Mémoire présenté le :

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

Par : Alice LESBATS

Titre Modélisation de la consommation en Santé par Machine Learning

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

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : Sham

Signature :

Directeur de mémoire en entreprise :

Nom : Sylvain CARACO

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Institut de Science Financière et d'Assurances
Université Claude Bernard Lyon 1

Modélisation de la consommation en Santé par Machine Learning

Mémoire de Master* écrit par Alice LESBATS.

Année 2020-2021

Tuteur entreprise : Sylvain Caraco
Tuteur ISFA : Nicolas Leboisne

∗. Ce mémoire est confidentiel et s'adresse uniquement aux membres du jury.

Table des matières

Introduction **8**

1 Mise en contexte **9**
1.1 Le système de protection sociale français . . . 9
1.1.1 Point d'histoire et définitions . . . 9
1.1.2 La Sécurité Sociale . . . 10
1.1.3 Les acteurs complémentaires . . . 10
1.1.4 Les types de contrats des complémentaires santé . . . 11
1.1.5 Les évolutions en Santé . . . 11
1.1.6 Le remboursement des prestations santé . . . 13
1.2 La tarification d'un contrat d'assurance . . . 14
1.3 Présentation des données disponibles . . . 15
1.4 Traitement des données . . . 16
1.4.1 Base Effectif . . . 16
1.4.2 Base Sinistres . . . 17
1.4.3 Base finale . . . 19

2 Analyse des données **22**
2.1 Description des portefeuilles . . . 22
2.1.1 Portefeuille 1 . . . 22
2.1.2 Portefeuille 2 . . . 24
2.2 Gestion des sinistres graves . . . 26
2.2.1 Observation des extrêmes . . . 26
2.2.2 Détermination du seuil . . . 29
2.2.3 Choix du seuil . . . 37

3 Approche théorique de la modélisation de la consommation **38**
3.1 L'apprentissage automatisé . . . 38
3.2 Mesures de performance . . . 39
3.3 Modèle linéaire généralisé . . . 39
3.3.1 Tarification en santé . . . 41
3.3.2 Estimation des paramètres . . . 42
3.3.3 Validation . . . 42
3.4 Autres modèles de machine learning . . . 43
3.4.1 Arbre de décision (CART) . . . 43
3.4.2 Boosting . . . 46
3.4.3 Descente de gradient . . . 48
3.4.4 Gradient Boosting Machine . . . 49
3.4.5 eXtreme Gradient Boosting . . . 51

4 Application des méthodes **55**
4.1 Etude des dépenses pour chaque variable . . . 55
4.2 Modèle linéaire généralisé . . . 67
4.2.1 Modélisation GLM Tweedie . . . 67
4.2.2 Modélisation GLM Fréquence - Coût moyen . . . 75
4.3 eXtreme gradient boosting . . . 91

4.3.1 Portefeuille 1 . . . 91
4.3.2 Portefeuille 2 . . . 91
4.4 Analyse et comparaison des résultats . . . 92
4.4.1 Portefeuille 1 . . . 92
4.4.2 Portefeuille 2 . . . 93
4.4.3 Conclusion sur la comparaison des résultats . . . 94

5 Application de tarification **95**
5.1 Paramètres . . . 95
5.2 Résultats . . . 96

Conclusion **97**

Bibliographie **98**

Glossaire **101**

Remerciements

Je souhaiterais remercier sincèrement toutes les personnes qui ont de près ou de loin participé au bon déroulement de mon alternance.

En premier lieu, je tiens à remercier toutes les personnes du Département Actuariat de SHAM pour leur accueil amical, les échanges enrichissants et leur coopération professionnelle.

Je voudrais remercier plus particulièrement :

Monsieur Sylvain CARACO mon tuteur entreprise pour son accueil, pour m'avoir permis de profiter de son expérience, pour ses conseils de qualité, pour sa disponibilité, pour la clarté de ses explications et la confiance qu'il a placée en moi en me permettant d'effectuer mon alternance au sein de SHAM dans le Département Actuariat.

Madame Chaima IDANI ma collègue de travail pour son accueil, son aide et son attention pour mon mémoire.

Je tiens à remercier le corps enseignant de l'Institut de Science Financière et d'Assurances pour le savoir et la connaissance que j'ai reçus depuis le début de ma formation.

Et plus particulièrement, j'adresse mes remerciements à Nicolas LEBOISNE mon tuteur pédagogique pour son encadrement et ses conseils.

Résumé

Dans ce mémoire, nous proposons d'utiliser différentes méthodologies pour déterminer la prime pure à des fins de tarification pour un assuré d'un contrat d'assurance en santé collective et de connaître les paramètres influents.

Cette étude est menée sur deux portefeuilles différents et sur deux types de montants (en frais réels et en montant remboursé par la mutuelle). Nous exploitons la théorie des modèles linéaires généralisés qui sont paramétriques et celle de l'algorithme de Machine Learning eXtreme Gradient Boosting qui est non-paramétrique.

Nous proposons deux méthodes pour notre modélisation avec les modèles linéaires généralisés. Dans un premier temps, nous modélisons avec une régression Tweedie tous postes confondus sans décomposition du risque ce qui revient à modéliser la prime pure. Dans un second temps, nous considérons que chaque sous-poste possède une consommation unique avec des paramètres influents différents. Ainsi, nous modélisons la prime pure avec une approche fréquence-coût moyen par sous-poste. En ce qui concerne la méthode non-paramétrique nous modélisons la prime pure avec l'eXtreme Gradient Boosting.

Dans un contexte d'utilisation en entreprise, nous avons développé une application R Shiny permettant de modéliser la prime pure pour un assuré en fonction de ses caractéristiques pour certaines méthodes choisies.

Mots clés : Assurance Santé, Modèles Linéaires Généralisés, Santé Collective, Régression Tweedie, Approche Fréquence-coût moyen, Machine Learning, GLM, XGBoost, R Shiny, Prime Pure, Remboursement de la part de la mutuelle, frais réels.

Abstract

In this case study, we propose using various methodologies to determine the loss cost for pricing a group health insurance policyholder and to know the influential parameters.

This study is carried out across two different portfolios and on two amount types (real expenses and amount reimbursed by the mutual insurance company). We exploit the theory of generalized linear models which are parametric and the Machine Learning's algorithm named eXtreme Gradient Boosting which is non-parametric.

We propose two methods for our modelling with generalized linear models. Firstly, we model with an all-act Tweedie regression without risk decomposition, which is equivalent to modelling the loss cost. Secondly, we consider that each sub-act has a unique consumption level with different influential parameters. Thus, we model the pure premium with a frequency-cost approach per act. As for the non-parametric method, we model the loss cost with the eXtreme Gradient Boosting.

In a business context, we have developed an R Shiny application allowing to model the pure premium for an insured according to his characteristics for some selected methods.

Keywords : Health insurance, Generalized Linear Models, Collective Health, Tweedie Regression, Frequency-cost method, Machine Learning, GLM, XGBoost, R Shiny, Pure Premium, Insurance company's reimbursement, real expenses.

Introduction

La santé est un domaine d'étude intéressant pour les assureurs, mais aussi pour les assurés. Il est parlant pour tout le monde, puisqu'il touche toute la population. Les moyens actuels et la masse de données disponibles permettent de réaliser de nombreuses études exploratoires. Il est important pour un organisme complémentaire de connaître les paramètres les plus influents pour la tarification de ses contrats. Dans un contexte juridique instable et en constante évolution, l'assureur doit avoir à l'esprit les fondamentaux de la consommation de ses assurés.

Ce mémoire va tenter de répondre à cette problématique pour SHAM en étudiant la consommation en santé de deux portefeuilles différents avec des modèles linéaires généralisés et un algorithme de machine learning. Le sujet est de pouvoir estimer la prime pure en terme de frais réels et de montant remboursé par la mutuelle avec trois méthodes. Une modélisation de la prime pure sera effectuée avec la loi Tweedie qui est couramment utilisée dans les pays anglo-saxons et par sous-poste en utilisant la méthode fréquence-coût moyen. Puis, avec une méthode alternative utilisant l'eXtreme Gradient Boosting qui fait partie de la famille des algorithmes de Machine Learning.

Dans un premier temps, nous ferons une mise en contexte avec les fondamentaux de la protection sociale en France et nous verrons la présentation d'un remboursement de frais de santé. Nous expliquerons le fonctionnement d'un contrat d'assurance, puis nous présenterons les données qui étaient disponibles pour cette étude et les traitements qu'elles ont pu subir.

Dans un deuxième temps, nous exposerons l'analyse des données avec une description des deux portefeuilles et la présentation de la gestion des sinistres graves. Dans un troisième temps, nous introduirons de façon théorique les modèles linéaires généralisés, ainsi que des algorithmes de machine learning plus complexes. Nous présenterons également les mesures de performance pour les comparer avec une application dans le cas de la tarification en santé collective.

Dans un quatrième temps, nous analyserons les résultats obtenus avec un cas pratique sur nos données pour les deux portefeuilles avec deux montants différents (en frais réels et en montant remboursé par la mutuelle). Finalement, nous nous intéresserons à l'application R Shiny créée pour ce mémoire permettant de visualiser nos résultats.

1 Mise en contexte

1.1 Le système de protection sociale français

Dans cette première section, nous allons présenter le système de protection sociale français. Tout d'abord, nous nous intéresserons à son origine, sa définition et ses objectifs en France. Puis, nous aborderons les différents éléments qui composent la Sécurité Sociale et son rôle au sein de la protection sociale. De plus, nous verrons les organismes complémentaires qui travaillent avec la Sécurité Sociale et les principales réformes en Santé. Finalement, nous expliquerons comment se déroule le remboursement d'une prestation santé.

1.1.1 Point d'histoire et définitions

La protection sociale regroupe deux grandes notions qui permettent de la définir : l'assurance et la solidarité. Pour l'assurance, il s'agit de se couvrir d'un risque qui peut se produire avec un aléa. Alors que la solidarité est une obligation d'aide, d'assistance ou de collaboration existant entre plusieurs personnes d'un groupe, ce qui ne couvre pas contre un risque.

Ainsi, la protection sociale est un ensemble d'institutions ayant pour objectif la protection des individus contre les conséquences de divers événements ou situations souvent qualifiés de « risques sociaux ». Elle offre une assistance financière aux bénéficiaires qui rencontrent des événements coûteux de la vie. Il peut s'agir du risque d'accident du travail, d'invalidité, de chômage, ou encore de maladies auxquelles nous allons particulièrement nous intéresser dans ce mémoire.

En 2020, les dépenses pour la protection sociale en France s'élevaient à 470 milliards d'euros soit plus que le budget de l'Etat (soit 350 milliards d'euros) et correspondant à environ 25% du PIB (qui était d'environ 2 000 milliards d'euros) ce chiffre étant en constante augmentation [3].

A partir du 17ème siècle, nous pouvons retrouver certaines notions de protection sociale avec la création du premier « régime de retraite » pour les marins en France (en 1673). Mais ce n'est qu'en 1898 que la protection sociale prend réellement son origine avec la première loi d'assurance sociale sur les accidents du travail.

En France, le système de protection sociale a pour objectif de créer une unité et un accès généralisé à la sécurité sociale, puis une extension des risques couverts. La notion que nous connaissons actuellement a été définie après la Seconde Guerre Mondiale en 1945 avec les ordonnances du 4 et 19 octobre créant ainsi l'organisation de la Sécurité Sociale. Ces ordonnances ont fusionné toutes les anciennes assurances (maladie, retraite,...) pour créer un réseau coordonné de caisses. Il persiste à y avoir tout de même des régimes spéciaux (fonctionnaires, marins, cheminots, mineurs, etc...) qui refusent à l'époque de s'intégrer au nouveau régime général. Un cadre qui était dit « transitoire » et qui perdure encore de nos jours.

1.1.2 La Sécurité Sociale

En France, le premier niveau de protection sociale est tenu par la Sécurité Sociale. Un régime est un ensemble de droits et obligations réciproques des Employés et leurs « ayant-droit », des Patrons et d'une Caisse de Sécurité Sociale. Il existe trois régimes de base qui sont les suivants :

- Général (géré par la CNAM – Caisse Nationale d'Assurance Maladie) : salariés et travailleurs assimilés à des salariés représentant environ 80% de la population française.
- Travailleurs non salariés non agricoles (RSI – SSI géré aussi par la CNAM) : artisans, commerçants et professions libérales.
- Agricole : exploitants et salariés agricoles, ainsi que certains secteurs rattachés à l'agriculture (géré par la MSA – Mutualité Sociale Agricole).

Il existe toujours des régimes spéciaux aussi gérés par la CNAM (qui refusent de s'intégrer au régime général) comme :

- Caisse prévoyance et retraite SNCF (CPR SNCF),
- Caisse d'assurance vieillesse, invalidité et maladie des cultes (CAVIMAC),
- Caisse de retraite et de prévoyance des clercs et employés de notaires (CRPCEN),
- Régime Alsace-Moselle,etc.

Chacun des régimes au sein de la Sécurité Sociale est organisé en branches. Par exemple, pour le régime général, il est organisé en quatre branches autonomes responsables de ses ressources et de ses dépenses :

- « Maladie » (maladie, maternité, paternité, invalidité, décès) gérée par l'Assurance Maladie,
- « Accidents du travail et Maladies professionnelles »,
- « Vieillesse et veuvage » (retraite),
- « Famille » (dont handicap, logement, RSA, etc.).

1.1.3 Les acteurs complémentaires

Le marché de l'assurance complémentaire est composé de différents acteurs comme les organismes assureurs, les entreprises de conseil, les gestionnaires et les courtiers. Par exemple en Santé, ils assurent la prise en charge d'une partie ou de la totalité des prestations santé ou viennent en supplément des prestations de l'Assurance Maladie obligatoire.

Au sein des organismes complémentaires, nous pouvons retrouver trois familles : les mutuelles, les Institutions de Prévoyance (IP) et les Sociétés d'assurance. Elles sont régies par des réglementations différentes, mais restent tout de même relativement similaires en terme de « technique et gestion financière ». Elles ont chacune des ensembles d'intervention privilégiés.

Les cabinets ou les entreprises d'audit et de conseil apportent une assistance à l'assureur ou au courtier et ne sont pas des intermédiaires au contrat. Leurs domaines d'activité sont variés comme la retraite, la prévoyance, la santé ou encore bien d'autres secteurs.

Les courtiers apporteurs sont des travailleurs indépendants, ils ne sont pas liés à une compagnie d'assurance et ils servent d'intermédiaires au détenteur du contrat plus précisément entre l'entreprise et l'organisme assureur. Ils sont rémunérés par commission d'apport.

En ce qui concerne les gestionnaires, ils s'occupent de la gestion admnistrative des contrats. Il existe deux types de gestionnaires, soit ils sont rattachés de près ou de loin à un cabinet de courtage, soit ils sont indépendants. Les gestionnaires ne sont pas des intermédiaires, mais des tierces-parties.

1.1.4 Les types de contrats des complémentaires santé

Dans cette sous-partie, nous présentons les différents types de contrat qui peuvent être souscrits auprès d'un organisme de complémentaire santé.

- **Les contrats individuels** : peuvent être directement souscrits par un individu. Ils concernent les étudiants, les fonctionnaires, les chômeurs, les retraités, les indépendants et les salariés du secteur privé souhaitant souscrire à une sur-complémentaire.
- **Les contrats collectifs :** sont souscrits par un employeur au titre de ses salariés.
  - **A adhésion obligatoire :** signifie que les salariés d'une entreprise ont l'obligation de souscrire au contrat proposé. Suite à l'Accord National Interprofessionnel du 14 juin 2013, il avait été décidé de mettre en place au 1er janvier 2016 ce type de contrat en obligeant les entreprises du secteur privé à proposer une couverture de complémentaire santé et que l'employeur prenne à sa charge un minimum de 50% de la cotisation. Dans certains cas particuliers, le salarié peut se voir dispenser de cette adhésion, mais ça ne remettra pas en question l'aspect collectif et obligatoire de ce type de contrat.
  - **A adhésion facultative (« sur-complémentaire ») :** signifie que le contrat est souscrit par l'entreprise, mais que l'adhésion reste à titre individuel.

1.1.5 Les évolutions en Santé

Dans cette-sous partie, nous présentons les principales réformes en Santé. Nous nous intéresserons à la portabilité des droits et aux contrats collectifs obligatoires. Puis, nous verrons l'évolution du cahier des charges des contrats « responsables et solidaires ».

- **Portabilité des droits :**
  A l'issue de l'Accord National Interprofessionnel (ANI) du 19 juin 2013, l'article L911-8 du Code de la Sécurité Sociale est venu ajouter la portabilité des droits des salariés ce qui leur permet de conserver le droit aux garanties prévues au contrat collectif de l'entreprise précédente.

- **Contrat collectif obligatoire :**
  La loi Fillon de 2003 a mis en place le caractère obligatoire des contrats collectifs. Elle oblige à avoir des garanties identiques à l'intérieur de chaque catégorie, d'obtenir une participation de l'employeur qui doit être à un taux ou montant uniforme pour l'ensemble

des salariés (sauf pour quelques cas particuliers comme les apprentis ou salariés en temps partiel par exemple), et de respecter la couverture de l'ensemble des salariés. Il y a plusieurs critères qui déterminent les catégories : la catégorie socio-professionnelle, la tranche de rémunération, la convention collective et d'autres encore.

- **La notion de contrats « responsables et solidaires » :**
  En 2002, les contrats « solidaires » ont été établis. Ils sont accessibles à tous, car ils ne doivent pas selon l'état de santé des assurés les discriminer, ou modifier leur cotisation. La loi Douste-Blazy du 13 août 2004 relative à l'Assurance Maladie a introduit les contrats dits « responsables ». Cette loi a défini le parcours de soin coordonnés autour des médecins traitants. Elle a aussi instauré des franchises et la contribution forfaitaire de 1€ par acte. Les contrats « responsables » sont entrés en vigueur au 1er avril 2015 avec le décret 2014-1374 paru le 18 novembre 2014.

  La mise en place des contrats « reponsables » pour les garanties individuelles et les contrats collectifs est apparue à partir du 1er janvier 2016. Par ailleurs, ces contrats doivent par exemple :

  - Prendre en charge l'intégralité du ticket modérateur pour tous les actes sauf pour l'homéopathie, les cures thermales et les médicaments remboursés à 30% et à 15%.
  - Ne pas prendre en charge la participation forfaitaire de 1€.
  - Prendre en charge intégralement et sans limite de durée le forfait journalier hospitalier.
  - Ne pas prendre en charge les franchises médicales.
  - Ne pas prendre en charge les majorations de participation sanctionnant l'absence de choix ou de recours au médecin traitant.
  - Ne pas prendre en charge les dépassements d'honoraires sur les actes cliniques et techniques autorisés si non respect du parcours de soins.
  - Prendre en charge tous les 2 ans une paire de lunette pour les adultes et tous les ans pour les mineurs avec justificatif. Des niveaux de garanties planchers et des plafonds par rapport au niveau de correction doivent être mis en place. De plus, le remboursement des montures ne doit pas dépasser 150€.
  - Il existe encore diverses obligations.

  Nous pouvons noter plusieurs avantages pour ce type de contrat. Tout d'abord, le taux de taxation des cotisations qui est de 13,27%, alors qu'il est de 20,27% pour les contrats non-responsables. Ensuite, la part patronale des cotisations est exonérée de charges sociales. Par ailleurs, nous pouvons citer l'avantage pour le salarié de la déduction de sa part de cotisation de son revenu imposable. De nos jours, quasiment la totalité des contrats de santé sont responsables et solidaires. Le gouvernement a beaucoup participé à leur généralisation avec leurs avantages fiscaux et sociaux.

- **Evolution du cahier des charges des contrats « responsables » :**
  La réforme 100% Santé a fait évoluer la notion de contrat « responsable » avec une obligation de reste à charge nul pour les actes du panier de soins dont les domaines sont

précisés dans l'article L. 871-1 du code de la Sécurité Sociale. Les postes de santé qui sont concernés sont l'optique, le dentaire et les aides auditives. L'objectif de cette réforme est de faciliter l'accès aux soins, mais aussi de simplifier la compréhension et la comparaison des garanties complémentaires santé.

Depuis le 1er janvier 2021, les évolutions notables pour les aides auditives sont observables avec un délai minimum avant renouvellement à 4 ans par oreille et un plafond de remboursement de 1 700€ par oreille. A partir de 2020, ses évolutions sur le poste optique ont été mises en place avec par exemple l'âge maximum des enfants abaissé à 16 ans, les modifications des conditions de renouvellement anticipé, ou encore l'abaissement de remboursement sur les montures à 100€. Finalement, pour le dentaire nous pouvons citer des améliorations comme la prise en charge des dépassements d'honoraires de soins dentaires prothétiques et des soins d'orthopédie dento faciale. Finalement, des paniers de soins ont été créés pour chacun de ces postes. Pour l'optique et les aides auditives, nous retrouvons deux paniers : le panier 100% santé et le panier à tarif libre. Puis, pour le dentaire, il y a trois paniers : le panier 100% santé, le panier maîtrisé et le panier à tarif libre.

1.1.6 Le remboursement des prestations santé

Dans le cadre de ce mémoire, nous nous focalisons sur le risque santé. Il est donc important de comprendre la composition d'un remboursement de prestation santé. Nous allons aborder quelques définitions importantes pour la compréhension des remboursements de frais de santé :

- Base de remboursement (BR) : tarif de base utilisé par la Sécurité Sociale pour effectuer le remboursement des honoraires et soins dispensés par les praticiens.
- Tarif de convention (TC) : correspond au tarif de responsabilité pour le secteur conventionné.
- Tarif d'autorité (TA) : correspond au tarif de responsabilité pour le secteur non conventionné, mais pour une valeur nettement inférieure.
- Ticket modérateur (TM) : part de la base de remboursement qui reste à la charge de l'assuré après le remboursement de la Sécurité Sociale.

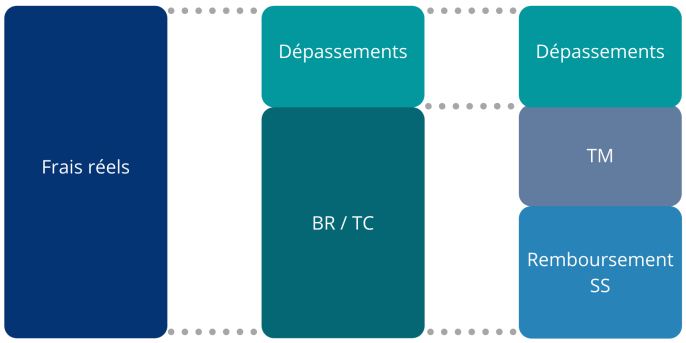

FIGURE 1 – Schéma général d'une prestation santé

Finalement, la base de remboursement est composée de deux éléments : le ticket modérateur et le remboursement sécurité sociale. La part de remboursement de la sécurité sociale varie d'une prestation à l'autre. Par exemple pour les actes médicaux, il s'agira de 70% de la base de remboursement, ou encore de 80% pour l'hospitalisation, etc. Le ticket modérateur et certains dépassements sont pris en charge par la mutuelle.

De plus, il existe une codification appelée « Nomenclature » (par exemple, NGAP - Nomenclature générale des actes professionnels, il en existe encore bien d'autres) afin de classer les différentes prestations. Par exemple, pour une analyse médicale en biologie (prise de sang), le code pour la nomenclature NGAP est « B ».

Prenons l'exemple d'une personne effectuant une consultation chez un généraliste conventionné. Les frais réels engendrés sont de 25 euros pour cette personne. Ces derniers sont composés de 24 euros de la base de remboursement auxquels s'ajoutent 1 euro de participation forfaitaire. La Sécurité Sociale remboursera 70% de la base de remboursement ce qui représente 16,5 euros. Le ticket modérateur vaudra 25 euros moins 17,5 euros, ce qui revient à 7,5 euros qui sera remboursé par la complémentaire santé. Finalement, il n'y a que 1 euro de reste à charge pour la personne ayant reçu les soins.

1.2 La tarification d'un contrat d'assurance

Un contrat (aussi appelé police) d'assurance possède un assuré qui est le détenteur du contrat et un assureur qui est le pourvoyeur du contrat et porteur du risque. Parfois, il peut y avoir une troisième partie avec un bénéficiaire. Ainsi, en échange de la couverture d'un risque par l'assureur, l'assuré verse une prime d'assurance. Donc, si un sinistre se produit, alors le bénéficiaire du contrat reçoit le montant contractuel prévu. Au départ, l'assuré supporte le risque économique, puis il sera transféré à l'assureur une fois que l'assuré aura souscrit à un contrat.

Dans un échantillon aléatoire d'une taille tendant vers l'infini, la loi des grands nombres permet de dire que nous nous rapprochons des caractéristiques statistiques de la population. En assurance, cette loi offre la possibilité de mutualiser les risques en souscrivant de nombreux contrats pour une même compagnie d'assurance.

Chaque portefeuille d'assurance couvre un risque précis, les pertes sont donc considérées de même loi de probabilité (indépendantes et identiquement distribuées). Ainsi, une tarification par garantie (par exemple en santé, pour les verres, puis les montures, etc) doit être réalisée. Néanmoins, les contrats sont a priori indépendants les uns des autres. L'assureur sera en capacité de prévoir avec une précision relative, les potentielles dépenses pour une certaine période.

Soit un portefeuille d'assurance contenant I polices. Notons la loi des dépenses du i^e contrat Y_i, et la loi des dépenses agrégées Y_I.

La loi forte des grands nombres nous dit que la convergence presque sûre de la moyenne empirique des pertes, notée $\bar{Y_I} = \frac{1}{I}\sum_{i=1}^{I} Y_i$ converge vers l'espérance de la loi soit :

$$\bar{Y_I} \to^{p.s.}_{n\to\infty} \mathbb{E}(Y_i) = \mu$$

Ce résultat pose les bases de la tarification avec la prime qui vaut au moins μ (appelée prime pure du contrat) et il peut être utilisé si nous avons un grand nombre d'observations pour appliquer la loi forte des grands nombres. Il s'agit de cette prime que nous modéliserons. La prime qu'un assuré paye s'appelle la prime commerciale.

La prime d'assurance aussi appelée prime commerciale est composée de plusieurs éléments :

- Prime technique correspond à la prime pure avec des chargements techniques.
- Prime commerciale (prime finale) englobe la prime technique à laquelle s'ajoute la rémunération des intermédiaires (courtiers, gestionnaires,etc.) comme des frais d'acquisition, de gestion, de rémunération, de distribution, d'administration, etc. Elle contient aussi une marge bénéficiaire définie par l'assureur.

L'assureur va essayer de créer une tarification segmentée, mais qui ne le sera pas trop pour conserver le principe de mutualisation.

1.3 Présentation des données disponibles

L'assureur possède deux portefeuilles en assurance santé collective qui sont distincts, le premier est à adhésion obligatoire et le deuxième à adhésion facultative. C'est pour cette raison que nous avons séparé l'étude en deux pour chacun des portefeuilles. Le premier est relatif au personnel hospitalier de droit privé. Tandis que le second est composé par des agents relevant de la Fonction Publique d'Etat. Par ailleurs, SHAM fait appel à de la gestion déléguée, ce qui explique que les données accessibles soient différentes pour chaque portefeuille, puisqu'ils n'ont pas le même gestionnaire. Dans les deux cas, nous avions accès à deux bases qui se rapportent à la période 2018-2020 et seulement aux actifs avec leurs ayants droit (puisque les individus partant à la retraite ont un comportement encore différent et la grande majorité est représentée par les actifs) :

- La base « Effectif » : contient l'ensemble des personnes assurées entre 2018 et 2020, ainsi que les informations relatives à l'assuré telles que : son sexe, sa date de naissance, puis son identifiant (permettant de distinguer les bénéficiaires de façon unique).
  - Pour le personnel de santé, nous pouvons également retrouver : le code NAF ou APE (attribué par l'Insee lors de l'immatriculation ou la déclaration d'activité de l'entreprise et qui change en fonction du secteur d'activité) de l'entreprise dans laquelle l'actif travaille, le département de résidence de l'assuré principal, le numéro de l'entreprise (enregistré dans notre système d'information) et la date mensuelle de présence de l'assuré dans le portefeuille.
  - Pour le personnel de la Fonction Publique d'Etat, nous avons la catégorie du bénéficiaire (actif, conjoint, ou ayant-droit), le niveau de garantie du contrat souscrit, la date de souscription au contrat, et la date de fin du contrat éventuellement.

- La base « Sinistres » : contient l'ensemble des prestations de santé engagées pendant la période 2018-2020 avec toutes les données nécessaires comme : un identifiant unique pour le bénéficiaire des soins, le code acte (permettant de coder les actes médicaux), le libellé décrivant l'acte effectué, la quantité d'actes consommée, le montant auquel s'élève les frais réels, le montant du remboursement de la part de la Sécurité Sociale, le montant du remboursement de la part de la mutuelle, nous pouvons retrouver aussi la date de soin et la date de paiement.
  - Pour le personnel de santé, il est ajouté une variable dans sa base qui est le numéro de l'entreprise associé au bénéficiaire ayant reçu les soins.
  - Pour le personnel de la Fonction Publique d'Etat, nous pouvons retrouver aussi le niveau de garantie, ainsi que la catégorie du bénéficiaire. Pour ce portefeuille, nous n'avons pas la date précise de soin ou de règlement, nous avons seulement accès au mois et à l'année.

1.4 Traitement des données

Dans cette partie, nous allons expliquer comment nous avons transformé nos données pour permettre leur exploration et pour la suite de notre étude.

1.4.1 Base Effectif

Dans cette sous-partie, nous nous focalisons sur la base « Effectif ». En premier lieu, nous avons enlevé des doublons dans notre base au niveau de différentes variables (tels que l'âge, le sexe, le département, et le code NAF de l'entreprise s'il était présent), car il pouvait y avoir des erreurs de saisie. Par ailleurs, pour le portefeuille du personnel de santé, nous avions créé une variable correspondant à la catégorie du bénéficiaire. Il existait un rang (1 : pour l'assuré principal, 2 : pour le conjoint et 3 : pour l'enfant) qui était contenu dans l'identifiant unique.

Ensuite, nous avons calculé l'exposition par assuré. L'exposition représente le temps en année durant lequel l'assuré cotise, c'est-à-dire le ratio du nombre de mois cotisés pour une année de couverture divisé par 12. Elle permettra de corriger par la suite la fréquence de consommation avec sa durée de couverture.

Par exemple, si un assuré n'est présent que 6 mois (soit 0.5 d'exposition) pendant l'année de couverture et va deux fois chez un spécialiste alors sa fréquence annuelle vaudra 4. Il faut savoir aussi que le calcul de l'exposition peut faire apparaître des valeurs extrêmes avec des personnes présentes très peu de temps dans le portefeuille. Comme pour une personne qui ne serait présente dans notre portefeuille que 3 mois (soit 0.25 d'exposition) et va 6 fois chez le médecin. Alors elle aura une fréquence annuelle de 24.

Finalement, nous avons aussi ajouté le nombre de bénéficiaires par entreprise lorsque nous en avions la possibilité.

1.4.2 Base Sinistres

Dans cette sous-partie, nous nous intéressons au traitement des bases « Sinistres ». Tout d'abord, nous retrouvons les frais réels, les remboursements de la part mutuelle et de la Sécurité Sociale, ainsi que la quantité consommée, tout ceci par assuré, par code acte et par année de soin. Quand le champ de numéro de sinistre est disponible dans la base de données, nous ajoutons un sous-total.

Il y a des sinistres de 2018 ou 2019 qui peuvent ne pas être encore déclarés ou qui n'ont pas été encore remboursés. En santé, il y a essentiellement une provision qui s'apparente à des IBNR (Incured But Not yet Reported), mais nous l'appelons généralement PSAP (Provisions pour Sinistres A Payer). Il existe un deuxième type de provision qui est très rare, la provision pour risque croissant (PRC), elle est destinée à couvrir des engagements pour des retraités lorsque le tarif est limité. La compagnie d'assurances SHAM n'est pas concernée par les PRC en Santé. Ainsi, nous pouvons définir les PSAP et IBNR de la façon suivante :

- PSAP : les sinistres qui sont connus par l'assureur, mais qui ne sont pas encore réglés.
- IBNR : les sinistres qui n'ont pas encore été déclarés à l'assureur.

Avant d'exposer le cas pratique sur nos données, nous allons présenter de façon théorique la méthode de Chain-Ladder qui va être utilisée par la suite. Nous supposons que les sinistres sont sur $n+1$ mois. Notons :

- i : mois d'origine ($i = 0, ..., n$) cela correspondra au mois de survenance du sinistre.
- j : mois de développement ($j = 0, ..., n$) cela correspond au mois de règlement du sinistre.
- x_{ij} : le coût de la sinistralité correspondant au mois d'origine i et au mois de développement j.

Les données x_{ij} se présentent sous forme matricielle dans un triangle de liquidation des sinistres. Si nous nous plaçons dans le cas pratique de nos données. Nous avons écarté la survenance 2020 de notre étude, car elle possède un comportement atypique dû aux différents confinements liés au Covid-19 et aux changements avec la réforme 100% santé. La période de soins qui a été étudiée correspond aux années 2018 et 2019. Notre calcul de triangle de liquidation des sinistres a été effectué sur 24 mois de développement. Dans le tableau 1, nous représentons de manière théorique nos triangles de règlements incrémentaux. Nous avons effectué un triangle de règlements incrémentaux par sous-poste pour chacun de nos portefeuilles. Notons que le $n+1$ vaudra 24.

