Sorbonne Université
Institut des Statisitiques de l'Université de Paris

Mémoire d'Actuariat

En vue de l'admission à l'Institut des Actuaires

Élaboration de profils de risques des contrats standards en santé collectif

---

La Mutuelle Générale

Réalisé par :
Alexandra BONHOMME

Sous la direction de :
Céline RELMONT GIRAUD
Olivier LOPEZ

Année universitaire
2021/2022

Résumé

Mots clés : assurance collective, complémentaire santé, ratio de rentabilité, GLM, sensibilité.

Dans un contexte de grande compétitivité sur le marché des assurances collectives, un organisme de complémentaire se doit d'avoir une bonne connaissance de son portefeuille et des différents profils recherchés pour maintenir sa rentabilité.

Ce sujet est d'autant plus essentiel pour La Mutuelle Générale sur le périmètre des contrats collectifs standards à destination des TPE/PME que la rentabilité des offres de ces contrats s'est nettement détériorée ces dernières années. Sur le périmètre santé standard collectif hors Loi Evin en gestion directe et adhésion obligatoire, le ratio Prestations sur Cotisations (P/C) nets de chargements pour frais et de Taxes en 2021 affiche une forte dégradation :

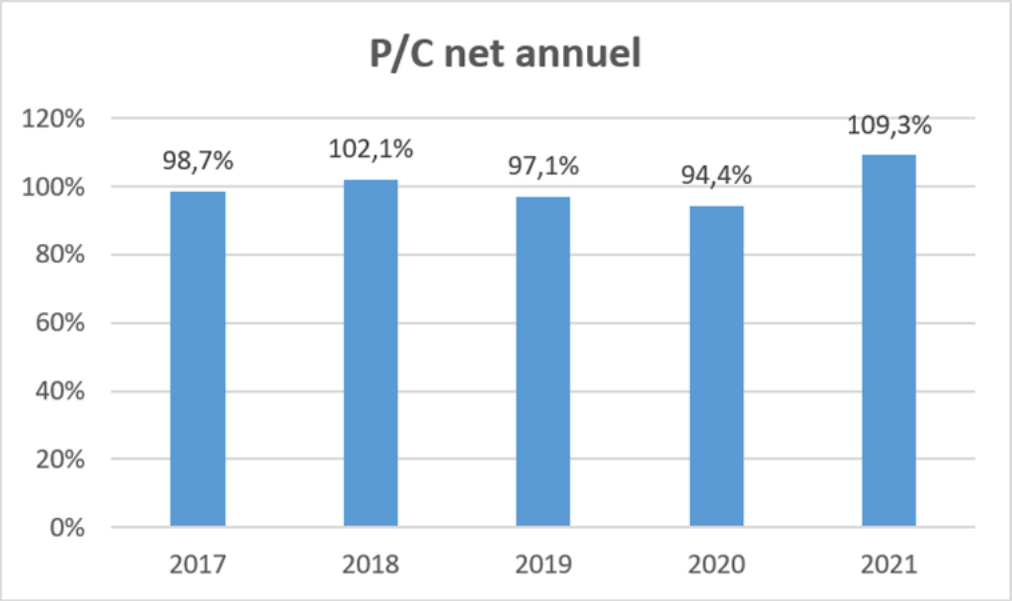


FIGURE 1 – Évolution du P/C annuel de LMG en santé collectif

La raison majeure du déficit de 2021 a été expliquée par une fraude commerciale découverte en cours d'année 2022 et une anticipation sous-estimée de la consommation du 100% Santé nouvellement mis en place. Cela dit, il reste primordial pour LMG de savoir identifier la typologie des contrats selon leur rentabilité, afin de cibler et affiner les variables tarifaires, pour obtenir un meilleur résultat global et également réaliser une meilleure prospection client.

L'objet du mémoire est de dégager des profils types de contrats qui se distinguent par leur niveau de rentabilité. Le périmètre de l'étude porte sur ce portefeuille de contrats standards collectifs en assurance complémentaire santé.

Quelles caractéristiques se rattachent donc aux contrats les plus et moins rentables ?

Dans un premier temps, ce mémoire exposera les différents défis qu'il a fallu relever sur la base de données pour fiabiliser et construire les variables nécessaires à la modélisation souhaitée de la rentabilité.
Il présentera une analyse descriptive fine de ces variables, afin de définir les modalités des caractéristiques principales et pertinentes des portefeuilles étudiés. Après ces premières étapes essentielles, une étude de corrélation sera réalisée pour sélectionner les variables les plus intéressantes à intégrer au modèle.
Ensuite, le choix d'un modèle GLM pour cette problématique d'élaboration de profils sera expliqué. Les données préalablement fiabilisées, seront exploitées pour lancer le GLM, qui modélisera l'indicateur du ratio de rentabilité P/C en fonction des paramètres liés aux contrats. Cette modélisation apportera une vision des profils des entreprises fonction de leur rentabilité.
Enfin, une analyse de sensibilité du modèle complétera cette dernière partie. Plusieurs scénarios permettront de tester la sensibilité et les limites du modèle.

Abstract

Key words : group health, complementary health, profitability ratio, GLM, sensitivity.

In a context of great competition in the group insurance market, a complementary healthcare insurer must be equipped with a thorough understanding of its portfolio and the characteristics of desirable profiles in order to maintain its profitability.

This subject is more than necessary for La Mutuelle Générale, which has seen a clear deterioration in its Benefits/Contributions (P/C) ratio in 2021 in the group health sector :

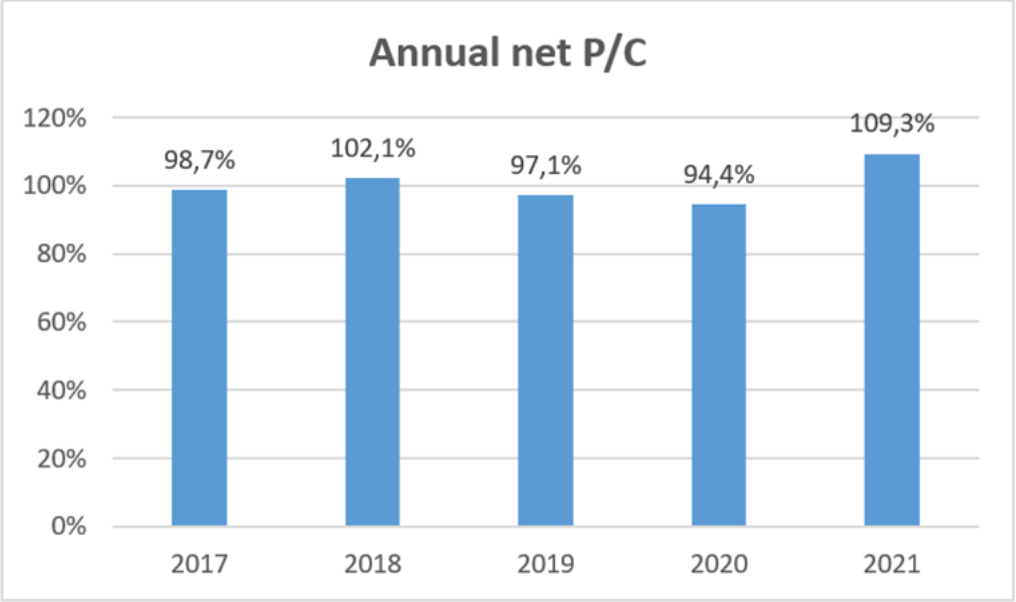


FIGURE 2 – Evolution of LMG's annual P/C in group health

If the reason for the 2021 deficit is mainly explained by a commercial fraud discovered during the year, it remains essential for LMG to know how to identify the profitable contracts in order to obtain a better global result and to carry out a better customer prospection.

The purpose of this report is to identify the typical profiles of group health contracts that are distinguished by their level of profitability : what characteristics are associated with the most and least profitable contracts ?

This thesis exposes the different challenges that had to be met on the database. It integrates a descriptive analysis of the variables that play a role in profitability in order to define the main characteristics of the portfolios of the studied perimeter. The information

collected will then be used to launch a GLM modeling the P/C indicator according to the variables linked to the contracts. This modeling will provide a vision on the profitability profiles of the companies according to the studied parameters. It will also be used to perform a sensitivity analysis of the variables.

Note de synthèse

Mots clés : assurance collective, complémentaire santé, ratio de rentabilité, GLM, sensibilité.

Contexte et objectif

Afin de faire face à la grande compétitivité du marché collectif, les organismes de complémentaires santé doivent maintenir une bonne connaissance de leur portefeuille. Identifier les différents profils de clients pour préserver la rentabilité devient alors une priorité. Notre étude a ainsi pour objectif de dégager une typologie des contrats selon leur rentabilité. Cette analyse est motivée par la volonté de mieux connaître les types de contrats déficitaires du portefeuille concerné, pour une meilleure prospection client par la suite et une éventuelle révision des variables tarifaires. Notre périmètre correspond aux contrats standards collectifs à adhésion obligatoire en gestion directe, et se constitue de 5 produits.

Démarche suivie

Pour satisfaire ces objectifs, nous étudions la rentabilité du portefeuille en mettant en place deux GLM, calibrés sur un historique de consommation couvrant la période 2017 – 2021. Notre variable réponse correspond au rapport prestations à cotisations de chaque contrat, noté P/C. Les variables dépendantes des GLM nous permettront de construire un profil de contrats déficitaires.

Données et fiabilisation

Les données de la base finale proviennent du retraitement et de la jointure de 5 sources différentes. Nous décrivons les contrats de notre base selon des données démographiques, des données de souscription et des données externes publiques.

Le niveau de couverture étant une variable essentielle à la caractérisation du contrat, nous fiabilisons la variable du niveau de garantie propre à chaque offre du périmètre, à l'aide d'une table de transcodification entre le code technique de garantie et le niveau de garantie. Nous harmonisons par la suite les niveaux de garanties entre les offres, en créant la variable niveau de gamme, indiquant si le contrat possède une couverture de type entrée de gamme, milieu de gamme ou haut de gamme.

Une fiabilisation des montants de prestations est également réalisée, basée sur une étude de boxplots des montants par module de soins et famille d'actes, afin d'identifier et corriger

les valeurs extrêmes de notre base.

Étude du P/C

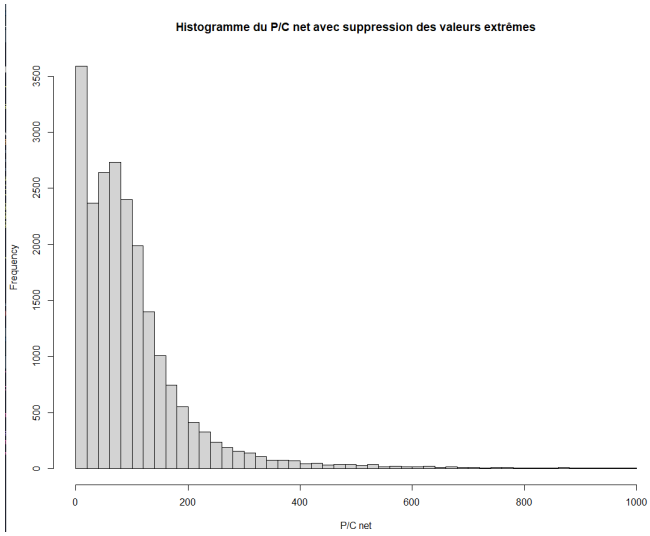


FIGURE 3 – Histogramme du P/C

L'histogramme du P/C met en lumière une partie des contrats sans prestations et donc de P/C nul. Il nous a paru intéressant de réaliser un GLM sur la base des P/C non nuls, et un second sur la base entière des P/C, et de sélectionner la meilleure modélisation pour dégager les profils de contrats selon le niveau de rentabilité.

Etude du profil de risque à l'aide d'un GLM Gamma

Le choix de la loi modélisant le P/C sur la base des P/C non nuls s'est réalisé par analyse graphiques et test d'ajustement de différentes lois candidates.

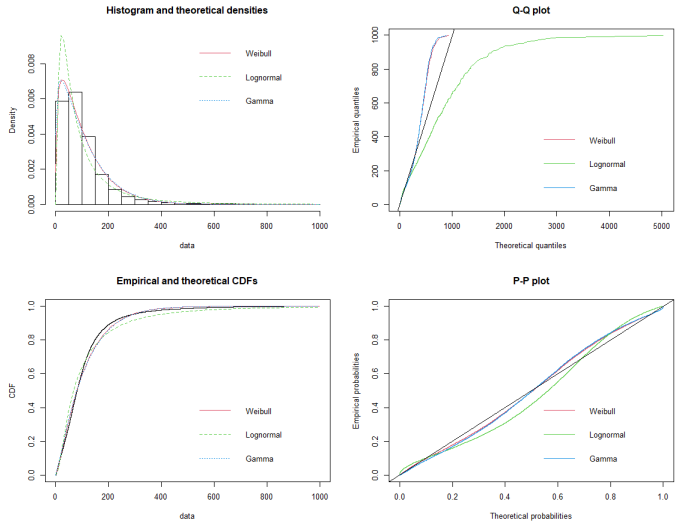
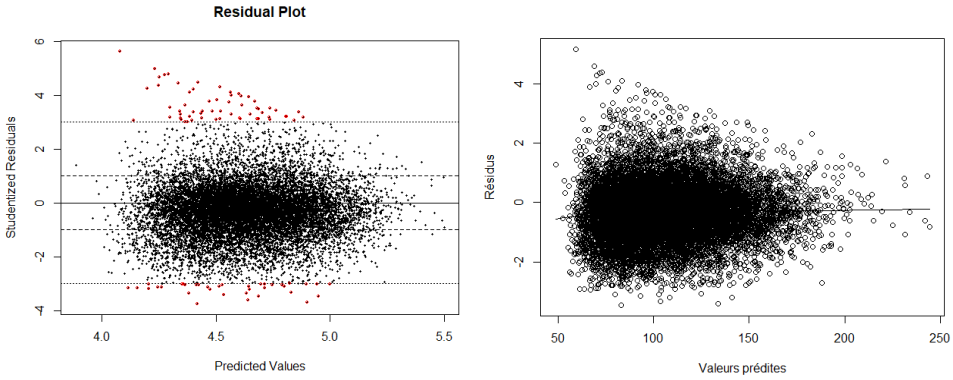


FIGURE 4 – Résultats des ajustements des lois

La sélection des variables pertinentes au sein du GLM s'est accompagnée d'une étude de liaisons entre les variables explicatives, ainsi que d'un test de sélection stepwise basé sur le BIC. Les variables retenues sont l'âge moyen des salariés, le nombre moyen d'enfant assuré par salarié, le niveau de gamme, l'ancienneté du contrat, la part de salariées femmes ainsi que la densité médicale du département de l'entreprise. Nous analysons les résidus de notre modélisation afin de discuter de la qualité d'ajustement, que nous jugeons convenable malgré la présence de résidus éloignés de l'axe nul.



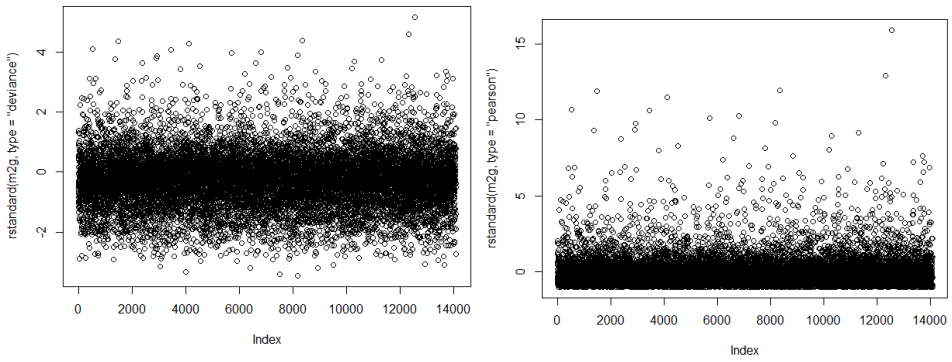


FIGURE 5 – Résidus du modèle Gamma

L'étude par la suite des coefficients des modalités des variables explicatives sélectionnées permet de dégager des conclusions quant à la pertinence de la tarification du portefeuille. Nous observons entre autres que les contrats ont un niveaux de rentabilité différents selon le niveau de gamme, ou bien selon l'âge moyen des salariés.

Étude du profil de risque à l'aide d'un GLM Tweedie

Les distributions de Tweedie font partie de la famille des lois exponentielles puisqu'il s'agit de lois de Poisson composées Gamma agrémentées d'une masse en zéro. L'étude de la fonction de log-vraisemblance de la densité de Tweedie permet de sélectionner le paramètre de puissance de la loi s'ajustant au mieux à la distribution du P/C.

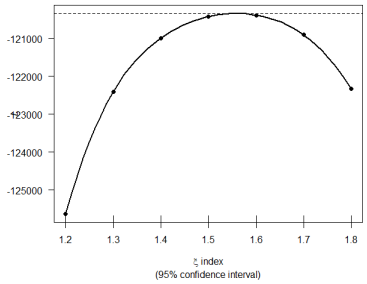


FIGURE 6 – Fonction de log-vraisemblance profil pour les P/C net des contrats

Nous combinons les tests de sélection de variables avec la connaissance du domaine afin de sélectionner les variables de notre nouvelle modélisation. Les variables retenues pour le GLM de Tweedie sont alors : l'âge moyen des salariés, l'âge moyen des conjoints, le collège, le nombre moyen d'enfant assuré par salarié, le niveau de gamme, l'ancienneté du contrat, la part de salariées femmes ainsi que la densité médicale du département de l'entreprise. Cependant, les résidus du modèle final Tweedie sont moins satisfaisants que ceux du GLM Gamma, dû à une présence de corrélation.

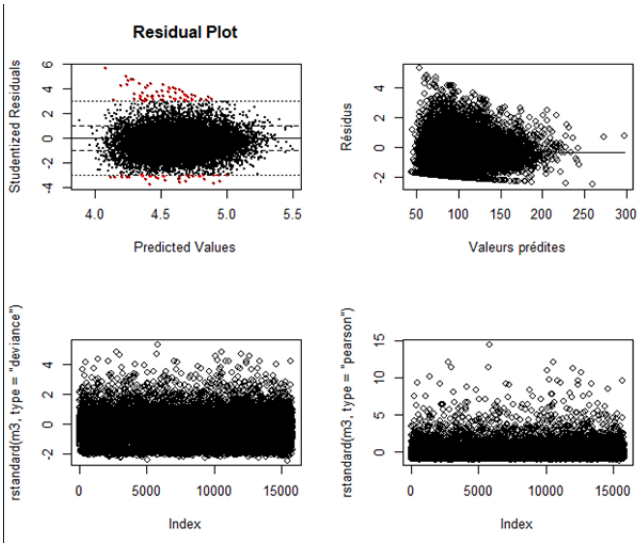


FIGURE 7 – Résidus du modèle Tweedie

Sensibilité

Finalement, nous jugeons une meilleure modélisation avec la loi Gamma et nous réalisons une étude de sensibilité basé sur ce modèle, en utilisant l'AIC comme indice de sensibilité. Le but est alors d'observer le niveau de contribution d'une variable explicative dans la modélisation, en comparant l'évolution de l'AIC d'un GLM sans cette variable par rapport à l'AIC du modèle scénario central. L'étude permet de conclure sur les relations existantes entre le P/C et différentes variables. Nous sommes arrivés à la conclusion que le profil de contrat le plus déficitaire concerne les contrats de type haut de gamme, et d'âge moyen de salarié élevé, constitué d'un grand nombre de salariées femmes, et d'enfants assurés par salarié, et où l'entreprise se situe dans une grande densité de médecins.

Sensibilité à la suppression d'une variable		
Variable supprimée	**AIC**	**Evolution par rapport au scénario central**
Age moyen des salariés	158145,55	246,9
Nombre moyen d'enfants par salarié	158139,86	241,21
Niveau de gamme	158088,32	189,67
Anciennete du contrat	158054	155,35
Proportion de salarié femme	157972,6	73,95
Densité médicale	157920,98	22,33

FIGURE 8 – Résultats de la sensibilité du modèle face à la suppression d'une variable

Conclusion

Le mémoire présente une démarche pour tenter d'élaborer un profil de risque des contrats standards collectifs. Les résultats peuvent être considérés comme convenables, mais ils

présentent plusieurs axes d'améliorations et de limites. Nous pouvons citer le taux de remise qui n'était pas assez fiable pour l'intégrer dans notre modélisation, mais aussi les résidus des modèles dont l'analyse mérite une amélioration.

Summary

Key words : group health, complementary health, profitability ratio, GLM, sensitivity.

In order to cope with the highly competitive group market, additional health insurance organizations must maintain a good knowledge of their portfolio. Identifying the different customer profiles to preserve profitability is therefore a priority. The objective of our study is to identify a typology of contracts according to their profitability. This analysis is motivated by the desire to gain a better understanding of the types of loss-making contracts in the portfolio concerned, in order to have a better client prospecting and a possible revision of pricing variables. Our scope corresponds to standard group contracts with compulsory membership under direct management, and consists of 5 products.

Action taken

To meet these objectives, we study the profitability of the portfolio by implementing two GLMs, calibrated on a consumption history covering the period 2017 - 2021. Our response variable is the benefit-to-contribution ratio of each contract, denoted P/C. The dependent variables of the GLMs will allow us to build a profile of loss-making contracts.

Data and reliability

The data in the final database comes from the reprocessing and joining of 5 different sources. We describe the contracts in our database according to demographic data, underwriting data and external public data.

As the level of coverage is an essential variable for the characterization of the contract, we improve its reliability for each offer of the perimeter, using a transcoding table between the technical code of coverage and the level of coverage. We then harmonize the levels of coverage between the offers, by creating the level of range variable, indicating whether the contract has entry-level, mid-range or high-end coverage.

A reliability check of the benefit amounts is also carried out, based on a boxplot study of the amounts per care module and medical procedure family, in order to identify and correct the extreme values of our database.

P/C study

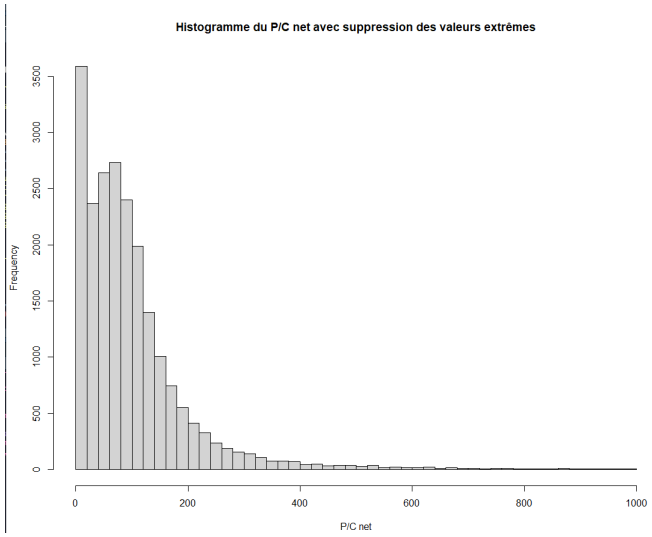


FIGURE 9 – Histogram of the P/C

The P/C histogram highlights a part of the contracts without benefits and therefore with a null P/C. It seemed interesting to us to carry out a GLM on the basis of the non-zero P/C, and a second one on the whole P/C basis, and to select the best modeling to identify the profiles of contracts according to the level of profitability.

Study of the risk profile using a GLM Gamma

The choice of the law modeling the P/C on the basis of non-zero P/C was made by graphical analysis and test fitting of different candidate laws.

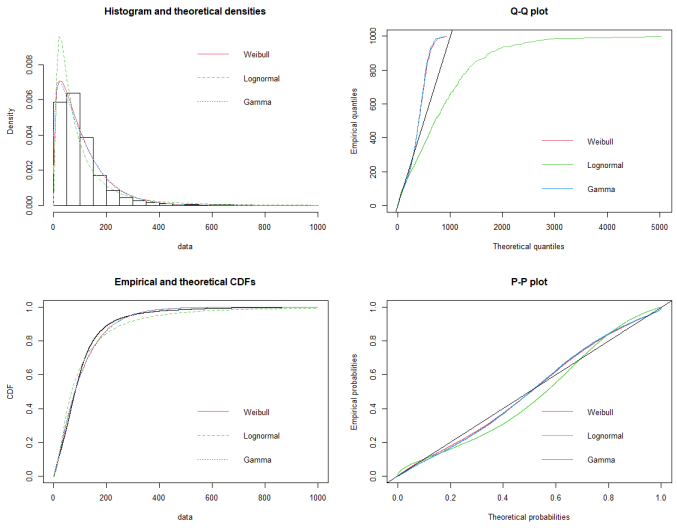


FIGURE 10 – Results of law adjustments

The selection of the relevant variables within the GLM was accompanied by a study of the links between the explanatory variables, as well as a stepwise selection test based on the BIC. The variables retained are the average age of employees, the average number of insured children per employee, the level of range, the length of the contract, the share of female employees and the medical density of the company's department. We analyze the residuals of our model to discuss the goodness of fit, which we consider to be adequate despite the presence of residuals far from the null axis.

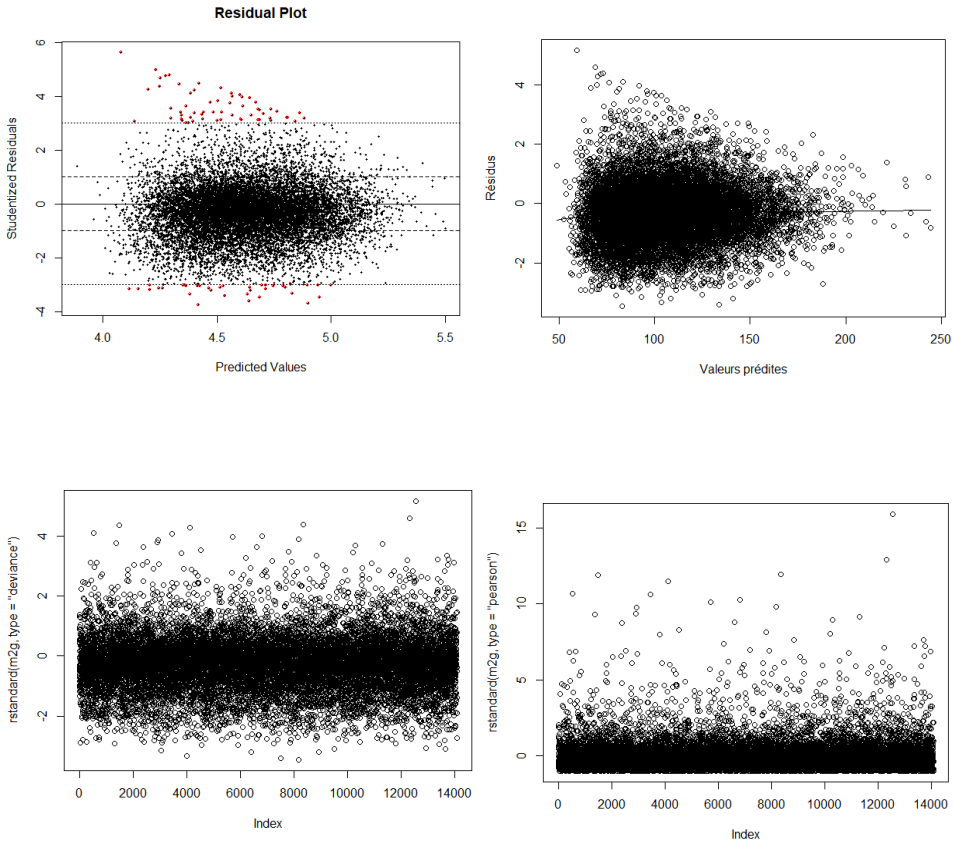


FIGURE 11 – Gamma model residuals

The subsequent study of the coefficients of the modalities of the selected explanatory variables allows us to draw conclusions about the relevance of the portfolio pricing. We observe, among other things, that the contracts have a different level of profitability according to the level of the range, or according to the average age of the employees.

Study of the risk profile using a Tweedie GLM

Tweedie distributions are part of the family of exponential laws since they are Poisson laws composed of Gamma with a mass in zero. The study of the log-likelihood function of the Tweedie density allows to select the power parameter of the distribution that best fits the P/C distribution.

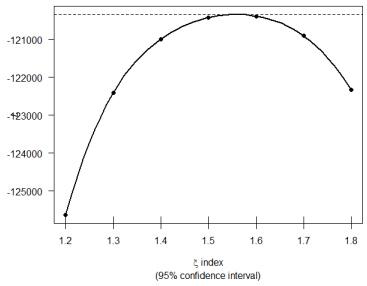


FIGURE 12 – Log likelihood profile function for net contract P/C

We combine variable selection tests with domain knowledge to select the variables for our new modeling. The variables selected for the Tweedie GLM are then : the average age of employees, the average age of spouses, the college, the average number of insured children per employee, the range level, the length of the contract, the share of female employees and the medical density of the company's department. However, the residuals of the Tweedie final model are less satisfactory than those of the GLM Gamma, due to the presence of correlation.

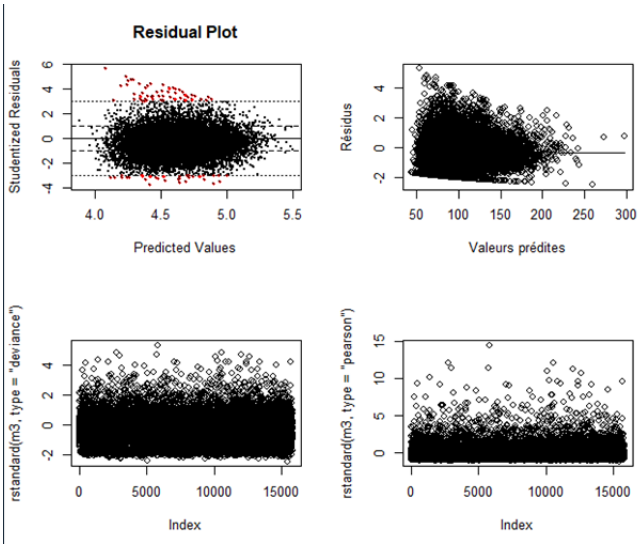


FIGURE 13 – Residue of Tweedie Model

Sensitivity

Finally, we judge a better modeling with the Gamma law and we realize a sensitivity study based on this model, using the AIC as sensitivity index. The aim is to observe the level of contribution of an explanatory variable in the model, by comparing the evolution of the AIC of a GLM without this variable with the AIC of the central scenario model. The study allows us to conclude on the existing relationships between the P/C

and different variables. We conclude that the most loss-making contract profile concerns high-end contracts, with a high average age of the employee, a large number of female employees, and insured children per employee, and where the company is located in a high density of doctors.

Sensibilité à la suppression d'une variable		
Variable supprimée	**AIC**	**Evolution par rapport au scénario central**
Age moyen des salariés	158145,55	246,9
Nombre moyen d'enfants par salarié	158139,86	241,21
Niveau de gamme	158088,32	189,67
Anciennete du contrat	158054	155,35
Proportion de salarié femme	157972,6	73,95
Densité médicale	157920,98	22,33

FIGURE 14 – Results of the sensitivity of the model to the suppression of a variable

Conclusion

This paper presents an approach to attempt to develop a risk profile of standard group contracts. The results can be considered as adequate, but they present several areas of improvement and limitations. We can mention the discount rate which was not reliable enough to be included in our modeling, but also the residuals of the models whose analysis deserves improvement. The thesis presents an approach to try to develop a risk profile of standard group contracts. The results can be considered as adequate, but there are several areas for improvement and limitations. We can mention the discount rate which was not reliable enough to be included in our modeling, but also the residuals of the models whose analysis deserves improvement.

Remerciements

Je souhaite adresser mes remerciements à toutes les personnes qui ont participé à la réalisation de ce mémoire.

Je remercie aussi Céline Relmont Giraud et Thu Huong Tran pour leur confiance et l'ensemble de leurs conseils prodigués. Je remercie grandement chacun des membres de l'équipe qui m'ont tous apporté leur soutien et leur aide à leur niveau.

Mes remerciements s'adressent aussi à mon tuteur académique Olivier Lopez.

Introduction

Le marché de l'Assurance Complémentaire Santé a connu des bouleversements ces dernières années, en raison de diverses décisions gouvernementales, l'évolution du contrat responsable et l'accord interprofessionnel (ANI) en 2016, la mise en place du 100% Santé en 2020, la taxe « covid ». Cela a impacté directement sur les résultats de cette branche d'assurance. Du point de vue de la Sécurité sociale, depuis de nombreuses années, les ressources se réduisent progressivement, produisant un recul régulier des remboursements de certains actes médicaux par celle-ci et un report de ces prises en charge sur les Organismes d'Assurance. En 2020, le système a en plus subi une dégradation abrupte et inédite des suites de la crise économique et sanitaire de la covid-19. En effet, la perte du régime général de la Sécurité sociale et du Fonds de solidarité vieillesse s'est établi à 38,6 milliards d'euros en 2020, soit une dépréciation de 36,6 milliards d'euros par rapport à 2019. En 2021, le déficit a diminué, bien qu'encore trop conséquent puisqu'il est ramené à 24,6 milliards d'euros. En somme, depuis 25 ans, la Sécurité sociale a été en excédent seulement sur 3 années, marquant la formation du « trou de la Sécu », qui correspond à l'accumulation sur les années du déficit de la Sécurité sociale.

Pour réduire ce déficit, le gouvernement a imposé aux organismes complémentaires santé une taxe « covid » en 2020, ou encore négocié des dispositifs de prise en charge d'actes tels que les consultations de psychologue en 2021 non remboursées alors par la Sécurité sociale. Le déremboursement de certains actes se reporte ainsi sur les complémentaires santé : des lois de finance de la Sécurité sociale imposent à ces organismes une prise en charge de soins de plus en plus importante, le dernier exemple étant la réforme 100% santé en 2021.

	Mutuelles	Sociétés d'assurances	Institutions de prévoyance	Parts de marché 2020 (en %)	Parts de marché 2011 (en %)
Top 10	4	3	3	41	29
Top 20	8	9	3	58	45
Top 50	21	20	9	78	69
Top 100	50	35	15	91	85

Note : Les plus grands organismes de 2020 ne sont pas nécessairement les mêmes que ceux de 2011.
Lecture : Les 10 plus grands organismes en matière de chiffre d'affaires en santé (4 mutuelles, 3 institutions de prévoyance et 3 sociétés d'assurances en 2020, soit 2,3 % de l'ensemble des organismes) représentent 41 % des cotisations collectées en 2020. En 2011, les 10 plus grands organismes concentraient 29 % des cotisations collectées.
Champ : Organismes assujettis à la taxe de solidarité additionnelle et contrôlés par l'ACPR au 31/12 de chaque année.
Source : ACPR.

FIGURE 15 – Parts de marchés des organismes de complémentaire santé

En 2019, la part prise en charge par les Organismes d'assurance des dépenses de Santé était de 13,5%. Avec la mise en place du 100% Santé et la reprise d'activité en 2021, cette part est passée de 12,2% en 2020 à 12,9% en 2021. Au niveau du marché des complémentaires santé, légèrement moins du quart des organismes représentaient 91% du marché en 2020 en termes de cotisations collectées.

