République Tunisienne
Ministère de l'Enseignement Scientifique et de la Recherche Scientifique
Université de Carthage - Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Rapport de Projet de Fin d'Etudes soumis afin d'obtenir le titre

d'Ingénieur en Statistique et Analyse de l'Information

Par

BOUGUILA Houcem

Modélisation et apprentissage statistique appliqués à la tarification en assurance non vie

---

Soutenu le 07 juillet 2021, devant le jury composé de:

Pr. Mokhtar KOUKI	Président
Mme. Ines ABDELJAOUED TEJ	Rapporteur
M. Hichem RAMMEH	Encadrant
M. Farouk MHAMDI	Examinateur

Projet de Fin d'Etudes fait à Hydatis Engineering

Résumé

Les modèles linéaires généralisés (GLM) constituent l'approche standard de la modélisation du processus de tarification en assurance non-vie depuis de nombreuses années et ont fait l'objet de développements récents visant à améliorer la qualité de la modélisation.

Les avantages des modèles GLM tels que leur capacité à modéliser des comportements non linéaires, en plus des avantages des statistiques paramétriques, permettent de surmonter les cas particuliers et quelques problèmes actuels du processus de tarification.

Les modèles linéaires généralisés présentent également certains inconvénients tels que l'imposition de contraintes sur la variable à modéliser et les variables explicatives, Cela a conduit certains assureurs à utiliser des méthodes d'apprentissage automatique qui sont devenues de plus en plus populaires. Grâce à leur nature non-paramétrique, ces méthodes peuvent s'affranchir des contraintes imposées par les GLM.

Dans ce mémoire, nous essayons d'étudier les enjeux de la tarification, d'aborder les méthodes utilisées sur un cas réel d'assurance santé tout en traitant le problème de la sur-dispersion des données.

Mots-clés – Tarification, Assurance santé, Modèles linéaires généralisés, GLM, Apprentissage statistique, Arbre de décision, Forêts aléatoires, Arbres boostés, Sur-dispersion, Modèle à inflation de zéros.

Abstract

Generalized Linear Models (GLM) have been the standard approach to modeling the non-life insurance pricing process for many years and have been the subject of recent developments aimed at improving modeling quality.

The advantages of GLM models, such as their ability to model non-linear behavior, in addition to the advantages of parametric statistics, make it possible to overcome the special cases and some of the current problems of the pricing process.

Generalized linear models also have some drawbacks such as imposing constraints on the variable to be modeled and the explanatory variables. This has led insurance companies to use machine learning methods which have become increasingly popular. Thanks to their non-parametric nature, these methods override the constraints imposed by GLM.

Our aim is to study the issues of pricing process, discuss the methods used on a real case of health insurance while dealing with the problem of over-dispersed data.

Keywords— Pricing, Health insurance, Generalized linear models, GLM, Machine learning, Decision tree, Random forest, Gradient boosting, Over-dispersion, Zero-inflated model

Remerciements

Je tiens à remercier les personnes suivantes, sans lesquelles je n'aurais pas pu mener à bien ce projet, et sans lesquelles je n'aurais pas réussi à finaliser mon cursus et à obtenir mon diplôme d'ingénieur !

L'équipe de Hydatis Engineering pour leur accueil et la confiance qu'ils m'ont accordée. Je tiens à exprimer ma gratitude à mon maître de stage Monsieur Alaeddine AZIZ pour l'opportunité qu'il m'a offerte ainsi que le suivi tout au long du stage.

Je tiens également à remercier Monsieur Hichem RAMMEH, mon encadrant universitaire, pour ses encouragements et ses remarques pertinentes lors du suivi du travail effectué.

Je suis également reconnaissant à ma famille et mes amis, pour leur soutien moral durant toutes les étapes de ma formation universitaire.

Table des Matières

Liste des Figures

Liste des Tableaux

I Introduction

L'évaluation du risque pour une compagnie d'assurance est au cœur de son travail quotidien. Pour ce faire, les assureurs appliquent la modélisation statistique, notamment les modèles de régression pour quantifier la relation entre le risque et les variables qui le décrivent. Après l'introduction des modèles linéaires généralisés (GLM) par Wedderburn et Nelder dans un article en 1972 [Nelder and Wedderburn(1972)], il n'a fallu que quelques années pour que les compagnies d'assurance commencent à introduire les modèles GLM, car ils présentaient une sophistication des modèles de régression simples, jusqu'à ce que les modèles linéaires généralisés deviennent les principaux modèles dans le travail d'évaluation des risques du secteur de l'assurance.

La souscription de polices d'assurance non-vie augmente d'année en année pour un certain nombre de raisons, notamment les exigences légales relatives aux polices d'assurance santé dans certains pays. Cette augmentation rend le marché de l'assurance très concurrentiel, ce qui a incité les assureurs à mettre en place diverses techniques de modélisation prédictive pour proposer des prix compétitifs en tenant compte de leurs engagements envers les assurés et de leur capacité à les rembourser.

La résolution de ce compromis entre prix compétitifs et capacité de remboursement est au cœur de la tarification de l'assurance non-vie. Dans ce sens, plusieurs techniques et méthodes sont développées dans le but d'une meilleure modélisation, notamment avec la disponibilité actuelle des données et le développement des techniques d'analyse.

La portée des méthodes d'apprentissage automatique atteint également les compagnies d'assurance. En effet, les compagnies d'assurance commencent à intégrer certains algorithmes d'apprentissage automatique dans leur processus de tarification. Cependant, les modèles GLM restent la norme du secteur en termes de développement de modèles analytiques de tarification, en raison de leur capacité à modéliser des comportements non linéaires et à bénéficier des avantages de la statistique paramétrique.

C'est dans ce cadre que ce projet est réalisé et notre objectif est de répondre aux questions suivantes : Quels sont les enjeux actuels de la tarification ? Les modèles GLM peuvent-ils surmonter les contraintes du processus de tarification tout en ayant une performance qui les maintient comme une norme dans la modélisation dans le secteur d'assurance, en particulier avec l'existence des méthodes d'apprentissage automatique ? Les méthodes d'apprentissage automatique confirment-elles leur succès dans le système de tarification ?

Ce rapport est divisée en trois parties principales, la première est consacrée à la contextualisation du projet dans laquelle le système de tarification en assurance non-vie est introduit et la méthode de calcul des primes est expliquée.

L'objectif de la deuxième partie est d'expliquer théoriquement la modélisation par les modèles linéaires généralisés. Nous nous concentrons également sur les méthodes d'apprentissage automatique, de la théorie générale et des arbres de décision aux méthodes plus avancées telles que les forêts aléatoires et les arbres boostés.

Dans la troisième partie, nous mettons en œuvre les méthodes discutées dans la partie théorique sur un cas réel d'assurance maladie où nous commençons par un traitement et une analyse des données collectées, cette étape est essentielle pour améliorer la qualité des données et comprendre les relations existantes entre les variables afin d'avoir une vision globale sur le portefeuille d'étude et de modéliser correctement par la suite.

La phase de modélisation est divisée en deux parties où nous commençons par modéliser la fréquence des sinistres. Durant cette phase, nous sommes contraints à l'un des problèmes majeurs de la souscription qui est l'excès de zéros dans le nombre de sinistres, nous montrons donc l'amélioration des performances en utilisant les modèles modifiés en zéro ou le modèle GLM avec la distribution Binomiale négative par rapport au modèle classique de Poisson. Une performance marquée est également obtenue pour le modèle d'arbre de régression et un modèle de l'algorithme gradient boosting. Une approche similaire dans la partie modélisation du coût moyen conduit à un modèle de forêt aléatoire plus performant que les autres.

Une dernière partie est consacrée à la mise en service des modèles étudiés en les déployant sur une plateforme web dans le cadre d'un projet de l'entreprise d'accueil.

II Contexte du projet

II.1 Entreprise d'accueil

Hydatis Engineering est une nouvelle entreprise innovante, spécialisée dans le conseil, le développement de services technologiques et la transformation digitale.

Elle dispose de deux affiliations, l'une à Tunis et l'autre à Courbevoie en France. L'objectif principal de Hydatis Engineering est de rendre ses clients plus innovants et productifs en leur fournissant de nouvelles solutions utilisant l'intelligence des données et en les guidant vers une transformation digitale qui conduit à plus de profit et à une meilleure performance.

Hydatis Engineering se caractérise par ses trois domaines d'expertises :

- **Expérience digitale** : Hydatis Engineering possède une expertise dans la transformation digitale en offrant des services de conseil et de gestion de la stratégie. Elle propose également des modèles orientés client avec une approche numérique.
- **Data intelligence** : Hydatis Engineering se concentre également sur les domaines de la science des données pour aider ses clients à extraire des informations de leurs données et les amener à prendre de meilleures décisions.
- **Outsourcing Nearshore** : En tant qu'ESN (entreprise de services numériques), Hydatis Engineering travaille également sur des projets d'externalisation ou outsourcing pour des éditeurs, des ESN et des intégrateurs de solutions de gestion.

II.2 Méthodologie de travail

Pendant le stage, la méthodologie de travail adoptée est Scrum, qui est une méthodologie du processus de développement logiciel Agile. Adopter une méthodologie Agile signifie que le travail est basé sur le développement itératif et que les exigences et les solutions évoluent grâce à la collaboration entre les différentes équipes de manière organisée. En outre, l'implication active du client et sa participation au projet sont considérées comme fondamentales dans les pratiques agiles.

La méthodologie Agile offre plusieurs implémentations, dont la plus connue est SCRUM, celle que nous avons adoptée dans notre travail.
Alors qu'Agile est une philosophie ou une orientation, Scrum est une méthodologie plus rigide et spécifique pour la gestion d'un projet. Elle est basée sur un processus qui identifie le travail, détermine

qui va effectuer chaque tâche, comment elle sera effectuée et quand elle sera terminée.
L'approche Scrum nécessite l'utilisation de cycles de développement appelés Sprints. À la fin de chaque sprint, une partie du projet est achevée et l'équipe identifie le travail à effectuer au cours du sprint suivant.
Scrum augmente considérablement la productivité et réduit le temps nécessaire pour obtenir des résultats par rapport aux processus traditionnels. Il est idéalement utilisé dans les projets dont les exigences évoluent rapidement.

II.3 Tarification en assurance non vie

II.3.1 Assurance non vie

L'assurance non-vie est généralement définie comme toute assurance qui n'est pas considérée comme une assurance-vie. Autrement dit, elle concerne les opérations d'assurance qui ne concernent pas la vie de l'assuré. Elle peut couvrir les personnes, les biens, l'assurance responsabilité civile ou les dettes.

L'assurance non-vie s'agit d'une police qui prévoit une indemnisation pour les pertes subies à la suite d'un sinistre. Cette police présente un document contractuel qui fixe les conditions d'un engagement entre une compagnie d'assurance et un assuré.

Dans le cadre de cet engagement et suite à la survenance d'un sinistre, la compagnie d'assurance doit rembourser à l'assuré sa perte sous certaines conditions et pendant une période déterminée en échange d'un paiement initial, appelé prime.

II.3.2 La tarification

La tarification est le processus d'estimation de la prime d'assurance que l'assuré doit payer pour bénéficier d'une couverture en cas de sinistre.

Le défi actuariel consiste à estimer ce montant sur la base de données historiques afin qu'il soit en phase avec l'évolution des risques. L'activité de tarification des compagnies d'assurance fait d'elles des sociétés de gestion des risques, puisqu'elles s'occupent tout simplement de la couverture des risques futurs de leurs assurés.

Afin de bien gérer ses risques, une compagnie d'assurance sépare toujours les différents contrats de la base de données en plusieurs catégories afin que dans chaque classe le risque soit le plus homogène possible.

On parle de classes a priori lorsqu'on segmente en fonction d'informations liées à l'assuré telles que l'âge, le sexe, le type de soins dans le cas de l'assurance maladie, etc.

On parle de classes a posteriori lorsque la segmentation utilise plutôt l'historique des sinistres.

Dans notre cas d'étude, nous nous intéressons particulièrement à la tarification a priori.

II.3.3 Calcul de la prime d'assurance

La prime d'assurance présente le montant qu'un assuré doit payer pour bénéficier de la couverture en cas du sinistre dans une période déterminée.

La valeur de la prime se compose de trois parties : une partie risque, une partie frais et une partie prestations.

La partie liée au risque est composé essentiellement de la prime pure qui présente le coût probable du sinistre, elle est considérée comme la partie principale de la valeur de la prime et sa détermination représente le principal défi du processus de tarification car elle est directement liée à l'évolution du risque dans un sens ou dans l'autre (par exemple, augmentation ou diminution des prix aux soins de santé ou du prix des médicaments).

Une charge de sécurité est ajoutée à la prime pure. Cette valeur est nécessaire pour que l'assureur puisse résister à la volatilité des sinistres.

Concrètement, l'évaluation de cette partie du risque est donc liée aux données historiques dont dispose l'assureur et à sa capacité de modélisation. A cette composante du risque, nous ajoutons des frais de gestion pour couvrir les frais de fonctionnement de l'assureur ainsi que les taxes.

Enfin, on ajoute au montant déjà établi, qui est entièrement technique, une part de bénéfice que l'assureur fixe en fonction de ses objectifs commerciaux.

Dans ce contexte, la meilleure estimation de la prime pure est le principal résultat du processus de souscription. Sa valeur représente la charge totale moyenne des sinistres payés par l'assureur. Mathématiquement, elle représente l'espérance des pertes.

Pour déterminer cette prime, il faut d'abord calculer la charge totale des sinistres S qui se définit comme suit :

$$S = Y_1 + \ldots + Y_N = \sum_{i=1}^{N} Y_i$$

Avec N : Variable aléatoire représentant le nombre des sinistres, à valeurs dans $\mathbf{N}$.
Y_i : Variable aléatoire représentant le coût d'un sinistre i, à valeurs dans $\mathbf{R}+$

Lorsque les coûts individuels Y_i sont i.i.d. et indépendants du nombre de sinistres N, La prime pure étant l'espérance des pertes peut être écrite [Charpentier(2010)] :

$$E(S) = E(N)E\left(Y_1\right)$$

dès que $E(N)$ et $E(Y_1)$ existent et sont finies.

II.3.4 Modèle Coût-Fréquence

Dans l'approche classique que nous abordons dans le calcul de la prime, les hypothèses présentées ci-dessus nous amènent à supposer que le coût total des sinistres est une somme aléatoire de variables

aléatoires indépendantes et identiquement distribuées, chacune représentant le coût d'un sinistre.

En d'autres termes, ce modèle suppose que la prime pure suit une distribution composée coût-fréquence. En pratique, cela se traduit par une distinction entre la modélisation de la fréquence des sinistres et celle du coût de chacun de ces sinistres.

Notre étude est donc composée de deux parties :

- La modélisation de la loi de fréquence.
- La modélisation de la loi de coût.

II.4 Revue de littérature

Les systèmes de tarification actuarielle sont généralement basés sur des avis d'experts ou sur une approche axée sur les données. Cette dernière approche tire parti des données pour effectuer une modélisation et générer des prédictions qui conduisent à des décisions. Les modèles linéaires généralisés (GLM) représentent l'approche standard de l'industrie pour développer des modèles analytiques de tarification [Haberman and Renshaw(1996)].

Les modèles GLM fournissent des résultats qui peuvent être facilement interprétés car un coefficient est calculé pour chaque facteur de risque utilisé dans la modélisation.

Les méthodes d'apprentissage statique sont moins faciles à interpréter, mais elles ont changé le paysage de la modélisation prédictive dans de nombreuses applications. Cela a attiré l'attention sur la possibilité de mettre en œuvre ces méthodes dans le domaine de l'assurance. Nous remarquons donc que le nombre d'articles de recherche dans cette direction augmente de plus en plus. En effet, en 2012, Guelman.L a publié un article [Guelman(2012)] pour comparer les GLMs et les modèles Gradient Boosting dans la prédiction du coût des pertes en assurance automobile. Wuthrich.M et Buser.C [Wuthrich and Buser(2020)] ont montré comment les méthodes basées sur les arbres de décision peuvent être adaptées pour modéliser la fréquence des pertes.

Ferrario.A, Noll.A et Wuthrich.M [Ferrario et al.(2020)Ferrario, Noll, and Wuthrich] montrent les aspects à considérer lors de l'exécution de modèles de réseaux de neurones sur des données de fréquence de sinistres.

II.5 Objectifs

Après la contextualisation du projet et la documentation des méthodes couramment utilisées en tarification, nous avons fixé les objectifs en deux grandes parties:

Une partie théorique : Dans cette partie nous visons à nous familiariser avec l'aspect théorique des modèles linéaires généralisés en étudiant les différents modèles qui peuvent être des solutions au problème causé par la structure des données en assurance non-vie.

Nous souhaitons également rappeler la théorie de l'apprentissage statistique supervisé et détailler les modèles basés sur des arbres de décision allant d'un simple arbre de régression à des méthodes plus

avancées telles que les méthodes d'agrégation et de boosting.

Une partie pratique : Notre objectif pour cette partie est de pouvoir appliquer les méthodes abordées dans la partie théorique sur un exemple réel de données d'assurance maladie en Tunisie.

L'objectif de la partie pratique n'est pas de prouver qu'un modèle est plus efficace qu'un autre mais plutôt d'essayer d'adapter chacun d'entre eux aux données et l'optimiser afin de modéliser correctement la fréquence de la sinistralité et le coût des sinistres pour finalement obtenir des prédictions sur la prime pure.

Un dernier objectif de la partie pratique est de convertir les méthodes développées en un processus de tarification déployé sur une plateforme web.

III Partie Théorique

III.1 Modélisation statistique

III.1.1 Contexte

Dans l'un des deux cas modélisation de la fréquence ou modélisation du coût le problème s'inscrit dans le cadre des problèmes de régression pour lesquels la variable d'intérêt la fréquence ou la charge de sinistres est quantitative, par opposition aux problèmes de classification, pour lesquels la variable réponse est qualitative.

Les modèles traditionnels répondant à ce type de question sont historiquement les modèles linéaires généralisés (GLM), qui étendent le cadre très limité de la régression linéaire multiple.
Les modèles linéaires ont été utilisés également pour modéliser la prime pure, cependant, dans la réalité la variable réponse n'est pas forcément gaussienne.
Ainsi, les modèles linéaires généralisés (GLM), comme leur nom l'indique, présentent une généralisation de la régression linéaire ordinaire puisque la variable réponse Y peut suivre une autre distribution de la famille exponentielle que la gaussienne.

Les modèles gaussiens linéaires étant le point de départ de l'étude des GLM, il est important de commencer par comprendre le principe du modèle de régression linéaire et d'en souligner les principaux aspects.

III.1.2 Modèle de régression linéaire

Les modèles linéaires permettent de décrire une variable de réponse en fonction d'une combinaison linéaire de variables prédictives.

Définition :

On appelle modèle linéaire un modèle statistique qui peut s'écrire sous la forme :

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \epsilon$$

- Y est une variable aléatoire réelle observée que l'on souhaite prédire appelée variable à expliquer ou variable réponse.
- Les variables $X_1,\ldots,X_k$ sont des variables réelles non aléatoires appelées variables explicatives ou prédicteurs vu qu'elles sont censées expliquer la variable Y.

- $\beta_0, \beta_1, ..., \beta_k$ sont les paramètres à estimer.
- ϵ est une variable aléatoire réelle non observée qui présente le terme d'erreur dans le modèle pour lequel on suppose :

$$E(\epsilon) = 0$$
$$\text{Var}(\epsilon) = \sigma^2 > 0$$

  Avec σ^2 un paramètre inconnu à estimer.
- Ces deux hypothèses sur ϵ impliquent les caractéristiques suivantes sur Y :

$$E(Y) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$$

$$\text{Var}(Y) = \sigma^2$$

III.1.3 Les modèles linéaires généralisés

Comme mentionné précédemment, les modèles linéaires généralisés (GLM) présentent une extension des modèles linéaires puisque la distribution de la variable réponse n'est pas limitée à la Normale mais peut être vue comme une réalisation de toute distribution de la famille exponentielle.

Contrairement aux modèles linéaires, les GLM peuvent également modéliser une dépendance non linéaire grâce aux fonctions de liaison (fonctions liens) qui associe linéairement l'image de la variable de réponse aux variables explicatives.

Chaque modèle GLM se base sur les trois éléments clés suivants :

- **Une composante aléatoire Y :**
  Dans le cadre de la GLM les observations de la variable aléatoire Y (variable réponse) sont supposés indépendantes et suivent une loi de probabilité de la famille exponentielle.

  La forme générale de la densité d'une loi de la famille exponentielle est donnée par :

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

  Où,
  - θ : paramètre de position.
  - ϕ : paramètre de dispersion.
  - a,b et c sont des fonctions réelles.

  Pour chaque problème, et lors de la conception de notre modèle, nous devons associer la distribution la plus appropriée de la famille exponentielle à notre variable de réponse Y.
- **Une composante déterministe :**
  Elle représente le prédicteur linéaire qui correspond à une combinaison linéaire des variables explicatives :

$$\eta = X^t\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Où,

$$X^t = \begin{pmatrix} 1 & X_1 & \dots & X_k \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

- **Une fonction lien g :** Cette fonction strictement monotone et différentiable représente la relation qui relie la composante aléatoire à la composante déterministe, en d'autres termes, c'est le lien qui associe linéairement l'image de la variable de réponse aux variables explicatives: Notons $\mu = E(Y)$ alors,

$$g(\mu) = \eta \text{ ou } \mu = g^{-1}(\eta) = g^{-1}\left(X^t\beta\right)$$

Il ressort clairement de cette équation que l'espérance de Y n'est rien d'autre qu'une transformation du prédicteur.