Mois d'origine	**0**	**1**	**...**	**j**	**...**	**n-i**	**...**	**n-1**	**n**
0	x_{00}	x_{01}	**...**	x_{0j}	**...**	$x_{0,n-i}$	**...**	$x_{0,n-1}$	x_{0n}
1	x_{10}	x_{11}	**...**	x_{1j}	**...**	$x_{1,n-i}$	**...**	x_{1n}	
...	...	...	...	...	...	...	...		
i	x_{i0}	x_{i1}	**...**	x_{ij}	**...**	$x_{i,n-i}$			
...	...	...	...	...	...				
n-j	$x_{n-j,0}$	**...**	**...**	$x_{n-j,n}$					
...	...	...	...						
n-1	$x_{n-1,0}$	$x_{n-1,1}$							
n	$x_{n,0}$								

TABLE 1 – Triangle de liquidation des sinistres.

Ainsi, les x_{ij} correspondent aux paiements incrémentaux, mais nous allons utiliser un triangle de paiements cumulés. Notons :

$$C_{ij} = \sum_{h=0}^{j} x_{ih} \tag{1}$$

Cela correspond aux paiements effectués jusqu'au mois de développement j pour les sinistres qui se sont déroulés pendant le mois i.

Avec nos données, nous avons réalisé notre étude avec la méthode Chain-Ladder sur les cadences de règlements cumulés des années 2018 et 2019. Nous avons réalisé ces triangles de règlements cumulés par sous-poste. Ainsi, si nous les présentons de façon théorique nous obtenons le triangle de règlements cumulés suivant :

Mois d'origine	**0**	**1**	**...**	**j**	**j+1**	**...**	**...**	**n-1**	**n**
0	C_{00}	C_{01}	**...**	C_{0j}	$C_{0,j+1}$	**...**	**...**	$C_{0,n-1}$	C_{0n}
1	C_{10}	C_{11}	**...**	C_{1j}	$C_{1,j+1}$	**...**	**...**	C_{1n}	
...	...	...	...	...	...	...	...		
i	C_{i0}	C_{i1}	**...**	C_{ij}	$C_{i,j+1}$	...			
...	...	...	...	...	...				
n-j-1	$C_{n-j-1,0}$	**...**	**...**	$C_{n-j-1,j}$	$C_{n-j-1,j+1}$				
n-j	$C_{n-j,0}$	**...**	**...**	$C_{n-j,j}$					
...	**...**	**...**	**...**						
n-1	$C_{n-1,0}$	$C_{n-1,1}$							
n	$C_{n,0}$								

TABLE 2 – Triangle de règlements cumulés.

Notre objectif était de calculer des facteurs de développement avec cette méthode, afin de les appliquer aux cadences de règlements des années 2019 et 2020 (tout ceci par sous-poste). Donc, la formule de calcul de ces facteurs de développement (noté f_j $\forall j = 0, ..., n-1$) est :

$$f_j = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}}, \forall j = 0, ..., n-1 \tag{2}$$

Finalement, vous trouverez ci-dessous comme exemple les coefficients de développement associés au triangle des règlements cumulés pour le sous-poste des consultations en médecine générale du portefeuille 1. La première ligne du tableau correspond au calcul des coefficients de développement f_j comme dans la formule présentée précédemment. La seconde ligne du tableau présente les coefficients de développement cumulés, c'est-à-dire multiplier les uns après les autres en partant de la droite vers la gauche. Il s'agira de la seconde ligne que nous appliquerons à nos dépenses.

Facteurs de développement	1,590	1,088	1,038	1,019	1,014	1,008	1,005	1,004	1,002	1,002	1,002	1,001	1,001	1,001	1,001	1,001	1,000	1,000	1,000	1,000	1,001	1,000	1,001
	1,915	1,204	1,107	1,067	1,047	1,032	1,024	1,019	1,015	1,013	1,011	1,008	1,007	1,006	1,005	1,004	1,003	1,003	1,003	1,002	1,002	1,001	1,001

FIGURE 2 – Coefficients de développement des soins courants pour le premier portefeuille

1.4.3 Base finale

Dans cette sous-partie, nous présentons la base finale pour chaque portefeuille que nous avons obtenu en fusionnant les deux bases « Effectif » et « Sinistres ». Dans cette base, une ligne complète représentera les dépenses d'un assuré pour un code acte (par exemple, des verres pour des lunettes pendant une année de couverture), sinon elle ne sera pas complète (quantité d'acte nulle, frais réels nuls, etc.), alors elle représentera un assuré n'ayant pas eu de sinistres pendant l'année de couverture.

Vous trouverez dans le tableau 3 la liste des variables qui composent la base obtenue pour chacun des portefeuilles. Nous appelerons par la suite portefeuille 1, celui qui contient le personnel hospitaliter et le portefeuille 2, celui qui représente la consommation du personnel de la Fonction Publique de l'Etat.

Par ailleurs, nous avons dû regrouper les codes actes par sous-poste et par poste, puisqu'il serait compliqué pour notre étude de déterminer une prime pure par code acte au vu de la volumétrie (plus de 300 pour le premier portefeuille et plus de 700 pour le second). Nous vous renvoyons à la définition choisie pour les postes et les sous-postes dans le tableau 5.

Pour le premier portefeuille, nous avons les variables décrites dans le tableau 3 ci-dessous :

Nom	**Description**
IDENTIFIANT	Identifiant unique permettant de distinguer les assurés.
CAT_BENEF	Identifiant pour le type de bénéficiaire : – *Assure* - Assuré principal, – *Conjoint* - Conjoint de l'assuré principal, – *Enfant* - Enfant de l'assuré principal.
AGE_BENEF	Âge du bénéficiaire.
SEXE_BENEF	Sexe du bénéficiaire.
DEPT	Département de résidence du bénéficiaire.
EXPOSITION	Exposition du bénéficiaire sur l'année de couverture.
AN_SOIN	Année de couverture.
CODACT	Code acte du soin, s'il y en a un pendant l'année de couverture.
POSTE	Poste de soin (Dentaire, Optique, etc.), s'il y en a un pendant l'année de couverture.
SOUS_POSTE	Sous-poste de soin (verre, monture, etc.), s'il y en a un pendant l'année de couverture.
QUANT	Nombre d'actes qu'il y a eu pendant l'année pour un code acte pour le bénéficiaire, s'il vaut 0 alors le bénéficiaire n'a pas consommé.
FRREL	Montant en frais réels dépensé pour un code acte pendant l'année de couverture pour le bénéficiaire.
MNTMUT	Montant remboursé par la mutuelle pour un code acte pendant l'année de couverture pour le bénéficiaire.
RBTSS	Montant remboursé par la Sécurité Sociale pour un code acte pendant l'année de couverture pour le bénéficiaire.

TABLE 3 – Liste des variables de la base pour le premier portefeuille.

Pour le second portefeuille, nous avons les variables décrites dans le tableau 4 ci-dessous :

Nom	**Description**
MEMBREID	Identifiant unique permettant de distinguer les assurés.
AGE	Âge du bénéficiaire.
SEXE	Sexe du bénéficiaire.
GARANTIE	Niveau de garantie du contrat qui peut prendre comme valeur : – *POP1A* - Niveau 1 de la garantie de la population 1, – *POP1B* - Niveau 2 de la garantie de la population 1, – *POP1C* - Niveau 3 de la garantie de la population 1, – *POP2A* - Niveau 1 de la garantie de la population 2, – *POP2B* - Niveau 2 de la garantie de la population 2, – *POP2C* - Niveau 3 de la garantie de la population 2, – *POP2D* - Niveau 4 de la garantie de la population 2, – *POP2E* - Niveau 5 de la garantie de la population 2.
CAT_BENEF	Identifiant pour le type de bénéficiaire : – *Assure* - Assuré principal, – *Conjoint* - Conjoint de l'assuré principal, – *Enfant* - Enfant de l'assuré principal.
EXPOSITION	Exposition du bénéficiaire sur l'année de couverture.
REGION	Région de résidence du bénéficiaire.
EXERCICE	Année de couverture du contrat.
CODACTE	Code acte du soin, s'il y en a un pendant l'année de couverture.
POSTE	Poste de soin, s'il y en a un pendant l'année de couverture.
SOUS_POSTE	Sous-poste de soin, s'il y en a un pendant l'année de couverture.
QUANT	Nombre d'actes qu'il y a eu pendant l'année pour un code acte pour le bénéficiaire, s'il vaut 0 alors le bénéficiaire n'a pas consommé.
FRREL	Montant en frais réels dépensé pour un code acte pendant l'année de couverture pour le bénéficiaire.
MNTMUT	Montant remboursé par la mutuelle pour un code acte pendant l'année de couverture pour le bénéficiaire.
RBTSS	Montant remboursé par la Sécurité Sociale pour un code acte pendant l'année de couverture pour le bénéficiaire.

TABLE 4 – Liste des variables de la base pour le second portefeuille.

Les postes sont la répartition assez classique, mais au niveau des sous-postes nous avons regroupé des code actes proches pour que le total des sous-postes fassent plus de 0,5% de la charge totale en frais réels, afin qu'ils représentent une part significative. Dans le tableau 5 ci-dessous, vous trouverez les regroupements effectués :

Poste	**Sous-poste**
Dentaire	Orthodontie Prothèses dentaires Radiologie (Dentaire) Soins dentaires
Divers	Autre (Divers)
Hospitalisation	Forfait hospitalier Frais de séjour Frais optionnels Soins hospitaliers Transport
Médecines douces	Médecines douces
Optique	Lentilles Monture Verres
Pharmacie	Pharmacie remboursement à 15% Pharmacie remboursement à 30% Pharmacie remboursement à 65% Autre (Pharmacie)
Prothèses auditives et autres	Autres prothèses et véhicules Petits matériels et pansements
Soins courants	Analyses médicales Auxiliaires médicaux Consultation spécialiste Consultation et visite en médecine générale Frais complémentaires de consultation Radiologie(soins courants) Autres soins courants (seulement pour le premier portefeuille)

TABLE 5 – Regroupement des actes par poste et sous-poste.

2 Analyse des données

Pendant une étude, il est essentiel de décrire et d'explorer les données disponibles pour en extraire des informations importantes. Pour cela, au préalable, il y a des étapes de nettoyage, de transformation et de modelage des données qui seront utilisées pour découvrir des hypothèses. Ainsi, nous passerons par ces étapes en analysant les données. Nous commencerons par décrire chaque portefeuille avec une analyse univariée et bivariée des variables explicatives. Une fois ces premières statistiques observées, nous réaliserons une étude des sinistres extrêmes.

Pour des questions de confidentialité, les échelles des figures 3 à 14 ont été retirées.

2.1 Description des portefeuilles

Dans cette partie, nous nous intéresserons à la composition de nos portefeuilles. Il faut noter que les variables contenues dans le portefeuille 1 nommées *DEPT*, *NB_BENEF* et *LIBELLE* sont enlevées, car elles n'apportent pas d'informations pour notre étude. Etant donné que ce portefeuille concerne surtout le personnel hospitalier, la grande majorité des individus se situent sur un seul département, puis une entreprise ayant un gros volume de personnes couvertes domine le portefeuille. Enfin dans notre analyse du portefeuille, nous comptons en terme d'exposition pour le nombre de bénéficiaires (c'est-à-dire une personne ne sera présente que 6 mois dans le portefeuille alors son exposition vaudra 0,5).

2.1.1 Portefeuille 1

D'après le graphique 3 représentant la pyramide des âges du portefeuille 1, nous pouvons remarquer qu'il y a beaucoup plus de femmes que d'hommes chez les adultes. L'âge moyen de ce portefeuille est d'environ 34 ans et la tranche d'âge la plus présente est 51-55 ans. De plus, il y a plus d'adultes que d'enfants.

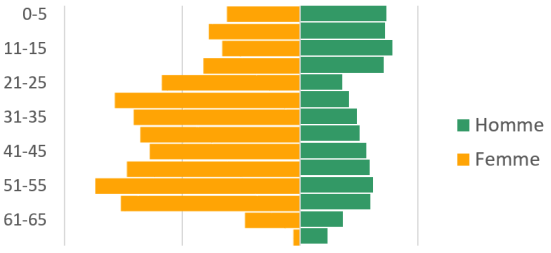

FIGURE 3 – Pyramide des âges du portefeuille 1.

FIGURE 4 – Répartition des individus par catégorie du bénéficiaire du portefeuille 1.

Dans le graphique 4, nous pouvons observer la répartition des individus par catégorie du bénéficiaire et nous notons que la catégorie la plus présente est l'assuré principal.

A l'aide du graphique représenté en figure 5, nous pouvons dire que le nombre de bénéficiaires par année de couverture est plutôt stable.

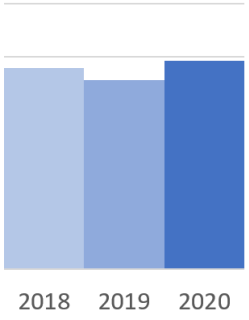

FIGURE 5 – Répartition des individus par année de couverture du portefeuille 1.

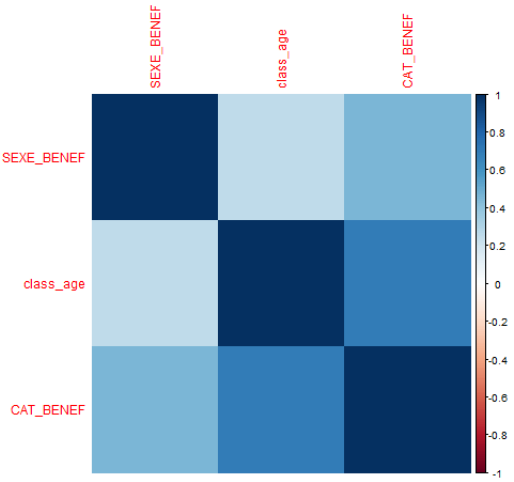

FIGURE 6 – Matrice de corrélation des variables qualitatives du portefeuille 1.

Nous avons calculé la matrice de corrélation des variables qualitatives à l'aide du V de Cramer qui se situe en figure 6. Ainsi, nous nous sommes rendus compte que les variables correspondant aux tranches d'âges et à la catégorie du bénéficiaire sont fortement corrélées environ à 70%. Donc, nous avons décidé de créer une nouvelle variable en les concaténant.

Dans le graphique suivant, vous trouverez la représentation de cette nouvelle variable qui est tout simplement le croisement des variables *CAT_BENEF* et *class_age*. Nous pouvons remarquer que les assurés sont significatifs entre 21 et 60 ans, mais que leur volume est faible pour les moins de 21 ans et les plus de 65 ans.

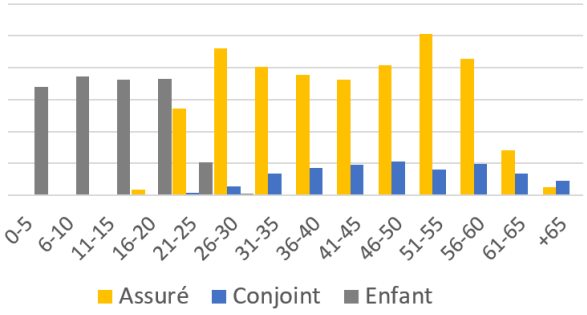

FIGURE 7 – Répartition par tranche d'âge et par catégorie du bénéficiaire du portefeuille 1.

Finalement, nous avons rempli notre condition d'indépendance entre nos variables explicatives pour la suite de notre étude. Donc, pour notre modélisation de la consommation nous aurons le choix entre plusieurs variables, comme la tranche d'âge, la catégorie du bénéficiaire ou le croisement des deux, et le sexe du bénéficiaire des soins.

2.1.2 Portefeuille 2

Le graphique 8 représente la pyramide des âges du second portefeuille. La répartition des individus entre les sexes semble équilibrée. Le portefeuille a un âge moyen de 33 ans et la tranche d'âges présente en majorité est de 45-50 ans. Par ailleurs, nous pouvons distinguer deux groupes à partir de la tranche 26-30 ans.

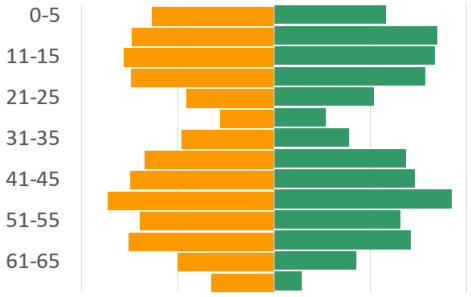

FIGURE 8 – Pyramide des âges du portefeuille 2.

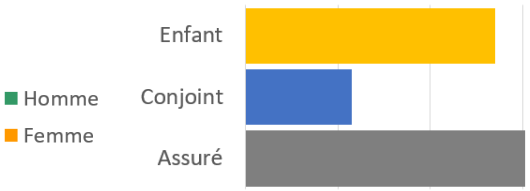

FIGURE 9 – Répartition des individus par catégorie du bénéficiaire du portefeuille 2.

Dans le graphique ci-dessus en figure 9 représentant la répartition des individus par catégorie du bénéficiaire, nous pouvons remarquer qu'il y a quasiment autant d'enfants que d'assurés principaux.

Le graphique 10 représente la répartition des individus par année de couverture. Il faut noter qu'il y a une forte croissance du portefeuille, car l'adhésion est facultative et les agents adhèrent progressivement au système depuis 2018.

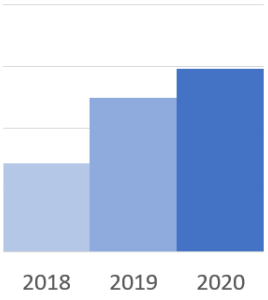

FIGURE 10 – Répartition des individus par année de couverture du portefeuille 2.

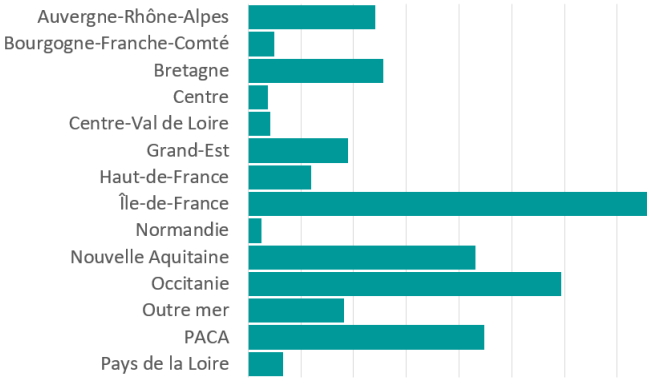

FIGURE 11 – Répartition des individus par région du portefeuille 2.

Le diagramme en barre en figure 11 montre la répartition par région. Les régions qui sont le plus présentes dans ce portefeuille sont : Île-de-France, Occitanie, Provence-Alpes-Côte-d'Azur et Nouvelle Aquitaine.

D'après le graphique ci-dessous en figure 12 décrivant la répartition des individus par niveau de garantie, nous remarquons que la grande majorité des individus appartient au niveau « POP1C » et que le niveau « POP2E » est quasiment inexistant.

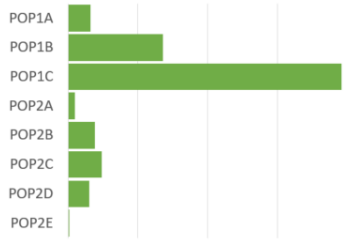

FIGURE 12 – Répartition des individus par niveau de garantie du portefeuille 2.

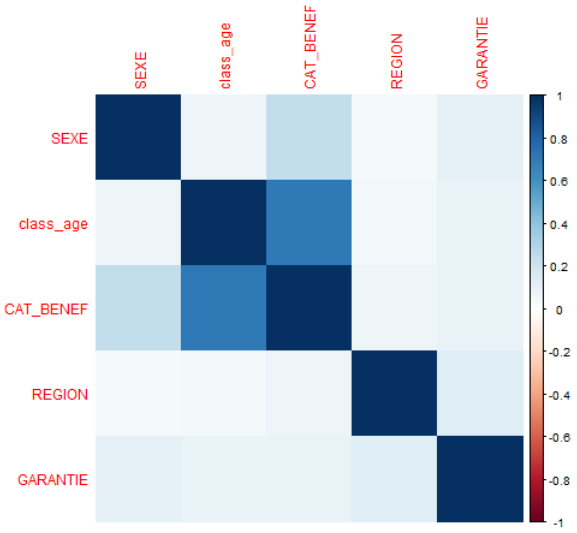

FIGURE 13 – Matrice de corrélation des variables qualitatives du portefeuille 2.

Nous avons effectué la même démarche que dans le portefeuille précédent en calculant la matrice de corrélation des variables qualitatives avec le V de Cramer se situant en figure 13. Nous en arrivons au même résultat et décidons d'agir de la même manière.

Dans le graphique ci-dessous, vous trouverez la représentation de la nouvelle variable qui est le croisement des variables *CAT_BENEF* et *class_age*. Nous notons que les assurés sont significatifs entre 21 et 65 ans, mais que leur volume est faible pour les moins de 21 ans et les plus de 65 ans.

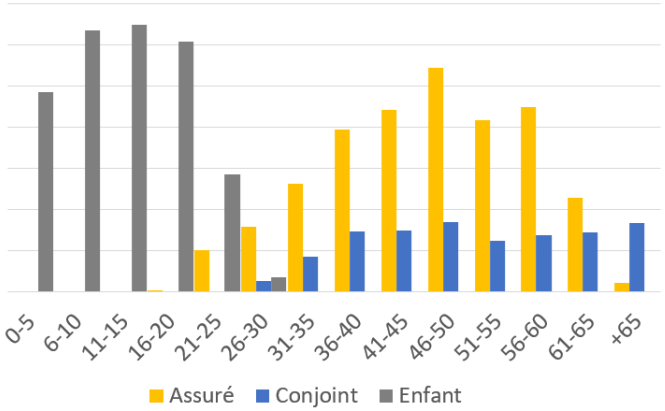

FIGURE 14 – Répartition par tranche d'âge et par catégorie du bénéficiaire du portefeuille 2.

Finalement, comme précédemment dans le portefeuille 1, nous avons rempli notre condition d'indépendance entre nos variables explicatives pour la suite de notre étude. Ainsi, nous pourrons choisir entre plusieurs variables explicatives. Par exemple, nous pourrons utiliser la tranche d'âge

ou la catégorie du bénéficiaire ou le croisement des deux, tout dépendra de la volumétrie. Ou encore, la variable représentant la région et le niveau de garantie de l'assuré principal.

2.2 Gestion des sinistres graves

En santé, comme dans d'autres secteurs telle que la protection des dommages aux biens, les assureurs ont besoin de gérer les sinistres extrêmes (aussi appelés graves). Ces sinistres peuvent de par leur fréquence d'apparition faible et leur intensité forte (leur montant va dépasser fortement la moyenne) troubler la modélisation que nous souhaitons réaliser.

Des méthodes statistiques issues de la théorie des valeurs extrêmes, nous offrent la possibilité d'étudier le comportement de ces sinistres contenus dans un échantillon et décrit par des variables aléatoires. Elles nous permettent d'avoir une meilleure approche qu'avec une simple étude statistique, en déterminant un seuil de montant de sinistre à partir duquel les sinistres seront considérés comme graves. Par conséquent, les sinistres ayant un montant en-dessous de ce seuil seront considérés comme attritionnels. De plus, la détermination de ce seuil s'effectue seulement sur les sinistres ayant un montant positif.

2.2.1 Observation des extrêmes

Tout d'abord, nous observons les données pour voir s'il y a des extrêmes dans les montants de sinistres, à l'aide des tableaux 6 et 7 qui résument les informations sur les sinistres de nos deux portefeuilles étudiés. En regardant de plus près, les valeurs prises par le coût des frais réels, nous notons que la médiane est très inférieure à la moyenne dans les deux portefeuilles. Par ailleurs, le maximum du coût d'un sinistre peut atteindre jusqu'à 64 313 € et 66 570 € respectivement dans les deux portefeuilles. Or, la moyenne du coût des sinistres est autour de 27 € et 35 € respectivement dans les portefeuilles. Classiquement, il s'agirait d'une distribution avec un grand nombre de petits sinistres et quelques-uns importants.

Minimum	Premier quantile	Médiane	Moyenne	Dernier quantile	Maximum
0,05	2,00	5,03	26,77	25,00	64 313

TABLE 6 – Répartition du montant de sinistres pour le premier portefeuille.

Minimum	Premier quantile	Médiane	Moyenne	Dernier quantile	Maximum
0,01	2,04	5,84	34,75	27,83	66 570

TABLE 7 – Répartition du montant de sinistres pour le second portefeuille.

Nous effectuons notre étude sur 373 691 sinistres pour le premier portefeuille et 1 109 432 sinistres pour le second. Les graphiques ci-après représentent le montant des sinistres par indice. Nous avons mis en évidence en rouge dans les nuages de points à gauche les montants de sinistres supérieurs à 20 000 €. Dans les graphiques à droite, nous avons représenté les sinitres par leur montant en enlevant ceux qui étaient au-dessus de 20 000 €. En observant les graphiques des nuages de points ci-dessous, nous pouvons dire que dans les deux cas les

données conviennent à la théorie des valeurs extrêmes, puisqu'il y a quelques sinistres extrêmes et la grande majorité est proche de 0.

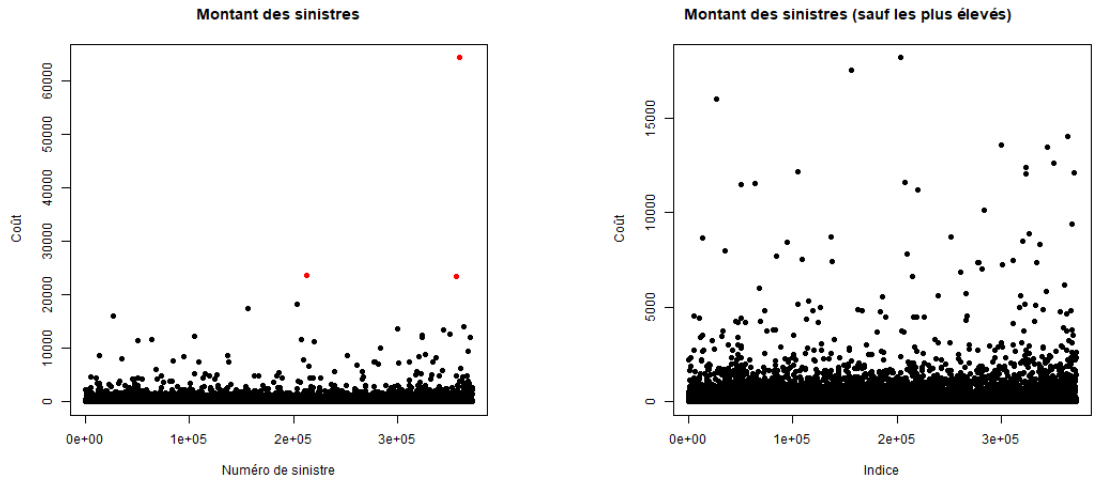

FIGURE 15 – Montant de sinistres pour le premier portefeuille.

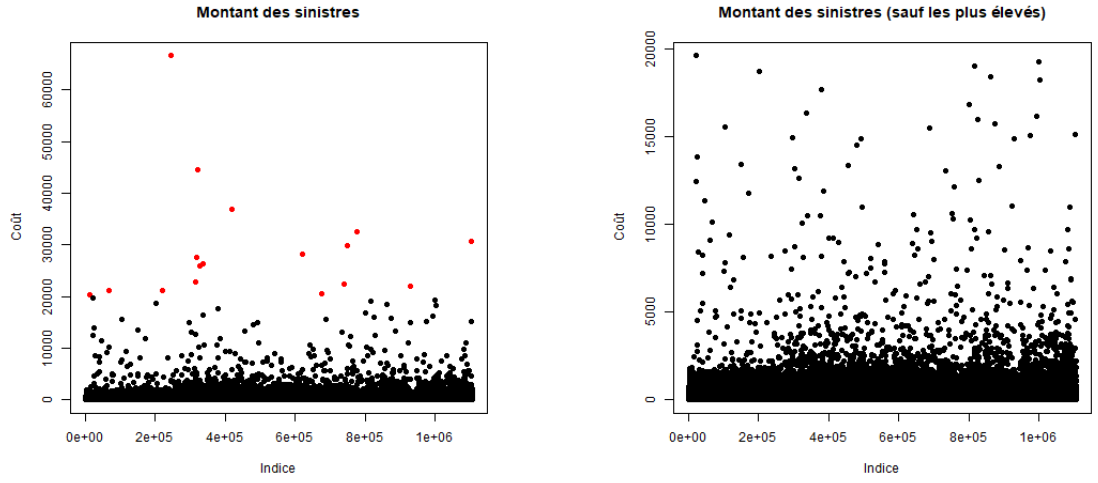

FIGURE 16 – Montant de sinistres pour le second portefeuille.

A l'aide des graphiques 17 et 18, nous avons bien des distributions de nos coûts en forme de coude dans les deux portefeuilles, ce qui est représentatif de distributions à queue lourde.

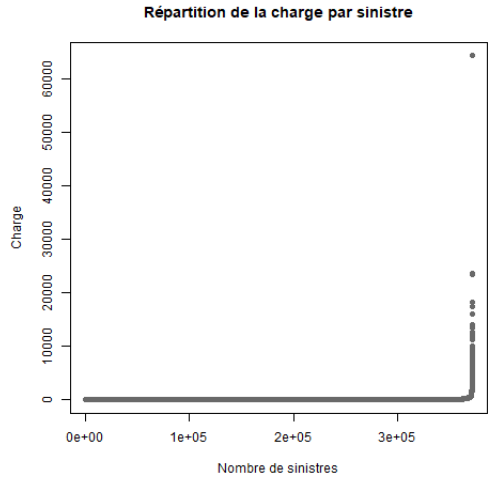

FIGURE 17 – Répartition de la charge en frais réels par sinistre triée pour le premier portefeuille.

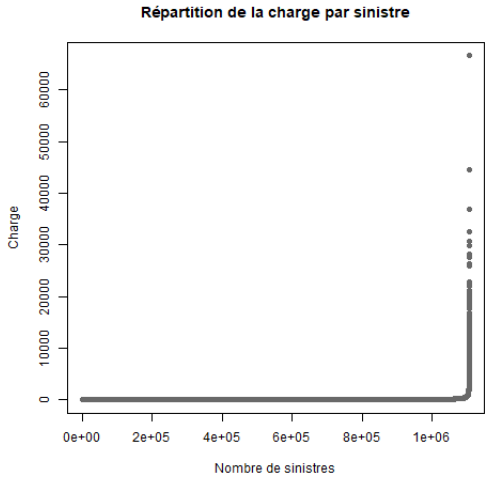

FIGURE 18 – Répartition de la charge en frais réels par sinistre triée pour le second portefeuille.

Normalement, dans une boîte à moustaches tous les points en dehors du dernier quantile sont considérés comme extrêmes. Or, celles représentées en figures 19 et 20 sont complètement écrasées en 0, ce qui est expliquée par une forte présence de sinistres à faible montant. Nous pouvons aussi le remarquer avec la moyenne qui est très loin du maximum dans les deux portefeuilles. Donc, nous allons déterminer un seuil pour avoir suffisamment de sinistres graves et qu'il ne soit pas trop bas pour ne pas prendre des sinistres attritionnels.

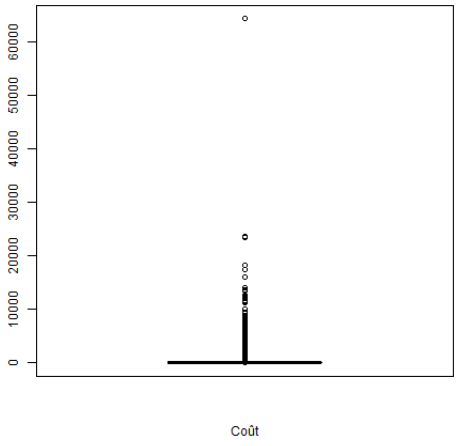

FIGURE 19 – Boîte à moustaches du coût (en frais réels) pour le premier portefeuille.

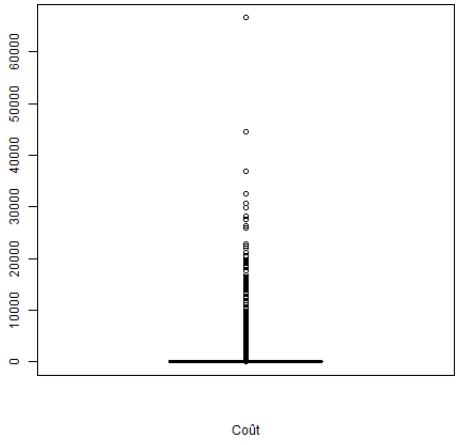

FIGURE 20 – Boîte à moustaches du coût (en frais réels) pour le second portefeuille.

Vous trouverez dans les tableaux 8 et 9 ci-dessous des mesures effectuées sur les sinistres par rapport au coût total (en frais réels) et au nombre de sinistres total avec différents seuils. Par exemple, si nous avions un seuil à 10 000 € pour le premier portefeuille, les sinistres graves représenteraient 3,23% du coût total et 0,0005% des sinistres.

Seuil	Pourcentage du coût total	Pourcentage du nombre de sinistres total
20 000 €	1,12%	0,0008%
10 000 €	3,23%	0,0005%
5 000 €	5,51%	0,014%
1 000 €	17,98%	0,20%

TABLE 8 – Mesures des sinistres sur la charge totale (en frais réels) pour le premier portefeuille.