Les contraintes règlementaires et la compétitivité impliquent une réduction progressive des marges au sein de ces structures. Par conséquent, la rentabilité des contrats demeure une problématique indéniable chez les organismes complémentaires de santé, en particulier sur le périmètre des contrats collectifs, où la concurrence rude amène parfois à des souscriptions de contrats structurellement déficitaires. C'est donc un point d'attention du Département Technique chez LMG.

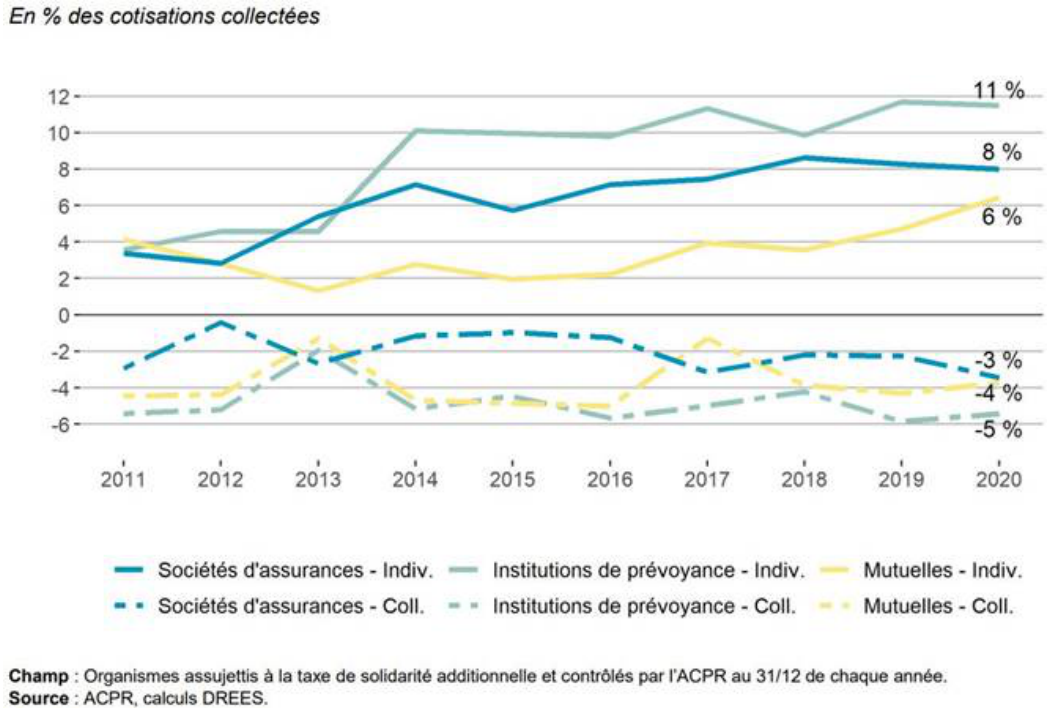


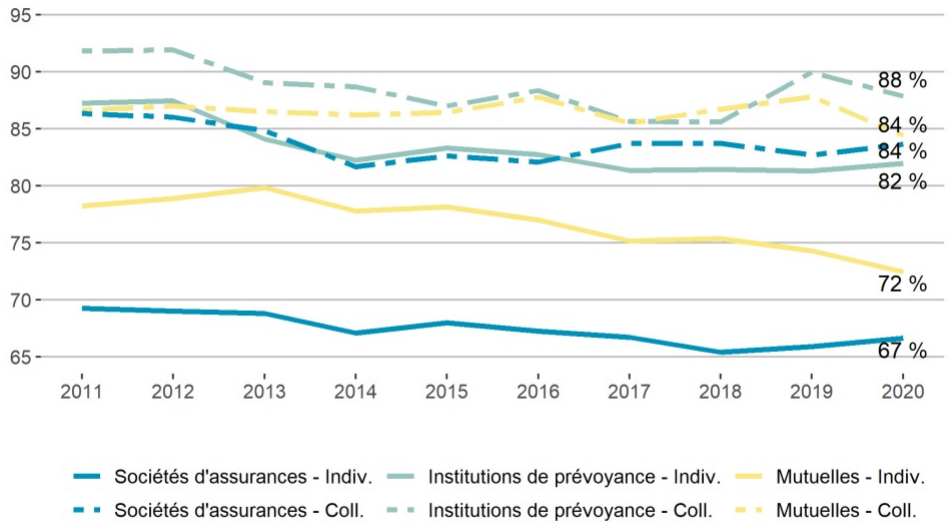
FIGURE 16 – Résultat technique en santé selon le type d'organisme et de contrats (source : DREES)

Le graphique ci-dessus permet notamment de mettre en lumière en 2016 l'impact du contexte législatif sur le résultat technique au sein des mutuelles collectives, suite à l'Accord National Interprofessionnel (ANI) qui impose à toutes les entreprises, même les TPE (Très Petites Entreprises) et PME (Petites et Moyennes Entreprises), de proposer une complémentaire santé obligatoire à leurs salariés. Ce décret entraîne un transfert des cotisations de l'assurance individuelle (principalement des mutuelles), caractérisé comme un marché rentable, vers celui de l'assurance collective, qui lui est bien moins rentable.

Par ailleurs, bien que les organismes complémentaires aient par le passé dégagé des excédents sur leur activité en santé, nous pouvons observer un déficit inhérent au marché de la santé collective, ces organismes cherchent désormais à mieux maitriser leurs résultats pour éviter qu'ils ne se dégradent plus. C'est pourquoi il demeure important d'analyser la rentabilité des contrats collectifs.

Graphique 3.5 – Prestations sur cotisations par type d'organisme et de contrat entre 2011 et 2020

En % des cotisations



Lecture : En 2020, les contrats individuels des mutuelles reversent aux assurés 72 % de leurs cotisations sous forme de prestations.
Champ : Organismes assujettis à la taxe de solidarité additionnelle et contrôlés par l'ACPR au 31/12 de chaque année.
Source : ACPR, calculs DREES.

FIGURE 17 – Evolution du P/C par type d'organismes entre 2011 et 2020

Le rapport de la DREES (Direction de la Recherche, des Études, de l'Évaluation et des Statistiques) sur la situation financière des organismes complémentaire santé met en lumière un ratio Prestations/Cotisations des contrats collectifs des mutuelles globalement plus élevé que celui des sociétés d'assurances, et moins élevé que celui des institutions de prévoyance. Ces résultats sont expliqués en partie par le poids du collectif, plus élevé pour les institutions de prévoyance et dans une moindre mesure pour les sociétés d'assurances, sans oublier la spécialisation de certaines mutuelles et institutions de prévoyance sur des professions ou des branches professionnelles. En 2020, le P/C des contrats collectifs des mutuelles, historiquement plus élevé que celui des sociétés d'assurances, rejoint ce dernier à 84%. Nous remarquons également que la baisse du ratio pour les mutuelles et la hausse pour les sociétés d'assurances en 2020, année particulière, se retrouvent autant dans les contrats collectifs que dans les contrats individuels.

Dans ce mémoire, nous nous concentrerons sur la rentabilité des contrats complémentaires Santé à destination de TPE et PME. L'objectif de ce mémoire est d'analyser la typologie des entreprises les moins rentables qui sont souscriptrices de ces gammes de produits. Ainsi, il pourrait être ensuite possible d'exploiter ces résultats afin de :

— ne plus proposer un certain type de contrats ;
— ne plus cibler un certain type d'entreprises ;
— essayer d'ajuster les tarifs des mauvais risques.

Pour ce faire, nous décrirons dans la Partie I le système de soin en France pour une bonne compréhension de l'assurance complémentaire Santé, puis le pré-traitement réalisé

sur les données à exploiter, et enfin les statistiques descriptives des différentes variables qui serviront à la modélisation de la Partie II.

L'étape du pré-traitement des données est un passage essentiel dans l'étude, pour une bonne utilisation des variables exploitées dans es analyses et les modèles, et pour retranscrire une interprétation cohérente des résultats des modèles. La partie II est consacrée à la modélisation des indicateurs de rentabilité par Entreprise, en particulier pour les plus déficitaires, afin de déterminer des caractéristiques exploitables selon leur niveau de rentabilité.

Table des matières

Première partie

Système de santé en France

Chapitre 1

La Sécurité Sociale

1.1 Régime obligatoire

Fondée en 1945 des suites de la Seconde Guerre mondiale, la Sécurité sociale (SS) est un ensemble d'Institutions avec une mission de Service Public. Elle a pour vocation de couvrir les individus actifs des conséquences de risques sociaux qui peuvent survenir tout au long de la vie. Le système est destiné à assister financièrement ses bénéficiaires qui rencontrent différents événements tels que la maladie, la maternité, l'invalidité, le décès, ou encore la vieillesse, le handicap, l'accès au logement.

Parmi les différentes institutions de la Sécurité sociale, les Caisses de Sécurité sociale sont des services de proximité qui prennent en charge le versement des prestations auprès de la population. Cette population d'actifs est rattachée à un des régimes de la Sécurité sociale, selon le secteur professionnel concerné. Chaque régime décrit l'ensemble des droits et obligations réciproques des Employés (et leurs « ayant-droit »), et des Patrons :

— Le Régime Général représente plus de 80% de la population. Il concerne les salariés et les travailleurs assimilés à des salariés, et a été étendu aux travailleurs indépendants depuis le 1er janvier 2018.
— Le Régime Local Alsace-Moselle s'adresse aux salariés des entreprises qui sont installées (ou ont leur siège) dans les départements du Bas-Rhin, du Haut-Rhin et de la Moselle. Il résulte du système de la Sécurité Sociale Allemande maintenue en Alsace-Moselle après la Seconde Guerre Mondiale, et propose des remboursements plus avantageux que le Régime Général ;
— L'ancien Régime Social des Indépendants (RSI) concernait les artisans, les commerçants et les professionnels libéraux, celui-ci est désormais fusionné au Régime Général ;
— Le Régime Agricole s'applique aux exploitants, salariés agricoles et d'autres secteurs rattachés à l'agriculture. Il est géré par la Caisse Centrale de la Mutualité Sociale Agricole (MSA) ;
— D'autres régimes spéciaux existent, comme le régime des marins, des militaires, de la SNCF, du Sénat, ...

6 branches autonomes composent actuellement le Régime Général, chacune responsable de ses ressources et dépenses :

— Branche maladie, gérée par la Caisse Nationale d'Assurance Maladie (CNAM) et son réseau qui se constitue entre autres des Caisses Primaires d'Assurance Maladie (CPAM), des Caisses Générales de Sécurité Sociale (CGSS) et des Directions Régionales du service Médical (DRSM) ;
— Branche accidents du travail et maladies professionnelles, gérée également par la CNAM ;
— Branche famille, gérée par la Caisse des Allocations Familiales (CAF) ;
— Branche retraite, gérée par l'Assurance retraite ;
— Branche recouvrement, gérée par l'Union de Recouvrement des Cotisations de Sécurité Sociale et d'Allocations Familiales (URSSAF) ;
— Branche autonomie, mise en place au 1er janvier 2021 et gérée par la Caisse nationale de Solidarité pour l'Autonomie (CNSA).

En France, le système de santé se rattache à la branche maladie de la Sécurité sociale.

Nous nous intéressons donc pour notre part à la branche maladie, appliquée au domaine de la santé, i.e. aux prestations correspondant à des remboursements de frais de santé comme la médecine, les soins pharmaceutiques, dentaire et optique ou encore l'hospitalisation. Nous verrons par la suite les mécanismes de remboursement de ces frais de santé, dans le cadre du régime obligatoire de la SS, et du régime complémentaire.

1.2 Mécanismes de remboursements en santé

Cette section se consacre au fonctionnement des remboursements relevant du régime de la Sécurité sociale, qui correspond au régime obligatoire, et des remboursements relevant du régime complémentaire qui s'appliquent dans le cadre d'assurance complémentaire santé.

Le montant de prise en charge des actes remboursés par la Sécurité Sociale (SS) d'une Complémentaire Santé est dépendant du montant de remboursement même de la SS.

1.2.1 Les nomenclatures des prestations de santé

Les nomenclatures définissent les actes, produits et prestations qui sont pris en charge totalement ou partiellement par l'assurance maladie obligatoire et les conditions de leur remboursement. Un code est défini pour chaque soin, chaque acte médical ou paramédical. À ce code correspond une BRSS et un taux de remboursement de la SS.

Par exemple, la nomenclature utilisée pour les actes dentaires est nommée Classification Commune des Actes Médicaux (CCAM). La CCAM est la codification de tous les actes techniques médicaux et bucco-dentaires et les gestes paramédicaux qui déterminent leurs tarifications ainsi que les majorations éventuelles (Exemple : ATC – actes chirurgie, ADA – anesthésie, ATM – actes techniques médicaux, ADI – imagerie).

La classification commune des actes médicaux (CCAM) répertorie donc par « code acte » l'ensemble des prestations médicales remboursées par le régime obligatoire. Pour chaque

code acte de soin, une base et un taux de remboursement sont définis par la SS, en accord avec les syndicats de praticiens.

1.2.2 La base de remboursement de la SS (BRSS)

En santé, chaque acte médical génère un coût appelé la dépense effective. La base de remboursement de la SS (BRSS) correspond au tarif maximal qui peut être prise en charge par la SS ; elle sert de base de calcul aux remboursements du régime obligatoire, et des organismes de complémentaires santé depuis 2006. Elle est fixée par convention entre les professionnels de santé et la Caisse nationale d'assurance maladie. La BRSS sert également de référence pour les organismes d'assurance complémentaire santé pour leurs remboursements.

La BRSS est composée de :

- De la part remboursée par le Régime Obligatoire (RO, la sécurité sociale)
- De la Participation Forfaitaire ou de la Franchise Médicale Ces montants sont une déduction des remboursements du RO à la charge de l'assuré. Ces montants ne sont pas couverts dans le cadre d'un contrat complémentaire Santé (sous peine de taxations supplémentaires pour les Entreprises et les assurés). En soins dentaires et en optique, il n'y a ni Participation Forfaitaire, ni Franchise Médicale.
- Du Ticket Modérateur (TM) : il s'agit de la différence entre la BRSS et le montant remboursé par le régime obligatoire (régime de la Sécurité sociale). Ce TM est à la charge de l'assuré s'il ne bénéficie pas d'une complémentaire santé.

$$TM = (1 - Taux\ de\ Remboursement\ de\ la\ SS) * BRSS$$

Le taux de remboursement du régime obligatoire, appelé taux RO, est appliqué à la BRSS pour définir le tarif qui sera pris en charge par la SS, que l'on nomme montant RO :

$$Montant\ RO = BRSS * Taux\ RO$$

1.2.3 Le dépassement d'honoraires

Si le prix pratiqué pour un acte médical est supérieur à la BRSS, il y a un dépassement d'honoraires égal au montant excédant la BRSS. Ces dépassements d'honoraires restent à la charge de l'assuré, en plus du TM. Les dépassements d'honoraires peuvent être pris en charge par la complémentaire santé.

Dans le cas de dépassements d'honoraires pratiqués par certains praticiens, le remboursement de la SS correspond toujours au montant RO. Le TM et le dépassement constituent la base du remboursement relevant de la complémentaire santé, à supposer que le patient est bénéficiaire d'un régime de complémentaire Santé. Le complément reste à la charge du patient.

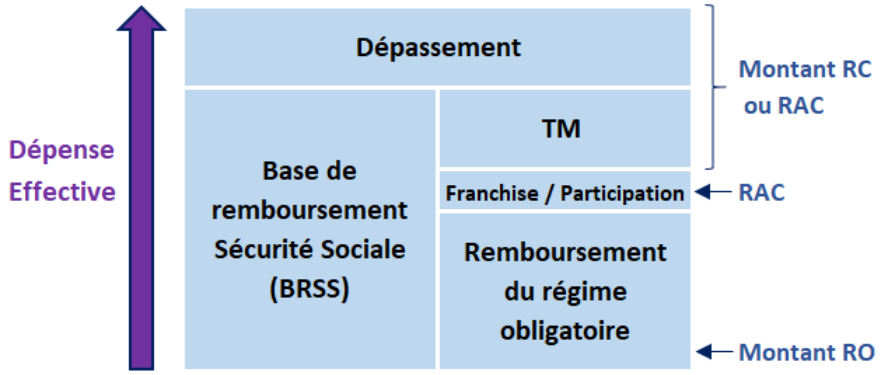


FIGURE 1.1 – Schéma de la dépense effective d'un soin

1.2.4 La dépense effective d'un acte de soin

La dépense effective correspond au montant réel dépensé par l'assuré pour un soin. C'est ainsi, pour un acte de soin, la somme du Remboursement de la SS + Franchise ou Participation forfaitaire (éventuelle) + TM + des éventuels Dépassements d'honoraires.

En règle générale, la SS rembourse les prestations des médecins conventionnés sur la base d'un tarif dit de convention. Pour les médecins qui pratiquent un tarif libre et n'adhèrent pas à la convention médicale, la SS rembourse les prestations sur la base d'un tarif d'autorité.

Chapitre 2

Régime complémentaire

La complémentaire santé est un contrat privé d'assurance, proposé par 4 organismes en France, chacun étant régulé par l'Autorité de Contrôle Prudentiel et de Résolution (ACPR), et réglementé par des Codes différents :

- Les mutuelles, régies par le Code de la mutualité ;
- Les sociétés d'assurance, régies par le Code des assurances ;
- Les institutions de prévoyance, régies par le Code de la Sécurité sociale ;
- Les mutuelles et institutions de prévoyance agricoles, régies par le Code rural et de la pêche maritime depuis 2010.

Contrairement aux sociétés d'assurances qui ont un objectif de bénéfice financier et dont les membres de la structure de gouvernance sont élus par l'assemblée générale des actionnaires, les mutuelles sont à but non lucratif, sans capital social, et la gouvernance est élue par les sociétaires. Par ailleurs, les mutuelles appliquent le principe de non-discrimination auprès de ses adhérents, et leur soumettent des cotisations qui sont indépendantes de leurs risques individuels. Ainsi, ces organismes n'ont pas le droit de sélectionner leurs membres selon leur état de santé, contrairement aux autres organismes.

Les institutions de prévoyance sont des sociétés de personnes à but non lucratif, gérant uniquement des contrats collectifs pour des salariés du secteur privé.

Les institutions de prévoyance et mutuelles agricoles sont spécialisées quant à elles dans les régimes complémentaires destinés aux bénéficiaires du régime agricole.

2.1 Classification des 6 modules de soins

Depuis 2019, les organismes assureurs ont signé un engagement pour la lisibilité des garan des contrats de complémentaire santé. Cet accord permet notamment d'harmoniser les noms de famille de soins et la diffusion d'exemples de calcul de remboursement. Cela a permis de créer des modules de soins communs aux organismes :

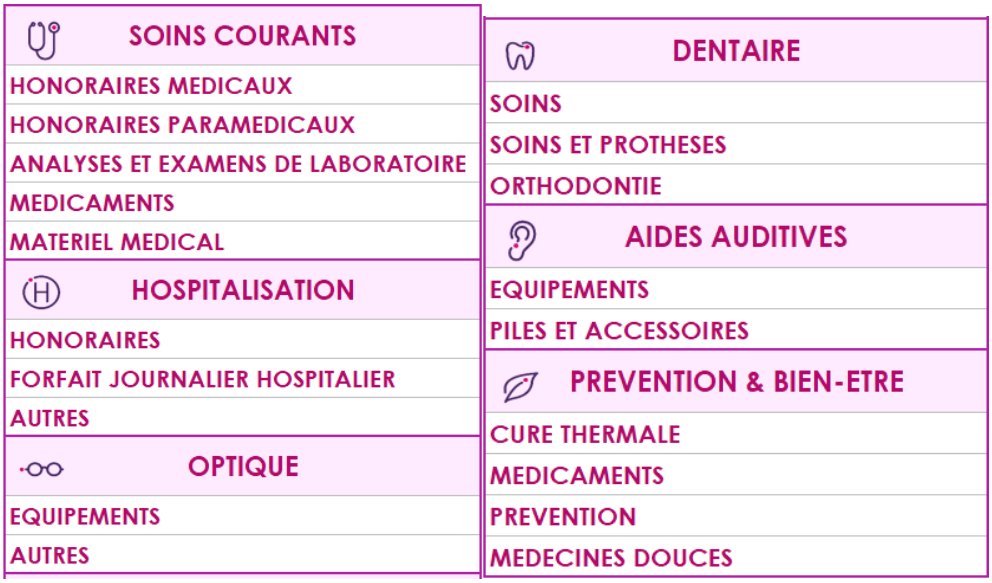


FIGURE 2.1 – Tableau des modules de soins LMG et leurs principaux actes

Nous avons :

1. Soins courants : ce module comprend les consultations, les soins de ville et les dépenses pharmaceutiques ;
2. Hospitalisation : ce module concerne les séjours à l'hôpital et les opérations chirurgicaux de santé ;
3. Optique : ce module est relatif aux lunettes, lentilles et chirurgie réfractive ;
4. Dentaire : ce module concerne les consultations et les appareils dentaires ;
5. Aides auditives : ce module est relatif aux audio prothèses et appareils auditifs.
6. Prévention & Bien être : ce module mis en place par LMG en plus des 5 autres, recouvre les cures thermales, des médicaments non pris en charge par la Sécurité sociale, les actes de prévention, ainsi que les médecines douces.

2.2 Niveaux de garantie

Les organismes de complémentaire santé proposent différents niveaux de garantie, afin de répondre au mieux aux besoins des assurés et d'offrir un tarif adapté selon leur niveau de consommation. Nous pouvons regrouper les niveaux de garantie de la façon suivant :

— Entrée de gamme : limite les remboursements au ticket modérateur, ne prend pas en charge les dépassements d'honoraires ;

— Milieu de gamme : prend en charge une partie des dépassements d'honoraires, ainsi que la plupart des dépenses peu remboursées par la Sécurité sociale, comme l'optique et le dentaire ;

— Haut de gamme : prend en charge une grande partie des dépassements d'honoraires, rembourse la quasi-totalité des prestations ;

2.3 Caractéristiques des contrats

L'organisme de complémentaire santé propose 2 types de contrats : le contrat individuel et le contrat collectif.

D'une part, le contrat individuel est un contrat passé entre 2 intervenants : l'organisme assureur et le souscripteur. L'assuré choisit ses niveaux de garanties (et possède la liberté de les adapter à ses besoins). Ces contrats peuvent concerner les étudiants, travailleurs non assimilés à des salariés, les chômeurs, mais aussi les salariés du privé souhaitant compléter leur contrat collectif avec une surcomplémentaire individuelle.

D'autre part, le contrat collectif fait intervenir un organisme assureur et une entreprise, association ou toute autre organisation professionnelle. L'adhésion est dite obligatoire lorsque le contrat s'impose à tout le personnel employé par l'entreprise. L'adhésion est facultative lorsque celle-ci est laissée au choix des salariés qui souhaitent adhérer à une surcomplémentaire proposée par l'employeur. De plus, en complément du socle de garanties prévues à titre obligatoire, le contrat collectif peut prévoir la souscription d'options. Celles-ci sont facultatives et destinées à compléter les garanties de base.

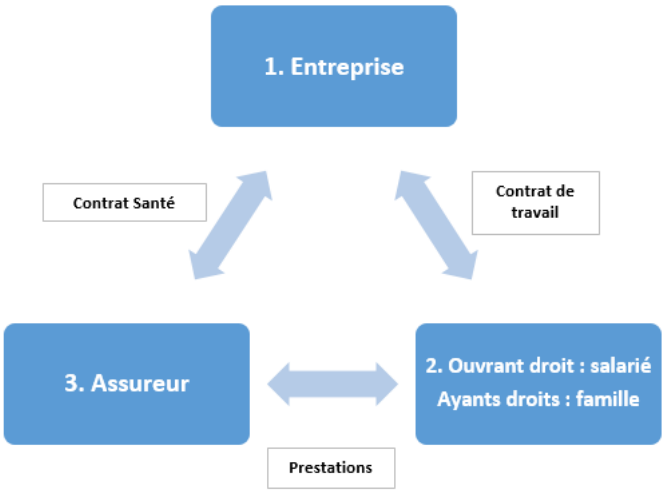


FIGURE 2.2 – Contrat Santé Collective obligatoire

Plus particulièrement, il existe deux types de contrats collectifs selon l'effectif de l'entreprise. Le contrat standard est à destination des entreprises de moins de 100 salariés et propose des garanties qui conviennent aux petites et moyennes entreprises, tandis que le contrat sur-mesure offre une couverture plus ciblée, puisqu'elles concernent les entreprises de plus de 100 salariés.

Nous pouvons également distinguer le type de gestion des contrats : les organismes d'aspeuvent procéder à une gestion interne dite directe de leurs contrats, ou bien confier tout ou une partie de cette gestion à un tiers externe, qui correspond alors à une délégation de gestion.

2.4 Exemple de remboursement

Le régime complémentaire vient en complément du régime obligatoire, et permet d'obtenir le remboursement de tout ou partie des frais de soins non pris en charge par les organismes de SS, dans le but de réduire le reste à charge de l'assuré.

L'Organisme d'Assurance Complémentaire reçoit des flux informatiques de la nomenclature SS. Ces flux sont réceptionnés dans des bases de données. Ces flux précisent notamment la nomenclature des actes remboursés par la SS avec la précision du montant remboursé, de la BRSS, de la dépense effective, l'assuré concerné. À partir de ces lignes de décompte par acte de la SS, l'Organisme d'Assurance Complémentaire imbrique dans son système de gestion des formules de calcul afin de réaliser la liquidation des actes selon le niveau de remboursement indiqué sur le régime de santé concerné (contrat) de l'individu qui reçoit le soin.

Nous présentons ci-dessous un exemple chiffré du mécanisme de remboursement d'une monture optique

- Dépense effective : 160€
- BRRS = 2,84€
- Taux RO = 60%
- Niveau de remboursement de la complémentaire santé = 150€
- Le montant remboursé par la Sécurité Sociale est de 2,84€*60% = 1,70€
- Le reste à charge assuré est de 160€ - 150€ - 1,70€ = 8,30€

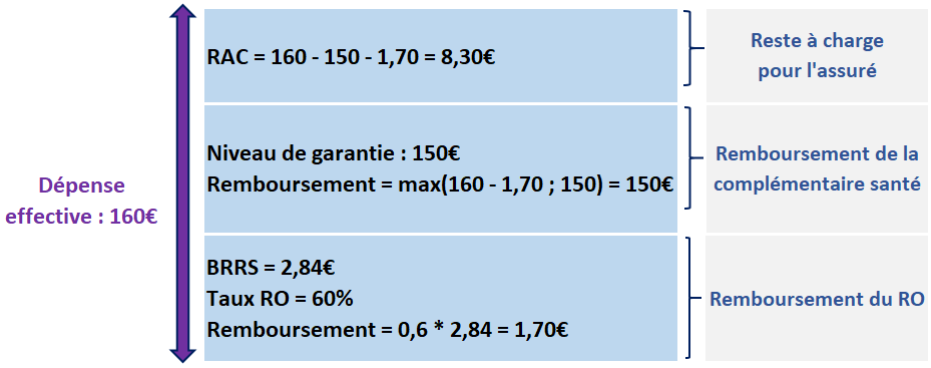


FIGURE 2.3 – Exemple de liquidation d'une monture optique

Pour les soins qui ne sont pas intégralement remboursés par le régime obligatoire, la définition du montant de remboursement du régime complémentaire (montant RC) dépend

de l'expression de garantie figurant sur le contrat. Ces remboursements peuvent s'exprimer en pourcentage de la BRSS comme dans l'exemple ci-dessus (60% du BRRS), ou encore en pourcentage du Plafond Mensuel de la Sécurité Sociale (PMSS).

L'expression du remboursement peut également être sous la forme d'un montant en euros. Ce mécanisme dit de « forfait », définit un montant maximum que l'assureur s'engage à rembourser à l'assuré.

Plus globalement, les dépenses du régime complémentaire santé sont influencées par certains facteurs. Parmi eux, nous pouvons citer les niveaux de garanties, le type de contrat, et les bénéficiaires couverts.

2.5 Réglementations

Cette section a pour objectif de présenter les réformes impactant l'assurance santé au fil des années.

2.5.1 Contrats responsables

Mis en place en 2016, les contrats responsables incitent l'assuré à adopter un comportement responsable dans sa consommation de soins en encadrant les dépenses de santé. Des règles sont ainsi fixées par décret concernant les niveaux de prise en charge, de plafonds et de planchers de certains frais médicaux. Parmi ces consignes, nous pouvons citer :

— Désigner un médecin traitant et le consulter en priorité ;
— Respecter le parcours de soins coordonné ;
— Éviter de consulter le professionnel pratiquant de forts dépassements d'honoraires, ...

Le respect de l'ensemble des critères permet notamment à l'assuré de ne pas avancer la part prise en charge par l'assurance maladie lors du règlement des frais médicaux : il s'agit du droit de tiers payant.

En optique par exemple, le remboursement appliqué dans le contrat responsable se limite à une paire de lunettes tous les deux ans (sauf pour les mineurs ou en cas d'évolution de la vue), et se comprend entre 420€ et 800€ pour les verres et 100€ pour les montures.

Les contrats responsables s'opposent à ceux dits « non responsables » : les garanties sont plus avantageuses et les plafonds de remboursements des contrats responsables sont dépassés.

Taxes

Des pénalités fiscales et sociales sont mises en place pour les assureurs et les entreprises afin d'inciter les acteurs de l'assurance à respecter les contrats responsables. Pour les

organismes complémentaires, une surtaxe (TSA) de 20,27% du montant des cotisations est appliquée aux contrats non responsables , contre 13,27% pour les contrats responsables.

2.5.2 Réforme 100% Santé

Dans cette partie, il nous a semblé important de présenter les point essentiels du 100% Santé, étant donné qu'il s'agit d'une réforme majeure qui a impacté dès 2020 les modules de soins Optique, Dentaire et Audio, ces derniers représentant un poids non négligeable de nos prestations. La réforme du 100% Santé est le fruit d'une promesse de campagne de mettre à disposition de tous une offre de soins dentaires, d'équipements optiques et de prothèses auditives sans reste à charge pour l'assuré. Le point commun entre ces trois postes de soins est un reste à charge important pour les assurés. Ceci est principalement dû au fait que la prise en charge de la sécurité sociale est historiquement moins élevée pour ces soins et que les prix de ces actes fixés librement par les praticiens sont déconnectés des montants de remboursement de la sécurité sociale. La conséquence est un taux de renoncement aux soins important pour ces trois postes.

Cette réforme vise ainsi à diminuer le renoncement aux soins dus aux coûts trop élevés des restes à charge dans le domaine de l'optique, du dentaire et de l'audiologie.Les principales mesures de cette réforme sont les suivantes :

- Création de prix limites de ventes (PLV) pour les actes du 100% Santé ;
- Nouvelles contraintes sur la prise en charge du remboursement complémentaire (RC) ;
- Augmentation de la prise en charge RO dans les trois paniers 100% Santé ;
- Évolution de la nomenclature CCAM de la SS.

Les organismes d'assurance complémentaire ont ainsi revu leurs garanties pour prendre en compte ces évolutions. La principale nouveauté apportée par la réforme 100% Santé est la mise en place de plusieurs paniers de soins présentant chacun des règles de facturation et de remboursement qui leur sont propres avec le panier RAC0 comme minimum à prendre en charge.

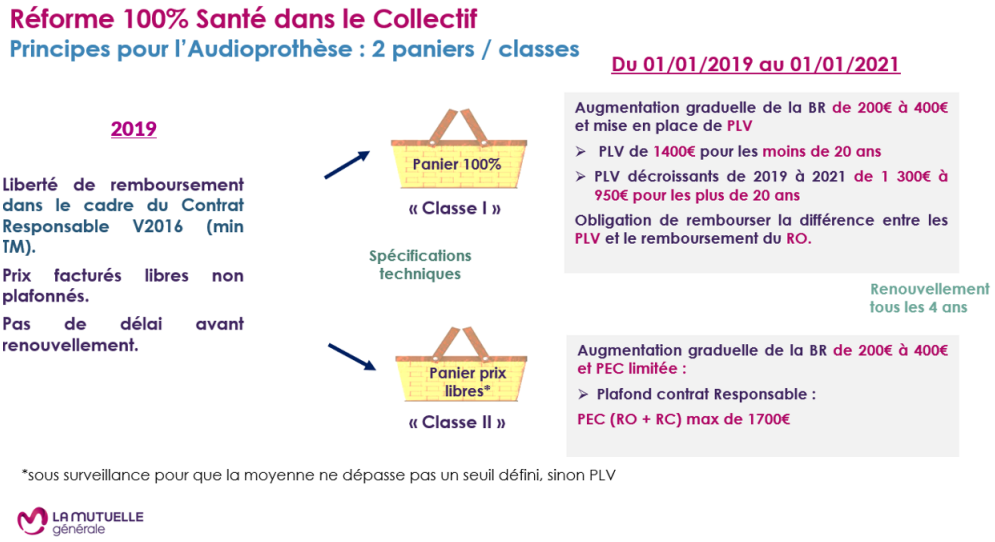


FIGURE 2.4 – Système de paniers pour le dentaire

En dentaire, trois paniers sont mis en place : le panier « RAC0 », le panier « Maîtrisé » et le panier « Libre ».

Le panier RAC0 présente des prix limites de vente (PLV) pour les praticiens et les complémentaires ont l'obligation de prendre en charge la différence entre la dépense réelle et le remboursement du régime obligatoire.

Le panier Maîtrisé n'a comme contraintes que certains PVL mais sans obligation de zéro reste à charge pour l'assuré.

Le panier Libre ne présente ni de PLV, ni d'obligation de rembourser la totalité du montant à la charge de l'assuré pour les actes de ce panier.

Les paniers Maitrisé et Libre ont toutefois comme contrainte de respecter les prises en charge minimums réglementaires du Contrat responsable et de l'Accord National Interprofessionnel (ANI) qui oblige les Entreprises à proposer un régime de soin à ses salariés respectant des niveaux de couverture minimums.