La liste des fonctions liens correspondantes aux lois de probabilité usuelles :

Loi de probabilité	Fonction de lien canonique
Normale	$g(x) = x$
Poisson	$g(x) = \log(x)$
Gamma	$g(x) = \frac{1}{x}$
Binomiale	$g(x) = \log(x) - \log(1-x)$

III.1.4 Modélisation de la prime pure

Nous rappelons à partir des sections précédentes que la prime pure qu'on souhaite déterminer représente la charge totale moyenne des sinistres et que dans le cadre du modèle collectif, cette prime suit une distribution composée coût-fréquence.

Ainsi, le problème s'agit d'une modélisation d'une loi composée de deux variables aléatoires (la fréquence et le coût). Par la suite, on doit étudier les lois des probabilités qui ajustent d'une manière indépendante ces variables.

III.1.5 Modélisation de la fréquence

La modélisation de la fréquence revient à la modélisation du « Nombre moyen de sinistres » qui présente une variable de comptage, elle doit donc être modélisée par une loi discrète de la famille exponentielle.

Les lois de type discrètes couramment utilisé dans le problème de tarification :

- Loi de Poisson
- Loi Binomiale négative (utilisée pour les données de comptage sur dispersées)

Loi de Poisson :

La loi de poisson est parfaitement adaptée à la modélisation de la fréquence puisqu'elle décrit le comportement du nombre d'événements (nombre de sinistres dans notre cas) au cours d'une période donnée.
Pour un nombre moyen d'occurrences égale à λ, alors la probabilité qu'il existe n occurrences est :

$$\mathbb{P}(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}, \lambda > 0$$

On dit alors que N suit une loi de Poisson de paramètre λ.

Les moments de premier ordre et second ordre :

$$E(N) = Var(N) = \lambda$$

Propriété d'équidispersion :

L'équidispersion est une caractéristique unique de la distribution de Poisson elle est représentée par le fait que la moyenne et la variance d'une variable distribuée par Poisson sont identiques.

Cette propriété d'équidispersion présente une hypothèse forte dans la théorie de la modélisation de la loi de poisson. En pratique, il est nécessaire d'analyser si un problème de sur-dispersion ou sous-dispersion existe.

La sous-dispersion est un phénomène rare, contrairement à la sur-dispersion présentée par le fait que la variance est supérieure à la moyenne, la variance dans ce cas est définie comme suit :

$$\text{Var}(N) = \phi E(N)$$

Avec $\phi \geq 1$, le paramètre de dispersion.

Dans le cas de l'assurance, les causes les plus probables de sur-dispersion sont : l'hétérogénéité du portefeuille ou la présence d'un grand nombre de valeurs nulles observées dans la variable "Nombre de sinistres", puisqu'en réalité la majorité des assurés ne subissent aucun sinistre, ce qui fait que le nombre de sinistres est gonflé à zéro.

Dans la littérature plusieurs méthodes pouvant corriger ce type de problème, dont les plus connus :

- Modèle quasi-Poisson.
- Modélisation par la loi Binomiale négative.
- Les modèles à inflation de zéro.

Modèle quasi-Poisson :

Ce modèle garantit de garder le cadre simple d'un modèle de Poisson, la seule différence est que ce modèle inclut un second paramètre utilisé dans l'estimation de la variance conditionnelle, appelé coefficient de dispersion.

Le modèle estimé avec cette correction suppose maintenant essentiellement une distribution d'erreur de type poisson avec une moyenne µ et une variance $\phi\mu$
Soit Y une variable aléatoire tel que :

$$E(Y) = \mu \quad \text{Var}(Y) = \phi\mu$$

Où, $\mu > 0$ et $\theta > 0$
Le fait que l'esperance et la variance ci-dessus sont reliées étroitement à l'esperance et la variance d'une distribution de Poisson, ainsi que l'utilisation d'une fonction de lien logarithmique justifie l'appellation de modèle "quasi-Poisson", et on note : $Y \sim \text{Poi}(\mu, \theta)$. [Ver Hoef and Boveng(2007)]

Les estimations des termes μ sont identiques aux estimations du modèle de Poisson et donne ainsi les mêmes estimations des paramètres β. Les deux modèles se diffèrent uniquement au niveau de l'expression de la variance.

Loi Binomiale négative :

Le modèle de régression Binomial-Négatif a été introduit pour répondre au phénomène de sur-dispersion lié aux modèles de Poisson.

Il y a plusieurs façons de construire la distribution Binomiale négative, mais la plus intuitive consiste à introduire un terme aléatoire d'hétérogénéité θ de moyenne 1 et de variance α dans le paramètre de la distribution de Poisson.

$$E(Y, \theta) = \exp\left\{X^t\beta + \epsilon\right\} = \mu\theta$$

Ainsi, la densité sera définie par :

$$f(y, \theta) = \exp\{-\mu\theta\} \frac{(\mu\theta)^y}{y!}$$

Si la variable θ suit une distribution Gamma, tel que les deux paramètres sont choisis pour être égaux à $\frac{1}{\alpha}$ ayant la densité suivante:

$$f(\theta) = \frac{(1/\alpha)^{1/\alpha}}{\Gamma(1/\alpha)} \theta^{1/\alpha - 1} \exp(-\theta/\alpha)$$

alors il est prouvé [Boucher et al.(2008)Boucher, Denuit, and Guillén] que l'espérance et la variance seront égales à :

$$E(Y) = \mu \quad \text{Var}(Y) = \mu + \alpha\mu^2$$

L'expression de la variance, montre que le modèle Binomiale négative suppose que la variance peut augmenter par le carré la moyenne, ce qui permet de montrer que ce modèle peut traiter les cas où la sur-dispersion est forte.

Critère de choix : Test d'ajustement du Khi-deux

Pour choisir la distribution la plus adéquate à une variable aléatoire discrète comme la fréquence, il convient de réaliser un test d'ajustement du Khi-deux.

Le but d'un test d'ajustement du Khi-deux est de déterminer si une variable est susceptible de provenir d'une distribution spécifiée ou non.

Données :

- Soit $(Y_1, \ldots, Y_n)$ l'échantillon des variables aléatoires présentant le nombre de sinistre dans les données.
- On suppose qu'on peut observer k valeurs de nombre de sinistres, et $n_1, ..., nk$ les effectifs respectives.
- On note $p_1, ..., pk$ les probabilités théoriques associés aux k valeurs.

Hypothèses :

H_0 **:** Les données observés suivent la distribution théorique.
H_1 **:** Les données observés ne suivent pas la distribution théorique.

Statistique du test :

$$D = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}$$

Si H_0 est vraie, alors D $\longrightarrow \chi^2_{(\mathrm{k-r-1})}$, avec k le nombre de valeurs qui peuvent être observées et r le nombre de paramètres estimés pour que la loi théorique soit définie.

Résultat :

En comparant la valeur de la statistique du test à la valeur critique de la table du Khi-deux, on peut décider d'accepter ou de rejeter le test.

Les modèles à inflation de zéro :

Comme indiqué précédemment, l'une des causes les plus probables du phénomène de surdispersion est la présence d'un grand nombre de valeurs nulles du nombre de sinistres dans le portefeuille, d'où l'intérêt du modèle à inflation zéro comme solution au problème.

Le modèle à inflation zéro est un modèle statistique basé sur une distribution de probabilité gonflé en zéro, c'est-à-dire une distribution qui permet des observations fréquentes à valeur nulle. Il permet de modéliser tout événement aléatoire contenant un excès de données à valeur nulle par unité de temps.

On estime la présence d'un excès de zéros quand le nombres de valeurs nulles observés par la variable réponse dépasse le nombre des zéros estimé par un ajustement avec la loi de poisson. Dans la littérature, deux modèles modifiés en zéro existent :

- Les modèles Zero inflated Poisson (ZIP).
- Les modèles Zero inflated Negative Binomial (ZINB).

Le modèle Zero inflated Poisson (ZIP) :

Ce modèle introduit par « Diane Lambert » mélange deux processus. Le premier consiste régi par une distribution par une loi de Bernoulli à détecter la présence ou non de sinistres. Le second est régi par une distribution de Poisson de paramètre μ qui génère le nombre de sinistres. Le mélange est décrit comme suit :

$$P(Y=0) = \pi + (1-\pi)\exp(-\mu)$$
$$P(Y=y) = (1-\pi)\exp(-\mu)$$

Où π présente la probabilité de non sinsitralité.

Le modèle Zero inflated Negative Binomial ZINB :

Comme le modèle ZIP, le modèle ZINB associe également deux processus. Le premier processus gère la génération des zéros. Le second génère les valeurs estimées par une distribution binomiale négative de paramètres μ et ν. Il est décrit comme suit [Karatekin(2014)] :

$$P\left(Y=y\right) = \begin{cases} \pi + (1-\pi)\left(1+\nu\mu\right)^{-1/\nu} \text{ si } y=0 \\ (1-\pi)\,\dfrac{\Gamma\left(y+1/\nu\right)}{\Gamma\left(y+1\right)\Gamma(1/v)}\left(\dfrac{1/\nu}{1/\nu+\mu}\right)^{1/\nu}\left(\dfrac{\mu}{1/\nu+\mu}\right)^{y} \text{ si } y>0 \end{cases}$$

III.1.6 Modélisation du coût

L'objet de cette partie est de modéliser coût moyen de sinistres. Pour cela, similairement à la modélisation de la fréquence nous étudierons les distributions qui ajustent mieux la variable coût.

La variable coût de sinistres ou encore montants de sinistres présente une variable aléatoire continue à valeurs réelles positives. Rappelons qu'on s'intéresse aux distributions appartenant à la famille exponentielle.

Pour cela, nous étudierons les lois continues couramment utilisés en tarification suivantes :

- Loi Gamma.

-Loi Log-Normale.

Le critère de sélection de la loi que nous utiliserons sera le graphique Q-Q plot.

Loi Gamma

La distribution Gamma est généralement utilisée pour les distributions asymétriques à droite, ce qui est le cas pour notre structure de coûts.

Densité de la loi Gamma :

Soit X une variable aléatoire qui suit la loi Gamma de paramètres α et β, alors sa fonction de densité est de la forme :

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, x > 0, \alpha, \beta > 0$$

Les moments de premier ordre et second ordre :

$$\mathbb{E}(X) = \frac{\alpha}{\beta}$$
$$\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$$

Loi Log-Normale

La loi Log-Nomale est également utilisé pour les distributions asymétriques à droite notamment en présence des valeurs moyennes faibles et des variances élevées.

Définition :

Une distribution log-normale est une distribution de probabilité continue d'une variable aléatoire dont le logarithme est normalement distribué. Ainsi, si la variable aléatoire X suit une distribution log-normale, alors Y = ln(X) suit une distribution normale.

Pour une log-normale de paramètres μ et σ^2, alors la fonction de densité de probabilité peut se mettre sous la forme :

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

Les moments de premier ordre et second ordre :

$$\mathbb{E}(X) = e^{\frac{\mu+\sigma^2}{2}}$$
$$\mathrm{Var}(X) = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}$$

Critère de choix: Q-Q plot

La fonction quantile-quantile ou **Q-Q plot** est un graphique permettant de comparer deux distributions de probabilité en traçant leurs quantiles l'un par rapport à l'autre.

Dans notre cas, il s'agit d'une représentation des valeurs ordonnées de l'echantillon en fonction des quantiles théoriques de la distribution Gamma ou Log-Normale.

Si les deux ensembles de quantiles proviennent de la même distribution, nous devrions voir les points former une ligne à peu près droite.

III.1.7 Estimation des paramètres

Une fois on a choisit la distribution qui s'adapte mieux a notre variable cible et qu'on a spécifié la fonction lien et les variables explicatives à intégrer dans le modèle, il nous reste qu'appliquer la GLM en estimant les coefficients de régression $\beta_0, ..., \beta_k$.

Pour les modèles GLM, l'estimation des coefficients se fait par l'estimateur du maximum de vraisemblance

Fonction de vraisemblance :

Soit Y la variable aléatoire réponse de densité $f(y,\theta)$, on exprime la fonction de vraisemblance comme étant la fonction de densité conjointe de l'échantillon.

$$L(y_1, \ldots, y_n, \theta) = \prod_{i=1}^{n} f(y_i, \theta)$$

Dans le cas du modèle linéaire généralisé où la variable réponse est issue de la famille exponentielle alors sa fonction de vraisemblance sera :

$$L(y_1, \cdot, y_n; \theta, \phi) = \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}.$$

Puisque la fonction logarithme est croissante alors la maximisation de la vraisemblance revient à maximiser son logarithme.

$$\log(L) = \sum_{i=1}^{n} \frac{y_i\theta_i - b\left(\theta_i\right)}{a_i(\phi)} + c\left(y_i, \phi\right)$$

Pour estimer les paramètres il faut maximiser la dernière fonction en calculant son dérivée en fonction de chaque paramètre β_i

$$\frac{\partial \log(L)(\boldsymbol{\beta_i})}{\partial \boldsymbol{\beta_i}} = \sum_{i=1}^{n} \frac{\left(y_i - b'\left(\theta_i\right)\right)}{a_i(\phi)} \frac{\partial \theta_i}{\partial \boldsymbol{\beta_i}}$$

III.1.8 Critères de validation et comparaison des modèles

Déviance

La déviance D se définit comme étant la différence entre le modèle proposé et le modèle saturé, c.-à-d, le modèle qui suppose que chaque point de données a ses propres paramètres, c'est le modèle parfait où tous les μ_i sont remplacés par y_i.

$$D := -2\left[\log(L)(\hat{\boldsymbol{\beta}}) - \log(L)_{saturé}\right]\phi$$

Si la déviance est faible, cela signifie que la différence de vraisemblance entre le modèle proposé et le modèle saturé est faible, c'est-à-dire les valeurs modélisées et les valeurs observées sont proches. En d'autres termes, plus la déviance est faible, plus le modèle proposé est performant en termes d'ajustement.

Test de rapport de vraisemblance

Le test du rapport de vraisemblance (Likelihood Ratio Test : LRT) est utilisé pour comparer la qualité de l'ajustement de deux modèles statistiques. Le test compare deux modèles emboîtés et détermine si l'augmentation de la complexité par l'ajout de paramètres supplémentaires rend le modèle significativement plus précis en termes d'adéquation aux données, ce qui est mesuré par la réduction de la déviance.

Remarque : Par "modèles emboîtés", nous entendons qu'il existe un modèle complexe et un modèle plus simple imbriqué qui diffèrent par le fait que le second contient moins de paramètres que le premier.

Hypothèses du test

H_0 **:** Pas de différence significative entre le modèle complexe et le modèle imbriqué.

H_1 **:** Il existe différence significative entre le modèle complexe et le modèle imbriqué.

Statistique du test

$$LR = 2 \cdot \ln\left(\frac{\mathcal{L}(Complexe)}{\mathcal{L}(Imbriqué)}\right) = 2(\ln \mathcal{L}(Complexe) - \ln \mathcal{L}(Imbriqué))$$

Intérêt

Nous utiliserons ce test pour vérifier l'intérêt d'inclure ou non des facteurs lors de la phase de sélection des variables. Le test sera également utile pour comparer notre modèle au modèle nul ne contenant que la constante afin d'étudier la significativité globale, ainsi que pour le comparer au modèle saturé afin d'étudier l'ajustement du modèle aux données.

AIC

L'AIC (Akaike Information Criterion) est un critère qui permet également d'évaluer la qualité d'un modèle et de le comparer à d'autres modèles.Ainsi, l'AIC fournit un moyen de sélection des modèles. Soit un modèle de k paramètres estimés et $\hat{L}$ le maximum du vraisemblance obtenue, alors l'AIC se définit par l'expression suivante :

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

Si on compare plusieurs modèles alors celui qui présente la valeur AIC minimale est le modèle préféré.

l'AIC récompense la qualité de l'ajustement mais il inclut également une pénalité qui croit avec le nombre de paramètres estimés k. Le but de la pénalisation des modèles ayant beaucoup de paramètres décourage les modèles de devenir très complexes et évite ainsi le sur-apprentissage.

BIC

Le BIC (Bayesian information criterion) est également une mesure de qualité et un critère de sélection des modèles. Il se définit comme suit :

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

Comme l'AIC, le critère BIC permet également un compromis entre la réduction du biais en ayant plus de paramètres et le découragement du sur-apprentissage en pénalisant le nombre élevé de paramètres. Cependant, le critère BIC effectue une pénalité plus sévère qui est fonction de la taille de l'échantillon car il s'est avéré que le critère AIC dépréciait les modèles avec un grand nombre de paramètres dans le cas d'un grand échantillon [Lebarbier and Mary-Huard(2004)].

III.2 Méthodes d'apprentissage statistique

III.2.1 Généralités et place de l'apprentissage statistique dans le domaine de l'assurance

Dans la littérature, les GLM sont de loin les modèles de tarification les plus populaires et les plus utilisés, même dans le secteur actuel. Cependant, les progrès technologiques ont déplacé le centre d'intérêt de la recherche vers l'apprentissage automatique et les techniques d'analyse de données massives, ce qui a également eu un impact sur le domaine de l'assurance.
En outre, certains travaux de recherche assez récents vont au-delà de la zone de confort actuarielle des GLM et montrent ainsi la performance des méthodes d'apprentissage dans ce domaine.

Dans cette partie, nous nous intéressons à l'aspect théorique de l'apprentissage statistique supervisé, nous commencerons par une explication mathématique générale, puis nous étudierons les arbres de décision jusqu'à arriver à des méthodes plus avancées comme le bagging, les forets aléatoires et le boosting.

III.2.2 Théorie de l'apprentissage statistique

L'apprentissage statistique est un domaine d'étude basé sur des approches mathématiques pour explorer les données à des fins explicatives ou prédictives. Ce domaine d'étude fait partie de l'intelligence artificielle, car il donne aux machines la capacité d'apprendre à partir de données.

Arthur Samuel[1] le définit comme étant le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés.

L'apprentissage statistique peut se diviser en deux catégories :

- ***L'apprentissage supervisé :*** C'est la situation d'apprentissage où on essaye de comprendre les relations entre les variables d'un ou plusieurs exemples annotés. Son but consiste à faire apprendre la machine à apprendre à partir de ces données annotées. Par exemple, un algorithme qui classe une nouvelle observation à un ensemble de classes prédéfinies.

- ***L'apprentissage non-supervisé :*** Contrairement à l'apprentissage supervisé, ce type d'apprentissage fait référence à la situation où les données ne sont pas annotées. Dans ce cas, la machine va essayer de trouver un schéma qui lui permet d'extraire des informations qui n'avaient pas été détectées auparavant. Par exemple, en se basant uniquement sur les observations d'une variable à expliquer, l'algorithme d'apprentissage non supervisé regroupe les observations en différentes classes homogènes.

Remarque : Remarque : Dans la suite de notre étude, on s'intéresse uniquement à l'apprentissage statistique supervisé vu qu'on va utiliser que des méthodes supervisées dans le problème de tarification.

[1]Un pionnier américain dans le domaine de l'intelligence artificielle qui a popularisé le terme "apprentissage automatique" en 1959

Concept de l'apprentissage supervisé

Soit Y une variable que l'on souhaite prédire par apprentissage supervisé. Contrairement à la modélisation classique (comme les modèles linéaires généralisés) qui nécessite d'imposer des hypothèses sur la distribution des données, l'apprentissage statistique se limite à une seule hypothèse liée au processus de génération des observations de la variable Y qui doivent être générées de manière identique et indépendante.

Le but de l'apprentissage consiste à avoir un algorithme qui va apprendre à prédire la valeur Y en fonction des variables explicatives X. Ce qui traduit mathématiquement par la meilleure décision en moyenne déduite à partir :

$$\mathrm{P}(\mathrm{Y} = . \mid X = x)$$

But : Utiliser l'information disponible présentée par un ensemble d'observations (X,Y) : (x_1,y_1),. . . , (x_n,y_n), pour trouver une stratégie permettant de prévoir une sortie Y associé à une nouvelle entrée X.

Formalisation [2] **:**

Ensemble de données d'entrainement : Dn := $\{(\mathrm{X}_1, \mathrm{Y}_1), \ldots, (\mathrm{X}_\mathrm{n}, \mathrm{Y}_\mathrm{n})\}$ avec :

- - X_i les variables d'entrées à valeur dans $\mathcal{X}$.
- - Y_i les variables de sortie à valeurs dans $\mathcal{Y}$.

Hypothèse : Les données sont indépendantes et identiquement distribuées.

Algorithme d'apprentissage : C'est une fonction $\widehat{f}$ appelée fonction de prédiction construite à partir de la règle d'apprentissage suivante :

$$\bigcup_{n\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$$
$$D_n \mapsto \hat{f}$$

Avec $\mathcal{F}$ l'ensemble des fonctions de prédictions. Ainsi, l'algorithme est la fonction optimale de la prédiction.

Pour expliquer le choix d'un algorithme « optimale », nous devons s'intéresser à la théorie de cette décision et définir ainsi une fonction coût.