Seuil	Pourcentage du coût total	Pourcentage du nombre de sinistres total
25 000 €	0,91%	0,0009%
18 000 €	1,6%	0,0002%
15 000 €	2,01%	0,005%
10 000 €	2,9%	0,013%
5 000 €	5%	0,015%
1 000 €	16,82%	0,25%

TABLE 9 – Mesures des sinistres sur la charge totale (en frais réels) pour le second portefeuille.

2.2.2 Détermination du seuil

Avant de présenter les résultats obtenus pour nos jeux de données au niveau de la sélection d'un seuil, nous allons au préalable établir un cadre théorique pour la théorie des valeurs extrêmes.

Il existe deux méthodes en pratique en théorie des valeurs extrêmes. Pour notre étude, nous utiliserons l'approche POT (*Peak-Over-Threashold*) qui semble être la plus adaptée, puisqu'elle permet d'extraire des sinistres graves avec un seuil sur une période. Nous n'avons pas choisi l'autre approche *Block Maxima* qui repose sur la division d'une période d'observation en blocks pour générer le maximum de chacun des blocks avec une loi des extrêmes.

Nous nous plaçons dans le cadre suivant, soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d) de fonction de distribution F et soit $M_n = max(X_1, ..., X_n)$. Par l'indépendance des variables, on a :

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, ..., X_n \leq x) = \mathbb{P}(X_1 \leq x) \times ... \times \mathbb{P}(X_n \leq x) = [F(x)]^n$$

Par la suite, nous allons devoir expliquer le terme de « lourdeur » de queue. Donc, nous avons besoin de définir le concept de domaine d'attraction, ainsi que le théorème Fisher-Tippett qui est un élément de base de la théorie des valeurs extrêmes.

Notons que :

- Si $F(x) < 1$, alors $\mathbb{P}(M_n \leq x) \to_{n\to\infty} 0$.
- Si x^F est le point extrêmal de F défini par :

$$x^F = sup_x\{F(x) < 1\},$$

Alors $M_n \to_{n\to\infty} x^F$. La distribution asymptotique de M_n est donc dégénérée.

Ainsi en 1928, le théorème de Fishet-Tippet est établi :

S'il existe des suites de réels $c_n > 0$ et d_n telles que :

$$\mathbb{P}\left(\frac{M_n - d_n}{c_n} \leq x\right) \to_{n\to\infty} G(x)$$

pour une distribution non-dégénérée G, alors G est du même type que l'une des trois distributions suivantes :

- *Fréchet ($\alpha > 0$) :*

$$\phi_\alpha(x) = \begin{cases} 0 \text{ si } x \leq 0, \\ e^{-x^{-\alpha}} \text{ si } x > 0. \end{cases} \tag{3}$$

- *Weibull ($\alpha > 0$) :*

$$\psi_\alpha(x) = \begin{cases} e^{-(-x^\alpha)} \text{ si } x \leq 0, \\ 1 \text{ si } x > 0. \end{cases} \tag{4}$$

- *Gumbel :*

$$\Lambda(x) = exp(e^{-x}),\ x \in \mathbb{R}. \tag{5}$$

Dans certains cas, il est préférable d'avoir une seule distribution qui unifie les trois précédentes. Ainsi, Von Mises (1954) et Jenkinson (1955) ont proposé la distribution des extrêmes généralisées (GEV) $GEV(\mu, \sigma, \xi)$:

$$G(x) = \begin{cases} exp\left(-\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right)\right) \text{ si } \xi \neq 0, \\ exp\left(-exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right) \text{ si } \xi = 0. \end{cases} \tag{6}$$

Où $y_+ = max(y, 0)$ et $\sigma > 0$. ξ est le paramètre de forme, μ le paramètre de position, σ le paramètre d'échelle.

On notera :

$$G_\xi(x) = exp\left(-(1 + \xi x)_+^{-\frac{1}{\xi}}\right). \tag{7}$$

Donc, nous aurons :

- La distribution de Gumbel correspond à $\xi = 0$, $GEV(0, 1, 0) =$ Gumbel,
- La distribution de Fréchet correspond à $\xi > 0$, $GEV(1, \alpha^{-1}, \alpha^{-1}) =$ Fréchet(α),
- La distribution de Weibull correspond à $\xi < 0$, $GEV(-1, \alpha^{-1}, -\alpha^{-1}) =$ Weibull(α).

Enfin, la notion de domaine d'attraction fait son apparition car, le théorème de convergence de la loi du maximum énonce que si :

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = (F(a_n x + b_n))^n \rightarrow_{n\to\infty} G(x) \tag{8}$$

où G est une distribution non dégénérée, alors G a une distribution des extrêmes généralisée.

Finalement, nous pouvons dire que F appartient au domaine d'attraction de $G(F \in D(G))$ s'il existe deux suites (a_n) et (b_n) telle que la convergence précédente ait lieu.

Notons que ξ appelé aussi indice de queue donne une information sur l'épaisseur des queues de distribution tel que :

- $\xi > 0$: « queue épaisse »,
- $\xi = 0$: « queue intermédiaire »,
- $\xi < 0$: « queue fine ».

Notre étude des valeurs extrêmes va se dérouler de la manière suivante :

1. Sélection du domaine d'attraction,
2. Sélection de seuils potentiels,
3. Choix du seuil .

- **<u>Sélection du domaine d'attraction</u>**
  Nous allons utiliser trois méthodes permettant de vérifier la sélection du domaine de Fréchet, car sa queue de distribution est épaisse comme celle de nos données.
  Soit n le nombre de sinistres, et notons les éléments de notre échantillon triés x_i :
  - Le ***Q-Q Plot*** permet de comparer les quantiles de la distribution empirique à ceux de la distribution théorique choisie. Si la représentation est *linéaire*, alors l'échantillon vient de la distribution théorique.

    Nous commençons avec la loi exponentielle. Pour l'élément x_j en abscisse, nous allons lui associer en ordonnée la fonction $-ln\left(1 - \frac{j}{n+1}\right)$.

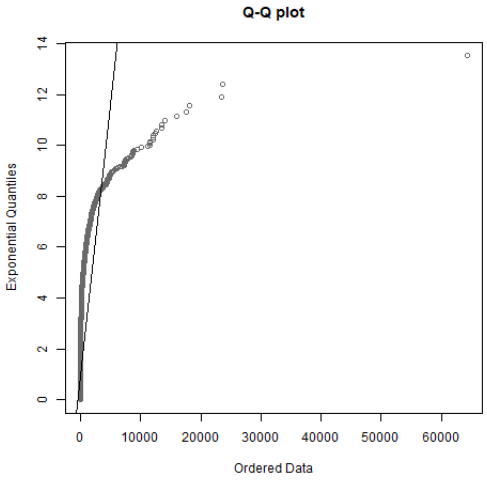

FIGURE 21 – Q-Q plot avec la distribution exponentielle pour le premier portefeuille.

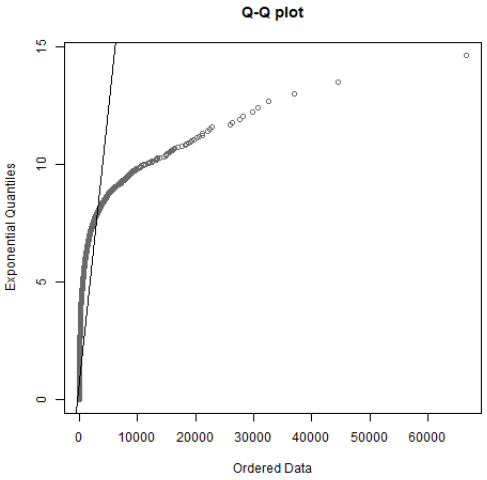

FIGURE 22 – Q-Q plot avec la distribution exponentielle pour le second portefeuille.

Nous pouvons observer distinctement pour les deux portefeuilles des débuts de courbes concaves se dessinant et se détachant de la droite. Nous pouvons l'interpréter comme le signe d'un comportement à queue lourde, ce qui semblerait être un domaine d'attration de type Fréchet.

– Une deuxième méthode utilise la ***fonction de répartition d'une loi sur*** $\mathbb{R}^+$:
  - Fréchet :
    $\phi_\alpha(x) = e^{-x^{-\alpha}}$ donc $ln\left(\frac{1}{\phi_\alpha(x)}\right)$ sera une fonction linéaire en x pour un α donné.

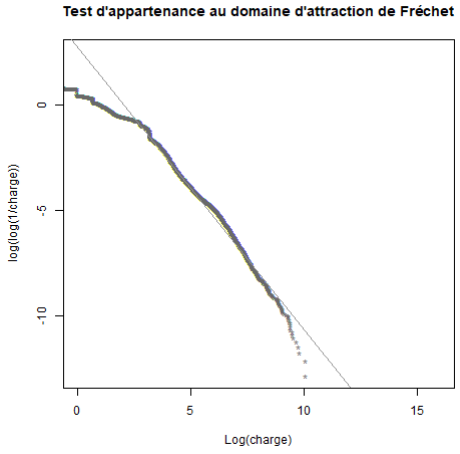

FIGURE 23 – Test d'appartenance au domaine d'attraction de Fréchet pour le premier portefeuille.

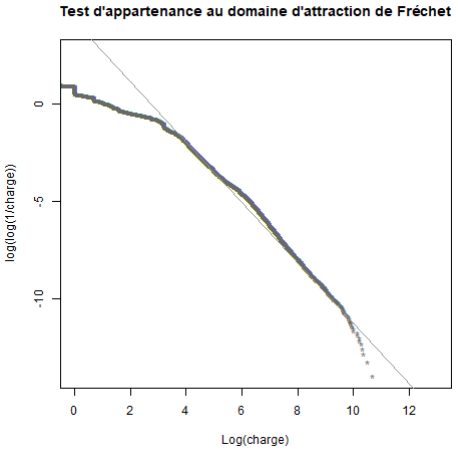

FIGURE 24 – Test d'appartenance au domaine d'attraction de Fréchet pour le second portefeuille.

Dans les deux graphiques 23 et 24 pour le test d'appartenance au domaine

de Fréchet, les courbes ont une structure linéaire, donc nous pouvons accepter l'hypothèse nulle $(H_0 : \xi > 0)$.

- Gumbel :
  $\Lambda(x) = exp\,(e^{-x})$ donc $ln\left(ln\left(\frac{1}{\Lambda(x)}\right)\right)$ sera une fonction linéaire en x.

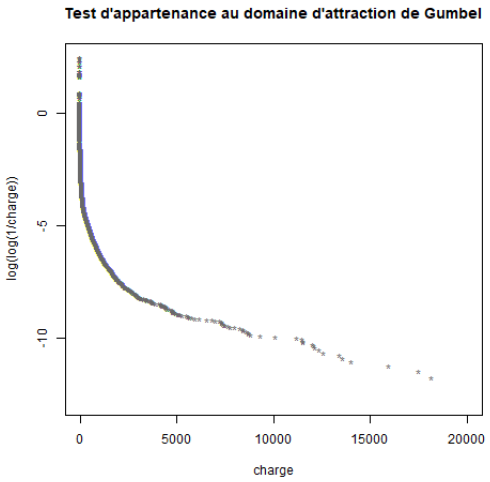

FIGURE 25 – Test d'appartenance au domaine d'attraction de Gumbel pour le premier portefeuille.

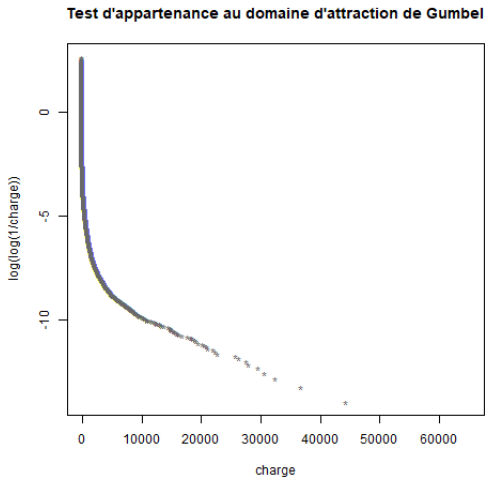

FIGURE 26 – Test d'appartenance au domaine d'attraction de Gumbel pour le second portefeuille.

Les tests d'appartenance au domaine de Gumbel représentés dans les graphiques 25 et 26, les courbes sont convexes, donc nous pouvons rejeter l'hypothèse nulle $(H_0 : \xi = 0)$.

- La dernière méthode est le ***Pareto Quantile Plot*** pour vérifier que le domaine d'attraction est celui de Fréchet. Nous devons regarder si la fin de notre courbe représentée est linéaire.

  Nous représentons pour le rang j, sur un graphique $ln\left(\frac{n+1}{j}\right)$ en abscisse et $ln(x_{n-j+1})$ en ordonnée où x_{n-j+1} représente la $n-j+1$ème observation.

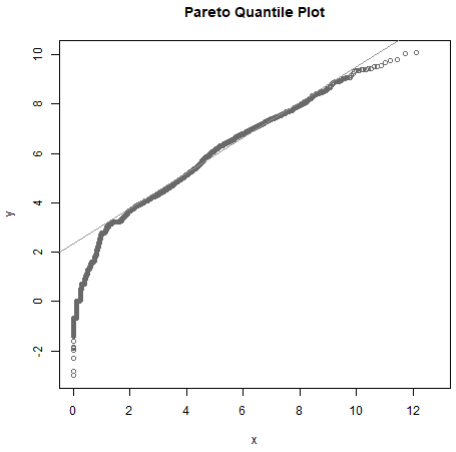

FIGURE 27 – Graphique du Pareto Quantile pour le premier portefeuille.

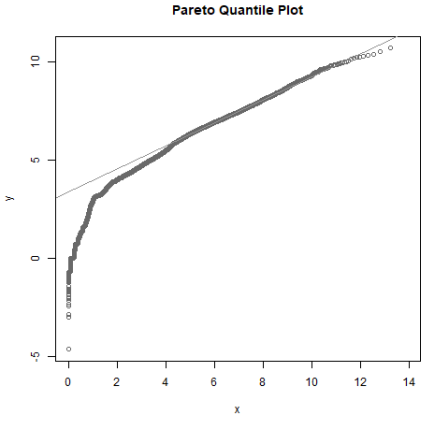

FIGURE 28 – Graphique du Pareto Quantile pour le second portefeuille.

D'après les figures ci-dessus, nous observons une structure linéaire pour la fin de la courbe. Si nous voulons être plus précis nous pouvons vérifier si c'est toujours le cas sur les cent derniers points du graphique.

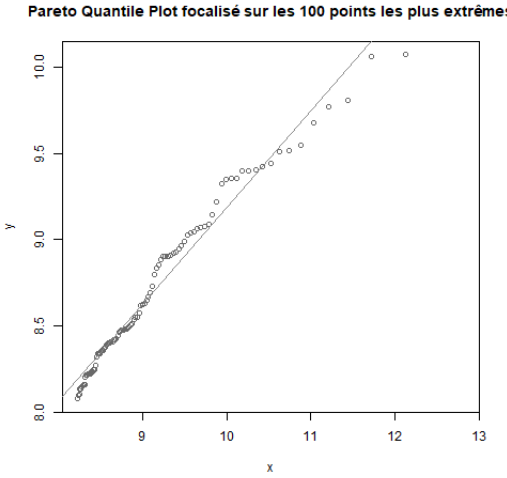

FIGURE 29 – Graphique du Pareto Quantile sur les 100 derniers points pour le premier portefeuille.

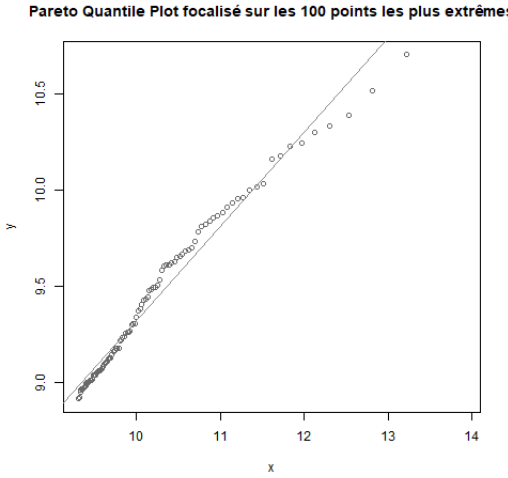

FIGURE 30 – Graphique du Pareto Quantile sur les 100 derniers points pour le second portefeuille.

Les graphiques semblent bien vérifier la tendance linéaire (avec un coefficient de détermination R^2 à 97% et 93%). Nous pouvons conclure après ce dernier test que nous acceptons l'hypothèse selon laquelle la distribution de nos jeux de données appartient à un domaine d'attraction de type Fréchet.

- **Sélection des seuils potentiels**

Tout d'abord, nous définissons la distribution de Pareto généralisée $GPD(\beta, \xi)$. Si X suit une loi GPD notons $X \sim GPD(\beta, \xi)$ avec $\beta > 0$ et $\xi \geq 0$ alors :

$$P(X \leq x) = G_{\xi,\beta}(x) = \begin{cases} 1 - [1 + \xi(x/\beta)]_+^{-\frac{1}{\xi}} & \text{si } \xi \neq 0, \\ 1 - e^{-\frac{x}{\beta}} & \text{si } \xi = 0. \end{cases} \tag{9}$$

La sélection d'un seuil adéquate peut être complexe, puisque si nous le choisissons trop élevé nous disposerons de très peu de données. Il faut aussi garder suffisamment de données afin d'avoir une bonne approximation asymptotique de la GPD (la distribution de Pareto généralisée). Pour cela nous avons différents outils à notre disposition qui sont décrits ci-dessous.

– **Mean Excess Plot (graphique de la fonction de dépassement moyen)** :
Nous pouvons utiliser l'une des propriétés des lois GPD qui est la fonction de dépassement moyen. Elle nous permet de dire que si $X \sim GPD(\beta, \xi)$, alors pour $\xi < 1$, $X - u | X > u \sim GPD(\beta + \xi u, \xi)$ et

$$\mathbb{E}(X - u | X > u) = \frac{\beta + \xi u}{1 - \xi}$$

Cette propriété permet d'avoir une relation linéaire entre le paramètre de forme ξ avec le seuil u, ainsi nous pouvons analyser le graphique de la *Mean Excess Plot* et choisir un seuil.

Finalement, pour nos deux portefeuilles nous obtenons les courbes suivantes :

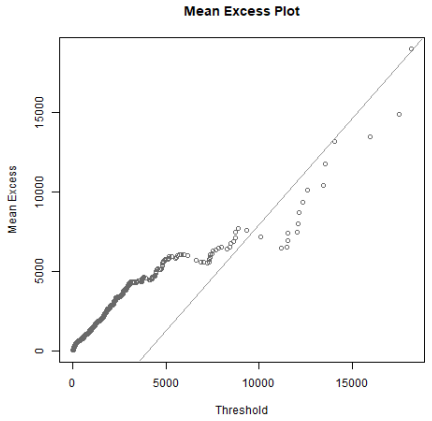

FIGURE 31 – Mean Excess Plot pour le premier portefeuille.

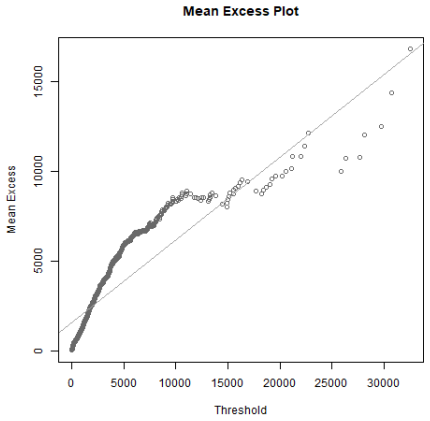

FIGURE 32 – Mean Excess Plot pour le second portefeuille.

La forte croissance de la courbe observée pour la première partie de la courbe confirme la conclusion que nous avions faite auparavant sur la queue lourde pour les deux portefeuilles.

Dans le cas du premier portefeuille, nous pouvons voir une linéarité se dessiner au-delà du seuil des graves 7 000 €, la courbe commence à se comporter comme une droite. Tandis que pour le second portefeuille, nous pourrions dire que cela se dessine en forme de droite à partir de 15 000 €.

Cette méthode reste tout de même peu précise au vu de la volatilité de la fin de la courbe, puisque les moyennes des excès sont effectuées sur très peu de points et de sa difficulté d'interprétation. Ainsi, il est difficile d'estimer de manière empirique un seuil avec ce graphique.

– **Estimateur de Hill** :
Cet estimateur a été introduit par Hill en 1975, il est défini de la manière suivante pour $\xi > 0$:

$$\xi_{k,n}^{Hill} = \frac{1}{k} \sum_{i=1}^{k} ln(X_{n-i+1,n}) - ln(X_{n-k,n}), k = 1, ..., n-1 \tag{10}$$

Avec n la taille de notre échantillon. Il se comprend comme la pente du Pareto Quantile plot représenté plus haut.

Nous avons pu estimer trois plages pour le premier portefeuille et deux pour le second qui correspondraient à des seuils potentiels. Ils sont détectés dès que la courbe semble se stabiliser.

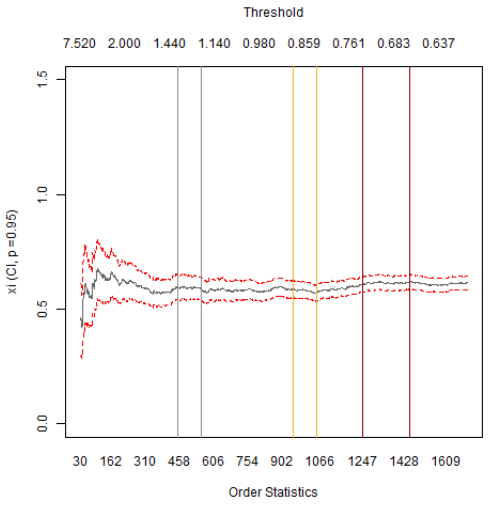

FIGURE 33 – Graphique de l'estimateur de Hill pour le premier portefeuille.

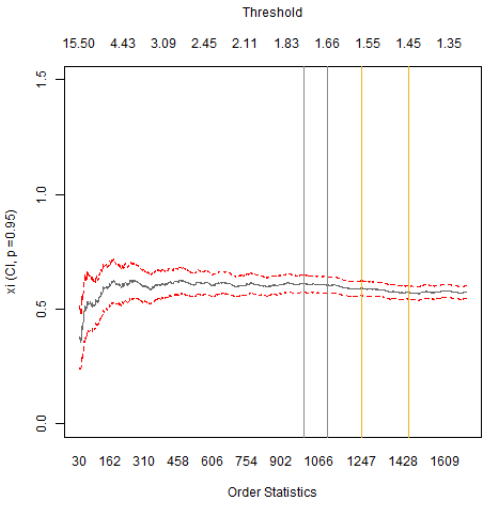

FIGURE 34 – Graphique de l'estimateur de Hill pour le second portefeuille.

Nous avons identifié trois plages pour le premier portefeuille :

Numéro de la plage	Intervalle		Nombre de sinistres en excès		Estimateur de Hill	
1	1 597,98€	1 354,33€	450	550	0,61	0,59
2	880,83€	842,16€	950	1050	0,583	0,585
3	734,22€	662,83€	1 250	1 450	0,59	0,62

TABLE 10 – Mesures des sinistres sur la charge totale pour le premier portefeuille.

Nous avons identifié deux plages pour le second portefeuille :

Numéro de la plage	Intervalle		Nombre de sinistres en excès		Estimateur de Hill	
1	1 742,24€	1 657,61€	1 000	1 100	0,59	0,60
2	1 557,68€	1 453,61€	1 250	1 450	0,620	0,621

TABLE 11 – Mesures des sinistres sur la charge totale pour le second portefeuille.

2.2.3 Choix du seuil

Après étude des sinistres graves, nous pouvons conclure que dans les deux portefeuilles nous retrouvons à peu près le même type de sinistres graves, ce qui est plutôt rassurant puisqu'il s'agit de sinistres de santé. Par ailleurs, après une réflexion, nous avons décidé de prendre un seuil à 5 000 €, puisque cela représentera 5% de la charge totale dans les deux cas. Ainsi, nous n'aurons pas les prothèses auditives et les hospitalisations de courte durée qui seront considérées comme des sinitres graves. Finalement, les sinistres graves pour le premier portefeuille correspondent à 51 sinistres et 174 sinistres pour le second, tout ceci pour trois années de couverture.

3 Approche théorique de la modélisation de la consommation

Dans ce chapitre, nous allons introduire les concepts fondamentaux de l'apprentissage automatisé qui sera la base de notre étude. Ensuite, nous présenterons de façon théorique les algorithmes principalement utilisés en actuariat et que nous utiliserons pour nos travaux.

3.1 L'apprentissage automatisé

L'apprentissage automatisé (machine learning) est un domaine de l'intelligence artificielle qui permet à un ordinateur d'apprendre à partir de données basées sur des approches mathématiques. Nous pouvons le découper en différents champs comme présenté ci-dessous dans la figure 35 :

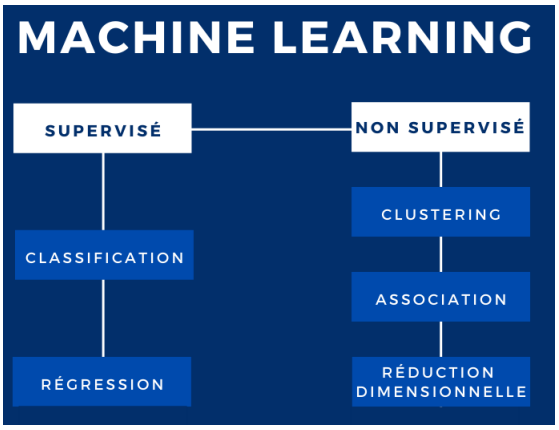

FIGURE 35 – Schéma explicatif du machine learning.

Il est donc séparé en deux types d'apprentissage :

- **Supervisé** signifiant que la machine a déjà accès à l'information, c'est-à-dire qu'elle possède déjà l'information sur la relation entre ce qu'elle doit modéliser et les données.
  - La classification signifie que nous modélisons une cible non numérique, comme le fait qu'une personne soit un « bon » ou « mauvais » risque.
  - La régression permet de modéliser une cible numérique, comme les dépenses moyennes pour un foyer par exemple.
- **Non supervisé** consiste à faire apprendre l'ordinateur sur des données sans cible particulière. Par exemple, reconnaître des communités dans un ensemble de personnes étant connectées entre elles.

Il existe des modèles classiques statistiques comme par exemple les modèles linéaires généralisés qui sont paramétriques, c'est-à-dire qu'ils ont des hypothèses sur la distribution de la variable à expliquer. Les algorithmes d'apprentissage automatisé quant à eux sont non paramétriques ce qui signifie que nous n'avons pas d'hypothèses sur la manière dont se distribue notre variable cible.

Lorsque nous utilisons du machine learning dans nos études, il est composé de deux phases, une première d'apprentissage durant laquelle nous prenons un échantillon de notre population

pour entraîner notre machine, puis une phase de validation en utilisant le reste de notre échantillon sur lequel nous allons tester notre modèle sur de nouvelles données. En règle générale, nous avons 80% de notre base de données choisie de façon aléatoire qui est dédiée à l'apprentissage et 20% à la validation.

Pour notre étude, nous utiliserons des algorithmes de type supervisé en régression. Nous verrons par la suite deux algorithmes de types différents avec un paramétrique le modèle linéaire généralisé et un non paramétrique l'eXtreme Gradient Boosting, afin de comparer leur performance.

3.2 Mesures de performance

Pour n'importe quel modèle de machine learning, il est important d'évaluer sa précision de prédiction. Nous allons présenter ci-dessous les mesures qui nous permettront d'évaluer et comparer nos modèles.

- **Mean Absolute Error (MAE)** est la moyenne en valeur absolue de la différence entre les valeurs prédites et observées. Elle s'écrit de la manière suivante :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{11}$$

  Où $\hat{y}_i$ correspond à la valeur prédite pour la i-ème observation, y_i correspond à la i-ème valeur observée et n la taille de notre échantillon.

- **Mean Square Error (MSE)** est la moyenne de la différence au carré entre les valeurs prédites et observées. Elle s'écrit comme :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

- **Root Mean Square Error (RMSE)** est la racine carrée de la moyenne de la différence au carré entre les valeurs prédites et observées. Elle s'écrit de la façon suivante :

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

3.3 Modèle linéaire généralisé

Pendant longtemps, les actuaires utilisaient des régressions linéaires. La complexité des problèmes statistiques les ont poussé à utiliser des méthodes plus complexes, comme le modèle linéaire généralisé (General Linear Model en anglais – GLM) qui est une généralisation d'une régression linéaire avec trois composantes. Il a été développé par John Neder et Robert Wedderburn en 1972. Cet algorithme permet d'expliquer la relation entre l'espérance d'une variable

cible et plusieurs variables explicatives. Soient pour un ensemble de n individus, p variables explicatives que nous notons $X_1, ..., X_n$ et une variable cible Y à expliquer. Notre modèle linéaire généralisé s'écrit de la manière suivante :

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \xi_i, \ i = 1, ..., n \tag{14}$$

- β_0 et β_k sont les paramètres inconnus du modèle de régression. Ces paramètres sont estimés par la méthode des moindres carrés ordinaires.
- ξ_i est un bruit.

Avec les hypothèses suivantes :

- $\mathbb{E}(\xi_i) = 0, \forall i$ (les erreurs sont centrées),
- $Var(\xi_i) = \sigma^2, \forall i$ (homoscédasticité des erreurs),
- $Cov(\xi_i, \xi_j) = 0, \forall i \neq j$ (non corrélation des erreurs).

Pour estimer les paramètres inconnus du modèle (β_k et σ la variance de l'erreur), nous utilisons la méthode des moindres carrés ordinaires (MCO), ou la méthode du maximum de vraisemblance (MV) à laquelle nous devons ajouter l'hypothèse que les erreurs suivent une loi normale. Pour estimer les paramètres, nous utilisons le maximum de vraisemblance. Par la méthode des moindres carrés ordinaires, nous obtenons :

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

Donc, l'estimateur de β s'écrit :

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

Par ailleurs, nous disons que $\hat{\beta}$ est le plus efficace des estimateurs sans biais (par le théorème de Gauss-Markov).Donc, $\xi \sim N_n(0, \sigma^2 I_n)$ et $Y \sim N_n(X\beta, I_n\sigma^2)$. Finalement, nous obtenons comme estimateur de β :

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

Avec comme hypothèse :

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

Finalement, notre modèle possède trois composantes qui sont les suivantes :

1. **Composante aléatoire** :
   Elle est représentée par des variables explicatives indépendantes $Y_1, ..., Y_n$ de densités appartenant à la famille exponentielle.
   Les lois normale, binomiale et poisson appartiennent à la famille de distributions exponentielle dont les densités s'écrivent sous la forme :

$$f(y, \theta, \phi) = exp\left(\frac{y\theta - b}{a(\phi)} + c(y, \phi)\right) \tag{15}$$

   Où ϕ est le paramètre de dispersion et θ le paramètre canonique.

2. **Composante déterministe** :
   Pour chaque $(Y_i)_{i=1,...n}$, nous avons un vecteur $(X_{1i}, ..., X_{pi})$ décrivant Y_i. Les vecteurs $X_1 = (X_{11}, ..., X_{1n})'$,...,$X_p = (X_{p1}, ..., X_{pn})$ sont des vecteurs explicatifs.

3. **Une fonction de lien** :
   Elle est représentée par g qui est littéralement la relation entre les composantes aléatoires et déterministes.

   g est une fonction déterministe strictement monotone définie sur $\mathbb{R}$ où :

$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{16}$$

   $\beta_0, ..., \beta_p$ paramètres inconnus qui sont des coefficients du modèle.
   Notons que : $\mu = \mathbb{E}(Y)$ et que $\beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ est un prédicteur linéaire.

   Chaque loi de probabilités de la famille exponentielle a une fonction de lien spécifique appelée « canonique » et définie pour $\theta = \eta$. Le lien canonique tel que $g(\mu_i) = \theta_i$, mais nous savons que pour une variable aléatoire Y dont la distribution est une forme exponentielle, nous avons :

$$\mathbb{E}(Y) = b'(\theta) \text{ et } Var(Y) = b''(\theta) \times a(\phi)$$

Loi de probabilité	**Fonction de lien canonique**
Normale	$\eta = \mu$
Poisson	$\eta = log(\mu)$
Gamma	$\eta = \frac{1}{\mu}$
Binomiale	$\eta = ln(\frac{\mu}{1-\mu})$
Gaussienne inverse	$\eta = \frac{1}{\mu^2}$

3.3.1 Tarification en santé

Nous pouvons rencontrer différentes utilisations des GLM en Santé comme :

- Modèle pour prédire le nombre de sinistres d'un assuré avec un GLM de loi Poisson ou binomiale négative.
- Modèle pour prédire le montant de sinistres d'un assuré avec un GLM de loi Gamma ou log-normale.
- Modèle *Tweedie* pour prédire la charge totale. Soit N le nombre de sinistres pour une police et $Z_1, ..., Z_n$ les coûts individuels. Le coût total vaut alors :

$$Y = \begin{cases} \sum_{j=1}^{N} Z_j \text{ si } N > 0, \\ 0 \text{ sinon.} \end{cases} \tag{17}$$

  Si $N \sim P(\lambda)$ et $Z_j \sim \Gamma$ indépendantes et N indépendant de Z, alors Y suit une loi de *Tweedie*. Cette loi fait partie de la famille exponentielle avec comme fonction variance supposée $Var(Y) = \phi\mu^p$ avec $1 < p < 2$ l'indice de puissance de la fonction variance et μ considérée comme l'espérance de la distribution. A noter que dans *R*, le paramètre p correspond à *var.power*.