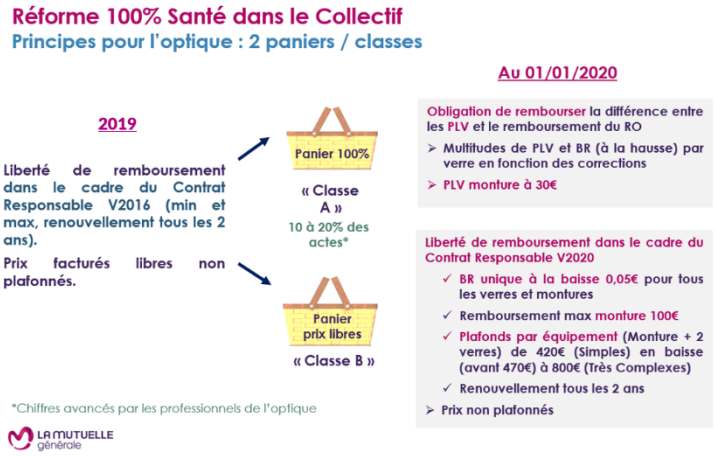


FIGURE 2.5 – Système de paniers pour l'optique

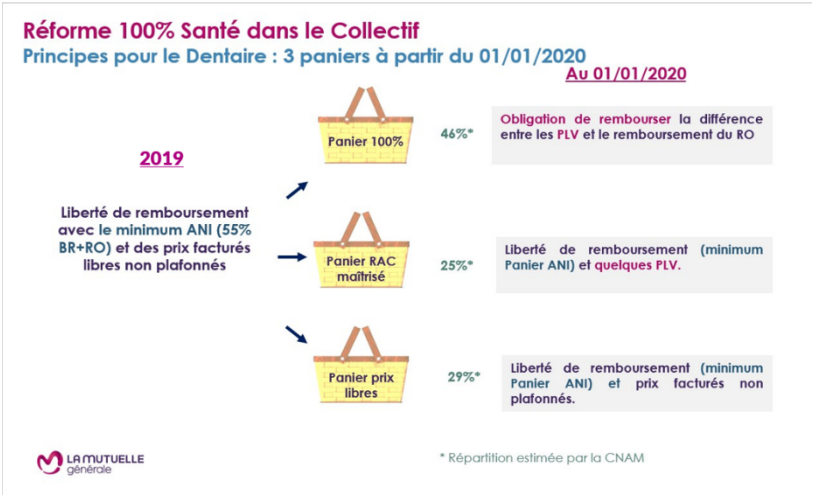


FIGURE 2.6 – Système de paniers pour l'audio

Pour les postes optique et audio, deux paniers sont proposés avec des règles similaires aux paniers RAC0 et Libre du dentaire. Les paniers RAC0 sont respectivement nommés « Panier A » et « Panier Classe I » pour l'optique et l'audio. Les paniers libres sont respectivement nommés « Panier B » et « Panier Classe II » pour l'optique et l'audio.

Pour ces trois postes, un planning d'évolution des Bases de Remboursement de la Sécurité Sociale (BRSS) et des PLV pour chaque acte entre 2019 et 2023 est disponible. Le remboursement complémentaire de l'assurance évolue donc en fonction de l'année.

2.5.3 Autres réglementations

Forfait patientèle

Le forfait patientèle désigne une rémunération forfaitaire versée au médecin traitant pour le suivi de ses patients. Ce forfait est mis en place depuis 2018 et les assurances complé-

mentaires le co-financent à hauteur de 0,8% des cotisations hors taxes.

Contribution covid

La taxe covid s'applique pour l'année d'exercice 2020 aux complémentaires et a pour objectif de compenser une partie de déficit de la SS, en transférant les fonds non déboursés des organismes de complémentaires, dus aux conséquences de la crise économique et sanitaire de la covid19. Cette taxe s'élève à 3,90% et se prélève à partir des cotisations hors taxes.

Deuxième partie

Présentation des données

Chapitre 3

Portefeuille des contrats étudiés

Cette partie s'intéresse à la présentation du portefeuille et du périmètre étudié. Les données utilisées concernent un portefeuille des contrats collectifs d'assurance complémentaire santé, souscrits au sein de plusieurs produits standards collectifs en gestion directe. Les données disponibles couvrent la période des 5 années d'observation de 2017 jusqu'à 2021.

Pour faciliter la lecture des produits étudiés, nous allons tout d'abord dérouler une liste numérotée rattachée à ces produits, puis nous allons en présenter les caractéristiques de chacun :

— Produit A : Santé Entreprise ;
— Produit B : Syntec ;
— Produit C : Esprit Collectif ;
— Produit D : FEDESAP & AIDEDOM ;
— Produit E - Regroupement de 3 produits destinés au secteur sportif : Sport et Santé, Tennis Club, ASPTT.

Historiquement, le marché du collectif de LMG réalise un bond en 2020 grâce à la commercialisation du nouveau produit A - Santé Entreprise, puisque ce dernier cible un large panel d'entreprises, dont celles qui auparavant se destinaient à l'un des 4 autres produits B, C, D et E. Ces derniers produits ne sont plus commercialisés, mais ont généré un portefeuille de contrats d'assurance durant leur période de vente, pour lequel la LMG continue de remplir ses engagements de mutuelle, jusqu'à la liquidation progressive du portefeuille. Le portefeuille rattaché aux produits B, C, D et E se dit ainsi "en run-off".

3.1 Produit A : Santé Entreprise

Regroupant plus de 13 000 assurés principaux, Santé Entreprise est le produit phare du marché collectif de la Mutuelle Générale, et le seul commercialisé actuellement. Il s'agit d'une offre santé lancée en 2019, dont la mise au point a été réalisée en amont de la réforme du 100% Santé. Sa tarification initiale est ainsi conforme à la réglementation rattachée à cette réforme.

Santé Entreprise se décline en 3 offres selon les clients ciblés

Offre Santé Entreprise Modulaire

L'offre Modulaire se destine aux petites PME de 10 salariés à 99 salariés, affiliés soit au Régime Général, soit au régime Alsace-Moselle.

L'offre Modulaire peut être proposée à différents groupes de salariés dits "collèges" : cadres, non-cadres, ou ensemble du personnel (cadre et non cadre confondus).

Elle propose une base de couverture obligatoire offrant près de 11 niveaux de garantie allant du socle ANI (remboursement du Ticket Modérateur), jusqu'aux plafonds du contrat responsable, intégrant les paniers du 100% Santé. Des options facultatives dont 10 responsables et 1 non responsable sont également mises à disposition : l'entreprise peut souscrire 2 options facultatives, au choix, dont la cotisation sera à la charge du salarié, s'il souhaite y adhérer.

L'offre se nomme "Modulaire" car le choix des niveaux de garanties sur la base obligatoire fonctionne selon le mécanisme du « -3/+3 ». En effet, l'entreprise choisit un niveau de garantie pour le module de prestations « Soins Courants ». Ce dernier correspond au module de référence qui va piloter la règle des –3/+3 niveaux. Pour chacun des 5 autres modules de prestations, l'entreprise pourra choisir un niveau de garantie dans un intervalle de 3 niveaux inférieurs et 3 niveaux supérieurs par rapport au niveau choisi pour le module Soins courants.

Par exemple, si l'entreprise souscrit au niveau 4 du module Soins Courants, alors celle-ci pourra choisir le module Hospitalisation au niveau 6, Optique au niveau 7, Dentaire au niveau 7, Audio au niveau 5 et Prévention & Bien-Être au niveau 3 :

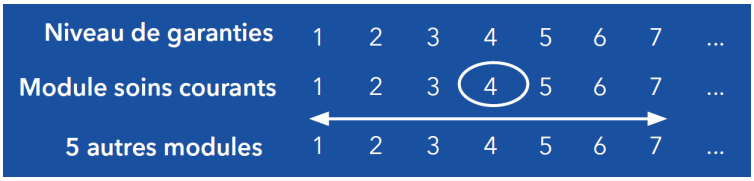


FIGURE 3.1 – Principe de l'offre Modulaire (source : LMG)

Offre Santé Entreprise Packagée

L'offre Packagée cible les entreprises de 1 à 19 salariés et met à disposition 5 formules de garanties au choix, qui propose une couverture de la formule ANI aux plafonds du contrat responsable, et intègre les paniers du 100% Santé. Ces 5 formules correspondent à des combinaisons prédéfinies des 11 niveaux de garanties présentées dans l'offre Modulaire.

Combinaison de niveaux de garantie selon les formules de l'offre Packagé

	FORMULE 1 Socle ANI	FORMULE 2	FORMULE 3	FORMULE 4	FORMULE 5
SOINS COURANTS	1	4	6	9	11
HOSPITALISATION	1	4	6	9	11
OPTIQUE	1	3	5	8	11
DENTAIRE	1	4	6	9	11
AIDES AUDITIVES	1	3	5	8	11
PREVENTION & BIEN-ETRE	1	2	5	8	11

FIGURE 3.2 – Principe de l'offre Packagée

Tout comme le Modulaire, le Packagé couvre le régime général et Alsace-Moselle, et les collèges couverts sont les cadres, non-cadres et l'ensemble du personnel. En plus de la base de couverture obligatoire, 4 options facultatives responsables et 1 non responsable : l'entreprise peut ainsi souscrire 2 options facultatives, dont la cotisation sera à la charge du salarié si celui-ci souhaite à adhérer à l'une d'elles.

Offre Santé Entreprise Dirigeant

L'offre Dirigeant est destinée aux dirigeants d'entreprise de 1 à 19 salariés, disposant du statut de Travailleurs Non Salarié et ayant souscrit un contrat collectif santé entreprise au profit de leur salarié. L'offre couvre ainsi la Sécurité sociale Indépendants, et propose la même base de couverture obligatoire que celle souscrite par les salariés. Le dirigeant peut également adhérer à une option facultative non responsable, venant renforcer la prise en charge sur les dépassements d'honoraires et les montures en optique.

3.2 Produit B : Syntec

Ce produit standard collectif en run-off est destiné aux salariés d'une branche professionnelle particulière dans le cadre d'une Convention collective nationale (CCN).

Une convention collective nationale est un accord signé entre les organisations d'employeurs et les organisations syndicales de salariés, sur le droit du travail d'une branche professionnelle spécifique. L'accord a pour objectif de prendre en compte les caractéristiques rattachées au secteur d'activité, en précisant notamment les conditions de travail, mais aussi les obligations en termes de complémentaire santé obligatoire. La complémentaire santé fournie par l'employeur doit ainsi respecter les règles imposées par sa convention collective.

Parmi ces CCN, nous pouvons retrouver SYNTEC, qui s'adresse aux bureaux d'études techniques et aux cabinets de conseil. Près de 80 000 entreprises dépendent de cette branche, et près de 900 000 salariés la constituent. Les secteurs d'activités concernés par SYNTEC sont entre autres :

— Programmation informatique ;

— Études de marchés et sondages ;
— Ingénierie, études techniques ;
— Traduction et interprétation ;
— Conseil en relation publique et communication...

LMG commercialisait ainsi un produit santé collectif offrant 2 formules de garanties, situées dans les tranches au milieu et en haut de gamme.

3.3 Produit C : EspritCo

EspritCo est une offre standard de complémentaire santé d'entreprise pour les TPE/PME, lancée en 2009 et actuellement en run-off.

La multitude de codes offres dans les bases de données qui sont rattachés à ce produit indique une grande évolution du Produit C - Esprit Collectif. Ce dernier s'est ainsi déployé en différentes générations au fil des années, chacune proposant de différentes offres.

Toutefois, la recherche d'historique et d'informations sur le réseau LMG et auprès de diffé directions a permis de mettre en lumière un manque de documentations concernant ce produit, notamment sur les caractéristiques de chacune des offres.

Nous présenterons ci-dessous les informations récupérées sur le produit ainsi que ces offres.

Esprit Collectif Options/Publicité/Cabinets médicaux

Esprit Collectif Option est la 1ère version du produit EspritCo, et a donné lieu par la suite à de nombreuses autres versions. Elle propose un contrat responsable (base + options) et 6 niveaux de couverture au choix, à un tarif adapté pour répondre à la démographie de l'entreprise. À partir du 1er janvier 2019, une uniformisation sera apportée sur les tableaux de garanties afin qu'il n'y ait plus qu'une seule version commune aux régimes général et local. La distinction entre les 2 régimes se fera uniquement sur les cotisations.

À partir de mai 2017, le produit Esprit Collectif Options propose des déclinaisons « ligne métier » cabinets médicaux et publicité, pour proposer un produit au secteur des cabinets médicaux et de la publicité.

Esprit Collectif Modulaire

Esprit Collectif Modulaire s'adresse aux TPE/PME de 10 à 50 salariés. Cette offre est une solution souple et plus compacte que Esprit Co Option : 2 modules Soins courants/-Hospitalisation et Optique/Dentaire sont mis en place, en offrant 5 niveaux de couverture et 19 combinaisons possibles.

Esprit Collectif Dirigeant

Il s'agit de l'offre proposée aux chefs d'entreprises de TPE/PME, avec 6 niveaux de couverture au choix.

3.4 Produit D : FEDESAP & AIDEDOM

Le produit cible les entreprises adhérant à la CCN Fédération Française des Services à la Personne et de Proximité. Lancé en 2016 et désormais en run-off, il concerne les entreprises du secteur des services à la personne implantées en France métropolitaine. Il offre une base employeur avec une option facultative pour différents salariés : soit une base obligatoire et 3 options facultatives qui sont financées par les salariés. La base propose 4 options de garanties positionnées dans les tranches entrée et milieu de gamme. Nous pouvons retrouver dans les bases les produits qui se nomment SAP/FESP/FEDESAP. Il s'agit en réalité du même produit en termes de garanties. Le produit de base est SAP. Les 2 autres sont des déclinaisons pour répondre à des fédérations/associations.

3.5 Produit E : Sport et Santé, Tennis Club, ASPTT

Ces 3 produits concernent des complémentaires santé liées à des CCN sportives.

Le produit Sport relève de la CCN Sport et est proposé aux entreprises des secteurs d'activités suivants :

- Organisation, gestion et encadrement d'activités sportives ;
- Gestion d'installation et d'équipements sportifs ;
- Enseignement, formation aux activités sportives et formation professionnelle aux métiers du sport ;
- promotion et organisation de manifestations sportives.

Tennis Club et ASPTT (Association sportive des postes, télégraphes et téléphones) sont 2 autres petits produits. l'ASPTT est originellement le nom donné aux associations sportives des postiers français. De ces associations est née une fédération : la fédération sportive des ASPTT. Il s'agit d'un héritage historique de LMG.

Il est difficile de bien segmenter l'ensemble des produits standards collectifs santé en raison de leur statut en run-off, et du fait qu'aucun document à jour ne soit mis à disposition. Malgré ce manque d'information sur ces produits, le bon périmètre pourra bien être choisi sur SAS grâce à la liste précise de l'ensemble des codes offres associées à ces produits.

Chapitre 4

Base de données

Nous manipulerons plusieurs sources de données :

- Sources de données relatives à l'outil de gestion interne de contrats « Active Infinite » (AI), dans laquelle nous trouvons les bases effectifs, prestations et cotisations ;
- Sources de données relatives à l'outil d'aide à la vente commerciale « EOLE », dans laquelle nous trouvons les différentes bases de données commerciales ;
- Table de données fournie par l'équipe Data et IA relatives aux informations des entreprises ayant souscrites chez LMG ;
- Source de données externe INSEE présentant la densité de professionnels de santé pour 100 000 habitants et par département (chiffres de 2018).

4.1 Bases EOLE

Les différentes bases EOLE regroupent toutes les données commerciales et contractuelles. Notons que lorsque les commerciaux rentrent de nouvelles données sur la plateforme EOLE concernant un nouveau compte ou contrat, les informations se stockent dans différentes sources selon la nature des informations. Nous pouvons citer comme sources : celles des opportunités commerciales, des devis commerciaux, ou encore des avenants aux contrats commerciaux... Ces différentes bases sont soumises à de nombreux traitements et jointures afin de construire une nouvelle table de travail lisible, combinant les données provenant des différentes vues. Ci-dessous un tableau des variables de la table construite EOLE qui présente par année le portefeuille de contrats commerciaux présents entre 2017 et 2021.

Variable	Description
ID_CTR	Numéro d'identification commercial du contrat
ID_Compte	Numéro d'identification commercial du compte
ID_Opport	Numéro d'identification commercial de l'opportunité
ID_devis	Numéro d'identification commercial du devis
ann	Année
NUM_DAC	Numéro de contrat commercial
Nom_compte	Nom de l'entreprise
SIREN	Numéro de SIREN
Nom_Produit	Nom du produit
DT_DB_CTR	Date de début du contrat
DT_RESIL_CTR	Date de résiliation du contrat
Tx_remise	Taux de remise commerciale appliquée sur les cotisations TTC
COLLEGE	Collège / Catégorie socio professionnelle
STRUCT_COT	Structure de cotisation
CD_APE	Code APE
LIB_APE	Libellé code APE
TOTAL_CHARGEMENT	Taux de chargement
CD_POSTAL	Code postal de l'entreprise

FIGURE 4.1 – Variables de la table EOLE

Les variables ID sont les clés qui nous ont permis de réaliser les jointures entre les différentes bases EOLE. Le num_dac correspond au numéro commercial caractérisant le contrat. Cette variable sera notre maille dans la base finale, avec l'année. L'entreprise possède un code et un libellé APE (Activité Principale Exercée) indiquant son secteur d'activité.

4.2 Base des effectifs

La base brute des effectifs est une base historisée de l'ensemble des bénéficiaires, qui détaille les informations sur le genre, date de naissance, numéro assuré et numéros de contrats, ou encore leurs différentes garanties. Cette base comprend une ligne pour chaque assuré et pour chacune de ses garanties souscrites. La base étant dupliquée de manière journalière, il est nécessaire de filtrer sur une date d'alimentation afin d'extraire une table avec toutes les connaissances que nous avons à cette date.

À partir de la base brute, nous construisons d'abord une table agrégée par année et par assuré. Nous récupérons le num_dac de la base EOLE grâce aux clés de jointure Voici un tableau récapitulatif des variables essentielles au sein de cette table.

Variable	Description
ann	Année
num_ass	Numéro assuré : propre à chaque assuré
NUM_CTR_INDIV	Numéro de contrat individuel : chaque salarié détient un contrat individuel pour lui et ses bénéficiaires
NUM_CTR_COL	Numéro de contrat collectif : unique pour une entreprise et un collège tarifé
LIB_CTR_COL	Libellé du contrat collectif
NUM_DAC	Numéro de contrat commercial
gamme	Nom de la génération du produit
CD_OFFRE	Code de l'offre reliée au produit
TYPE_ASS	Type de l'assuré
CD_SEXE	Code sexe de l'assuré
age	Age de l'assuré
dte_effet_ctr_col	Date d'effet du contrat collectif
dte_fin_ctr_col	Date de fin du contrat collectif
dte_deb_ctr_indiv	Date de début du contrat individuel
dte_fin_ctr_indiv	Date de fin du contrat individuel
dte_deb_couverture	Date de début de couverture de l'assuré
dte_fin_couverture	Date de fin de couverture de l'assuré
tx_presence	Taux de présence de l'assuré à l'année : proportion de l'année pendant laquelle il a été présent
Region_DT	Groupement de régions utilisée par la direction technique
Zone_STD	Zonier de consommation du périmètre standard construit par la direction technique
dt_ali	Date d'alimentation

FIGURE 4.2 – Variables de la table des effectifs agrégés par assuré

Un contrat collectif possède un numéro de contrat commercial (NUM_DAC) unique qui le distingue des autres. Il bénéficie aussi d'un code offre (CD_OFFRE).

.Lorsqu'un salarié d'une entreprise souscrit à un contrat collectif, LMG créée un contrat individuel qui est rattaché au contrat collectif. Ceci permet de renseigner la date de début du contrat de l'assuré en question qui peut être embauché après la signature du contrat collectif. De plus, les ayants droits du salarié sont rattachés au même contrat individuel : : il peut s'agir du conjoint, des enfants ou des ascendants. La variable « type assuré » (TYPE_ASS) donne cette information dans la base.

Le zonier se base sur une étude de corrélation entre le niveau de dépenses en complémentaire santé des salariés et le département de l'entreprise. La zone 1 correspond au niveau de consommation le plus élevé, et la zone 5 au plus faible. Les tarifs proposés par LMG sont alors indexés selon le niveau de consommation de la zone dans laquelle se trouve l'entreprise.

4.3 Base des prestations

La base des prestations présente les remboursements effectués auprès des bénéficiaires et contient les informations liées à l'acte de soins remboursé. Nous notons qu'un acte de soins est un ensemble d'actions et de pratiques mises en œuvre pour participer au rétablissement ou à l'entretien de la santé d'une personne [1]. Il peut se décomposer en actes médicaux définis et limités. Nous construisons une première table de prestations

1. Code de la santé publique

annuelle par numéro de décompte du soin accompli.

Variable	Description
NUM_DAC	Numéro de contrat commercial
gamme	Nom de la génération du produit
CD_OFFRE	Code de la génération de produit
NUM_CTR_COL	Numéro de contrat collectif
num_decompte_DT	Numéro de décompte
NUM_CTR_INDIV	Numéro du contrat individuel
NUM_ASS	Numéro assuré
TYPE_ASS	Type de l'assuré
Type_ass_DT	Sexe de l'assuré
module	Module de soins
BRANCHE	Branche de soins
cd_acte	Code acte
SMT_LB_ACT	Libéllé du code acte
SMT_GRP_DT_N4	Nom du groupement d'acte
SMT_PANIER	Panier de soins (renseigné si le soin est concerné par la réforme 100% Santé)
CD_OPTION_COORDINATION	Code option de pratique tarifaire maitrisé
DTE_SURV	Date de survenance
AAAA_DTE_SURV	Année de survenance
AAAAMM_DTE_SURV	Année et mois de survenance
PRIX_UNITAIRE	Prix unitaire de l'acte
MNT_BR	Montant Base de remboursement
NB_ACTES	Nombres d'actes
MNT_DEPENSE	Montant de dépense
MNT_RO	Montant remboursé par le régime obligatoire
MNT_RC	Montant remboursé par le régime complémentaire
RAC	Reste à charge
dt_ali	Date alimentation

FIGURE 4.3 – Variables de la table des prestations

Chaque acte réalisé possède un libellé et un code acte (CD_ACTE) unique qui est

rattaché à une nomenclature officielle de la SS. Cela permet de retrouver pour chaque soin son montant de BRSS et son taux RO. Notons que le code acte est un répertoire interne à LMG.

Pour un soin accompli, le praticien envoie à la Sécurité sociale une feuille de soin avec le détail des actes réalisés. Celle-ci se traduit numériquement par un numéro de décompte (NUM_DECOMPTE) comprenant une date et un numéro. Il peut apparaître plusieurs lignes (NUM_LIG_DECOMPTE) à un même numéro de décompte puisqu'il peut y avoir plusieurs actes au cours d'une même consultation ou des régulations (annulation par exemple).

La date de survenance correspond à la date de décompte de l'acte. Cette date est celle à laquelle l'acte a été payé, en format jours+mois+année ou seulement mois+année.

Le nombre d'actes (NB_ACTE) indique la quantité d'actes consommés pour la même ligne de prestations. Pour l'achat de deux boîtes d'un même médicament au même moment, le nombre d'actes est de deux. Le nombre d'actes RC (NB_ACTES_RC) désigne le nombre d'actes pour lequel la mutuelle a dû effectuer un remboursement non nul. C'est le même principe pour le nombre d'actes RO (NB_ACTES_RO).

Le montant de dépense (MNT_DEPENSE) représente la dépense effective totale.

Rappelons que chaque soin possède un montant de BRSS (MNT_BASE_REMBOURSE et un taux RO (TAUX_RO). Les taux de remboursements RO et RC - basés sur le

mon- tant de BRSS - permettent de retrouver les montants de remboursement de régimes obligatoires (MNT_RO) et complémentaires (MNT_RC).

Nous rappelons que nous disposons d'un historique de 5 années de sinistralité. Afin de prendre en compte les prestations survenues en 2021 mais pas encore enregistrées dans la base en 2021, nous choisissons une date d'alimentation à mi 2022 pour l'extraction des prestations de la survenance 2017-2021.

4.4 Base des cotisations

La base des cotisations décrit les cotisations émises par les assurés, en précisant la date d'émission et les montants liés à la cotisation (hors taxe, toutes taxes comprises...). En assurance collective, la cotisation représente le montant versé par l'employeur à l'assureur en échange de la couverture complémentaire. Ces cotisations sont calculées hors taxes (HT) et avec taxes (TTC). Ces données sont des montants observés.

Nous construisons une première table des cotisations par assuré et par période d'émission.

Variable	Description
ANN	Année
CD_OFFRE	code de la génération du produit
NUM_DAC	Numéro de contrat commercial
NUM_CTR_COL	Numéro de contrat collectif
CD_PRODUIT	Code de la génération du produit
LIB_PRODUIT	Libéllé du code de la génération du produit
CD_GAT	Code de la garantie technique
LIB_GT	Libellé de la garantie technique
DTE_DEB_CTR_COL	Date de début du contrat collectif
DTE_FIN_CTR_COL	Date de fin du contrat collectif
AAAA_DTE_EMISSION_COT	Année d'émission de la cotisation
MM_DTE_EMISSION_COT	Mois d'émission de la cotisation
AAAAMM_DTE_ALIMENTATION	Année et Mois de la date d'alimentation
MNT_HT_POUR_DT	Montant Hors Taxes de la cotisation utilisé par la Direction Technique
MNT_TTC_COT_EMISES	Montant Toutes Taxes comprises de la cotisations

FIGURE 4.4 – Variables de la table des cotisations

4.5 Données externes à la Direction Technique

Nous avons tenté de récupérer des informations supplémentaires permettant de décrire le contrat et son entreprise, pouvant avoir un impact sur le niveau de rentabilité.

4.5.1 Base INSEE de la densité médicale par département

Afin d'enrichir notre étude, nous avons décidé de nous procurer des données issues de l'INSEE décrivant la densité médicale 2018 par département pour 100 000 habitants. La table se présente comme ceci :

Variable	Description
ANN	Année
CD_OFFRE	code de la génération du produit
NUM_DAC	Numéro de contrat commercial
NUM_CTR_COL	Numéro de contrat collectif
CD_PRODUIT	Code de la génération du produit
LIB_PRODUIT	Libéllé du code de la génération du produit
CD_GAT	Code de la garantie technique
LIB_GT	Libellé de la garantie technique
DTE_DEB_CTR_COL	Date de début du contrat collectif
DTE_FIN_CTR_COL	Date de fin du contrat collectif
AAAA_DTE_EMISSION_COT	Année d'émission de la cotisation
MM_DTE_EMISSION_COT	Mois d'émission de la cotisation
AAAAMM_DTE_ALIMENTATION	Année et Mois de la date d'alimentation
MNT_HT_POUR_DT	Montant Hors Taxes de la cotisation utilisé par la Direction Technique
MNT_TTC_COT_EMISES	Montant Toutes Taxes comprises de la cotisations

FIGURE 4.5 – Variables de la table INSEE

Il nous a semblé intéressant d'intégrer dans notre étude une information représentant l'accès aux soins : nous pouvons nous demander si un accès de soins facilité au niveau de la zone de l'entreprise pourrait influencer la consommation des assurés et de ce fait le P/C du contrat collectif.

4.5.2 Base de la Direction Stratégique Digitale & Data (DSDD)

Nous nous sommes également rapprochés de l'équipe DSDD qui possède des données historiques supplémentaires sur les entreprises clientes de LMG ; nous avons pu ainsi nous procurer les informations suivantes :

Variable	Description
SIREN	Numéro de SIREN
EOLE_Debut_couverture	Date de début de couverture du premier contrat de l'entreprise
dte_creation_entreprise	Date de création de l'entreprise

FIGURE 4.6 – Variables de la table fournie par la DSDD

Nous pourrons donc étudier l'éventuelle influence sur la rentabilité du contrat de l'âge de l'entreprise, et de l'ancienneté historique de l'entreprise chez LMG.

Troisième partie

Nettoyage et fiabilisation des données

Chapitre 5

Pré-traitement de la base de données

Ce chapitre a pour but d'expliquer les différentes étapes de traitement que nous réalisons afin de compléter nos données et d'y apporter des corrections. Dans la suite de l'étude, les termes de "produit" et "gamme" correspondent à la même notion ; ainsi la gamme Santé Entreprise désigne le produit Santé Entreprise précédemment présenté. L'assimilation de ces 2 termes provient du fait que le nom de la variable caractérisant le nom du produit (Santé Entreprise, EspritCo, Syntec, etc) dans les bases de données s'intitule "gamme".

5.1 Variable Niveau de garantie

Le traitement ci-dessous consiste à récupérer depuis la table du portefeuille le niveau de garantie souscrit par chaque assuré au sein de leur contrat.

La variable niveau_garantie de la base EOLE qui transmet cette information étant très peu renseigné, nous avons réfléchi à un mode de récupération de la modalité. Nous avons réalisé un traitement sur cette variable en récupérant la variable cd_gt de la vue portefeuille. Cette variable renseigne un code garantie rattaché à l'assuré. Plusieurs codes garantie peuvent être rattachés à un assuré dans la vue portefeuille. En effet, le cd_gt renseigne les garanties, mais également les forfaits et services auxquels l'assuré a droit, comme le forfait maternité et le service client téléphonique. La particularité de ce traitement réside dans le fait que les différents produits du périmètre standard collectif ne proposent pas les mêmes formules de garantie. Chaque produit du standard collectif dispose de ses propres niveaux de garantie, ainsi les modalités du cd_gt sont différentes entre ces produits.

Le code de la garantie technique n'est donc pas commun à l'ensemble du périmètre. Nous avons donc analysé pour chaque produit les modalités du cd_gt ainsi que la variable lib_gt qui complète le code avec le libellé. La variable niveau_garantie étant bien renseignée seulement pour la gamme Santé Entreprise, nous avons pu réaliser le traitement sur tous les autres produits.

Voici donc les étapes pour la création de la variable niveau de garantie :

- Rapatriement du « code garantie technique » de l'assuré à partir du numéro assuré
- Construction d'une table de transcodification entre le code garantie et un niveau de garantie simplifié : cette étape a nécessité une recherche manuelle dans l'outil de gestion sur une centaine de codes.
- Ajout de la variable niveau de garantie dans la base de données

Voici un exemple pour la gamme Syntec : chaque assuré a bien un cd_gt qui renseigne la formule souscrite, nous pouvons donc à partir de cette variable compléter la variable niveau_garantie, en récupérant la dernière lettre du cd_gt.

De même, une autre règle de décision a été construite pour les autres offres de la gamme FEDESAP&AIDEDOM, ainsi que AUTRE (ccnsport, tennisante, asptt) : un assuré a au plus 3 cd_gt qui lui sont rattachés, dont 1 renseigne si l'assuré a souscrit à la base employeur seule (« BA » à la fin de la combinaison du cd_gt) ou bien s'il a souscrit en plus à une option (BA+ le numéro de l'option souscrite).

num_ass	TYPE_ASS	CD_OFFRE	CD_PRODUIT	CD_GT	LIB_GT
9105	ASSPRI	FEDESAP	AS2015FSAP	CMC173BA2	Complémentaire Maladie Collective 173 Base 2
9105	ASSPRI	FEDESAP	AS2015FSAP	MAT175OP2	Forfait Maternité Collectif Option 2
9105	ASSPRI	FEDESAP	AS2015FSAP	CMC175OP2	Complémentaire Maladie Collective 175 Option 2
320481	ASSPRI	FEDESAP	AS2015FSAP	CMC173BA	Complémentaire Maladie Collective 173 Base Seule
354962	ASSPRI	FEDESAP	AS2015FSAP	CMC173BA1	Complémentaire Maladie Collective 173 Base 1
354962	ASSPRI	FEDESAP	AS2015FSAP	CMC174OP1	Complémentaire Maladie Collective 174 Option 1
445757	ASSPRI	FEDESAP	AS2015FSAP	CMC173BA	Complémentaire Maladie Collective 173 Base Seule
469832	ASSPRI	FEDESAP	AS2015FSAP	CMC173BA	Complémentaire Maladie Collective 173 Base Seule
476978	ASSPRI	FEDESAP	AS2015FSAP	CMC173BA3	Complémentaire Maladie Collective 173 Base 3
476978	ASSPRI	FEDESAP	AS2015FSAP	MAT176OP3	Forfait Maternité Collectif 176 Option 3
476978	ASSPRI	FEDESAP	AS2015FSAP	CMC176OP3	Complémentaire Maladie Collective 176 Option 3

FIGURE 5.1 – Illustration de la construction de la variable cd_gt pour l'offre FEDESAP

Un développement de ce traitement et du cheminement d'étude sont exposées en Annexes.

5.2 Jointure dans la base des prestations du niveau de garantie souscrit par l'assuré

La construction de la base finale des P/C par entreprise nécessite une jointure entre la base des prestations et des effectifs.

Nous avons rencontré un problème tenace de doublons sur une jointure entre la base des prestations et des effectifs : nous remarquons que des assurés possèdent différents niveaux de garanties dus à des changements de contrat, ce qui crée des doublons lorsqu'on souhaite joindre le niveau de garantie aux lignes de prestations de chaque assuré.

Pour y remédier, nous avons donc créé des tables mensuelles des effectifs : étant donné qu'un assuré ne peut pas avoir plusieurs niveaux de garanties sur un même mois, cela permet d'avoir une transcodification du niveau de garantie pour chaque assuré par mois.

Nous avons ensuite réalisé une jointure sur la table des prestations qui, elle aussi, a été mensualisée, ce qui devrait permettre d'éviter tout doublon.

Nous avons ensuite ré-agrégé la table des prestations mensuelles pour obtenir une table annuelle. Suite à ce retraitement, nous observons désormais un écart d'une quarantaine de lignes entre la table initiale des prestations et la table des prestations obtenue par jointure, contre 10 000 lignes d'écart avant la modification de la requête. Les doublons restants apparaissent lorsque l'assuré change de contrat en milieu de mois ; nous considérons dans ce cas-là que la prestation enregistrée pour ce mois s'associe avec le nouveau niveau de garantie souscrit.

5.3 Variable Niveau de gamme

Le périmètre standard collectif propose de multiples formules de garantie pour chaque offre, ce qui rend difficile la comparaison des niveaux de garanties entre les offre et produit.

