEAU <= 1.50 | | | |--- NB_PP_MOY <= 0.62 | | | | |--- value: [11.68] | | | |--- NB_PP_MOY > 0.62 | | | | |--- value: [35.01] | | |--- NIVEAU > 1.50 | | | |--- NB_PP_MOY <= 0.84 | | | | |--- value: [17.94] | | | |--- NB_PP_MOY > 0.84 | | | | |--- value: [59.57] | |--- NIVEAU > 2.50 | | |--- NIVEAU <= 3.50 | | | |--- NB_PP_MOY <= 0.90 | | | | |--- value: [39.63] | | | |--- NB_PP_MOY > 0.90 | | | | |--- value: [82.99] | | |--- NIVEAU > 3.50 | | | |--- ANNEE_ANC <= 2.50 | | | | |--- value: [142.31] | | | |--- ANNEE_ANC > 2.50 | | | | |--- value: [100.45]	|--- NIVEAU <= 2.50 | |--- NIVEAU <= 1.50 | | |--- AGE_EX <= 48.94 | | | |--- AGE_EX <= 30.85 | | | | |--- value: [0.20] | | | |--- AGE_EX > 30.85 | | | | |--- value: [1.36] | | |--- AGE_EX > 48.94 | | | |--- NB_PP_MOY <= 0.58 | | | | |--- value: [0.84] | | | |--- NB_PP_MOY > 0.58 | | | | |--- value: [4.08] | |--- NIVEAU > 1.50 | | |--- AGE_EX <= 41.44 | | | |--- AGE_EX <= 28.79 | | | | |--- value: [1.84] | | | |--- AGE_EX > 28.79 | | | | |--- value: [10.36] | | |--- AGE_EX > 41.44 | | | |--- NB_PP_MOY <= 0.87 | | | | |--- value: [13.76] | | | |--- NB_PP_MOY > 0.87 | | | | |--- value: [32.26] |--- NIVEAU > 2.50 | |--- AGE_EX <= 40.95 | | |--- AGE_EX <= 28.29 | | | |--- CD_TYPE_PP <= 1.50 | | | | |--- value: [19.88] | | | |--- CD_TYPE_PP > 1.50 | | | | |--- value: [1.61] | | |--- AGE_EX > 28.29 | | | |--- NIVEAU <= 5.50 | | | | |--- value: [33.20] | | | |--- NIVEAU > 5.50 | | | | |--- value: [116.38] | |--- AGE_EX > 40.95 | | |--- NIVEAU <= 3.50 | | | |--- NB_PP_MOY <= 0.50 | | | | |--- value: [26.40] | | | |--- NB_PP_MOY > 0.50 | | | | |--- value: [58.38] | | |--- NIVEAU > 3.50 | | | |--- NIVEAU <= 4.50 | | | | |--- value: [95.42] | | | |--- NIVEAU > 4.50 | | | | |--- value: [138.21]

Plusieurs différences peuvent être constatées entre le modèle 2020 et le modèle 2019 :

- Le classement des variables selon leurs importantes a évolué : en 2020 la variable la plus discriminante est l'âge alors qu'en 2019 c'est le niveau de garantie. Le tableau ci-dessous permet de disposer de l'importante des variables par année :

Données d'entrainement 2020	***Données d'entrainement 2019***

Données d'entrainement 2020

	Variable	Importance 2020
2	AGE_EX	0.56
0	NIVEAU	0.36
3	NB_PP_MOY	0.05
6	ANNEE_ANC	0.02
1	CODE_DEPT	0.00
4	CD_SEXE	0.00
5	CD_TYPE_PP	0.00

Données d'entrainement 2019

	Variable	Importance 2019
0	NIVEAU	0.60
2	AGE_EX	0.36
3	NB_PP_MOY	0.01
6	ANNEE_ANC	0.01
5	CD_TYPE_PP	0.00
1	CODE_DEPT	0.00
4	CD_SEXE	0.00

- Une augmentation du montant des prestations par classe d'âge et par niveau de garantie. Par exemple : en 2019, le montant des prestations attendu pour un assuré avec le premier niveau de garantie (en jaune) ne dépasse pas 4.08 quelles que soient les autres caractéristiques. Pour le même assuré le montant 2020 peut atteindre dans certains cas 35. Un autre exemple concerne les assurés avec des garanties d'un niveau égal à 3 et dont l'âge dépasse 40 ans (en vert), ce montant ne dépasse pas 58.38 en 2019 alors qu'il peut atteindre 82.99 en 2020.

Enfin, la modélisation du montant de prestations à l'aide des données des assurés ayant eu recours aux soins permet aussi de mettre en évidence l'évolution de la différence du coût selon le niveau de garantie :

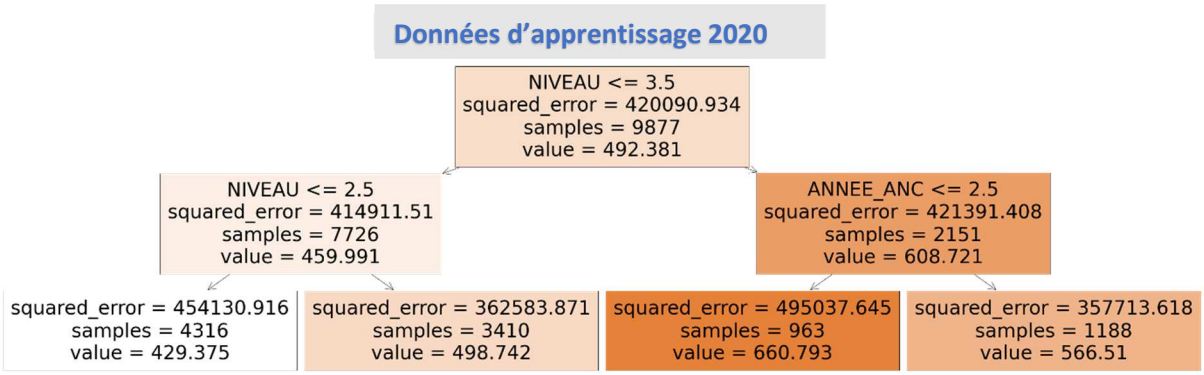


Figure 63 : Arbre de décision avec données d'apprentissage 2020

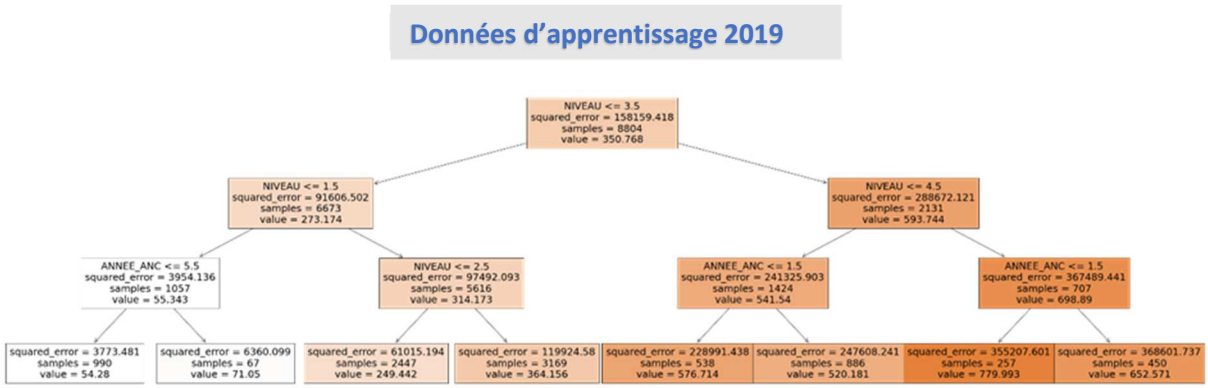


Figure 64 : Arbre de décision avec données d'apprentissage 2019

Le modèle entrainé avec les données 2019 distingue bien la charge par niveau de garantie. La charge prédite associée au niveau 1 est différente de celle du niveau 2, qui diffère de celle du niveau 3 et ainsi de suite. Seuls les niveaux 5 et 6 auront des charges identiques. Cette différence est moins présente dans les règles de décision obtenues à partir du modèle entrainé avec les données 2020. La charge prédite associée au niveau 1 est identique à celle du niveau 2. Celles prédites pour les niveaux 4 ,5 et 6 sont aussi identiques. Ces constats s'ils se généralisent à d'autres postes de soins peuvent conduire à une simplification des grilles tarifaires voir à une refonte des gammes.

Conclusion

La maîtrise de risque couvert nécessite un suivi régulier de la sinistralité et des mesures d'impacts de toute évolution réglementaire pouvant occasionner une dégradation ou au contraire une amélioration de cette dernière. Un tel suivi a pour objectif d'anticiper ou détecter toute évolution des engagements de l'assureur ainsi que de maintenir des grilles tarifaires en adéquation avec le risque assuré. En effet, une dégradation de la sinistralité peut avoir plusieurs impacts :

- Des impacts sur le résultat avec une baisse de la marge technique accompagnée d'une hausse des frais de gestion (dans le cas ou par exemple le nombre d'actes augmente sensiblement) ;
- Des impacts sur les indicateurs de solvabilité avec SCR en augmentation sous l'effet de l'augmentation du SCR de souscription accompagnée d'une diminution des fonds propres sous l'effet de l'augmentation du BE. Ces facteurs combinés conduiront in fine à une dégradation du ratio de couverture.