3.3.2 Estimation des paramètres

Dans cette section, nous allons voir comment le GLM estime le vecteur β et ϕ (si ils sont inconnus) grâce à la maximisation de la fonction de vraisemblance qui est :

$$L(Y, \theta(\beta), \phi) = \prod_{i=1}^{n} f(y_i, \theta_i, \phi_i) \tag{18}$$

Finalement, nous pouvons estimer les β_i en résolvant le système à p équations suivant :

$$\frac{\partial ln(L(Y, \theta(\beta), \phi))}{\partial \beta_k} = 0 \tag{19}$$

Où $Y = (y_1, ..., y_n)$, $\beta = (\beta_1, ..., \beta_n)$, avec n qui est le nombre d'observations et p le nombre de paramètres.

Puisque nous avons un large nombre de données en pratique, la résolution du système d'équations est faite avec des techniques numériques (par exemple, avec la méthode de Newton-Raphtson ou la méthode du score de Fisher).

Parfois, (comme dans notre étude), nous pouvons utiliser un paramètre dit « offset », puisqu'il permet de considérer qu'il peut y avoir des groupes de tailles différentes.

3.3.3 Validation

Pour valider notre modèle, nous avons besoin d'en connaître sa qualité, il y a des statistiques pour cela qui sont les suivantes :

- **Deviance :** est une mesure qui compare la différence entre les valeurs prédites et celles observées. Elle peut être définie comme :

$$D = 2ln\left(\frac{L_{SAT}}{L}\right) \tag{20}$$

  Où L_{SAT} est la vraisemblance du modèle saturé et L la vraisemblance de notre modèle. Nous supposons que nous avons estimé un modèle avec n observations et p le nombre de variables explicatives avec $p < n$. En pratique, nous dirions que si $\frac{D}{n-p-1}$ n'est pas plus grand que 1, alors le modèle a une bonne qualité d'estimation.

- **La statistique de Pearson** : compare les valeurs prédites $\hat{\mu_i}$ et celles observées y_i.

$$\chi^2 = \sum_{i=1}^{n} \frac{y_i \hat{\mu_i}}{Var(y_i)} \tag{21}$$

- **AIC (Akaike Information Criterion)** :

$$AIC = -2ln(y_i \hat{\mu_i}) + 2\frac{n}{n-p-1} \tag{22}$$

  Elle utilise les mêmes notations que la statistique de Pearson.

Pour toutes les statistiques, nous supposons que $p < n$. Pour la validation de notre modèle, nous utiliserons la AIC en la minimisant le plus possible.

3.4 Autres modèles de machine learning

D'autres modèles d'apprentissage automatisé sont aussi utilisés en actuariat tel que les arbres de décision (Classification and Regression Tree – CART), ou encore les forêts aléatoires, mais ils sont non paramétriques. Depuis quelques années, un algorithme est en vedette dans les compétitions de machine learning, il s'agit de l'eXtreme Gradient Boosting (XGBoost). Il possède une implémentation open source (ce qui signifie que son code est accessible au public). Il est une optimisation de l'algorithme d'arbre de décision gradient boosting. Il a été conçu pour sa vitesse et sa performance. Il réfère à l'objectif technique de repousser la limite des ressources de calcul pour les algorithmes d'arbres de boosting. Cette algorithme a été créé en 2014 par Tianqi Chen, c'est une mise en oeuvre d'algorithmes de gradient boosting.

Nous pourrions nous interroger sur les raisons pour lesquelles nous devrions utiliser cet algortihme. La première raison est sa vitesse d'exécution comparée à d'autres algorithmes. Nous pouvons ajouter sa performance qui a été démontrée plusieurs fois lors de compétitions de Machine Learning. Sa capacité de prédiction fait aussi partie de ces raisons. Puisque, il fonctionne aussi bien dans le cadre d'une classification, que d'une régression.

Dans cette partie, nous présentons les algorithmes de machine learning à la base de la création du XGBoost. Nous commencerons par l'explication des arbres de décision. Ensuite, nous verrons les algorithmes du boosting et de la descente de gradient. Puis, nous terminerons par la Gradient Boosting Machine avant la présentation de l'eXtreme Gradient Boosting.

3.4.1 Arbre de décision (CART)

Les arbres de décision ont été introduits en 1984 par Breimann et Freidmann. Il existe deux types de prédiction avec les arbres de décision, nous pouvons les utiliser autant pour de la régression pour prédire une quantité que pour de la classification pour créer des classes hétérogènes d'individus.

Cet algorithme partira de variables explicatives (par exemple, l'âge ou le sexe) pour prédire une variable d'intérêt aussi appelée variable cible (par exemple, le salaire). Il est très utilisé pour la prise de décision en entreprise.

Un arbre se présente de la façon suivante avec un noeud qui est la racine et d'autres noeuds appelés noeuds fils, puis se termine par des feuilles comme vous pouvez le voir dans la figure 36. :

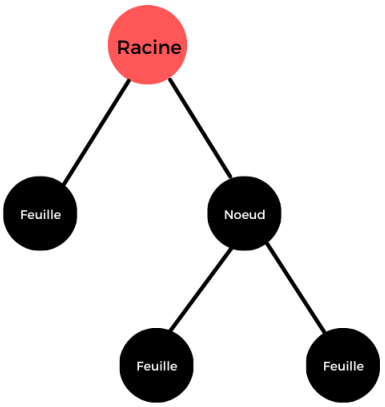

FIGURE 36 – Illustration d'un arbre de décision.

Ainsi, le principe de l'algorithme consiste à partir d'un ensemble d'individus de les séparer en deux sous-groupes en utilisant la variable qui séparera le mieux la population.

Puis, il s'arrêtera quand il n'y a plus qu'un seul individu dans un sous-groupe ou lorsque l'ensemble des individus contenus dans le dernier sous-groupe est homogène.

- Racine : contient la totalité des individus ou observations à séparer en groupe.
- Noeuds : correpondent aux noeuds ayant des fils et des descendants.
- Troncs, branches ou arêtes : correspondent aux règles de séparation pour classer la population.
- Feuille : est un sous-groupe homogène qui est créé lors de la séparation (les sous-groupes sont disjoints).

Par exemple, vous trouverez ci-dessous un exemple simple :

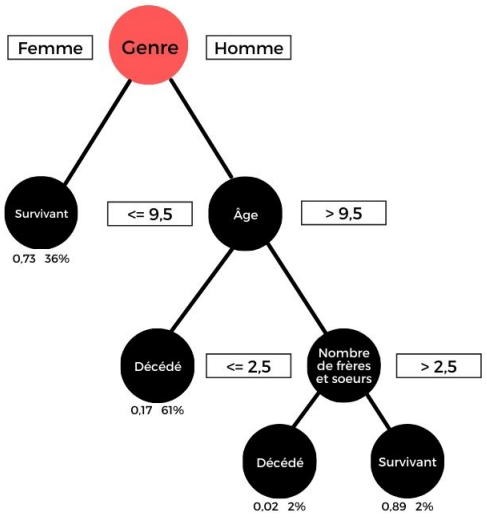

FIGURE 37 – Exemple d'un arbre de décision.

Pour cet exemple, nous nous concentrons sur l'estimation de la probabilité de survie des passagers du Titanic. Si nous prenons l'exemple d'un homme ayant 23 ans et une soeur, alors il survivra dans 98% des cas.

De façon plus formelle, soit X notre échantillon d'apprentissage ayant p variables explicatives et soit Y la variable cible. Nous voulons prédire cette variable à partir des variables explicatives $X_1, ..., X_p$. Prenons l'individu $x = (x_1, ..., x_p)$ avec p caractéristiques.

Lors de la construction de l'arbre de décision, au départ à la racine nous avons la totalité des individus. Ensuite, à chaque étape, nous séparons en deux groupes notre échantillon avec une des variables explicatives. Quand il s'agit d'une variable dite quantitative, alors nous définissons des intervalles. Tandis que lorsque c'est une variable qualitative, nous séparons en deux les modalités. Pour trouver la variable explicative qui nous permettra d'avoir la meilleure séparation, l'algorithme cherche à maximiser l'homogénéité des fils du noeud courant, ce qui revient à minimiser l'hétérogénéité qui peut s'exprimer de la façon suivante :

$$\hat{H}_n = \sum_{j=1, j\in\mathbb{N}}^{p_n} (y_j - \bar{y}_n)^2 - \left(\sum_{j=1, j\in\mathbb{N}_g}^{p_{Fg}} (y_j - \bar{y}_{Fg}) + \sum_{j=1, j\in\mathbb{N}_d}^{p_{Fd}} (y_j - \bar{y}_{Fd}) \right) \tag{23}$$

Où y_j est la valeur de la variable cible pour le ième individu, n représente le noeud courant, $\bar{y}_n$ est la moyenne empirique du noeud n, puis, p représente le nombre de caractéristiques étudiées, enfin, Fg et Fd représente respectivement les fils gauche et droite du noeud n. Cette mesure est le gain d'information obtenu par la séparation.

Ensuite, l'algorithme passe au noeud suivant que nous pouvons noter n_2. Ainsi, le gain d'information sera de nouveau calculé pour chacune des variables explicatives. De plus, la variable qui minimisera l'hétérogénéité sera choisie, comme vu précédemment. Finalement, l'algorithme d'arbre de décision se terminera quand les derniers individus seront dans la même classe.

3.4.2 Boosting

Le Boosting est une méthode adaptative améliorant l'ajustement à chaque étape avec une construction adaptative séquentielle d'estimateurs. Plus précisément, l'algorithme agrégera et combinera des modèles de façon séquentielle pour éviter le sur-apprentissage. Chaque nouveau modèle sera construit à partir de l'ancien en le corrigeant au fur et à mesure avec des poids sur les observations.

Si nous nous plaçons dans un cas pratique avec des données et que nous connaissons la manière dont nous voulons construire une théorie autour de celles-ci. Alors par exemple, nous créons un premier modèle en l'entraînant sur une partie des données appelée échantillon d'apprentissage, puis nous validerons ce premier modèle avec un échantillon test. Ensuite, nous analyserons les erreurs qui proviendront du premier modèle. A noter que c'est à partir de cette étape que le boosting entre en scène. L'algorithme de Boosting construira un nouveau modèle qui se concentrera sur les erreurs du précédent afin de les minimiser. Nous utiliserons ces erreurs lors d'une prochaine prédiction ce qui s'appelle une méthode d'agrégation de modèle ou d'ensemble. Il n'est pas obligatoire de s'arrêter à deux modèles, nous pouvons en faire autant que nous le désirons.

Si nous nous plaçons dans un cadre théorique avec la création de deux modèles, nous avons :

- X une variable explicative,
- Y notre variable cible,
- n le nombre d'observations.

Nous pouvons ainsi noter :

- f_0 notre modèle de départ qui peut être vu comme une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$ représentant les valeurs prédites de notre modèle,
- $\epsilon_i = y_i f_j(x_i)$ l'erreur de prédiction avec x_i une observation de la variable X et y_i l'observation cible de la variable Y pour le $j$ème modèle.
- g_0 représente les erreurs de prédiction de notre modèle f_0, g_0 est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$.

Par la suite, nous construisons un nouveau modèle noté $g_k(x_i) \sim \epsilon_i$ dont l'objectif est de prédire les observations où il y a des erreurs. Ainsi, nous obtiendrons dans le cadre d'une régression le modèle de boosting suivant :

$$f_{k+1}(x) = f_k(x) + g_k(x) \tag{24}$$

La méthode de Boosting peut être généralisée de la manière suivante. Soit notre modèle f_k que nous entraînons ayant comme résidu $\epsilon_i = -\frac{\delta l_i}{\delta f(x_i)}$. Notre prochain modèle corrigera les

erreurs du précédent nous le noterons $g_k(x_i) \sim \epsilon_i$, il minimisera les résidus avec une fonction de perte (par exemple, Mean Square Error MSE comme présentée dans la partie des mesures de performances) entre y et $f(x)$.

Précédemment, nous avons additionné les modèles en considérant qu'ils ont les mêmes poids, mais ça n'est pas forcément le cas. Donc, nous ajoutons un poids au modèle sur les erreurs, ce qui revient à écrire :

$$f_{k+1} = f_k(x) + \gamma_k g_k(x) \tag{25}$$

Ainsi, les poids nous permettront de donner plus ou moins d'importance au modèle. Nous avons pris l'exemple d'une fonction de perte MSE, mais nous pouvons aussi utiliser par exemple la Mean Absolute Error (MAE). Le Boosting permet d'obtenir un modèle plus performant. L'algorithme qui a été créé à ces fins se nomme l'algorithme de Boosting (Adaboost). Vous trouverez ci-dessous la description de cet algorithme.

Description de l'algorithme :
Notons :

- T : le nombre de modèles utilisés,
- X : la matrice de données avec les variables explicatives,
- y : notre variable cible,
- p_i : le poids de la $i$ème observation,
- f_0 : le premier modèle.

De plus, notons que : $\sum_{i=1}^{n} p_i > 0$.
Etape 1 : Initialisation des poids p_i, création de f_0 et création d'une liste L qui contiendra les modèles.

1 $p_i \leftarrow 1/n \ \forall i$;
2 $f_0(X) = g_0(X) = 0$;
3 $L \leftarrow [f_0]$

Etape 2 : Utilisation de boucles itératives pour la création des modèles.

pour t *allant de* 0 *à* $T-1$ **faire**
- Création aléatoire d'un échantillon d'entraînement noté $test$ à partir de X en utilisant des poids p_i;
- Construction d'un modèle d'apprentissage g_t pour prédire les erreurs sur les données de X qui ne sont pas dans $test$;
- Calcul du taux d'erreur commis noté $e_t = \frac{\sum_{i=1}^{n}(e_i \cdot p_i)}{\sum_{i=1}^{n} p_i}$;
- **si** $e_t > 0.5$ *ou* $e_t = 0$ **alors**
  - Arrêt de l'algorithme;
- **sinon**
  - Création du paramètre : $\Delta_t = \frac{1}{2} ln\left(\frac{1-e_t}{e_t}\right)$ permettant de définir la taille des étapes du modèle adaptatif boosting et de définir le poids de chaque modèle qui va être créé avec le taux d'erreur commis e_t;
  - Mise à jour du poids suivant $p_i = p_i e^{-\Delta_t}$;
  - Normalisation des poids p_i pour avoir une somme qui vaut 1.;
  - Création du nouveau modèle $f_t = f_{t-1} + \Delta_t g_t$;
- **fin**

fin

Ainsi, à la fin de l'algorithme nous obtiendrons le modèle finale suivant :

$$f_T(x) = sign\left(\sum_{t=0}^{T-1} \Delta_t . g_t\right) \tag{26}$$

Avec x correspondant à un vecteur des variables explicatives et $f_T(x)$ la prédiction faite de la variable cible y à partir des données d'entrée par notre modèle de Boosting. En règle générale, les modèles utilisés pour avoir un modèle d'apprentissage final sont les arbres de décision.

3.4.3 Descente de gradient

Dans cette partie, nous verrons la formalisation mathématique de l'algorithme de la descente du gradient. Cette algorithme permet d'approcher de façon itérative une solution d'un problème d'optimisation convexe.

Soit f une fonction deux fois différentiable avec $f : \mathbb{R}^n \to \mathbb{R}$ ($\mathbb{R}^n$ est muni de la norme $||.||$ et du produit scalaire noté $< . >$). Nous souhaitons minimiser la fonction f. Notons $\delta f(x)$ la différentielle de f en x et $\nabla_f(x) = \frac{\delta f}{\delta x_i}(x)$ pour chaque i le gradient de f en x. Nous avons la formule générale de la descente du gradient qui est la suivante :

$$x_{n+1} = x_n - \tau_n \nabla_f(x_n) \tag{27}$$

Avec τ_n représentant le pas de la descente et ∇_f la direction de descente.

A chaque itération, l'algorithme minimise la fonction $f : x \in \mathbb{R}^n \to f(x) \in \mathbb{R}$ différentiable pour obtenir $f(x_{n+1}) \leq f(x_n)$.

Description de l'algorithme :
Etape 1 : Initialisation

- Choix d'un point de départ $x_0 \in \mathbb{R}^n$,
- $s \geq 0$ un seuil qui est un critère d'arrêt pour l'algorithme.

Etape 2 : Boucle itérative avec critère d'arrêt.

tant que $||\nabla_f(x_n)|| \leq s$ **faire**
    Calcul de $\nabla_f(x_n)$;
    Mise à jour des coordonnées avec $x_{n+1} = x_n - \tau_n \nabla_f(x_n)$;
fin

Finalement, si nous appliquons cet algorithme de descente de gradient à la fonction de perte de notre modèle. Notons l la fonction de perte représentant l'écart entre y_i (la valeur réelle observée) et $\hat{y_i}$ (la valeur prédite à partir de x_i avec f). Donc, notre fonction de perte globale s'écrira :

$$\xi(f) = \sum_{i=1}^{n} l(y_i, f(x_i)) \tag{28}$$

L'objectif de l'algorithme de la descente du gradient est de minimiser la fonction ξ avec la fonction f qui représente le modèle agrégé. Ainsi, si nous écrivons la valeur du modèle à l'itération k de la manière suivante :

$$f_k(x) = f_{k-1}(x) - \tau_k \nabla_l(y, f(x))$$

En remplaçant $\nabla_l(y, f(x))$ par sa valeur, nous obtenons :

$$f_k(x) = f_{k-1}(x_j) - \tau_k \frac{\delta l(y_i, f(x_i))}{\delta f(x_i)}$$

Nous pouvons noter que Adaboost utilise une fonction de perte exponentielle. Ainsi, la fonction de perte de notre modèle s'écrira :

$$\xi(f) = \sum_{i=1}^{n} e^{-y_i f(x_i)}$$

3.4.4 Gradient Boosting Machine

Le Gradient Boosting Machine (GBM) est la combinaison de plusieurs modèles de type arbre de décision offrant la possibilité d'obtenir un modèle plus performant. L'objectif de cet algorithme est de construire un ensemble d'arbres basés sur les gradients de la fonction de perte et au même moment leur apprendre à résoudre des problèmes plus complexes.

Description de l'algorithme :

Prenons $(x_i, y_i)_{i=1,\ldots,n}$ le couple des données de la $i$ème observation de notre jeu de données. Soit x_i représente le vecteur des données de nos variables explicatives indépendantes et y_i la valeur de notre variable cible tout ceci pour la $i$ème observation.

Dans notre cas, la fonction de perte choisie est la Mean Square Error (MSE) ou encore appelé écart quadratique. La fonction de perte globale s'écrira de la manière suivante :

$$\Lambda(y_i, f(x_i)) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

Avec n représentant le nombre d'observations.

Etape 1 : Initialisation

1 $\Lambda(y_i, \nu) = x_i$;
2 $f_0(x) = \underset{\nu}{\operatorname{argmin}} \sum_{i=1}^{n} \Lambda(y_i, \nu)$;
3 A représentant le nombre d'arbres;

Etape 2 : Boucles itératives

1 **pour** *a allant de 1 à A* **faire**
2 Calcul des résidus : $e_{i,a} = -\left(\frac{\delta\Lambda(y_i, f(x_i))}{\delta f(x_i)}\right)_{f(x)=f_{a-1}(x)} \quad \forall i \in 1, n.$;
3 Ajustement du modèle de base sur les résidus, notons ce modèle $m_a(x)$ et est entraîné sur les données $(x_i, e_{i,a})\forall i$;
4 Calcul du paramètre de : $\nu_a = \underset{\nu}{\operatorname{argmin}} \sum_{i=1}^{n} \Lambda(y_i, f_{a-1})(x_i) + \nu m_a(x_i)$;
5 Création du nouveau modèle : $f_a(x) = f_{a-1}(x) + \nu_a$;
6 **fin**

3.4.5 eXtreme Gradient Boosting

L'eXtreme Gradient Boosting (XGBoost) construit de manière itérative des nouveaux modèles et les combine pour créer un modèle d'ensemble. Au départ, il construit un premier modèle et calcule les erreurs de chaque observation dans l'échantillon d'apprentissage. Puis, il construit un nouveau modèle pour prédire les résidus (erreurs). Finalement, il ajoute ces prédictions au modèle d'ensemble final. Le XGBoost est supérieur en le comparant à un algorithme de gradient boosting, car il permet d'avoir un bon équilibre entre le biais et la variance. L'algorithmme du Gradient Boosting optimise seulement la variance, car il a tendance à trop bien s'adapter à l'échantillon d'apprentissage. Ainsi, le XGBoost est une façon plus rapide et optimisée de l'algorithme de Gradient Boosting avec des arbres de décision. L'algorithme qui est ce cas particulier du Gradient Boosting se nomme le Tree Gradient Boosting.

Nous allons désormais expliquer de manière théorique et algorithmique l'eXtreme Gradient Boosting toujours pour le cas d'une régression. Il s'agit d'un système d'apprentissage avec un ensemble d'arbres. Notons qu'il utilise une fonction objective pour s'adapter aux données, nous pouvons l'écrire de la façon suivante :

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{a=1}^{A} \Omega(f_a)$$

Avec :

- La partie gauche correspondant à la fonction de perte que l'on notera par la suite $L(\Theta)$ qui mesure comment le modèle s'adapte aux données de l'échantillon d'apprentissage. Où :
  - $l(y_i, \hat{y}_i)$ représente la mesure permettant de connaître l'écart entre la prédiction et la réalité.
  - n représentant le nombre d'observations de notre échantillon d'apprentissage.
  - $\hat{y}_i$ représente l'observation prédite
  - y_i correspond à l'observation réelle.
- La partie droite correspondant à la complexité des arbres. Où :
  - $\Omega(f_a)$ représente la complexité des arbres du modèle à l'itération a.
  - A représente le nombre de modèles.

Dans le cadre de l'algorithme XGBoost, la fonction de perte se note de la manière suivante :

$$L(\Theta) = \sum_{i} (\hat{y}_i - y_i)^2$$

Où $l(y_i, \hat{y}_i) = (\hat{y}_i - y_i)^2$, avec $\hat{y}_i$ représente l'observation prédite et y_i l'observation réelle.

Ensuite, nous nous intéressons au calcul de la compléxité des arbres qui s'écrit finalement comme ci-dessous :

$$\Omega(f_a) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2$$

Avec :

- Le nombre de noeuds J.
- La norme 2 des poids des feuilles w_j.
- γ et λ sont des hyperparamètres.

Ainsi, nous voyons que la valeur prédite contenue dans la fonction de perte lors de l'apprentissage est comme un modèle additif. Ainsi, cela donnerait :

$$\hat{y_i}^{(a)} = \sum_{k=1}^{a} f_k(x_i) = \hat{y_i}^{(a-1)} + f_a(x_i)$$

Où :

- $\hat{y_i}^{(a)}$ représente le modèle finale,
- $\hat{y_i}^{(a-1)}$ correspond au modèle précédent,
- $f_a(x_i)$ est le nouveau modèle créé à l'itération a.

Notons qu'un compromis s'opère au niveau de l'optimisation. Ainsi, la fonction objective optimisera les pertes entre l'apprentissage et la complexité des arbres. Généralement, lorsque nous optimisons la perte d'apprentissage, alors cela encourage les modèles prédictifs à avoir une meilleure précision sur l'échantillon d'apprentissage. Alors que si nous optimisons la complexité des arbres, cela donne des modèles plus simples.

Finalement, nous pouvons réécrire notre fonction objective en utilisant les notations précédentes ce qui donnerait :

$$Obj^{(a)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(a-1)} + f_t(x_i)\right) + \Omega(f_a)$$

Dans la suite, nous allons utiliser l'approximation de Taylor qui est utilisée pour approximer une fonction plusieurs fois dérivable au voisinage d'un point par un polynôme dont les coefficients dépendent uniquement des dérivées de la fonction en ce point. Supposons qu'une fonction f soit de classe C^n sur I qui est un intervalle de $\mathbb{R}$. Alors, pour tout $\Delta \in \mathbb{R}$ tel que $x + \Delta x$ appartienne à I. Nous avons pour la formule de Taylor d'ordre 2 l'expression suivante :

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2$$

Si nous appliquons ce théorème aux deux termes de la fonction objective, alors cela nous donne :

$$Obj^{(a)} \approx \sum_{i=1}^{a} \left[l\left(y_i, \hat{y_i}^{(a-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right]$$

Où :

- $g_i = \delta_{\hat{y}^{(a-1)}}(\hat{y}^{(a-1)} - y_i)^2$
- $h_i = \delta^2_{\hat{y}^{(a-1)}}(y_i - \hat{y}^{(a-1)})^2$

Désormais, nous additionnons la fonction de perte d'apprentissage et la complexité des arbres, nous obtenons le résultat suivant :

$$Obj^{(a)} = \sum_{i=1}^{a} \left[g_i f_a(x_i) + \frac{1}{2} h_i f_a^2(x_i) \right] + \Omega(f_a)$$

Suite à cela, nous pouvons simplifier l'expression de la fonction objective par :

$$Obj^{(a)} = \sum_{j=1}^{n} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma a$$

Où les valeurs g_i et h_i représentent la valeur de chaque feuille dans l'arbre :

- $G_j = \sum_{i \in I_j} g_i$
- $H_j = \sum_{i \in I_j} h_i$

Désormais, nous sommes capables de dériver chaque fonction quadratique de la somme qui est de la forme suivante :

$$G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2$$

Afin de trouver un poids optimal noté :

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

Finalement, si nous injectons cette valeur optimale dans la fonction objective d'origine nous obtenons une nouvelle fonction objective ne contenant pas de w_j que nous appellerons la fonction objective minimale :

$$minObj = -\frac{1}{2} \sum_{j=1}^{A} \frac{G_j^2}{H_j + \lambda} + \gamma A$$

Description de l'algorithme :
Soient $Y \in \mathbb{R}$ la variable cible et $X \in \mathbb{R}$ un vecteur de p variables explicatives et indépendantes.

Etape 1 : Initialisation

- $f_0(x_i)$ un premier modèle constant,
- A le nombre d'arbres qui vont être construits.

Etape 2 : Boucles itératives

pour *a allant de 1 à A* **faire**
  Création d'un nouveau modèle $f_a(x_i)$;
  Ajout du nouveau modèle au précédent;
  Apprentissage du modèle $f_a(x_i)$;
  Recherche de l'objectif minimum avec $minObj = -\frac{1}{2}\sum_{j=1}^{a}\frac{G_j^2}{H_j+\lambda} + \gamma a$ permettant de mesurer la qualité de la structure de l'arbre;
  L'arbre est construit d'après le précédent algorithme présenté pour le Gradient Boosting adapté aux arbres;
  pour *k allant de* 0 *à* n **faire**
    $Gain_k = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_l+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$;
    si $Gain_k < 0$ **alors**
      Arrêt de la boucle;
    sinon
      Ajout de la meilleure partition;
    fin
  fin
  si $minObj > 0,5$ *ou* $minObj = 0$ **alors**
    Arrêt de l'algorithme;
  sinon
    $f_{(a+1)}(x_i) = f_a(x_i) + f_{(a-1)}(x_i)$
  fin
fin

Sachant que G_L et G_R sont respectivement les noeuds fils gauche et droite du noeud courant.

Le modèle final est une combinaison de tous les modèles contenus dans la boucle et il s'agira du modèle d'ensemble XGBoost qui peut s'écrire de la façon suivante :

$$\hat{y_i} = \sum_{a=1}^{A} f_a(x_i)$$

4 Application des méthodes

Dans ce chapitre, nous nous plaçons dans le cadre d'une tarification en santé collective. Pour notre partie application, nous allons désormais séparer l'analyse en deux pour les deux portefeuilles. L'objectif de notre étude est de calculer la prime pure en frais réels et en remboursement de la part de la mutuelle à partir des données qui nous étaient accessibles et dont nous avons pu faire une première présentation dans le chapitre 1. Pour cela, nous réaliserons des modèles linéaires généralisés (GLM) avec la loi Tweedie sur la totalité des données sur chacun de nos portefeuilles avec deux montants différents (en frais réels et en montant remboursé par la mutuelle). Ensuite, nous utiliserons une méthode fréquence-coût moyen avec des GLM pour calculer la prime pure. Finalement, nous ferons une dernière estimation de la prime pure en utilisant un dernier algorithme le l'eXtreme Gradient Boosting (XGBoost) sur la totalité de nos données. Il faut noter que les données utiliser pour ces estimations de primes pures représentent 95% des sinistres dont les 5% restant constituent les graves.

Pour chacun de nos modèles, nous séparons notre échantillon en deux sous-échantillons, un premier pour s'entraîner (qui contiendra 80% de notre jeu de données) et un second pour valider notre modèle (contenant 20% de notre échantillon).

4.1 Etude des dépenses pour chaque variable

Nous étudions les dépenses moyennes par modalité de variables pour chacun des montants (en frais réels, et en remboursement de la part de la mutuelle) pour comprendre leur impact et voir si les variables ont un lien statistique (de dépendance) avec notre variable d'intérêt.

1. Pour le premier portefeuille :
    - Répartition des dépenses par poste :

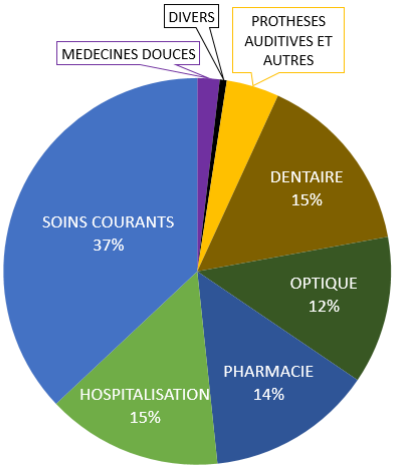

FIGURE 38 – Répartition des dépenses en frais réels par poste pour le premier portefeuille.

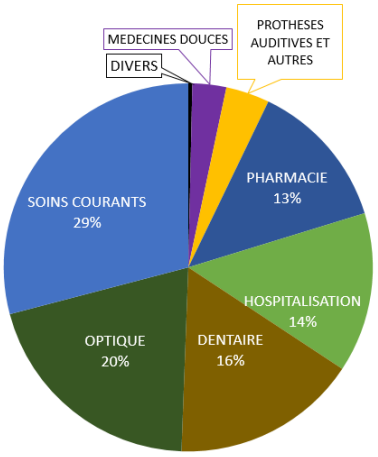

FIGURE 39 – Répartition des dépenses de la mutuelle par poste pour le premier portefeuille.

Les postes qui représentent la plus grande part de dépenses en frais réels sont les soins courants, l'hospitalisation, et le dentaire. Tandis que pour les dépenses de la part de la mutuelle ce sont les soins courants, l'optique et le dentaire.

- Dépenses moyennes annuelles pour un assuré par poste :

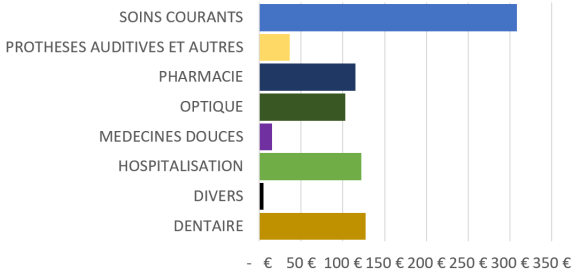

FIGURE 40 – Dépenses moyennes annuelles par assuré par poste (en frais réels) du portefeuille 1.

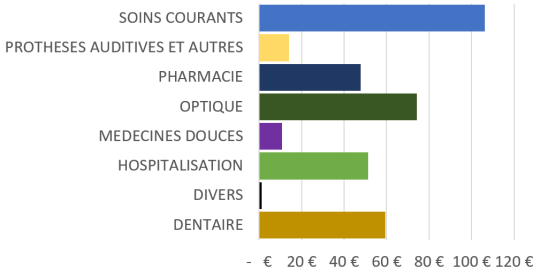

FIGURE 41 – Dépenses moyennes annuelles par assuré par poste (pour la mutuelle) du portefeuille 1.

Nous pouvons remarquer que le poste où il y a le plus de dépenses en moyenne en frais réels ou en remboursement de la part de la mutuelle sont les soins courants.