GAMME	OFFRES	FORMULE de GARANTIE	OPTIONS FACULTATIVES
SANTE ENTREPRISE	MODULAIRE	Modulaire de 1 à 11	10 resp + 1 non resp
	PACKAGE	Formule de 1 à 5	4 resp + 1 non resp
	DIRIGEANT		
ESPRITCO	ESPRITCO	Niveau et Modulaire de 1 à 6	niveaux 3,4,5 et 6 peuvent être souscrit par le salarié en tant qu'option si l'employeur a souscrit une base ANI,3 ou 4
	DIRIGEANT	Niveau de 1 à 6	
	PREMS	Niveau et Modulaire de 1 à 6	
FEDESAP & AIDEDOM	FEDESAP	1 Base	3 options
	AIDEDOM		
	AIDEPERS		
SYNTEC	SYNTEC	Formule A et B	X
AUTRE	ASPTT	1 Base	2 options
	CCNSPORT	1 Base	2 options
	FESP	1 Base	3 options
	TENNISANTE	1 Base	2 options

FIGURE 5.2 – Présentation des formules de garanties par offre

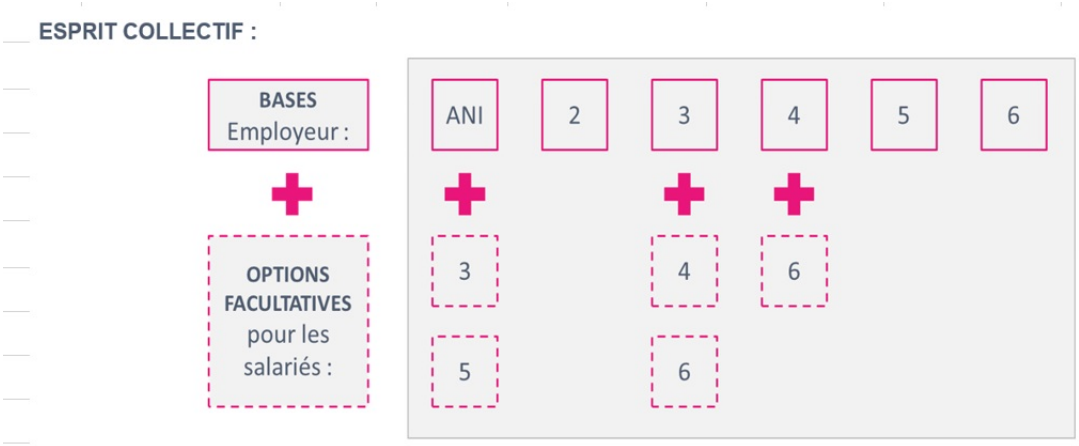


FIGURE 5.3 – Illustration de la formule de garantie pour le produit EspritCo

Afin d'avoir une idée sur le niveau de couverture proposé par chaque offre, nous avons réalisé un second traitement, permettant de classifier le niveau de couverture des différentes formules de garanties au sein des offres. L'idée est de pouvoir harmoniser les niveaux de garanties de l'ensemble des produits, en créant une variable niveau de gamme, qui renseignerait si les niveaux de garantie au sein de l'offre s'apparentent à de l'Entrée de gamme, du Milieu de gamme, ou du Haut de gamme.

Pour cela, nous nous sommes basés sur les montants RC moyen par acte observés dans la gamme Santé Entreprise en 2020 et 2021. Pour rappel, au sein de la gamme SE, 11 niveaux de garanties sont proposés. Les niveaux de garantie sont regroupés par 6 grandes familles de soins, appelées modules de soins. Les niveaux de module dépendent du choix initial du niveau du module « Soins Courants ». Les autres modules choisis peuvent être d'un niveau supérieur ou inférieur de +3-3 au maximum par rapport au module « Soins Courants ». Nous appelons le

— Bas de gamme : les régimes choisis avec un niveau de module SC de 1 à 4
— Milieu de gamme : les régimes choisis avec un niveau de module SC de 5 à 8
— Haut de gamme : les régimes choisis avec un niveau de module SC de 9 à 11.

Les niveaux de gamme étant déjà identifiés pour SE, nous pouvons ainsi utiliser cette référence pour classifier les différents niveaux de garantie des autres offres. Pour cette étude, nous avons donc décidé de travailler sur les modules les plus prépondérants en termes de prestations, c'est pourquoi nous nous sommes intéressés aux modules Optique, Dentaire, Soins Courants et Hospitalisation ; les montants de remboursement des modules Aides Auditives étant trop faibles. Au sein de chaque module de soins sélectionné, nous avons choisi les actes les plus représentatifs en termes de prestations, puis nous avons calculé le montant RC moyen de chacun de ces actes, pour les niveaux considérés comme Entrée de gamme, Milieu de gamme et Haut de gamme.

$$Montant\ RC\ moyen^{1} = \frac{Montant\ RC}{Nombre\ d'actes^{2}}$$

OPTIQUE		SANTE ENTREPRISE					
		Entrée de gamme		Milieu de gamme		Haut de gamme	
EQUIPEMENTS A TARIF LIBRE (5)	CD_GRP_DT	Remboursement moyen par acte	Poids de l'acte	Remboursement moyen par acte	Poids de l'acte	Remboursement moyen par acte	Poids de l'acte
Monture	MON	92 €	31%	98 €	23%	98 €	21%
Verre complexe	VER_COM	145 €	49%	218 €	52%	245 €	53%
Verre simple	VER_SIM	62 €	21%	102 €	24%	119 €	26%
GLOBAL NIVEAU DE REMBOURSEMENT		111 € [0;111]		161 € [111;161]		181 € [161;181]	
	Poids des prestations du module	16%		19%		17%	

FIGURE 5.4 – Illustration du montant RC moyen observé par niveau de gamme pour le module Optique

En pondérant ces montants RC moyen avec le poids des prestations de ces actes au sein du module, nous obtenons un montant moyen de remboursement observé, pour les 3 niveaux de gamme. Ces montants nous permettent d'obtenir un classement de niveau de remboursement du module selon le niveau de gamme. Nous calculons ainsi un niveau de remboursement moyen observé dans SE, pour les 6 modules et les 3 niveaux de gamme. Nous pondérons ces 6 niveaux de remboursement avec le poids des prestations des modules, afin d'obtenir un niveau de remboursement moyen global pour les 3 niveaux de gamme.

2021	Entrée de gamme	Milieu de gamme	Haut de gamme	
OPTIQUE	111 €	161 €	181 €	Montant de couverture moyen
	16%	19%	17%	Poids des prestations du module
DENTAIRE	263 €	371 €	445 €	
	77%	72%	74%	
SOINS COURANTS	11 €	12 €	14 €	
	3%	2%	2%	
HOSPITALISATION	43 €	53 €	69 €	
	5%	6%	6%	
GLOBAL SE Montant de couverture moyen global	222 €	303 €	366 €	

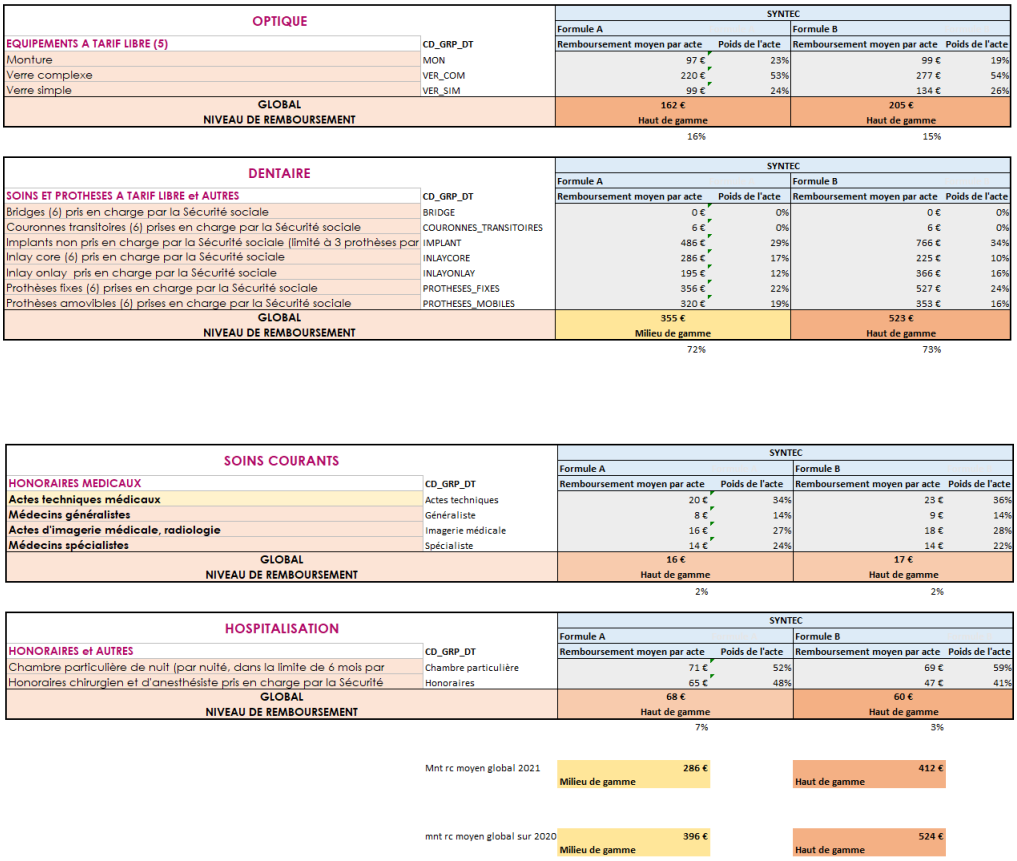
Moyenne 2020/2021			
GLOBAL SE Montant de couverture	311 €	479 €	480 €

FIGURE 5.5 – Montant RC moyen observé par niveau de gamme, par module et au global

Ces 3 niveaux de remboursements moyens globaux servent ainsi de références pour classifier les niveaux de remboursement moyens globaux des autres offres. Nous avons réalisé cette technique sur 2020 et 2021 puis nous avons fait une moyenne sur ces deux années.

Voici en guise d'exemple le tableau réalisé pour la gamme SYNTEC en 2021 :

1. Montant RC moyen total enregistré pour l'acte de soins
2. Nombre d'actes médicaux enregistré pour l'acte de soins



OPTIQUE		SYNTEC			
		Formule A		Formule B	
EQUIPEMENTS A TARIF LIBRE (5)	CD_GRP_DT	Remboursement moyen par acte	Poids de l'acte	Remboursement moyen par acte	Poids de l'acte
Monture	MON	97 €	23%	99 €	19%
Verre complexe	VER_COM	220 €	53%	277 €	54%
Verre simple	VER_SIM	99 €	24%	134 €	26%
GLOBAL NIVEAU DE REMBOURSEMENT		162 € Haut de gamme		205 € Haut de gamme	
		16%		15%	

DENTAIRE		SYNTEC			
		Formule A		Formule B	
SOINS ET PROTHESES A TARIF LIBRE et AUTRES	CD_GRP_DT	Remboursement moyen par acte	Poids de l'acte	Remboursement moyen par acte	Poids de l'acte
Bridges (6) pris en charge par la Sécurité sociale	BRIDGE	0 €	0%	0 €	0%
Couronnes transitoires (6) prises en charge par la Sécurité sociale	COURONNES_TRANSITOIRES	6 €	0%	6 €	0%
Implants non pris en charge par la Sécurité sociale (limité à 3 prothèses par	IMPLANT	486 €	29%	766 €	34%
Inlay core (6) pris en charge par la Sécurité sociale	INLAYCORE	286 €	17%	225 €	10%
Inlay onlay pris en charge par la Sécurité sociale	INLAYONLAY	195 €	12%	366 €	16%
Prothèses fixes (6) prises en charge par la Sécurité sociale	PROTHESES_FIXES	356 €	22%	527 €	24%
Prothèses amovibles (6) prises en charge par la Sécurité sociale	PROTHESES_MOBILES	320 €	19%	353 €	16%
GLOBAL NIVEAU DE REMBOURSEMENT		355 € Milieu de gamme		523 € Haut de gamme	
		72%		73%	

SOINS COURANTS		SYNTEC			
		Formule A		Formule B	
HONORAIRES MEDICAUX	CD_GRP_DT	Remboursement moyen par acte	Poids de l'acte	Remboursement moyen par acte	Poids de l'acte
Actes techniques médicaux	Actes techniques	20 €	34%	23 €	36%
Médecins généralistes	Généraliste	8 €	14%	9 €	14%
Actes d'imagerie médicale, radiologie	Imagerie médicale	16 €	27%	18 €	28%
Médecins spécialistes	Spécialiste	14 €	24%	14 €	22%
GLOBAL NIVEAU DE REMBOURSEMENT		16 € Haut de gamme		17 € Haut de gamme	
		2%		2%	

HOSPITALISATION		SYNTEC			
		Formule A		Formule B	
HONORAIRES et AUTRES	CD_GRP_DT	Remboursement moyen par acte	Poids de l'acte	Remboursement moyen par acte	Poids de l'acte
Chambre particulière de nuit (par nuité, dans la limite de 6 mois par	Chambre particulière	71 €	52%	69 €	59%
Honoraires chirurgien et d'anesthésiste pris en charge par la Sécurité	Honoraires	65 €	48%	47 €	41%
GLOBAL NIVEAU DE REMBOURSEMENT		68 € Haut de gamme		60 € Haut de gamme	
		7%		3%	
	Mnt rc moyen global 2021	286 € Milieu de gamme		412 € Haut de gamme	
	mnt rc moyen global sur 2020	396 € Milieu de gamme		524 € Haut de gamme	

FIGURE 5.6 – Calcul du niveau de gamme observé par formule de garantie et par module pour le produit SYNTEC

Finalement, nous obtenons les niveaux de gamme suivants :

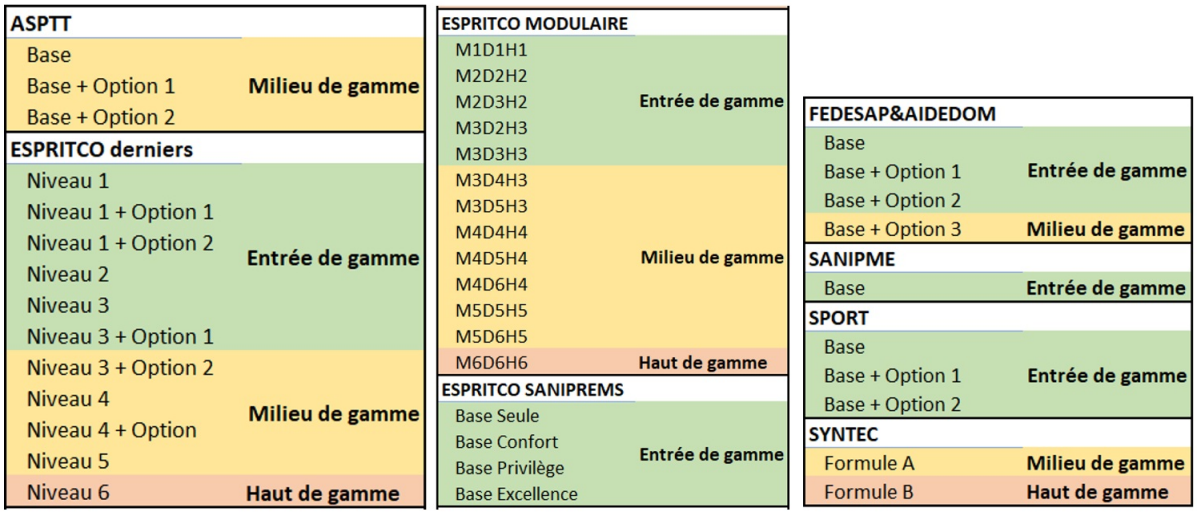


FIGURE 5.7 – Niveau de gamme de chaque formule de garantie selon les offres du périmètre standard collectif

Nous observons que les produits FEDESAP&AIDEDOM ainsi que les contrats complémentaires adaptés aux clubs sportifs s'apparentent à de l'Entrée de Gamme. L'ASPTT a globalement une couverture Milieu de gamme, ce qui peut s'expliquer avec l'historique de LMG, qui proposait des couvertures à l'administration responsable des postes et télégraphes. Syntec et EspritCo proposent quant à eux des niveaux de couvertures allant de l'Entrée de gamme pour les faibles niveaux de garantie, jusqu'à Haut de gamme pour les hauts niveaux de garanties.

5.4 Traitements supplémentaires

Nous avons également réalisé d'autres traitements sur les bases de données afin de corriger les erreurs observées sur les différentes variables.

5.4.1 Base EOLE

— Chaque offre du périmètre standard collectif propose des structures de cotisations différentes, qui peuvent contenir très peu d'effectifs selon les produits. Celles-ci peuvent avoir des noms différents entre les offres, mais s'apparente à la même structure ou peuvent se regrouper ensemble de par leur similarité. Pour alléger et harmoniser cette variable, nous avons réalisé des regroupements de structures de cotisations.

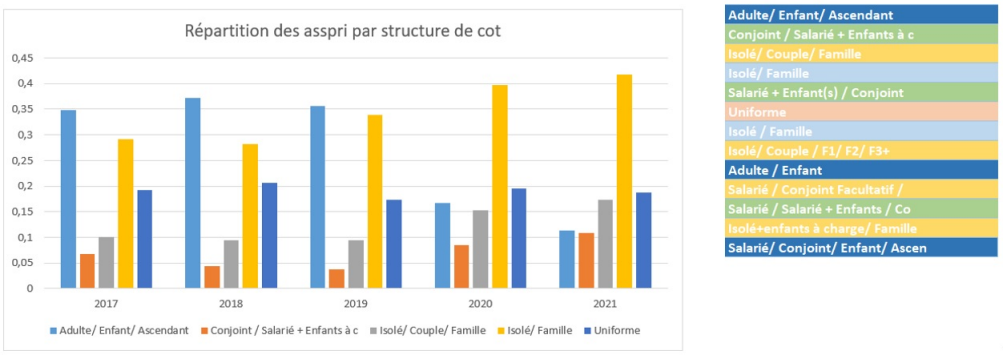


FIGURE 5.8 – Nouvelles modalités de la structure de cotisation

— La variable du taux de remise commerciale n'étant pas renseignée dans les bases EOLE avant l'année 2020, nous avons été contraints de ne pas l'inclure dans notre étude, malgré son potentiel impact sur la rentabilité du contrat. Cela constitue une limite opérationnelle à notre étude.

— Les codes APE ont été regroupé en 5 domaines afin de créeer une variable globale renseignant le domaine d'activité de l'entreprise. Les domaines APE sont : Agriculture, Commerce, Construction, Industrie, Service.

5.4.2 Base des effectifs

— Des anomalies sur le code sexe ont été constatées, nous avons donc retraité la variable cd_sexe en récupérant le premier chiffre du numéro de sécurité sociale associé à l'assuré.

— En analysant la fréquence des âges des assurés sur la base des effectifs, nous avons corrigé également des âges trop aberrants.

— Dans la base des effectifs agrégés, des doublons de numéros assurés étaient présents, avec notamment des numéros de contrat individuel qui se rattachent aux mêmes num_ass, avec échange de type_ass des bénéficiaires entre les 2 contrats (l'assuré principal du premier contrat devient le conjoint du second contrat et vice-versa). Ces anomalies provenant probablement d'une erreur de gestion, nous décidons d'éliminer un des doublons afin d'éviter des doublons de prestations.

5.4.3 Base des prestations

— Afin de mieux tenir compte de la sinistralité de l'année 2021, nous avons récupéré les données de provisions estimés des montants de cotisation et prestations de la survenance 2021

— Dans le cadre la réforme 100% Santé, des plafonds limites de vente ont été mis en place depuis 2020 sur des actes Optique, Dentaire et Audio. Ces plafonds restreignent le remboursement réalisé par la complémentaire santé à un certain montant en fonction du panier, tel que 100€ pour l'acte monture en panier Libre. Nous avons ainsi retraité les montants de remboursements de ces actes en Optique de manière ce que les montants ne dépassent pas les plafonds. Pour les autres modules, nous avons décidé de ne retraiter seulement les montants de remboursements aberrants par rapport à l'ensemble des montants RC au sein du module étudié. En analysant les boxplots des montants RC de chaque famille d'actes par module, nous avons pu identifier les montants aberrants et les corriger en une autre valeur retraitée. Cette valeur maximale correspond au double de la distance entre le 1er et le 3ème quartile des données de montants RC observées dans la famille d'acte étudiée. Cette correction permet alors de conserver l'information d'un remboursement élevé, tout en homogénéisant la base pour une meilleure modélisation par la suite.

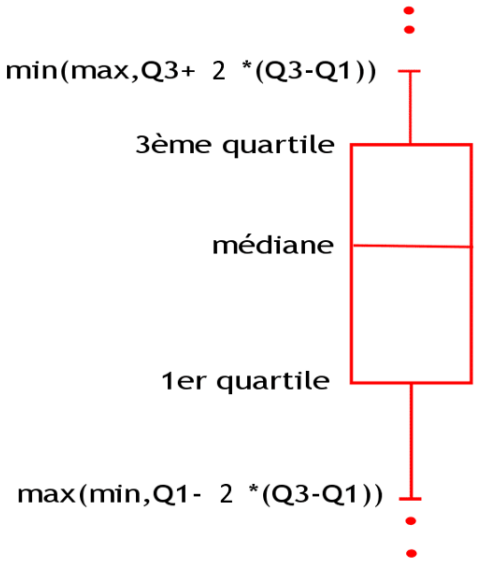


FIGURE 5.9 – Identification des outliers par la méthode du boxplot

Ces valeurs retraitées ne représentent cependant qu'une petite partie des données : nous supposons que les montants RC renseignés dans la base sont ceux effectivement réglés par LMG, il est donc important de ne pas trop transformer ces valeurs.

Chapitre 6

Base finale

Finalement, nous agrégeons les bases des effectifs, prestations et cotisations par contrat, en récupérant la variable du num_dac sur chacune de ces bases. Nous effectuons par la suite la jointure de ces bases avec la base EOLE finale enrichie des données DSDD et Insee. Nous illustrons par un schéma les volumétries des bases et les grandes étapes permettant la construction de la base finale.

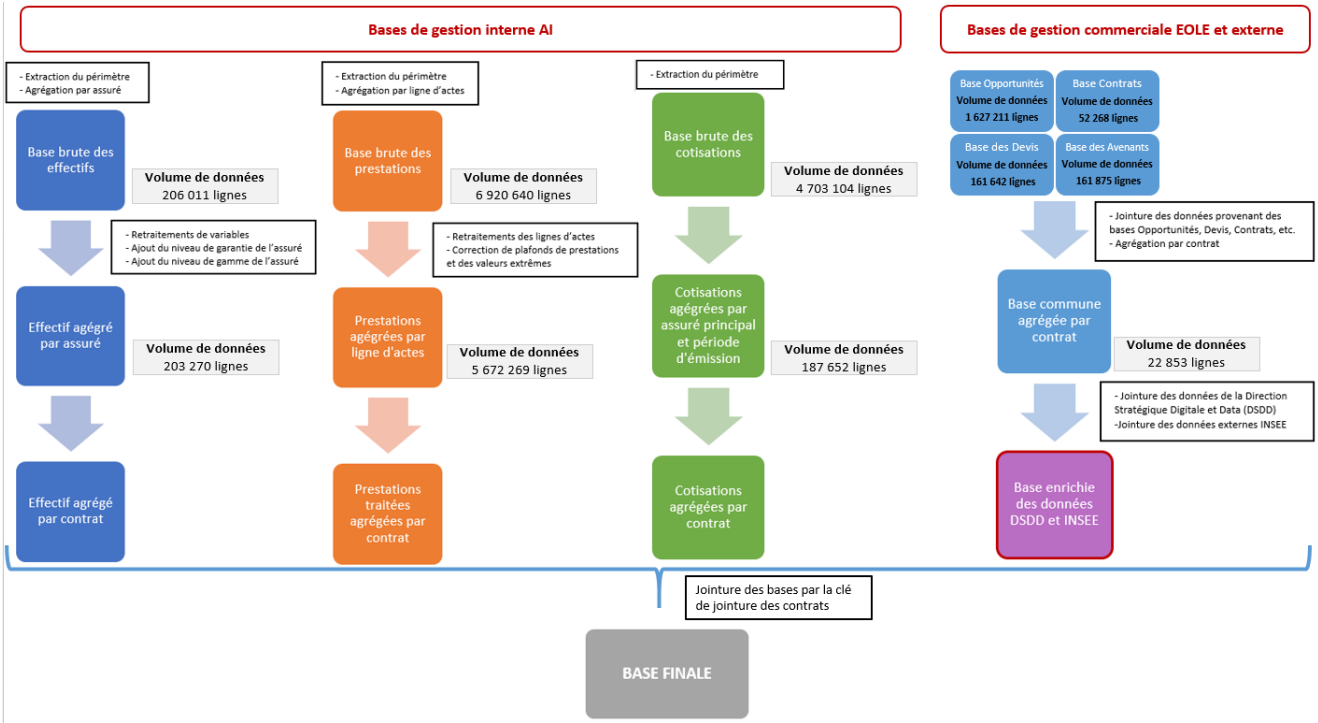


FIGURE 6.1 – Schéma de la construction de la base finale

Les variables de la base finale caractérisant le contrat sont présentées ci-dessous :

Variable	Description
ANN_EXO	Année
Nom_compte	Nom de l'entreprise
NUM_DAC	Numéro de contrat commercial
Gamme_DT	Produit : Santé Entreprise, Syntec, EspritCo, Fedesap&Aidedom, Autre
DT_DB_CTR	Date de début de contrat commercial
age_entreprise	Age de l'entreprise
ancienneté_LMG_entreprise	Ancienneté de l'entreprise chez LMG
ancienneté_LMG_contrat	Ancienneté du contrat chez LMG
COLLEGE	Collège : Cadre, Non-Cadre, Ensemble du personnel, TNS
struct_cot_ret	Structure de cotisation : Isolé|Couple|Famille, Isolé|Famille, Salarié|Conjoint|Enfant, Salarié|Salarié+Enfants|Conjoint, Uniforme
TOTAL_CHARGEMENT	Taux de chargement
Region_DT	Région globalisée
Zone_STD	Zonier standard de consommation
ape_domaine	Domaine APE : Agriculture, Commerce, Construction, Industrie, Service
niv_gamme	Niveau de gamme : Entrée, Milieu, Haut
Densite_nb_med_tot	Dénsité des professionnels de santé du département
densite_nb_med_gen	
densite_nb_med_spe	
densite_nb_dentiste	
densite_nb_infirmier	
densite_nb_pharma	
eff_assure	Effectif assuré
expo_assure	Exposition assuré
eff_salarie	Effectif salarié
expo_salarie	Exposition salarié
repart_salarie_h	Part de salariés hommes sur la base des salariés
repart_salarie_f	Part de salariées femmes sur la base des salariés
age_moy_expo_salarie	Age moyen des salariés
repart_conj	Part des conjoint sur la base des assurés
pctage_conj_par_salarie	Pourcentage moyen de conjoint assuré par salarié
age_moy_expo_conj	Age moyen des conjoints
nb_enf_par_salarie	Nombre moyen d'enfants assuré par salarié
CA_brut	Chiffre d'affaires brut de taxes et chargements
CA_PC	Chiffre d'affaires net de taxes, chargements et du forfait patientèle
PRESTATIONS_PC	Prestations y compris taxe covid
PC_NET	P/C net

FIGURE 6.2 – Variables de la table finale

Nous avons calculé à cet effet des indices démographiques du contrat, tels que la part de salariées femmes, le nombre moyen d'enfants assurés par salariés, ou encore le pourcentage moyen de conjoints assurés. La variable réponse P/C net est alors construite à partir du ration des prestations à cotisations du contrat.

Quatrième partie

Analyses descriptives

Chapitre 7

Variables et P/C

Avant de modéliser nos données, nous allons nous intéresser davantage aux variables de nos bases, et leur éventuel relation avec le P/C..

7.1 Produit

Nous modélisons ci-dessous la répartition des adhérents par produit au fil des années.

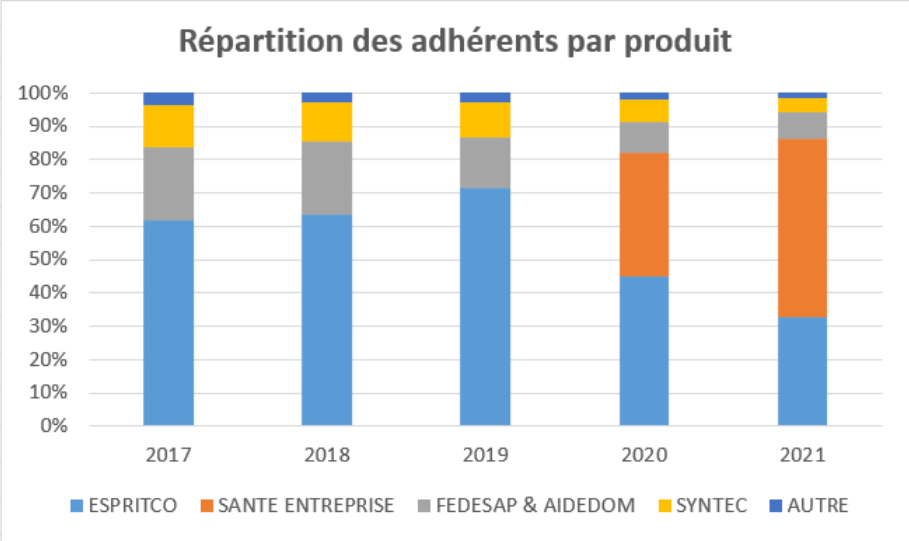


FIGURE 7.1 – Analyse descriptive des produits

Entre 2017 et 2019, nous observons une proportion grandissante d'adhérents au sein du produit ESPRITCO, et une diminution progressive d'adhérents aux produits FEDESAP&AIDEDOM et SYNTEC. À partir de 2020, le nouveau produit SE est commercialisé en masse et est souscrit par 31% des adhérents en 2020 et 46% en 2021.

7.2 Sexe

Nous modélisons ci-dessous la répartition des assurés principaux par sexe au fil des années.

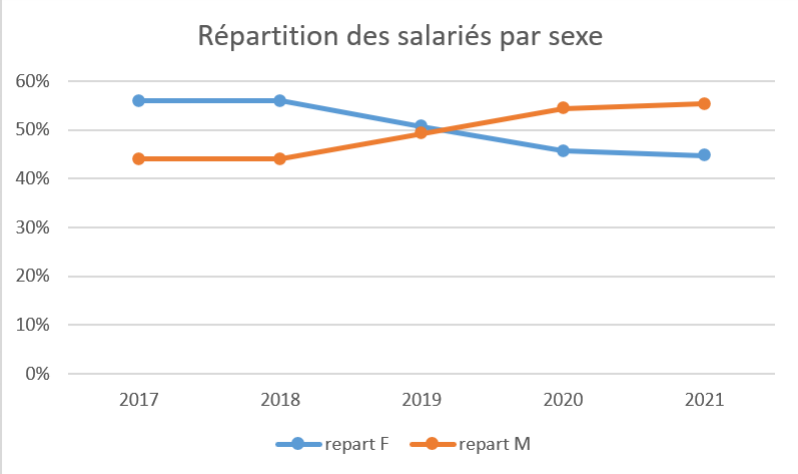


FIGURE 7.2 – Analyse descriptive de la variable sexe

Nous observons entre 2017 et 2021 un inversement de la population majoritaire chez les assurés principaux : les femmes sont majoritaires en 2017 avec près de 56% des assurés principaux, la population de femme décroît ensuite progressivement pour laisser place à une majorité d'homme en 2020, et atteignant 55% en 2021. Cette inversion de tendance s'explique par les produits proposés par LMG au fil de ces 5 années. Les entreprises souscrivant à FEDESAPAIDEDOM ont une grande population de femmes. Avec l'arrivée de SE en 2020, les entreprises visées travaillent dans les secteurs de l'ingénierie, études techniques, ou encore du conseil ; présentant une grande population d'hommes.

7.3 Type assuré

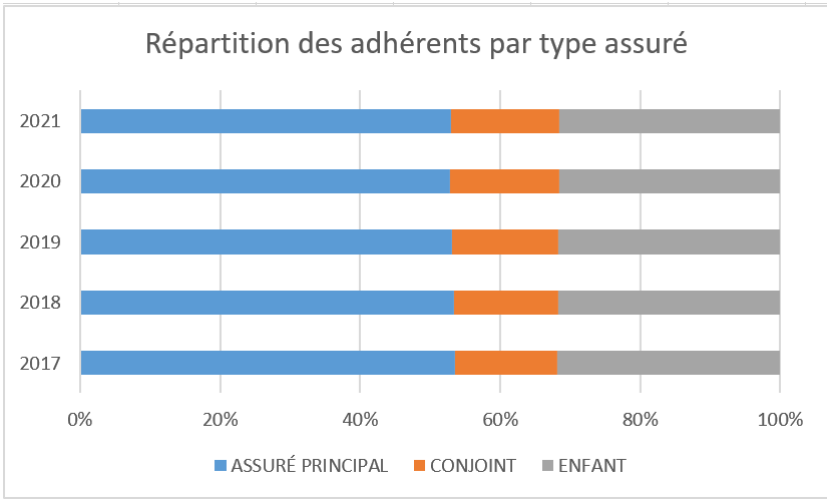


FIGURE 7.3 – Analyse descriptive du type assuré

Les types assurés se répartissent en 3 groupes : assuré principal, conjoint et enfant. Cette répartition est semblable au fil des 5 années d'observation, avec 53% des assurés princi-

paux, 32% d'enfants, et 15% de conjoints. Nous observons un plus grand nombre d'enfants assurés que de conjoints, cela peut s'expliquer par le fait que les conjoints ont la possibilité de souscrire à la complémentaire santé de leurs entreprises en tant que salarié.

7.4 Tranche d'âge moyen des salariés

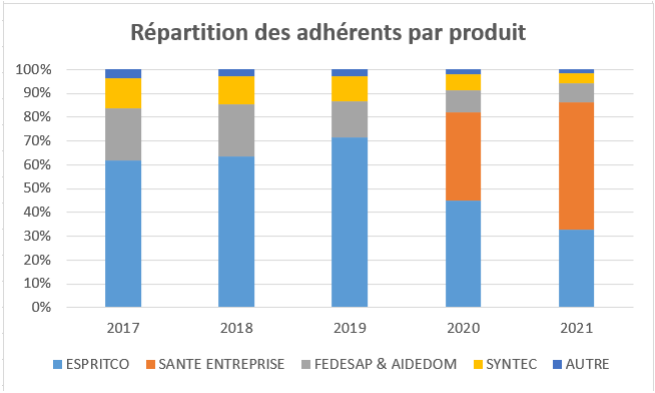


FIGURE 7.4 – Analyse descriptive des tranches d'âge moyen des salariés

Nous remarquons que l'âge moyen des salariés se distribue normalement autour de la tranche d'âge 41-45 ans. En effet, l'âge moyen des assurés principaux du portefeuille se positionne globalement autour de 42 ans sur les 5 années d'étude.

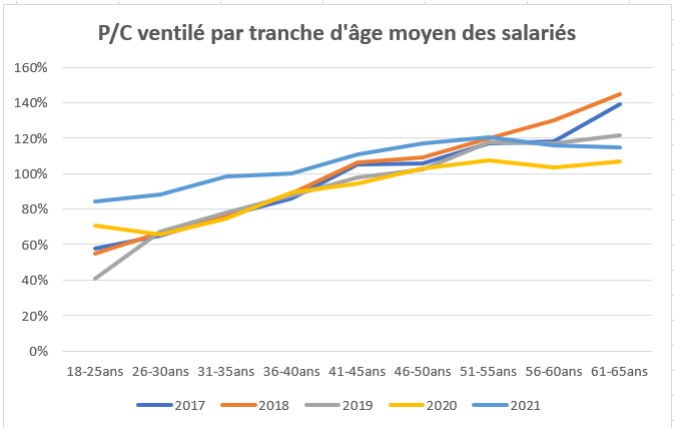


FIGURE 7.5 – P/C ventilé par tranche d'âge moyen des salariés

Le graphique ci-dessous met en lumière la dérive du P/C d'un contrat en fonction de l'âge moyen des salariés : plus les salariés ont un âge moyen élevé, plus ils consomment davantage d'actes médicaux.