Le travail effectué s'inscrit dans ce cadre afin de disposer des éléments nécessaires à une juste évaluation de la situation. Le périmètre de l'étude présente néanmoins deux inconvénients :

- Une étude menée sur la consommation de l'année 2020 fortement impactée par la crise sanitaire qui a eu pour conséquence un recours aux soins moins important du fait du confinement et des craintes liées à la transmission du virus ;
- Une réforme pas totalement déployée à la date de la réalisation de l'étude. En effet, la prise en charge totale des aides auditives n'a abouti qu'en 2021 alors que les données exploitées concernent des soins antérieurs à cette année.

Malgré ces deux inconvénients, les analyses ont permis tout de même d'identifier une évolution dans des indicateurs de sinistralité associés à certains soins et sur certains segments du portefeuille :

- Les différents tests statistiques ont permis de conclure à une augmentation de la fréquence des recours aux soins, du nombre d'actes et des montants des prestations. Ces augmentations s'accompagnent d'une baisse du reste à charge. Ces différentes variations concernent principalement les soins de type « prothèses dentaires », les assurés équipés de garanties « entrée de gamme » et, dans une moindre mesure, les assurés avec des garanties « milieu de gamme ». Les conséquences de telles évolutions sont diverses ; une augmentation conséquente du nombre d'actes peut par exemple aboutir à une augmentation des dossiers à traiter par les équipes de gestion, ce qui peut modifier les délais de traitement et impacter les cadences des paiements utilisés dans le calcul des provisions et l'estimation de la charge ultime ;
- Les modèles GLM utilisés dans le calcul de la charge totale sous l'hypothèse du scénario 2 (la baisse des recours aux soins représente un mois de consommation), ont conduit à une estimation d'une charge en évolution de 74% pour les soins de type « prothèses dentaires » et de 2.3% pour les soins de type « équipements optiques ». Ces deux variations combinées représentent une augmentation de 5.5% de la charge globale tous soins confondus. Ces éléments s'accompagnent aussi par une estimation d'un reste à charge en baisse. Une charge réelle correspondant à ses estimations, dégradera les équilibres techniques et interrogera sur la suffisance des tarifs pratiqués. Elle pourra aussi conduire à une augmentation de l'exigence du capital pour le risque de primes et de réserve qui est un sous module du risque de souscription :
    - Une augmentation des prestations payées nécessitera la constitution d'un montant de provisions plus important ce qui conduira à une augmentation du volume des réserves ;

- Le maintien de l'équilibre technique par augmentation tarifaire influera lui sur le volume des primes qui augmentera aussi.

Les estimations obtenues devront toutefois être pris avec précaution. Des variations plus aux moins importantes peuvent être constatées selon le scénario retenu concernant l'impact de la crise sanitaire sur la consommation. Un impact qui reste difficile à chiffrer. Les données réelles 2021 de la sinistralité du portefeuille vues à fin octobre 2021 viennent confirmer cet aspect comme le montre le tableau suivant qui porte sur les prestations liées à des soins de type « prothèses dentaires » :

	Montant des prestations réel 2021 (vu au 31/10/2021)	*Montant prédit * probabilité de survenance*		
		Scenario 1	*Scenario 2*	*Scenario 3*
Niveau de gamme	*Montant*	*Montant*	*Montant*	*Montant*
Entrée de gamme	4 255 776	2 894 667	3 322 247	3 575 259
Milieu de gamme	4 465 215	3 743 456	3 979 229	4 452 583
Haut de gamme	773 181	728 100	860 228	885 890
Total	***9 494 172***	***7 366 223***	***8 161 704***	***8 913 732***

Tableau 82 : estimation du montant des prestations de l'année 2021 des prothèses dentaires

L'écart constaté entre les estimations et le réel constaté peut avoir plusieurs causes :

- Les différents scénarii sous-estiment l'impact du COVID et, dans ce cas, il aurait fallu retenir dans la construction du modèle, une exposition beaucoup plus faible pour obtenir des résultats plus satisfaisants ;
- Un effet rattrapage qui a pour conséquence la réalisation en 2021 des soins non effectués en 2020.

L'analyse des prestations réglées en 2021 jusqu'au 31/10/2021 qui sont liées à des soins de type « équipements optiques » montre que le scénario 3 retenu précédemment conduit à une surestimation de ce montant. Cette analyse décrit une absence d'effet de la réforme 100% santé comme l'indique le tableau suivant :

	Montant des prestations réel 2019	*Montant des prestations réel 2021 (vu au 31/10/2021)*	*Montant prédit * probabilité de survenance*		
			Scenario 1	*Scenario 2*	*Scenario 3*
Niveau de gamme	*Montant*	*Montant*	*Montant*	*Montant*	*Montant*
Entrée de gamme	*2 543 531*	2 429 125	2 367 471	2 594 515	2 846 099
Milieu de gamme	*3 987 499*	3 215 238	3 154 897	3 578 194	3 996 678
Haut de gamme	*715 157*	580 710	689 393	842 379	829 451
Total	***7 246 186***	***6 225 073***	***6 211 761***	***7 015 088***	***7 672 228***

Tableau 83 : estimation du montant des prestations de l'année 2021 des équipements optiques

Pour rappel :

- Le premier scénario représente une absence d'effet de la crise sanitaire sur la consommation avec un offset calculé à partir de l'exposition réelle ;
- Dans le deuxième scénario, nous allons considérer que la baisse du recours aux soins du fait de la pandémie représente un mois de prestations ;
- Dans le troisième scénario, nous allons considérer que cette baisse représente deux mois de prestations.

Une des explications possibles des différences constatées entres les soins « prothèses dentaires » et les « soins équipements optiques » est la mise en avant des paniers 100% santé qui diffère d'une profession à l'autre. En effet, ces paniers sont souvent proposés par les chirurgiens-dentistes car ils s'y sont engagés à travers la convention nationale 2018-2023 dont l'un des axes principaux est le renforcement de la qualité des soins dentaires en rendant le tarif accessible à tous, alors que ces paniers restent peu proposés par les opticiens. Un changement dans la commercialisation de cette offre par cette profession peut conduire à des évolutions significatives de la sinistralité liée à ses soins.

Enfin, l'exploration des méthodes de type machine learning a permis de disposer d'un outil supplémentaire dans l'analyse des comportements des assurés. En effet, grâce à ces dernières, il a été possible d'établir des règles de décisions quant aux recours aux soins et aux paniers 100% santé. Ces règles peuvent, par exemple, être exploitées dans la segmentation des gammes. A titre d'exemple, ces méthodes ont montré que la majorité des assurés équipés de garanties dont le niveau est inférieur à 4 ont recours à des paniers 100% santé.

L'étude réalisée a permis aussi de mettre en évidence l'impact de la crise sanitaire sur le recours aux soins. D'autres conséquences pourront dans le futur, être associées à cette crise. Par exemple, quel impact aura la généralisation du télétravail (imposée par la crise) sur la santé des français ? Est-ce que cela réduirait la transmission de certains virus comme la grippe ou la gastro-entérite et conduirait à une baisse des piques de consultations constatés durant les phases épidémiques ? Est-ce que cela aggraverait la sédentarité de certaines populations avec une dégradation de leurs états de santé ?

Bibliographie

- Fatemeh ABDOLLAHI (2017) : Tarification d'une complémentaire santé à destination des séniors, modulaire par poste de garanties et l'impact sur la solvabilité ;

- Edouard PLUNET (2021) : Modification d'un zonier et conséquences opérationnelles – Application sur une gamme de produits individuels en complémentaire santé ;

- Thomas SENNE (2015) : Apport de l'analyse prospective au pilotage d'un portefeuille santé individuel fermé à la souscription au sein d'une mutuelle ;

- Antoine PAGLIA, Martial V. PHELIPPE-GUINVARC'H : Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique ; publié au BULLETIN FRANÇAIS D'ACTUARIAT, Vol. 11, n°22, juillet - décembre 2011, pp. 49 – 81 ;

- Rapport 2020 sur la solvabilité et la situation financière du groupe AESIO ;

- Site internet : https://www.ameli.fr/chirurgien-dentiste/textes-reference/convention/convention-nationale-2018-2023 ;

- Rapport 2020 : La situation financière des organismes complémentaires santé, DREES ;

- Rapport 2019 : La situation financière des organismes complémentaires santé, DREES ;

- Edition 2019 : La complémentaire santé : Acteurs, bénéficiaires, garanties, DREES ;

- Edition 2020 : Les chiffres clés de la Sécurité sociale 2019, Direction de la Sécurité ;sociale
- Actualité et dossier en santé publique (n° 102 mars 2018) : Reste à charge et santé ;

- Arthur CHARPENTIER, Christophe DUTANG (décembre 2012) : l'actuariat avec R ;

- Arthur CHARPENTIER : Statistique de l'assurance, STT 6705V, Statistique de l'assurance II : partie 1 - assurance non-vie d tarification & provisionnement ;

- Olivier SAUTORY : La statistique descriptive avec le système SA , INSEE GUIDES N°1-2 ;

- http://cedric.cnam.fr/vertigo/Cours/ml2/;

- SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357 Submitted 09/01; published 06/02.