Théorie de la décision :

- Soit f une fonction de prédiction.
- Soient les variables aléatoires (X,Y) présentant une paire de données de test.
- $\mathcal{A}$: l'ensemble d'actions ou de prédictions possibles.

[2] Cette partie est basée sur le cours : Introduction au cadre de l'apprentissage supervisé de Guillaume Obozinski.

- Soit ℓ une fonction de $\mathcal{A} \times \mathcal{Y}$ à valeurs dans R, appelé fonction de coût ou fonction de perte. Cette fonction quantifie le coût à payer pour avoir choisi l'action a quand la variable Y prend une valeur donnée y.

A partir de la fonction de coût, nous estimons la qualité de la fonction de prédiction f pour le problème de décision lié à X,Y et la fonction de coût ℓ par l'erreur de généralisation ou simplement le risque défini par :

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)]$$

Au vu de cette définition, on peut alors définir la fonction optimale de prédiction $f^\star$ comme une fonction de $\mathcal{F}$ qui minimise le risque R :

$$f^\star = \text{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$$

III.2.3 Arbres de décision

Généralités

Dans cette partie, on s'intéresse aux arbres de décision l'une des méthodes de l'apprentissage supervisé utilisée dans le domaine du data mining et qui prouve son utilité dans le problème de tarification.

L'objectif est de créer un modèle qui prédit la valeur d'une variable cible (ou à prédire) en fonction de plusieurs variables d'entrée dites explicatives ou prédictives. Le but de ce modèle consiste à construire un partitionnement de groupes de données les plus homogènes possibles du point de vue de la variable cible en utilisant les variables prédictives.

L'apprentissage par arbre de décision trouve son nom du fait qu'il s'agit d'une méthode structurée hiérarchiquement comme un arbre. En effet, Ce terme arbre est lié à la notion de récursivité dans le fonctionnement de l'algorithme représenté en une série de tests sur les variables explicatives en vue de prédire un résultat ou une classe.

Dans la représentation d'un arbre de décision, on utilise des nœuds, des branches et des feuilles. Chaque nœud représente une conjonction dans laquelle un test sur les variables explicatives est effectué. En fonction de la réponse au test, on se déplace le long d'une branche vers des nœuds suivants qui peuvent être terminaux, ils sont appelés dans ce cas des feuilles. Chaque branche présente donc une valeur de la variable prédictive traitée dans le nœud parent. Chaque feuille représente une classe à part entière.

Une fois l'arbre est construit, On comprend que pour chaque nouveaux individu (ou observation) qu'on souhaite attribuer à une classe ou de lui affecter une valeur, l'algorithme teste les variables décrivant cet individu dans les nœuds de l'arbre jusqu'à arriver aux feuilles où les décisions sont prises.

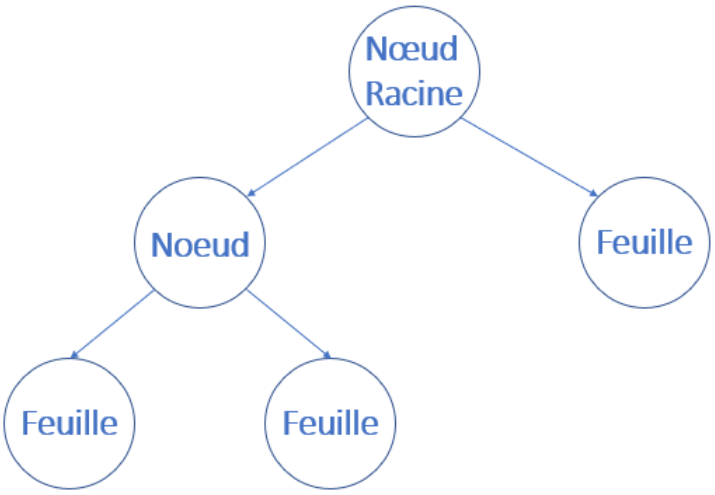


Figure III.1: Schéma simplifié d'un arbre CART

Dans l'approche des arbres de décision deux types d'estimation existent et qui sont distingués par la nature de la variable cible.

Arbres de classification : lorsque la variable à prédire est qualitative. Le but étant donc de prévoir la classe à laquelle va appartenir la réponse.

Arbres de régression : lorsque la variable à prédire est quantitative. Le but étant de déterminer sa valeur.

Dans ce qui suit, nous n'étudions que les arbres de régression puisque nous ne nous intéressons pas aux arbres de classification dans notre modélisation des coûts ou des fréquences.

Construction d'un arbre de décision : Méthode CART

L'algorithme CART (Classification And Regression Trees), a été introduit en 1984 par Leo Breiman, il s'agit d'une méthode très utilisée dans la construction des arbres de décision, elle peut traiter à la fois les taches de classification et de régression.

La construction d'un arbre de type CART consiste à créer les noeuds, où chaque noeud est déterminé par le choix de la variable explicative qui sera utilisé pour faire une division.

La division est déterminé par un seuil de la variable explicative utilisé si elle est de type quantitative ou un regroupement de modalités si la variable explicative est qualitative.

Le choix de la variable explicative pour faire la division dans un noeud se base sur un critère d'hétérogénéité que l'algorithme essaye de le minimiser pour chaque division.

La construction de l'arbre se termine à l'aide d'une règle d'arrêt. En effet, en suivant cette règle l'algorithme considère qu'un noeud sera divisé ou devient une feuille.

Arbre de régression

Pour les problèmes de régression, la fonction de coût ou encore la fonction d'hétérogénéité qu'on souhaite minimiser lors de chaque division est la somme des erreurs au carré.

$$\sum (Y_{observé} - Y_{prédite})^2$$

Arbre Optimal et élagage des arbres

La méthode CART commence par construire un grand arbre (dite saturé) obtenu par divisions successives et des nouvelles coupures de sorte que la fonction coût ou l'impureté soit minimale.

Une fois l'arbre est construit, CART applique un élagage dans un but d'optimiser la performance, d'éviter le phénomène de sur-apprentissage et d'améliorer ainsi la robustesse du modèle.

L'objectif de l'élagage consiste à trouver l'arbre optimal entre l'arbre saturé et la moins complexe (la plus petite : à une seule feuille).
Plusieurs méthodes sophistiquées peuvent être utilisées, telles que l'élagage de type « coût-complexité ».

III.2.4 Méthodes ensemblistes

Contrairement aux modèles robustes tels que les GLM, les algorithmes d'apprentissage sont parfois très instables et peuvent donner de fausses estimations auxquelles on ne veut pas se fier lorsqu'on doit prendre une décision importante. C'est pourquoi nous utilisons des méthodes ensemblistes, qui sont basées sur l'agrégation de plusieurs modèles afin d'obtenir un meilleur résultat. La divergence des modèles dans ce modèle ensembliste devient un avantage puisque chaque modèle est spécialisé sur une partie des données.

Lorsque nous parlons de méthodes ensemblistes, nous parlons principalement des méthodes parallèles ou d'agrégation et des méthodes séquentielles comme le modèle Gradient Boosting.

Méthodes parallèles : Agrégation

Bagging

Le Bagging est un ensemble de méthodes d'agrégation introduites par L.Brieman. Le principe du Bagging consiste à sous-échantillonner les données d'entraînement en sous-échantillons appelés échantillons bootstrap. La construction d'un estimateur agrégé performant consiste à rassembler plusieurs estimateurs appelées « classifieurs faibles » faibles chacun construit sur un sous-échantillon.

Le sous-échantillonnage est effectué de manière indépendante, ce qui permet d'atténuer la dépendance entre les estimateurs. Si les estimateurs sont indépendants et ont la même distribution (iid), le biais de l'estimateur agrégé sera le même que celui des estimateurs mais sa variance diminuera. De ce résultat, nous déduisons l'avantage du Bagging des estimateurs puisqu'elle résout les principaux

problèmes de l'apprentissage supervisé liés au compromis biais-variance, comme le problème du sur-apprentissage.

Cependant en pratique rien ne garantit la performance de l'estimateur agrégé vu que la réduction de la variance dépend de la corrélation des estimateurs et de leur variance qui peut être très élevée.

Forêts aléatoires

La méthode des forêts aléatoires est une technique similaire au Bagging, comme son nom l'indique elle s'agit de l'agrégation de plusieurs arbres de classification ou de régression. La seule différence entre l'algorithme de forêt aléatoire et l'algorithme Bagging est que dans le Bagging, nous utilisons toutes les variables explicatives pour tester toutes les divisions possibles d'un nœud, puis nous choisissons une division optimale parmi celles-ci, tandis que dans l'algorithme de forêt aléatoire, nous n'utilisons qu'un nombre déterminé de variables tirées aléatoirement de l'ensemble des variables.

L'ajout de ce caractère aléatoire à la construction du modèle agrégé permettra une plus grande indépendance entre les arbres et réduira ainsi la variance de l'estimation.

Méthodes séquentielles

Contrairement aux méthodes parallèles et au principe du Bagging où les estimateurs sont construits séparément les uns des autres, l'idée des méthodes séquentielles est que la construction se fait de manière itérative de sorte que chaque modèle utilise les résultats du précédent et corrige ainsi les défauts. Dans ce contexte, on parle du principe de Boosting.

Boosting

Le principe du boosting diffère de celui du bagging en ce que les classifieurs faibles ne sont pas construits indépendamment mais plutôt séquentiellement en introduisant des poids à chaque étape. Tous les classifieurs sont construits sur les mêmes données d'entraînement sans utiliser d'échantillonnage. Le principe du boosting consiste à ajouter des poids aux observations des données d'apprentissage de sorte que les observations mal prédites par un classifieur soient surpondérées et que le classifieur suivant prête attention à cette pondération pour corriger les défauts. Le processus est complété de cette manière séquentielle d'un modèle à un autre jusqu'à ce qu'un classificateur performant soit atteint.

Le principal avantage du boosting est qu'il peut transformer un ensemble de classifieurs faibles en un classifieur puissant. En effet, contrairement au Bagging où les estimateurs sont identiquement distribués, permettant d'avoir le même biais à la fin, l'aspect séquentiel des corrections apportées par le Boosting permet d'avoir un biais significativement réduit.

Gradient boosting machine: GBM

Le Gradient Boosting est une technique de boosting géneralement utilisé sur les arbres de décision, elle se présente comme étant une aggrégation entre le boosting et la technique du descente du gradient.

- Descente du gradient : C'est un algorithme d'optimisation itératif permettant de trouver un minimum local d'une fonction. En apprentissage automatique il est utilisé pour déterminer les paramètres en minimisant la fonction de perte.

Dans le cas de la régression, chaque modèle généré utilise les résidus obtenus par le modèle précédent et tente de minimiser la fonction de perte. Ainsi, la construction successive des modèles permet de réduire les résidus obtenus par chaque modèle par le suivant, ce qui fait que les modèles GBM se caractérisent par leur réduction significative des erreurs de prédiction.

IV Étude de cas pratique

IV.1 Produit étudié

Notre étude se concentre sur le cas de l'assurance maladie collective. Le contrat d'assurance maladie est un instrument financier permettant de faire face aux impacts économiques qui peuvent déstabiliser le patrimoine suite à des problèmes de santé, principalement en cas d'hospitalisations, d'accidents ou de maladies graves.
Une assurance maladie est dite collective ou de groupe lorsqu'elle présente un contrat qui offre une couverture à un groupe de membres, généralement composé d'employés d'une entreprise ou de membres d'une organisation.

IV.2 Analyse des données

IV.2.1 Les données d'étude

Pour chaque étude réalisée en assurance il est important de bien définir une période d'étude car généralement les phénomènes étudiés n'ont pas forcement la même périodicité. Ainsi, dans notre étude nous nous basons sur des données relatives à l'année 2017.

Les données dont on dispose consiste en deux tables de données d'assurance santé collective. La première contenant les informations sur les contrats et les assurés ordonnées et identifiés par leurs clés (identifiant), les clés des organisations dont ils sont liés, ainsi que et les numéros de contrats réalisés. Parmi les informations disponibles de cette table, on cite les variables suivantes :

- Numéro Contrat
- Id organisation (à laquelle appartient l'assuré)
- Nom organisation
- Classe de l'organisation (classe 1 ou A)
- Id assuré
- Date début contrat
- Date d'expiration contrat

- Statut : L'assuré peut être l'assuré principal ou un dépendant (Conjoint ou enfant).
- Date de naissance
- Sexe

On dispose également d'une table de données sinistres. Ces données contiennent les actes médicaux survenus lors de la période. Pour chaque observation, on dispose des informations sur le type d'acte, le montant demandé, le montant remboursé, etc. Parmi les variables disponibles on cite :

- Numéro Contrat
- Id organisation (à laquelle appartient l'assuré)
- Id assuré
- Poste de garantie : Hospitalisation, Pharmacie, Soins courants, Dentaire, Optique.
- Nom du donneur du certificat ou autre document: Médecin, hôpital, clinique ...
- Évaluation spécifique : un commentaire sur l'acte maladie.
- Date de déclaration.
- Date de remboursement.
- Montant demandé
- Montant remboursé

Enfin, nous regroupons les deux tableaux pour avoir des informations à la fois sur les sinistres et les assurés. Le regroupement par identifiant permet de montrer le nombre (fréquence) et le coût moyen des sinistres par assuré en fonction de chaque poste de garantie.

IV.2.2 Analyse par poste de garantie

Une bonne pratique en matière de tarification d'assurance consiste à ne pas mélanger les sinistres de remboursement de différentes catégories de couverture, car elles se comportent différemment en termes de fréquence et de coût. En effet, il est clair qu'il existe une grande différence entre la fréquence et le coût des soins courants (par exemple, les visites chez le médecin) et l'hospitalisation, qui est moins fréquente et plus coûteuse si elle est effectuée dans le secteur privé. Une analyse par poste de garantie est alors nécessaire car nous essayons toujours de segmenter l'information pour affiner toujours plus les tarifs.

Description des Postes

Dans le tableau ci-dessous on présente les différentes catégories d'actes (ou postes de garanties) étudiés. Pour chaque poste, on cite pour chaque catégorie les principaux exemples d'actes existants dans nos jeux de données. Nous affichons également le nombre d'actes présent dans les données pour chaque catégorie.

Postes de garanties	Exemples d'actes	Nombre d'actes
Soins Courants	- Consultation Médecin (généraliste ou spécialiste)	7188 (40.7%)
	- Analyses en laboratoires	
	- Centres de traitements et diagnostics	
	- Imagerie médicale	
Dentaire	- Consultation et visite chez les dentistes	636 (3.6 %)
	- Prothèses dentaires	
Pharmacie	- Médicaments	7433 (42.2 %)
Hospitalisation	- Frais d'interventions médicaux ou chirurgicaux.	2021 (11.5 %)
	- Frais de Séjour	
	- Maternité	
Optique	- Frais opticien (verres, lentilles, etc.)	350 (2%)

Table IV.1: Les différentes actes des postes des garantie

IV.2.3 Traitement des données

Cette étape est essentielle avant toute étude, étant donné l'importance de la qualité des données sur la modélisation, puis sur les résultats et les décisions. Il est donc important de choisir les informations à conserver et celles à rejeter et de gérer les valeurs manquantes et les erreurs de saisie.

Un premier traitement effectué consiste à éliminer les variables que nous jugeons inutiles pour notre étude. Ainsi, nous ne gardons que les variables mentionnées ci-dessus à l'exception des dates de début de contrat et d'échéance, puisque tous les contrats sont lancés début janvier 2017 et arrivent à échéance début janvier 2018. Dans ce cas, nous ne nous intéressons pas à la période d'exposition, car il n'y a pas de renouvellement puisqu'il s'agit d'une seule période de contrats et parmi les variables existantes, il n'y a pas d'information indiquant la résiliation d'une police.

Un deuxième traitement nécessaire consiste à convertir la variable "Date de naissance" en une variable âge, puisque les variables de type date sont inutiles dans la construction des modèles.

Grâce à cette variable âge, nous avons pu identifier et corriger les informations manquantes et fausses dans les données. En effet, pour la variable "statut de l'assuré" il y a 49 valeurs manquantes. La présence de la variable d'âge a montré que les observations avec ces valeurs manquantes sont des enfants. Nous avons donc pu remplacer ces valeurs manquantes par la valeur "Enfant".

De plus, lorsque nous analysons les observations des personnes de moins de 18 ans, nous remarquons qu'il y a 42 bénéficiaires avec le statut "Principal" et 119 avec le statut "Conjoint". Il est donc clair que ces valeurs sont des erreurs de saisie et doivent être corrigées avec la valeur "Enfant".

IV.2.4 Statistiques descriptives

Après traitement des données, nous avons un tableau de 89040 observations et 11 variables. Il est important de voir la structure de ces données en décrivant les caractéristiques de manière quantitative afin d'avoir une vision globale du portefeuille d'étude.

La structure des variables qualitatives est la suivante :

Variable qualitative	Modalités et pourcentage
Sexe	· Femme : 51.24 % · Homme 48.76 %
Statut	· Principal 37.4 % · Conjoint 26.5 % · Enfant 36.1 %
Situation matrimoniale	· Divorcé(e) 0.08 % · Marié(e) 53.41 % · Célibataire 46.45 % · Veuf(ve) 0.06 %
Poste garantie	· Soins courants 40.7 % · Dentaire 3.6 % · Pharmacie 42.2 % · Hospitalisation 11.5 % · Optique 2%
Classe	· Classe 1 50.86 % · Classe A 49.14 %

Table IV.2: Statistiques descritptives des variables qualitatives

La structure des variables quantitatives est la suivante :

Variable	Minimum	Médiane	Moyenne	Maximum
Age	0	34	29.96	79
Nombre d'actes	0	0	0.19	10
Montant demandé	0	45	80.33	8441.52
Montant remboursé	0	40	64.51	3990.61

Table IV.3: Statistiques descriptives des variables quantitatives

Variable Poste de garantie La structure de la variable poste de garantie est extraite que des données sinistrés vu que le type de poste est déterminé après la déclaration du sinistre.

De cette structure, nous pouvons déduire que les catégories pharmacie et soins courants sont majoritaires dans les données, tandis que les actes optiques et dentaires sont les moins fréquents.

L'analyse de la structure de cette variable est très importante pour connaître la distribution des sinistres par catégorie d'actes dans le portefeuille d'étude, et pour connaître la nature de la division que nous allons faire sur les données puisque nous allons les analyser par poste de garantie.

Variable Sexe

Le portefeuille de l'étude est caractérisé par une population féminine supérieure de 51,24% à la population masculine de 48,76%. Cette structure reste la même pour la catégorie des enfants, alors que dans la catégorie de l'assuré principal les hommes présentent la majorité de la population et inversement pour la catégorie du conjoint où le nombre de femmes est beaucoup plus important.

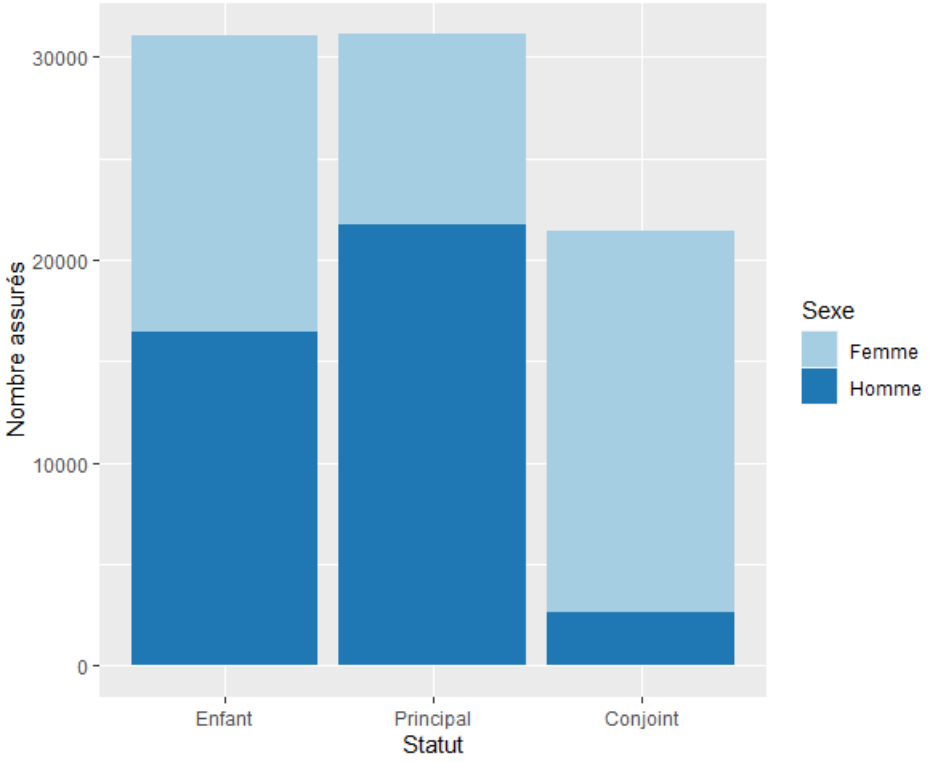


Figure IV.1: Répartition des assurés par statut et sexe

Variable Age

L'âge moyen des assurés est presque égal à 30 ans, cependant, on peut voir sur la pyramide des âges IV.2 que les groupes d'âge les plus représentés sont les enfants de moins de 9 ans et le groupe d'âge entre 27 et 36 ans pour les femmes. Cela peut être dû au fait que cette catégorie est liée aux actes de santé relatifs à la maternité, d'où une souscription élevée de contrats pour cette catégorie.