7.5 Démographie du portefeuille

Le tableau ci-dessous illustre les données démographiques du portefeuille ventilé sur les 5 années d'observations. Nous remarquons une augmentation du nombre d'entreprises clientes LMG et de ce fait des assurés (appelés membres participants) au fil des années, dont une hausse de 42% entre 2019 et 2020, avec la commercialisation du nouveau produit Santé Entreprise. Nous observons également une légère hausse du pourcentage des conjoints au sein des assurés, passant de 28% à 31% entre 2017 et 2021. L'âge moyen de 42 ans des salariés ainsi que le nombre moyen de 0,6 enfant par salarié restent quant à eux constants sur les 5 années. Comme discuté plus haut, le portefeuille laisse davantage de place aux hommes assurés jusqu'à 56% en 2021.

Démographie	2017	2018	2019	2020	2021
Nb entreprise	2 876	3 439	3 674	4 406	5 063
Membre participant	9 917	13 330	15 235	21 685	25 519
Age moyen	42	42	42	42	42
% conjoint	*28%*	*29%*	*30%*	*31%*	*31%*
Nb enfant moyen	0,6	0,6	0,6	0,6	0,6
% Hommes	*45%*	*45%*	*51%*	*55%*	*56%*

FIGURE 7.6 – Indices démographiques par année du portefeuille étudié

7.6 Répartition des entreprises selon l'effectif salarié

Nous constatons que la majorité du portefeuille est constitué de TPE de moins de 5 salariés, et que la proportion de PME au sein du portefeuille se développe au fil des années.

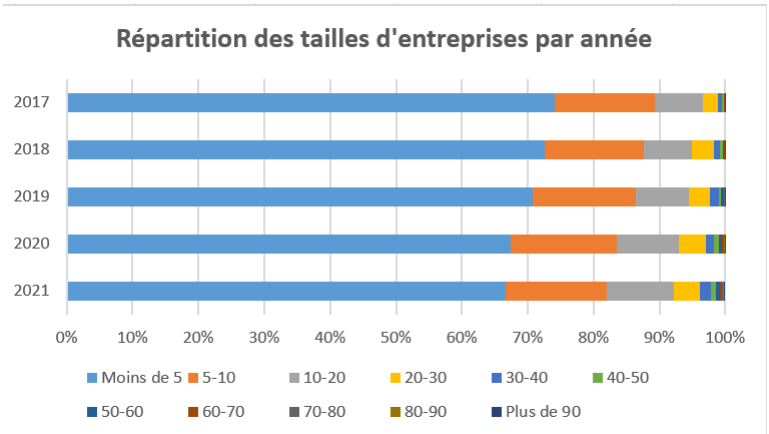


FIGURE 7.7 – Analyse descriptive de la taille d'entreprises

7.7 Collège ou CSP tarifé

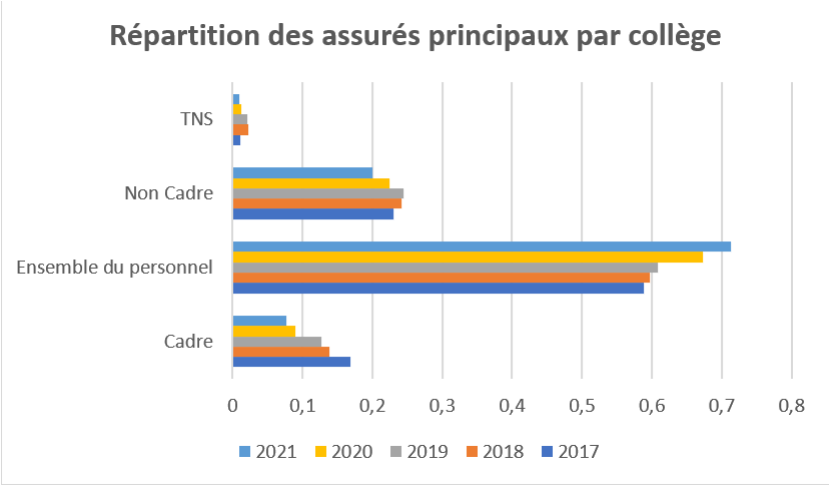


FIGURE 7.8 – Analyse descriptive du collège

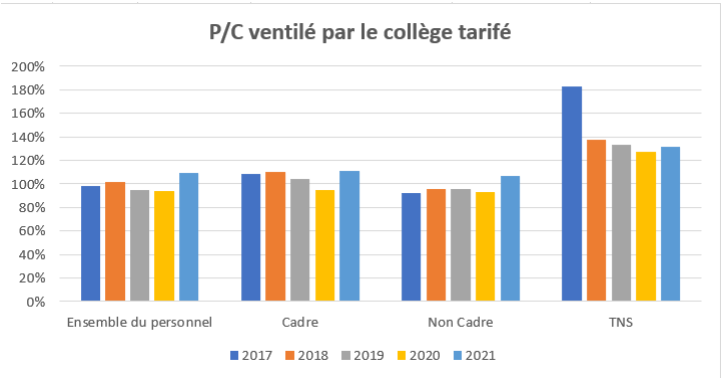


FIGURE 7.9 – P/C ventilé par tranche d'âge moyen des salariés

Sur les 5 années d'observation, plus de la moitié des assurés principaux sont rattachés à des contrats tarifés pour l'ensemble du personnel. Les contrats les plus tarifés sont ensuite les non-cadres, puis les cadres. Enfin, les contrats tarifés pour les TNS sont très minoritaires, car ils ne concernent que les dirigeants des entreprises.

7.8 P/C par année

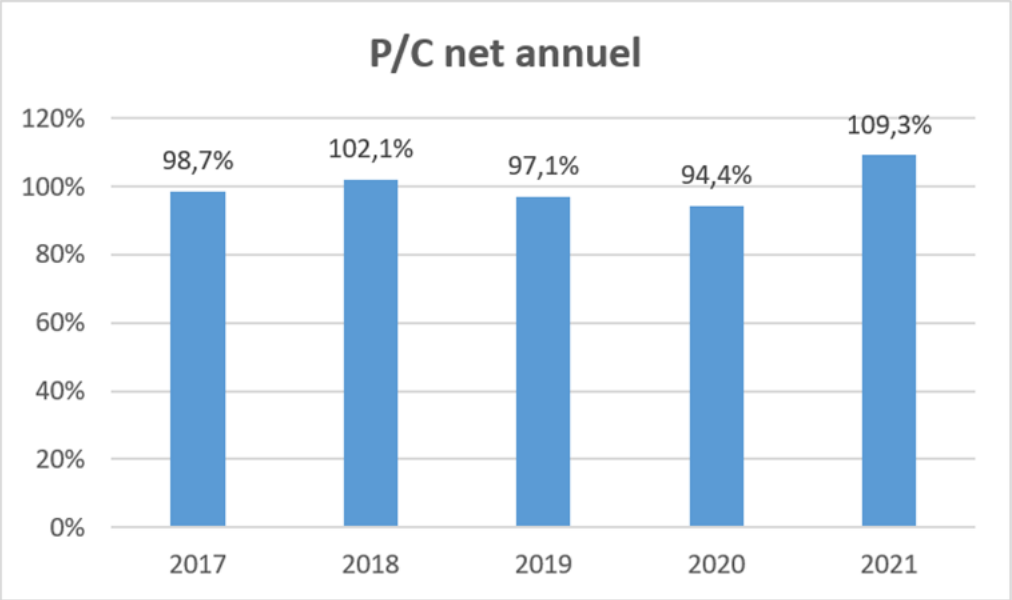


FIGURE 7.10 – Evolution du P/C annuel

Concernant le P/C annuel, nous pouvons observer que celui de 2021 dépasse de loin ceux des années précédentes. Ce ratio élevé se fait d'autant plus remarquer par rapport à celui de 2020 : année atypique dû à la crise sanitaire, le P/C se montre très bas, mais de manière artificielle. Le confinement en 2020 a effectivement entraîné une baisse massive des prestations. À l'inverse, l'année 2021 subit le contrecoup de l'année 2020 : les personnes ayant consommé peu d'actes médicaux en 2020, auront tendance à "rattraper" ces actes qui n'ont pas pu être réalisés en 2020, lors de l'année suivante. Ce facteur peut en partie expliquer le haut ratio du P/C en 2021. Nous pouvons également observer un P/C élevé en 2018 avec une plus grande commercialisation du produit ESPRITCO, puis se stabilise en 2019.

7.9 Structure de cotisation

La structure de cotisation (exemple : "Isolé/Famille") désigne un ensemble de sous structures de cotisation (exemple : Isolé et Famille) choisi par l'employeur et proposé aux salariés. Les salariés ont donc le choix de sélectionner une sous structure de cotisation parmi celles indiquées au sein de la structure de cotisation. La structure de cotisation dite Uniforme est indépendante du nombre d'ayants droit. Ainsi, que le salarié soit seul bénéficiaire ou qu'il le soit avec sa famille, le montant de la cotisation sera le même.

Si nous reprenons l'exemple, selon le choix de la sous structure Isolé ou Famille, le salarié choisit soit de se couvrir seul, soit de couvrir toute sa famille. La consom- en sera fortement impactée. Le challenge de l'actuaire est alors de proposer le tarif adéquat de manière à prendre en compte cette différence de niveau de consommation au sein de la même

structure de cotisation. Une mauvaise tarification engendrerait entre autres un mauvais P/C pour cette structure de cotisation.

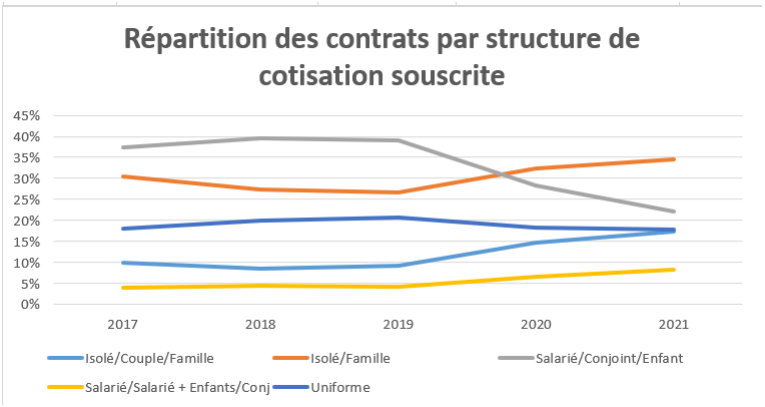


FIGURE 7.11 – Analyse descriptive de la structure de cotisation

Nous observons une évolution des structures de cotisation les plus représentées au sein du portefeuille au fil des années : la proportion des structures Salarié/Conjoint/Enfant et Uniforme diminue à partir de 2020, pour laisser place au développement des structures Isolé/Famille, Isolé/Couple/Famille et Salarié/Salarié+Enfants/Conjoint.

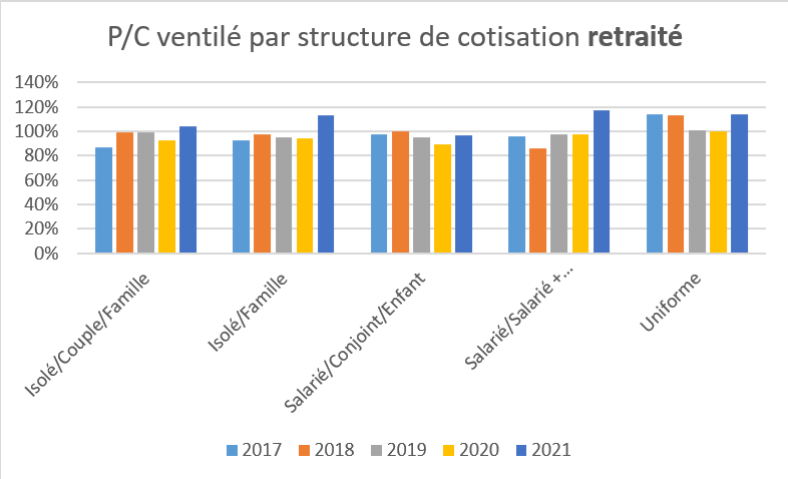


FIGURE 7.12 – P/C ventilé par structure de cotisation

Nous constatons que le P/C par la variable retraitée de la structure de cotisation ventilée par année ne permet pas de ressortir une structure de cotisation très influente dans le P/C, étant donné qu'aucun des P/C (hormis le Non Renseigné) pour chaque structure ne dépasse globalement les 100% sur les 5 années d'observations. Cela traduirait de plus une bonne tarification des produits selon la structure de cotisation. De ce fait, nous pouvons nous demander si une modélisation du P/C en agrégeant la table sans tenir compte des distinctions de structure de cotisation pourrait permettre de construire un modèle plus robuste.

7.10 Zonier de consommation LMG du périmètre standard collectif

Le zonier de consommation du périmètre standard collectif a été construit en 2016 et divise la France métropolitaine en 5 zones selon le département et son niveau de consommation, comme suit :

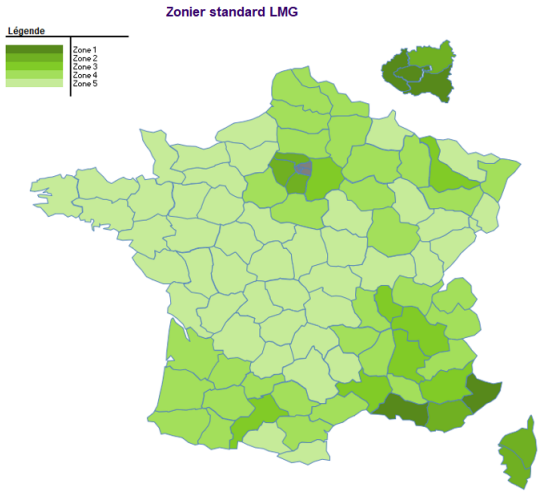


FIGURE 7.13 – Zonier standard sur la carte de France

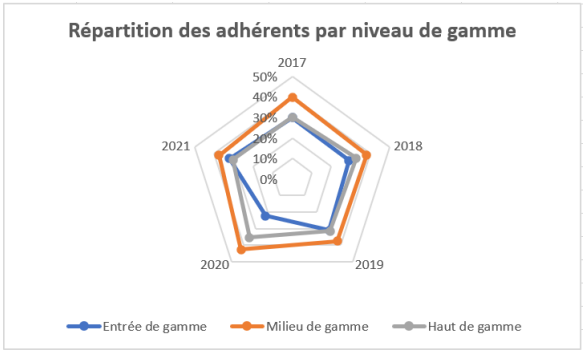


FIGURE 7.14 – Analyse descriptive du zonier standard

Nous constatons que la répartition est semblable entre 2017 et 2021 : les zones les moins consommatrices 4 et 5 concentrent le plus de contrats, suivies de la zone la plus consommatrice 1 puis intermédiaire 3 et 2.

7.11 Région

Les régions utilisées par la Direction Technique de LMG sont les suivantes :

— Auvergne-Rhône Alpes et Bourgogne-Franche-Comté

— Centre et Pays de la Loire

— Grand Est et Hauts-de-France
— Normandie et Bretagne
— Nouvelle Aquitaine
— Occitanie
— Provence-Alpes-Côte d'Azur
— Ile-de-France

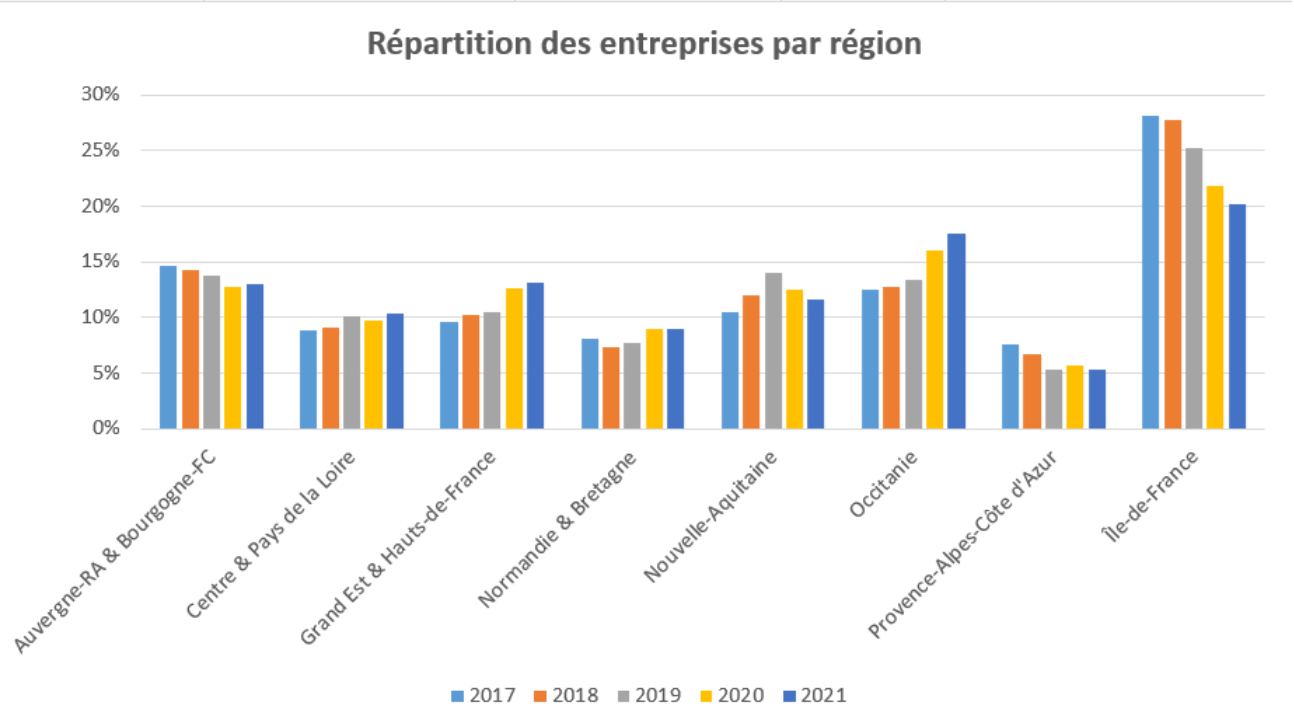
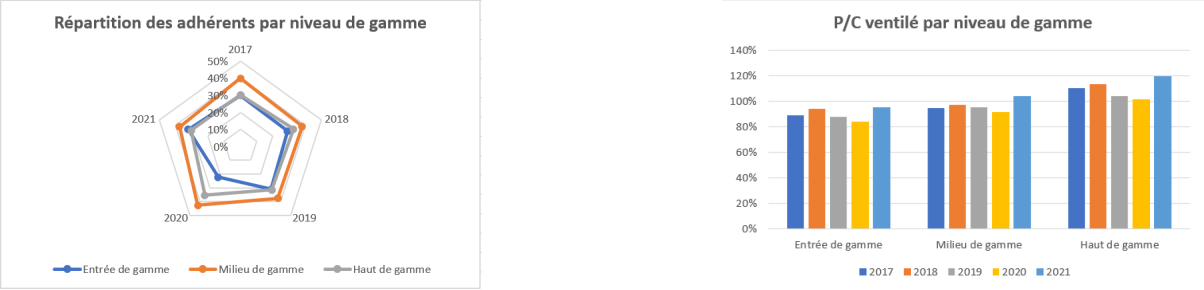


FIGURE 7.15 – Analyse descriptive des régions DT

Nous observons que la région dominante des entreprises présentes dans le portefeuille est l'Île-de-France. Cependant, nous notons que cette proportion diminue au fil des années, et qu'une augmentation des entreprises s'observe en Occitanie et Grand Est/Hauts-de-France.

7.12 Ancienneté du contrat chez LMG

Nous remarquons qu'en 2019, le portefeuille est constitué de plus de 60% de contrats souscrits en 2016, qui correspond à l'année de mise en place de la mutuelle d'entreprise obligatoire. Cela nous indique qu'une grande partie des entreprises sont restées couvertes par LMG et n'ont pas résilié. En 2020, nous retrouvons la commercialisation du produit SE via les 30% de contrats de moins d'1 an d'ancienneté et les 45% de contrats d'1 à 2 ans d'ancienneté en 2021.



(a) (b)

FIGURE 7.17 – Analyse descriptive du niveau de gamme

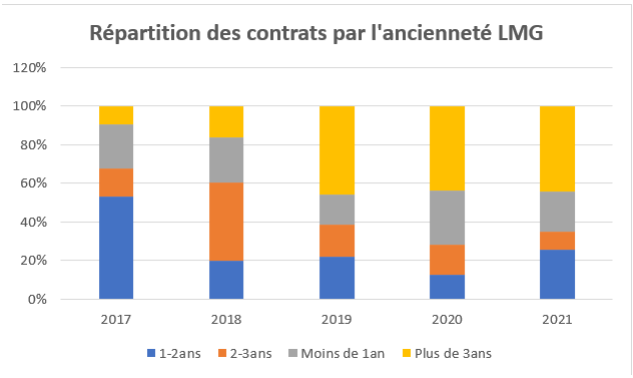


FIGURE 7.16 – Analyse descriptive de l'ancienneté du contrat

7.13 Niveau de gamme du contrat

Les contrats milieu de gamme sont les plus représentés au sein du portefeuille sur les 5 années d'études. Nous observons par ailleurs que les contrats d'entrée de gamme et haut de gamme ont globalement la même répartition entre 2017 et 2021, sauf en 2020 où le haut de gamme est davantage présent. Le graphique du P/C selon le niveau de gamme illustre bien le fait que plus le contrat est considéré haut de gamme, plus le P/C a tendance à se dégrader. Nous nous attendons donc lors de l'étude des GLM à une forte influence de la variable niveau de gamme sur la variable réponse.

7.14 Démographie des contrats

Au sein d'un contrat, le pourcentage de conjoint assuré et le nombre d'enfants moyen par salarié sont des paramètres influençant la tarification de l'offre. Nous pouvons supposer que ces 2 indices démographiques peuvent jouer sur le niveau de rentabilité d'une entreprise. Notre portefeuille se constitue de contrats dont près de la moitié ne couvre aucun conjoint de salarié, et un quart couvre plus de 65% de conjoint. Ensuite, 40% des contrats en 2021 ne couvre aucun enfant.

Les graphiques du P/C ventilé par ces 2 paramètres témoignent de l'importance de ces

(a) (b)

FIGURE 7.18 – Analyse descriptive du nombre moyen de conjoints assurés

derniers sur la rentabilité d'un contrat.

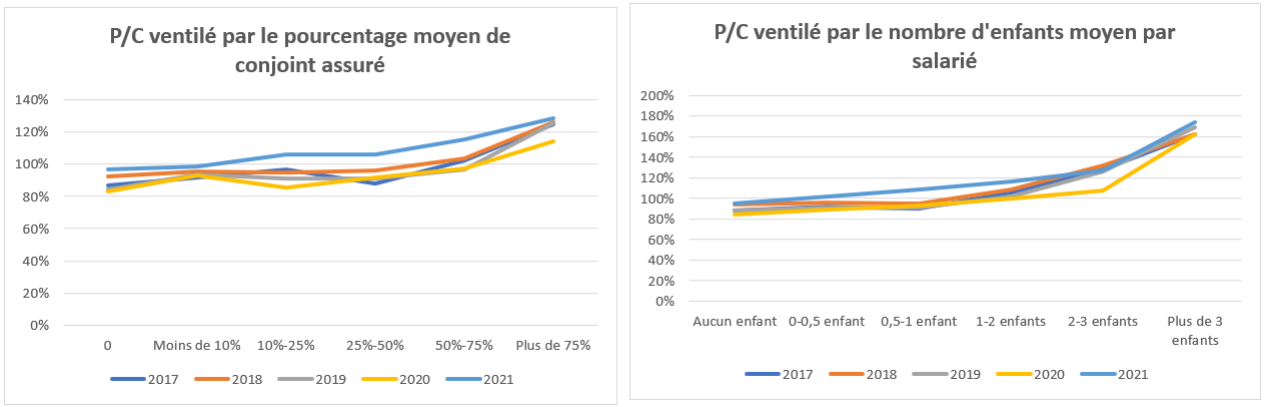


FIGURE 7.19 – P/C ventilé par les indices démographiques sur les conjoints et enfants assurés

Nous trouverons en Annexes des graphiques illustrant des statistiques univariées supplémentaires des variables de l'étude.

Cinquième partie

Méthodologies

Chapitre 8

Théorie derrière les modèles linéaires généralisés

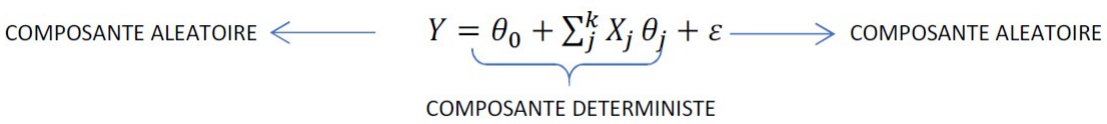
La conception du modèle linéaire classique repose sur le fait qu'une variation constante d'une variable explicative entraîne une variation constante de la variable réponse. Or, en santé et dans de nombreux autres domaines, les phénomènes modélisés peuvent être davantage complexes, et l'influence des variables explicatives peut être non linéaire. Nous avons pu voir dans les analyses descriptives les relations entre le P/C et des indices démographiques n'étaient pas linéaires. La théorie des modèles linéaires généralisés (GLM) a été introduite par Nelder et Wedderburn en 1972. La classe des GLM est une extension des modèles linéaires traditionnels.

8.1 Rappels sur les modèles linéaires

Les modèles linéaires permettent de décrire la relation linéaire qui existe entre une variable aléatoire dite réponse Y et un vecteur déterministe de variables explicatives (ou régresseurs) $X = (X_1, X_2, X_3, \ldots, X_k)$. Nous avons :

$$Y = X\Theta + \epsilon$$

Où Θ est le vecteur des paramètres inconnus du modèle, et ϵ est un vecteur aléatoire contenant les erreurs du modèle. Le modèle peut s'écrire sous la forme suivante :



Où :

— Y est la variable aléatoire réponse. Il s'agit de la variable que nous observons et que nous souhaitons expliquer ;

— $(X_1, X_2, X_3, \ldots, X_k)$ sont les variables explicatives, celles-ci sont déterministes et observées ;

— $(\theta_1, \theta_2, \theta_3, \ldots, \theta_k)$ sont les coefficients du modèle, ils ne sont pas observés et doivent donc être estimés ;

— ϵ est une variable aléatoire non observée, appelée résidus, et traduit toute l'information non modélisée, i.e. la variation dans la variable réponse qui reste inexpliquée par les variables explicatives. Les résidus représentent ainsi l'écart entre la valeur observée et la valeur estimée par le modèle. En particulier, les résidus satisfont les hypothèses Gauss-Markov :

  — Les erreurs ont une espérance nulle : $\mathbb{E}(\epsilon_i) = 0$

  — Homoscédasticité : les erreurs ont la même variance (inconnue) σ^2 : $\mathbb{V}(\epsilon_i) = \sigma^2$

  — Les résidus sont non corrélés : $Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$

  Finalement, les résidus sont homogènes, indépendants, et suivent une loi normale de moyenne nulle et de variance résiduelles. Matriciellement, nous obtenons $\mathbb{E}(\epsilon) = 0$ et $\mathbb{V}(\epsilon) = \sigma^2 I_k$ où In est la matrice identité k x k

Ces hypothèses impliquent que la variable réponse doit être approximativement distribuée selon la loi normale, avec une variance égale à celle des résidus, et une espérance s'exprimant comme une combinaison linéaire des variables explicatives : $(formule(a))$

$$\mathbb{E}(Y) = \sum_{j=1}^{p} \beta_j X^{\mathrm{j}}$$

$$\mathbb{V}(Y) = \sigma^2$$

Autrement dit, la variable réponse moyenne est expliquée en sommant les effets des variables explicatives.

8.2 Modèles linéaires généralisés

Les GLM permettent d'étendre les idées de modélisation linéaire à une classe plus large de type de variable de réponse, dont la méthodologie de modélisation reste commune. Cette classe de modèles linéaires s'affranchit de la contrainte de normalité des résidus. Contrairement aux modèles linéaires standards qui cherchent à modéliser la variable réponse, les GLM modélisent une fonction de l'espérance de cette variable, appelée fonction de lien. Les 3 éléments fondamentaux sur lesquels reposent les GLM sont :

— Le prédicteur linéaire ;

— La fonction de lien ;

— La structure des erreurs.

8.2.1 Le prédicteur linéaire

Le prédicteur linéaire désigne la combinaison linéaire des prédicteurs (variables explicatives) qui permettent d'expliquer en partie la variable réponse (ne prend pas en compte

les erreurs). Le prédicteur linéaire est ainsi déterminé par $\eta = X\beta$, où :

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}_{n\times 1} \qquad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n\times p} \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p\times 1}$$

8.2.2 La fonction de lien

Si nous posons la moyenne μ_i comme l'espérance d'une réponse observée $\mu_i = \mathbb{E}(Y_i)$, alors la fonction de lien correspond à la transformation mathématique qui est appliquée à la moyenne, afin d'obtenir le prédicteur linéaire de cette observation. Autrement dit, dans le cadre du GLM, les valeurs du prédicteur linéaire s'obtiennent en transformant au préalable les valeurs observées par la fonction de lien, contrairement aux modèles linéaires où les valeurs du prédicteur linéaire correspondent directement aux valeurs observées moyennes (cf $formule(a)$) :

$$\eta_i = g(\mu_i) \qquad i = 1, \dots, n$$

Avec $\mu_i = \mathbb{E}(Y_i)$; $i = 1, 2, \dots, n$. L'objectif de la fonction de lien est d'obliger les valeurs prédites à avoir des contraintes identiques aux valeurs observées, notamment en termes d'échelle, afin de fournir des prédictions cohérentes.

8.2.3 La structure d'erreur

La structure d'erreurs désigne la famille de distribution des résidus, et se distingue généralement pour chaque fonction de lien. Elle permet notamment de mettre en relation de façon claire la moyenne et la variance.

8.2.4 Le type de réponse

La nature de la variable réponse permet de sélectionner de manière adéquate la fonction de lien ainsi que la structure d'erreur inhérente au GLM. En effet, lorsque les réponses sont de type comptage (données entières positives et bornées en 0), alors la variance des résidus augmente de façon proportionnelle avec les réponses prédites (les données de comptage et leurs résidus suivent théoriquement une distribution de Poisson de paramètre Lambda, donc la variance des résidus n'est pas constante, mais égale à la moyenne elle-même égale a lambda). La fonction de lien utilisée est alors logarithmique, et la structure d'erreur est poissonnienne. On parle de régression de Poisson. Si les réponses sont binaires ou représentent des proportions, alors les données suivent une loi Binomiale de paramètres n et p ; la structure d'erreur est binomiale et la fonction de lien utilisée est logistique.

Type des réponses (et des erreurs)	Domaine de définition des réponses	Distribution des erreurs (et des réponses)	Nom de la fonction de lien	Fonction de Lien	Fonction de la moyenne	Fonction de la variance
Quantitatif continu	Réel $]-\infty ; +\infty[$	Gaussienne	Identité	$\sum_{j=1}^{p} x_{ij}\beta_j = \mu$	$\mu = \sum_{j=1}^{p} x_{ij}\beta_j$	$var(y_i) = cste$
Comptage	Entier $[0 ; +\infty[$	Poisson	Log	$\sum_{j=1}^{p} x_{ij}\beta_j = \ln(\mu)$	$\mu = exp\left(\sum_{j=1}^{p} x_{ij}\,\beta_j\right)$	$var(y_i) = \propto \mu$
Binaire (oui/non)	Entier $[0 ;1]$	Binomiale	Logit	$\sum_{j=1}^{p} x_{ij}\beta_j = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{1}{1+\exp(-\sum_{j=1}^{p} x_{ij}\beta_j}$	$var(y_i) = \propto \frac{\mu(n-\mu)}{n}$

Ainsi, un modèle linéaire classique a une fonction de lien « identité » et une structure d'erreur gaussienne.

Chapitre 9

Méthodologies pour la modélisation de la rentabilité

9.1 Choix de la loi modélisant le P/C

9.1.1 Observations graphiques

Grâce à la fonction fitdist sur R, nous pouvons sélectionner une famille de distribution suivant laquelle nous souhaitons modéliser les données empiriques de notre variable réponse ; avec notamment l'estimation des paramètres de la loi, l'histogramme de la densité et de la fonction de répartition de la variable réponse par la loi théorique.

9.1.2 Test de Kolmogorov Smirnov

Le test d'hypothèse de Kolmogorov Smirnov permet de déterminer si un échantillon suit bien une loi connue par sa fonction de répartition, ou bien si deux échantillons suivent la même loi. Les hypothèses de ce test sont :

— $H_0 : F = F_0$; la répartition des observations (notée F) s'intègre bien dans une distribution donnée (notée F_0)

— $H_1 : F \neq F_0$; la répartition des observations ne s'intègre pas bien dans une distribution donnée.

Le test de Kolmogorov Smirnov consiste ainsi à mesurer l'écart maximal existant entre la fonction de densité cumulée observée empirique et celle connue théorique (ou inconnue sous forme analytique). Cette distance séparant les deux fonctions est testée pour déterminer si l'écart est considéré importante compte tenu de l'effectif. Pour un échantillon de X de taille n $(X_1, X_2, \ldots, X_n)$, la distance Δ_n est déterminée par :

$$\Delta_n = sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

Où $F_n(x)$ est la fonction de répartition empirique, elle représente la part des observations pour lesquelles la valeur est inférieure ou égale à x.

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} I_{]-\infty, x]}(X_k)$$

Si nous posons $U_n = \sqrt{n}\Delta_n$, alors nous pouvons démontrer que lorsque H_0 n'est pas vraie, U_n tend vers $+\infty$, et inversement U_n suit asymptotiquement une loi sur $\mathbb{R}^+$ définie par sa fonction de répartition lorsque H_0 est vraie. La région critique du test est ainsi constituée des grandes valeurs de Δ_n. Le test étant effectué au niveau α qui est donné, nous pouvons définir la valeur critique selon 2 manières :

— La loi de Δ_n étant tabulée, son fractile $C_n(1-\alpha)$ d'orde $1-\alpha$ fournit un test exact de niveau α en rejetant l'hypothèse H_0 si la valeur observée δ_n de Δ_n dépasse $C_n(1-\alpha)$, et en l'acceptant le cas contraire. Ce test est valable pour toute taille n de l'échantillon.

— La loi limite étant tabulée, son fractile $Cn(1-\alpha)$ d'orde $1-\alpha$ fournit un test asymptotique de niveau α en rejetant l'hypothèse H_0 si la valeur observée δ_n de Δ_n dépasse $C_n(1-\alpha)$, et en l'acceptant le cas contraire.

Nous pouvons noter le test de Cramer-Von Mises, qui est aussi un test statistique utilisé pour évaluer la qualité de l'ajustement d'une loi continue, qui cherche également à vérifier l'adéquation suivante, sous l'hypothèse nulle :

$$H_0 : F = F_n$$

avec F la fonction de répartition théorique de la loi à tester et F_n la fonction de répartition empirique de nos données observées.