5 Annexe

5.1 Annexe 1 : dictionnaire des données

Les deux tableaux qui suivent contiennent respectivement les dictionnaires de données pour la table effectifs et la table prestation :

Nom de la variable	Description
NUM_PP	Numéro de personne (identifiant de la personne protégée)
TYPE_PP	Type de la personne protégée : assuré principal, conjoint, enfant…
CODE_GARANTIE	Code garantie
PRODUIT	Code produit
TYPE_PRODUIT	Individuel ou collectif
DATE_NAISS	Date de naissance de la personne protégée
SEXE	Sexe de la personne protégée
CODE_DEPT	Code du département de la résidence de la personne protégée
DATE_ADH_GAR	Date d'adhésion à la garantie
DATE_RAD_GAR	Date de radiation de la garantie
DATE_ANCIENNETE	Date de la première souscription
ANNEE_NAISS	Année de naissance de la personne protégée
MOIS_NAISS	Mois de naissance de la personne protégée
ANNEE_MOIS_NAISS	Année et mois de naissance de la personne protégée
ANNEE_ADH_GAR	Année de l'adhésion de la personne protégée
MOIS_ADH_GAR	Mois d'adhésion de la personne protégée
ANNEE_MOIS_ADH_GAR	Année et mois d'adhésion de la personne protégée
ANNEE_RAD_GAR	Année de radiation de la personne protégée
MOIS_RAD_GAR	Mois de radiation de la personne protégée
ANNEE_MOIS_RAD_GAR	Année et mois de la personne protégée
ANNEE_ANC	Année de la première souscription
MOIS_ANC	Mois de la première souscription
ANNEE_MOIS_ANC	Année et mois de la première souscription
NB_MOIS_ANC	Ancienneté en mois
NIVEAU_GAR	Niveau de garantie de la personne protégée
NIVEAU_GAM	Niveau de gamme de la personne protégée : entrée de gamme, milieu de gamme
DT_DEB_EX	Date du début de l'exercice : correspond au 1er Janvier de l'exercice
DT_FIN_EX	Date de fin de l'exercice : correspond au 31 décembre de l'exercice
AGE_EX	Age exacte de l'assurée en année
ANNEE	Année de l'exercice
AGE_ARD	Age arrondi à l'unité inférieur
TRANCHE_AGE_1	Tranche d'âge par pas de 5 ans
TRANCHE_AGE_2	Tranche d'âge : "0;19"/"19;26"/"26;46"/"46;60"/"60 ans et plus"
NB_PP	Compteur des assurés toujours égal 1
NB_PP_MOY	Exposition sur l'exercice. Ex : si l'assuré est présent dans le portefeuille uniquement six mois de l'exercice ce champs prendra la valeur 0,5

Nom de variable	Description
NUM_PP	Numéro de personne (identifiant de la personne protégée)
CODE_GARANTIE	Code garantie
PRODUIT	Code produit
ANNEE	Exercice de survenance
DT_DEB_SOIN	Date de début des soins
DT_FIN_SOIN	Date de fin des soins
DT_PAIEMENT	Date de paiement du sinistre
DT_TRAITEMENT	Date du traitement du sinistre (date de connaissance du sinistre)
ACTE_RO	Code acte prenant du régime d'origine
ACTE_RC	Code acte de la mutuelle
LIB_ACTE_RC	Libellé des soins
RISQ_RC	Nature du risque : Dentaire, optique…
MT_DEPE_ASSURE	Montant dépensé par l'assuré
MT_AUTRE_MUT	Montant remboursé par d'autres mutuelles
MT_DEPA	Montant du dépassement par rapport au tarif de convention
MT_TICKET_MOD	Montant du ticket modérateur
TX_RO	Taux de remboursement de la sécurité sociale
TX_RC	Taux de remboursement de la complémentaire
MT_RO	Montant remboursé par la sécurité sociale
MT_RC	Montant remboursé par la mutuelle

Le tableau ci-dessous présente la codification des variables explicatives :

Variables explicatives	Modalité	Code modalité
Tranche d'âge	[0,9]	1
	[10,19]	2
	[20,24]	3
	[25,29]	4
	[30,34]	5
	[35,39]	6
	[40,44]	7
	[45,49]	8
	[50,54]	9
	[55,64]	10
	[65,74]	11
	[75,84]	12
	Plus de 85	13
Sexe	Homme	1
	Femme	2
Type personnes protégées	Assuré principal	1
	Conjoint	2
	Enfant	3
	Autre	4

5.2 Annexe 2 : Compléments Fréquence de consommation

Pour rappel, les valeurs de la statique Z obtenue par niveau de couverture pour les soins dentaires sur une période d'un an, sont les suivantes :

	Entrée de gamme	***Milieu de gamme***	***haut de gamme***
Z 2018 --2019	*0,00*	*0,11*	*-3,03*
Z 2019 --2020	*7,26*	*-2,58*	*-2,96*

Ces statistiques permettent de conclure que :

- Dans le cas bilatéral :
  - Pour les assurés avec une couverture « entrée de gamme », il existe une différence significative entre la proportion 2019 et celle de 2020. Cette affirmation n'est pas vérifiée en comparant 2018 à 2019 ;
  - Pour les assurés avec une couverture « milieu de gamme », il existe une différence significative entre la proportion 2019 et la proportion 2020. Cette affirmation n'est pas vérifiée en comparant 2018 à 2019 ;
  - Pour les assurés avec une couverture « haut de gamme », il existe une différence significative entre les proportions 2018 et 2019 comparées respectivement aux proportions 2019 et 2020.
- Dans le cas unilatéral à droite :
  - Pour les assurés avec une couverture « entrée de gamme », la proportion 2019 est significativement inférieur à celle de 2020 ;
  - Pour les assurés avec une couverture « milieu de gamme », la proportion 2020 est significativement inférieur à celle de 2019 ;
  - Pour les assurés avec une couverture « haut de gamme », les proportions 2018 et 2019 sont respectivement supérieures à celles de 2019 et 2020

Les statistiques obtenues pour les deux premiers mois de chaque année, sont les suivantes :

	Entrée de gamme	***Milieu de gamme***	***Haut de gamme***
Z 2018 --2019	*0,00*	*-0,31*	*-1,12*
Z 2019 --2020	*4,75*	*1,65*	*-1,25*

L'interprétation de ces statistiques est la suivante :

- Dans le cas bilatéral :
  - Pour les assurés avec une couverture « entrée de gamme », il existe une différence significative entre la proportion 2019 et celle de 2020. Cette affirmation n'est pas vérifiée en comparant 2018 à 2019 ;
  - Pour les assurés avec une couverture « milieu de gamme » et « haut de gamme », les valeurs prises par la statistiques Z permettent de conclure à aucune différence entre les proportions : 2018 comparée à 2019 et 2019 comparée à 2020.
- Dans le cas unilatéral à droite :
  - Pour les assurés avec une couverture « entrée de gamme », la proportion 2019 est significativement inférieure à celle des assurés y ayant eu recours en 2020. La valeur prise par la statistiques Z permet de conclure à l'absence différence entre les proportions 2018 comparées à 2019 ;

- Pour les assurés avec une couverture « milieu de gamme », la proportion 2020 est significativement supérieure à celle de 2019 ;
- Pour les assurés avec une couverture « haut de gamme », les valeurs prises par la statistique Z permettent de conclure à l'absence de différence entre les proportions 2018 comparée à 2019 et 2019 comparée à 2020

Les deux tableaux suivants permettent de disposer de la répartition du nombre d'actes et du montant des prestations par type de panier pour les soins de type « prothèses dentaires » :

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Nombre d'actes		Nombre d'actes		Nombre d'actes	
		Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
ANNEE	PANIER 100% SANTE						
2020	NON	560	21.70 %	1 170	30.10 %	279	43.05 %
	OUI	2 020	78.29 %	2 717	69.89 %	369	56.94 %
	Total année	2 580	100.00 %	3 887	100.00 %	648	100.00 %

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Montant des prestations		Montant des prestations		Montant des prestations	
		Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
ANNEE	PANIER 100% SANTE						
2020	NON	33 794	9.92 %	128 100	21.30 %	41 957	36.65 %
	OUI	306 556	90.07 %	473 233	78.69 %	72 495	63.34 %
	Total année	340 350	100.00 %	601 333	100.00 %	114 452	100.00 %

Nous pouvons constater que le panier 100% santé est plus utilisé par les adhérents « entrée de gamme » que par les adhérents « haut de gamme ».