Nous affichons ci-dessous la pyramide des âges par sexe de l'assuré. La visualisation du graphique nous a conduit à construire la variable «Tranche d'âge »

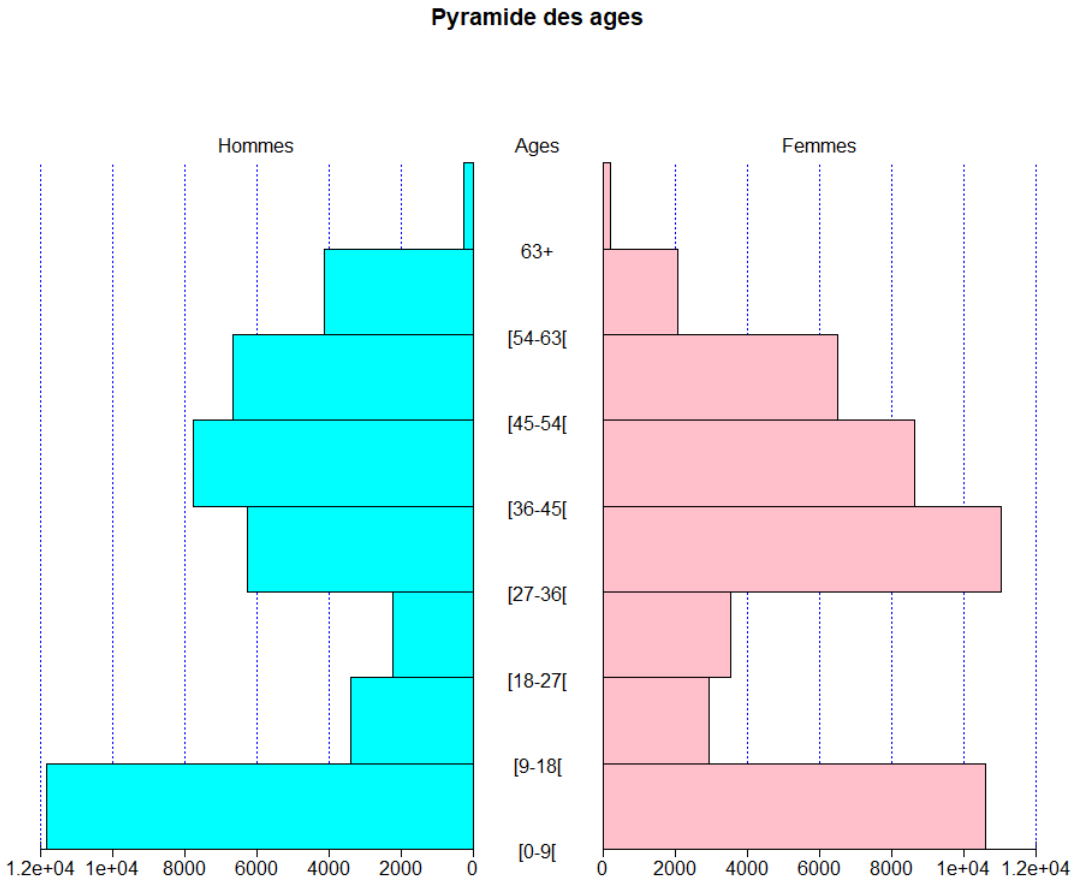


Figure IV.2: Pyramide des âges par sexe

IV.2.5 Analyse de la fréquence

La durée d'exposition des assurés dans les données est la même, c'est-à-dire qu'ils sont tous présents pour la même période d'un an, ceci est lié au fait que les contrats sont des contrats de groupe qui sont généralement standards pour la même période et commencent en début de l'année.
On rappelle que :

$$\text{Fréquence} = \frac{Nombredesinistres}{Exposition}$$

Or, l'exposition est toujours égale à 1 dans notre cas. Parler de fréquence revient donc à parler du nombre de sinistres.

La variable du nombre de sinistres (ou fréquence) est notre variable cible dans la première étape de la résolution du problème, qui est la modélisation de la fréquence. Il est donc important de décrire sa structure dans le portefeuille et sa relation avec les autres variables.

Dans un premier temps, nous avons commencé par créer une nouvelle variable appelée "Sinistre" qui indique si le bénéficiaire a déclaré au moins un sinistre ou non. La structure de cette variable dans le portefeuille est la suivante :

Sinistre	Nombre d'observations
Oui (au moins 1)	7273

Table IV.4: Structure de la variable Sinistre

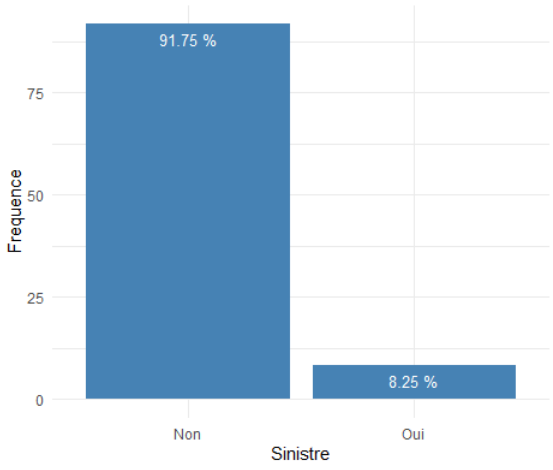


Figure IV.3: Structure de la variable Sinistre

Nous passons ensuite à l'analyse de la variable du nombre de sinistres elle-même. Le tableau ci-dessous montre le nombre d'observations dans les données associées à chaque nombre de sinistres.

Nombre d'actes	0	1	2	3	4	5	6	7
Nombre d'observations	82610	1832	4916	1787	1523	607	453	262

Nombre d'actes	8	9	10	11	12	13	14	18
Nombre d'observations	122	86	78	13	4	6	4	1

Table IV.5: Répartition de la fréquence des sinistres

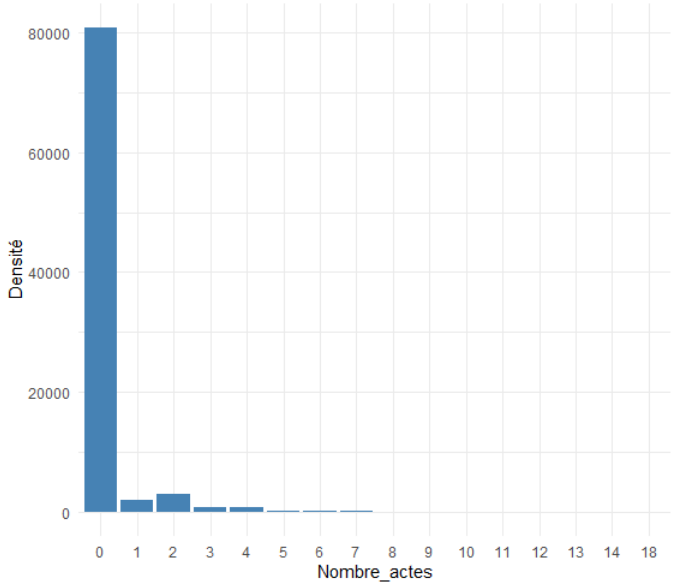


Figure IV.4: Répartition de la fréquence des sinistres

Il ressort du tableau des valeurs numériques et du diagramme de la série statistique fréquence de sinistres que la valeur zéro est largement dominante dans le portefeuille en raison du fait que l'occurrence des sinistres est généralement rare ou de la non-déclaration des bénéficiaires, en particulier dans le cas de sinistres non graves et de faible montant.

Analyse bivariée

En plus de la visualisation de la distribution de fréquence ainsi que des graphiques du nombre d'actes en fonction des caractéristiques de l'assuré, il est également important d'appliquer les tests statistiques d'indépendance dans une analyse bivariée afin d'expliquer les éventuelles relations statistiques avec la variable Fréquence des sinistres.

Tout d'abord, il est intéressant de voir quelles sont les variables disponibles qui ont un lien statistique ou non avec l'occurrence des sinistres, et donc avec la variable "Sinistre" que nous avons créé.

Variable	P-Valeur
Tanche d'âge	0.3538
Statut	0.03383
Situation matrimoniale	0.1183
Sexe	0.2129
Classe	0.8123

Table IV.6: Test d'indépendance Khi-deux avec la variable Sinistre

A partir des résultats du test de Khi-2, on remarque que seulement le caractère statut de l'assuré a un lien statistique avec l'occurrence des sinistres vu que la p-value obtenu est inférieur au seuil de 5% ce qui implique le rejet de l'hypothèse de l'indépendance. Pour les autres variables les valeurs obtenus ne permettent pas de prouver la non indépendance avec l'occurrence des sinistres, cependant, elles peuvent avoir un lien avec la variable nombre de sinistres.

Variable	P-Valeur
Tanche d'âge	<2.2e-16
Statut	<2.2e-16
Situation matrimoniale	6.56e-11
Sexe	0.006
Classe	2.2e-16

Table IV.7: Test d'indépendance Khi-deux avec la variable Nombre de sinistres

Pour toutes les variables explicatives on a obtenu une p-valeur très faible ce qui montre qu'il une certaine interaction avec la variable Fréquence vu que l'hypothèse d'indépendance est rejeté.

IV.2.6 Analyse du coût

Notre étude se concentre maintenant sur la consommation de l'assuré, c'est-à-dire que nous nous intéressons à la variable "Montant des sinistres" qui représente le coût réel des remboursements effectués par l'assurance. L'objectif de la deuxième partie du problème de tarification est de modéliser le remboursement moyen d'un sinistre. La variable "Montant des sinistres" dont nous disposons indique le coût total de tous les actes consommés par l'assuré, nous devons donc créer une nouvelle variable "coût moyen" en divisant par le nombre d'actes consommés.

Lors de la modélisation du coût, il n'est plus utile d'avoir des valeurs nulles dues à des observations sans sinistres, il est donc conseillé de les éliminer afin de ne pas se tromper dans les résultats.

Avant de s'intéresser au coût moyen il est aussi important d'avoir une idée sur la somme globale remboursée par l'assurance et les taux de couverture pour chaque poste de garantie.

Poste de garantie	Charge totale remboursée
Dentaire	49626.4
Hospitalisation	183884.8
Optique	40183.5
Pharmacie	285454
Soins courants	261440.4

Table IV.8: Somme totale remboursée par l'assurance

Taux de couverture

Le taux de couverture est la capacité d'un assureur à couvrir le montant demandé par un assuré. Il s'agit d'un pourcentage de remboursement allant de 0 à 100 % du montant réclamé. Plusieurs facteurs peuvent influencer la décision du taux de couverture pour chaque sinistre. La valeur du taux de couverture peut également être indiquée au moment de la souscription du contrat.

$$\text{Taux couverture} = \text{montant remboursé} \,/\, \text{montant demandé (frais réels)}$$

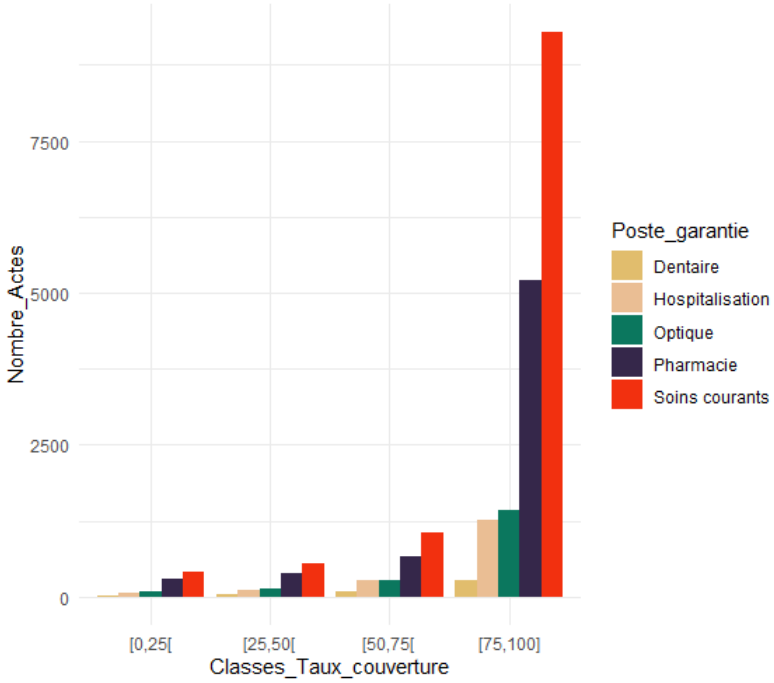


Figure IV.5: Taux de couverture par poste de garantie

La figure ci-dessus IV.5 montre la distribution du taux par taux de couverture. On peut voir que la majorité des actes ont eu une couverture de plus de 75%. Les actes non remboursées ou remboursées avec moins de 25% du montant demandé sont une minorité dans le portefeuille.

Distribution coût moyen

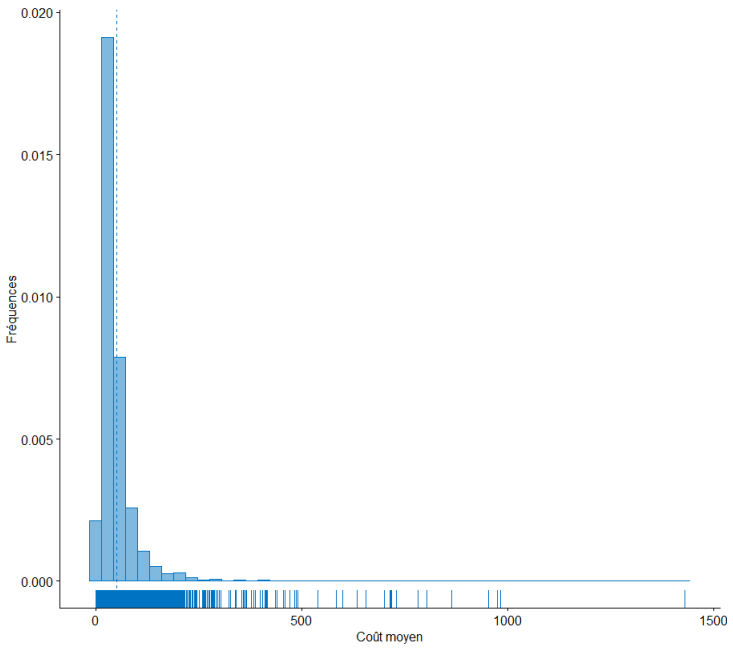


Figure IV.6: Distribution du coût moyen

Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
0,185	25,628	36,687	50,516	56,084	1428

Table IV.9: Statistiques descriptive du coût moyen

Il ressort de l'histogramme du coût moyen IV.6 et des valeurs du tableau IV.9 que la variable présente une forte asymétrie vers la droite en raison de la présence de coûts élevés qui atteignent 1428 Dt. La présence de ces coûts est attestée par le fait que dans chaque portefeuille d'assurance il y a des sinistres graves mais ils sont très rares. La majorité des données est centrée autour de 36 Dt avec une moyenne égale à 50 Dt.

Distribution coût moyen par poste de garantie

Pour montrer l'importance d'une analyse par poste de garantie on dessine l'histogramme du coût moyen pour les différents postes.

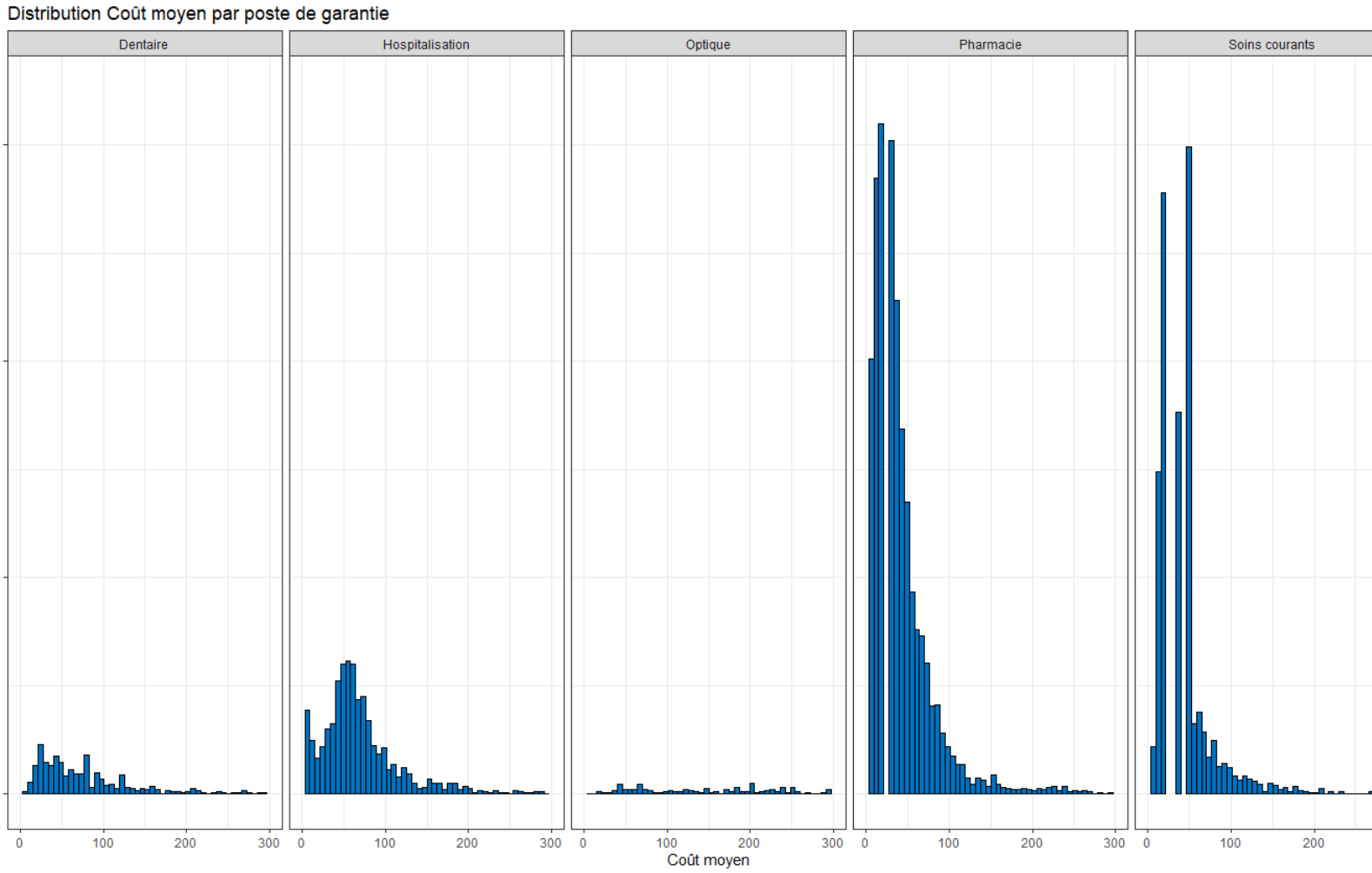


Figure IV.7: Distribution du coût moyen par poste de garantie

Le graphique montre qu'il y a une différence entre les formes des histogrammes des postes de garantie, ce qui montre une différence dans la structure des coûts entre les différentes catégories.

Analyse en fonction des variables explicatives

Coût moyen en fonction de la variable Classe

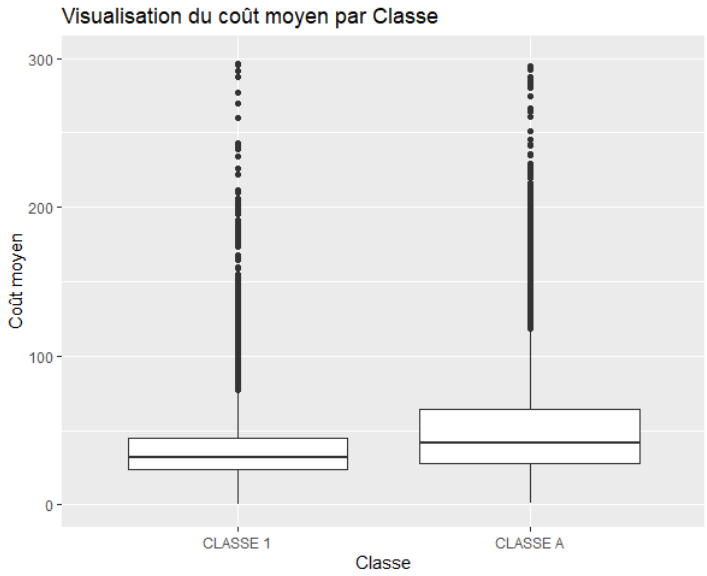


Figure IV.8: Répartition selon la variable classe

Il ressort de la boite à moustaches que les coûts pour la classe 1 sont moins réparties que les coûts de la classe A. On note également que la valeur de la médiane de la classe A est supérieur à celle de la classe 1.