- Tableau récapitulatif des dépenses par sous-poste :

Poste	**Sous-poste**	**Charge annuelle moyenne en frais réels par assuré**	**Charge annuelle moyenne de la mutuelle par assuré**	**Pourcentage du coût total (en frais réels)**	**Pourcentage du coût total (en dépenses mutuelle)**
Dentaire	Orthodontie	28€	15€	3,4%	4,2%
	Prothèses dentaires	66€	34€	7,9%	9,4%
	Radiologie (Dentaire)	4€	1€	0,5%	0,3%
	Soins dentaires	29€	9€	3,5%	2,4%
Divers	Autre (Divers)	5€	1€	0,6%	0,4%
Hospitalisation	Forfait hospitalier	8€	8€	0,9%	2,1%
	Frais de séjour	53€	14€	6,4%	3,9%
	Frais optionnels	17€	16€	2,1%	4,3%
	Soins hospitaliers	39€	12€	4,7%	3,2%
	Transport	5€	2€	0,6%	0,5%
Médecines douces	Médecines douces	16€	11€	1,9%	3%
Optique	Lentilles	9€	8€	1,1%	2,1%
	Monture	29€	21€	3,5%	5,8%
	Verres	64€	45€	7,7%	12,4%
Pharmacie	Pharmacie remboursement 15%	4€	4€	0,5%	1%
	Pharmacie remboursement 30%	12€	8€	1,5%	2,3%
	Pharmacie remboursement 65%	76€	26€	9,2%	7,2%
	Autre (Pharmacie)	22€	9€	2,6%	2,5%
Prothèses auditives et autres prothèses	Autres prothèses et véhicules	8€	3€	0,9%	0,7%
	Petits matériels et pansements	29€	11€	3,5%	3,1%
Soins courants	Analyses médicales	47€	18€	5,6%	4,9%
	Auxiliaires médicaux	61€	24€	7,3%	6,5%
	Consultation spécialiste	68€	25€	8,2%	6,9%
	Consultation et visite en médecine générale	45€	13€	5,5%	3,6%
	Frais complémentaires de consultations	10€	3€	1,2%	0,9%
	Radiologie (Soins courants)	23€	7€	2,7%	1,8%
	Autres soins courants	54€	16€	6,5%	4,5%

TABLE 12 – Tableau récapitulatif des dépenses et des proportions par sous-poste pour le premier portefeuille.

Le sous-poste représentant le plus de dépenses en frais réels est la pharmacie remboursée à 65% par la Sécurité Sociale. Alors que ceux représentant le plus de dépenses en moyenne par assuré pour la mutuelle sont les prothèses dentaires et les lunettes.

- Charge annuelle moyenne pour un assuré par sexe :

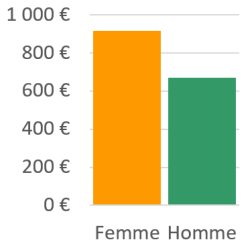

FIGURE 42 – Charge annuelle moyenne pour un assuré par sexe (en frais réels) du portefeuille 1.

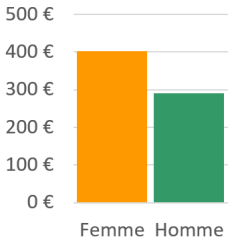

FIGURE 43 – Charge annuelle moyenne pour un assuré par sexe (en remboursement mutuelle) du portefeuille 1.

Dans les graphiques 42 et 43, nous pouvons faire les mêmes observations, c'est-à-dire qu'en moyenne les femmes ont tendance à peu plus consommer que les hommes en frais de santé. Ceci peut être expliqué par le fait que le portefeuille 1 est composé de beaucoup plus de femmes que d'hommes. Ainsi, nous pouvons voir que le sexe a un effet sur la charge moyenne.

- Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire :

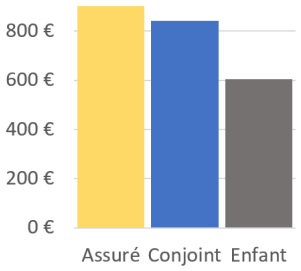

FIGURE 44 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en frais réels) du portefeuille 1.

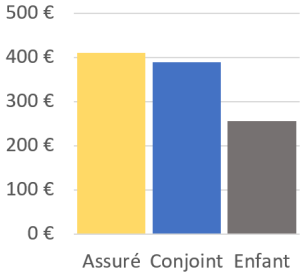

FIGURE 45 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 1.

Les graphiques ci-dessus montrent qu'en moyenne les assurés sont ceux qui dépensent le plus en frais de santé. Nous pouvons aussi noter que la charge moyenne pour les enfants est beaucoup plus basse que le reste des modalités, ce qui peut montrer que cette variable a un effet sur la charge moyenne.

- Charge annuelle moyenne pour un assuré par tranche d'âges :

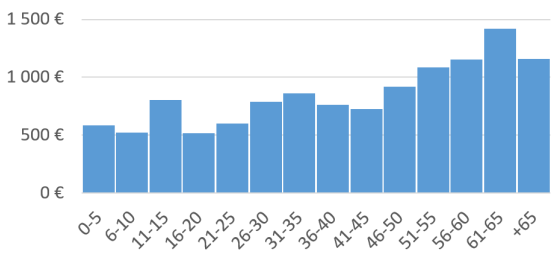

FIGURE 46 – Charge annuelle moyenne pour un assuré par tranche d'âges (en frais réels) du portefeuille 1.

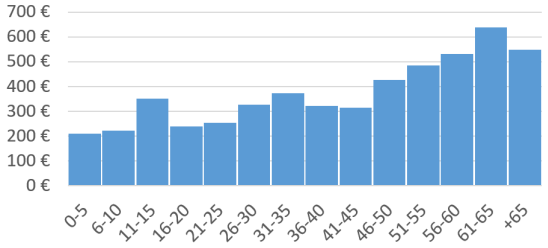

FIGURE 47 – Charge annuelle moyenne pour un assuré par tranche d'âges (en remboursement mutuelle) du portefeuille 1.

A l'aide des graphiques précédents, nous pouvons qu'en moyenne la tranche d'âge entre 11 et 15 est celle qui consomme le plus avant 21 ans ce qui peut être expliqué par l'orthodontie. De plus, nous pouvons voir l'effet de l'âge, plus les assurés vieillissent plus ils consomment en frais de santé.

- Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire :

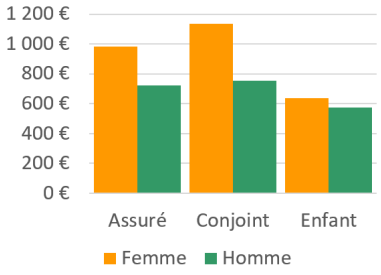

FIGURE 48 – Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en frais réels) du portefeuille 1.

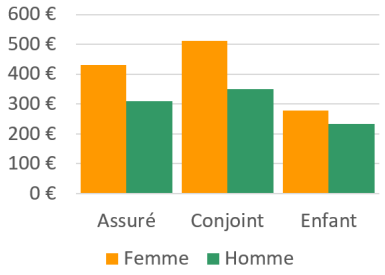

FIGURE 49 – Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 1.

D'après les graphiques précédents, nous pouvons dire que les individus qui consomment le plus en moyenne sont les conjointes.

- Charge annuelle moyenne pour un assuré par sexe et par tranche d'âges :

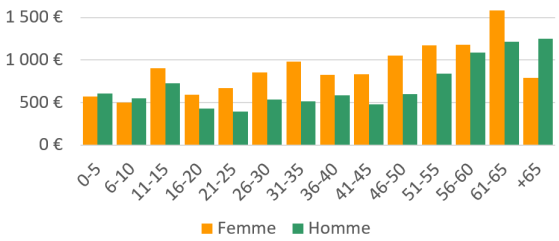

FIGURE 50 – Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en frais réels) du portefeuille 1.

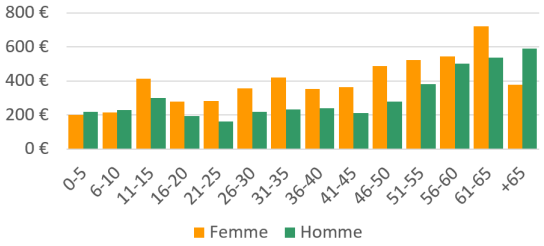

FIGURE 51 – Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en remboursement mutuelle) du portefeuille 1.

En moyenne pour les frais réels, les personnes les plus consommant d'après le graphique 50 sont les femmes entre 61 et 65 ans. Puis en moyenne pour les remboursements mutuelle, les personnes consommant le plus sont les femmes entre 46 et 65 ans.

- Charge annuelle moyenne pour un assuré par sexe, catégorie du bénéficiaire et tranche d'âges :

D'après le graphique 52 ci-dessus, nous pouvons en conclure que les personnes consommant le plus sont les individus entre 61 et 65 ans en terme de frais réels moyens.

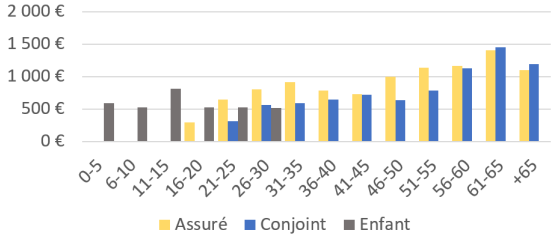

FIGURE 52 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en frais réels) du portefeuille 1.

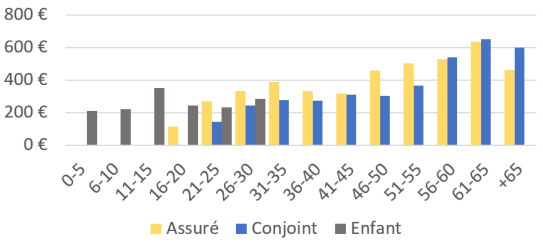

FIGURE 53 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en remboursement mutuelle) du portefeuille 1.

Le graphique 53 nous permet de dire qu'en moyenne les personnes consommant le plus en terme de remboursement de la part de la mutuelle sont les conjoints de plus de 61 ans en terme de remboursement de la part de la mutuelle.

2. Pour le second portefeuille :

   - Répartition des dépenses par poste :

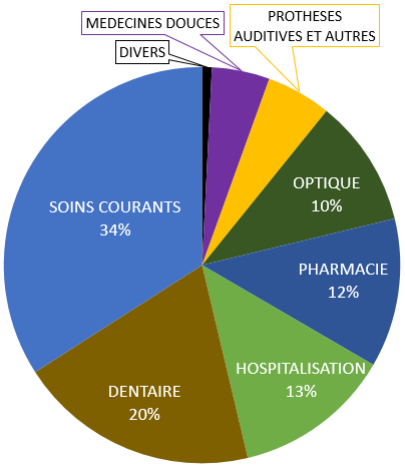

FIGURE 54 – Répartition des dépenses en frais réels par poste pour le second portefeuille.

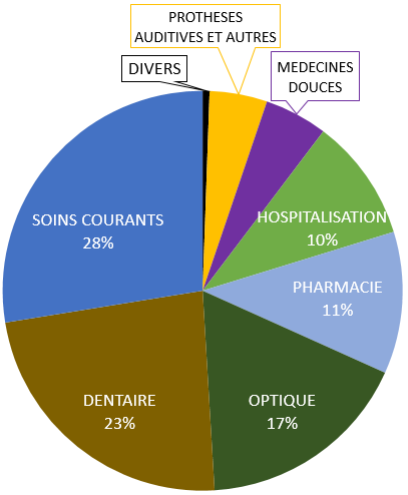

FIGURE 55 – Répartition des dépenses de la mutuelle par poste pour le second portefeuille.

Pour les dépenses en frais réels, les postes étant les plus présents sont les soins courants, le dentaire et l'hospitalisation. Tandis que pour les dépenses de la mutuelle, nous avons les soins courants, le dentaire et l'optique.

   - Dépenses moyennes annuelles pour un assuré par poste :

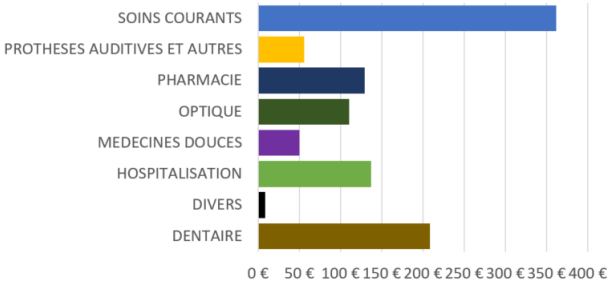

FIGURE 56 – Dépenses moyennes annuelles par assuré par poste (en frais réels) du portefeuille 2.

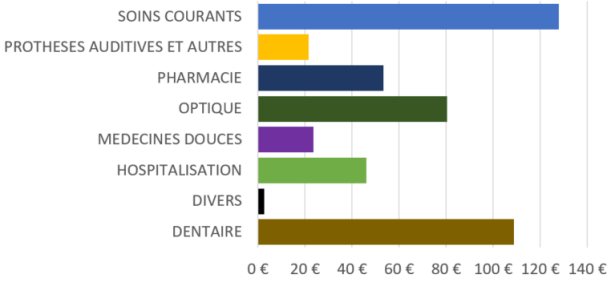

FIGURE 57 – Dépenses moyennes annuelles par assuré par poste (pour la mutuelle) du portefeuille 2.

Nous observons que le poste ayant les dépenses les plus élevées en moyenne pour les frais réels ou pour la mutuelle sont les soins courants.

- Tableau récapitulatif des dépenses par sous-poste :

Poste	**Sous-poste**	**Charge annuelle moyenne en frais réels par assuré**	**Charge annuelle moyenne de la mutuelle par assuré**	**Pourcentage du coût total (en frais réels)**	**Pourcentage du coût total (en dépenses mutuelle)**
Dentaire	Orthodontie	61€	35€	5,8%	7,5%
	Prothèses dentaires	109€	64€	10,3%	13,8%
	Radiologie (Dentaire)	2€	1€	0,2%	0,1%
	Soins dentaires	36€	9€	3,4%	2%
Divers	Autre (Divers)	8€	3€	0,8%	0,6%
Hospitalisation	Forfait hospitalier	7€	7€	0,7%	1,6%
	Frais de séjour	60€	11€	5,6%	2,4%
	Frais optionnels	17€	13€	1,6%	2,9%
	Soins hospitaliers	48€	13€	4,6%	2,8%
	Transport	4€	2€	0,4%	0,3%
Médecines douces	Médecines douces	50€	24€	4,7%	5,1%
Optique	Lentilles	9€	2€	0,9%	0,3%
	Monture	37€	25€	3,5%	5,3%
	Verres	64€	54€	6%	11,7%
Pharmacie	Pharmacie remboursement 15%	6€	5€	0,5%	1%
	Pharmacie remboursement 30%	13€	9€	1,3%	2%
	Pharmacie remboursement 65%	68€	23€	6,4%	4,8%
	Autre (Pharmacie)	43€	17€	4%	3,7%
Prothèses auditives et autres prothèses	Autres prothèses et véhicules	30€	11€	2,8%	2,3%
	Petits matériels et pansements	27€	11€	2,5%	2,4%
Soins courants	Analyses médicales	45€	17€	4,3%	3,6%
	Auxiliaires médicaux	69€	27€	6,5%	5,9%
	Consultation spécialiste	105€	38€	9,8%	8,1%
	Consultation et visite en médecine générale	76€	25€	7,2%	5,3%
	Frais complémentaires de consultations	12€	4€	1,1%	0,8%
	Radiologie (Soins courants)	55€	18€	5,2%	3,8%

TABLE 13 – Tableau récapitulatif des dépenses et des proportions par sous-poste pour le second portefeuille.

Nous remarquons que le sous-poste prothèses dentaires représente le plus de dépenses autant en frais réels que pour la mutuelle, mais aussi en charge moyenne annuelle par assuré.

- Charge annuelle moyenne pour un assuré par sexe :

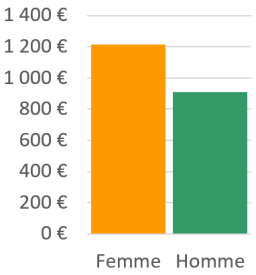

FIGURE 58 – Charge annuelle moyenne pour un assuré par sexe (en frais réels) du portefeuille 2.

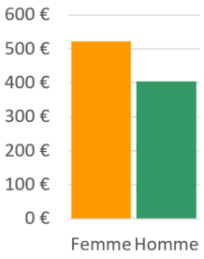

FIGURE 59 – Charge annuelle moyenne pour un assuré par sexe (en remboursement mutuelle) du portefeuille 2.

A l'aide des informations contenues dans les graphiques précédents, nous pouvons en conclure qu'en moyenne en terme de frais réels ou de dépenses mutuelle les personnes consommant un peu plus sont les femmes.

- Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire :

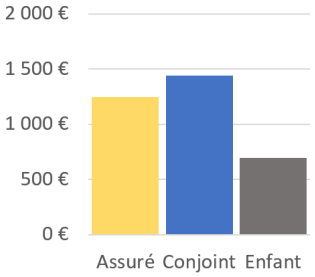

FIGURE 60 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en frais réels) du portefeuille 2.

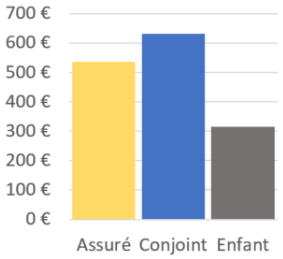

FIGURE 61 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 2.

Les informations que nous pouvons tirer des graphiques ci-dessus sont les personnes consommant le plus en moyenne qui sont les conjoints et le moins étant les enfants.

- Charge annuelle moyenne pour un assuré par tranche d'âges :

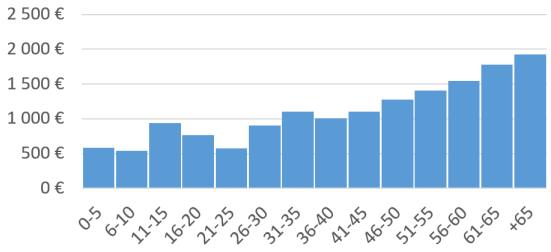

FIGURE 62 – Charge annuelle moyenne pour un assuré par tranche d'âges (en frais réels) du portefeuille 2.

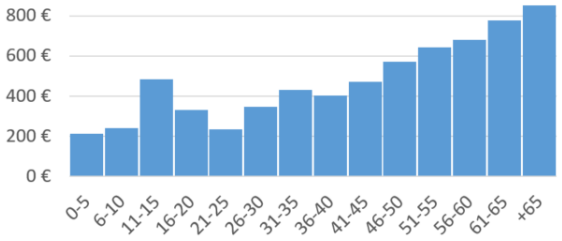

FIGURE 63 – Charge annuelle moyenne pour un assuré par tranche d'âges (en remboursement mutuelle) du portefeuille 2.

Au moyen des éléments contenus dans les figures précédentes, nous observons l'influence de l'âge sur les charges moyennes.

- Charge annuelle moyenne pour un assuré par région :

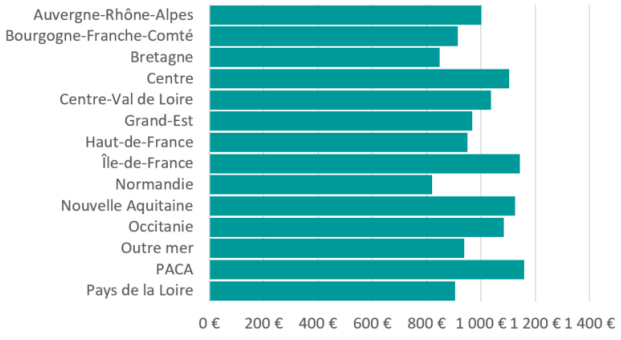

FIGURE 64 – Charge annuelle moyenne pour un assuré par région (en frais réels) du portefeuille 2.

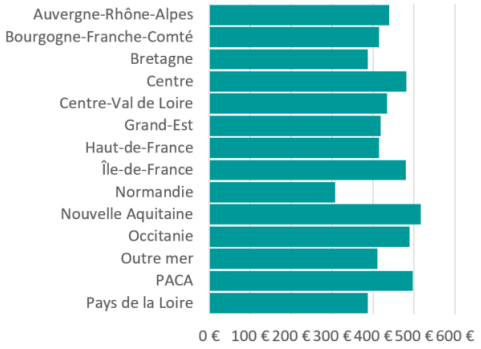

FIGURE 65 – Charge annuelle moyenne pour un assuré par région (en remboursement mutuelle) du portefeuille 2.

Les figures 64 et 65 mettent en exergue les régions où l'on consomme le plus en moyenne, comme en PACA, en Nouvelle Aquitaine et en Île-de-France.

- Charge annuelle moyenne pour un assuré par niveau de garantie :

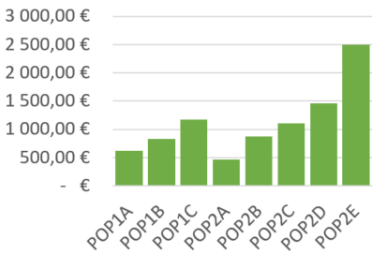

FIGURE 66 – Charge annuelle moyenne pour un assuré par niveau de garantie (en frais réels) du portefeuille 2.

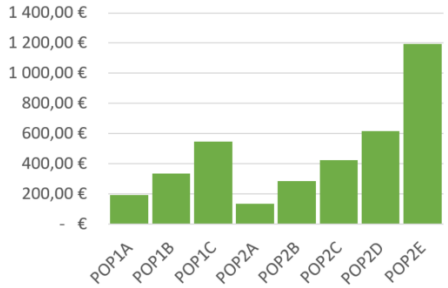

FIGURE 67 – Charge annuelle moyenne pour un assuré par niveau de garantie (en remboursement mutuelle) du portefeuille 2.

Avec les graphiques ci-dessus, nous voyons l'effet des niveaux de garantie. Plus le niveau de garantie est élevé, plus les personnes sont couvertes ce qui explique l'augmentation de la consommation.

- Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire :

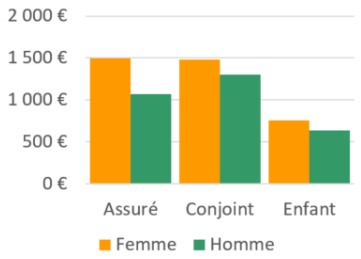

FIGURE 68 – Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en frais réels) du portefeuille 2.

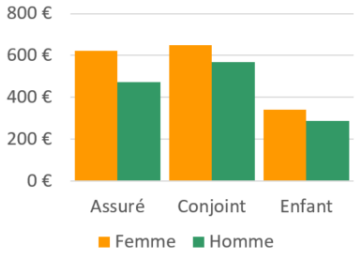

FIGURE 69 – Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 2.

D'après les graphiques 68 et 69, en moyenne, les femmes ont tendance à consommer légèrement plus que les hommes. De plus, nous pouvons remarquer que chez les hommes ce sont les conjoints qui consomment le plus.

- Charge annuelle moyenne pour un assuré par sexe et par tranche d'âges :

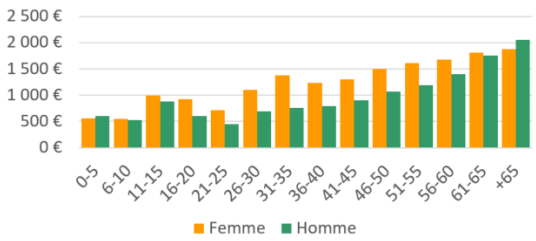

FIGURE 70 – Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en frais réels) du portefeuille 2.

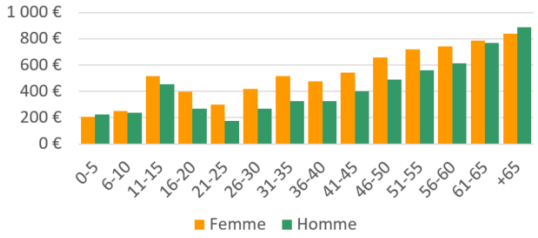

FIGURE 71 – Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en remboursement mutuelle) du portefeuille 2.

Avec les graphiques ci-dessus, nous pouvons dire qu'en moyenne entre 11 et 65 ans les femmes consomment un peu plus que les hommes.

- Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et par tranche d'âges :

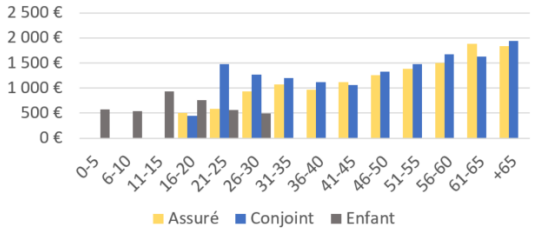

FIGURE 72 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en frais réels) du portefeuille 2.

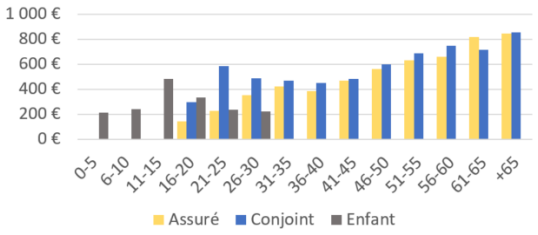

FIGURE 73 – Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en remboursement mutuelle) du portefeuille 2.

D'après les graphiques précédents, nous pouvons noter qu'il y a une antisélection des conjoints entre 21-25 ans. De plus, les conjoints consomment nettement plus que les assurés principaux entre 25 et 36 ans. Nous pouvons bien observer le pic de l'orthodontie entre 11 et 15 ans.

4.2 Modèle linéaire généralisé

Dans cette section, nous verrons comment nous avons modélisé notre prime pure avec des modèles linéaires généralisés. Dans un premier temps, nous modélisons la prime pure par assuré au niveau global pour les deux portefeuilles et pour les deux types de montants (en frais réels et en remboursement de la part de la mutuelle) avec la loi Tweedie.

Dans un deuxième temps, nous modélisons par sous-poste (verre, monture,etc.) avec la méthode fréquence-coût moyen. Certains sous-postes ne représentant que très peu de données (moins de 3% des dépenses en frais réels), nous les modélisons avec des GLM Tweedie par poste (dentaire, optique, etc.) auxquels nous appliquons un coefficient.

Finalement, nous terminons avec une troisième méthode avec le XGBoost pour modéliser les dépenses en frais réels et montant remboursé par la mutuelle. Chaque modèle est effectué sur deux types de montants différents sur les frais réels dépensés, puis sur les montants remboursés par la mutuelle, mais aussi sur deux portefeuilles différents. Par soucis de simplification et de compréhension, nous avons décidé que pour la présentation des modélisations fréquence-coût moyen, nous exposerons les résultats seulement d'un poste pour un portefeuille.

4.2.1 Modélisation GLM Tweedie

Dans le cas de cette modélisation, nous avons décidé d'effectuer cette méthode tous postes confondus pour la prime pure directement en utilisant la distribution Tweedie. L'hypothèse d'indépendance entre la fréquence et le coût moyen est remis en cause dans ce cas-là, puisque nous mélangeons des postes totalement différents.

- **Premier portefeuille :**

  D'après le graphique suivant, la valeur estimée la plus adéquate pour les frais réels semble être sélectionnée 1,6 pour *var.power* et 1,5 dans le cas des montants remboursés par la mutuelle. Nous vous renvoyons à la définition de ce paramètre contenu dans la partie concernant la théorie.

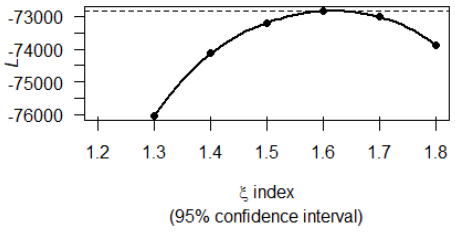

FIGURE 74 – Estimation du paramètre tweedie *var.power* (frais réels) pour le modèle global du portefeuille 1.

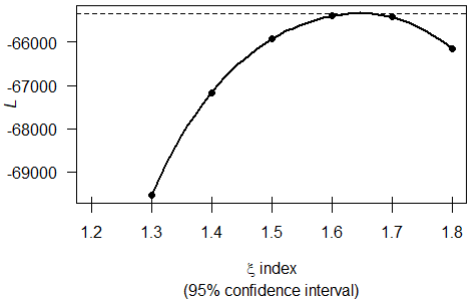

FIGURE 75 – Estimation du paramètre tweedie *var.power* (remboursement mutuelle) pour le modèle global du portefeuille 1.

Ensuite, nous réalisons un modèle dit saturé, c'est-à-dire contenant les variables explicatives suivantes *SEXE_BENEF* et *cat_class*. Par ailleurs, la variable d'exposition *EXPOSITION* est en *offset*. L'année de couverture n'est pas significative, donc nous avons décidé de ne pas la prendre en compte dans nos modèles. Puisqu'en santé nous avons les mêmes besoins d'une année sur l'autre, alors qu'en dommage aux biens par exemple cela peut varier. Finalement, notre modèle s'écrit de la manière suivante :

$$\mathbb{E}(Y_{cout}) = exp(\beta_0 + \beta_{SexeHomme} X_{SexeHomme} + \beta_{cat\_classAssure\_16-20} X_{cat\_classAssure\_16-20} + \dots$$

$$+ \beta_{cat\_classEnfant\_26-30} X_{cat\_classEnfant\_26-30} + offset(log(EXPOSITION)))$$

D'après le tableau ci-dessous, nous pouvons noter que les modèles ne possèdent pas une bonne qualité d'ajustement étant donné que $\chi^2 < D$, comme nous pouvons voir dans le tableau ci-dessous :

Modèle	Deviance	χ^2	AIC
Frais réels	257 028	$\chi^2_{(0.95,13393)} = 13663.34$	144 837,7
Remboursement mutuelle	309 006	$\chi^2_{(0.95,13393)} = 13663.34$	131 437,4

TABLE 14 – Mesures de performance du modèle GLM Tweedie au niveau global du portefeuille 1.

Dans le tableau suivant, vous pouvez retrouver les résultats des deux modèles avec leur coefficients estimés, ainsi que leurs significativité. Par exemple pour le modèle des frais réels, toute chose égale par ailleurs en moyenne, les hommes ont tendance à avoir relativement moins de dépenses que les femmes en santé. Nous pouvons aussi remarquer avec les coefficients estimés qu'ils semblent y avoir une augmentation des dépenses avec l'âge chez les assurés et les conjoints. Par ailleurs en moyenne, nous pouvons ajouter qu'il semble y avoir toute chose égale par ailleurs une augmentation des dépenses entre 11 et 15 ans par rapport au reste des enfants que nous pouvons imaginer être dû à l'orthodontie.

Variables	Modèle frais réels			Modèle remboursement mutuelle		
	Coefficient estimé	*pvalue*	*Significativité*	*Coefficient estimé*	*pvalue*	*Significativité*
Intercept	6.728	<2e-16	***	5.866	<2e-16	***
SEXE_BENEFHomme	-0.282	1.13e-11	***	-0.323	1.16e-15	***
cat_classAssure_16-20	-1.084	0.001	**	-1.182	0.001	***
cat_classAssure_21-25	-0.265	0.007	.	-0.251	0.020	**
cat_classAssure_31-35	0.159	0.063		0.153	0.064	.
cat_classAssure_36-40	-0.092	0.306		-0.071	0.418046	
cat_classAssure_41-45	-0.070	0.440		-0.033	0.705169	
cat_classAssure_46-50	0.226	0.008	**	0.334	3.60e-05	***
cat_classAssure_51-55	0.351	1.13e-05	***	0.410	9.05e-08	***
cat_classAssure_56-60	0.355	1.87e-05	***	0.445	1.74e-08	***
cat_classAssure_61-65	0.615	1.90e-08	***	0.692	1.75e-11	***
cat_classAssure_+65	0.330	0.199		0.394	0.11	
cat_classConjoint_21-25	-1.721	0.085	.	-0.704	0.157	
cat_classConjoint_26-30	-0.138	0.595		-0.068	0.790	
cat_classConjoint_31-35	-0.217	0.214		-0.066	0.696	
cat_classConjoint_36-40	-0.022	0.88		-0.008	0.958	
cat_classConjoint_41-45	-0.038	0.803		0.022	0.884	
cat_classConjoint_46-50	-0.137	0.366		-0.004	0.976	
cat_classConjoint_51-55	0.155	0.33		0.266	0.078	.
cat_classConjoint_56-60	0.509	0.001	***	0.689	7.44e-08	***
cat_classConjoint_61-65	0.762	1.22e-06	***	0.852	6.60e-09	***
cat_classConjoint_+65	0.669	0.001	***	0.885	8.90e-08	***
cat_classEnfant_0-5	-0.098	0.2931		-0.271	0.004054	**
cat_classEnfant_6-10	-0.372	8.58e-05	***	-0.348	0.0001	***
cat_classEnfant_11-15	0.057	0.5330		0.117	0.188	
cat_classEnfant_16-20	-0.341	0.0004	***	-0.242	0.010	***
cat_classEnfant_21-25	-0.544	0.001	***	-0.408	0.010	***
cat_classEnfant_26-30	-0.595	0.289		-0.420	0.439	

TABLE 15 – Résultats du modèle GLM Tweedie au niveau global pour le portefeuille 1.

Dans les graphiques ci-dessous sont représentés les différents résidus des modèles. Dans la grande majorité des cas, les points représentant les résidus de Pearson standardisés se situent autour de 0. Il y a quelques points vers 100 et deux autour de 250. Nous pouvons dire que cela reste satisfaisant et pour valider les modèles.

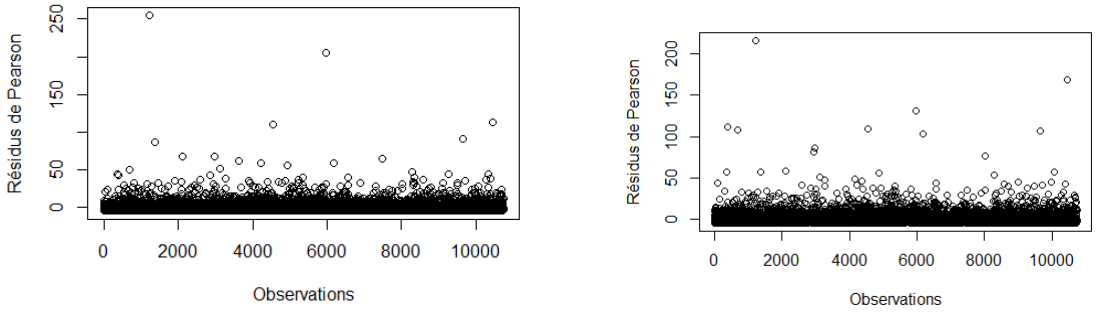

FIGURE 76 – Résidus du modèle de régression Tweedie (frais réels à gauche et montant remboursé par la mutuelle à droite) tous postes confondus pour le portefeuille 1.