Les deux tableaux suivants permettent de disposer de la répartition du nombre d'actes et du montant des prestations par type de panier pour les soins de type « équipements optiques » :

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Montant des prestations		Montant des prestations		Montant des prestations	
		Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
ANNEE	PANIER 100% SANTE						
2020	NON	382 131	80.54 %	659 379	96.78 %	112 099	99.81 %
	OUI	92 291	19.45 %	21 905	3.21 %	205	0.18 %
	Total année	474 422	100.00 %	681 284	100.00 %	112 304	100.00 %

		GAMME					
		1_ENTREE		2_MILIEU		3_HAUT	
		Nombre d'actes		Nombre d'actes		Nombre d'actes	
		Nombre	Fréquence	Nombre	Fréquence	Nombre	Fréquence
ANNEE	PANIER 100% SANTE						
2020	NON	5 452	76.40 %	5 972	94.15 %	783	99.49 %
	OUI	1 684	23.59 %	371	5.84 %	4	0.50 %
	Total année	7 136	100.00 %	6 343	100.00 %	787	100.00 %

5.3 Annexe 3 : Compléments Test des rangs signés de Wilcoxon

- Les résultats du test appliqués au nombre d'actes annuels pour les soins de type « équipements optiques » sont les suivants :

		Comparaison sur une année entière				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	3 398 655	Pr >= |S|	0,0002
	2_MILIEU	Rang signé	S	-3 995 000	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	-67 515	Pr >= |S|	0,0549
2019-2018	1_ENTREE	Rang signé	S	2 431 780	Pr >= |S|	0,0027
	2_MILIEU	Rang signé	S	1 604 073	Pr >= |S|	0,0388
	3_HAUT	Rang signé	S	-24 217	Pr >= |S|	0,5223

L'interprétation des différentes statistiques et P-values est la suivante :

- Comparaison sur une année entière :
  - La comparaison 2019 à 2020 révèle une différence significative pour les segments « entrée de gamme » et « milieu de gamme » ; le premier connait une augmentation et le second est en baisse ;
  - La comparaison 2018 à 2019 révèle une différence significative pour les segments « entrée de gamme » et « milieu de gamme. Les deux connaissent une augmentation. Les statistiques des tests sont toutefois moins significatives que celles du précèdent test.

- Les résultats du test appliqués au montant moyen annuel des prestations pour les soins de type « prothèses dentaires » sont les suivants :

		Comparaison sur une année entière				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	144553	Pr >= |S|	<,0001
	2_MILIEU	Rang signé	S	267589	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	7262,5	Pr >= |S|	<,0001
2019-2018	1_ENTREE	Rang signé	S	-7430,5	Pr >= |S|	0,3379
	2_MILIEU	Rang signé	S	-18423	Pr >= |S|	0,3007
	3_HAUT	Rang signé	S	-3758	Pr >= |S|	0,0059

Les statistiques et les P-values indiquent :

- Une différence significative est identifiée entre les années 2020 et 2019 pour les trois segments ;
- La comparaison des années 2018 et 2019, ne montre aucune différence.

- Les résultats du test appliqués au montant moyen annuel des prestations pour les soins de type « équipements optiques » sont les suivants :

		Comparaison sur une année entière				
		Test	Statistique		p-value	
2020-2019	1_ENTREE	Rang signé	S	-173248	Pr >= |S|	<,0001
	2_MILIEU	Rang signé	S	-178169	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	-1487,5	Pr >= |S|	0,1309
2019-2018	1_ENTREE	Rang signé	S	47704	Pr >= |S|	<,0001
	2_MILIEU	Rang signé	S	46779,5	Pr >= |S|	<,0001
	3_HAUT	Rang signé	S	2348,5	Pr >= |S|	0,0027

Les statistiques et les P-values indiquent :

- Une différence significative est identifiée entre les années 2020 et 2019 pour les segments « entrée de gamme » et « milieu de gamme ». Ces différences se matérialisent par une baisse des prestations ;.
- La comparaison des années 2018 et 2019, montre une différence pour tous les segments avec une tendance à la hausse.

- Les résultats du test appliqués au montant moyen annuel du reste à charge pour les soins de type « prothèses dentaires » sont les suivants :

		Comparaison sur une année entière				
		Test	*Statistique*		*p-value*	
2020-2019	*1_ENTREE*	*Rang signé*	*S*	-26506	*Pr >= |S|*	0,0035
	2_MILIEU	*Rang signé*	*S*	-60221	*Pr >= |S|*	0,0005
	3_HAUT	*Rang signé*	*S*	852	*Pr >= |S|*	0,5196
2019-2018	*1_ENTREE*	*Rang signé*	*S*	20686,5	*Pr >= |S|*	0,0136
	2_MILIEU	*Rang signé*	*S*	57069,5	*Pr >= |S|*	0,0022
	3_HAUT	*Rang signé*	*S*	-1634,5	*Pr >= |S|*	0,2486

L'interprétation de ces résultats est la suivante :

- Il existe une différence significative entre 2020 et 2019 pour les segments « entrée de gamme » et « milieu de gamme » avec une tendance à la baisse. Concernant le segment « haut de gamme » aucune différence n'est identifiée ;
- Il existe une différence significative entre 2019 et 2018 pour les segments « entrée de gamme » et « milieu de gamme » avec une tendance à la hausse. Concernant le segment « haut de gamme » aucune différence n'est identifiée.

- Les résultats du test appliqués au montant moyen annuel du reste à charge pour les soins de type « équipements optiques » sont les suivants :

		Comparaison sur une année entière				
		Test	*Statistique*		*p-value*	
2020-2019	*1_ENTREE*	*Rang signé*	*S*	105416	*Pr >= |S|*	<,0001
	2_MILIEU	*Rang signé*	*S*	224068,5	*Pr >= |S|*	<,0001
	3_HAUT	*Rang signé*	*S*	7437	*Pr >= |S|*	<,0001
2019-2018	*1_ENTREE*	*Rang signé*	*S*	64008	*Pr >= |S|*	<,0001
	2_MILIEU	*Rang signé*	*S*	98350,5	*Pr >= |S|*	<,0001
	3_HAUT	*Rang signé*	*S*	1374,5	*Pr >= |S|*	0,2052

L'interprétation de ces résultats est la suivante :

- La comparaison 2019 à 2020 révèle une différence significative pour les trois segments :ils connaissent tous une hausse ;
- La comparaison 2018 à 2019 révèle une différence pour les deux segments « entrée de gamme » et « milieu de gamme » ; les deux connaissent une hausse.

5.4 Annexe 4 : Résultats Test de Wilcoxon - Mann – Whitney

❖ ***Prothèses dentaires***

- **Nombre actes (Année entière) :**

		2020-2019	2019-2018
1_ENTREE	Statistique	1025849061	1178728735
	Z	-5,8048	-3,3743
	Unilatéral Pr	<,0001	0,0004
	Bilatéral Pr > |Z|	<,0001	0,0007
2_MILIEU	Statistique	196154834,5	257189873
	Z	1,6848	0,3756
	Unilatéral Pr	0,046	0,3536
	Bilatéral Pr > |Z|	0,092	0,7072
3_HAUT	Statistique	4793330	6101379,5
	Z	5,0956	1,8
	Unilatéral Pr	<,0001	0,0359
	Bilatéral Pr > |Z|	<,0001	0,0719

- Concernant les assurés avec des garanties « entrée de gamme » : les résultats des tests montrent une différence entre le nombre d'actes de l'année N et celui de l'année N-1. Cette différence se traduit par une augmentation sur deux années successives. Les p-values de la comparaison 2020 à 2019 sont toutefois plus significatives que celles de la comparaison de 2018 à 2019 ;
- Concernant les assurés avec des garanties « milieu de gamme » : les résultats du test unilatéral montrent une différence entre le nombre d'actes de l'année 2019 et celui de l'année 2020. Cette différence se traduit par une baisse de ce dernier. La p-value de la comparaison de l'année 2019 à 2020 est toutefois peu significative ;
- Concernant les assurés avec des garanties « haut de gamme » : les résultats du test unilatéral montrent une différence entre le nombre d'actes de l'année N et celui de l'année N-1. Cette différence se traduit par une baisse de ce dernier. La p-value de la comparaison de l'année 2018 à 2019 est toutefois peu significative.
- **Montant moyen des prestations et montant moyen du reste à charge (Année entière) :**

		Montant moyen des prestations		montant moyen du reste à charge	
		2020-2019	2019-2018	2020-2019	2019-2018
1_ENTREE	Statistique	1162033	1760678.5000	1876876	1571844,5
	Z	-15,6602	5.2364	16,1417	-3,0152
	Unilatéral Pr	<,0001	<.0001	<,0001	0,0013
	Bilatéral Pr > |Z|	<,0001	<.0001	<,0001	0,0026
2_MILIEU	Statistique	1990995	3359051.5000	2615641,5	3131392,5
	Z	-11,4643	7.6491	11,4504	1,2154
	Unilatéral Pr	<,0001	<.0001	<,0001	0,1121
	Bilatéral Pr > |Z|	<,0001	<.0001	<,0001	0,2242
3_HAUT	Statistique	75685	140547.5000	65959,5	137284
	Z	1,3768	1.7970	-3,945	0,751
	Unilatéral Pr	0,0843	0.0362	<,0001	0,2263
	Bilatéral Pr > |Z|	0,1686	0.0723	<,0001	0,4526

- Concernant les assurés avec des garanties « entrée de gamme » : les résultats des tests montrent une différence entre l'année N et celui de l'année N-1 pour les deux variables : montant moyen des prestations et montant moyen du reste à charge. Cette différence se traduit par une augmentation des prestations en 2020 contre une baisse en 2019 et une baisse du reste à charge en 2020 contre une hausse en 2019. Les p-values de la comparaison 2020 à 2019 sont toutefois plus significatives que celles de la comparaison de 2018 à 2019 ;

- Concernant les assurés avec des garanties « milieu de gamme , les s résultats des tests montrent une différence entre l'année N et celui de l'année N-1 pour les deux variables : montant moyen des prestations et montant moyen du reste à charge (la différence est constatée uniquement entre 2020 et 2019). Cette différence se traduit par une augmentation des prestations en 2020 contre une baisse en 2019 et une baisse du reste à charge en 2020 contre une stabilité en 2019. Les p-values sont toutes très significatives ;
- Concernant les assurés avec des garanties « haut de gamme » : les résultats du test unilatéral montrent une différence entre le reste à charge de l'année 2020 comparé à celui de l'année 2019 et le montant des prestations de l'année 2019 comparé à celui de 2018 (avec une p-value peu significative). Ces différences se traduisent par une baisse du montant des prestations en 2019 et une hausse du reste à charge en 2020.