Coût moyen en fonction de la variable Situation matrimoniale

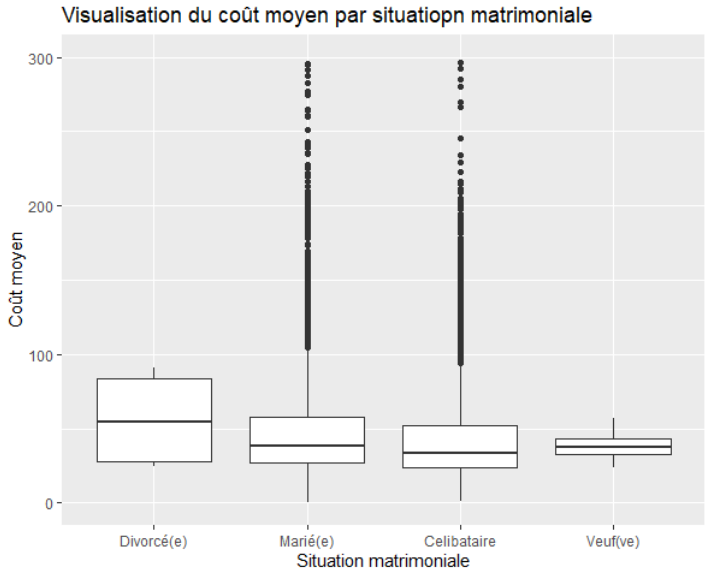


Figure IV.9: Répartition selon la situation matrimoniale

Sur ce graphique, on remarque que les coûts sont plus élevés et plus répartis au fur et à mesure que l'on se déplace de la droite du graphique vers la gauche. Ceci montre une différence de la structure des coûts entre les modalités de la variable situation matrimoniale.

Coût moyen par statut et sexe

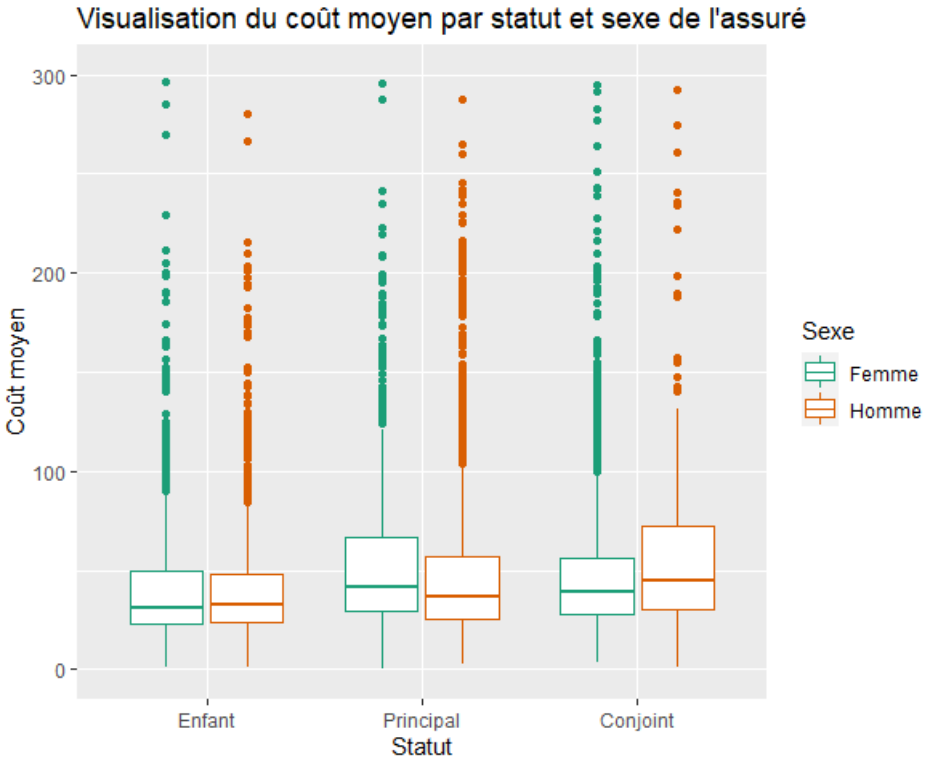


Figure IV.10: Répartition selon le statut et le sexe de l'assuré

La longueur des boites à moustaches est comparable dans ce cas, ce qui montre une similarité dans la distribution. Nous déduisons également que pour le groupe des hommes assurés principaux et le groupe des femmes conjointes, les coûts sont centrés autour d'une valeur plus élevée que les autres groupes.

Coût moyen par tranche d'âge et statut

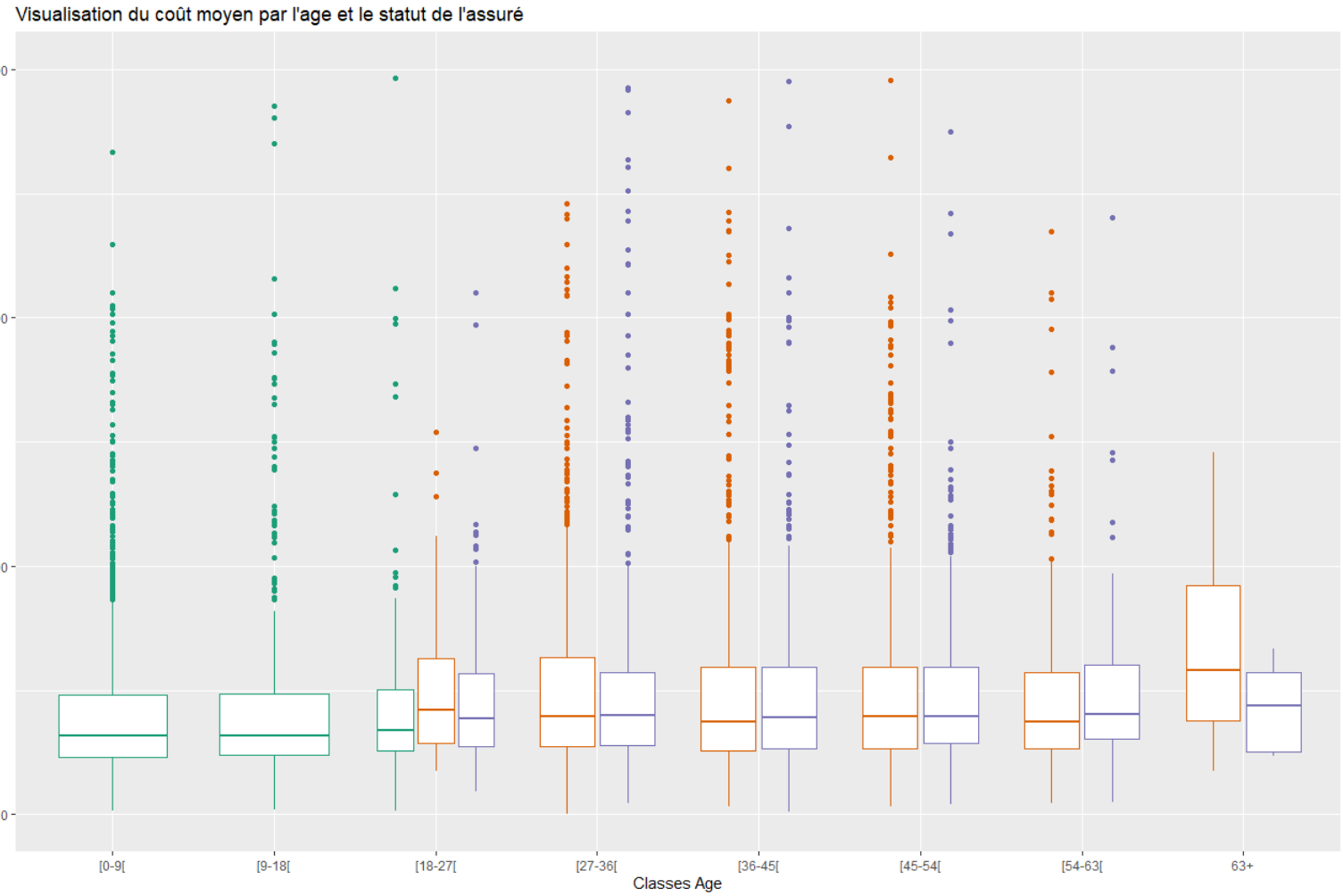


Figure IV.11: Répartition selon la tranche d'âge et le statut de l'assuré

L'information frappante de ce graphique est que le groupe des assurés principaux de plus de 63 ans a une structure de coûts plus distribuée et une valeur médiane plus élevée que tous les autres groupes. La catégorie des plus de 63 ans ne présente pas de valeurs de coûts extrêmes.

IV.2.7 Échantillonage

Une étape importante qui précède la modélisation est l'échantillonnage. Il consiste à diviser les données en trois sous-tableaux indépendants : un tableau d'apprentissage, un tableau de validation et un tableau de test.

En général, on garde 70% ou plus des données pour l'échantillon d'apprentissage puisqu'il sera utilisé pour construire le modèle et estimer les paramètres.

Les 30% restants sont répartis entre l'échantillon de validation et l'échantillon de test. Le premier est utilisé pour optimiser les paramètres déjà obtenus après la phase d'apprentissage, le second permet

de tester la performance du modèle.

Dans ce qui suit, nous ne nous intéressons qu'aux données relatives à la catégorie "Soins courants". Nous avons collecté 4820 observations, chacune représentant un assuré ayant un nombre donné d'actes en soins courants. Ce tableau sera utilisé pour la modélisation du coût moyen alors que pour la modélisation de la fréquence nous devons le concaténer avec les observations n'ayant aucun sinistre.

Pour cela nous avons récupéré 40,7% des données sans sinistres des données initiales afin de garder la même structure vue que les actes en soins courants présents également 40,7% des actes.

Enfin, nous obtenons 35882 observations pour la table « Soins Courants », dont 31062 n'ont aucune réclamation, ce qui équivaut à 86.5% des données.

La répartition de ce tableau est la suivante :

Table	Nombre d'observations
Train	25046 (70%)
Validation	7106 (20%)
Test	3517 (10%)

Table IV.10: Échantillonage des données Soins courants

IV.3 Modélisation de la fréquence

IV.3.1 Modèles linéaires généralisés

Dans cette section, nous nous concentrons sur l'application du modèle linéaire généralisé (GLM) à notre exemple de données d'assurance maladie.

Comme indiqué dans la partie théorique, la construction des modèles linéaires généralisés repose sur le choix de trois éléments clés : La distribution de la variable à prédire, les variables explicatives (le prédicteur linéaire) et la fonction lien.

Choix de la distribution

Nous avons détaillé dans la partie théorique que la loi de poisson est la plus adéquate pour ajuster les variables de comptage et qu'elle est principalement choisie pour la modélisation de la fréquence dans le problème de la tarification avec des corrections possibles selon les cas.

Pour cela, nous étudions le comportement de la fréquence de sinistre dans la catégorie Soins courants et nous effectuons les techniques nécessaires pour vérifier l'ajustement de la loi de poisson ainsi que la loi Binomiale négative.

Ajustement par la loi de Poisson

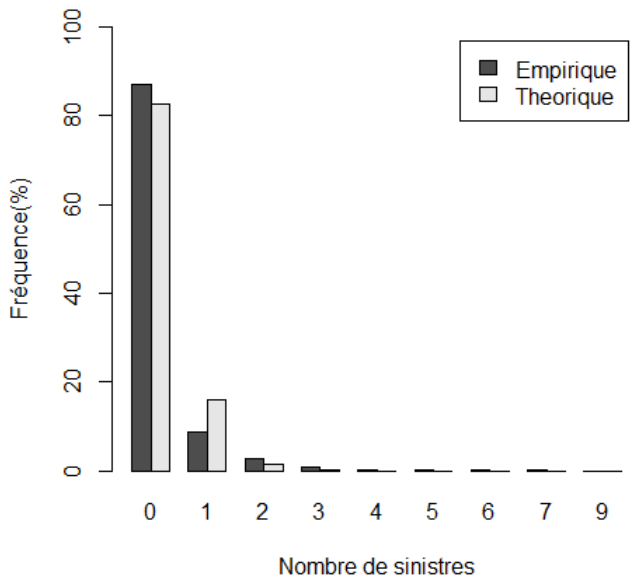


Figure IV.12: Ajustement par la loi de Poisson

Variable\Nombre sinistres	0	1	2	3	4	5	6	7	8	9
Observé	23162	2317	713	244	80	39	8	8	0	1
Estimé	21920	4211	404	25	1,2	4,7	1,5	4,2	1	2,1

Il est clair qu'il y a une différence significative entre les valeurs prédites par le modèle poissonien et les valeurs observées. On peut vérifier ces résultats par le recours aux tests d'ajustement.

Test d'ajustement Khi-deux

Statistique du test	Degré de liberté	p-valeur
151962	8	<2.2e-16

Table IV.11: Test d'ajustement Khi-deux pour la loi de poisson

Les résultats du test montrent une -valeur inférieure à 0.05, nous rejetons donc l'hypothèse nulle du test indiquant que la fréquence suit une distribution en poisson.
Ce résultat est très probable puisque la distribution en poisson requiert l'hypothèse d'égalité entre la moyenne et la variance, alors que dans notre cas, nous avons une légère sur-dispersion de la variable de fréquence puisque la variance égale à 0,754 est supérieure à la moyenne qui est égale à 0,192.

Ajustement par la loi Binomiale négative

Nous avons vu dans la partie théorique que la distribution Binomiale négative peut-être une alternative à la distribution de poisson dans le cas de la sur-dispersion puisqu'elle utilise un terme d'hétérogénéité pour résoudre le problème.
Pour ce faire, nous allons étudier si la distribution Binomiale négative s'adapte à notre échantillon de données.

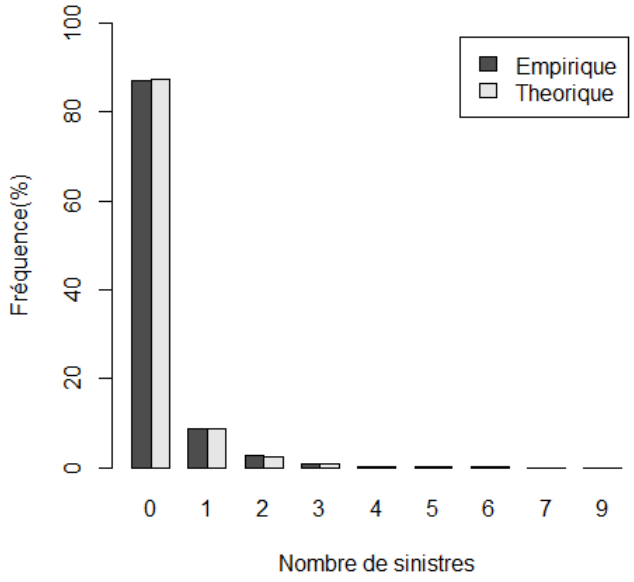


Figure IV.13: Ajustement par la loi Binomiale négative

Nous remarquons sur le graphique de la distribution théorique et empirique que la distribution binomiale négative s'adapte parfaitement aux données de la fréquence des sinistres.

Nous confirmons ce résultat avec le test d'adéquation du Khi-deux :

Test d'ajustement Khi-deux

Statistique du test	Degré de liberté	p-valeur
9.2578	8	0.321

Table IV.12: Test d'ajustement Khi-deux pour la loi Binomiale négative

Choix du prédicteur linéaire

Le prédicteur linéaire (les variables explicatives) présente la composante déterministe dans un modèle GLM, pour cela, l'étape de sélection des variables doit être bien étudiée en raison de son impact direct sur les résultats. En effet, le choix de variables non appropriées au cas étudié peut influencer négativement la performance du modèle et ainsi fausser les prédictions obtenues.

Pour les données d'assurance maladie, il n'y a généralement pas assez de données utiles pour effectuer une analyse détaillée de la sélection des variables. En effet, seules quelques caractéristiques des assurés sont disponibles. Il est clair qu'une grande partie des données ne peut être utilisée car les dates de survenue des sinistres sont antérieures à la disponibilité de ces informations et il n'est pas approprié d'expliquer ces sinistres avec des informations obtenues ultérieurement.

Malgré l'inconvénient du petit nombre de variables disponibles, nous avons essayé d'optimiser autant que possible notre sélection de variables.

Pour la sélection des variables à introduire dans le modèle, nous utilisons le test du rapport des vraisemblances (LRT). Nous avons détaillé dans la partie théorique que le test LRT compare deux modèles emboîtés, c'est-à-dire un modèle complet incluant les variables explicatives et un modèle imbriqué qui n'inclut qu'un groupe de variables utilisées par le premier.

Nous allons utiliser ce test dans notre procédure de sélection, nous comparons un modèle complet contenant toutes les variables explicatives avec un modèle contenant les mêmes variables sauf une qui est exclue. Les résultats du test nous diront si le petit modèle est significativement différent du modèle complet en termes de déviance. Si c'est le cas, cela signifie que l'existence de la variable dans le modèle est importante.

En testant des modèles GLM avec la loi de Poisson on obtient :

Variable	P-valeur
Sexe	0 ,098
Tranche âge	4,40 e-09
Statut	0,207
Situation matrimoniale	2,9 e-05
Classe	<2,2 e-16

D'après les p-valeurs, nous pouvons voir qu'en éliminant la variable sexe ou la variable statut, nous n'avons aucune preuve que le modèle contenant ces variables est meilleur. Cependant, en éliminant les deux variables en même temps et en refaisant le test, nous constatons qu'il existe une différence significative avec le modèle complet, ce qui s'explique par le fait que les deux variables ne conduisent pas à une amélioration individuelle mais que leur présence ensemble est importante. La cause en est la présence d'une interaction entre les deux variables, ce qui est confirmé par le graphique de Répartition des assurés selon le sexe et le statut dans le portefeuille IV.1. Pour modéliser cette interaction, nous ajouterons au modèle GLM une variable qui croise le statut et le sexe.

Variables retenues pour la modélisation de la fréquence par GLM-Poisson
Sexe
Statut
Tranche âge
Classe
Situation matrimoniale
Statut*Sexe : Variable qui croise sexe et statut

Table IV.13: Les variables retenues pour la modélisation de la fréquence par GLM-Poisson

Pour la modélisation par loi Binomiale négative nous utilisons une autre méthode de sélection de variables qui est la procédure de sélection stepwise en employant la méthode d'élimination backward (descendante) qui commence par un modèle saturé contenant toutes les variables et élimine une par une pour optimiser la performance du modèle selon un certain critère. Généralement on utilise le critère d'information d'Akaike (AIC).

GLM – Poisson

Malgré le fait qu'une sur-dispersion existe dans les données, nous appliquons d'abord le modèle Poissonien car c'est le modèle trivial pour modéliser des variables discrètes. Nous effectuerons également un test de sur-dispersion sur le modèle résultant.

L'application du modèle nous donne finalement les résultats suivants :

```
Call:
glm(formula = Nombre_actes_poste ~ Sexe + Statut + Classes_age +
    Classe + Situation_matrimoniale + Statut * Sexe, family = poisson(link = log),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0008  -0.6577  -0.6103  -0.5661   6.0434

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.792135   0.582077  -3.079  0.00208 **
SexeHomme                        0.037677   0.048396   0.779  0.43627
StatutPrincipal                  0.082171   0.122405   0.671  0.50203
StatutConjoint                   0.171161   0.121272   1.411  0.15813
Classes_age[9-18[               -0.187338   0.063429  -2.954  0.00314 **
Classes_age[18-27[              -0.069104   0.082286  -0.840  0.40102
Classes_age[27-36[              -0.127245   0.118915  -1.070  0.28460
Classes_age[36-45[              -0.251056   0.120622  -2.081  0.03740 *
Classes_age[45-54[              -0.264548   0.121826  -2.172  0.02989 *
Classes_age[54-63[              -0.002574   0.127072  -0.020  0.98384
Classes_age63+                   0.611407   0.194415   3.145  0.00166 **
ClasseCLASSE A                  -0.277208   0.028759  -9.639  < 2e-16 ***
Situation_matrimonialeMarié(e)   0.479390   0.578565   0.829  0.40734
Situation_matrimonialeCelibataire  0.260976   0.580831   0.449  0.65320
Situation_matrimonialeVeuf(ve)   1.976435   0.767483   2.575  0.01002 *
SexeHomme:StatutPrincipal       -0.110024   0.070083  -1.570  0.11643
SexeHomme:StatutConjoint        -0.342789   0.103721  -3.305  0.00095 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 21537  on 25116  degrees of freedom
Residual deviance: 21343  on 25100  degrees of freedom
AIC: 29036

Number of Fisher Scoring iterations: 6
```

Dans le tableau obtenu dans les résultats on s'interesse aux valeurs obtenus dans le tableau des

coefficients. Dans ce tableau on a 4 colonnes Etimate, std Error , z value et Pr(>|z|).

- **Estimate** : C'est la colonne des coefficients estimés par le modèle.
- **Std error** : C'est l'écart-type de l'estimation du coefficient dans le modèle. Il s'agit d'une mesure de l'incertitude concernant cette estimation, si elle est trop grande, cela signifie que l'estimation du point de coefficient a été calculée avec beaucoup d'imprécision.
- **Pr(>|z|)** : la p-valeur du test permettant de déterminer si l'estimation ponctuelle du coefficient est significativement différente de 0.

Les résultats affiche également les valeurs de la déviance résiduelle et Déviance nulle.

La valeur **Déviance nulle** présente la différence en déviance entre le modèle saturé (un modèle qui suppose que chaque point de données a ses propres paramètres) et le modèle nul (un modèle qui contient seulement la valeur de l'intercept).