9.2 Estimation des paramètres du GLM

L'estimation des paramètres du GLM se fait par « maximum de vraisemblance » dont nous rappelons ici la méthode. Rappelons tout d'abord que la vraisemblance est, par définition, un produit de fonctions de densité. Pour en déterminer le maximum, il suffit de déterminer la valeur du paramètre de la fonction qui l'annule tout en gardant la dérivée seconde négative. Pour des raisons pratiques, on préfère dériver le logarithme de la vraisemblance. On dérive ainsi une somme plutôt qu'un produit.

De plus, comme la fonction logarithme est strictement croissante, maximiser le logarithme de la fonction équivaut à maximiser la fonction. Considérons p variables explicatives dont les observations sont rangées dans la matrice de plan d'expérience X, β un vecteur de p paramètres et le prédicteur linéaire à n composant

$$\eta = X\beta$$

La fonction canonique g est supposée monotone différentiable telle que : $g(\eta_i) = \theta_i$

Pour n observations supposées indépendantes et en tenant compte que θ dépend de Z, la log vraisemblance s'écrit :

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \ln f(y_i, \theta_i, \varphi) = \sum_{i=1}^{n} \ell(y_i, \theta_i, \varphi)$$

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{[y_i - b'(\theta_i)]}{a(\varphi)} = \frac{[y_i - \mu_i]}{a(\varphi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = b''(\theta_i) = \frac{Var(Y)}{a(\varphi)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} \text{ dépend de la fonction lien } \eta_i = g(\mu_i)$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \text{ car } \eta_i = x_i'\beta$$

Les équations de la vraisemblance sont :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \qquad j = 1, 2, \ldots, p$$

Ce sont des équations non-linéaires en Z dont la résolution requise des méthodes itératives dans lesquelles interviennent le Hessien (pour Newton-Raphson) ou le matrice d'information (pour les Scores de Fischer). La matrice d'information est la matrice : $\mathcal{F} = X'WX$

De terme général

$$[\mathcal{F}]_{jk} = \mathbb{E}\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Où W est la matrice diagonale de « pondération »

$$[W]_{jk} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Dans le cas particulier où la fonction lien du modèle linéaire généralisé utilisée est la fonction lien canonique associée à la structure exponentielle alors plusieurs simplifications interviennent :

$$\eta_i = \theta_i = x_i^{'}\beta$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b^{'}(\theta_i)}{\partial \theta_i} = b^{''}(\theta_i)$$

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{Var(Y_i)} b^{''}(\theta_i) x_{ij} = \frac{(y_i - \mu_i)}{a(\varphi)} x_{ij}$$

De plus, comme les termes $\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k}$ ne dépendent plus de yi, on montre que le Hessien est égal à la matrice d'information et donc les méthodes de résolution du score de Ficher de Newton-Raphson coïncident. Si, de plus, $a(\Phi)$ est constante pour les observations, les équations de vraisemblance deviennent :

$$X'y = X'\mu$$

Ainsi, dans le cas gaussien, le modèle décrivant $\mu = X\beta$ avec la fonction de lien canonique identité, on retrouve la solution :

$$\beta = (X'X)^{-1}X'y$$

Qui coïncide avec celle obtenue par minimisations des moindres carrées. Les estimations des $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ nous seront données par la suite grâce à la procédure GENMOD sur SAS et glm sur R ;

9.3 Sélection des variables

Notre problème considère une variable à expliquer et un ensemble de n variables explicatives. Il s'agit de déterminer un sous-ensemble de l'ensemble des variables explicatives réalisant un compromis entre le souhait que le modèle sélectionné contienne peu de paramètres et le souhait que ce modèle bénéficie d'un pouvoir explicatif suffisant.

Plus un modèle contient de variables explicatives, plus il est précis, mais moins il est robuste. À l'inverse, moins un modèle contient de variables explicatives, plus il est robuste, mais moins il est précis. En effet, l'ajout d'une nouvelle variable explicative apporte des informations supplémentaires sur la variable à expliquer, mais impose de fait une contrainte supplémentaire au modèle.

Dans un GLM, il est donc essentiel de déterminer la combinaison de variables explicatives qui permettra d'obtenir le meilleur compromis entre précision et robustesse.

9.3.1 Mesure de la corrélation : V de Cramer

Le test du Khi-deux ne nous donne pas d'information sur la force de liaison entre les variables qualitatives. Pour mesurer cette force, nous utilisons le V de cramer qui permet d'indiquer les liaisons plus ou moins fortes entre deux variables, à partir d'un tableau de contingence déterminé. Cette mesure de corrélation est très utilisée en pratique puisqu'elle ne dépend pas de la taille de l'échantillon contrairement au test du Khi-deux.

Règle de décision :

— si le V de Cramer est proche de 1 : forte liaison des variables ;

— si le V de Cramer est proche de 0 : faible liaison des variables.

9.3.2 Critères de sélection du modèle

Résidus de Pearson

Les résidus obtenus en comparant valeurs observées yi et valeurs prédites $\hat{y}_i$ sont pondérés par leur précision estimée par l'écart-type : $\hat{\sigma}_i$ de $\hat{y}_i$. Ceci définit les résidus de Pearson :

$$r_i^p = \frac{y_i - \hat{y}_i}{\hat{\sigma}_i}$$

Ces résidus mesurent donc la contribution de chaque observation à la significativité du test découlant de cette statistique. Ces résidus ne sont pas de variance unité et sont donc difficiles à interpréter. Une estimation de leurs écarts-types conduit à la définition des résidus de Pearson Standardisés :

$$r_i^{p_s} = \frac{y_i - \hat{y}_i}{\sigma_i \sqrt{h_{ii}}}$$

Faisant intervenir le terme diagonal de la matrice H (hat matrix), qui est construit :

$$H = W^{\frac{1}{2}} X (X'WX)^{-1} X') W^{\frac{1}{2}}$$

Relative au produit scalaire de matrice W, sur le sous-espace engendré par les variables explicatives. Les termes diagonaux de cette matrice supérieure à $\frac{3p}{n}$ indiquent des valeurs potentiellement influentes.

Résidus de Déviance

Les résidus de Déviance mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. Les résidus de déviance sont définis par :

$$d_i = signe(y_i - \hat{y}_i) \sqrt{2(\mathcal{L}(y, y) - \mathcal{L}(y, \hat{y}))}$$

Où : $\mathcal{L}(y, y)$: fonction de log-vraisemblance du modèle saturé, c'est-à-dire le modèle possédant autant de paramètre que d'observations et estimant donc exactement les données. $\mathcal{L}(y, \hat{y})$: fonction de log-vraisemblance du modèle estimé.

Voici la définition des résidus de Déviance Standardisés :

$$d_i^s = signe(y_i - \hat{y}_i) \sqrt{\frac{2(\mathcal{L}(y, y) - \mathcal{L}(y, \hat{y}))}{1 - h_{ii}}}$$

Ainsi, plus les résidus sont proches de 0 (ou résidus standardisés proches de 1), mieux le modèle estime les observations.

La variable introduite dans le modèle est celle qui entraîne la plus grande diminution des résidus ou qui amène les résidus standardisés proches de la valeur 1.

Nous allons présenter ci-dessous 2 méthodes principales servies à la sélection de variables :

Forward

Dans la méthode « forward », il s'agit de démarrer avec le modèle contenant une seule variable, nous cherchons ensuite la variable qui, associée à la première, engendre la plus grande perte de résidus du modèle.

Backward

Dans la méthode « backward », il s'agit de démarrer avec le modèle complet (c'est-à-dire toutes les variables ayant un effet significatif sur le risque) puis de retirer la variable la moins probante, autrement dit celle dont l'élimination engendre peu d'impact sur la diminution de résidus du modèle.

L'introduction des variables est stoppée à partir du moment où leur effet sur le modèle n'est plus significatif, où la perte en résidus est jugée négligeable.

En pratique, le logiciel SAS fournit les déviances de vraisemblance par l'analyse « type 1 » et l'analyse « type 3 », fréquemment utilisées pour tester la pertinente des variables explicatives.

9.3.3 Analyse du type 1 et type 3 sous SAS

Type 1

Effectuer les calculs pour les modèles comportant uniquement la variable 1, puis la variable 1 et la variable 2,... en suivant l'ordre indiqué dans la saisie du modèle. À chaque étape, elle fournit « -2x le logarithme du rapport de vraisemblance » correspondant aux hypothèses suivantes :

— H_0 : le modèle contient les variables 1 à j avec j compris entre 1 et le nombre de variables explicatives.

— H_1 : le modèle contient les variables 1 à $j+1$ et les paramètres associés à la variable $j+1$ qui ne sont pas tous nuls.

Dans l'analyse de type 1, on tient compte de l'ordre dans lequel les variables entrent dans le modèle. Pour supprimer la contrainte sur l'ordre d'entrée des variables, on recourt à l'analyse de type 3.

Type 3

Pour chaque variable j spécifiée dans le modèle, l'analyse du « type 3 » est effectuée sous 2 hypothèses suivantes :

— H_0 : le modèle contient toutes les variables sauf la variable j

— H_1 : le modèle contient toutes les variables et les paramètres associés à j ne sont pas tous nuls.

Les deux analyses « type 1 » et « type 3 » fournissent la p-value associée au test du rapport des vraisemblances pour un niveau de confiance de 95% par défaut. On accepte les variables qui donnent une p-value inférieure à 5%. A contrario, nous éliminons les variables dont leurs p-values sont supérieures à 5%.

9.3.4 Validation du modèle

Validation graphique

La validation du modèle est jugée bonne si les résidus observés se situent autour de l'axe des abscisses (proches de 0) et avec une variance constante, autrement dit si le nuage de points est de forme cylindrique autour de l'axe des abscisses.

Sixième partie

Résultats

Chapitre 10

Distribution du P/C

La partie Résultats se destine à appliquer des GLM afin d'établir la modélisation la plus pertinente valide du P/C et de déterminer les critères influençant la rentabilité d'un contrat. Pour ce faire, nous allons appliquer la théorie explicitée dans le paragraphe précédent.

Dans un premier temps, nous présenterons la démarche permettant de choisir la distribution selon laquelle le P/C sera modélisée. Ensuite, nous étudierons la dépendance et la corrélation entre les variables quantitatives et qualitatives de la base finale pour une meilleure sélection des variables explicatives dans le cadre du GLM. Nous exposerons par la suite la recherche du modèle le plus robuste et précis. Enfin, nous tenterons de valider le modèle choisi.

L'objectif de cette partie est de sélectionner les lois usuelles possibles modélisant au mieux nos données observées, afin de modéliser GLM du P/C selon ces distributions.

Le graphique ci-dessous représente tout d'abord l'histogramme de la variable pc_net, qui n'a pas été retraitée. Cette dernière contient des valeurs extrêmes qui peuvent impacter de manière notoire les estimateurs lors de l'ajustement d'une distribution type. L'unité de la variable est le %.

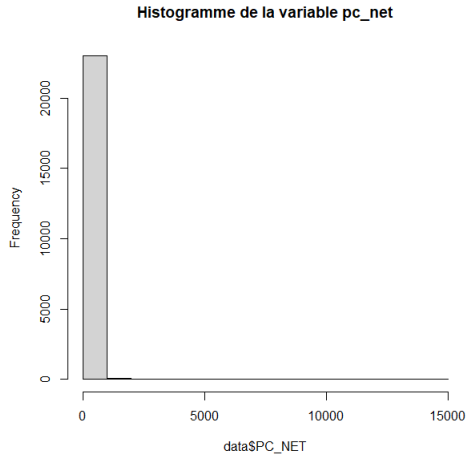


FIGURE 10.1 – Histogramme du P/C

L'histogramme met en lumière la présence de P/C très élevés, empêchant une bonne lecture des P/C « typiques » de notre base.

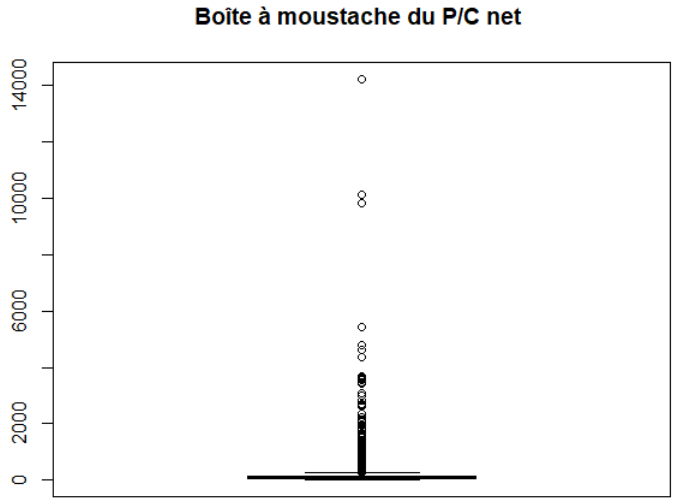


FIGURE 10.2 – Boxplot du P/C

Le boxplot du P/C net montre une grande quantité de valeurs observées extrêmes.

Nous décidons d'étudier les P/C supérieurs à 1000%. Ils représentent 0,3% de notre base.

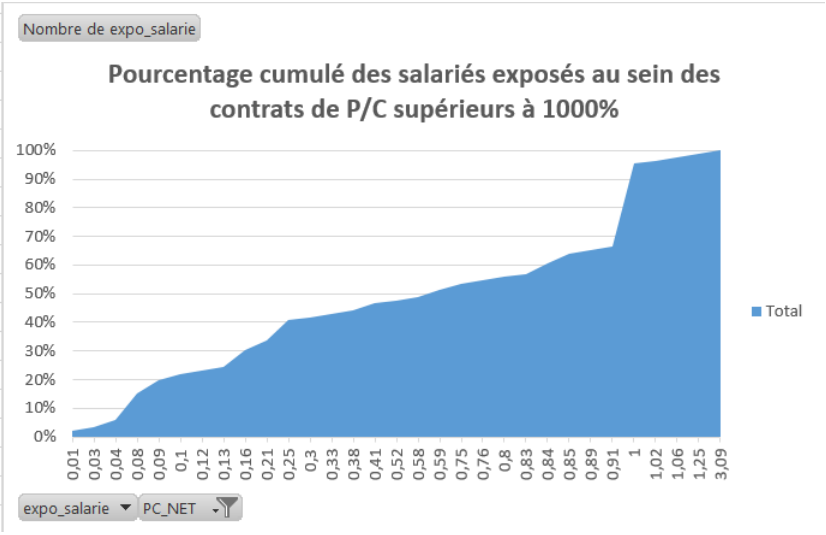


FIGURE 10.3 – Étude des P/C nuls en fonction de l'exposition des salariés

Plus de 95% de ces contrats couvrent un effectif salarié dont l'exposition ne dépasse pas 1. Nous en déduisons que la majorité des P/C extrêmes sont produits par des contrats

couvrant 1 salarié (et ses éventuels bénéficiaires), et dont les dépenses sont disproportionnées par rapport à la cotisation individuelle du salarié sur sa période de couverture. Le fait de plus que le salarié ne soit pas présent sur toute l'année observée ne donne pas un P/C annuel fiable et correctement représentatif. Voici le détail des 5 contrats restants couvrant plus de 1 salarié en exposition.

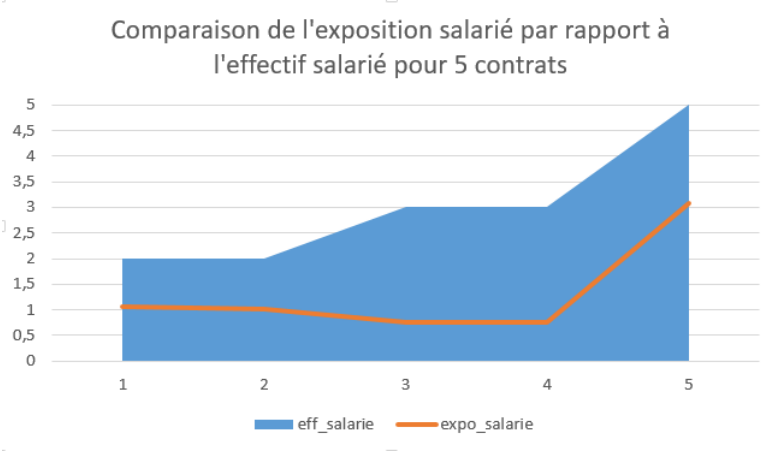


FIGURE 10.4 – Étude des P/C nuls pour les contrats de plus de 1 salarié exposé

Le taux d'exposition reste assez faible par rapport à l'effectif couvert. Étant donné que les résultats du GLM peuvent être très fortement influencées par une ou plusieurs valeurs trop extrêmes considérées outliers, nous décidons de supprimer de la base tous les contrats de P/C supérieur à 1000%.

Nous présentons l'histogramme du P/C net et ses quantiles suite aux suppressions de P/C extrêmes :

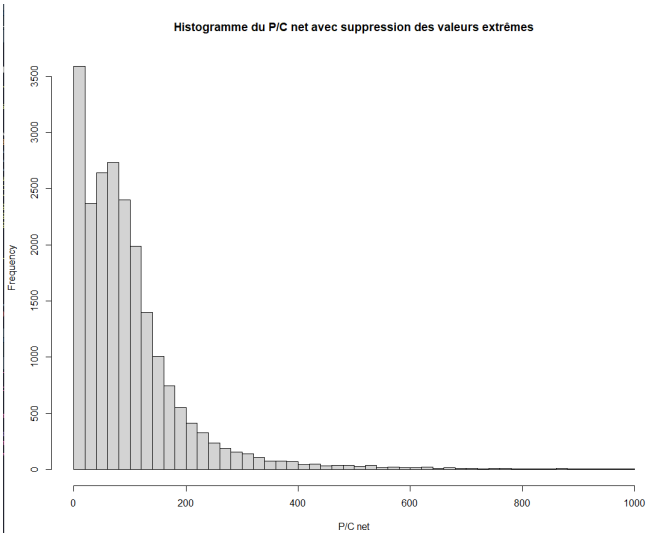


FIGURE 10.5 – Histogramme du P/C

Description de la nouvelle variable P/C net					
Min	1er Quantile	Médiane	Moyenne	3ème Quantile	Max
0,00	35,60	76,10	99,47	127,00	997,50

FIGURE 10.6 – Quantiles du P/C retraité

Nous notons par ailleurs une concentration de P/C nul : notre base se constitue de 5,6% de P/C à 0. Ces contrats concernent en moyenne des petits effectifs, n'ayant pas consommé sur la période de couverture.

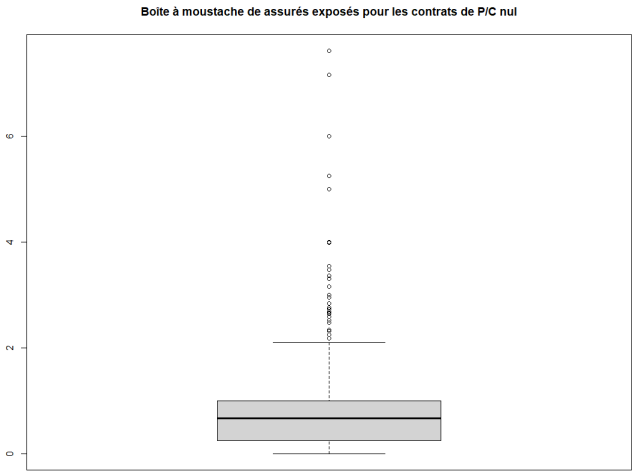


FIGURE 10.7 – Boxplot de l'exposition assuré pour le contrat de P/C nul

Cette proportion de données nulles nous permettent de nous interroger sur l'intérêt de l'utilisation d'une loi de Tweedie pour ajuster nos données. En effet, les distributions de Tweedie appartiennent à la classe des modèles de dispersion exponentielle. La famille de distributions Tweedie comprend des distributions continues telles que la distribution Normale et Gamma, la distribution de Poisson exclusivement discrète, et la classe de distributions composées mixtes Poisson-Gamma qui ont une quantité importante de zéros.

La distribution de nos P/C n'étant pas normale, les GLM se présentent comme une méthode adaptée de modélisation, puisqu'elles ne supposent pas de distribution normale sur la variable réponse.

Nous décidons de réaliser 2 GLM : un sur la base des P/C non nuls, et un second sur la base entière des P/C. Nous dégagerons à la suite de chaque GLM la typologie de contrat rentable et non rentable, et nous observerons les éventuelles différences de résultats selon les 2 modèles.

Chapitre 11

Étude des liaisons entre les variables

Dans cette section, nous étudierons les corrélations et forces de liaisons entre les différentes variables. En effet, dans le contexte d'une régression, les variables explicatives sont également appelées variables indépendantes. Le terme "indépendant" indique qu'elles sont autonomes et que les autres variables du modèle ne les influencent pas. Une trop forte corrélation entre des variables explicatives pourrait poser des problèmes lors de l'ajustement du modèle et de l'interprétation des résultats.

11.1 Variables quantitatives

Nous allons tout d'abord décrire à l'aide d'une matrice de corrélation de Pearson, la force de corrélation entre chaque variable numériques 2 à 2.

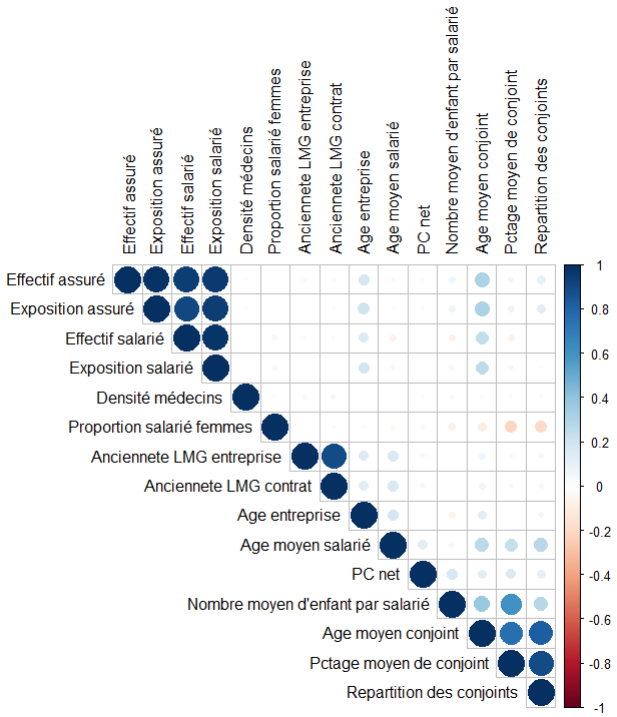


FIGURE 11.1 – Matrice de corrélation entre les variables quantitatives

Plus le cercle est grand et foncé, plus la corrélation entre les 2 variables testées est forte.

Les variables représentant l'effectif et l'exposition des salariés et assurés s'expliquent entre elles et présentent de ce fait une grande corrélation : notre choix pour la modélisation se tournera vers une conservation de la variable d'exposition des salariés. En effet, l'exposition donne une information plus précise sur la présence réelle des personnes au sein du contrat, et l'information du nombre d'assurés est capturée dans les variables démographiques des conjoints et des enfants.

De même, les variables proportion de conjoint et pourcentage de conjoint expliquent la même information ; nous décidons de conserver la variable proportion de conjoint assuré, puisque celle-ci présente moins de corrélation avec le pourcentage d'enfant assuré, et avec l'exposition des salariés.

De plus, l'ancienneté de l'entreprise et du contrat présentent trop de corrélation pour sélectionner les deux dans le modèle. Nous choisissons de réaliser la modélisation avec la variable expliquant l'ancienneté du contrat, puisque nous cherchons à déterminer une typologie au niveau des contrats plutôt qu'au niveau des entreprises.

Enfin, nous n'avons représenté sur la matrice qu'une variable de la densité médicale, afin de faciliter la lecture. Néanmoins, les différentes densités médicales de notre base sont toutes très corrélées. En effet, l'accès aux soins pour chaque département reste globalement uniforme Quel que soit le domaine médical. En règle générale, les départements présentant une grande densité de médecins présentent également une grande densité de pharmacies, d'infirmiers, ou encore de dentistes. Il conviendra alors de ne sélectionner dans la modélisation qu'une variable représentant la densité médicale. Ce choix peut se réaliser en lançant séparément des modèles intégrant une des densités, et de comparer la qualité d'ajustement entre tous les modèles testés. Nous choisissons par défaut la densité totale de médecins pour la première modélisation.

Les variables quantitatives pouvant être sélectionné pour le GLM sont donc :

- Age de l'entreprise
- Exposition des salariés
- Densité de médecins
- Proportion de salariées femmes
- Nombre moyen d'enfants assurés par salarié
- Proportion de conjoint
- Ancienneté du contrat
- Age moyen des salariés
- Age moyen des conjoints

11.2 Variables qualitatives

Afin de mesurer la corrélation entre les variables qualitatives, nous allons étudier les résultats des du V de Cramer entre chaque variable par paire.

Pour rappel, les règles de décisions du V de Cramer sont les suivantes :

— V de Cramer proche de 1 : forte liaison des variables ;
— V de Cramer proche de 0 : faible liaison des variables.

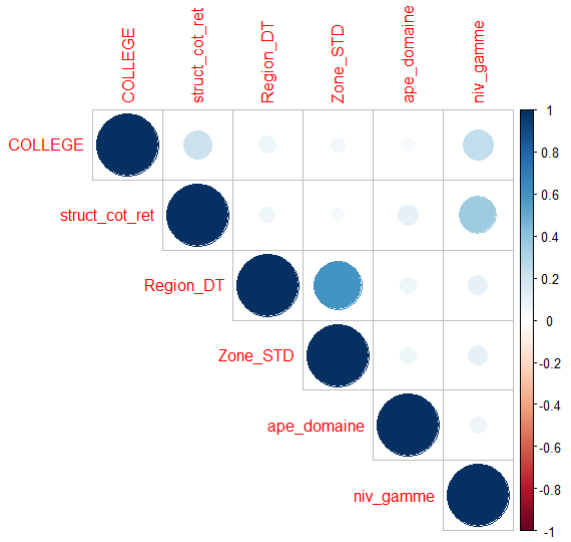


FIGURE 11.2 – Représentation des V de Cramer pour chaque variable qualitatives

Nous observons une liaison très forte entre la région_DT et la zone_std puisqu'il s'agit de variables de localisation de l'entreprise. Nous décidons pour la modélisation de ne conserver qu'une des 2 variables. En l'occurrence, les zones du zonier standards sont segmentées selon le niveau de consommation du lieu. Nous décidons d'incorporer le zonier dans le modèle afin de tester son influence sur le P/C des contrats. Si le zonier est bien segmenté, la variable devrait ressortir comme non significative dans l'explication du P/C. La deuxième liaison la plus forte est celle entre le niveau de gamme et la structure de cotisation, mais celle-ci est de 0.36. Nous pouvons la considérer comme correct pour incorporer les deux variables dans le modèle, sous réserve des résultats de la modélisation.

Les variables qualitatives pouvant être sélectionnées pour le GLM sont donc :

— Collège
— Structure de cotisation
— Zone STD ou Région DT
— Domaine APE
— Niveau de gamme

Liaison avec la variable réponse

Nous souhaitons à présent évaluer les liaisons entre la variable réponse et les variables explicatives. Nous allons croiser les variables explicatives avec le P/C net afin de se faire une idée du pouvoir discriminant de chaque variable. Pour ce faire, nous utilisons le test de Kruskal-Wallis. Il est utilisé pour comparer au moins trois modalités d'une variable suivant une variable quantitative, et tester l'hypothèse nulle suivant laquelle les différentes modalités sont issus de distributions de même médiane. Si la valeur de p est inférieure ou égale au seuil de signification, nous pouvons rejeter l'hypothèse nulle et conclure que toutes les médianes de modalités ne sont pas égales. Ici, nous rejetons H0 au seuil α de 5%.

Variable	Statistique	Degré de liberté	P valeur
Niveau de gamme	511	2	2.2e-16
Structure de cotisation	423	4	2.2e-16
Collège	73	3	7.276e-16
Domaine APE	79	5	1.225e-15
Région DT	35	7	8.229e-06
Zone STD	17	4	0.001386

FIGURE 11.3 – Résultats des tests de Kruskal Wallis

La variable qui rejette l'hypothèse nulle avec le plus de certitude est le niveau de gamme, puis la structure de cotisation. Nous avons observé une corrélation entre ces deux variables, mais celle-ci ne dépassait pas 0.5, donc nous sélectionnerons toujours ces 2 variables au sein du modèle.

11.3 Postulats sur la modélisation

Dans la section suivante, nous interpréterons les coefficients du modèle. Ce faisant, nous devons garder à l'esprit que toute conclusion que nous tirons concerne le modèle que nous construisons, plutôt que le véritable processus génératif des données.C'est pourquoi nous restons prudents sur les conclusions de cette section.

11.4 Postulats de validité du modèle

Pour rappel, la modélisation GLM postule que les erreurs sont indépendantes et identiquement distribuées. Cette hypothèse implique les points suivants :

— Indépendance : les erreurs sont indépendantes

— Linéarité : l'espérance des erreurs est nulle.

— Homoscédasticité : la variance des erreurs est constante

— Normalité : les termes d'erreurs suivent une loi normale

Chapitre 12

Première modélisation GLM : base des P/C non nuls

Nous commençons par modéliser le P/C sur la base des P/C non nuls.

12.1 Choix de la loi modélisant le P/C

Dans un premier temps, nous présenterons la démarche permettant de choisir la distribution selon laquelle le P/C sera modélisée. Nous exposerons par la suite la recherche du modèle le plus robuste et précis. Enfin, nous tenterons de valider le modèle choisi et de dégager des profils de contrats selon la rentabilité.

Nous souhaitons d'abord réaliser une modélisation du P/C sur la base des contrats ayant consommé Les contrats considérés comme rentables correspondent aux contrats dont le P/C ne dépasse pas 100%. Nous nous concentrons donc sur les données de P/C strictement positifs. Cette base est constituée de 20 441 observations, et l'histogramme est présenté ci-dessous.

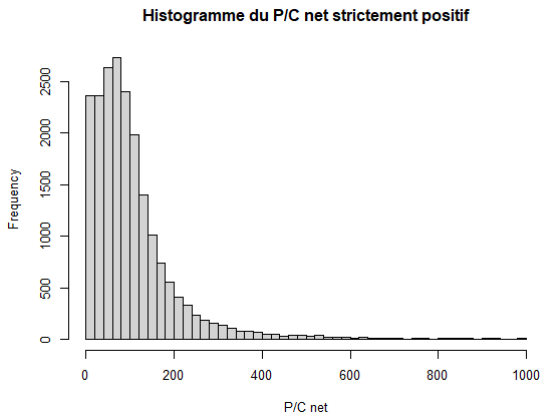


FIGURE 12.1 – Histogramme du P/C non nul

Nous testons l'ajustement de la distribution du P/C avec les lois Gamma, Log-Normale et Weibull à l'aide de la fonction fitdist sur R :

Gamma : r=1.299 θ=0.012

Log-normale : μ=4.226 σ=1.058

Weibull : β=1.136 θ=110.648

Le graphique ci-dessous représente l'ajustement de la densité, fonction de répartition, QQ plot et PP plot du P/C net retraité par la loi Gamma, Log-Normale et Weibull.

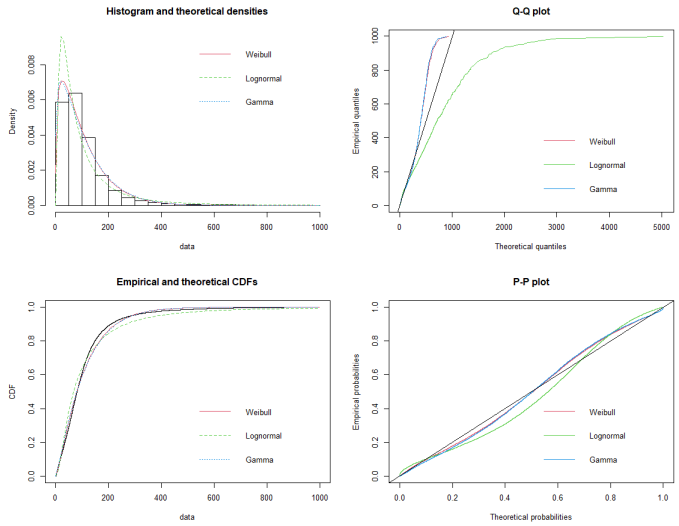


FIGURE 12.2 – Résultats des ajustements des lois

En observant le graphique de l'ajustement de la densité du P/C par les différentes lois, il semble que les lois Gamma et Weibull ajustent mieux les observations que la loi Log-Normal.

Le graphique de la fonction de répartition permet de constater que la loi Log normale surestime davantage les P/C inférieurs à 100%, par rapport à Gamma et Weibull, qui surestiment légèrement moins. Pour les P/C supérieurs à 100%, la loi Log-normale sous-estiment les valeurs, tandis que les lois Gamma et Weibull semblent se rapprocher plus conformément des valeurs, bien qu'elles sous-estiment aussi les P/C entre 100% et 300%.

Nous pouvons dégager les mêmes constats sur les diagrammes Quantile-Quantile et Probabilité-Probabilité, Les données empiriques des trois lois se détachent de celles théoriques ; Log-normale semble offrir le moins bon ajustement.

Globalement, nous constatons toujours que les lois Gamma et Weibull ajustent mieux les observations que la loi Log-Normale.

Nous réalisons les tests d'adéquation de Kolmogorov Smirnov et Cramer Von Mises afin de déterminer la loi qui convient le mieux à nos données de P/C.

Notons que les deux tests concluent sur un rejet de l'hypothèse nulle, selon laquelle la fonction de répartition théorique de la loi à tester est égale à la fonction de répartition empirique. Ce rejet peut s'expliquer par la grande quantité de données observées, qui influe sur la règle de rejet de l'hypothèse nulle. Nous nous basons donc sur les statistiques de test pour faire notre choix de loi.

Pour choisir la loi qui servira à la modélisation, conservons la loi dont les statistiques de test sont les plus faibles.

Test d'adéquation à la loi Gamma		
Test	**Statistique**	**Valeur**
Kolmogorov Smirnov	D	0.042
Cramer Von Mises	Omega2	12.636

Test d'adéquation à la loi Log-normale		
Test	**Statistique**	**Valeur**
Kolmogorov Smirnov	D	0.092
Cramer Von Mises	Omega2	56.336

Test d'adéquation à la loi Weibull		
Test	**Statistique**	**Valeur**
Kolmogorov Smirnov	D	0.048
Cramer Von Mises	Omega2	17.934

FIGURE 12.3 – Résultats des tests d'adéquation

Nous constatons donc que la loi Gamma présente la meilleure adéquation. Nous modéliserons donc le P/C net avec cette loi. Pour la fonction de lien, nous choisissons la fonction log. En effet, la forme multiplicative donne des interprétations simplifiées dans le cas de modèles multiples.