D'après le tableau 16, la prédiction du modèle de Tweedie obtient les moyennes sont plutôt proches entre les observations et les prédictions.

Prime pure	**Observé**	**Prédiction**
Frais réels	835,40€	817,83€
Remboursement mutuelle	359,16€	359,16€

TABLE 16 – Résumé des moyennes du modèle GLM Tweedie tous postes confondus du portefeuille 1.

Dans le tableau 17, nous présentons différentes mesures de performance qui nous permettront de comparer les modèles GLM entre eux. Pour les modèles frais réels et les montants remboursés de la mutuelle tous postes confondus, il a été retenu qu'ils ont relativement une mauvaise qualité d'ajustement. Néanmoins, ils possèdent des valeurs faibles de la mesure MAE ce qui tend à montrer une bonne prédiction de leur part. Finalement, nous pouvons ajouter que nos modèles ont une forte volatilité au vu de la grandeur des mesures de performance MSE et RMSE connues pour être sensibles aux valeurs aberrantes.

Modèle	*MAE*	*RMSE*	*MSE*	*AIC*	*Deviance*	χ^2
Frais réels	608,7	1093,31	1 195 340	144 837,7	257 028	13 662,33
Montants de la mutuelle	277,23	486,14	236336,4	131 435,6	309 010	13 661,32

TABLE 17 – Mesure de précision du GLM Tweedie tous postes confondus du portefeuille 1.

- **Second portefeuille :**

Dans un premier temps, nous estimons la valeur optimale pour le paramètre *var.power* du modèle *Tweedie* à l'aide des graphiques ci-dessous nous retenons 1,6 (pour les deux montants) :

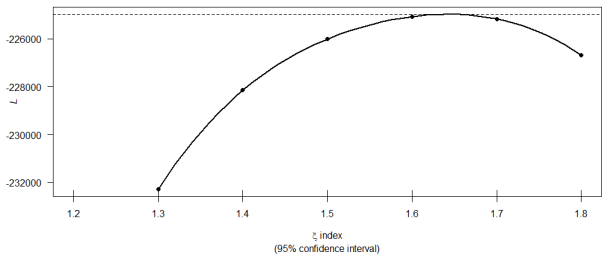

FIGURE 77 – Estimation du paramètre tweedie (frais réels) *var.power* du modèle GLM Tweedie tous postes confondus du portefeuille 2.

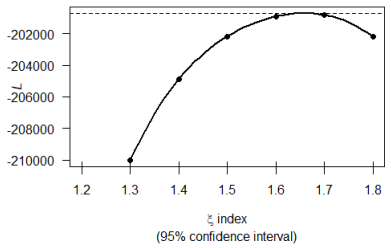

FIGURE 78 – Estimation du paramètre tweedie (mutuelle) *var.power* du modèle GLM Tweedie tous postes confondus du portefeuille 2.

Ensuite, nous réalisons un modèle avec les variables explicatives suivantes *SEXE*, *cat_class*, *REGION* et *GARANTIE* ? Nous utilisons l'exposition en *offset*. L'année de couverture n'est pas significative, donc nous ne la considérons pas pour nos modèles. Finalement, notre modèle s'écrit de la manière suivante :

$$\mathbb{E}(Y_{cout}) = exp(\beta_0 + \beta_{SexeHomme} X_{SexeHomme} + \beta_{cat\_classAssure\_16-20} X_{cat\_classAssure\_16-20} + \dots$$
$$+ \beta_{cat\_classEnfant\_26-30} X_{cat\_classEnfant\_26-30} + \beta_{garantiePOP1A} X_{garantiePOP1A} + \dots$$
$$+ \beta_{garantiePOP2E} X_{garantiePOP2E} + \beta_{REGIONARA} X_{REGIONARA} + \dots$$
$$+ \beta_{REGIONPDL} X_{REGIONPDL} + offset(log(EXPOSITION)))$$

Dans le tableau 18, nous présentons certains résultats du modèle. Nous pouvons dire que les modèles sont relativement de mauvaise qualité d'ajustement, car $\chi^2 < D$.

Modèle	Deviance	χ^2	AIC
Frais réels	548 504	$\chi^2_{(0.95,37040)} = 37489,83$	449 125,5
Remboursement mutuelle	412 159	$\chi^2_{(0.95,37040)} = 37489,83$	401 345,3

TABLE 18 – Mesures de performance du modèle GLM Tweedie tous postes confondus du portefeuille 2.

Le tableau 19 représente les résultats de nos modèles pour le second portefeuille. Nous pouvons y retrouver notamment les coefficients estimés, ainsi que la significativité des modalités des variables :

Variables	Modèle frais réels			Modèle remboursement mutuelle		
	Coefficient estimé	*pvalue*	*Significativité*	*Coefficient estimé*	*pvalue*	*Significativité*
Intercept	7.09473	<2e-16	***	6.435640	< 2e-16	***
SEXE_BENEFHomme	-0.25560	<2e-16	***	-0.240	<2e-16	***
cat_classAssure_16-20	-0.216	0.468		-0.611	0.051	
cat_classAssure_21-25	-0.274	0.001	***	-0.464	1.16e-09	***
cat_classAssure_26-30	0.133	0.022	*	-0.093	0.106	
cat_classAssure_31-35	0.223	2.36e-06	***	0.021	0.657	
cat_classAssure_36-40	0.113	0.007	**	-0.098	0.016	*
cat_classAssure_41-45	0.225	1.59e-08	***	0.045	0.238	
cat_classAssure_46-50	0.340	< 2e-16	***	0.212	2.54e-09	***
cat_classAssure_51-55	0.413	< 2e-16	***	0.321	< 2e-16	***
cat_classAssure_56-60	0.499	< 2e-16	***	0.341	< 2e-16	***
cat_classAssure_61-65	0.705	< 2e-16	***	0.535	< 2e-16	***
cat_classAssure_+65	0.818	2.05e-10	***	0.696	2.02e-08	***
cat_classConjoint_16-20	-0.625	0.635		-0.329	0.781	
cat_classConjoint_21-25	0.194	0.590		-0.014	0.970	
cat_classConjoint_26-30	0.368	0.002	**	0.174	0.140	
cat_classConjoint_31-35	0.190	0.009	**	0.041	0.571	**
cat_classConjoint_36-40	0.099	0.090	.	-0.136	0.017	*
cat_classConjoint_41-45	0.013	0.83		-0.150	0.009	**
cat_classConjoint_46-50	0.276	2.32e-073	***	0.157	0.002	**
cat_classConjoint_51-55	0.435	6.31e-13	***	0.329	1.42e-08	***
cat_classConjoint_56-60	0.508	< 2e-16	***	0.375	5.01e-12	***
cat_classConjoint_+65	0.65	< 2e-16	***	0.47	3.61e-10	***
cat_classConjoint_61-65	0.506	< 2e-16	***	0.334	3.61e-10	***
cat_classEnfant_0-5	-0.396	< 2e-16	***	-0.697	< 2e-16	***
cat_classEnfant_6-10	-0.478	< 2e-16	***	-0.596	< 2e-16	***
cat_classEnfant_16-20	-0.194	< 2e-16	***	-0.370	< 2e-16	***
cat_classEnfant_21-25	-0.491	< 2e-16	***	-0.692	6.46e-10	***
cat_classEnfant_26-30	-0.655	9.21e-08	***	-0.727	9.21e-08	***
GARANTIEPOP1A	-0.600	< 2e-16	***	-0.970	< 2e-16	***
GARANTIEPOP1B	-0.284	< 2e-16	***	-0.416	< 2e-16	***
GARANTIEPOP2A	-0.939	< 2e-16	***	-1.346	< 2e-16	***
GARANTIEPOP2B	-0.312	< 2e-16	***	-0.647	< 2e-16	***
GARANTIEPOP2C	-0.058	0.063	***	-0.225	2.60e-13	***
GARANTIEPOP2D	0.113	0.002	**	0.007	0.841	
GARANTIEPOP2E	0.6648	8.38e-07	***	0.755	3.52e-09	***
REGIONARA	-0.122	0.0001	***	-0.064	0.056	.
REGIONBFC	-0.258	0.0002	***	-0.175	0.009	**
REGIONBR	-0.228	2.42e-11	***	-0.157	2.26e-06	***
REGIONC	-0.080	0.301		-0.081	0.278	
REGIONCVL	-0.238	0.001	**	-0.198	0.007	**
REGIONGE	-0.121	0.001	**	-0.094	0.010	*
REGIONHF	-0.078	0.001	.	-0.025	0.580	
REGIONN	-0.239	0.009	**	0.020	0.456	
REGIONNA	-0.049	0.076	.	-0.002	0.949	.
REGIONO	-0.061	0.015	*	-0.061	0.015	*
REGIONOUTRE	-0.236	1.05e-09	***	-0.221	4.87e-09	***
REGIONPACA	0.039	0.144		0.039	0.144	
REGIONPDL	-0.186	0.002	**	-0.129	0.144	

TABLE 19 – Résultats du modèle GLM Tweedie tous postes confondus du portefeuille 2.

D'après le tableau ci-dessus des coefficients estimés, nous remarquons que notre modèle a beaucoup de modalités de variables significatives. De plus, nous pouvons effectuer les mêmes remarques que pour le portefeuille au niveau du sexe. Les mêmes remarques

peuvent être faites sur l'effet de l'âge chez les assurés et les conjoints. Pour la catégorie de bénéficiaire enfant, nous en venons aux mêmes conclusions que le portefeuille précédent.Toute chose égale par ailleurs en moyenne la modalité enfant entre 11-15 ans un peu plus que les autres chez les enfants. En ce qui concerne les régions, toute chose égale par ailleurs en moyenne les régions Île de France et PACA ont tendance à avoir plus de dépenses. Finalement, si nous nous intéressons au niveau de garantie nous pouvons observer une augmentation en moyenne de la consommation en santé plus nous allons dans les niveaux de garantie élevés.

Dans les graphiques ci-dessous sont représentés les résidus des modèles. La grande majorité des points pour les résidus de Pearson se situent vers 0. Quelques points sont autour de 40 et 100. Nous pouvons dire que les modèles sont satisfaisants et nous les validons.

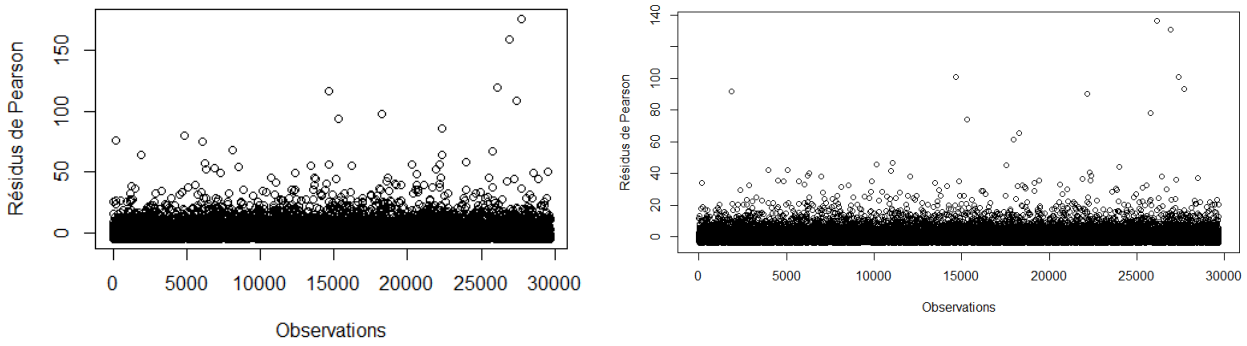

FIGURE 79 – Résidus de Deviance du modèle de régression Tweedie (frais réels et montant remboursé par la mutuelle) tous postes confondus du portefeuille 2.

Le tableau 20 représente la comparaison des moyennes entre les échantillons modélisés et la réalité. Les moyennes entre les deux sont assez proches avec une dépense moyenne autour de 466,05€ prédit pour le remboursement mutuelle contre 473,45 € en réalité. Puis, une prime pure moyenne annuelle de 1 062,66€ prédite en frais réels avec ce modèle contre 1 073,17€ en réalité.

Prime pure	**Observé**	**Prédiction**
Frais réels	1073,17€	1062,66€
Remboursement mutuelle	473,45€	466,05€

TABLE 20 – Résumé des moyennes de prédiction du modèle Tweedie tous postes confondus du portefeuille 2.

Le tableau 21 représente les mesures de performance obtenues pour nos modèles. Nous avons une plutôt mauvaise qualité d'ajustement dans nos modèles avec $D > \chi^2$. Néanmoins, ils ont tout de même une bonne qualité de prédiction, puisque les mesures MAE ont des valeurs plutôt faibles. Par contre, nous pouvons noter que nos modèles ont une forte volatilité avec des mesures de performance MSE et RMSE plutôt élevées.

Modèle	*MAE*	*RMSE*	*MSE*	*AIC*	*Deviance*	χ^2
Frais réels	759,6	1 264,7	1 599 444	449 125,5	548 504	37 489,8
Montants de la mutuelle	341,7	576,1	331 882,5	401 283,1	412 461	37 489,8

TABLE 21 – Mesures de précision du modèle GLM Tweedie tous postes confondus du portefeuille 2.

Finalement, à l'aide des ces premiers modèles au niveau global, nous pouvons remarquer qu'ils sont de mauvaise qualité d'ajustement. Donc, il sera intéressant de pouvoir observer à des niveaux plus fins les dépenses en santé. Ainsi, notre deuxième méthode se verra utile afin de modéliser par sous-poste la prime pure annuelle.

4.2.2 Modélisation GLM Fréquence - Coût moyen

Dans cette partie, nous allons nous intéresser à la modélisation fréquence - coût moyen par sous-poste. Certains sous-postes ne représentent que très peu de volumétrie en terme de coût (inférieur à 3% de la charge totale), nous les modéliserons avec le modèle GLM avec la loi Tweedie du poste en appliquant un coefficient. Puis, pour les sous-postes qui représentent plus de 3% de la charge totale, nous avons jugé qu'il y avait suffisamment de données pour les étudier en détail avec un modèle GLM fréquence - coût moyen, afin d'affiner la modélisation.

Pour une raison de clarté et de compréhension, nous présenterons dans cette partie seulement les résultats du poste Optique du portefeuille 2 pour ne pas perdre le lecteur. Il faut noter qu'il a été fait un modèle Tweedie pour chaque poste (dentaire, optique, hospitalisation, etc.), pour chaque portefeuille et pour les deux montants étudiés (remboursement mutuelle et frais réels). Mais aussi, nous avons effectué un modèle fréquence - coût sur tous les sous-postes suffisamment représentatifs, pour chaque portefeuille et chaque montant. Vous trouverez en annexe pour chaque portefeuille la liste exhaustive des sous-postes et la modélisation que nous avons choisie pour chacun d'entre eux.

Tout d'abord, nous allons présenter les résultats obtenus par le modèle GLM avec la loi Tweedie pour le poste Optique. Nous estimons la valeur optimale pour le paramètre *var.power*. Nous avons trouvé 1,2 pour les frais réels et les montants remboursés par la mutuelle. Ensuite, nous réalisons un modèle avec les variables explicatives suivantes *SEXE*, *cat_class*, *REGION* et *GARANTIE*, puisqu'il s'agissait des variables les plus significatives. De plus, la variable d'exposition est en *offset*. L'année de couverture n'est pas prise en compte pour nos modèles. Notre modèle s'écrit de la façon suivante :

$$
\begin{aligned}
Y_{cout} = \beta_0 + \beta_{SexeHomme} X_{SexeHomme} + \beta_{cat\_classEnfant\_age0-5} X_{cat\_classEnfant0-5} + \dots \\
+ \beta_{cat\_classAssure+65} X_{cat\_classAssure+65} + \beta_{garantiePOP1A} X_{garantiePOP1A} + \dots \\
+ \beta_{garantiePOP2E} X_{garantiePOP1E} + \beta_{REGIONARA} X_{REGIONARA} + \dots \\
+ \beta_{REGIONPDL} X_{REGIONPDL} + offset(log(EXPOSITION))
\end{aligned}
$$

Dans le tableau 18, nous présentons certains résultats du modèle. Nous pouvons dire que les modèles sont relativement de mauvaise qualité d'ajustement, car $\chi^2 < D$.

Modèle	Deviance	χ^2	AIC
Frais réels	261 6245	$\chi^2_{(0.95,37040)} = 37488,82$	124 189.3
Remboursement mutuelle	3 302 865	$\chi^2_{(0.95,37040)} = 37488,82$	138 410.2

TABLE 22 – Mesure de performance du modèle GLM Tweedie du poste Optique du portefeuille 2.

Le tableau 23 représente les résultats de nos modèles. Nous pouvons observer le coefficient estimé et la significativité pour chaque modalité de variable :

Variables	Modèle frais réels			Modèle remboursement mutuelle		
	Coefficient estimé	*pvalue*	*Significativité*	*Coefficient estimé*	*pvalue*	*Significativité*
Intercept	4,63	<2e-16	***	4,31	<2e-16	***
SEXE_BENEFHomme	-0,25	<2e-16	***	-0.,22	<2e-16	***
cat_classAssure_16-20	-0,77	0.35		-0,35	0,63	
cat_classAssure_21-25	0,46	<2,25e-05	***	0,34	0,003	**
cat_classAssure_26-30	0,53	<3,47e-10	***	0,43	1,58e-06	***
cat_classAssure_31-35	0,23	0,002	**	0,13	0,07	.
cat_classAssure_36-40	0,12	0,08	.	0,01	0,88	
cat_classAssure_41-45	0,40	5,62e-12	***	0,35	2,02e-09	***
cat_classAssure_46-50	0,86	<2e-16	***	0,87	<2e-16	***
cat_classAssure_51-55	0,89	<2e-16	***	0,90	<2e-16	***
cat_classAssure_56-60	0,86	<2e-16	*	0,87	<2e-16	***
cat_classAssure_61-65	0,83	<2e-16	***	0,84	<2e-16	***
cat_classAssure_+65	0,74	2,19e-05	***	0,71	2,92e-05	***
cat_classConjoint_16-20	1,28	0,32		1,64	0,14	
cat_classConjoint_21-25	0,18	0,75		0,08	0,89	
cat_classConjoint_26-30	0,47	0,01	**	0,34	0,05	.
cat_classConjoint_31-35	0,06	0,59		0,02	088	
cat_classConjoint_36-40	-0,06	0,55		-0,09	0,31	
cat_classConjoint_41-45	0,44	2,62e-08	***	0,44	1,54e-08	***
cat_classConjoint_46-50	0,77	<2e-16	***	0,81	<2e-16	***
cat_classConjoint_51-55	0,84	<2e-16	***	0,86	<2e-16	***
cat_classConjoint_56-60	0,76	<2e-16	***	0,76	<2e-16	***
cat_classConjoint_61-65	0,67	<2e-16		0,72	<2e-16	***
cat_classConjoint_+65	0,60	5,74e-16	***	0,65	<2e-16	***
cat_classEnfant_0-5	-1,33	<2e-16	***	-1,24	<2e-16	***
cat_classEnfant_6-10	-0,24	<7,40e-05	***	-0,17	0,003	**
cat_classEnfant_16-20	0,02	0,77		-0,05	0,34	
cat_classEnfant_21-25	-0,01	0,93		0,01	0,88	
cat_classEnfant_26-30	-0,15	0,41		-0,36	0,07	.
GARANTIEPOP1A	-0,96	<2e-16	***	-1,34	<2e-16	***
GARANTIEPOP1B	-0,44	<2e-16	***	-0,58	<2e-16	***
GARANTIEPOP2A	-1.20	<2e-16	***	-1,5	<2e-16	***
GARANTIEPOP2B	-0,64	<2e-16	***	-1,02	<2e-16	***
GARANTIEPOP2C	-0,33	1,13e-12	***	-0,55	<2e-16	***
GARANTIEPOP2D	-0,18	0,001	***	-0,32	1,88e-09	***
GARANTIEPOP2E	0,34	0,04	*	0,40	0,01	*
REGIONARA	-0,07	0,17		0,02	0,61	
REGIONBFC	-0,10	0,3		-0,02	0,83	
REGIONBR	0,07	0,16		0,14	0,002	**
REGIONC	-0,19	0,09	.	-0,18	0,1	.
REGIONCVL	-0,14	0,17		-0,05	0,62	
REGIONGE	0,01	0,91		0,09	0,08	.
REGIONHF	-0,06	0,36		0,02	0,74	
REGIONN	-0,09	0,50		0,08	0,54	
REGIONNA	-0,005	0,90		0,06	0,11	
REGIONO	0,04	0,29		0,11	0,002	**
REGIONOUTRE	-0,07	0,29		-0,006	0,90	
REGIONPACA	-0,12	0,002	**	-0,08	0,04	*
REGIONPDL	-0,03	0,73		0,03	0,72	

TABLE 23 – Résultats du modèle GLM Tweedie du poste Optique du portefeuille 2.

D'après le tableau 23, nous pouvons en déduire que la variable région n'est pas significative pour le modèle Tweedie en Optique. Par contre, nous pouvons remarquer qu'en moyenne plus le niveau de garantie augmente et plus les assurés ont tendance à plus dépenser. Pour l'âge, nous pouvons faire les mêmes constats qu'au niveau global avec en moyenne une augmentation des dépenses avec l'âge pour l'optique pour toutes les catégories de bénéficiaire. Finalement, toute chose égale par ailleurs en moyenne les hommes ont tendance à un peu moins consommer en optique que les femmes dans ce portefeuille. Car, le coefficient estimé des hommes est négatif et que la modalité de référence est le sexe féminin.

Dans les graphiques ci-dessous sont représentés les résidus des modèles pour chaque montant. Les points des résidus de Pearson et de la Deviance sont plutôt centrés autour de 0. Nous pouvons dire que les modèles sont satisfaisants et nous les validons.

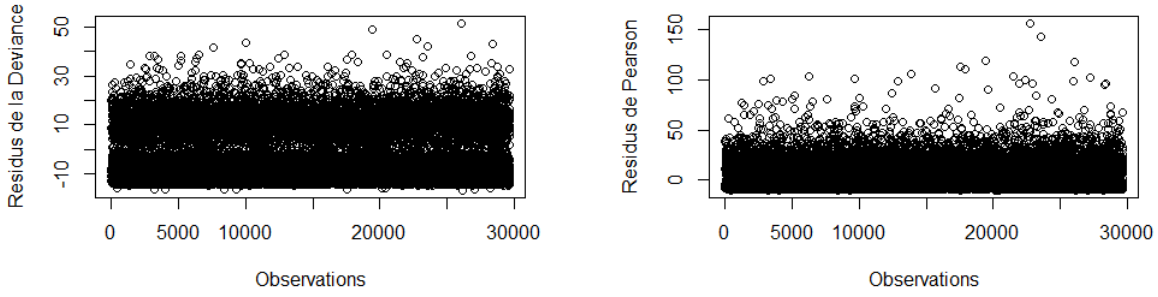

FIGURE 80 – Résidus des modèles de régression Tweedie (frais réels) pour le poste optique du portefeuille 2.

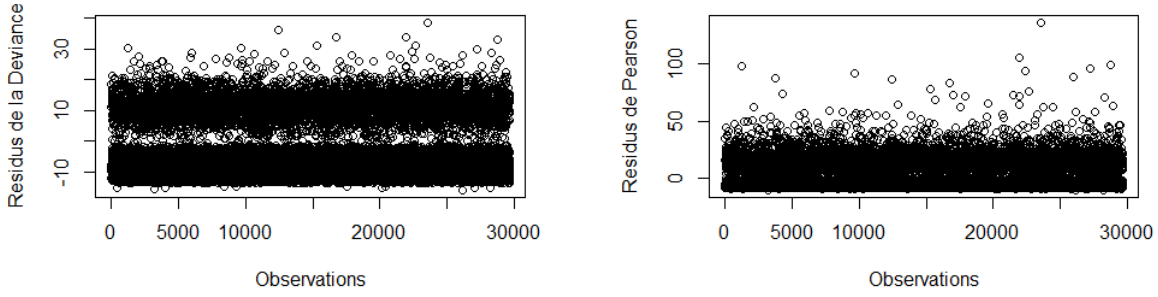

FIGURE 81 – Résidus du modèle de régression Tweedie (remboursement mutuelle) pour le poste optique du portefeuille 2.

Le tableau 24 ci-dessous résume les informations sur les moyennes de l'échantillon de validation. Ainsi, en moyenne notre modèle a prédit sur notre échantillon de validation qu'un

bénéficiaire dépensera sur une année autour d'une centaine d'euros en frais réels et autour de soixante-quatorze euros en remboursement mutuelle.

Prime pure	**Moyenne observée**	**Moyenne prédite**
Frais réels	101,06 €	102,22 €
Remboursement mutuelle	73,44 €	74,29 €

TABLE 24 – Résumé des moyennes de prédictions du modèle GLM Tweedie du poste optique pour le portefeuille 2.

Le tableau 25 suivant représente les différentes mesures de performance obtenues pour notre modèle GLM Tweedie sur le poste Optique du portefeuille 2. Nous pouvons dire que le modèle est plutôt bon, puisque les valeurs des mesures MAE, RMSE et MSE sont peu élevées. Nous pouvons noter que nos modèles sont de bonne qualité, puisque les valeurs de MAE sont faibles. Par ailleurs, nos modèles ont une faible volatilité au vu des valeurs des mesures MSE et RMSE qui sont peu élevées.

Modèle	*MAE*	*RMSE*	*MSE*	*AIC*
Frais réels	140	195	38 017	138 410
Montants de la mutuelle	105	140	19 527	124189

TABLE 25 – Mesures de précision pour les modèles Tweedie pour le poste optique du portefeuille 2.

Après avoir réalisé les modèles par poste, nous leur appliquons un coefficient pour qu'ils puissent représenter un sous-poste. Nous effectuons cette démarche sur les sous-postes les moins volimuneux en terme de coût, c'est-à-dire représentant moins de 3% de la charge totale (en frais réels).

Pour les autres sous-postes, nous utilisons des GLM avec une méthode coût-fréquence. Finalement, pour l'étude du poste optique, nous avons étudié trois sous-postes : les lentilles, les montures et les verres. Le sous-poste des lentilles ne représentant que 0,9% de la charge totale en frais réels et 0,3% de la charge total en dépenses mutuelle, nous avons décidé de simplifier la modélisation de ce sous-poste avec la méthode décrite précédemment.

Autrement dit, lorsque nous voudrons prédire ce sous-poste, nous utiliserons le modèle Tweedie du poste Optique présenté ci-dessus. Puis, nous appliquerons un coefficient 0,9% pour les frais réels et 0,3% pour les remboursements mutuelles aux prédictions effectuées par nos modèles.

Désormais, nous allons nous occuper des sous-postes montures et verres. Nous retrouvons les sous-postes concernant les montures et les verres. Dans ces cas, la prime pure sera modélisée avec des GLM en utilisant la méthode fréquence-coût moyen.

Avant d'effectuer le modèle linéaire généralisé, il faut déterminer la distribution des variables d'intérêt qui est le coût moyen et la fréquence. Dans notre étude, nous modéliserons qu'un seul modèle pour la fréquence, puisqu'elle ne change pas entre les montants en frais réels et les montants remboursés par la mutuelle. Nous avons comparé les distributions des lois Gamma et log-normale pour le coût moyen, puis les loi Poisson et Binomiale Négatove pour la fréquence. Après réflexion, nous avons choisi de prendre les loi Gamma et binomiale négative (il s'agit des mêmes lois pour tous les sous-postes dans les deux portefeuilles).

- **Les montures :**

  Tout d'abord, nous présentons les résultats pour les modèles du coût moyen (en frais réels et dépenses mutuelle). Vous trouverez ci-dessous les représentations des distributions des lois pour le coût moyen en frais réels et en remboursement mutuelle. Nous pouvons voir que les dépenses en frais réels et remboursement mutuelle au niveau du coût moyen ont plutôt la même distribution.

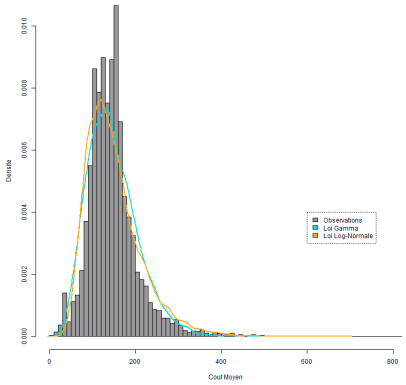

FIGURE 82 – Graphiques des fonctions de densité des lois pour le coût moyen en frais réels du sous-poste monture pour le portefeuille 2.

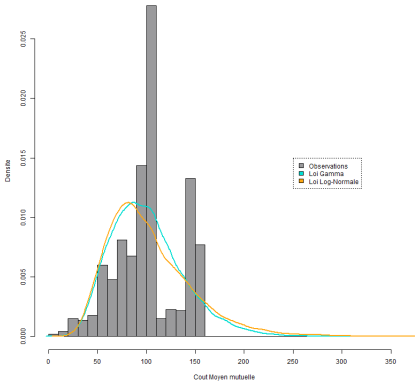

FIGURE 83 – Graphiques des fonctions de densité des lois pour le coût moyen du remboursement mutuelle du sous-poste monture pour le portefeuille 2.

Vous trouverez ci-dessous le graphique des fonctions de répartition des lois pour la fréquence.

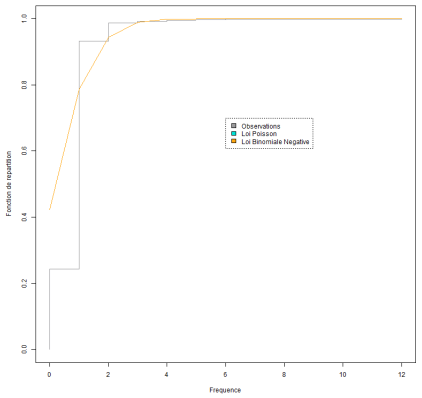

FIGURE 84 – Graphique des fonctions de répartition des lois pour la fréquence du sous-poste monture pour le portefeuille 2.

Une fois les lois choisies, nous réalisons trois modèles. Un premier sur le coût moyen en frais réels. Il s'écrira de la façon suivante :

$$Y_{\text{cout moyen}} = \beta_0 + \beta_{SexeHomme} X_{SexeHomme} + \beta_{cat\_classEnfant\_0-5} X_{cat\_classEnfant\_0-5} + \dots$$
$$+ \beta_{cat\_classAssure\_+65} X_{cat\_classAssure\_+65} + \beta_{garantiePOP1A} X_{garantiePOP1A} + \dots$$
$$+ \beta_{garantiePOP2E} X_{garantiePOP2E} + \beta_{REGIONARA} X_{REGIONARA} + \dots$$
$$+ \beta_{REGIONPDL} X_{REGIONPDL} + offset(log(EXPOSITION))$$

Nous avons représenté dans le tableau 26 les résultats obtenus pour le coût moyen en frais réels et en remboursement mutuelle. Nous pouvons remarquer que toute chose égale par ailleurs en moyenne les hommes et les femmes dépensent à peu près la même chose en terme de monture de lunette dans ce portefeuille, puisque leurs coefficients estimés sont très proche.

Nous pouvons aussi observer qu'en moyenne avec l'âge, les bénéficiaires ont tendance à dépenser plus. Puis, à partir de 60 ans les bénéficiaires de type assuré ont une tendance à la baisse. En ce qui concerne les régions, nous pouvons remarquer des différences avec des dépenses (en frais réels) plus élevées que la moyenne en outre mer et Pays de la Loire.

Finalement, si nous nous concentrons sur le niveau de garantie, alors nous pouvons effectuer les mêmes remarques qu'à un niveau global. Ce modèle nous permettra néanmoins d'être plus précis au niveau des dépenses en coût moyen.