❖ ***Equipements optique :***

- **Nombre actes (Année entière) :**

		2020-2019	2019-2018
1_ENTREE	Statistique	1023295937	1172576491
	Z	-4,9493	-5,3465
	Unilatéral Pr	<,0001	<,0001
	Bilatéral Pr > |Z|	<,0001	<,0001
2_MILIEU	Statistique	197607712	258049456,5
	Z	3,7817	0,9967
	Unilatéral Pr	<,0001	0,1595
	Bilatéral Pr > |Z|	0,0002	0,3189
3_HAUT	Statistique	4737965	6097924,5
	Z	2,1293	1,476
	Unilatéral Pr	0,0166	0,07
	Bilatéral Pr > |Z|	0,0332	0,1399

- Concernant les assurés avec des garanties « entrée de gamme » : les résultats des tests montrent une différence entre le nombre d'actes de l'année N et celui de l'année N-1. Cette différence se traduit par une augmentation sur deux années successives ;
- Concernant les assurés avec des garanties « milieu de gamme » ou « haut de gamme » : les résultats du test unilatéral montrent une différence entre le nombre d'actes de l'année 2019 et celui de l'année 2020. Cette différence se traduit par une baisse de ce dernier.
- **Montant moyen des prestations et montant moyen du reste à charge (Année entière)**

		Montant moyen des prestations		montant moyen du reste à charge	
		2020-2019	2019-2018	2020-2019	2019-2018
1_ENTREE	Statistique	17111809,5	17060753,5	15765598	16729227
	Z	11,2748	0,6968	0,6464	-3,1604
	Unilatéral Pr	<,0001	0,243	0,259	0,0008
	Bilatéral Pr > |Z|	<,0001	0,4859	0,518	0,0016
2_MILIEU	Statistique	11549788	14365091	9210975,5	14002429,5
	Z	16,0975	0,0516	-11,9795	-3,2105
	Unilatéral Pr	<,0001	0,4794	<,0001	0,0007
	Bilatéral Pr > |Z|	<,0001	0,9588	<,0001	0,0013
3_HAUT	Statistique	340016,5	404005	288057	391156,5
	Z	5,6194	1,9622	-3,036	0,1318
	Unilatéral Pr	<,0001	0,0249	0,0012	0,4476
	Bilatéral Pr > |Z|	<,0001	0,0497	0,0024	0,8951

- Concernant les assurés avec des garanties « entrée de gamme », les résultats des tests montrent une différence entre l'année 2020 et ceux de l'année 2012 pour le montant moyen des prestations. Cette différence se traduit par une baisse en 2020. Le montant moyen du reste à charge de l'année 2019 est en augmentation par rapport à celui de 2018 ;
- Concernant les assurés avec des garanties « milieu de gamme » , les résultats des tests montrent une différence entre l'année 2020 et ceux de l'année 2019 pour les deux variables : montant moyen des prestations et montant moyen du reste à charge. Cette différence se traduit par une baisse des prestations en 2020 accompagnée d'une hausse du reste à charge. Une hausse que nous constatons aussi en comparant 2018 à 2019 ;
- Concernant les assurés avec des garanties « milieu de gamme » , les résultats des tests montrent une différence entre l'année 2020 et ceux de l'année 2019 pour les deux variables : montant moyen des prestations et montant moyen du reste à charge. Cette différence se traduit par une baisse des prestations en 2020 (baisse constatée aussi en 2019) accompagnée d'une hausse du reste à charge.

5.5 Annexe 5 : Coefficient d'asymétrie, coefficient d'aplatissement et méthode des noyaux

Le coefficient d'asymétrie ainsi que le coefficient d'aplatissement font simultanément appel aux moments d'ordre 3 et 4. Pour rappel, le moment centré d'ordre k d'une variable aléatoire X est défini de la manière suivante :

$$\mu_k = E[(X - E(X))^k].$$

Pour une suite de N observations de la variable X notées : $x_1, x_2, \dots\dots, x_N$, le moment empirique centrée d'ordre k est lui noté :

$$\bar{\mu}_k = \sum_{i=1}^{N} (x_i - \bar{x})^k \; ; \; \bar{x} \;\; \acute{e}tant\ la\ moyenne\ empirique\ .$$

❖ Le coefficient d'asymétrie appelé aussi coefficient de Skewness est défini par : $\gamma_1 = \frac{\mu_3}{{\mu_2}^{\frac{3}{2}}}$ .Il permet de mesurer le degré de symétrie d'une distribution :
- Un coefficient de nul veut dire que la distribution est symétrique ;
- Un coefficient positif veut dire que la distribution possède une forte queue vers la droite ;
- Un coefficient négatif veut dire que la distribution possède une forte queue vers la gauche.

❖ Le coefficient d'aplatissement proposé par Pearson est défini par : $\beta_2 = \frac{\mu_4}{{\mu_2}^2}$. Le coefficient appelé Kurtosis proposé par Fisher et généralement affiché par la plupart des logiciels est le coefficient γ_2 défini par : $\gamma_2 = \beta_2 - 3$, car il permet de faire référence à la distribution d'une loi normale pour laquelle : $\beta_2 = 3$.

Le coefficient d'aplatissement renseigne sur le degré d'aplatissement de la courbe de fréquences d'une distribution. L'aplatissement est jugé en se référant au modèle de la courbe de densité de la loi normale. On dira qu'une courbe de fréquences est plus ou moins aplatie que le modèle de la loi normale. Ces valeurs sont à interpréter de la manière suivante :

- Si $\gamma_2 = 0$, la courbe de fréquences est comparable à celle de la loi normale ;
- Si $\gamma_2 > 0$, la courbe de fréquences est plus pointue que celle de la loi normale ;
- Si $\gamma_2 < 0$, la courbe de fréquences est plus aplatie que celle de la loi normale.

❖ La méthode d'estimation par noyaux est une méthode d'estimation non paramétrique, qui a été initiée par Rosenblatt en 1956 et développée par Parzen en 1962. Elle s'appuie sur deux éléments que sont la fonction noyau et le paramètre de lissage appelé aussi fenêtre.

Pour un échantillon $X_1, X_2, \dots\dots, X_n$issu d'une v.a. X de fonction de densité de probabilité f. Un estimateur de la densité selon cette méthode est de la forme :

$$\widehat{f_h}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x - X_i}{h}),$$

où h est un paramètre appelé paramètre de lissage qui satisfait : $\lim_{n\to\infty} h(n) = 0$. La fonction $K(y)$ appelée noyau vérifie :

- $K(y) \geq 0$;
- $\int_{-\infty}^{+\infty} K(y)dy = 1$;
- $\sup_{K\in\mathbb{R}} K(y) < \infty$;
- $K(y) = K(-y)$;
- $\lim_{y\to\infty} |yK(y)| = 0.$

Parmi les différentes fonctions noyaux disponibles, nous pouvons citer :

- Le noyau gaussien défini par : $K(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2}$;
- Le noyau logistique défini par :$K(y) = \frac{e^{-|y|}}{(1+e^{-|y|})^2}$.

A noter que le choix de la fonction noyau a peu d'impact sur la qualité d'estimation (à condition que ce noyau soit lisse, comme par ex. le noyau gaussien ou le noyau logistique ci-dessus), le paramètre de lissage lui a un impact très significatif sur les résultats.

5.6 Annexe 6 : Modélisation de la variable recours aux équipements optiques

Nous allons présenter dans cette section les statistiques d'ajustement qui permettent de valider le modèle retenu.

Entrée de gamme :

Les critères AIC du modèle complet et du modèle réduit (modèle retenu) ainsi que le résultat du test du rapport de la vraisemblance sont résumés dans le tableau suivant :

	Statistique d'ajustement du modèle		*Test du rapport de la vraisemblance*	
	Modèle sous contrainte	*Modèle complet*		
AIC	60381,449	60385,204	D	0,245
-2 Log L	60323,449	60323,204	P-value	0,885

Nous pouvons constater que le modèle complet possède un critère AIC plus grand et que la p-value du test du rapport de la vraisemblance est supérieur à 0.05 ce qui ne permet de rejeter l'hypothèse nulle. Ces deux informations confirment la pertinence de retenir le modèle réduit.