La valeur **Déviance résiduelle** présente la différence en déviance entre le modèle saturé et le modèle proposé.

Plus la déviance nulle est petite, plus le modèle nul explique assez bien les données. De même pour la déviance résiduelle par rapport au modèle proposé.

Les variables significatives :

Les variables ayant des * dans le tableau sont les variables significatives du modèle, c'est-à-dire elles ont eu une p-valeur < 0.05, ce qui montre que leurs coefficients sont significativement non nuls.

On rappelle que la fonction lien g choisi est la fonction logarithme donc on détermine la valeur prédite de la fréquence par :

$$\mathrm{E}[N] = g^{-1}\left(x_0 + x_1\beta_1 + \cdots + x_J\beta_J\right) = \exp\left(x_0 + x_1\beta_1 + \cdots + x_J\beta_J\right)$$

Avec N : Nombre de sinistres (fréquence), les x_i sont les valeurs des variables explicatives et les β_i les coefficients estimés.

Une fois les coefficients estimés et connus, nous pouvons interpréter l'importance des variables sur le nombre de sinistres. Prenons l'exemple du coefficient significatif -0.1873 qui correspond à la variable "Classe d'âge [9,18[" il s'explique par le fait que si l'assuré appartient à cette classe d'âge alors la valeur de cette variable est égale à 1 (puisqu'il s'agit d'une variable indicatrice) et le nombre de sinistres augmente de $\exp(-0.1873) = 0.829$

Test de significativité globale

Le modèle indique que la majorité des variables sont significatives, cependant, pour confirmer la significativité de notre modèle nous allons le comparer au modèle nul, qui ne contient aucune variable explicative, c'est a dire on estime qu'un seul paramètre qui est la constante (Intercept).

La confirmation du résultat est indiqué par la réalisation d'un test de Khi-2 ayant les données suivante :

H0 : Les prédictions du modèle proposé sont proches du modèle nul.

Statistique du Test : [Déviance(modèle nul) - Déviance(moèle proposé)]

Degré de liberté : [Ddl(moèle nul) - Ddl(moèle proposé)]

Les résultats du test, présente une valeur de p égale à **6.880498e-32** $\approx$ **0**, ce qui montre que le modèle est globalement significatif.

Détection de la sur-dispersion :

Pour détecter la sur-dispersion, nous utilisons la fonction *dispersiontest* du package **AER**.
La fonction réalise un test sur le coefficient de dispersion alpha

$$VAR[y] = (1 + \text{ alpha })^{*}mu = \text{ dispersion }^{\star}mu.$$

Si alpha est >0 c'est-à-dire la dispersion > 1 alors une sur-dispersion est prouvée, et c'est le cas de notre modèle.

```
Overdispersion test

data:  step.modelpoisson
z = 20.22, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.760661
```

Les résultats du test montre bien l'existence d'une sur-dispersion avec un coefficient de sur-dispersion égale à 1.76.

GLM – Binomiale négative

Un modèle avec la loi Binomiale négative peut être une approche alternative pour modéliser la sur-dispersion dans les données de comptage, vu qu'il ajoute un effet aléatoire au modèle poissonien pour représenter l'hétérogénéité non observée.

En appliquant la procédure de sélection par étapes, à savoir la méthode descendante (backward),nous commençons par un modèle saturé contenant toutes les variables et éliminons à chaque itération la variable qui diminue le critère de sélection (nous avons choisi l'AIC dans notre cas) jusqu'à ce que nous atteignions une combinaison de variables où il n'y a plus d'amélioration du critère.

```
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)                          -0.84511    0.25277  -3.343 0.000827 ***
‘Classes_age_[0-9[‘                  -0.70067    0.26729  -2.621 0.008757 **
‘Classes_age_[9-18[‘                 -0.88026    0.27480  -3.203 0.001359 **
‘Classes_age_[18-27[‘                -0.71472    0.26870  -2.660 0.007816 **
‘Classes_age_[27-36[‘                -0.72372    0.25531  -2.835 0.004587 **
‘Classes_age_[36-45[‘                -0.86975    0.25504  -3.410 0.000649 ***
‘Classes_age_[45-54[‘                -0.88147    0.25565  -3.448 0.000565 ***
‘Classes_age_[54-63[‘                -0.62420    0.26013  -2.400 0.016413 *
Statut_Principal                     -0.09556    0.04936  -1.936 0.052884 .
Situation_matrimoniale_Celibataire -0.24833    0.07190  -3.454 0.000553 ***
‘Classe_CLASSE 1‘                     0.28278    0.03862   7.322 2.44e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.2463) family taken to be 1)

    Null deviance: 10811  on 25116  degrees of freedom
Residual deviance: 10714  on 25106  degrees of freedom
AIC: 26241

Number of Fisher Scoring iterations: 1


              Theta:  0.24627
          Std. Err.:  0.00934

 2 x log-likelihood:  -26216.64300
```

La procédure de sélection dans le modèle Binomiale négative a éliminé beaucoup de variable et n'a gardé que 10.

Test de significativité globale

Pareil comme le modèle de poisson, on commence par tester la significativité globale du modèle proposé par rapport au modèle nul.

La valeur de p obtenue égale à **2.884457e-16 $\approx$ 0**, nous pouvons donc conclure que le modèle est globalement significatif.
Le modèle indique que la majorité des variables sont significatives, cependant, pour confirmer la significativité de notre modèle nous allons le comparer au modèle nul, qui ne contient aucune variable explicative, c'est a dire on estime qu'un seul paramètre qui est la constante (Intercept).

En comparaison avec les valeurs obtenus du modèle de poisson, le modèle Binomiale négative présente beaucoup moins de Déviance nul grace à la significativité de l'intercept et beaucoup moins de déviance résiduelle ce qui montre une meilleure explication des données. Les valeurs des indicateurs AIC et BIC sont également améliorés.

Modèle	Déviance Nulle	Déviance Résiduelle	AIC	BIC
Poisson	21537	21355	29036.34	29174.57
Binomiale négative	10811	10714	26240.64	26338.22

Table IV.14: Comparaison modèle Poisson et modèle Binomiale négative

Modèles à inflation de zéro

Comme nous l'avons mentionné dans la partie théorique, les modèles à inflation de zéro peuvent être une alternative à un simple modèle poisson pour prendre en compte le phénomène de sur-dispersion. Nous avons vu dans le graphique de l'ajustement de la loi poisson IV.13 que les zéros observés dépassent les zéros estimés par l'ajustement poisson, ce qui montre l'existence d'un excès de zéros. Cet excès peut être résolut par l'application des modèles à inflation de zéro.
Dans cette section, nous allons modéliser le nombre de sinistres par les deux modèles les plus utilisés qui sont le modèle ZIP (zero inflated poisson) et le modèle binomial négatif zero inflated.

La fonction *zeroinfl* du package *pscl* du langage R, permettent d'obtenir les résultats suivants :

Zero Inflated Poisson : ZIP

```
Call:
zeroinfl(formula = Nombre_actes_poste ~ Sexe + Classes_age + Statut + Situation_matrimonia
    Classe, data = train_freq, dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5111 -0.3762 -0.3343 -0.3160 12.2865

Count model coefficients (poisson with log link):
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.28394    0.95784  -0.296  0.76689
SexeHomme                       -0.12808    0.05508  -2.325  0.02006 *
Classes_age[9-18[               -0.41656    0.13491  -3.088  0.00202 **
Classes_age[18-27[              -0.15069    0.15874  -0.949  0.34247
Classes_age[27-36[              -0.22960    0.20328  -1.129  0.25870
Classes_age[36-45[              -0.36909    0.20612  -1.791  0.07335 .
Classes_age[45-54[              -0.40826    0.20767  -1.966  0.04931 *
Classes_age[54-63[              -0.09734    0.21375  -0.455  0.64882
Classes_age63+                   0.13773    0.28604   0.482  0.63016
StatutPrincipal                  0.53136    0.20187   2.632  0.00848 **
StatutConjoint                   0.50258    0.20632   2.436  0.01485 *
Situation_matrimonialeMarié(e)   0.10593    0.95519   0.111  0.91170
Situation_matrimonialeCelibataire -0.02780   0.95758  -0.029  0.97684
Situation_matrimonialeVeuf(ve)   0.19756    1.07914   0.183  0.85474
ClasseCLASSE A                  -0.01840    0.04830  -0.381  0.70328

Zero-inflation model coefficients (binomial with logit link):
```

```
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.13607    1.08934   1.043  0.29700
SexeHomme                        -0.10723    0.06910  -1.552  0.12074
Classes_age[9-18[                -0.35101    0.19746  -1.778  0.07547 .
Classes_age[18-27[               -0.11817    0.21362  -0.553  0.58013
Classes_age[27-36[               -0.14319    0.26463  -0.541  0.58843
Classes_age[36-45[               -0.14985    0.26775  -0.560  0.57571
Classes_age[45-54[               -0.17325    0.26991  -0.642  0.52096
Classes_age[54-63[               -0.08405    0.27685  -0.304  0.76143
Classes_age63+                   -0.52288    0.39825  -1.313  0.18921
StatutPrincipal                   0.67111    0.25935   2.588  0.00966 **
StatutConjoint                    0.53854    0.26503   2.032  0.04216 *
Situation_matrimonialeMarié(e)   -0.44930    1.08477  -0.414  0.67873
Situation_matrimonialeCelibataire -0.33367   1.08833  -0.307  0.75916
Situation_matrimonialeVeuf(ve)  -13.69363  240.18356  -0.057  0.95453
ClasseCLASSE A                    0.35331    0.05930   5.958 2.55e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 37
Log-likelihood: -1.312e+04 on 30 Df
```

Zero Inflated Negatif Binomial : ZINB

Remarque : Les détails des résultats du code du modèle ZINB sont affichés en Annexe.

Les variables significatives obtenus :

Processus 1 : Modèle Binomiale négative		Processus 2 : Régression logistique	
Variables significatives	Coefficients	Variables significatives	Coefficients
Classes_age[9-18[	-0.5	StatutPrincipal	1.06
StatutPrincipal	0.609	ClasseCLASSE A	0.458
StatutConjoint	0.577		

On remarque que pour les deux cas (ZIP ou ZINB) on a deux processus qui sont exécutés, le premier pour la modélisation du nombre de sinistres sans les zéros, ie, on ne modélise que sur les données sinistrés de la table d'apprentissage. Le deuxième correspond à une régression logistique, ie, une régression binaire pour prédire la survenance ou non d'un sinistre. Cette distinction des processus permet de mieux modéliser le nombre de sinistre et résout le problème d'excès de zéros non ajusté par la distribution de poisson.

Le fonctionnement et l'interprétation du premier processus est le même que les modèles déjà réalisés, il s'agit de la génération de valeurs estimées par une distribution de Poisson ou Binomiale Négative, cependant, le deuxième processus est complètement différent puisque le fonctionnement est basé sur

l'utilisation de la loi de Bernoulli pour estimer la probabilité de non sinistralité.

Remarque : Dans notre cas on a utilisé les mêmes variables explicatives pour les deux processus, cependant, ces derniers peuvent avoir différentes variables explicatives.

On a vu dans la partie théorique que l'espérance de la variable réponse dans le modèle ZIP ou ZINB et donnée par :

$$E\left(Y\right) = \left(1 - \pi\right)\mu$$

Avec π, la probabilité d'avoir 0 sinistre, et μ, l'espérance de la variable réponse de l'autre processus (Poisson ou Binomiale négative sur les données sinistrés).

Pour une régression logistique on estime π par :

$$\hat{\pi} = \frac{\exp\left\{\hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \cdots + \hat{\gamma}_q x_q\right\}}{1 + \exp\left\{\hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \cdots + \hat{\gamma}_q x_q\right\}}$$

et l'estimation de dans les deux cas poisson ou binomiale négative est donnée par :

$$\hat{\mu} = \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p\right\}.$$

Le résultat de l'estimation sera :

$$\widehat{E(Y)} = (1 - \hat{\pi})\hat{\mu}$$

Comparaison modèle ZIP et ZINB :

	Log-vraisemblance	Nombre de paramètres	AIC	BIC
ZIP	-13120	30	26294,07	26538,01
ZINB	-13060	31	26182,05	26434,12

Table IV.15: Comparaison des modèles ZIP et ZINB

Le modèle ZINB a une log-vraisemblance plus élevée que le modèle ZIP, ce qui montre qu'il a moins de biais. En se basant sur les critères AIC et BIC, nous pouvons conclure que la qualité du modèle ZINB est meilleure puisqu'il présente des valeurs inférieures dans ces critères.

Comparaison des modèles GLM de la fréquence

Modèle	Degré de liberté	AIC	BIC
Poisson	17	29036,34	29174,57
Binomiale négative	11	26240,64	26338,22
ZIP	30	26294,07	26538,01
ZINB	31	26182,05	26434,12

Table IV.16: Tableau comparatif des modèles GLM de la fréquence

La valeur minimale du critère AIC corresspond au modèle ZINB alors que la valeur minimale du critère BIC est associé au modèle Binomiale négative.

Le modèle linéaire généralisé Poisson sans aucune correction possède les valeurs les plus elevés en AIC et BIC, ceci montre l'importance de la correction de la sur-dispersion par le modèle binomiale négative ou par les modèles modifiés en zéro.

IV.3.2 Arbre de régression de la fréquence

La variable de fréquence des sinistres à modéliser est une variable quantitative, nous procédons donc à un arbre de régression.

En utilisant le package rpart du langage R, nous pouvons construire notre arbre en utilisant toutes les variables explicatives qui sont toutes qualitatives. Dans ce cas, nous n'avons pas besoin de les convertir en variables indicatrices.

Les principaux hyperparamètres à prendre en compte lors de la construction de l'arbre sont :

- **minsplit** : Le nombre minimum d'observations qui doivent exister dans un noeud pour effectuer la séparation.
- **minbucket** : C'est un paramètre lié au premier puisqu'il identifie le nombre minimum d'observations qui doivent exister dans chaque feuille.
- **cp** : C'est le coût de complexité, la valeur de ce paramètre est considérée comme un seuil pour les variables explicatives, en fait, on procède à un split avec une variable explicative dans un nœud seulement si la variable parvient à minimiser l'erreur par une valeur supérieure à ce seuil de complexité. Ce paramètre a un impact direct sur la taille de l'arbre. Nous l'utiliserons dans la suite pour trouver l'arbre optimal après élagage.
- **Xval** : C'est le nombre de fois que l'algorithme effectue la technique de validation croisée.

Une bonne pratique dans la construction des arbres consiste à choisir les hyperparamètres de manière à recevoir un arbre saturé pour l'optimiser parfaitement par la suite.

Nous choisissons les hyperparamètres suivants :

xval	Cp	Minbucket	Maxdepth
6	0.0001	1255 = 5% des données	20

Table IV.17: Hyperparamètres Arbre de regression - Fréquence

Dans le cas suivant, on a beaucoup réduit le coût de complexité pour avoir une arbre saturé contenant plus de variables explicatives mais on a choisit minbucket à 5 % des données pour que l'arbre soit lisible.

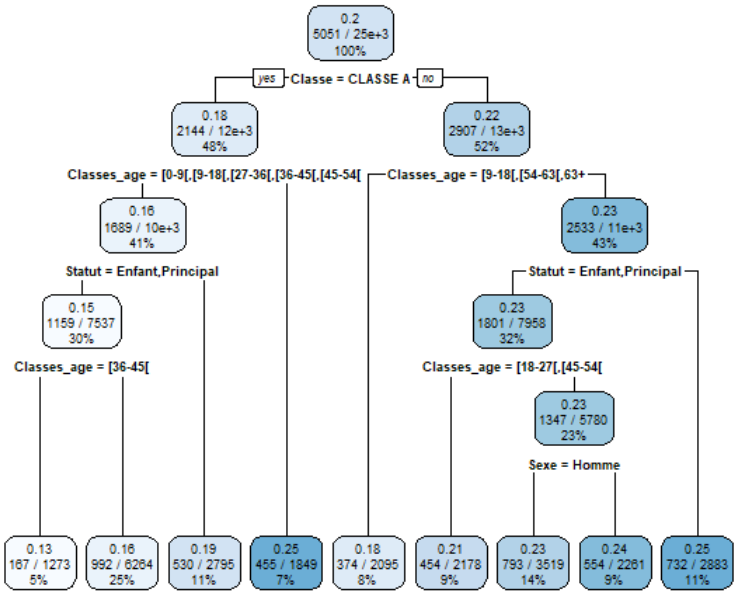


Figure IV.14: Arbre saturé Fréquence

L'arbre saturé risque d'avoir un sur-apprentissage en s'adaptant bien aux données d'apprentissage ce qui implique de mauvaises performances sur les nouvelles données.

Il est donc important d'élaguer l'arbre en augmentant le coût de complexité.

Dans le tableau suivant, nous calculons les erreurs pour chaque coût de complexité, il faut donc choisir celui qui minimise le plus l'erreur (x-error).

```
        CP nsplit rel error  xerror     xstd
1 0.0101417      0   1.00000 1.00052 0.052031
2 0.0059182      1   0.98986 1.00033 0.052000
3 0.0052431      2   0.98394 1.00278 0.052165
4 0.0039747      3   0.97870 0.99920 0.051903
5 0.0034139      4   0.97472 0.99649 0.051743
6 0.0030477      6   0.96789 0.99387 0.051656
7 0.0015072      7   0.96485 0.99027 0.051560
8 0.0010761      8   0.96334 0.99088 0.051705
9 0.0010000      9   0.96226 0.99137 0.051690
```

On constate donc que la valeur 0.0015072 donne le meilleur arbre en termes de minimisation de l'erreur.

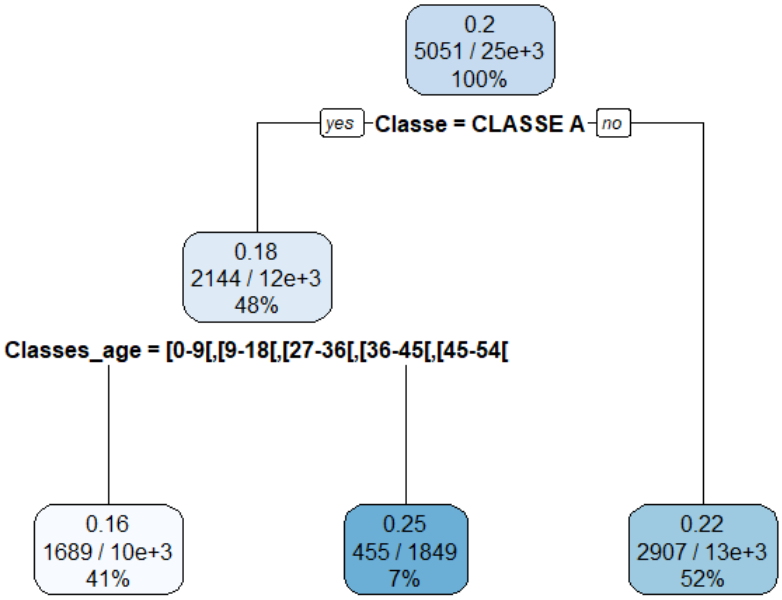


Figure IV.15: Arbre élagué Fréquence

Pour prédire la fréquence des sinistres, nous commençons par le premier nœud (la racine), qui contient tous les points de l'échantillon. Nous concluons qu'en considérant toutes les données, la fréquence moyenne est égale à 0,2. Cette estimation est notre meilleure prédiction pour les données, étant donné qu'il n'y a pas d'autres informations sur les caractéristiques de l'assuré.

En descendant dans l'arbre, le modèle commence à inclure des informations provenant des prédicteurs. L'algorithme choisit le prédicteur et le point de coupure qui réduisent les sommes des erreurs au carré dans chaque partition. En d'autres termes, une partition est effectuée en décidant quelle variable créerait les groupes les plus homogènes par rapport à la variable de résultat.
Parmi les résultats obtenus, on peut déduire de la première partition que les assurés de la classe A ont une chance de 18% de faire un sinistre pendant un an alors que ce pourcentage est égal à 22% pour les assurés de la classe 1. D'autre part, pour la catégorie de la classe A, si les assurés ont moins de 54 ans alors ils ont une chance de 16% contrairement à ceux de plus de 54 ans qui sont plus risqués avec un pourcentage de 25%.

IV.3.3 GBM – Poisson

L'algorithme Gradient Boosting Machine (GBM) est une méthode ensembliste qui utilise la technique de **Boosting** aggrégant ainsi d'une facon séquentielles les arbres, sa spécificité est le fait qu'il utilise la technique du descente de gradient pour minimiser la fonction de perte.

L'algorithme GBM, nous permet également de choisir la distribution de la variable réponse. Vu qu'on traite la variable fréquence on choisit la loi de poisson comme distribution.
L'algorithme GBM est caractérisé par son avantage à réduire les erreurs de prédiction.