Variables	Modèle frais réels			Modèle remboursement mutuelle		
	Coefficient estimé	*pvalue*	*Significativité*	*Coefficient estimé*	*pvalue*	*Significativité*
(Intercept)	4,899	<2e-16	***	4,265	<2e-16	***
SEXEHomme	-0,022	0,03	*	0,007	0,3	
cat_classAssure_16-20	-0,307	0,44		-0,011	0,95	
cat_classAssure_21-25	0,035	0,72		-0,031	0,628	
cat_classAssure_26-30	0,025	0,8		-0,037	0,52	
cat_classAssure_31-35	0,072	0,44		0,018	0,75	
cat_classAssure_36-40	0,114	0,21		-0,012	0,84	
cat_classAssure_41-45	0,0581	0,52		-0,032	0,56	
cat_classAssure_46-50	0,038	0,67		-0,039	0,48	
cat_classAssure_51-55	0,046	0,61		-0,045	0,41	
cat_classAssure_56-60	0,054	0,549		-0,05	0,4	
cat_classAssure_61-65	-0,0003	0,99		-0,05	0,37	
cat_classConjoint_+65	-0,008	0,93		-0,073	0,2	
cat_classConjoint_16-20	-0,373	0,347		0,0007	0,99	
cat_classConjoint_21-25	-0,501	0,08	.	-0,06	0,65	
cat_classConjoint_26-30	0,107	0,34		-0,04	0,58	
cat_classConjoint_31-35	0,037	0,71		0,017	0,78	
cat_classConjoint_36-40	0,079	0,41		0,015	0,8	
cat_classConjoint_41-45	-0,041	0,66		-0,053	0,36	
cat_classConjoint_46-50	-0,035	0,7		-0,057	0,32	
cat_classConjoint_51-55	-0,015	0,87		-0,087	0,13	
cat_classConjoint_56-60	0,017	0,85		-0,009	0,88	
cat_classConjoint_61-65	0,003	0,98		-0,019	0,74	
cat_classEnfant_0-5	-0,276	0,004	**	-0,289	6,79E-07	***
cat_classEnfant_6-10	-0,282	0,002	**	-0,313	1,76E-08	***
cat_classEnfant_11-15	-0,174	0,05	.	-0,287	2,11E-07	***
cat_classEnfant_16-20	-0,071	0,44		-0,156	0,005	**
cat_classEnfant_21-25	-0,037	0,7		-0,025	0,66	
cat_classEnfant_26-30	0,18	0,15		-0,052	0,5	
REGIONBFC	0,042	0,35		0,017	0,56	
REGIONBR	0,073	0,004	**	0,029	0,08	.
REGIONC	0,061	0,23		-0,003	0,92	
REGIONCVL	0,075	0,14		0,031	0,32	
REGIONGE	0,054	0,05	*	0,021	0,24	
REGIONHF	0,02	0,55		0,006	0,78	
REGIONIDF	0,073	0,0005	***	0,012	0,39	
REGIONN	0,015	0,8		0,031	0,41	
REGIONNA	0,04	0,08	.	0,012	0,42	
REGIONO	0,055	0,011	*	0,011	0,42	
REGIONOUTRE	0,178	2,37E-10	***	0,032	0,08	.
REGIONPACA	0,019	0,41		-0,003	0,85	
REGIONPDL	0,111	0,004	**	0,009	0,7	
GARANTIEPOP1B	0,046	0,2		0,32	<2e-16	***
GARANTIEPOP1C	0,117	8,70E-05	***	0,555	<2e-16	***
GARANTIEPOP2A	-0,206	0,0004	***	-0,28	2,75E-14	***
GARANTIEPOP2B	-0,133	0,0003	***	-0,229	<2e-16	***
GARANTIEPOP2C	-0,054	0,116		0,02	0,36	
GARANTIEPOP2D	0,027	0,45		0,244	<2e-16	***
GARANTIEPOP2E	0,185	0,03	*	0,605	<2e-16	

TABLE 26 – Résultats des modèles coût moyen pour les montures en frais réels et en remboursement mutuelle du portefeuille 2.

Les figures 85 et 86 ci-dessous représentent les résidus de Déviance de notre modèle. Nous pouvons dire que notre modèle est bon, puisqu'ils sont plutôt homogènes et centrés entre -2 et 2.

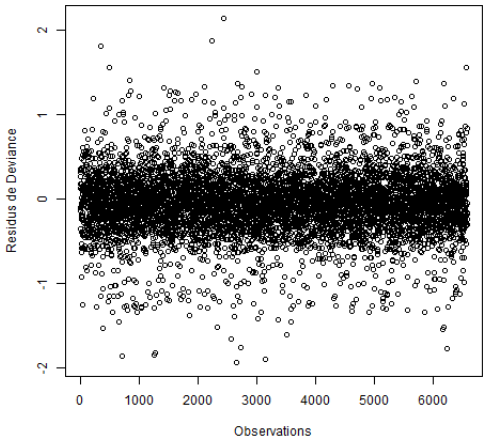

FIGURE 85 – Résidus de déviance du modèle coût moyen en frais réels pour les montures du portefeuille 2.

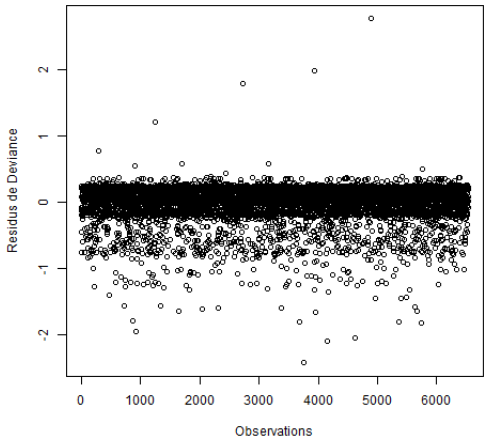

FIGURE 86 – Résidus de déviance du modèle coût moyen du remboursement mutuelle pour le sous-poste monture du portefeuille 2.

Enfin, nous modélisons la fréquence avec notre modèle qui s'écrira de la manière suivante :

$$Y_{\text{frequence}} = \beta_0 + \beta_{SexeHomme} X_{SexeHomme} + \beta_{cat\_classEnfant\_0-5} X_{cat\_classEnfant\_0-5} + \dots$$
$$+ \beta_{cat\_classAssure\_+65} X_{cat\_classAssure\_+65} + \beta_{garantiePOP1A} X_{garantiePOP1A} + \dots$$
$$+ \beta_{garantiePOP2E} X_{garantiePOP2E} + \beta_{REGIONARA} X_{REGIONARA} + \dots$$
$$+ \beta_{REGIONPDL} X_{REGIONPDL} + offset(log(EXPOSITION))$$

Vous trouverez dans le tableau 27 ci-dessous les coefficients estimés de notre modèle GLM de fréquence avec la loi Binomiale Négative. D'après les résultats, nous pouvons en déduire qu'en moyenne, les femmes ont tendance à changer plus souvent leur monture de lunette que les hommes.

De plus, toute chose égale par ailleurs en moyenne, les assurés et les conjoints ont tendance à changer plus souvent leur monture avec l'âge jusqu'à 60 ans. Ensuite, nous pouvons observer une tendance à la baisse. En moyenne, les régions dans lesquelles il y a plus de changement de montures sont : Bretagne, Occitanie, ou encore, en Normandie. Pour conclure, nous pouvons observer une augmentation de la fréquence avec le niveau de garantie.

Variables	*Coefficient estimé*	*pvalue*	*Significativité*
(Intercept)	-1,707	1,99e-14	***
SEXEHomme	-0,23	<2e-16	***
cat_classAssure_16-20	-0,507	0,44	
cat_classAssure_21-25	0,091	0,7	
cat_classAssure_26-30	0,083	0,7	
cat_classAssure_31-35	-0,156	0,5	
cat_classAssure_36-40	-0,284	0,2	
cat_classAssure_41-45	-0,039	0,9	
cat_classAssure_46-50	0,325	0,12	
cat_classAssure_51-55	0,282	0,2	
cat_classAssure_56-60	0,267	0,2	
cat_classAssure_61-65	0,184	0,4	
cat_classConjoint_+65	0,01	0,96	
cat_classConjoint_16-20	1,32	0,25	
cat_classConjoint_21-25	0,542	0,3	
cat_classConjoint_26-30	0,09	0,7	
cat_classConjoint_31-35	-0,327	0,2	
cat_classConjoint_36-40	-0,3	0,2	
cat_classConjoint_41-45	0,044	0,8	
cat_classConjoint_46-50	0,308	0,2	
cat_classConjoint_51-55	0,294	0,2	
cat_classConjoint_56-60	0,185	0,4	
cat_classConjoint_61-65	0,028	0,9	
cat_classEnfant_0-5	-1,166	1,06e-07	***
cat_classEnfant_6-10	-0,131	0,5	
cat_classEnfant_11-15	-0,042	0,8	
cat_classEnfant_16-20	-0,179	0,4	
cat_classEnfant_21-25	-0,343	0,11	
cat_classEnfant_26-30	-0,454	0,1	
REGIONBFC	-0,113	0,3	
REGIONBR	0,126	0,04	*
REGIONC	-0.139	0,3	
REGIONCVL	0,01	0,9	
REGIONGE	0,072	0,3	
REGIONHF	0,008	0,9	
REGIONIDF	-0,01	0,8	
REGIONN	0,095	0,5	
REGIONNA	0,062	0,3	
REGIONO	0,098	0,1	.
REGIONOUTRE	-0,122	0,1	.
REGIONPACA	-0,087	0,12	
REGIONPDL	-0,006	0,9	
GARANTIEPOP1B	0,286	6,40e-05	***
GARANTIEPOP1C	0,524	1,46e-14	***
GARANTIEPOP2A	0,11	0,4	
GARANTIEPOP2B	0,254	0,003	**
GARANTIEPOP2C	0,448	1,55e-08	***
GARANTIEPOP2D	0,518	1,06e-09	***
GARANTIEPOP2E	0.789	0.0001	***

TABLE 27 – Résultats du modèle GLM pour la fréquence du sous-poste monture du portefeuille 2.

D'après ce tableau 28, nous observons une bonne performance des modèles au niveau de la prédiction avec des mesures relativement basses. De plus, nous remarquons qu'ils ont une bonne qualité d'ajustement. Nous pouvons aussi ajouter que nos modèles ont une bonne qualité de prédiction avec des valeurs de MAE qui sont faibles. Et nous avons une faible volatilité de nos résultats étant donné la valeur de la mesure MSE et RMSE.

Modèle	*MAE*	*RMSE*	*MSE*	*AIC*	*Deviance*	χ^2
Coût moyen (frais réels)	38,3	54,5	2 972,8	70 764	949,5	8 374,3
Coût moyen (mutuelle)	17,5	23,5	552,3	61 660	529	8 327,7
Fréquence	0,36	0,6	0,35	34 444	15 794	37 488,8

TABLE 28 – Mesures de précision des modèles GLM coût moyen et fréquence du sous-poste monture du portefeuille 2.

Dans le tableau 29, nous pouvons voir qu'en moyenne dans notre échantillon de validation une monture de lunette en frais réels coûterait autour de 147 € en prédiction et 145 € en réalité. Tandis qu'en moyenne, une monture de lunette couterait autour de 100 € à la mutuelle. Pour le modèle de fréquence, nous pouvons dire qu'en moyenne les bénéficiaires changent leur monture tous les 4 ans.

Modèle	**Moyenne observée**	**Moyenne prédite**
Coût moyen (frais réels)	144,6 €	147,24 €
Coût moyen (remboursement mutuelle)	100,72 €	100,76 €
Fréquence	0,25	0,23

TABLE 29 – Résumé des moyennes de prédiction des modèles coût moyen et fréquence du sous-poste monture du portefeuille 2.

- **Les verres :**

Désormais, nous nous intéressons aux résultats obtenus sur le sous-poste des verres. Nous avons utilisé la même méthode que pour le sous-poste des montures. Ainsi, vous trouverez ci-dessous les graphiques des distributions des lois pour le coût moyen en frais réels et pour le montant remboursé par la mutuelle. Nous pouvons remarquer que pour les deux montants les distributions ont la même allure. Ainsi, nous avons choisi la loi Gamma pour les deux types de montant.

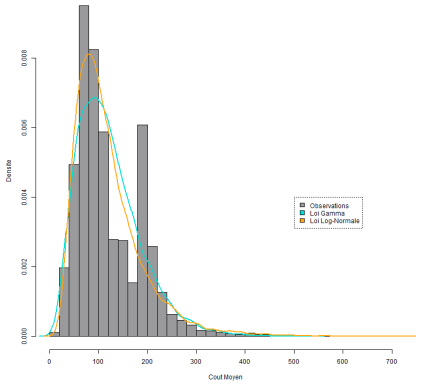

FIGURE 87 – Graphiques des fonctions de densité des lois pour le coût moyen en frais réels du sous-poste verre du portefeuille 2.

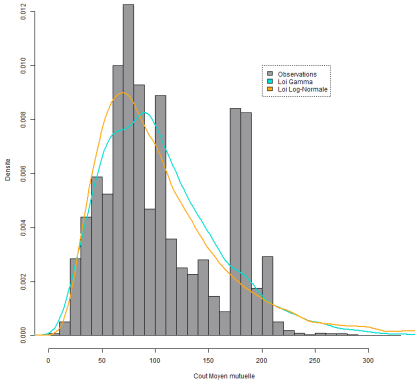

FIGURE 88 – Graphiques des fonctions de densité des lois pour le coût moyen du remboursement mutuelle du sous-poste verre du portefeuille 2.

Vous trouverez ci-dessous le graphique des fonctions de répartition des lois pour la fréquence. Nous pouvons voir que les deux lois suivent plutôt bien la fonction de répartition des observations. Nous avons choisi de prendre la loi Binomiale Négative.

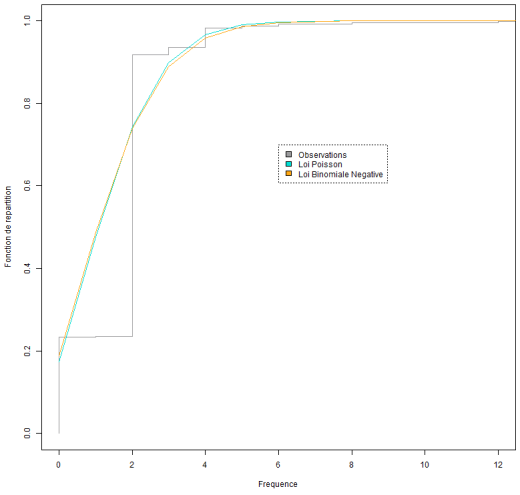

FIGURE 89 – Graphique des fonctions de répartition des lois pour la fréquence du sous-poste verre du portefeuille 2.

Une fois les lois choisies, nous réalisons trois modèles. Un premier sur le coût moyen en frais réels et un second sur le coût moyen en montant remboursé par la mutuelle. Dans ces modèles nous utilisons les variables de sexe, de classe d'âge, de région et de niveau de garantie de l'assuré. Nos modèles pour le coût moyen s'écriront comme :

$$Y_{\text{cout moyen frais réels}} = \beta_0 + \beta_{class\_age0-5} X_{class\_age0-5} + \dots$$
$$+\beta_{class\_age+65} X_{class\_age+65} + \beta_{garantiePOP1A} X_{garantiePOP1A} + \dots$$
$$+\beta_{garantiePOP2E} X_{garantiePOP2E} + \beta_{REGIONARA} X_{REGIONARA} + \dots$$
$$+\beta_{REGIONPDL} X_{REGIONPDL} + offset(log(EXPOSITION))$$

$$Y_{\text{cout moyen mutuelle}} = \beta_0 + \beta_{SEXEHomme} X_{SEXEHomme} + \beta_{class\_age0-5} X_{class\_age0-5} + \dots$$
$$+\beta_{class\_age+65} X_{class\_age+65} + \beta_{garantiePOP1A} X_{garantiePOP1A} + \dots$$
$$+\beta_{garantiePOP2E} X_{garantiePOP2E} + \beta_{REGIONARA} X_{REGIONARA} + \dots$$
$$+\beta_{REGIONPDL} X_{REGIONPDL} + offset(log(EXPOSITION))$$

D'après les résultats de nos modèles contenus dans le tableau ci-dessous, si nous nous concentrons sur le niveau de garantie, alors nous pouvons effectuer les mêmes remarques qu'à un niveau plus global. Ce modèle nous permettra néanmoins d'être plus précis au niveau des dépenses en coût moyen. En observant les résultats, nous pouvons noter que la variable classe d'âge est très significative dans nos deux modèles. Toute chose égale par ailleurs en moyenne avec l'âge, les dépenses des verres ont tendance à augmenter. La variable qui représente la région n'est pas très significative avec quelques modalités qui ressortent comme la Bretagne, les départements outre mer et les pays de la loire.

Variables	Modèle frais réels			Modèle remboursement mutuelle		
	Coefficient estimé	*pvalue*	*Significativité*	*Coefficient estimé*	*pvalue*	*Significativité*
(Intercept)	4,61	<2e-16	***	4,1	<2e-16	***
SEXEHomme				-0,01	0,5	
class_age0-5	-0,71	<2e-16	***	-0,74	<2e-16	***
class_age0-5	-0,71	<2e-16	***	-0,74	<2e-16	***
class_age6-10	-0,73	<2e-16	***	-0,72	<2e-16***	
class_age11-15	-0,68	<2e-16	***	-0,68	<2e-16	***
class_age16-20	-0,62	<2e-16	***	-0,68	<2e-16	***
class_age21-25	-0,56	<2e-16	***	-0,57	<2e-16	***
class_age26-30	-0,55	<2e-16	***	-0,50	<2e-16	***
class_age31-35	-0,51	<2e-16	***	-0,48	<2e-16	***
class_age36-40	-0,46	<2e-16	***	-0,48	<2e-16	***
class_age41-45	-0,28	<2e-16	***	-0,25	<2e-16	***
class_age51-55	0,1	<2e-16	***	0,08	1,44e-06	***
class_age56-60	0,12	<2e-16	***	0,08	2,23e-07	***
class_age61-65	0,13	<2e-16	***	0,09	4,30e-06	***
class_age+65	0,12	<2e-16	***	0,02	0,001	***
REGIONBFC	-0,01	0,36		-0,02	0,5	
REGIONBR	-0,02	0,95	**	-0,07	0,12	
REGIONC	0,03	0,91		-0,01	0,83	
REGIONCVL	-0,03	0,55		0,01	0,70	
REGIONGE	0,003	0,91	*	0,02	0,51	
REGIONHF	-0,002	0,94		0,01	0,68	
REGIONIDF	0,03	0,16		0,01	0,68	
REGIONN	0,03	0,58		0,003	0,4951	
REGIONNA	-0,02	0,418	.	-0,01	0,47	
REGIONO	0,02	0,41	*	0,01	0,62	
REGIONOUTRE	0,1	0,0004	***	0,04	0,07	.
REGIONPACA	-0,03	0,18		-0,03	0,11	
REGIONPDL	-0,01	0,84	**	-0,01	0,72	
GARANTIEPOP1B	0,22	4,41e-13	***	0,51	<2e-16	***
GARANTIEPOP1C	0,51	<2e-16	***	0,92	<2e-16	***
GARANTIEPOP2A	-0,13	0,03	*	-0,02	0,63	
GARANTIEPOP2B	0,02	0,5	***	0,28	<2e-16	***
GARANTIEPOP2C	0,12	0,0004	***	0,49	<2e-16	***
GARANTIEPOP2D	0,28	1,23e-15	***	0,65	<2e-16	***
GARANTIEPOP2E	0,53	1,50e-10	***	0,98<2e-16	***	

TABLE 30 – Résultats du GLM pour le coût moyen en frais réels et en remboursement mutuelle pour les verres du portefeuille 2.

Les figures ci-dessous représentent les résidus de déviance de nos modèles pour le coût moyen. Nous pouvons dire que nos modèles sont bons, puisqu'ils sont homogènes et centrés entre -2 et 2.

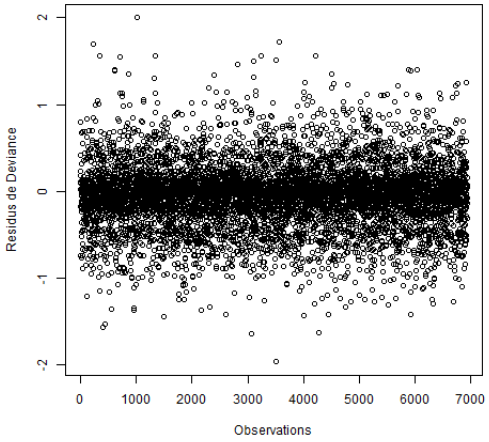

FIGURE 90 – Résidus de déviance du modèle coût moyen en frais réels du sous-poste verre pour le portefeuille 2.

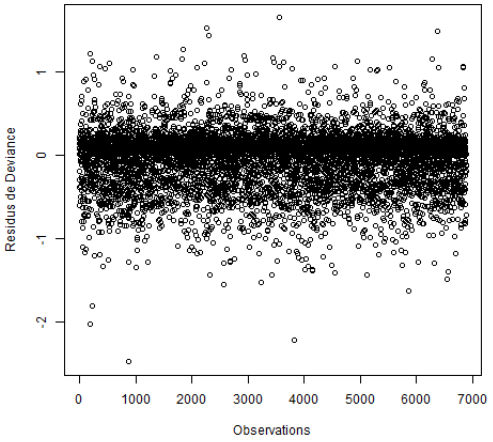

FIGURE 91 – Résidus de déviance du modèle coût moyen du remboursement mutuelle du sous-poste verre pour le portefeuille 2.

Ensuite, nous modélisons la fréquence avec la loi Binomiale Négative et nous pouvons écrire notre modèle de la façon suivante :

$$Y_{\text{frequence}} = \beta_0 + \beta_{SexeHomme} X_{SexeHomme} + \beta_{cclass\_age\_0-5} X_{class\_age\_0-5} + \dots$$
$$+\beta_{class\_age\_+65} X_{class\_age\_+65} + \beta_{garantiePOP1A} X_{garantiePOP1A} + \dots$$
$$+\beta_{garantiePOP2E} X_{garantiePOP2E} + \beta_{REGIONARA} X_{REGIONARA} + \dots$$
$$+\beta_{REGIONPDL} X_{REGIONPDL} + offset(log(EXPOSITION))$$

Vous trouverez dans le tableau 31 les coefficients estimés de notre modèle GLM de fréquence avec la loi Binomiale Négative. A l'aide des résultats sur la fréquence, nous observons qu'en moyenne les hommes changent un peu moins souvent de verres que les femmes. Les variables représentant le croisement de la variable classe d'âge et la catégorie du bénéficiaire et la variable de région ne sont pas très significatives. Tandis que la variable correspondant au niveau de garantie de l'assuré est très significative. Ainsi, toute chose égale par ailleurs en moyenne dans ce portefeuille, plus le niveau de garantie de l'assuré augmente plus les individus auront tendance à changer souvent leurs verres.

Variables	*Coefficient estimé*	*pvalue*	*Significativité*
(Intercept)	-0,97	8,32e-06	***
SEXEHomme	-0,23	<2e-16	***
cat_classAssure_16-20	0,01	0,98	
cat_classAssure_21-25	0,14	0,52	
cat_classAssure_26-30	0,03	0,9	
cat_classAssure_31-35	-0,17	0,42	
cat_classAssure_36-40	-0,30	0,14	
cat_classAssure_41-45	-0,08	0,67	
cat_classAssure_46-50	0,30	0,14	
cat_classAssure_51-55	0,26	0,21	
cat_classAssure_56-60	0,22	0,28	
cat_classAssure_61-65	0,18	0,38	
cat_classConjoint_+65	0,07	0,75	
cat_classConjoint_16-20	1,17	0,34	
cat_classConjoint_21-25	0,11	0,85	
cat_classConjoint_26-30	0,2	0,44	
cat_classConjoint_31-35	-0,3	0,19	
cat_classConjoint_36-40	-0,42	0,05	.
cat_classConjoint_41-45	0,03	0,90	
cat_classConjoint_46-50	0,28	0,23	
cat_classConjoint_51-55	0,29	0,17	
cat_classConjoint_56-60	0,19	0,36	
cat_classConjoint_61-65	0,04	0,86	
cat_classEnfant_0-5	-1,19	2,16e-08	***
cat_classEnfant_6-10	-0,17	0,39	
cat_classEnfant_11-15	-0,04	0,86	
cat_classEnfant_16-20	-0,19	0,35	
cat_classEnfant_21-25	-0,33	0,15	
cat_classEnfant_26-30	-0,59	0,03	*
REGIONBFC	0,04	0,68	
REGIONBR	0,18	0,003	**
REGIONC	-0,008	0,49	
REGIONCVL	0,005	0,97	
REGIONGE	0,09	0,16	
REGIONHF	0,04	0,56	
REGIONIDF	0,008	0,86	
REGIONN	0,13	0,34	
REGIONNA	0,11	0,04	*
REGIONO	0,11	0,03	*
REGIONOUTRE	-0,08	0,25	
REGIONPACA	-0,08	0,13	
REGIONPDL	0,007	0,94	
GARANTIEPOP1B	0,31	5,43e-06	***
GARANTIEPOP1C	0,53	7,70e-16	***
GARANTIEPOP2A-	0,07	0,57	
GARANTIEPOP2B	0,23	0,006	**
GARANTIEPOP2C	0,498	1,48e-10	***
GARANTIEPOP2D	0,59	5,22e-13	***
GARANTIEPOP2E	0,79	0.0001	***

TABLE 31 – Résultats du modèle GLM pour la fréquence du sous-poste verre pour le portefeuille 2.

Le tableau 32 représente les différentes mesures de performance des modèles. D'après celui-ci, nous observons une bonne performance des modèles au niveau de la prédiction avec des mesures MAE, MSE et RMSE qui sont basses. De plus, nous remarquons que les modèles pour les coûts moyens ont une bonne qualité d'ajustement puisque $\chi^2 > D$. Alors que pour le modèle de la fréquence, ça n'est pas le cas.

Modèle	*MAE*	*RMSE*	*MSE*	*AIC*	*Deviance*	χ^2
Coût moyen (frais réels)	32,7	50,83	2 583,8	71 122	1020	8788.5
Coût moyen (mutuelle)	25	34,9	1215	67338	836	8788.5
Fréquence	0,75	1,1	1,3	54 611	27 386	8788.5

TABLE 32 – Mesures de performance des modèles GLM coût-fréquence du sous-poste verre pour le portefeuille 2.

Dans le tableau 33, nous pouvons retrouver les résultats obtenus en prédiction dans l'échantillon de validation pour chacun des modèles avec les résultats réels. Ainsi, en moyenne dans cet échantillon, il a été prédit qu'un verre coûterait environ 123 € et en réalité cela coûte autour de 122 € en frais réels. Puis, en moyenne pour le montant remboursé par la mutuelle, il est autour de 105 € en prédiction et en réalité. Finalement, en moyenne les verres sont changés tous les quatre ans.

Modèle	**Moyenne observée**	**Moyenne prédite**
Coût moyen (frais réels)	122,11 €	121,47 €
Coût moyen (remboursement mutuelle)	105,74 €	105,19 €
Fréquence	0,25	0,25

TABLE 33 – Résumé des moyennes prédiction des modèles GLM coût-fréquence du sous-poste verre pour le portefeuille 2.

4.3 eXtreme gradient boosting

Dans cette partie, nous allons présenter les résultats obtenus avec l'algorithme XGBoost pour nos deux portefeuilles et les deux montants modélisés. Notons que les données utilisées pour les modélisations ne représentent que 95% des sinistres et les 5% restant constituent les graves.

4.3.1 Portefeuille 1

Pour le premier portefeuille, nous avons utilisé deux variables explicatives qui sont l'âge et la variable représentant le croisement de la catégorie de l'assuré (assuré, conjoint, ou enfant) avec la classe d'âge. Dans un premier temps, nous avons tout simplement observé la moyenne des montants prédits et observés. Nous pouvons observer qu'en moyenne, les dépenses annuelles pour un individu se situe autour de 830€ en frais réels et autour de 360€ en montant remboursé par la mutuelle.

Prime pure	***Moyenne observée***	***Moyenne prédite***
Frais réels	835 €	828 €
Remboursement SHAM	359 €	366 €

TABLE 34 – Comparaison des moyennes prédites et réelles pour le portefeuille 1 avec le modèle XGBoost.

Si nous observons la valeur de la mesure MAE qui représente la moyenne des résidus de la valeur réelle et de la valeur prédite en prenant leur différence en valeur absolue. La valeur de cette mesure contenue dans le tableau ci-dessous n'est pas très élevée ce qui peut suggérer que notre modèle a une bonne qualité de prédiction. Car, chaque résidu est proportionnel à l'erreur totale, et les erreurs importantes sont linéaires à l'erreur totale. En règle générale, nous utilisons la mesure de performance MAE, car elle est plus robuste aux valeurs aberrantes. Tandis que la mesure MSE exprime les résidus au carré. Donc, c'est pour cette raison qu'elle possède une valeur élevée. Nous pouvons voir que cela est le cas dans notre jeu de données, puisque la mesure MSE et RMSE sont élevées. Nous effectuons les mêmes commentaires sur les résultats pour les deux types de montants.

Prime pure	***MAE***	***RMSE***	***MSE***
Frais réels	612	1096	1 200 488
Remboursement SHAM	275	483	233 144

TABLE 35 – Mesures de performances du modèle XGBoost pour le portefeuille 1.

4.3.2 Portefeuille 2

Pour le second portefeuille, plus de variables explicatives étaient disponibles ce qui nous a permis d'utiliser le sexe, la variable représentant le croisement de la catégorie du bénéficiaire avec la classe d'âge, la variable représentant la région où réside l'assuré et son niveau de garantie. Nous pouvons remarquer qu'en moyenne les dépenses annuelles d'un individu en frais réels sont d'environ 1 060€ et de 470 € en montant remboursé par la mutuelle.

Prime pure	*Moyenne observée*	*Moyenne prédite*
Frais réels	1073 €	1060 €
Remboursement SHAM	474 €	463,56 €

TABLE 36 – Comparaison des moyennes prédites et réelles pour le portefeuille 2 avec le modèle XGBoost.

A cet effet, nous utilisons les mesures qui sont présentées dans le tableau ci-dessous. Nous pouvons ainsi en conclure que notre modèle est de plutôt bonne qualité avec une MAE de valeur faible. Néanmoins, notons que nous avons une forte volatilité de nos résultats avec la valeur de la mesure des MSE et RMSE.

Prime pure	*MAE*	*RMSE*	*MSE*
Frais réels	768	1287	1 656 601
Remboursement SHAM	342	579	335 546

TABLE 37 – Mesures de performances du modèle XGBoost pour le portefeuille 2.

4.4 Analyse et comparaison des résultats

Dans cette partie, nous allons comparer les résultats des différents modèles réalisés entre eux. Nous effectuerons ces analyses sur les deux portefeuilles.

4.4.1 Portefeuille 1

Le tableau ci-dessous regroupe les moyennes observées dans la réalité et pour chaque modèle. La moyenne ne nous permet pas de comparer la performance et la précision des modèles entre eux. Nous pouvons dire qu'en moyenne les dépenses annuelles d'un individu sont autour de 830 € en frais réels et 360 € en montant remboursé par la mutuelle.

Prime pure	**Observé**	**Modèle coût-fréquence**	**Tweedie**	**XGBoost**
Frais réels	835 €	818 €	817 €	828 €
Remboursement mutuelle	360 €	360 €	359 €	366 €

TABLE 38 – Comparaison des moyennes prédites par les modèles avec la réalité pour le portefeuille 1.

Dans le tableau suivant, nous donnons les mesures de performance des modèles pour pouvoir les comparer. Ainsi, nous notons qu'en matière de frais réels, les modèles effectués ont quasiment tous les mêmes performances. Nous pourrions dire que l'algorithme XGBoost est un peu moins performant que les GLM dans ce cas-ci. Néanmoins, ils ont tous une bonne qualité de prédiction étant donné que leurs valeurs de MAE sont faibles. Mais, ils ont une certaine volatilité avec des valeurs un peu hautes pour les mesures MSE et RMSE. Dans le cas du montant remboursé par la mutuelle, le modèle Tweedie est sous-performant par rapport aux autres avec des mesures

beaucoup plus élevées. Pour ce cas, il s'agit du modèle coût-fréquence qui se distingue avec des valeurs de mesures plus basses.

Modèle	MAE	RMSE	MSE
Modèle coût-fréquence (frais réels)	609	1090	1 188 253
Tweedie (frais réels)	609	1093	1 195 340
XGBoost (frais réels)	612	1096	1 200 488
Modèle coût-fréquence (mutuelle)	273	479	229 266
Tweedie (mutuelle)	569	1177	1 386 129
XGBoost (mutuelle)	275	483	233 144

TABLE 39 – Comparaison des mesures de performance des modèles pour le portefeuille 1.

4.4.2 Portefeuille 2

Désormais, nous nous concentrons sur les résultats du deuxième portefeuille. Le tableau ci-après regroupe les moyennes observées avec les moyennes des observations prédites de chaque modèle sur l'échantillon de validation. En moyenne, les dépenses annuelles d'un individu sont autour de 1060€ en frais réels et de 470 € en remboursement de la part de la mutuelle.

Prime pure	Observé	Modèle coût-fréquence	Tweedie	XGBoost
Frais réels	1073 €	1063€	1067 €	1060 €
Remboursement mutuelle	474 €	464 €	489 €	464 €

TABLE 40 – Comparaison des moyennes observées et prédites par les modèles pour le portefeuille 2.

Maintenant, nous nous intéressons aux mesures de performance des modèles qui se situent dans le tableau 41 pour pouvoir les comparer. Dans les résultats, nous observons des mesures de performance assez similaires entre les modèles. Mais, nous pouvons tout de même sélectionner le modèle coût-fréquence qui obtient les meilleurs résultats de performance. Dans le cas des frais réels, il semblerait que les modèles ont une forte volatilité au vu des valeurs des mesures MSE et RMSE ce qui tend à moins être le cas pour le montant remboursé par la mutuelle.

Modèle	MAE	RMSE	MSE
Modèle coût-fréquence (frais réels)	759	1 261	1 590 038
Tweedie(frais réels)	760	1 265	1 599 444
XGBoost(frais réels)	768	1 287	1 656 601
Modèle coût-fréquence (mutuelle)	348	574	329 171
Tweedie(mutuelle)	342	576	331 883
XGBoost(mutuelle)	342	579	335 546

TABLE 41 – Comparaison des mesures de performance des modèles pour le portefeuille 2.

4.4.3 Conclusion sur la comparaison des résultats

Pour en conclure sur ces résultats de mesures de performance, nous avons remarqué que le modèle avec la méthode fréquence-coût moyen est celui qui arrivait à obtenir les meilleurs résultats. L'algorithme XGBoost n'était pas très loin derrière. Puis, le dernier en terme de performance était le modèle GLM avec la loi Tweedie tous postes confondus. Il semblerait que le fait d'avoir séparé par sous-poste les modèles cela ait eu un effet sur la performance et la précision du modèle.