L'analyse des effets par niveau de variable est présentée dans le tableau suivant :

Estimation du rapport de cotes			
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_SEXE 1 vs 2	0.712	0.683	0.743
ANNEE_ANC 0 vs 8	0.890	0.674	1.174
ANNEE_ANC 1 vs 8	1.195	0.906	1.575
ANNEE_ANC 2 vs 8	1.204	0.912	1.590
ANNEE_ANC 3 vs 8	1.234	0.934	1.630
ANNEE_ANC 4 vs 8	1.205	0.911	1.594
ANNEE_ANC 5 vs 8	1.099	0.830	1.455
ANNEE_ANC 6 vs 8	1.122	0.846	1.488
ANNEE_ANC 7 vs 8	1.073	0.804	1.432
CD_TRANCHE_AGE 1 vs 13	0.725	0.611	0.860
CD_TRANCHE_AGE 2 vs 13	1.205	1.017	1.428
CD_TRANCHE_AGE 3 vs 13	0.986	0.827	1.175
CD_TRANCHE_AGE 4 vs 13	0.918	0.772	1.091
CD_TRANCHE_AGE 5 vs 13	0.802	0.670	0.960
CD_TRANCHE_AGE 6 vs 13	0.857	0.713	1.029
CD_TRANCHE_AGE 7 vs 13	1.103	0.919	1.322
CD_TRANCHE_AGE 8 vs 13	1.903	1.600	2.262
CD_TRANCHE_AGE 9 vs 13	2.027	1.703	2.413
CD_TRANCHE_AGE 10 vs 13	1.971	1.678	2.314
CD_TRANCHE_AGE 11 vs 13	2.151	1.837	2.519
CD_TRANCHE_AGE 12 vs 13	1.884	1.586	2.238
NIVEAU 1 vs 2	0.711	0.682	0.742
CD_ZONE 1 vs 7	0.936	0.680	1.289
CD_ZONE 2 vs 7	0.934	0.684	1.276
CD_ZONE 3 vs 7	0.945	0.697	1.283
CD_ZONE 4 vs 7	0.861	0.634	1.171
CD_ZONE 5 vs 7	1.128	0.782	1.627
CD_ZONE 6 vs 7	0.983	0.722	1.339

Ci-dessous l'interprétation de ces statistiques :

- Les hommes ont moins de chances d'avoir recours à ses soins ;
- Comparés aux assurés avec 8 ans d'ancienneté, les nouveaux adhérents de l'année ont moins de chance d'avoir recours à ses soins alors que les autres en ont plus ;
- Comparés à la dernière tranche d'âge, les tranches 1, 3,4,5,6 ont moins de chance d'avoir recours à ces soins alors que les autres en ont plus. Par exemple, un assuré de la tranche 11 aura deux fois plus de chance d'avoir recours à un équipement optique ;
- Exceptés les assurés de la cinquième zone géographique, tous les autres auront plus de chance d'avoir recours à ces soins.

Milieu de gamme

Les critères AIC du modèle complet et du modèle réduit (modèle retenu) ainsi que le résultat du test du rapport de la vraisemblance sont résumés dans le tableau suivant :

	Statistique d'ajustement du modèle		*Test du rapport de la vraisemblance*	
	Modèle sous contrainte	*Modèle complet*		
AIC	44147,982	44149,701	D	0,281
-2 Log L	44087,982	44087,701	P-value	0,869

Nous pouvons constater là encore que le modèle complet possède un critère AIC plus grand et que la p-value du test du rapport de la vraisemblance est supérieur à 0.05, ce qui ne permet de rejeter l'hypothèse nulle. Ces deux informations confirment la pertinence de retenir le modèle réduit.

Les statistiques décrivant les effets des niveaux de chaque variable explicative sont résumées dans le tableau suivant :

	Estimation du rapport de cotes		
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_TYPE_PP 2 vs 4	0.960	0.898	1.027
CD_TYPE_PP 3 vs 4	0.699	0.565	0.864
CD_SEXE 1 vs 2	0.790	0.753	0.828
ANNEE_ANC 0 vs 8	1.116	0.843	1.476
ANNEE_ANC 1 vs 8	1.285	0.972	1.698
ANNEE_ANC 2 vs 8	1.281	0.969	1.695
ANNEE_ANC 3 vs 8	1.282	0.969	1.696
ANNEE_ANC 4 vs 8	1.340	1.013	1.774
ANNEE_ANC 5 vs 8	1.235	0.933	1.634
ANNEE_ANC 6 vs 8	1.236	0.932	1.639
ANNEE_ANC 7 vs 8	1.249	0.937	1.664
CD_TRANCHE_AGE 1 vs 13	1.674	1.117	2.511
CD_TRANCHE_AGE 2 vs 13	1.725	1.164	2.557
CD_TRANCHE_AGE 3 vs 13	1.501	1.051	2.144
CD_TRANCHE_AGE 4 vs 13	1.232	0.875	1.733
CD_TRANCHE_AGE 5 vs 13	1.188	0.844	1.670
CD_TRANCHE_AGE 6 vs 13	1.015	0.719	1.433
CD_TRANCHE_AGE 7 vs 13	1.320	0.937	1.861
CD_TRANCHE_AGE 8 vs 13	1.997	1.425	2.798
CD_TRANCHE_AGE 9 vs 13	1.879	1.342	2.632
CD_TRANCHE_AGE 10 vs 13	1.798	1.296	2.493
CD_TRANCHE_AGE 11 vs 13	1.958	1.413	2.712
CD_TRANCHE_AGE 12 vs 13	1.678	1.190	2.367
CD_ZONE 1 vs 7	1.329	0.863	2.048
CD_ZONE 2 vs 7	1.328	0.866	2.038
CD_ZONE 3 vs 7	1.301	0.851	1.987
CD_ZONE 4 vs 7	1.191	0.778	1.823
CD_ZONE 5 vs 7	1.042	0.613	1.770
CD_ZONE 6 vs 7	1.309	0.852	2.012

Ces statistiques peuvent être interprétées de la manière suivante :

- Les hommes ont moins de chance d'avoir recours à ces soins ;
- Les enfants et les conjoints ont moins de chance d'avoir recours à ces soins ;

- Comparés aux assurés avec une ancienneté de 8 ans, les autres assurés ont plus de chance d'avoir recours à ces soins ;
- Comparés aux assurés de la tranche d'âge 13, les autres assurés ont plus de chance d'avoir recours à ces soins ;
- Comparés aux assurés de la zone géographique 7, les autres assurés ont plus de chance d'avoir recours à ces soins.

Haut de gamme

Les critères AIC du modèle complet et du modèle réduit (modèle retenu) ainsi que le résultat du test du rapport de la vraisemblance sont résumés dans le tableau suivant :

	Statistique d'ajustement du modèle		*Test du rapport de la vraisemblance*	
	Modèle sous contrainte	*Modèle complet*		
AIC	6032,375	6040,641	D	0,274
-2 Log L	5978,645	5978,371	P-value	0,872

Les mêmes remarques sont encore valables pour ce segment.

Les statistiques décrivant les effets des niveaux de chaque variable explicative sont résumées dans le tableau suivant :

	Estimation du rapport de cotes		
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_SEXE 1 vs 2	0.830	0.732	0.942
CD_TRANCHE_AGE 1 vs 12	0.380	0.218	0.663
CD_TRANCHE_AGE 2 vs 12	0.488	0.300	0.792
CD_TRANCHE_AGE 3 vs 12	0.624	0.361	1.080
CD_TRANCHE_AGE 4 vs 12	0.373	0.214	0.649
CD_TRANCHE_AGE 5 vs 12	0.352	0.202	0.615
CD_TRANCHE_AGE 6 vs 12	0.352	0.196	0.630
CD_TRANCHE_AGE 7 vs 12	0.517	0.300	0.888
CD_TRANCHE_AGE 8 vs 12	0.492	0.286	0.845
CD_TRANCHE_AGE 9 vs 12	0.660	0.390	1.116
CD_TRANCHE_AGE 10 vs 12	0.694	0.428	1.126
CD_TRANCHE_AGE 11 vs 12	0.648	0.400	1.049
CD_TRANCHE_AGE 13 vs 12	<0.001	<0.001	>999.999
CD_ZONE 1 vs 7	2.160	0.724	6.445
CD_ZONE 2 vs 7	2.165	0.746	6.280
CD_ZONE 3 vs 7	2.280	0.793	6.559
CD_ZONE 4 vs 7	2.078	0.720	5.997
CD_ZONE 5 vs 7	4.670	1.479	14.741
CD_ZONE 6 vs 7	2.794	0.960	8.132

L'interprétation de ces statistiques est la suivante :

- Les hommes ont moins de chance d'avoir recours à ces soins ;
- Comparés aux assurés de la tranche d'âge 12, les autres assurés ont plus de chance d'avoir recours à ces soins ;
- Comparés aux assurés de la zone géographique 7, les autres assurés ont plus de chance d'avoir recours à ces soins.