Les paramètres du modèle à prendre en compte lors de l'implémentation sont :

- **Distribution** : La distrubution de la variable réponse.
- **n.trees** : Le nombre d'arbres construites.
- **interaction.depth** : Contrôle la profondeur maximale de l'arbre qui sera créé. Elle peut également être décrite comme la longueur du chemin le plus long entre la racine de l'arbre et une feuille.
- **Learning rate** : C'est le Taux d'apprentissage
- **cv.folds** : Nombre de blocs de la validation croisée

Pour le choix de ces paramètres, la technique de recherche par grille (grid search) est utilisée pour tester les valeurs suivantes pour chaque paramètre :

Learning rate	Interaction depth	ntrees
0.1, 0.01, 0.05, 0.03, 0.001	2, 3, 4, 6	50, 100, 300, 500

Table IV.18: Tuning des hyperparamètres GBM-Poisson

La combinaison optimale obtenu est la suivante :

Learning rate	Interaction depth	ntrees
0.1	3	300

Table IV.19: Combinaison optimale des hyperparamètres GBM-Poisson

Cependant, il est nécessaire de représenter l'évolution de la déviance en fonction du nombre d'arbre pour éviter le risque de sur-apprentissage.

On note que la courbe en noir est l'évolution de la déviance calculée pendant la phase d'apprentissage et la courbe en vert pendant la phase de validation dans le processus de validation croisée.

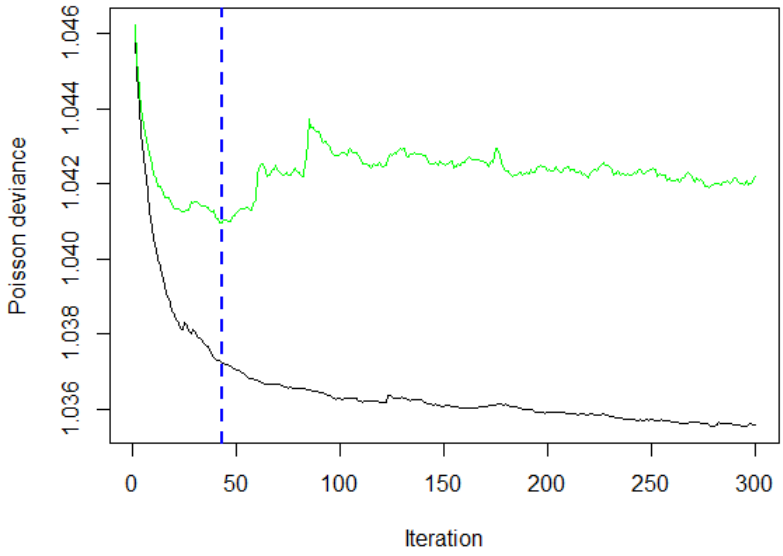


Figure IV.16: Evolution de la déviance GBM-Poisson

Le nombre d'arbres optimales est 43, une valeur supérieure à ce nombre cause le sur-apprentissage.

Interprétation

Comme les foréts aleatoires les modeles gbm présentent également le problème de la « boite boire » vu le manque de réponse à la question « Comment on a obtenu les prédictions ? ».

Pour cela, on represente le graphique de l'influence relative des variables qui est mesuré par l'augmentation en erreur de prédiction après la permutation des valeurs de la variable considéré.

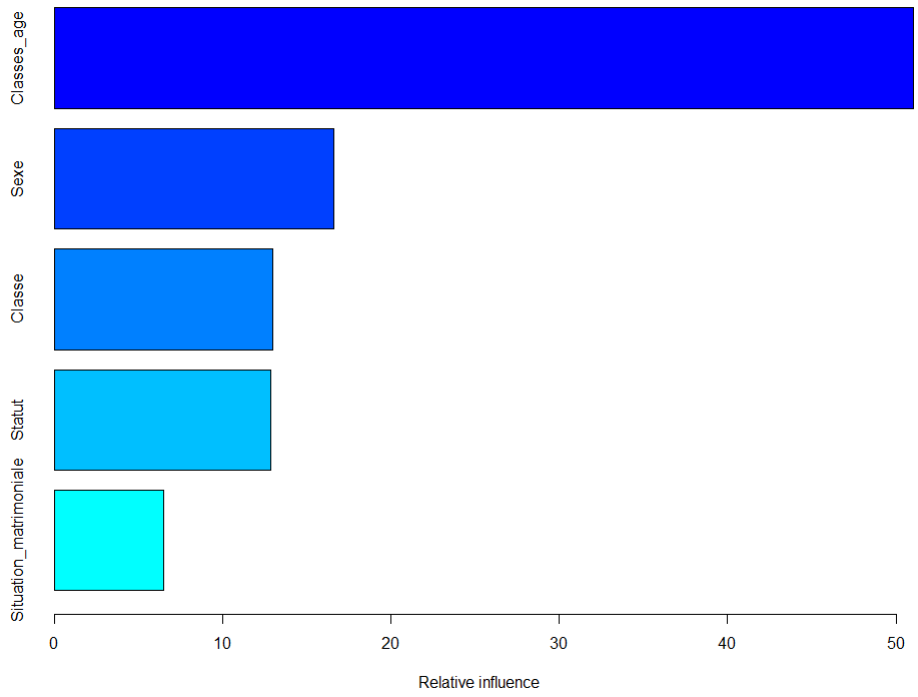


Figure IV.17: Influence relative des variables GBM-Poisson

Comparaison avec l'arbre de régression

	MAE	MSE	RMSE	Moyenne observée (Test)	Prédiction (Test)
CART	0,341	0,385	0.620	0,196	0,246
GBM	0,344	0,376	0,613	0,196	0,201

Table IV.20: Comparaison entre GBM et CART

La racine carré de l'erreur quadratique moyenne est plus faible pour le modèle GBM, cependant l'erreur absolue moyenne MAE est en faveur du modèle de l'arbre.

Le choix du meilleure modèle se fait pour le modèle GBM vu qu'il est plus robuste. La technique de boosting adapté par le modèle GBM le rend normalement plus performant, cependant le nombre réduit des variables explicatives influent négativement sur le modèle. De plus, le modèle GBM nécessite plus de travail sur le paramétrage.

IV.4 Modélisation du coût

IV.4.1 Modèles linéaires généralisés

L'objectif de cette section est d'appliquer des modèles linéaires généralisés pour modéliser le coût moyen d'un sinistre. Comme nous l'avons fait pour la modélisation des fréquences, nous devons choisir de manière appropriée les éléments de base d'un GLM, à savoir le choix de la distribution, des variables explicatives et de la fonction lien.

Choix de la distribution

La variable Coût moyen est une variable continue et positive, et les distributions les plus utilisées dans le cas la tarification sont les lois gamma et log-normale. Pour cela, nous étudions le comportement du coût dans la même catégorie d'acte : Soins courants et nous effectuons les techniques nécessaires pour vérifier l'ajustement des lois étudiés.

Dans le domaine de l'assurance, l'apparition de sinistres à coût élevé est très rare, mais leur présence dans les données crée une asymétrie dans la distribution, de sorte que les compagnies d'assurance les modélisent généralement séparément. Dans notre jeu de données, les sinistres dont le coût moyen est supérieur à 150 ne sont présents que dans 0,4% du jeu de données, nous avons donc choisi de les éliminer afin qu'ils ne perturbent pas la modélisation.

La distribution du coût moyen remboursé sera :

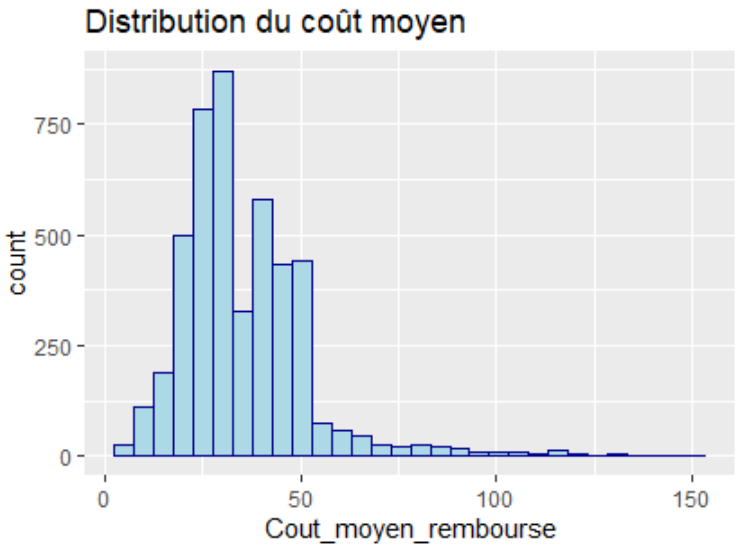


Figure IV.18: Distribution du coût moyen pour le poste Soins courants

Ajustement de la distribution Gamma

Grâce à la fonction *fitdist* du package *fitdistrplus*, on peut voir la distribution du coût moyen observé (empirique) par rapport à la distribution théorique Gamma.

La construction de la distribution théorique se réalise en choisissant les paramètres de la loi Gamma .Ces paramètres de la loi Gamma seront estimés par la fonction *fitdist*.

Les paramètres estimés de la loi :

Paramètre de forme k = 5.2431079 Paramètre d'intensité $\beta = 0.1482126$

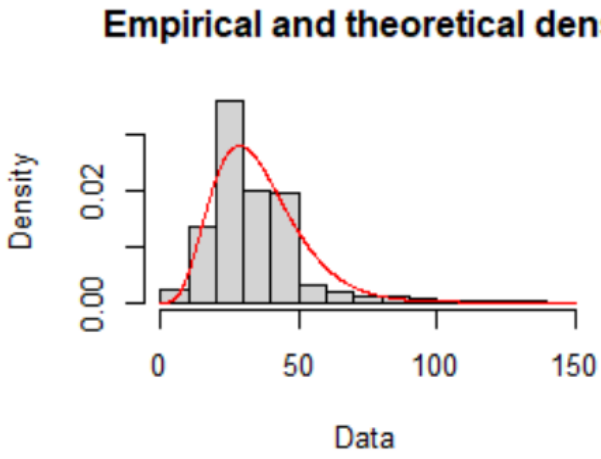


Figure IV.19: Distribution de la loi empirique et théorique - Gamma

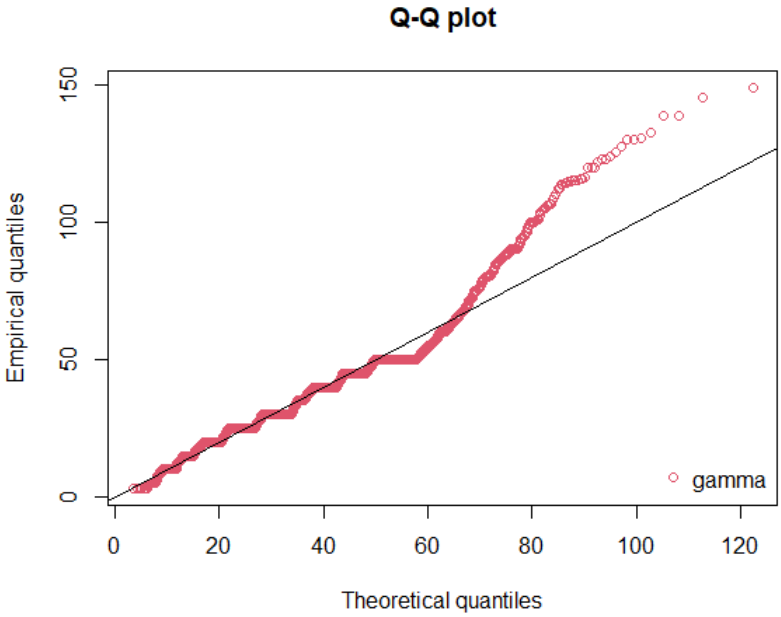


Figure IV.20: Q-Q plot pour la loi Gamma

Ajustement de la distribution Log-normale

Les paramètres estimés de la loi :

Esperance du logarithme de la variable aléatoire $\mu = 3.4677591$ k $= 5.2431079$
Ecart-type du logarithme de la variable aléatoire $\sigma = 0.4531026$

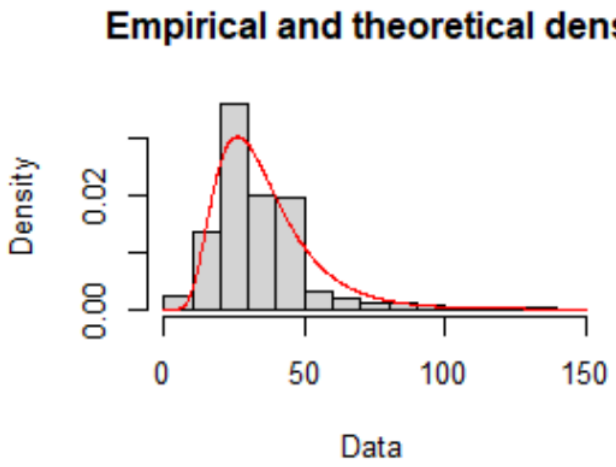


Figure IV.21: Distribution de la loi empirique et théorique - Log-Normale

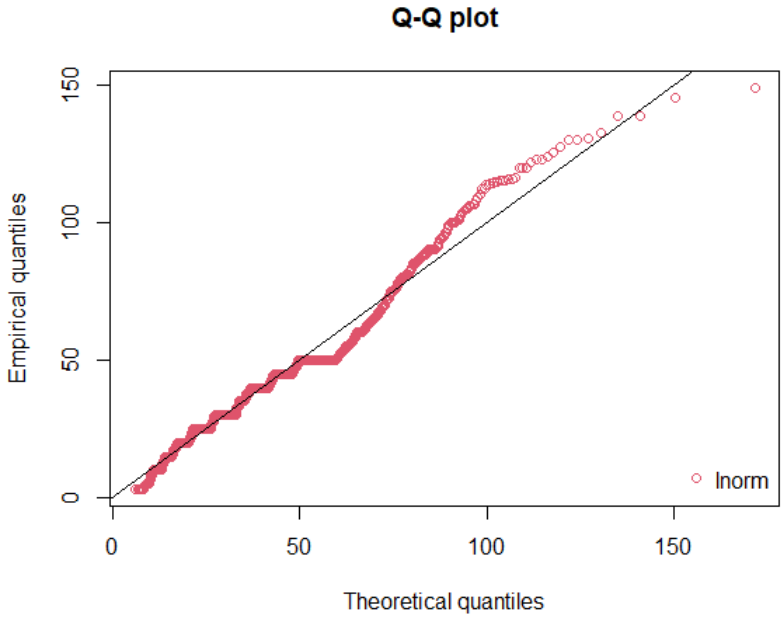


Figure IV.22: Q-Q plot pour la loi Log-Normale

A partir des graphiques de l'histogramme et la densité de la loi théorique on ne peut pas vraiment savoir qui ajuste mieux les données, cependant, on recours au le diagramme quantile-quantile (Q-Q plot).

Nous rappelons qu'en analysant le diagramma Q-Q plot on obtient un bon ajustement lorsuqye on a un alignement des points sur la droite.

Pour le diagramme de la loi Gamma les quantiles commencent par être alignés mais détachent complétement de la droite d'une façon progressive ce qui est un signe de mauvais ajustement pour les valeurs extremes du coût.

Cependant, les quantiles de la loi Log-Normale sont globalement alignés malgré le mauvais ajustement de quelques points.

Choix du prédicteur linéaire

Pour la modélisation du coût moyen on refait le même traitement fait lors de la modélisation de la fréquence, on intégre toutes les variables explicatives disponibles : Sexe, Tranche Age, Statut assuré, Classe et situation matrimoniale. La sélection des variables se fait grace à la méthode de sélection pas à pas (stepwise) en employant la méthode d'élimination descendante (backward) qui commence par un modèle saturé contenant toutes les variables et élimine une par une pour optimiser la performance du modèle selon le critère AIC.

Choix de la fonction lien

Dans la suite nous intéressons à un modèle avec la distribution Gamma, pour cela, il convient d'appliquer la fonction inverse $x \mapsto \frac{1}{x}$ comme fonction lien vu qu'elle est la plus utilisé dans ce cas.

Remarque : Dans notre cas, la structure du coût moyen est ajustée par les deux distributions Gamma et Log-Normale quasiment de la même manière, les résultats de la modélisation sont également proches. Pour cela, nous ne détaillons que la modélisation par la distribution Gamma, les coefficients obtenus par la distribution Log-Normale sont affichés en annexe.

GLM – Gamma

En appliquant le modèle GLM sur R avec l'utilisation la méthode descendante comme procédure de sélection on obtient les résultats suivants :

```
Call:
glm(formula = Cout_moyen_rembourse ~ Sexe_Femme + `Classes_age_[0-9[` +
    `Classes_age_[36-45[` + `Classes_age_[45-54[` + Statut_Enfant +
    `Classe_CLASSE 1`, family = Gamma(link = "inverse"), data = train_dummies)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.74932  -0.31075  -0.08614   0.20843   1.80416
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.0267184  0.0005206  51.324  < 2e-16 ***
Sexe_Femme            -0.0022553  0.0004519  -4.991 6.34e-07 ***
‘Classes_age_[0-9[‘   -0.0013721  0.0008520  -1.610   0.1074
‘Classes_age_[36-45[‘  0.0020519  0.0006725   3.051   0.0023 **
‘Classes_age_[45-54[‘  0.0017173  0.0006912   2.484   0.0130 *
Statut_Enfant          0.0036150  0.0008280   4.366 1.31e-05 ***
‘Classe_CLASSE 1‘      0.0023078  0.0004513   5.114 3.34e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 0.2040708)

    Null deviance: 641.76  on 3281  degrees of freedom
Residual deviance: 623.18  on 3275  degrees of freedom
AIC: 26726

Number of Fisher Scoring iterations: 5
```

Contrairement à la modélisation des fréquences, les variables Sexe et Statut sont significatives dans la modélisation des coûts, mais tous les modalités de la variable Situation matrimoniale n'ont pas été pris.

Parmi les 6 variables gardés par la procédure de sélection nous obtenons 5 variables significatives (les p-valeurs sont faibles inférieur à 5%), cela nous amène à conclure qu'il existe des preuves solides que les coefficients ne sont pas nuls.

De plus les résultats montre que l'intercept est statistiquement significative avec une valeur de la statistique du test assez élevée ce qui montre son importance dans l'explication de la variance de notre variable dépendante.

Pour confirmer la qualité d'adaptation du modèle aux données, on peut analyser la déviance.
Test d'adéquation de la déviance

Sous l'hypothèse H0 (le modèle s'adapte bien aux données), la déviance suit une Khi-deux à un certain nombre de degrés de liberté (égale à 3275 dans notre cas).

La p-valeur obtenue est proche de 1 ce qui nous amène que le modèle obtenue s'adapte bien aux données.

IV.4.2 Arbres de régression pour le coût

Pour la modélisation du coût, nous procédons de la même manière à la construction de l'arbre de régression que pour la modélisation de la fréquence. On commence par la construction de l'arbre saturé, on fait varier le coût de complexité pour élaguer l'arbre et obtenir une somme d'erreurs plus faible.

L'arbre optimale après élagage est le suivant :

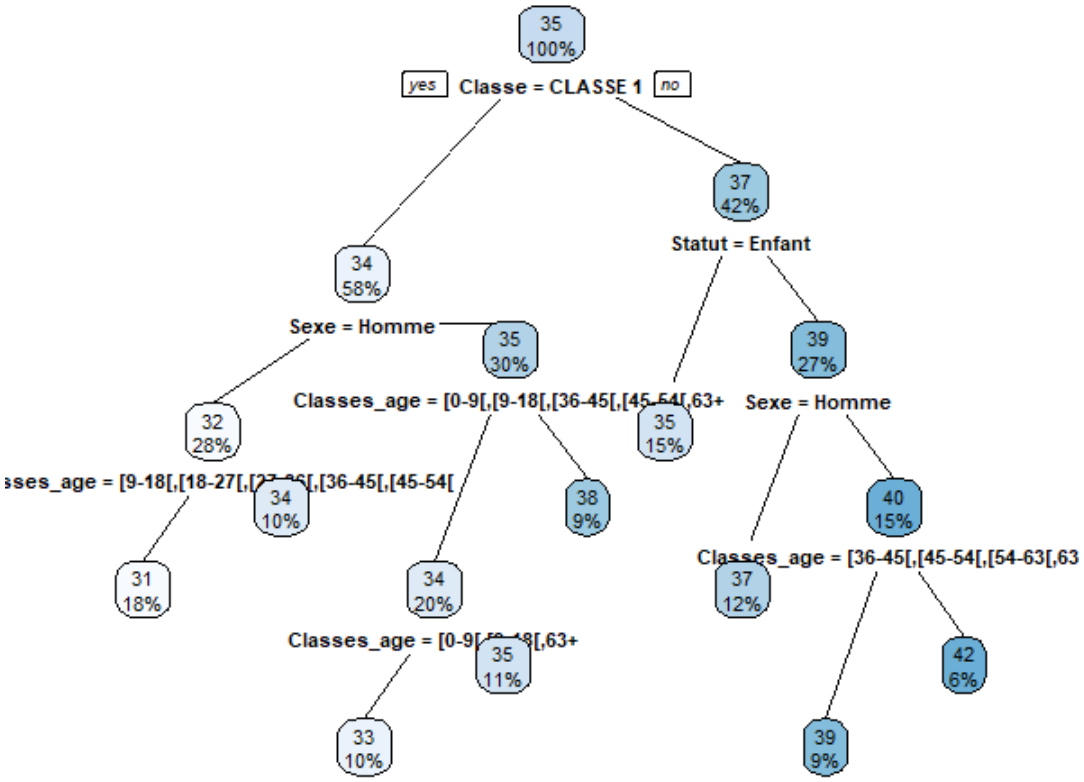


Figure IV.23: Arbre de régression après élagage pour le coût

L'arbre des fréquences nous a permis de conclure que les assurés appartenant aux compagnies de la classe 1 ont généralement une fréquence de sinistres plus élevée que ceux de la classe A, tandis que l'arbre des coûts montre que les compagnies de la classe 1 ont un coût moyen des sinistres plus faible, ce qui est attendu puisque les sinistres les moins chers sont les plus fréquents.