5 Application de tarification

Dans cette partie, nous présentons une application qui a été développée en R avec le package Shiny. Elle permet de visualiser les résultats des modèles. L'application offre la possibilité de changer tous les paramètres d'un assuré et de pouvoir observer la prime pure prédite par les modèles pour un individu avec les caractéristiques renseignées. Elle a été développée pour les deux portefeuilles et les deux types de montants. Il faut noter que cette application n'est pas utilisable pour un tarif individuel, car il n'est pas autorisé de segmenter par le sexe (Gender directive). Il s'agit d'un exemple d'outil réalisable, mais qui devra être adapté ultérieurement pour la tarification collective, en tenant compte du niveau des garanties, qui est très influant sur la consommation en santé.

5.1 Paramètres

Dans cette partie, nous observons le premier onglet de l'application qui permet de renseigner tous les paramètres (les caractéristiques du bénéficiaire) nécessaires. Vous trouverez ci-dessous la présentation de cet onglet pour le deuxième portefeuille. Nous pouvons retrouver les mêmes variables étudiées lors des modèles comme la classe d'âge, la catégorie du bénéficiaire, la région, le niveau de garantie et le sexe.

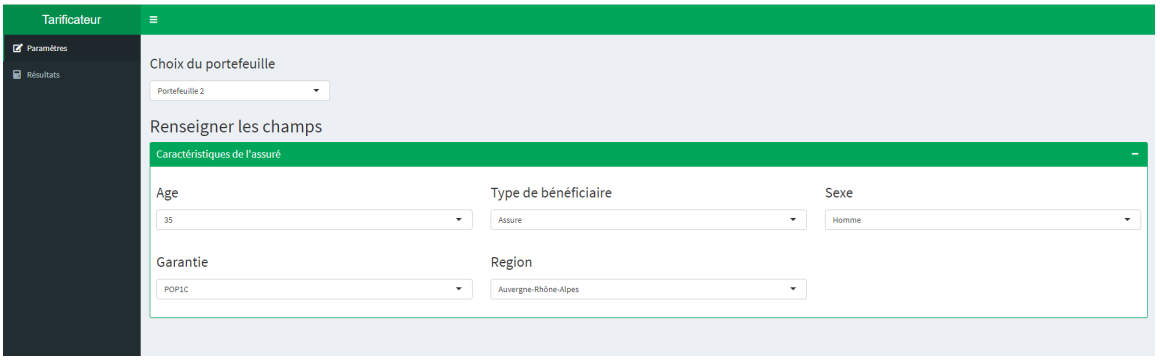

FIGURE 92 – Onglet de paramétrage de l'application.

Dans cet onglet, nous pouvons choisir tout d'abord le portefeuille sur lequel nous souhaitons travailler. Ensuite, nous pouvons choisir les caractéristiques du bénéficiaire des soins. Si nous choisissons d'étudier le portefeuille 2 et que nous prenons l'exemple présenté dans la figure ci-dessus, c'est-à-dire un homme de 35 ans qui appartient à la catégorie bénéficiaire de l'assuré, qui possède un niveau de garantie POP1C et habitant en région Auvergne-Rhônes-Alpes. Finalement, nous verrons les résultats s'affichent dans l'onglet Résultats.

5.2 Résultats

Ainsi, nous avons un deuxième onglet dans notre application permettant de visualiser les résultats de nos modèles. Il y a deux montants s'affichant qui correspondent au montant en frais réels prédit par le modèle à gauche, puis le montant remboursé par la mutuelle prédit à droite. De plus, à l'aide du menu déroulant en haut à gauche vous pouvez passer des résultats d'un modèle à l'autre.

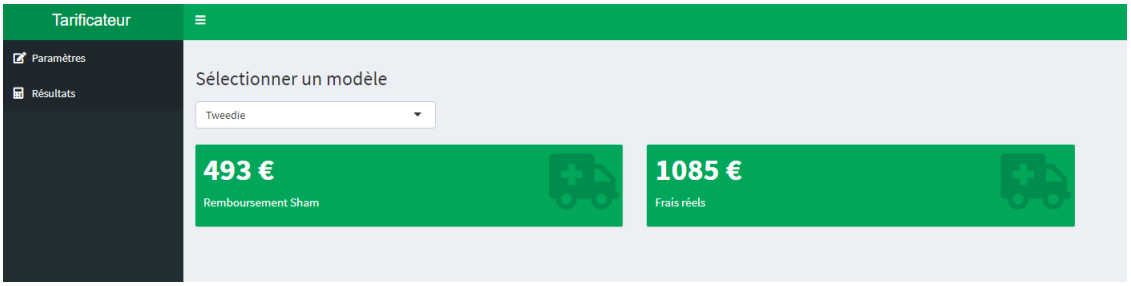

FIGURE 93 – Onglet des résultats de l'application.

D'après l'exemple présenté précédemment, nous obtenons les résultats dans l'onglet ci-dessus. Ainsi, notre assuré pour le modèle Tweedie au niveau global prédira une dépense de 493€ pour le remboursement de la mutuelle et 1 085 € en frais réels.

Conclusion

Afin d'offrir la possibilité à SHAM d'avoir des méthodes permettant d'étudier les paramètres influençant sur la tarification en santé collective, une étude sur la consommation des assurés semblait nécessaire. Au travers de ce mémoire nous avons essayé de répondre à cette problématique.

Au préalable, nous sommes passés par une mise en contexte du sujet d'étude avec l'introduction du système de protection sociale français, la description de la tarification d'un contrat d'assurance et l'exposition du périmètre d'étude qui consiste en la présentation des données disponibles et les traitements qu'elles ont subis. Par la suite, une première analyse des données sur leur ensemble a permis d'observer les relations entre les variables explicatives et les liens entre les variables explicatives avec la variable d'intérêt. Ainsi, nous avons identifié les variables pertinentes. Puis, une analyse des sinistres graves a été réalisée à l'aide de la théorie des valeurs extrêmes pour définir leur seuil.

Une fois que toutes ces étapes étaient terminées, nous avons développé différents modèles. Dans une premier temps, nous avons utilisé comme méthode de modélisation le GLM avec la loi Tweedie pour nous permettre d'avoir une idée de combien vaut en moyenne la prime pure (en ajoutant 5% pour les sinistres graves) et connaître les paramètres importants pour la consommation tous postes confondus. Nous avons aussi remarqué que ces modèles Tweedie ne sont pas de très grande qualité d'ajustement pour nos données. Dans un deuxième temps, nous nous sommes servis de modèles linéaires généralisés avec la méthode fréquence-coût moyen par sous-postes (les lois choisies étaient Gamma pour le coût moyen et binomiale négative pour la fréquence) que nous avons couplé avec des GLM Tweedie par poste pour approcher la consommation des assurés de façon plus précice. Cette méthode nous a permis de pouvoir obtenir une vision plus détaillée de ces dépenses.

Enfin, nous avons comparé tous ces résultats à un modèle un peu plus complexe l'eXtreme Gradient Boosting (XGB) qui est connu pour sa robustesse et pour son efficacité dans les compétitions de machine learning. Ces modèles avaient sensiblement les mêmes résultats en terme de moyenne et de mesures de performance. Néanmoins, nous avions quelques différences en fonction de la taille du portefeuille, du volume disponible de données et du type de montants étudiés. Nous pouvons tout de même noter une petite distinction pour le modèle GLM fréquence-coût moyen par sous-poste qui obtenait de meilleurs résultats. Chacun des modèles apportent une vision différente de la consommation en santé et des paramètres influents. Par ailleurs, l'application R Shiny qui a été développée nous permet de comparer en fonction des caractéristiques paramétrées les résultats ce qui est un bon outil de vision de l'impact des paramètres sur la prime.

A la suite de cette étude, ce mémoire a permis à SHAM de commencer un travail de recherche des paramètres ayant une influence sur la consommation en santé des assurés avec des méthodes telles que les GLM et des algorithmes de machine learning.

Bibliographie

1. *Solidarité Santé* [En ligne]. [Art. Présentation de la sécurité sociale du 16 juillet 2021]. Disponible sur : https ://solidarites-sante.gouv.fr/affaires-sociales/securite-sociale/article/presentation-de-la-securite-sociale
2. *Sécurité Sociale* [En ligne]. [Art. Les grandes dates]. Disponible sur : https ://www.securite-sociale.fr/la-secu-cest-quoi/histoire/les-grandes-dates
3. *Assurance Maladie Ameli* [En ligne]. [Art. Notre histoire du 21 août 2021]. Disponible sur : https ://assurance-maladie.ameli.fr/qui-sommes-nous/histoire/histoire# : :text=Les%20ordonnances%20des%204%20et,famille%20dans%20des%20conditions%20d%C3%A9centes
4. *La Sécurité Sociale* [En ligne]. [Art. Chiffres clés]. Disponible sur : https ://www.securite-sociale.fr/la-secu-cest-quoi/chiffres-cles
5. **Akshita Chugh.** *Medium Analytics Vidhya* [En ligne]. Akshita Chugh [Art. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better ? du 8 décembre 2020]. Disponible sur : https ://medium.com/analytics-vidhya/mae-mse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e
6. **Sunil Ray.** *Analytics Vidhya* [En ligne]. Sunil Ray [Art. A complete comprehensive Guide to Data Exploration du 10 Janvier 2016]. Disponible sur : https ://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/
7. **Collins Ayuya.** *Section* [En ligne]. Collins Ayuya [Art. Parametric versus Non-Parametric Models du 22 février 2021]. Disponible sur : https ://www.section.io/engineering-education/parametric-vs-nonparametric/
8. **Bastien L.** *LeBigData* [En ligne]. Bastien L. [Art. Machine Learning et Big Data : définition et explications du 3 février 2021]. Disponible sur : https ://www.lebigdata.fr/machine-learning-et-big-data
9. *Droit Travail France* [En ligne]. [Art. La protection sociale]. Disponible sur : https ://www.droit-travail-france.fr/protection-sociale.php# : :text=La%20protection%20sociale%20vise%20tous,cons%C3%A9quences%20financi%C3%A8res%20des%20risques%20sociaux.&text=La%20protection%20sociale%2C%20via%20la,compenser%20la%20perte%20de%20revenu
10. *Sécurité Sociale* [En ligne]. [Art. Les grandes dates]. Disponible sur : https ://www.securite-sociale.fr/la-secu-cest-quoi/histoire/les-grandes-dates
11. *Vie Publique* [En ligne]. [Art. Qu'est-ce que la protection sociale ?]. Disponible sur : https ://www.vie-publique.fr/fiches/24109-quest-ce-que-la-protection-sociale
12. **Jake Hoare**. *DisplayR* [En ligne]. [Art. Gradient Boosting Explained – The Coolest Kid on The Machine Learning Block]. Disponible sur : https ://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/
13. **Dimitri Delcaillau.** *Mémoire d'actuariat : Contrôle et Transparence des modèles complexes en actuariat.*, 2019.
14. **David Monnier.** *Mémoire d'actuariat : Modèles linéaires généralisés et assurance santé individuelle : Tarification et évaluation des engagements sous solvabilité II.*, 2016.
15. **François-Henri Toutain.** *Mémoire d'actuariat : Création d'un outil de tarification santé*, 2018.

16. **Riad Bessioud.** *Mémoire d'actuariat : Modélisation de la partie attritionnelle de la provision RC médicale*, 2015.
17. **Jordan Marie-Rose.** *Mémoire d'actuariat : Modélisation comportementale : impact de la couverture santé sur les dépenses en dentaire*, 2018.
18. **Ghita Bouchta.** *Mémoire d'actuariat : Mise en oeuvre de méthodes innovantes de tarification*, 2017.
19. **Arthur Charpentier.** *Computational Actuarial Science with R*, 2016.
20. **Marc-Antoine Cottignies, Alice Lesbats, Manon Le Touzo.** *Research work (Travail d'Etudes et de Recherche)*, 2020.
21. **Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone.** *Classification and Regression Trees*, 1984.
22. **Benoit Cayla** *datacorner* [En ligne]. [Art. La star des algorithmes de ML : XGBoost du 31 mai 2019]. Disponible sur : https ://www.datacorner.fr/xgboost/
23. **Jason Brownlee** *machinelearningmastery* [En ligne]. [Art. A Gentle Introduction to XGBoost for Applied Machine Learning du 17 août 2016]. Disponible sur : https ://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
24. **Jason Brownlee** *machinelearningmastery* [En ligne]. [Art. A Gentle Introduction to XGBoost for Applied Machine Learning du 17 août 2016]. Disponible sur : https ://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
25. **Graphique arbre de décision** *Wikipedia* [En ligne]. Disponible sur : https ://en.wikipedia.org
26. **Guillaume P.** *datascientest* [En ligne]. [Art. Algorithme de descente de gradient du 20 juillet 2020]. Disponible sur : https ://datascientest.com/descente-de-gradient
27. *Wikipedia* [En ligne]. [Art. Algorithme du gradient]. Disponible sur : https ://fr.wikipedia.org/wiki/Algorithme_du_gradient
28. **Krishna Kumar Mahto** *towardsdatascience* [En ligne]. [Art. Demystifying Maths of Gradient Boosting du 25 février 2019]. Disponible sur : https ://towardsdatascience.com/demystifying-maths-of-gradient-boosting-bd5715e82b7c# :~ :text=The%20idea%20is%20simple%2D%20form,suitable%20number%20of%20base%20learners.
29. **Julien Damon** *Telos* [En ligne]. [Art. Protection sociale : enjeux présidentiels du 22 novembre 2021]. Disponible sur : https ://www.telos-eu.com/fr/societe/protection-sociale-enjeux-presidentiels.html

30. *AG2R La Mondiale* [En ligne]. [Art. Protection sociale : enjeux présidentiels du 21 février 2020]. Disponible sur : https ://www.ag2rlamondiale.fr/sante-prevoyance/mutuelle-sante/conseil-le-contrat-responsable-pour-une-mutuelle-sante-qu-est-ce-que-c-est

31. *Institut National de la Consommation* [En ligne]. [Art. Le contrat d'assurance complémentaire santé du 27 février 2019]. Disponible sur : https ://www.inc-conso.fr/content/assurance/le-contrat-dassurance-complementaire-sante

32. *Energie mutuelle* [En ligne]. [Art. Contrats Responsables et Solidaires, de quoi s'agit-il ? de janvier 2016]. Disponible sur : https ://www.energiemutuelle.fr/actualites/contrats-responsables-solidaires

33. *Mutuelle MGC* [En ligne]. [Art. Qu'est-ce qu'un contrat responsable et solidaire ? du 5 août 2019]. Disponible sur : https ://www.mutuellemgc.fr/faq/qu-est-ce-qu-un-contrat-responsable-et-solidaire/# : :text=Comme%20son%20nom%20l'indique,un%20cahier%20des%20charges%20strict

Glossaire

GLM : Generalized Linear Model.
SHAM : Société hospitalière d'assurance mutuelle.
SAS : Statistical Analysis System.
XGBoost : eXtreme Gradient Boosting.
BR : Base de remboursement.
TC : Tarif de convention.
TA : Tarif autorisé.
TM : Ticket modérateur.
CNAM : Caisse nationale de l'assurance maladie.
RSI : Régime social des indépendants.
MSA : Sécurité sociale agricole (Mutualité Sociale Agricole).
CPR SNCF : Caisse de prévoyance et de retraite du personnel de la Société nationale des chemins de fer français.
CAVIMAC : Caisse d'assurance vieillesse, invalidité et maladie des cultes.
CRPCEN : Caisse de retraite et de prévoyance des clercs et employés de notaires.
RSA : Revenu de solidarité active.
CSS : Code de la Sécurité Sociale.
IP : Institutions de Prévoyance.
NGAP : Nomenclature générale des actes professionels.
Code NAF/APE : immatribulation d'activité de l'entreprise qui change en fonction du secteur d'activité.
PSAP : Provisions pour Sinistres A Payer.
PRC : Provisions pour risque croissant.
IBNR : Incurred But Not Reported.
POT : Peak-Over-Threashold.
GEV : Generalized extreme value.
MAE : Mean Absolute Error.
MSE : Mean Square Error.
RMSE : Root Mean Square Error.
MCO : Méthodes des moindres carrés ordinaires.
MV : Maximum de vraisemblance.
AIC : Akaike Information Criterion.
CART : Classification and Regression Tree.
AdaBoost : Adaptative boosting.
GBM : Gradient Boosting Machine.

Annexe

Poste	Sous-poste	Modèle (frais réels)	Modèle (mutuelle)
Dentaire	Orthodontie	GLM/Tweedie pondération 0,22, car le sous-poste représente 22% des dépenses du poste Dentaire.	GLM/Tweedie pondération 0,26.
	Prothèses dentaires	GLM/Tweedie pondération 0,52.	GLM/Tweedie pondération 0,58.
	Radiologie (Dentaire)	GLM/Tweedie pondération 0,03.	GLM/Tweedie pondération 0,02.
	Soins dentaires	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Hospitalisation	Forfait hospitalier	GLM/Tweedie pondération 0,06.	GLM/Tweedie pondération 0,15.
	Frais de séjour	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Frais optionnels	GLM/Tweedie pondération 0,14.	GLM/Tweedie pondération 0,31.
	Soins hospitaliers	GLM/Tweedie pondération 0,32.	GLM/Tweedie pondération 0,23.
	Transport	GLM/Tweedie pondération 0,04.	GLM/Tweedie pondération 0,03.
Médecines douces	Médecines douces	GLM/Tweedie.	GLM/Tweedie.
Optique	Lentilles	GLM/Tweedie pondération 0,09.	GLM/Tweedie pondération 0,01.
	Monture	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Verres	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Pharmacie	Pharmacie remboursement 15%	GLM/Tweedie pondération 0,04.	GLM/Tweedie pondération 0,08.
	Pharmacie remboursement 30%	GLM/Tweedie pondération 0,11.	GLM/Tweedie pondération 0,18.
	Pharmacie remboursement 65%	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Autre (Pharmacie)	GLM/Tweedie pondération 0,19.	GLM/Tweedie pondération 0,19.
Prothèses auditives et autres	Autres prothèses et véhicules	GLM/Tweedie pondération 0,79.	GLM/Tweedie pondération 0,8.
	Petits matériels et pansements	GLM/Tweedie pondération 0,21.	GLM/Tweedie pondération 0,2.
Soins courants	Analyses médicales	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Auxiliaires médicaux	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Consultation spécialiste	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Consultation et visite en médecine générale	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Frais complémentaires de consultations	GLM/Tweedie pondération 0,03.	GLM/Tweedie pondération 0,03.
	Radiologie (Soins courants)	GLM/Tweedie pondération 0,07.	GLM/Tweedie pondération 0,06.
	Autres soins courants	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Divers	Autre (Divers)	GLM/Tweedie.	GLM/Tweedie.

TABLE 42 – Tableau récapitulatif des modèles par sous-poste pour le premier portefeuille.

Poste	Sous-poste	Modèle (frais réels)	Modèle (mutuelle)
Dentaire	Orthodontie	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Prothèses dentaires	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Radiologie (Dentaire)	GLM/Tweedie pondération 0,009.	GLM/Tweedie pondération 0,005.
	Soins dentaires	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Hospitalisation	Forfait hospitalier	GLM/Tweedie pondération 0,05.	GLM/Tweedie pondération 0,16.
	Frais de séjour	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Frais optionnels	GLM/Tweedie pondération 0,13.	GLM/Tweedie pondération 0,29.
	Soins hospitaliers	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Transport	GLM/Tweedie pondération 0,03.	GLM/Tweedie pondération 0,03.
Médecines douces	Médecines douces	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Optique	Lentilles	GLM/Tweedie pondération 0,08.	GLM/Tweedie pondération 0,02.
	Monture	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Verres	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Pharmacie	Pharmacie remboursement 15%	GLM/Tweedie pondération 0,04.	GLM/Tweedie pondération 0,09.
	Pharmacie remboursement 30%	GLM/Tweedie pondération 0,1.	GLM/Tweedie pondération 0,17.
	Pharmacie remboursement 65%	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Autre (Pharmacie)	GLM/Tweedie pondération 0,33.	GLM/Tweedie pondération 0,32.
Prothèses auditives et autres	Autres prothèses et véhicules	GLM/Tweedie pondération 0,79.	GLM/Tweedie pondération 0,8.
	Petits matériels et pansements	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Soins courants	Analyses médicales	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Auxiliaires médicaux	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Consultation spécialiste	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Consultation et visite en médecine générale	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
	Frais complémentaires de consultations	Modèle GLM avec la loi Tweedie du poste Soins courants avec un coefficient 0,03.	Modèle GLM avec la loi Tweedie du poste Soins courants avec un coefficient 0,03.
	Radiologie (Soins courants)	GLM/fréquence-coût moyen.	GLM/fréquence-coût moyen.
Divers	Autre (Divers)	GLM/Tweedie.	GLM/Tweedie.

TABLE 43 – Tableau récapitulatif des modèles par sous-poste pour le second portefeuille.

Table des figures

1 Schéma général d'une prestation santé . . . 13
2 Coefficients de développement des soins courants pour le premier portefeuille . . 18
3 Pyramide des âges du portefeuille 1. . . . 22
4 Répartition des individus par catégorie du bénéficiaire du portefeuille 1. 22
5 Répartition des individus par année de couverture du portefeuille 1. 23
6 Matrice de corrélation des variables qualitatives du portefeuille 1. 23
7 Répartition par tranche d'âge et par catégorie du bénéficiaire du portefeuille 1. . 23
8 Pyramide des âges du portefeuille 2. . . . 24
9 Répartition des individus par catégorie du bénéficiaire du portefeuille 2. 24
10 Répartition des individus par année de couverture du portefeuille 2. 24
11 Répartition des individus par région du portefeuille 2. . . . 24
12 Répartition des individus par niveau de garantie du portefeuille 2. 25
13 Matrice de corrélation des variables qualitatives du portefeuille 2. 25
14 Répartition par tranche d'âge et par catégorie du bénéficiaire du portefeuille 2. . 25
15 Montant de sinistres pour le premier portefeuille. . . . 27
16 Montant de sinistres pour le second portefeuille. . . . 27
17 Répartition de la charge en frais réels par sinistre triée pour le premier portefeuille. 28
18 Répartition de la charge en frais réels par sinistre triée pour le second portefeuille. 28
19 Boîte à moustaches du coût (en frais réels) pour le premier portefeuille. . . . 28
20 Boîte à moustaches du coût (en frais réels) pour le second portefeuille. 28
21 Q-Q plot avec la distribution exponentielle pour le premier portefeuille. 32
22 Q-Q plot avec la distribution exponentielle pour le second portefeuille. 32
23 Test d'appartenance au domaine d'attraction de Fréchet pour le premier portefeuille. . . . 32
24 Test d'appartenance au domaine d'attraction de Fréchet pour le second portefeuille. 32
25 Test d'appartenance au domaine d'attraction de Gumbel pour le premier portefeuille. . . . 33
26 Test d'appartenance au domaine d'attraction de Gumbel pour le second portefeuille. . . . 33
27 Graphique du Pareto Quantile pour le premier portefeuille. . . . 34
28 Graphique du Pareto Quantile pour le second portefeuille. . . . 34
29 Graphique du Pareto Quantile sur les 100 derniers points pour le premier portefeuille. . . . 34
30 Graphique du Pareto Quantile sur les 100 derniers points pour le second portefeuille. 34
31 Mean Excess Plot pour le premier portefeuille. . . . 35
32 Mean Excess Plot pour le second portefeuille. . . . 35
33 Graphique de l'estimateur de Hill pour le premier portefeuille. . . . 36
34 Graphique de l'estimateur de Hill pour le second portefeuille. . . . 36
35 Schéma explicatif du machine learning. . . . 38
36 Illustration d'un arbre de décision. . . . 44
37 Exemple d'un arbre de décision. . . . 45
38 Répartition des dépenses en frais réels par poste pour le premier portefeuille. . . 55
39 Répartition des dépenses de la mutuelle par poste pour le premier portefeuille. . 55
40 Dépenses moyennes annuelles par assuré par poste (en frais réels) du portefeuille 1. . . . 56

41 Dépenses moyennes annuelles par assuré par poste (pour la mutuelle) du portefeuille 1. . . . 56
42 Charge annuelle moyenne pour un assuré par sexe (en frais réels) du portefeuille 1. 58
43 Charge annuelle moyenne pour un assuré par sexe (en remboursement mutuelle) du portefeuille 1. . . . 58
44 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en frais réels) du portefeuille 1. . . . 58
45 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 1. . . . 58
46 Charge annuelle moyenne pour un assuré par tranche d'âges (en frais réels) du portefeuille 1. . . . 59
47 Charge annuelle moyenne pour un assuré par tranche d'âges (en remboursement mutuelle) du portefeuille 1. . . . 59
48 Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en frais réels) du portefeuille 1. . . . 59
49 Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 1. . . . 59
50 Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en frais réels) du portefeuille 1. . . . 60
51 Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en remboursement mutuelle) du portefeuille 1. . . . 60
52 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en frais réels) du portefeuille 1. . . . 60
53 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en remboursement mutuelle) du portefeuille 1. . . . 60
54 Répartition des dépenses en frais réels par poste pour le second portefeuille. . . . 61
55 Répartition des dépenses de la mutuelle par poste pour le second portefeuille. . . . 61
56 Dépenses moyennes annuelles par assuré par poste (en frais réels) du portefeuille 2. . . . 61
57 Dépenses moyennes annuelles par assuré par poste (pour la mutuelle) du portefeuille 2. . . . 61
58 Charge annuelle moyenne pour un assuré par sexe (en frais réels) du portefeuille 2. 63
59 Charge annuelle moyenne pour un assuré par sexe (en remboursement mutuelle) du portefeuille 2. . . . 63
60 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en frais réels) du portefeuille 2. . . . 63
61 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 2. . . . 63
62 Charge annuelle moyenne pour un assuré par tranche d'âges (en frais réels) du portefeuille 2. . . . 64
63 Charge annuelle moyenne pour un assuré par tranche d'âges (en remboursement mutuelle) du portefeuille 2. . . . 64
64 Charge annuelle moyenne pour un assuré par région (en frais réels) du portefeuille 2. . . . 64
65 Charge annuelle moyenne pour un assuré par région (en remboursement mutuelle) du portefeuille 2. . . . 64

66 Charge annuelle moyenne pour un assuré par niveau de garantie (en frais réels) du portefeuille 2. 65
67 Charge annuelle moyenne pour un assuré par niveau de garantie (en remboursement mutuelle) du portefeuille 2. 65
68 Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en frais réels) du portefeuille 2. 65
69 Charge annuelle moyenne pour un assuré par sexe et par catégorie du bénéficiaire (en remboursement mutuelle) du portefeuille 2. 65
70 Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en frais réels) du portefeuille 2. 66
71 Charge annuelle moyenne pour un assuré par sexe et tranche d'âges (en remboursement mutuelle) du portefeuille 2. 66
72 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en frais réels) du portefeuille 2. 66
73 Charge annuelle moyenne pour un assuré par catégorie du bénéficiaire et tranche d'âges (en remboursement mutuelle) du portefeuille 2. 66
74 Estimation du paramètre tweedie *var.power* (frais réels) pour le modèle global du portefeuille 1. 68
75 Estimation du paramètre tweedie *var.power* (remboursement mutuelle) pour le modèle global du portefeuille 1. 68
76 Résidus du modèle de régression Tweedie (frais réels à gauche et montant remboursé par la mutuelle à droite) tous postes confondus pour le portefeuille 1. . . 70
77 Estimation du paramètre tweedie (frais réels) *var.power* du modèle GLM Tweedie tous postes confondus du portefeuille 2. 71
78 Estimation du paramètre tweedie (mutuelle) *var.power* du modèle GLM Tweedie tous postes confondus du portefeuille 2. 71
79 Résidus de Deviance du modèle de régression Tweedie (frais réels et montant remboursé par la mutuelle) tous postes confondus du portefeuille 2. 73
80 Résidus des modèles de régression Tweedie (frais réels) pour le poste optique du portefeuille 2. 77
81 Résidus du modèle de régression Tweedie (remboursement mutuelle) pour le poste optique du portefeuille 2. 77
82 Graphiques des fonctions de densité des lois pour le coût moyen en frais réels du sous-poste monture pour le portefeuille 2. 79
83 Graphiques des fonctions de densité des lois pour le coût moyen du remboursement mutuelle du sous-poste monture pour le portefeuille 2. 79
84 Graphique des fonctions de répartition des lois pour la fréquence du sous-poste monture pour le portefeuille 2. 80
85 Résidus de déviance du modèle coût moyen en frais réels pour les montures du portefeuille 2. 82
86 Résidus de déviance du modèle coût moyen du remboursement mutuelle pour le sous-poste monture du portefeuille 2. 82
87 Graphiques des fonctions de densité des lois pour le coût moyen en frais réels du sous-poste verre du portefeuille 2. 85
88 Graphiques des fonctions de densité des lois pour le coût moyen du remboursement mutuelle du sous-poste verre du portefeuille 2. 85

89 Graphique des fonctions de répartition des lois pour la fréquence du sous-poste verre du portefeuille 2. 86
90 Résidus de déviance du modèle coût moyen en frais réels du sous-poste verre pour le portefeuille 2. 88
91 Résidus de déviance du modèle coût moyen du remboursement mutuelle du sous-poste verre pour le portefeuille 2. 88
92 Onglet de paramétrage de l'application. 95
93 Onglet des résultats de l'application. 96

Liste des tableaux

1 Triangle de liquidation des sinistres. 17
2 Triangle de règlements cumulés. 18
3 Liste des variables de la base pour le premier portefeuille. 19
4 Liste des variables de la base pour le second portefeuille. 20
5 Regroupement des actes par poste et sous-poste. 21
6 Répartition du montant de sinistres pour le premier portefeuille. 26
7 Répartition du montant de sinistres pour le second portefeuille. 26
8 Mesures des sinistres sur la charge totale (en frais réels) pour le premier portefeuille. 29
9 Mesures des sinistres sur la charge totale (en frais réels) pour le second portefeuille. 29
10 Mesures des sinistres sur la charge totale pour le premier portefeuille. 37
11 Mesures des sinistres sur la charge totale pour le second portefeuille. 37
12 Tableau récapitulatif des dépenses et des proportions par sous-poste pour le premier portefeuille. 57
13 Tableau récapitulatif des dépenses et des proportions par sous-poste pour le second portefeuille. 62
14 Mesures de performance du modèle GLM Tweedie au niveau global du portefeuille 1. 68
15 Résultats du modèle GLM Tweedie au niveau global pour le portefeuille 1. . . . 69
16 Résumé des moyennes du modèle GLM Tweedie tous postes confondus du portefeuille 1. 70
17 Mesure de précision du GLM Tweedie tous postes confondus du portefeuille 1. . 70
18 Mesures de performance du modèle GLM Tweedie tous postes confondus du portefeuille 2. 71
19 Résultats du modèle GLM Tweedie tous postes confondus du portefeuille 2. . . 72
20 Résumé des moyennes de prédiction du modèle Tweedie tous postes confondus du portefeuille 2. 73
21 Mesures de précision du modèle GLM Tweedie tous postes confondus du portefeuille 2. 74
22 Mesure de performance du modèle GLM Tweedie du poste Optique du portefeuille 2. 75
23 Résultats du modèle GLM Tweedie du poste Optique du portefeuille 2. 76
24 Résumé des moyennes de prédictions du modèle GLM Tweedie du poste optique pour le portefeuille 2. 78
25 Mesures de précision pour les modèles Tweedie pour le poste optique du portefeuille 2. 78
26 Résultats des modèles coût moyen pour les montures en frais réels et en remboursement mutuelle du portefeuille 2. 81
27 Résultats du modèle GLM pour la fréquence du sous-poste monture du portefeuille 2. 83
28 Mesures de précision des modèles GLM coût moyen et fréquence du sous-poste monture du portefeuille 2. 84
29 Résumé des moyennes de prédiction des modèles coût moyen et fréquence du sous-poste monture du portefeuille 2. 84
30 Résultats du GLM pour le coût moyen en frais réels et en remboursement mutuelle pour les verres du portefeuille 2. 87

31 Résultats du modèle GLM pour la fréquence du sous-poste verre pour le portefeuille 2. 89
32 Mesures de performance des modèles GLM coût-fréquence du sous-poste verre pour le portefeuille 2. 90
33 Résumé des moyennes prédiction des modèles GLM coût-fréquence du sous-poste verre pour le portefeuille 2. 90
34 Comparaison des moyennes prédites et réelles pour le portefeuille 1 avec le modèle XGBoost. 91
35 Mesures de performances du modèle XGBoost pour le portefeuille 1. 91
36 Comparaison des moyennes prédites et réelles pour le portefeuille 2 avec le modèle XGBoost. 92
37 Mesures de performances du modèle XGBoost pour le portefeuille 2. 92
38 Comparaison des moyennes prédites par les modèles avec la réalité pour le portefeuille 1. 92
39 Comparaison des mesures de performance des modèles pour le portefeuille 1. . . 93
40 Comparaison des moyennes observées et prédites par les modèles pour le portefeuille 2. 93
41 Comparaison des mesures de performance des modèles pour le portefeuille 2. . . 94
42 Tableau récapitulatif des modèles par sous-poste pour le premier portefeuille. . . 102
43 Tableau récapitulatif des modèles par sous-poste pour le second portefeuille. . . 103