5.7 Annexe 7 : Modélisation à partir des données antérieures à 2020 :

Dans ce qui suit du nous allons présenter les différentes statistiques à l'image de celles produites dans le cadre des modèles réalisés à partir des données de l'année 2020. Ces statistiques concernent aussi bien la variable T_i que la variable $X_i * 1_{T_i=1}$

Variable recours aux soins T_i

- Prothèses dentaires

Entrée de gamme

Sélection des variables explicatives :

Récapitulatif sur la sélection Stepwise

Etape	Effet Saisi	Effet Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
1	CD_TRANCHE_AGE		12	1	3050.1447		<.0001
2	NIVEAU		1	2	363.9062		<.0001
3	ANNEE_ANC		7	3	35.5205		<.0001
4	CD_ZONE		6	4	32.1370		<.0001

Critères d'ajustement du modèle

	Statistique d'ajustement du modèle			
	Modèle sous contrainte	*Modèle complet*		*Test du rapport de la vraisemblance*
AIC	25071,866	25075,809	D	-4,06
-2 Log L	25013,809	25017,866	P-value	0,24

Effets des différents niveaux des variables explicatives

	Estimation du rapport de cotes		
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
ANNEE_ANC 0 vs 7	0.702	0.487	1.013
ANNEE_ANC 1 vs 7	0.892	0.619	1.286
ANNEE_ANC 2 vs 7	0.932	0.646	1.344
ANNEE_ANC 3 vs 7	0.890	0.616	1.286
ANNEE_ANC 4 vs 7	0.947	0.655	1.369
ANNEE_ANC 5 vs 7	0.919	0.635	1.332
ANNEE_ANC 6 vs 7	0.898	0.614	1.314
CD_TRANCHE_AGE 1 vs 13	<0.001	<0.001	>999.999
CD_TRANCHE_AGE 2 vs 13	0.033	0.020	0.054
CD_TRANCHE_AGE 3 vs 13	0.105	0.075	0.147
CD_TRANCHE_AGE 4 vs 13	0.174	0.131	0.229
CD_TRANCHE_AGE 5 vs 13	0.342	0.266	0.440
CD_TRANCHE_AGE 6 vs 13	0.375	0.289	0.486
CD_TRANCHE_AGE 7 vs 13	0.531	0.414	0.681
CD_TRANCHE_AGE 8 vs 13	0.537	0.419	0.687
CD_TRANCHE_AGE 9 vs 13	0.751	0.592	0.954
CD_TRANCHE_AGE 10 vs 13	0.931	0.756	1.145
CD_TRANCHE_AGE 11 vs 13	1.302	1.064	1.592
CD_TRANCHE_AGE 12 vs 13	1.447	1.162	1.800
NIVEAU 1 vs 2	0.496	0.460	0.535
CD_ZONE 1 vs 7	0.667	0.391	1.138
CD_ZONE 2 vs 7	0.736	0.439	1.234
CD_ZONE 3 vs 7	0.798	0.481	1.321
CD_ZONE 4 vs 7	0.873	0.526	1.450
CD_ZONE 5 vs 7	1.051	0.582	1.897
CD_ZONE 6 vs 7	0.980	0.589	1.630

Score de concordance (échantillon d'apprentissage)

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	78.3	D de Somers	0.569
Pourcentage discordant	21.4	Gamma	0.570
Pourcentage lié	0.3	Tau-a	0.054
Paires	251927232	c	0.784

Score de concordance (échantillon de validation)

		Consommateurs prédits *NON* *Nombre d'assurés*	
		Nombre	*Fréquence*
Consommateurs réels	*Bien classé*		
NON	*OUI*	30 023	100
	Total	30 023	95
OUI	*Bien classé*		
	NON	1 530	100
	Total	1 530	5

<u>Milieu de gamme</u>

Sélection des variables explicatives

Récapitulatif sur la sélection Stepwise						
Etape	*Effet Saisi*	*Effet Supprimé*	*DDL*	*Nombre dans*	*Khi-2 du score*	*Khi-2 de Wald*	*Pr > khi-2*
1	CD_TRANCHE_AGE		12	1	1682.4324		<.0001
2	NIVEAU		1	2	66.4870		<.0001
3	CD_SEXE		1	3	11.4801		0.0007

Critères d'ajustement du modèle

	Statistique d'ajustement du modèle			
	Modèle sous contrainte	*Modèle complet*		*Test du rapport de la vraisemblance*
AIC	27679,492	27683,462	D	-28,03
-2 Log L	27621,462	27649,492	P-value	0,24

Effets des différents niveaux des variables explicatives

Estimation du rapport de cotes			
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_SEXE 1 vs 2	0.898	0.843	0.956
CD_TRANCHE_AGE 1 vs 13	<0.001	<0.001	>999.999
CD_TRANCHE_AGE 2 vs 13	0.022	0.012	0.040
CD_TRANCHE_AGE 3 vs 13	0.174	0.111	0.272
CD_TRANCHE_AGE 4 vs 13	0.275	0.184	0.412
CD_TRANCHE_AGE 5 vs 13	0.419	0.284	0.619
CD_TRANCHE_AGE 6 vs 13	0.494	0.335	0.729
CD_TRANCHE_AGE 7 vs 13	0.660	0.449	0.968
CD_TRANCHE_AGE 8 vs 13	0.764	0.523	1.114
CD_TRANCHE_AGE 9 vs 13	0.774	0.531	1.128
CD_TRANCHE_AGE 10 vs 13	0.990	0.691	1.419
CD_TRANCHE_AGE 11 vs 13	1.170	0.817	1.675
CD_TRANCHE_AGE 12 vs 13	1.307	0.893	1.914
NIVEAU 3 vs 4	0.755	0.706	0.808

Score de concordance (échantillon d'apprentissage)

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	65.5	D de Somers	0.362
Pourcentage discordant	29.4	Gamma	0.381
Pourcentage lié	5.1	Tau-a	0.073
Paires	180528552	c	0.681

Score de concordance (échantillon de validation)

		Consommateurs prédits NON	
		Nombre d'assurés	
		Nombre	*Fréquence*
Consommateurs réels	*Bien classé*		
NON	*OUI*	16 087	100
	Total	16 087	89
OUI	*Bien classé*		
	NON	2 065	100
	Total	2 065	11

Haut de gamme

Sélection des variables explicatives

Récapitulatif sur la sélection Stepwise						
Etape	*Saisi*	*Effet Supprimé*	*DDL*	*Nombre dans*	*Khi-2 du score*	*Khi-2 de Wald*	*Pr > khi-2*
1	CD_TRANCHE_AGE		12	1	363.3114		<.0001
2	NIVEAU		1	2	8.1791		0.0042
3	ANNEE_ANC		7	3	14.3514		0.0453

Critères d'ajustement du modèle

	Statistique d'ajustement du modèle			
	Modèle sous contrainte	*Modèle complet*	*Test du rapport de la vraisemblance*	
AIC	3893,972	3902,689	D	-9,28
-2 Log L	3842,689	3851,972	P-value	0,24

Effets des différents niveaux des variables explicatives

Estimation du rapport de cotes			
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
ANNEE_ANC 0 vs 7	2.232	0.773	6.447
ANNEE_ANC 1 vs 7	2.227	0.772	6.423
ANNEE_ANC 2 vs 7	1.898	0.657	5.487
ANNEE_ANC 3 vs 7	1.973	0.680	5.726
ANNEE_ANC 4 vs 7	1.519	0.526	4.386
ANNEE_ANC 5 vs 7	1.583	0.544	4.612
ANNEE_ANC 6 vs 7	1.695	0.574	5.008
CD_TRANCHE_AGE 1 vs 13	<0.001	<0.001	>999.999
CD_TRANCHE_AGE 2 vs 13	0.019	0.002	0.208
CD_TRANCHE_AGE 3 vs 13	0.287	0.032	2.601
CD_TRANCHE_AGE 4 vs 13	0.847	0.101	7.127
CD_TRANCHE_AGE 5 vs 13	0.884	0.105	7.468
CD_TRANCHE_AGE 6 vs 13	0.937	0.111	7.903
CD_TRANCHE_AGE 7 vs 13	1.430	0.171	11.945
CD_TRANCHE_AGE 8 vs 13	1.679	0.201	13.992
CD_TRANCHE_AGE 9 vs 13	1.991	0.240	16.535
CD_TRANCHE_AGE 10 vs 13	1.963	0.239	16.110
CD_TRANCHE_AGE 11 vs 13	1.969	0.240	16.155
CD_TRANCHE_AGE 12 vs 13	2.911	0.333	25.446
NIVEAU 5 vs 6	0.779	0.662	0.916

Score de concordance (échantillon d'apprentissage)

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	71.2	D de Somers	0.434
Pourcentage discordant	27.8	Gamma	0.439
Pourcentage lié	1.0	Tau-a	0.103
Paires	3560720	c	0.717

Score de concordance (échantillon de validation)

		Consommateurs prédits *NON* *Nombre d'assurés*	
		Nombre	*Fréquence*
Consommateurs réels	*Bien classé*		
NON	*OUI*	2 074	100
	Total	2 074	87
OUI	*Bien classé*		
	NON	322	100
	Total	322	13