12.2 Sorties GLM

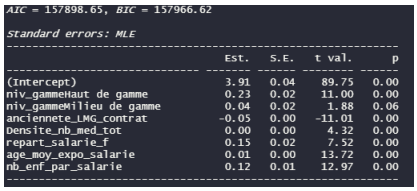
Pour finaliser la pré-sélection des variables du modèle, nous réalisons une sélection pas à pas mixte grâce à la fonction bestglm() sur R. Cet algorithme combine la sélection et l'élimination de variable afin de déterminer les variables ayant l'apport le plus significatif au modèle. L'algorithme base sa sélection selon un critère d'ajustement qu'il faut choisir. Nous optons pour le BIC, car nous souhaitons construire un modèle combinant fort pouvoir explicatif et parcimonie. Le BIC pénalisant plus le nombre de variables présents dans le modèle, nous le choisissons au critère AIC.

```
BIC
Best Model:
                          Df    Sum Sq Mean Sq F value   Pr(>F)
niv_gamme                  2   3466969 1733485  167.56  < 2e-16 ***
anciennete_LMG_contrat     1    754633  754633   72.94  < 2e-16 ***
Densite_nb_med_tot         1    120924  120924   11.69 0.000631 ***
repart_salarie_f           1    315465  315465   30.49 3.41e-08 ***
age_moy_expo_salarie       1   1754005 1754005  169.54  < 2e-16 ***
nb_enf_par_salarie         1   2173358 2173358  210.08  < 2e-16 ***
Residuals              14067 145528116   10345
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 12.4 – Résultats de la sélection stepwise

Le modèle retenu par la procédure contient le niveau de gamme, l'ancienneté du contrat chez LMG, la densité médicale, la proportion de salariée femme, l'âge moyen des salariés ainsi que le nombre d'enfants moyen par salarié.

Nous présentons ci-dessous les résultats de l'adéquation de ce nouveau modèle.



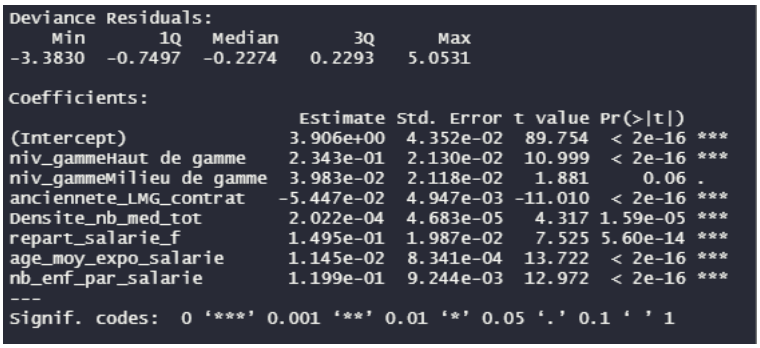
AIC = 157898.65, *BIC* = 157966.62

Standard errors: MLE

	Est.	S.E.	t val.	p
(Intercept)	3.91	0.04	89.75	0.00
niv_gammeHaut de gamme	0.23	0.02	11.00	0.00
niv_gammeMilieu de gamme	0.04	0.02	1.88	0.06
anciennete_LMG_contrat	-0.05	0.00	-11.01	0.00
Densite_nb_med_tot	0.00	0.00	4.32	0.00
repart_salarie_f	0.15	0.02	7.52	0.00
age_moy_expo_salarie	0.01	0.00	13.72	0.00
nb_enf_par_salarie	0.12	0.01	12.97	0.00

FIGURE 12.5 – Critères d'ajustement après la sélection stewise

Nous décrivons ci-dessous les résultats du modèle, présentant pour chaque modalité, l'estimation du paramètre, son erreur standard, la t value correspondant à l'estimateur / std error, et la p value associée à la t value. Si la p value est inférieure à 0.05, le paramètre possède une relation statistiquement significative avec le P/C.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3830  -0.7497  -0.2274   0.2293   5.0531

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.906e+00  4.352e-02  89.754  < 2e-16 ***
niv_gammeHaut de gamme   2.343e-01  2.130e-02  10.999  < 2e-16 ***
niv_gammeMilieu de gamme 3.983e-02  2.118e-02   1.881     0.06 .
anciennete_LMG_contrat  -5.447e-02  4.947e-03 -11.010  < 2e-16 ***
Densite_nb_med_tot       2.022e-04  4.683e-05   4.317 1.59e-05 ***
repart_salarie_f         1.495e-01  1.987e-02   7.525 5.60e-14 ***
age_moy_expo_salarie     1.145e-02  8.341e-04  13.722  < 2e-16 ***
nb_enf_par_salarie       1.199e-01  9.244e-03  12.972  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 12.6 – Critères d'ajustement après la sélection stepwise

On constate que tous les coefficients sont significatifs, excepté celui du Milieu de gamme. Ce modèle conclut que le passage d'entrée de gamme à milieu de gamme, toutes choses égales par ailleurs, ne modifie pas la valeur du P/C de manière significative.

Globalement, les variables sélectionnées par l'algorithme paraissent cohérentes. En effet, nous pouvons penser que le niveau de gamme influence la consommation des assurés du contrat : l'assuré d'un contrat haut de gamme peut modifier son comportement et s'exposer davantage aux risques étant donné son niveau de couverture élevé. Ce phénomène peut s'expliquer par l'aléa moral. De plus, les jeunes contrats du portefeuille n'ayant pas été soumis à des indexations tarifaires, ils peuvent présenter un déficit lors de la première année. Ensuite, tout comme l'effet du niveau de gamme, une grande densité médicale peut entrainer la modification du comportement de l'assuré, qui aura l'accès aux soins et est donc susceptible de consommer. Enfin, la proportion d'assurés femmes, l'âge moyen des salariés et le nombre d'enfants par salarié étant des paramètres tarifaires, le fait que ces variables influent le P/C n'est pas surprenant.

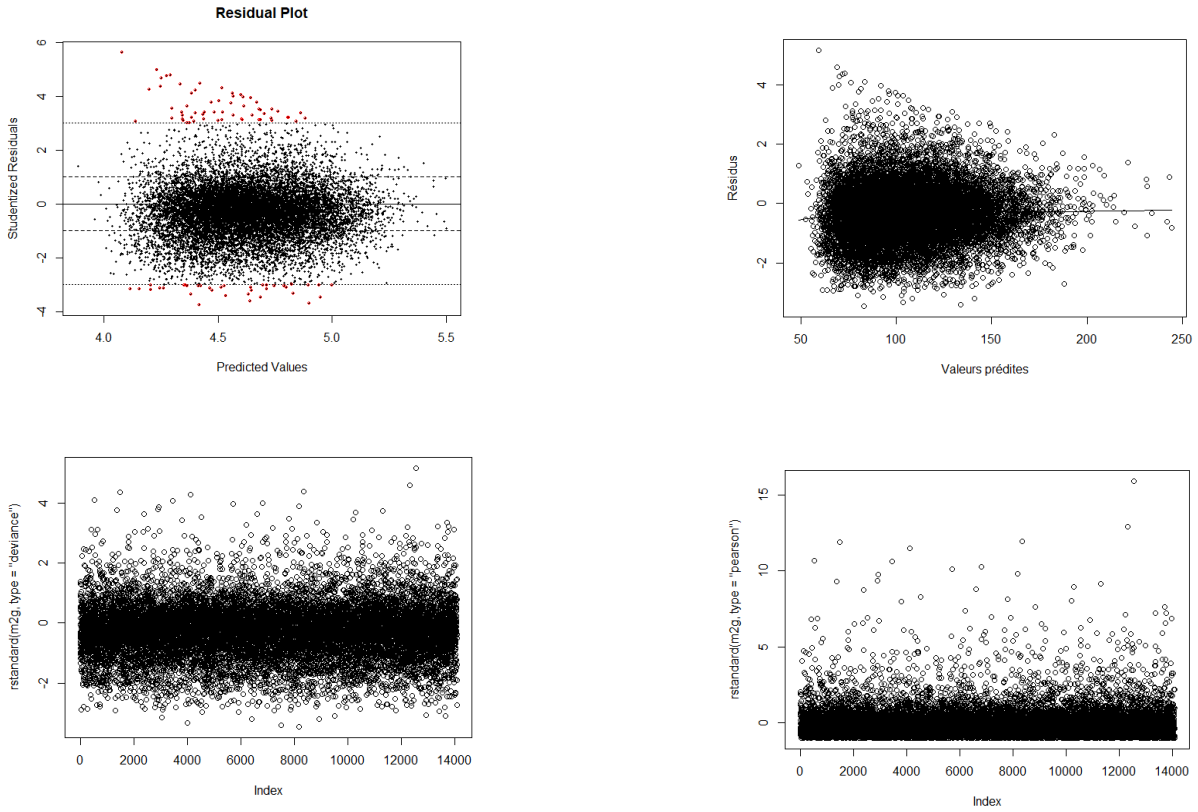


FIGURE 12.7 – Résidus du modèle Gamma

12.3 Analyse des résidus

Nous étudions les résidus afin de juger du pouvoir explicatif du modèle :

Nous pouvons voir que les résidus sont bien centrés autour de l'axe 0 et que le nuage de point ne montre pas de schéma précis. Pour les résidus de Pearson standardisés, la grande majorité des résidus de Pearson standardisés se situent proches de 0, avec quelques points entre 5 et 10 et un point autour de 15.

Nous notons en revanche que la dispersion des résidus a tendance à diminuer légèrement au fur et à mesure que la valeur ajustée augmente ; la variance des erreurs n'est pas tout à fait constante pour toutes les observations. Ce phénomène pourrait réduire la précision des estimations dans la régression.

Pour notre premier modèle, nous pouvons juger le diagnostic convenable et nous validons le modèle.

12.4 Interprétation des résultats

Cette partie a pour but de déterminer à partir du GLM l'effet de chaque variable explicative sur le P/C. Pour cela, nous pouvons nous baser sur les coefficients obtenus par la

régression.

Cependant, nous notons que les différentes variables ont des échelles naturelles différentes, en raison de leur unité de mesure différente. En effet, la proportion de salariées femmes varie entre 0 et 1, tandis que la densité totale de médecin varie plutôt entre 100 et 1000. La comparaison des coefficients peut être incorrecte.

Pour rendre la comparaison possible, nous décidons de standardiser nos coefficients afin de les ramener à la même échelle. Ces nouveaux coefficients indiquent alors la variation moyenne du P/C en fonction d'une variation d'un écart-type d'une variable explicative.

Les coefficients standardisés sont présentés ci-dessous :

Standard errors: MLE

	Est.	S.E.	t val.	p
(Intercept)	4.53	0.02	293.98	0.00
niv_gammeHaut de gamme	0.23	0.02	11.00	0.00
niv_gammeMilieu de gamme	0.04	0.02	1.88	0.06
anciennete_LMG_contrat	-0.09	0.01	-11.01	0.00
Densite_nb_med_tot	0.04	0.01	4.32	0.00
repart_salarie_f	0.06	0.01	7.52	0.00
age_moy_expo_salarie	0.12	0.01	13.72	0.00
nb_enf_par_salarie	0.11	0.01	12.97	0.00

FIGURE 12.8 – Coefficients standardisés

Concernant la variable qualitative du niveau de gamme, nous observons qu'effectivement plus le niveau de gamme est élevé, plus la variable a un effet négatif sur le P/C. En effet, pour les contrats milieu de gamme, dont le coefficient e^{β} est 1.04, leur P/C sera en moyenne 4% plus élevée que celui du portefeuille de référence en entrée de gamme, toutes choses égales par ailleurs. Les contrats du type haut de gamme ont un coefficient e^{β} de 1.26, signifiant qu'en moyenne, leur P/C sera 26% plus fort que celui du portefeuille en entrée de gamme.

Cette différence marquée entre les coefficients traduit une inégalité tarifaire des niveaux de gammes. En effet, nous aurions dû observer une égalité de l'influence des différentes modalités si les niveaux de garantie étaient tarifés à leur juste sinistralité.

Pour l'analyse des coefficients des variables quantitatives, la fonction utilisée fournit un coefficient β par variable, mais il serait intéressant de représenter l'influence de cette variable en différents points remarquables sur la variable réponse du P/C.

De ce fait, si X est le vecteur des observations de la variable quantitative, alors le vecteur des coefficients relatifs à chaque observation est $e^{\beta * X}$.

Nous présentons les coefficients relatifs à chaque variable quantitative :

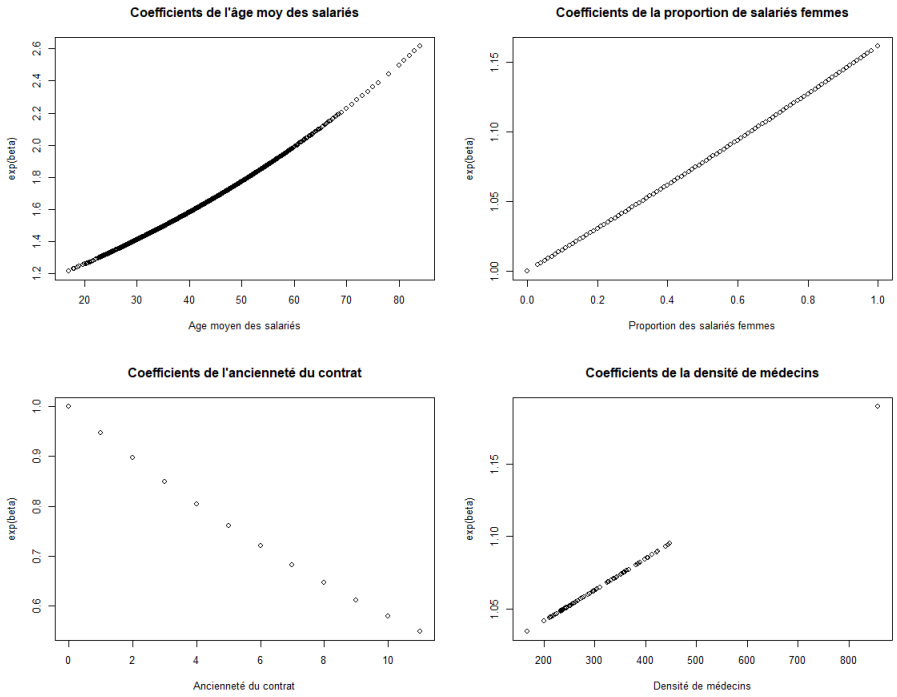


FIGURE 12.9 – Coefficients des variables quantitatives

L'interprétation des coefficients de la régression met en avant une corrélation positive entre le P/C et les variables de l'âge moyen des salariés, de la proportion de salariées femmes, et de la densité médicale. L'ancienneté du contrat est quant à elle corrélée négativement au P/C.

Suivant cette première modélisation, nous pouvons en déduire que plus l'âge moyen des salariés est élevé, plus le P/C moyen du portefeuille étudié se dégrade. De même pour la densité médicale et la proportion de salariées femmes qui entrainent une dégradation du P/C en augmentant. Pour l'ancienneté du contrat, il en ressort que les contrats présentant plus d'ancienneté, aura un P/C plus faible que les contrats du portefeuille sans ancienneté, toutes choses égales par ailleurs.

Chapitre 13

Seconde modélisation GLM : intégration des contrats de P/C nuls

Nous modélisons à présent le P/C sur sa base entière selon une loi de Tweedie. Nous travaillons sur une base de 21 749 observations. Nous souhaitons tout d'abord ajuster une loi de Tweedie à la distribution de notre P/C

13.1 Loi de Tweedie et GLM

Les distributions de Tweedie correspondent à une famille de distributions incluant les lois normale,gamma, de Poisson et Poisson-gamma, agrémentées d'une masse en zéro. Pour toute variable aléatoire Y suivant une distribution de la famille de Tweedie, nous avons la relation :

$$V(Y) = aE(Y)^p$$

avec a et p des constantes positives.

Pour p = 0, nous obtenons une distribution normale, pour p = 1, une distribution de Poisson, pour p = 2, une distribution gamma.

Nous nous intéressons au cas où p prend sa valeur entre 1 et 2. En effet, la distribution du P/C se compose de données non négligeables en 0, et est très asymétrique à droite : ces aspects orientent notre choix de distribution pour la modélisation du P/C vers cette loi de Tweedie.

La distribution de Tweedie faisant partie de la famille exponentielle, peut s'écrire sous la forme suivante.

$$f(y|\theta,\phi) = a(y,\phi)exp(\frac{y\theta - k(\theta)}{\phi})$$

Avec :

- θ un paramètre réel dit naturel ;
- ϕ un paramètre de dispersion strictement positif, qui contrôle la variance
- $a()$ et $k()$ des fonctions, $k()$ déterminant la distribution de la variable réponse

Nous avons, d'après la propriété des moments des familles exponentielles, $\mu = E(Y) = k'(\theta)$ et $V(Y) = \phi k''(\theta)$.

Soit $V(Y) = \phi\mu^p$ avec $1 < p < 2$, alors la distribution de Tweedie peut se réecrire sous la forme suivante

$$f(y|\theta, \phi) = a(y, \phi)exp(\frac{1}{\phi}(\frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}))$$

Ainsi, nous pouvons observer qui si $p = 1$ nous avos une loi de Poisson, et si $p = 2$ nous avons une loi Gamma.

13.2 Estimation des paramètres de la loi de Tweedie

Afin d'ajuster le modèle Poisson-Gamma aux données, nous devons estimer les paramètres inconnus. Nous avons trois paramètres à estimer, la moyenne μ, le paramètre de dispersion ϕ et l'indice de la distribution p. La fonction tweedie.profile, dans R (Dunn , 2013), donne une estimation des paramètres utilisant la fonction de log-vraisemblance de la densité Tweedie.

Nous présentons ci-dessous la fonction de log-vraisemblance profil en fonction de p, qui nous permet de voir la valeur de p qui maximise la log-vraisemblance.

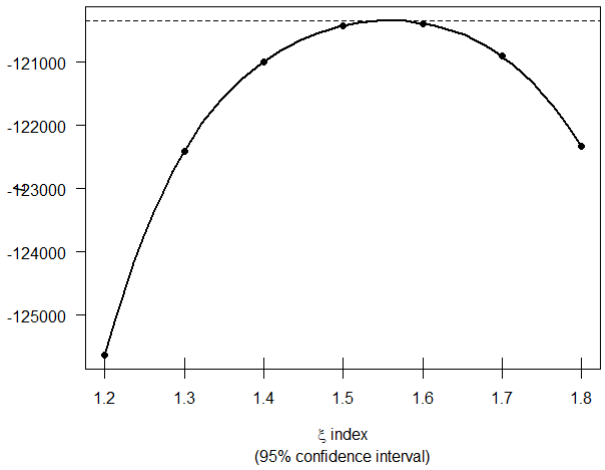


FIGURE 13.1 – Fonction de log-vraisemblance profil pour les P/C net des contrats

Voici les estimations de l'indice de puissance et du paramètre de dispersion selon cette technique :

$p = 1.55$ $\phi = 6.20$.

Le troisième paramètre étant la moyenne μ, nous choisissons la moyenne empirique de nos observations du P/C net : $\mu = 97.5$.

La distribution du P/C est donc ajustée à la loi de Tweedie de paramètre ($p = 1.55$, $\phi = 6.20$, $\mu = 97.5$).

13.3 Sorties GLM

Nous commençons la modélisation avec toutes les variables explicatives suivantes : structure de cotisation, Collège, Zone STD, Domaine APE, Niveau de gamme, Age de l'entreprise, Ancienneté du contrat chez LMG, Densité de médecins du département, Exposition des salariés, Proportion de salariés Femme, Age moyen des salariés, Pourcentage de conjoint par salarié, Proportion de conjoint assuré, age moyen des conjoints, Nombre d'enfants moyen par salariés.

La sortie GLM est la suivante, avec un $AIC = 174982$

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             3.454e+00  1.456e-01  23.727  < 2e-16 ***
struct_cot_retIsolé/Famille            -5.237e-02  2.852e-02  -1.836 0.066344 .
struct_cot_retSalarié/Conjoint/Enfant   1.254e-02  2.972e-02   0.422 0.673051
struct_cot_retSalarié/Salarié + Enfants/Conj  2.682e-03  4.307e-02   0.062 0.950349
struct_cot_retUniforme                 -5.902e-02  3.463e-02  -1.704 0.088330 .
COLLEGEEnsemble du personnel            1.107e-01  2.398e-02   4.618 3.91e-06 ***
COLLEGENon Cadre                        9.002e-02  2.741e-02   3.284 0.001027 **
COLLEGETNS                              5.720e-02  4.529e-02   1.263 0.206662
Zone_STDZone 2                          3.182e-02  4.053e-02   0.785 0.432371
Zone_STDZone 3                          2.266e-02  3.452e-02   0.656 0.511544
Zone_STDZone 4                         -6.707e-02  3.298e-02  -2.034 0.041976 *
Zone_STDZone 5                         -6.457e-02  3.514e-02  -1.838 0.066154 .
ape_domaineCOMMERCES                    2.577e-01  1.265e-01   2.037 0.041664 *
ape_domaineCONSTRUCTION                 1.457e-01  1.309e-01   1.113 0.265677
ape_domaineINDUSTRIES                   1.833e-01  1.287e-01   1.425 0.154297
ape_domaineNR                           1.574e-01  3.171e-01   0.496 0.619557
ape_domaineSERVICES                     2.624e-01  1.262e-01   2.079 0.037657 *
niv_gammeHaut de gamme                  3.024e-01  2.453e-02  12.326  < 2e-16 ***
niv_gammeMilieu de gamme                9.392e-02  2.364e-02   3.973 7.13e-05 ***
age_entreprise                         -7.579e-04  5.433e-04  -1.395 0.162997
anciennete_LMG_contrat                 -4.189e-02  5.189e-03  -8.072 7.39e-16 ***
Densite_nb_med_tot                      7.649e-05  7.192e-05   1.064 0.287544
expo_salarie                            1.848e-03  1.356e-03   1.363 0.172895
repart_salarie_f                        1.627e-01  2.141e-02   7.599 3.14e-14 ***
age_moy_expo_salarie                    1.196e-02  9.112e-04  13.130  < 2e-16 ***
repart_conj                            -4.698e-02  1.212e-01  -0.388 0.698214
age_moy_expo_conj                       2.418e-03  7.313e-04   3.306 0.000948 ***
nb_enf_par_salarie                      1.559e-01  1.080e-02  14.429  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

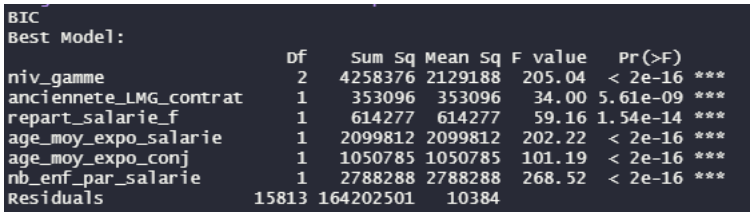
FIGURE 13.2 – Critères d'ajustement

D'après les p value de cette régression, retrouvons également comme les variables significatives, celles du niveau de gamme, de l'ancienneté, de la proportion de salariées femmes et de l'âge moyen des salariés.

Le nombre moyen d'enfants assurés ressort comme une variable numérique significative. En revanche, le modèle ne considère pas la densité de médecins comme significative. Pour les variables catégorielles, nous notons que les coefficients pour le collège Ensemble du personnel et Cadre, la zone 3 et le domaine APE du commerce sont inférieurs à 0.05.

Au vu des résultats, nous sommes d'avis de tester le modèle sans la structure de cotisation, la zone, le domaine APE, l'âge entreprise, l'exposition salarié et la proportion de conjoints.

Pour compléter notre jugement, nous lançons sur R l'algorithme de sélection stepwise basé sur le BIC.



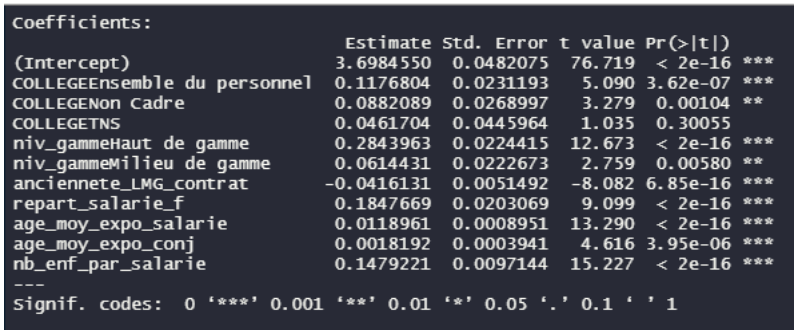
```
BIC
Best Model:
                        Df    Sum Sq Mean Sq F value   Pr(>F)
niv_gamme                2   4258376 2129188  205.04  < 2e-16 ***
anciennete_LMG_contrat   1    353096  353096   34.00 5.61e-09 ***
repart_salarie_f         1    614277  614277   59.16 1.54e-14 ***
age_moy_expo_salarie     1   2099812 2099812  202.22  < 2e-16 ***
age_moy_expo_conj        1   1050785 1050785  101.19  < 2e-16 ***
nb_enf_par_salarie       1   2788288 2788288  268.52  < 2e-16 ***
Residuals            15813 164202501   10384
```

FIGURE 13.3 – Résultat de la sélection stepwise

Bien que le modèle retenu ne contienne pas le collège, nous lançons un nouveau glm contenant toutes les variables sélectionnées par l'algorithme stepwise, et le collège en plus.

Second GLM

Les résultats d'ajustement du deuxième glm sont présentés ci-dessous, avec $AIC = 175049.8$:



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.6984550  0.0482075  76.719  < 2e-16 ***
COLLEGEEnsemble du personnel 0.1176804  0.0231193   5.090 3.62e-07 ***
COLLEGENon Cadre             0.0882089  0.0268997   3.279  0.00104 **
COLLEGETNS                   0.0461704  0.0445964   1.035  0.30055
niv_gammeHaut de gamme       0.2843963  0.0224415  12.673  < 2e-16 ***
niv_gammeMilieu de gamme     0.0614431  0.0222673   2.759  0.00580 **
anciennete_LMG_contrat      -0.0416131  0.0051492  -8.082 6.85e-16 ***
repart_salarie_f             0.1847669  0.0203069   9.099  < 2e-16 ***
age_moy_expo_salarie         0.0118961  0.0008951  13.290  < 2e-16 ***
age_moy_expo_conj            0.0018192  0.0003941   4.616 3.95e-06 ***
nb_enf_par_salarie           0.1479221  0.0097144  15.227  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 13.4 – Résultat de la sélection stepwise

Nous constatons que tous coefficients sont significatifs, excepté pour le contrat TNS, qui est peu représenté dans notre portefeuille.

Pour vérifier la légitimité du collège dans notre modèle, nous allons comparer 2 modèles qui diffèrent par la présence ou non du collège uniquement. Nous obtenons les résultats suivants :

$$AIC_{college} = 175049.8$$
$$AIC_{sans_college} = 175084$$

L'AIC du modèle qui inclut le collège étant meilleur, nous décidons de garder la variable du collège.

Enfin, nous avons vu qu'il n'y avait pas assez de preuve pour vérifier la significativité de la densité totale de médecins dans la modélisation Tweedie du P/C. Nous pouvons tester si une autre variable de la densité médicale s'ajuste mieux dans la régression. Nous ajoutons au dernier modèle la densité de médecins généralistes.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.6011625  0.0595081  60.516  < 2e-16 ***
COLLEGEEnsemble du personnel 0.1168573  0.0231049   5.058 4.29e-07 ***
COLLEGENon Cadre             0.0890530  0.0268799   3.313 0.000925 ***
COLLEGETNS                   0.0481671  0.0445681   1.081 0.279823
niv_gammeHaut de gamme       0.2818841  0.0224430  12.560  < 2e-16 ***
niv_gammeMilieu de gamme     0.0628567  0.0222617   2.824 0.004756 **
densite_nb_med_gen           0.0006049  0.0002186   2.767 0.005670 **
ancienneteLMG_contrat       -0.0420472  0.0051490  -8.166 3.42e-16 ***
repart_salarie_f             0.1824241  0.0203102   8.982  < 2e-16 ***
age_moy_expo_salarie         0.0119004  0.0008944  13.306  < 2e-16 ***
age_moy_expo_conj            0.0018311  0.0003939   4.649 3.37e-06 ***
nb_enf_par_salarie           0.1492100  0.0097126  15.362  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 13.5 – Critères d'ajustement suite à l'ajout de la densité de médecins généralistes

La p-value du coefficient de la densité médicale nous indique cette nouvelle variable est significative. L'AIC de 175040 est également plus faible.

13.4 Analyse des résidus

Nous étudions les résidus afin de juger notre modèle final :

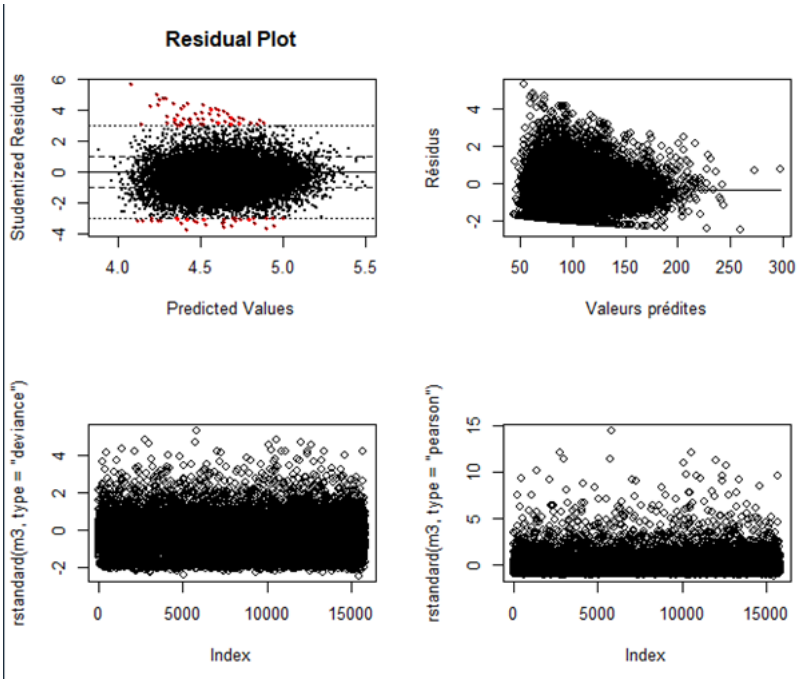


FIGURE 13.6 – Résidus du modèle Tweedie

Nous pouvons voir que les résidus studentisés sont centrés autour de l'axe 0 et la présence d'outliers. Pour les résidus de Pearson standardisés, la grande majorité des résidus de Pearson standardisés se situent proches de 0, avec quelques points entre 5 et 10 et un point autour de 15. En revanche, comparé aux résidus de la régression Gamma, le nuage de résidus suivant les valeurs prédites montre un schéma du point longiligne centré en -2.

Pour les P/C prédits entre 50% et 150%, une partie des valeurs prédites ont des résidus compris entre -2 et 2 de manière aléatoire, mais une autre partie des valeurs présentent des résidus systématiques proches de -2.

Ce tracé des résidus indique que les variables indépendantes ne capturent pas la totalité de la composante déterministe. Une partie de l'information explicative a été transférée à l'erreur supposée aléatoire.

En représentant graphiquement les résidus standardisés en fonction de chaque variable explicatives, nous remarquons ce même schéma non aléatoire sur le graphique selon l'âge moyen des salariés :

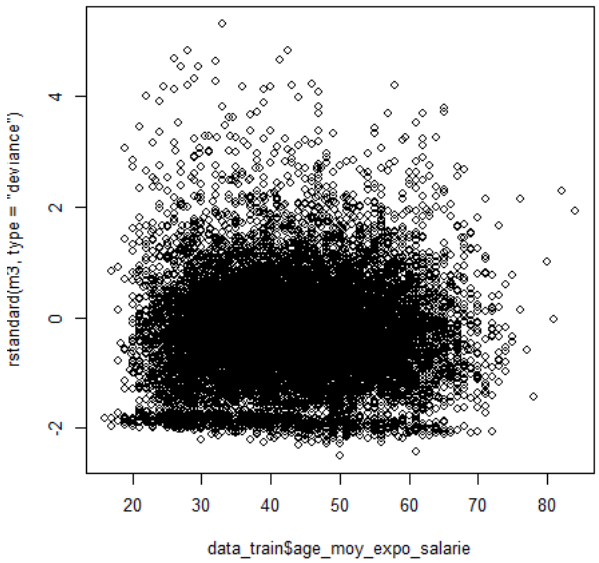


FIGURE 13.7 – Résidus standardisés en fonction de l'âge moyen des salariés

Ce problème peut être lié aux variables confusionnelles et entraîne un biais de variable omise. Il s'agit d'un point d'amélioration de ce modèle. Nous pouvons conclure globalement que le modèle GLM Tweedie n'est pas aussi performant que celui de Gamma, au vu des résidus. L'intégration des contrats de P/C nuls ne permet de pas un meilleur ajustement de la régression. Nous décidons d'évaluer les coefficients de la régression de Tweedie afin de les comparer à ceux du GLM de Gamma. Nous restons prudents sur l'interprétation étant donné une présence de corrélation des résidus, bien que la majorité soient aléatoires centrés en 0.

13.5 Interprétation des résultats

Nous standardisons les coefficients afin de pouvoir les comparer. Les coefficients des variables qualitatives sont représentés dans le graphique ci-dessous :

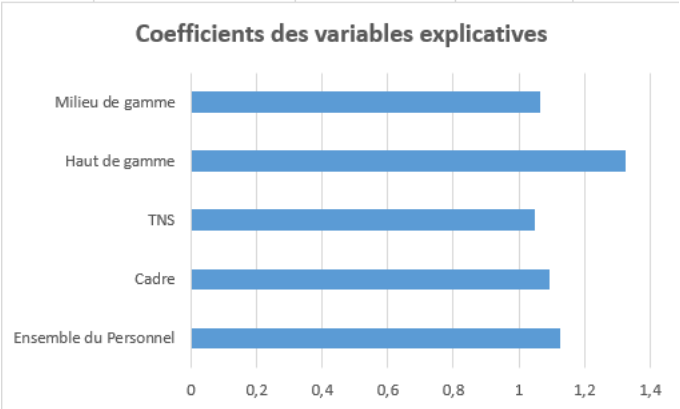


FIGURE 13.8 – Coefficients des variables qualitatives retenues

Le haut de gamme ressort également comme une modalité ayant un effet dégradant sur le P/C. Le P/C des contrats hauts de gammes seront considérés en moyenne 23% plus élevés que ceux des contrats entrée de gamma, ce qui rejoint l'interprétation des coefficients du modèle Gamma. Concernant le collège, qui n'a pas été selectionné dans le modèle Gamma, la modélisation Tweedie traduit le fait que les contrats tarifés pour l'ensemble du personnel et les cadres seront respectivement 12% et 9% en moyenne plus élevés que les contrats de référence tarifés pour les non-cadres. Les cadres correspondant à une catégorie socioprofessionnelle (CSP) plus avantageuse, les contrats tarifés pour les cadres peuvent offrir de meilleurs remboursements. Les contrats pour l'ensemble du personnel couvrent les cadres et non cadres et se basent sur la mutualisation des risques entre les 2 CSP pour construire leur tarif. Nous avons observé dans la partie de l'analyse descriptive que le P/C pour les contrats non-cadres était globalement moins élevés que le P/C des autres collèges. Les coefficients seraient cohérents avec les observations du portefeuille. La différence de coefficients pourrait traduire une inégalité tarifaire entre les collèges tarifés. Pour atteindre un P/C cible identique à tous les collèges, l'organisme complémentaire pourrait légèrement augmenter les cotisations des contrats pour l'ensemble du personnel et les cadres.

Nous présentons les coefficients relatifs à chaque variable quantitative :

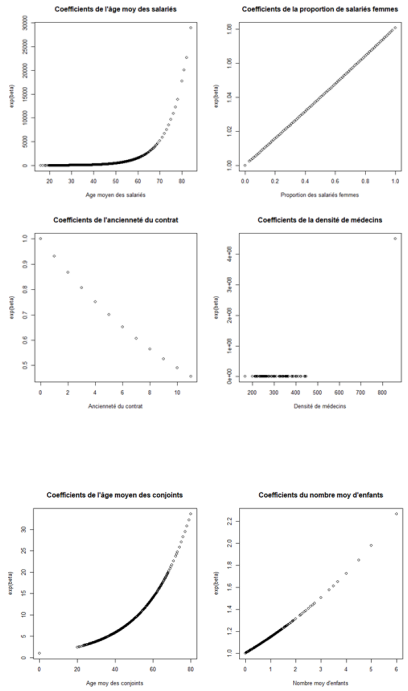


FIGURE 13.9 – Coefficients des variables quantitatives

Nous remarquons que les coefficients sont peu interprétables, particulièrement ceux relatifs à l'âge moyen des salariés et des conjoints. En effet, les coefficients estimés par le GLM pour ces 2 variables sont très élevés. La robustesse de ces coefficients n'est pas garantie, ce qui ne rend pas l'interprétation possible. De non-cadres effets compensatoires entre les variables peuvent rentrer en jeu. Le vecteur des coefficients relatifs aux coefficients de la répartition des salariées femmes et de l'ancienneté des contrats sont eux comparables à ceux du modèle Gamma. Nous observons une légère dégradation du P/C moyen au fur et à mesure que la répartition des salariées femmes augmente, ainsi qu'une baisse moyenne du P/C lorsque que le contrat est plus ancien que le portefeuille de contrats moyen, toutes choses égales par ailleurs. Les coefficients sortis par la modélisation GLM indiquent que l'ajustement est de plus mauvaise qualité que le modèle Gamma. Nous considérons donc que le modèle à retenir entre les deux est celui modélisé par la loi Gamma.

Chapitre 14

Sensibilité de la modélisation

Nous souhaitons à présent tester la sensibilité de notre modèle le plus robuste. Ces tests permettront une analyse critique du modèle et de la participation de chaque variable dans l'explication du P/C. Plus globalement, nous allons la robustesse et la variabilité du modèle GLM face à la suppression de variables explicatives, ou bien encore face à la modification du périmètre (réduction des années d'observations) ou à la modification des modalités de variables.

Pour ce faire, nous partons d'un scénario central, qui correspond au dernier modèle GLM sélectionné avec les variables explicatives significatives. Nous étudierons le GLM Gamma en raison de ces meilleurs résidus. Nous travaillerons sur l'indice AIC du modèle pour tester la sensibilité. Notre but est d'observer le niveau de contribution d'une variable explicative dans la modélisation, en comparant l'évolution de l'AIC d'un GLM sans cette variable par rapport à l'AIC du scénario central.

Pour rappel, l'AIC détermine la valeur d'information relative du modèle en utilisant l'estimation du maximum de vraisemblance et le nombre de paramètres (variables indépendantes) dans le modèle. La formule de l'AIC est la suivante :

$$AIC = 2K - 2ln(L)$$

K est le nombre de variables indépendantes utilisées et L est l'estimation de la log-vraisemblance, soit la probabilité que le modèle ait pu produire les valeurs observées.

$$AIC_{scnariocentral} = 157898$$

Pour comparer les modèles à l'aide de l'AIC, nous calculons l'AIC de chaque modèle. Si un modèle est inférieur de plus de 2 unités AIC à un autre, il est considéré comme significativement meilleur que ce dernier.

Notons que l'AIC compare les modèles entre eux, mais ne nous dit pas si le "meilleur" modèle est bien ajusté aux données. Dans notre cas, le meilleur modèle correspond à celui dont les résidus sont les plus satisfaisants.

14.1 Sensibilité à l'omission de variable

Nous testons la sensibilité de chaque variable au sein de la modélisation.

Sensibilité à la suppression d'une variable		
Variable supprimée	**AIC**	**Evolution par rapport au scénario central**
Age moyen des salariés	158145,55	246,9
Nombre moyen d'enfants par salarié	158139,86	241,21
Niveau de gamme	158088,32	189,67
Anciennete du contrat	158054	155,35
Proportion de salarié femme	157972,6	73,95
Densité médicale	157920,98	22,33

FIGURE 14.1 – Résultats de la sensibilité du modèle à la suppression d'une variable

Nous constatons que l'omission de chaque variable entraîne une baisse significative de la qualité d'ajustement du modèle, ce qui peut traduire de la bonne sélection des variables explicatives finales. Le tableau étant trié de la plus grande variation positive à la négative, nous constatons que notre modélisation de la rentabilité est très sensible à 2 variables : L'âge moyen des salariés et le nombre moyen d'enfants par salariés, avec une différence d'AIC de plus de 200 unités par rapport au scénario central. Le test du niveau de gamme et de l'ancienneté du contrat témoigne également d'une grande significativité au sein du modèle. Les dernières variables les plus impactantes sont la proportion de salariées femmes et la densité médicale. Ces résultats restent cohérents avec les coefficients de chaque variable. Nous pouvions effectivement voir que la variation moyenne du P/C selon la proportion de salariées femmes ou la densité médicale est moins forte que celle selon l'âge moyen des salariés ou du nombre d'enfants par salarié.

14.2 Sensibilité à la réduction du périmètre

Nous testons également la variabilité du modèle face à la modification du périmètre étudié.

Sensibilité à la réduction du périmètre			
Périmètre supprimé	**Réduction des observations**	**AIC**	**Evolution par rapport au scénario central**
Année 2020	-26%	122838,83	35059,82
Produit Santé Entreprise	-21%	130323,99	27574,66

FIGURE 14.2 – Résultats de la sensibilité du modèle à la réduction du périmètre

Le modèle semble davantage sensible à la suppression de l'année d'observation 2020 que la suppression du produit Sante Entreprise. En effet, les observations du scénario sans l'année 2020 sont réduits de 26% et offre une différence de plus de 35000 unités d'AIC, tandis que le scénario sans le produit Santé Entreprise présente une différence de 27000 unités pour une réduction de 21%.

Septième partie

Conclusion

Le mémoire présente une approche mathématique pour déceler des profils de contrats de complémentaire santé collective des plus au moins rentables, en se basant sur le principe du GLM. L'étude met en avant les caractéristiques des entreprises les moins rentables du portefeuille sélectionné. Cela permettra de mieux cibler les Entreprises plus rentables et d'ajuster les normes tarifaires également.

Pour l'application du GLM, la variable du P/C doit suivre une des lois de la famille exponentielle. Nous testons la modélisation GLM selon 2 lois, en se basant d'une part sur la base des P/C non nuls, et d'autre part sur la base entière des P/C. La première loi retenue est la loi Gamma ; choisie par une méthode graphique et sur les résultats du test de Cran Von Mises. La seconde loi est la loi de Tweedie de paramètre de puissance compris entre 1 et 2, qui permet intégrer nos observations de P/C nuls. L'estimation des paramètres de la loi s'est réalisée en maximisant la fonction de log-vraisemblance de la densité Tweedie.

Les GLM finaux nous permettent notamment de déterminer des variables explicatives qui influencent sur le niveau de rentabilité d'un contrat. Les variables retenues pour le GLM Gamma sont : l'âge moyen des salariés, le nombre moyen d'enfants par salarié, le niveau de gamme choisi, l'ancienneté du contrat, la part de salariées femmes, la densité du nombre de médecins dans le département de l'entreprise. Concernant le GLM de Tweedie, les variables retenues par la modèle qui sont communes au 2 modèles sont l'âge moyen des salariés, le nombre d'enfants par salarié, le niveau de gamme, l'ancienneté du contrat et la part de salariées femmes. Le GLM Tweedie inclut également l'âge moyen des conjoints, le collège tarifé ainsi que la densité de médecins généralistes.

En revanche, la validation du modèle de Tweedie n'est pas aussi ferme que celle du modèle Gamma : les résidus de glm par Tweedie sont peu satisfaisants en raison d'une présence d'une corrélation.

Les résidus du modèle Gamma sont quant à eux plus satisfaisants, mais nous notons tout de même la présence de valeurs plus extrêmes dans les résidus.

Nous sommes arrivés à la conclusion que le profil de contrat le plus déficitaire concerne les contrats de type haut de gamme, et d'âge moyen de salarié élevé, constitué d'un grand nombre de salariées femmes, et d'enfants assurés par salarié, et où l'entreprise se situe dans une grande densité de médecins.

Globalement, plus le niveau de gamme du contrat est élevé, plus la rentabilité du contrat se dégrade en moyenne, ce phénomène peut s'expliquer par le fait qu'une bonne couverture entraîne un niveau de consommation plus libre. L'âge moyen des salariés du contrat influe également le résultat moyen étant donné que les personnes âgées ont tendance à avoir plus de sinistres et de consommation. En termes de localisation, les résultats indiquent une relation positive entre la densité médicale du département de l'entreprise et le P/C. Une densité élevée traduit un meilleur accès aux soins, et aussi le fait que les assurés ne renoncent pas aux soins et donc consomment. Nous avons également observé que les jeunes contrats sont en moyenne moins rentables que les contrats anciens ; l'historique de ces

contrats au fil des ans permet une meilleure gestion du risque et un meilleur redressement du résultat, grâce aux indexations annuelles des cotisations. Enfin, plus le contrat couvre de conjoints et d'enfants par salarié, plus la rentabilité de celui-ci se dégrade en moyenne. Ce fait n'est pas non plus choquant, puisque les bénéficiaires profitent de la couverture de l'adhérent pour consommer.

Nous pouvons conclure que si des typologies de contrats déficitaires peuvent se dessiner à partir de notre périmètre, alors une révision des normes tarifaires peut être envisagé, notamment pour les contrats de niveau de couverture haut de gamme. Ensuite, l'organisme peut ce plus cibler une typologie de contrats afin de diminuer la détention de mauvais risques au sein du portefeuille.

Nous jugeons les résultats obtenus convenables mais l'étude peut être enrichie de plusieurs manières. Tout d'abord, la validité du modèle peut être davantage analysé, notamment en construisant la modélisation sur 80% de notre base, appelée « base d'apprentissage », puis en testant la qualité de cette modélisation une la base restante appelée « base de test ». Cela permet d'apporter de l'information sur la manière dont nos prédictions s'ajustent aux valeurs empiriques. De plus, une étude plus poussée sur les résidus permettra de valider ou non la faibilité des coefficients sortis.

Les résultats peuvent être optimisés par les études supplémentaires telles que :

- Une étude plus affinée sur les contrats de très petite taille ainsi que sur les observations de P/C extrêmes
- L'éventuel impact du 100% Santé sur les types de contrats plus rentables ou non
- Une étude zoomée sur le seul produit phare encore commercialisée

Annexes

14.3 Étude de la variable du code de garantie technique

Voici les analyses et remarques que nous avons pu soulever en étudiant vue brute du portefeuille et la variable « cd_gt » imputée dans celle-ci. Tout d'abord, le code de garantie technique renseigne les différents garanties et services couverts par LMG dans le cadre d'un contrat collectif. Ainsi, un salarié ayant souscrit au contrat de complémentaire santé de son entreprise se verra bénéficier d'une garantie de complémentaire santé, et de services tels qu'un service téléphonique et un forfait maternité. Dans la vue portefeuille, un assuré a au moins une ligne par garantie et par service, le cd_gt transcrivant la garantie ou le service.

Nous rappelons que la gamme ESPRITCO est plus complexe à bien appréhender en raison du grand nombre de générations du produit et des offres qui en découlent. De plus, les TdG (Tableau de garantie) peuvent différer selon ces offres, et ont évolué jusqu'à l'uniformisation début 2019.

Nous avons constaté que les offres liées au produit ESPRITCO se distinguent comme suit :

- MODULAIRE (le code offre contient « MOD »)
- DIRIGEANT (le code offre contient « DIR ») ;
- PREMS (le code offre contient « PREMS ») ;
- ESPRITCO dernières versions : il s'agit des dernières distributions du produit ESPRITCO qui ne font partie ni du modulaire, ni du dirigeant, ni du PREMS. Cela concerne 2 codes offres (ECOPTCTG, ECOPTLMG) ;
- ESPRITCO anciennes versions : il s'agit des codes offres restantes.

Pour les offres DIRIGEANT et ESPRITCO dernières versions, un niveau de garantie unique est souscrit sur l'ensemble des modules de soins. Le cd_gt pour la garantie complémentaire santé, semble également renseigner le niveau souscrit grâce aux caractères « N » + le numéro de niveau (et « O » + le numéro de l'option souscrite si l'assuré en a souscrite une), disposés en fin de code.

num_ass	TYPE_ASS	CD_OFFRE	CD_PRODUIT	CD_GT	LIB_GT
333606	ASSPRI	ECDIRLMG	ECDIRLMGN6	CMC208N6	Complémentaire Maladie Collective 208 N6
333606	ASSPRI	ECDIRLMG	ECDIRLMGN6	LICLAIRE	Ligne Claire (Services SMS)
333606	ASSPRI	ECDIRLMG	ECDIRLMGN6	MAT208N6	Forfait Maternité Collective 208 N6
777621	ASSPRI	ECDIRLMG	ECDIRLMGN5	CMC207N5	Complémentaire Maladie Collective 207 N5
777621	ASSPRI	ECDIRLMG	ECDIRLMGN5	LICLAIRE	Ligne Claire (Services SMS)
777621	ASSPRI	ECDIRLMG	ECDIRLMGN5	MAT207N5	Forfait Maternité Collectif 207 N5
861314	CONJOI	ECDIRLMG	ECDIRLMGN3	CMC203BN3	Complémentaire Maladie Collective 203 Base N3
861314	CONJOI	ECDIRLMG	ECDIRLMGN3	LICLAIRE	Ligne Claire (Services SMS)
861314	CONJOI	ECDIRLMG	ECDIRLMGN3	MAT203BN3	Forfait Maternité Collectif 203 Base N3
1478247	CONJOI	ECDIRLMG	ECDIRLMGN3	CMC203BN3	Complémentaire Maladie Collective 203 Base N3
1478247	CONJOI	ECDIRLMG	ECDIRLMGN3	LICLAIRE	Ligne Claire (Services SMS)
1478247	CONJOI	ECDIRLMG	ECDIRLMGN3	MAT203BN3	Forfait Maternité Collectif 203 Base N3

FIGURE 14.3 – Illustration de la variable cd_gt pour l'offre Dirigeant d'EspritCo

Pour MODULAIRE, le cd_gt indique une suite de 3 chiffres qui semblent renseigner dans l'ordre la modulation des niveaux de garanties entre 3 modules de soins « M », « D » et « H » :

num_ass	TYPE_ASS	CD_OFFRE	CD_PRODUIT	CD_GT	LIB_GT
312393	ASSPRI	ECMODLMG	ECMODLMGN4	CMC444MDH	Complémentaire Maladie Collective Niv 444
312393	ASSPRI	ECMODLMG	ECMODLMGN4	LICLAIRE	Ligne Claire (Services SMS)
443283	ASSPRI	ECMODLMG	ECMODLMGN5	CMC565MDH	Complémentaire Maladie Collective Niv 565
443283	ASSPRI	ECMODLMG	ECMODLMGN5	LICLAIRE	Ligne Claire (Services SMS)
465954	ASSPRI	ECMODLMG	ECMODLMGN5	CMC565MDH	Complémentaire Maladie Collective Niv 565
465954	ASSPRI	ECMODLMG	ECMODLMGN5	LICLAIRE	Ligne Claire (Services SMS)
473962	ASSPRI	ECMODLMG	ECMODLMGN3	CMC343MDH	Complémentaire Maladie Collective Niv 343
473962	ASSPRI	ECMODLMG	ECMODLMGN3	LICLAIRE	Ligne Claire (Services SMS)
558504	ASSPRI	ECMODLMG	ECMODLMGN4	CMC464MDH	Complémentaire Maladie Collective Niv 464
558504	ASSPRI	ECMODLMG	ECMODLMGN4	LICLAIRE	Ligne Claire (Services SMS)
805339	CONJOI	ECMODLMG	ECMODLMGN4	CMC454MDH	Complémentaire Maladie Collective Niv 454
805339	CONJOI	ECMODLMG	ECMODLMGN4	LICLAIRE	Ligne Claire (Services SMS)

FIGURE 14.4 – Illustration de la variable cd_gt pour l'offre Modulaire d'EspritCo

D'après nos recherches, ces modules correspondraient à Médecine de ville, Dentaire+Optique et Hospitalisation. En effet, cela reste cohérent avec ce que nous avons pu observer au sein des offres ESPRITCO anciennes versions, pour lesquelles chaque assuré possède 1 cd_gt pour chacun de ces 3 modules.

Cela resterait également cohérent avec l'ancien TdG d'ESPRITCO de 2017 que nous nous sommes procuré, car ce dernier distinguait ces 4 modules : Soins Courants/Médicaments, Dentaire, Optique et Hospitalisation.

num_ass	TYPE_ASS	CD_OFFRE	CD_PRODUIT	CD_GT	LIB_GT
11852	ASSPRI	ESPRITCOL2	AS2013580S	CMC666DO2	Compl Maladie Esprit Coll V2 Dentaire Optique 666
11852	ASSPRI	ESPRITCOL2	AS2013580S	CMC666HO2	Compl Maladie Esprit Coll V2 Hospitalisation 666
11852	ASSPRI	ESPRITCOL2	AS2013580S	CMC666ME2	Compl Maladie Esprit Coll V2 Médecine de Ville 666
11852	ASSPRI	ESPRITCOL2	AS2013580S	LICLAIRE	Ligne Claire (Services SMS)
310290	ASSPRI	ESPRITCOL2	LM37500066	CMC666DO2	Compl Maladie Esprit Coll V2 Dentaire Optique 666
310290	ASSPRI	ESPRITCOL2	LM37500066	CMC666HO2	Compl Maladie Esprit Coll V2 Hospitalisation 666
310290	ASSPRI	ESPRITCOL2	LM37500066	CMC666ME2	Compl Maladie Esprit Coll V2 Médecine de Ville 666
416603	CONJOI	ESPRITCOL2	LM27500066	CMC666DO2	Compl Maladie Esprit Coll V2 Dentaire Optique 666
416603	CONJOI	ESPRITCOL2	LM27500066	CMC666HO2	Compl Maladie Esprit Coll V2 Hospitalisation 666
416603	CONJOI	ESPRITCOL2	LM27500066	CMC666ME2	Compl Maladie Esprit Coll V2 Médecine de Ville 666
416603	CONJOI	ESPRITCOL2	LM27500066	LICLAIRE	Ligne Claire (Services SMS)

FIGURE 14.5 – Illustration de la variable cd_gt pour une ancienne génération du produit d'EspritCo

Cela resterait également cohérent avec l'ancien TdG d'esprit co de 2017 que nous nous sommes procuré, car ce dernier distinguait ces 4 modules : Soins Courants/Médicaments, Dentaire, Optique et Hospitalisation.

Enfin pour PREMS, nous avons de la même manière que MODULAIRE une suite à 3 chiffres qui semblent renseigner les 3 niveaux souscrits des 3 modules

num_ass	TYPE_ASS	CD_OFFRE	CD_PRODUIT	CD_GT	LIB_GT
264768	ASSPRI	PREMSTPE4	AS2015PRIF	CMC666PR2	Compl Maladie PREMS V2 666
264768	ASSPRI	PREMSTPE4	AS2015PRIF	LICLAIRE	Ligne Claire (Services SMS)
264768	ASSPRI	PREMSTPE4	AS2015PRIF	MAT666PR2	Forfait Maternité PREMS V2 666
264769	ENFANT	PREMSTPE4	AS2015PRIF	CMC666PR2	Compl Maladie PREMS V2 666
264769	ENFANT	PREMSTPE4	AS2015PRIF	LICLAIRE	Ligne Claire (Services SMS)
264769	ENFANT	PREMSTPE4	AS2015PRIF	MAT666PR2	Forfait Maternité PREMS V2 666
290317	ASSPRI	PREMSTPE4	AS2015PRAE	CMC444PR2	Compl Maladie PREMS V2 444
290317	ASSPRI	PREMSTPE4	AS2015PRAE	LICLAIRE	Ligne Claire (Services SMS)
290317	ASSPRI	PREMSTPE4	AS2015PRAE	MAT444PR2	Forfait Maternité PREMS V2 444
302770	ASSPRI	PREMSTPE3	AS2013PRUN	CMC666PRE	Compl Maladie PREMS 666
302770	ASSPRI	PREMSTPE3	AS2013PRUN	LICLAIRE	Ligne Claire (Services SMS)
302770	ASSPRI	PREMSTPE3	AS2013PRUN	MAT666PRE	Forfait Maternité PREMS 666

FIGURE 14.6 – Illustration de la variable cd_gt pour l'offre PREMS d'EspritCo

Toutefois, il semblerait qu'au sein des offres PREMS, une offre aurait une expression de garantie différente des autres car le cd_gt indique une souscription au Niveau « Base », « Confort », « Privilège » et « Excellence ». Cette offre « SANIPREMS » concerne d'après le cd_produit associés les assurés dans le cadre ANI.

num_ass	TYPE_ASS	CD_OFFRE	CD_PRODUIT	CD_GT	LIB_GT
32280	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
326017	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
396425	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
465894	ENFANT	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
665905	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
729782	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BPR	Compl. Maladie Collective 146 Base Privilège
729782	ASSPRI	SANIPREMS	AS14SOCANI	CMC147PRI	Compl. Maladie Collective 147 Option Privilège
729782	ASSPRI	SANIPREMS	AS14SOCANI	MAT147PRI	Forfait Maternité Collectif 147 Option Privilège
906488	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
997800	ENFANT	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
1258846	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
1364947	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule
1580838	ASSPRI	SANIPREMS	AS14SOCANI	CMC146BA	Compl. Maladie Collective 146 Base Seule

FIGURE 14.7 – Illustration de la variable cd_gt pour l'offre SANIPREMS d'EspritCo

14.4 Analyse descriptive de variables

Les graphiques suivants décrivent le portefeuille selon les domaines APE, les densités médicales du département de l'entreprise et l'âge de l'entreprise.

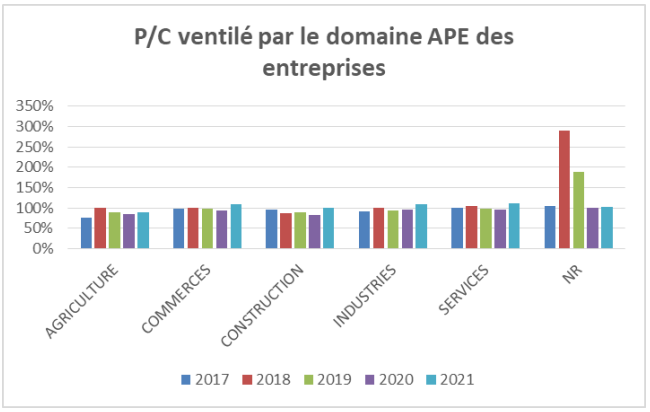


FIGURE 14.8 – P/C ventilé par domaine APE

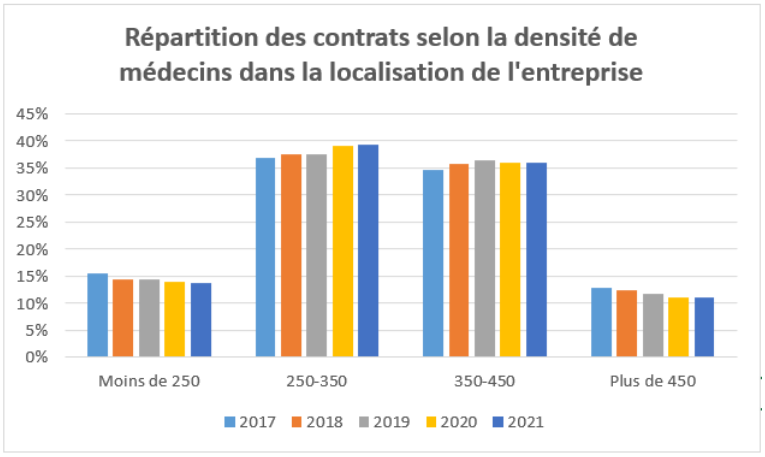


FIGURE 14.9 – Analyse descriptive de la densité totale de médecins

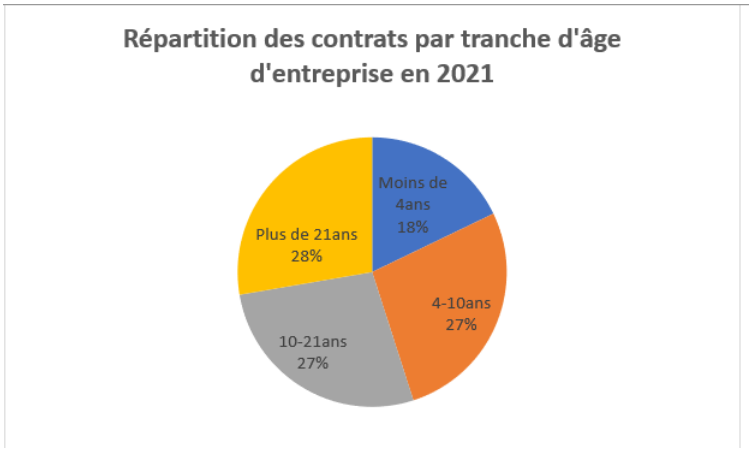


FIGURE 14.10 – Analyse descriptive de l'âge de l'entreprise

14.5 Représentation des résidus du GLM Tweedie en fonction des variables explicatives

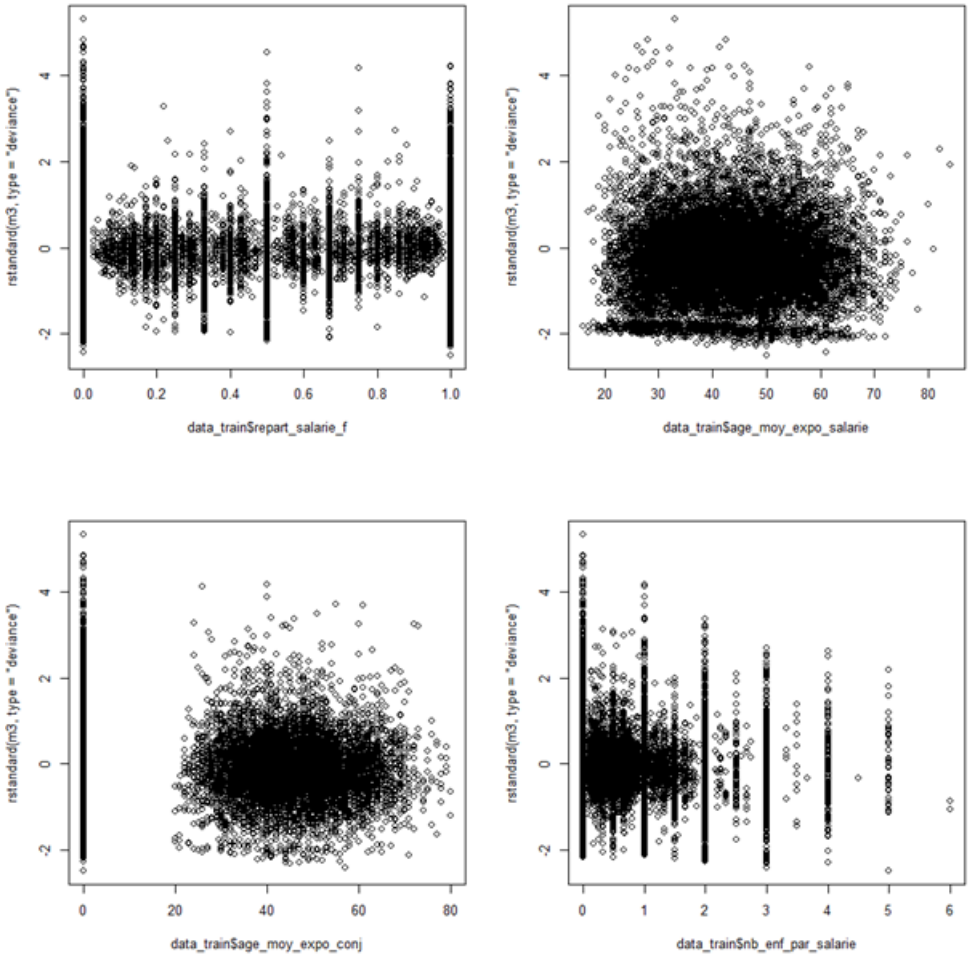


FIGURE 14.11 – Représentation graphique des résidus en fonction des variables explicatives

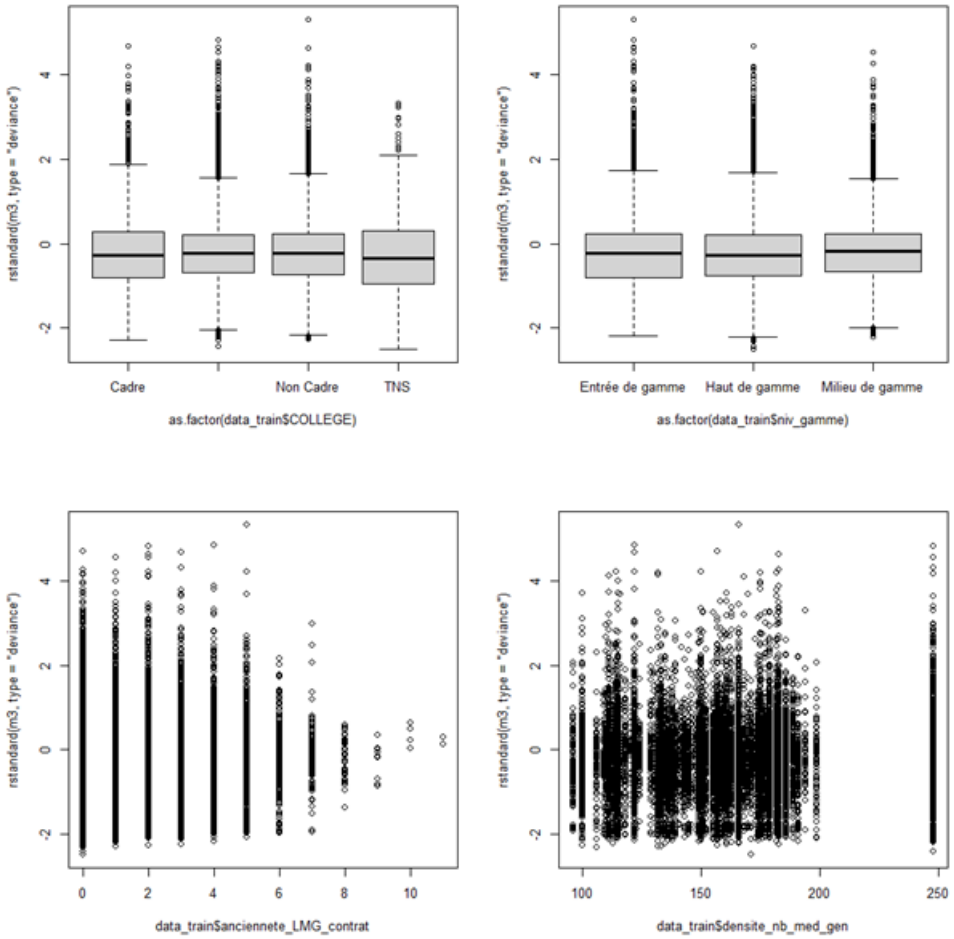


FIGURE 14.12 – Représentation graphique des résidus en fonction des variables explicatives

Liste des abréviations

LMG	La Mutuelle Générale
GLM	Generalized Linear Model
SS	Sécurité Sociale
BR	Base de Remboursement
CSBM	Consommation de biens met soins médicaux
CCN	Convention Collective Nationale
PLV	Plafond Limite de Vente
PMSS	Plafond Mensuel de la Sécurité Sociale
RAC	Reste A Charge
TM	Ticket Modérateur
ANI	Accord National Interprofessionnel
TPE	Très Petites Entreprises
PME	Petites et Moyennes Entreprises
APE	Activité Principale Exercée
DREES	Direction de la Recherche, des Études, de l'Évaluation et des Statistiques
RO	Régime Obligatoire
RC	Remboursement Complémentaire
CCAM	Classification Commune des Actes Médicaux
ACPR	Autorité de Contrôle Prudentiel et de Résolution
INSEE	Institut National de la Statistique et des Études Économiques
SE	Santé Entreprise
CAH	Classification Ascendante Hiérarchique
AIC	Akaike Information Criterion

Table des figures

Bibliographie

1. NGUYEN N. [2013] « Construction de bases de tarification pour des contrats complémentaires santé collectifs par le Modèle Linéaire Généralisé ». Mémoire d'actuaire - ISFA.

2. GOURLIER S. [2014] « Analyse de la rentabilité d'un produit en santé individuelle ». Mémoire d'actuaire - Paris Dauhpine.

3. BONNIFAIT C. [2019] « Optimisation d'un outil de tarification santé destiné au pilotage des grands comptes et Branches professionnelles ». Mémoire d'actuaire - EURIA.

3. ABDOLLAHI F. [2017] « Tarification d'une complémentaire santé à destination des séniors, modulaire par poste de garanties et l'impact sur la solvabilité ». Mémoire d'actuaire - ISUP.

4. VAUTRIN M. [2010] « Élaboration d'une méthode de tarification avec indicateurs de risque pour des contrats complémentaires santé collectifs ». Mémoire d'actuaire - ISUP.

5. EBER V. [2017] « Pilotage et étude de rentabilité d'un contrat de prévoyance collective ». Mémoire d'actuaire - DUAS.

6. PESNEAUD A. [2019] « Création de zoniers en asssurance habitation à l'aide de variables externes et de méthodes de Data Science ». ISUP.

7. DRESS (2021). Comptes nationaux de la santé.

8. CHARPENTIER A. & DENUIT M. « Mathématiques de l'assurance non-vie, Tarification et provisionnement » - ECONOMICA.

9.- Tweedie, M. C. K. (1984), An index which distinguishes between some important exponential families. Statistics : Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference (Eds. J. K. Ghosh and J. Roy), pp. 579-604. Calcutta : Indian Statistical Institute.

10. Agresti, A. (2015), Foundations of Linear and Generalized Linear Models. Wiley series in Probability and Statistics

11. MONNIER D. Modèles linéaires généralisés et assurance santé individuelle : Tarification et évaluation des engagements sous solvabilité II