- Equipements optiques

<u>Entrée de gamme</u>

Sélection des variables explicatives

Récapitulatif sur la sélection Stepwise						
Etape	*Saisi*	*Effet Supprimé*	*DDL*	*Nombre dans*	*Khi-2 du score*	*Khi-2 de Wald*	*Pr > khi-2*
1	CD_TRANCHE_AGE		12	1	1485.3422		<.0001
2	NIVEAU		1	2	434.2784		<.0001
3	CD_SEXE		1	3	274.8114		<.0001
4	ANNEE_ANC		7	4	192.4680		<.0001
5	CD_ZONE		6	5	13.2172		0.0397

Critères d'ajustement du modèle

	Statistique d'ajustement du modèle			
	Modèle sous contrainte	*Modèle complet*		*Test du rapport de la vraisemblance*
AIC	58598,44	58598,994	D	-5,45
-2 Log L	58536,994	58542,44	P-value	0,24

Effets des différents niveaux des variables explicatives

	Estimation du rapport de cotes		
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_SEXE 1 vs 2	0.696	0.667	0.726
ANNEE_ANC 0 vs 7	0.733	0.576	0.933
ANNEE_ANC 1 vs 7	1.081	0.849	1.375
ANNEE_ANC 2 vs 7	1.018	0.799	1.297
ANNEE_ANC 3 vs 7	1.104	0.865	1.407
ANNEE_ANC 4 vs 7	1.026	0.803	1.310
ANNEE_ANC 5 vs 7	1.009	0.789	1.290
ANNEE_ANC 6 vs 7	1.007	0.783	1.295
CD_TRANCHE_AGE 1 vs 13	0.734	0.615	0.875
CD_TRANCHE_AGE 2 vs 13	1.214	1.019	1.447
CD_TRANCHE_AGE 3 vs 13	1.047	0.876	1.253
CD_TRANCHE_AGE 4 vs 13	0.924	0.774	1.104
CD_TRANCHE_AGE 5 vs 13	0.770	0.640	0.927
CD_TRANCHE_AGE 6 vs 13	0.846	0.700	1.022
CD_TRANCHE_AGE 7 vs 13	0.980	0.811	1.184
CD_TRANCHE_AGE 8 vs 13	1.641	1.370	1.967
CD_TRANCHE_AGE 9 vs 13	1.839	1.535	2.205
CD_TRANCHE_AGE 10 vs 13	1.658	1.401	1.962
CD_TRANCHE_AGE 11 vs 13	1.949	1.652	2.301
CD_TRANCHE_AGE 12 vs 13	1.582	1.318	1.898
NIVEAU 1 vs 2	0.651	0.623	0.680
CD_ZONE 1 vs 7	1.288	0.867	1.913
CD_ZONE 2 vs 7	1.375	0.932	2.028
CD_ZONE 3 vs 7	1.322	0.901	1.941
CD_ZONE 4 vs 7	1.282	0.872	1.885
CD_ZONE 5 vs 7	1.558	1.009	2.405

Score de concordance (échantillon d'apprentissage)

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	78.3	D de Somers	0.569
Pourcentage discordant	21.4	Gamma	0.570
Pourcentage lié	0.3	Tau-a	0.054
Paires	251927232	c	0.784

Score de concordance (échantillon de validation)

		Consommateurs prédits *NON* *Nombre d'assurés*	
		Nombre	*Fréquence*
Consommateurs réels	*Bien classé*		
NON	*NON*	2 074	87
	Total	2 074	87
OUI	*Bien classé*		
	NON	322	13
	Total	322	13

<u>Milieu de gamme</u>

Sélection des variables explicatives

Récapitulatif sur la sélection Stepwise						
Etape	*Effet Saisi*	*Effet Supprimé*	*DDL*	*Nombre dans*	*Khi-2 du score*	*Khi-2 de Wald*	*Pr > khi-2*
1	CD_TRANCHE_AGE		12	1	330.5852		<.0001
2	CD_SEXE		1	2	109.1625		<.0001
3	NIVEAU		1	3	22.1262		<.0001
4	CD_TYPE_PP		3	4	20.4103		0.0001
5	ANNEE_ANC		7	5	29.7667		0.0001

Critères d'ajustement du modèle

	Statistique d'ajustement du modèle			
	Modèle sous contrainte	*Modèle complet*	*Test du rapport de la vraisemblance*	
AIC	46644,097	46645,466	D	-10,63
-2 Log L	46583,466	46594,097	P-value	0,24

Effets des différents niveaux des variables explicatives

	Estimation du rapport de cotes		
Effet	*Estimation du point*	*Intervalle de confiance de Wald à95%*	
CD_TYPE_PP 1 vs 4	0.608	0.269	1.373
CD_TYPE_PP 2 vs 4	0.631	0.279	1.428
CD_TYPE_PP 3 vs 4	0.386	0.168	0.885
CD_SEXE 1 vs 2	0.785	0.750	0.822
ANNEE_ANC 0 vs 7	0.816	0.649	1.028
ANNEE_ANC 1 vs 7	0.918	0.729	1.155
ANNEE_ANC 2 vs 7	0.969	0.770	1.220
ANNEE_ANC 3 vs 7	0.986	0.783	1.242
ANNEE_ANC 4 vs 7	0.930	0.738	1.171
ANNEE_ANC 5 vs 7	0.978	0.775	1.233
ANNEE_ANC 6 vs 7	0.927	0.730	1.176
CD_TRANCHE_AGE 1 vs 13	1.328	0.901	1.957
CD_TRANCHE_AGE 2 vs 13	1.747	1.200	2.543
CD_TRANCHE_AGE 3 vs 13	1.351	0.963	1.895
CD_TRANCHE_AGE 4 vs 13	1.164	0.840	1.613
CD_TRANCHE_AGE 5 vs 13	0.937	0.675	1.301
CD_TRANCHE_AGE 6 vs 13	0.905	0.650	1.260
CD_TRANCHE_AGE 7 vs 13	1.067	0.767	1.485
CD_TRANCHE_AGE 8 vs 13	1.513	1.093	2.095
CD_TRANCHE_AGE 9 vs 13	1.559	1.127	2.157
CD_TRANCHE_AGE 10 vs 13	1.506	1.099	2.063
CD_TRANCHE_AGE 11 vs 13	1.543	1.127	2.112
CD_TRANCHE_AGE 12 vs 13	1.482	1.062	2.070
NIVEAU 3 vs 4	0.884	0.841	0.929

Score de concordance (échantillon d'apprentissage)

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	57.0	D de Somers	0.145
Pourcentage discordant	42.5	Gamma	0.145
Pourcentage lié	0.5	Tau-a	0.054
Paires	330653295	c	0.572

Score de concordance (échantillon de validation)

		Consommateurs prédits *NON* *Nombre d'assurés*	
		Nombre	*Fréquence*
Consommateurs réels	*Bien classé*		
NON	*NON*	2 074	87
	Total	2 074	87
OUI	*Bien classé*		
	NON	322	13
	Total	322	13

Haut de gamme

Sélection des variables explicatives

Récapitulatif sur la sélection Stepwise						
Etape	Effet Saisi	Supprimé	DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2
1	CD_TRANCHE_AGE		12	1	81.0281		<.0001
2	CD_TYPE_PP		3	2	13.7169		0.0033
3		CD_TYPE_PP	3	1		7.2274	0.0650

Critères d'ajustement du modèle

Statistique d'ajustement du modèle				
	Modèle sous contrainte	Modèle complet	Test du rapport de la vraisemblance	
AIC	6175,103	6162,037	D	-49,07
-2 Log L	6100,037	6149,103	P-value	0,24

Effets des différents niveaux des variables explicatives

Estimation du rapport de cotes			
Effet	Estimation du point	Intervalle de confiance de Wald à95%	
CD_TRANCHE_AGE 1 vs 13	0.500	0.094	2.660
CD_TRANCHE_AGE 2 vs 13	0.601	0.116	3.118
CD_TRANCHE_AGE 3 vs 13	0.708	0.133	3.759
CD_TRANCHE_AGE 4 vs 13	0.703	0.133	3.711
CD_TRANCHE_AGE 5 vs 13	0.546	0.103	2.894
CD_TRANCHE_AGE 6 vs 13	0.727	0.137	3.847
CD_TRANCHE_AGE 7 vs 13	0.725	0.137	3.829
CD_TRANCHE_AGE 8 vs 13	1.048	0.199	5.502
CD_TRANCHE_AGE 9 vs 13	1.214	0.231	6.364
CD_TRANCHE_AGE 10 vs 13	1.077	0.208	5.581
CD_TRANCHE_AGE 11 vs 13	1.071	0.207	5.551
CD_TRANCHE_AGE 12 vs 13	0.816	0.144	4.624

Score de concordance (échantillon d'apprentissage)

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	50.7	D de Somers	0.152
Pourcentage discordant	35.5	Gamma	0.176
Pourcentage lié	13.7	Tau-a	0.057
Paires	5732055	c	0.576

Score de concordance (échantillon de validation)

		Consommateurs prédits *NON* *Nombre d'assurés*	
		Nombre	*Fréquence*
Consommateurs réels	*Bien classé*		
NON	*OUI*	2 074	100
	Total	2 074	87
OUI	*Bien classé*		
	NON	322	100
	Total	322	13