Le coût moyen estimé le plus élevé par l'arbre est 42Dt qui est associé au groupe d'assurés appartenant à des compagnies de classe A, ayant le statut de principal ou de conjoint, de sexe féminin et âgés de 18 à 36 ans ou de plus de 63 ans.

IV.4.3 Forêts aléatoires pour le coût

Nous avons vu dans la partie théorique que l'algorithme des forêts aléatoires est une méthode d'agrégation d'arbres qui se distingue par la sélection d'un nombre déterminé de variables explicatives dans l'étape des divisions à chaque nœud. Étant une technique basée sur l'agrégation (bagging), les forets aléatoires peuvent réduire la variance des arbres de régression qui sont faits et ainsi optimiser la qualité de la prédiction.

Les principaux paramètres de tuning de l'algorithme sont les suivants :

- **ntree :** c'est le nombre d'arbres à réaliser lors de la construction du modèle final du foret alatoire.
- **nodesize :** La taille minimale des feuilles.
- **mtry :** Le nombre de variables sélectionnés aléatoirement pour être testées à chaque division lors de la construction des arbres.

Une première étape lors de la construction de la forêt est de déterminer le nombre d'arbres dans le modèle pour éviter le sur-apprentissage. Pour ce faire, nous commençons par un modèle contenant 200 arbres et nous visualisons la convergence du taux d'erreur en fonction du nombre d'arbres.

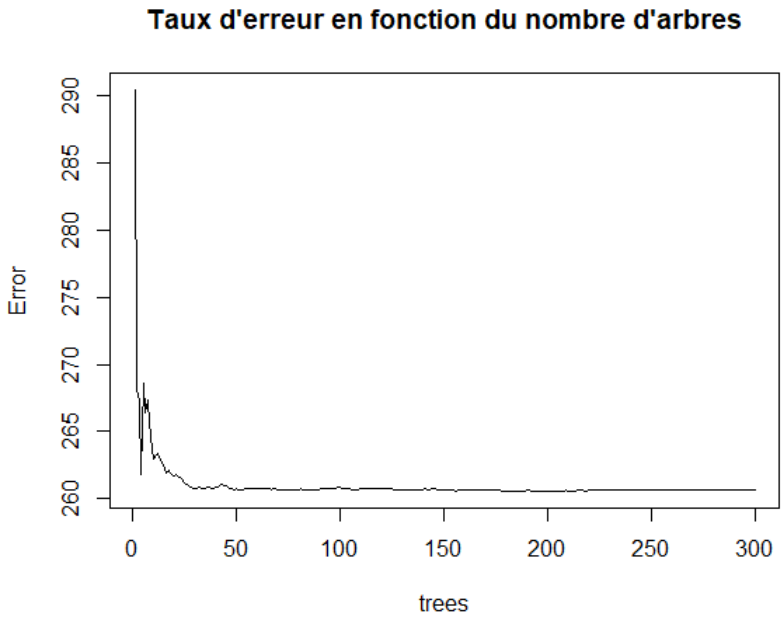


Figure IV.24: Taux d'erreur en fonction du nombre d'arbres

En se basant sur ce graphique et sur le tableau des taux erreurs, nous remarquons qu'à partir de 190 arbres, la courbe d'évolution de l'erreur est stable.

Encore dans la phase du paramétrage du modèle nous avons réalisé la technique de « *grid search* » pour rechercher la valeur optimale du paramètre (mtry) du nombre de variables explicatives testées dans les divisions.

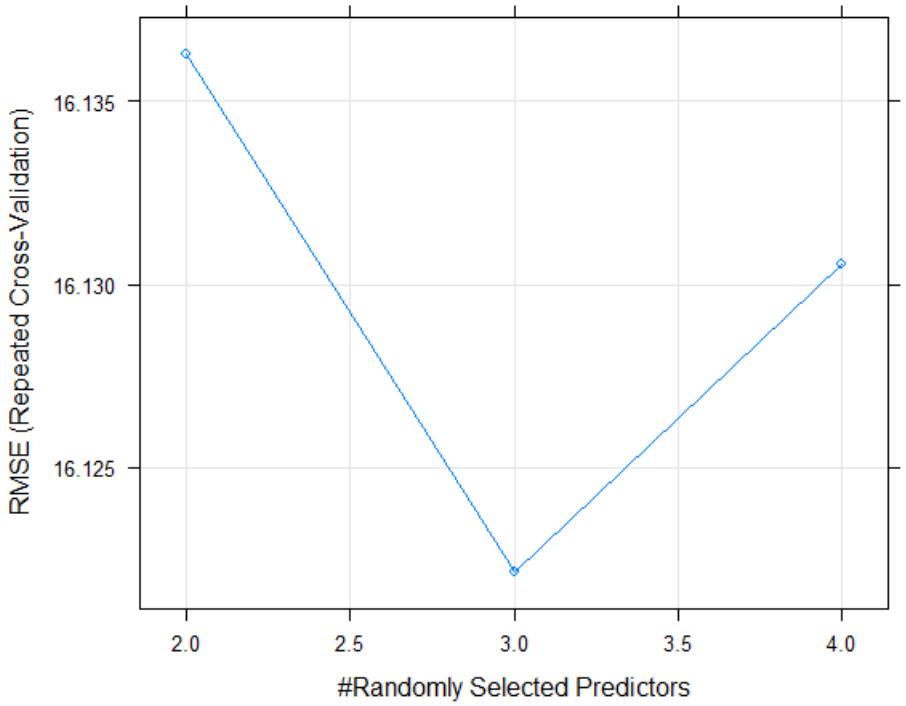


Figure IV.25: La racine carrée de la moyenne des erreurs quadratiques en fonction du paramètre mtry

Nous remarquons que la valeur 3 minimise le plus la valeur de la racine de l'erreur quadratique moyenne (RMSE), pour cela, on choisit cette valeur dans la construction de notre modèle

Enfin, nous imposons que les feuilles des arbres contiennent au moins 164 observations = 5% des données afin que les arbres ne soient pas trop grands et que nous évitions le sur-apprentissage.

Performance du modèle sur les données Test : **MSE** = 259.51, **RMSE** = 16.10

IV.4.4 GBM – Gamma

Pareil aux autres modèles l'implémentation de la gbm nécessite un bon choix de paramètres nous avons procédé à cette phase de tuning des hyperparametres avec la technique grid search. Les valeurs téstés pour les differents hyperparametres sont les suivants :

Hyperparametres	Description	Valeurs testées
Learning rate	Taux d'apprentissage	0.01, 0.05 et 0.1
Max depth	Profondeur des arbres	3, 5 ,9 et 12
Ntrees	Nombre d'arbres construits	50, 100 et 300
Distribution	Distribution de la variable réponse	Gamma

Table IV.21: Tuning des hyperparamètres GBM-Gamma

Durant la phase du Tuning 48 modèles ont été exécutés pour tester la meilleure combinaison[1] d'hyperparamètres en, pour finalement obtenir la combinaison suivante :

Learning rate	Max depth	Ntrees
0.05	3	300

Table IV.22: Combinaison optimale des hyperparamètres GBM-Gamma

Performance du modèle

	MSE	MAE	RMSE	Mean Residual Deviance
Données d'apprentissage	255.25	11.60	15.98	9.12
Données de validation	287.99	11.96	16.94	9.15
Données Test	260.41	11.69	16.14	9.11

Table IV.23: Performance du modèle GBM-Gamma

Interprétation

L'importance des variables est déterminée en calculant l'influence relative de chaque variable : si cette variable a été sélectionnée pour être divisée pendant le processus de construction de l'arbre, et de combien l'erreur quadratique (sur tous les arbres) s'est améliorée (diminuée) en conséquence.

Les résultats sont affichés dans l'annexe, un graphique résumant l'ordre des variables selon leur importance est ci-dessous :

[1] On a choisit MSE comme étant le critère pour le choix de la meilleure combinaison

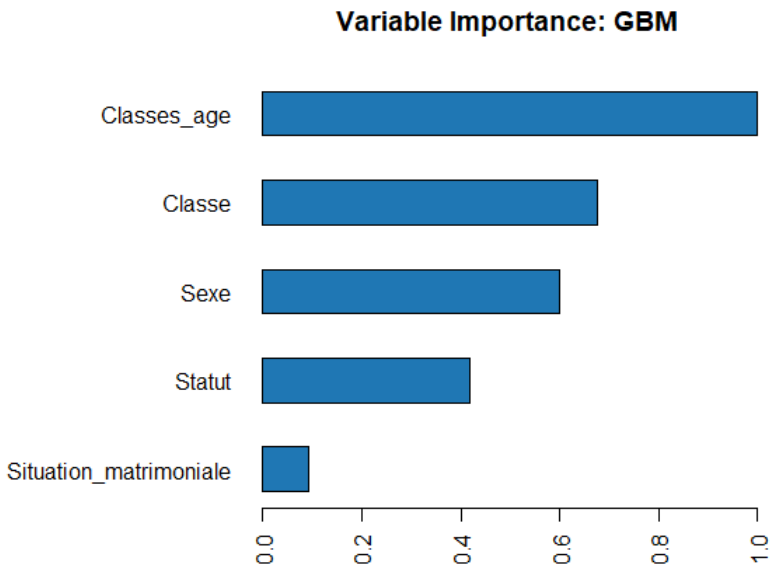


Figure IV.26: Importance des variables GBM-Gamma

IV.4.5 Performance des modèles du coût moyen

Modèle	MAE	MSE	RMSE
GLM-Gamma	11,824	261,895	16,183
CART	11,732	259,691	16,114
Forêt aléatoire	11,688	259,516	16,1
GBM	11,69	260,41	16,14

Table IV.24: Performance des modèles du coût moyen

Le tableau ci-dessus contient les indicateurs d'écart entre les prédictions et les observations à savoir L'erreur absolue moyenne (MAE), le carré moyen des erreurs (MSE) et son racine carré RMSE.

Tous les indicateurs placent le modèle de Forêt aléatoire en premier, le modèle CART se place en deuxième position par rapport au critère MAE, cependant le modèle GBM possède une valeur MSE plus faible que le modèle CART ce qui les rend comparables en termes de performances.

Par contre le dernier modèle en terme de performance est le seule modèle paramétrique GLM avec la distribution Gamma.

IV.5 Calcul de la prime pure

Arrivons a ce stade où les modèles sont consruits, il nous reste qu'a prédire la valeur de la prime en fonction des valeurs prédites de la fréquence et du coût sur les données test.

Nous rappelons que la prime pure est égale au produit de la fréquence et le coût moyen du sinistre.

Les résultats des prédictions pour deux observations des données de test sur uniquement la catégorie "Soins courants" sont les suivants :

Prédiction de la prime pure pour la catégorie soins courants							
Caractéristiques assuré					Prime pure prédite		
Sexe	Statut	Situation matrimoniale	Tranche age	Classe	GLM	CART	GBM
Femme	Principal	Mariée	>63 ans	A	15,96	19,2	24,95
Homme	Enfant	Célibataire	[9-18[	A	4,57	5,33	4,92

Table IV.25: Prédiction de la prime pure pour la catégorie des soins courants

Avec,

- Prime (GLM) = fréquence (GLM-Binomiale-négative) × coût moyen (GLM-Gamma)
- Prime (CART) = fréquence (CART) × coût moyen (CART)
- Prime (GBM) = fréquence (GBM-Poisson) × coût moyen (GBM-Gamma)

Les résultats des prédictions sur uniquement la catégorie "Pharmacie" sont les suivants :

Prédiction de la prime pure pour la catégorie Pharmacie							
Caractéristiques assuré					Prime pure prédite		
Sexe	Statut	Situation matrimoniale	Tranche age	Classe	GLM	CART	GBM
Femme	Principal	Mariée	>63 ans	A	18,27	11,14	26,55
Homme	Enfant	Célibataire	[9-18[	A	4,99	7,17	6,11

Table IV.26: Prédiction de la prime pure pour la catégorie pharmacie

Remarque : Les résultats des modèles de la catégorie pharmacie sont affichés en Annexe.

Après l'agrégation de tous les postes de garantie on peut estimer la prime pure de ces deux types d'assurés par :

Prédiction de la prime pure							
Caractéristiques assuré					Prime pure prédite		
Sexe	Statut	Situation matrimoniale	Tranche age	Classe	GLM	CART	GBM
Femme	Principal	Mariée	>63 ans	A	69.46	62.68	104.10
Homme	Enfant	Célibataire	[9-18[	A	20.14	25.02	22.08

Table IV.27: Prédiction de la prime pure

Nous rappelons que pour trouver la prime réelle payé par l'assuré il convient d'ajouter à la prime pure la charge de sécurité, les frais de gestion, les taxes ainsi que la marge bénéfice de l'assureur.

IV.6 Avantages et inconvénients des modèles

	Avantages	Inconvénients
GLM	Capacité à modéliser des comportements non linéaires à travers la fonction lien. Peut s'adapter à différentes données grâce à la variété des distributions que le modèle peut utiliser. Possibilité de faire des tests statistiques pour comparer ou analyser la qualité du modèle.	Problème d'interactions des variables explicatives : Les modèles GLM imposent à ce que les effets des variables explicatives soient additives, c-à-d, ils doivent être indépendantes. Nécessite la connaissance de la loi de la variable dépendante.
Arbres CART	N'impose pas d'hypothèses ni sur la structure ni sur la distribution des données. Peut extraire des interactions complexe entre les variables.	Susceptible au problème de sur-apprentissage. Difficulté d'interprétation (effet de la boite noire).
Forêts aléatoires	L'optimisation des paramètres ne demande pas de validation croisée grâce à la façon auquel l'erreur out of bag est calculé. Permet d'éviter le problème de sur-apprentissage.	N'est pas très efficace en nombre réduit de variables explicatives. Comme il s'agit d'une méthode de Bagging, le temps de calcul peut être élevé.
GBM	Réduction de l'erreur. Meilleure performance. Beaucoup de flexibilité : peu se faire sur différentes fonctions de perte et fournit plusieurs réglages d'hyperparamètres.	Le paramétrage est plus complexe que les forêts aléatoires. Phase d'apprentissage prend généralement plus de temps.

Table IV.28: Avantages et inconvénients des modèles

IV.7 Partie projet : Déploiement

La modélisation et la construction de modèles d'apprentissage automatique sont au cœur de l'analyse et de la prédiction des informations pour comprendre les phénomènes et prendre des décisions. Dans ce que nous avons fait, nous avons vu l'impact de ces méthodes dans un cas réel qui est la tarification. Cependant, dans le monde technologique actuel, l'objectif n'est pas d'exécuter et d'expérimenter une fois mais de déployer les modèles afin qu'ils soient en production et disponibles pour les utilisateurs finaux à tout moment de manière simple et entièrement automatisée.

L'objectif de cette section est de présenter le travail effectué sur le processus de déploiement des modèles de tarification déjà abordés. Ces travaux sont réalisés dans le cadre du projet « **Digitarif** » au sein de l'équipe RD de Hydatis Engineering .

IV.7.1 Projet Digitarif

Dans le cadre de ses travaux de digitalisation, Hydatis Engineering se concentre cette fois sur l'intégration de la modélisation scientifique et de l'intelligence artificielle dans le secteur de l'assurance. En effet, la nouvelle solution Digitarif est une plateforme conçue pour automatiser et accélérer le processus traditionnel de tarification à l'aide d'algorithmes d'apprentissage automatique.

IV.7.2 Objectif

En développement informatique on appelle API (Application Programming Interface) la méthode permettant la communication entre les programmes sans l'intervention de l'utilisateur. En d'autres termes plus pratiques, un API est un ensemble de fonctions qui permettent d'accéder aux données et fonctionnalités d'une application existante. Les plateformes web et les applications en géneral utilise les API pour accéder aux fonctionnalités du logiciel produisant la solution (serveur back-end).

La construction de l'API est basée sur le développement de ce que l'on appelle des points de terminaison (EndPoints) qui sont des méthodes permettant d'effectuer différentes taches ou actions dans l'application.

L'objectif poursuivi est de mettre en œuvre les techniques et les modèles déjà abordés dans la tarification en assurance santé sous la forme d'un ensemble des endpoints, créant ainsi un processus de déploiement complet, de l'importation des données brutes à la prédiction et à l'analyse des résultats.

IV.7.3 Bibliothèques utilisées

FLASK

FLASK est un Framework de développement web open source développé en python. Il s'agit d'une collection de modules et bibliothèques permettant la création des applications web d'une façon simple sans se soucier des détails comme l'intégration du système d'authentification ou la gestion de threads. FLASK est le cœur du travail sur la construction des endpoints et le déploiement des modèles sur le web.

H2O

H2O est une plateforme open source d'apprentissage automatique développé en JAVA, elle supporte les algorithmes statistiques les plus utilisés ainsi que des outils de manipulation et traitement de données. En plus de la variété des outils statistiques en sciences de données qu'elle propose, elle est caractérisée par sa rapidité de traitement et efficacité vu qu'elle se base sur un calcul distribué en mémoire, c'est-à-dire que tous les exécutions sont faites sur une mémoire distribuée dans les ressources de H2O (cluster H2O).

Grâce au mécanisme des API, on peut accéder à ses fonctionnalités via plusieurs langages de programmation. Dans notre cas, nous utilisons python pour connecter H2O à notre API créée dans FLASK.

MLflow

MLflow est une plateforme de gestion du cycle de vie d'un projet d'apprentissage automatique. Elle permet le suivi des dernières exécutions, la reproduction et la gestion des versions des modèles déployés.

MLflow offre plusieurs fonctionnalités intéressantes dans la gestion des projets d'apprentissage automatique, parmi lesquelles la fonctionnalité de suivi (MLflow Tracking).

- **MLflow Tracking** : Un API pour enregistrer les paramètres, le code et les résultats des expériences d'apprentissage automatique et les comparer à l'aide d'une interface utilisateur interactive.

IV.7.4 Les fonctions développées

L'étape	Les fonctions développées
Traitement des données	- Importation des données - Mapping (adaptation des données à une structure spécifique) - Imputation des valeurs manquantes - Normalisation - Discrétisation - Echantillonnage (division train et test)
Construction des modèles	*) Modélisation de la fréquence - GLM - GBM - XGBOOST
	*) Modélisation du coût moyen - GLM - GBM - XGBOOST
Prédiction	- Affichage des prédictions - Analyse des performances
Interprétabilité	*) Techniques de l'intelligence artificielle explicable : - SHAP - Visualisation de l'importance des variables. - PDP
Tracking avec MLflow	- Suivi des paramètres des modèles. - Enregistrement des données d'apprentissage et test. - Enregistrement des modèles.

Table IV.29: Fonctions développées pour la partie déploiement

V Conclusion

La tarification en assurance a été et sera toujours un sujet d'étude et d'amélioration étant donné son impact direct sur la compétitivité du marché. Dans ce projet, nous avons étudié la tarification sur un cas réel sur des données d'assurance santé. L'objectif était de rappeler les bases théoriques et d'appliquer les modèles linéaires généralisés présentant l'approche standard en matière de tarification, ainsi que certaines méthodes d'apprentissage qui intègrent de plus en plus le domaine ces derniers temps.

L'idée du travail n'était pas de prouver la performance d'un modèle par rapport aux autres, mais plutôt d'étudier le processus de réalisation de chaque modèle et d'essayer de l'optimiser pour qu'il soit capable de résoudre les problèmes découlant des données, tels que la présence d'une asymétrie dans la structure des coûts, une hétérogénéité du portefeuille et la présence d'une grande proportion de zéro dans le nombre de sinistres causés principalement par la non-déclaration.

Nous avons discuté au début de la manière dont nous allons calculer la prime d'assurance pure, ce qui nous ramène à la division du travail en deux parties indépendantes : une modélisation de la fréquence des sinistres et une modélisation du coût d'un sinistre.

Pour la partie fréquence, nous avons commencé par le GLM où nous avons spécifié la nécessité d'utiliser des lois de probabilité discrètes, pour cela nous avons étudié l'ajustement de la distribution de Poisson et de la distribution Binomiale négative où cette dernière a prouvé son intérêt comme solution à la sur-dispersion. Les modèles d'inflation zéro ont également été utilisés comme alternatives à la distribution de Poisson.

Avec le même objectif de modélisation de la fréquence de sinistralité, nous avons appliqué les modèles basés sur des arbres, où nous avons commencé par l'arbre de régression de type CART. Nous avons détaillé la manière dont nous avons construit l'arbre en partant d'un arbre saturé pour arriver à un arbre optimal en utilisant la technique d'élagage. L'arbre de régression nous a permis non seulement de prédire la fréquence pour les nouvelles observations mais aussi de visualiser les groupes existants dans la population en suivant la structure de ces branches. Un autre algorithme utilisant la technique du boosting a été réalisé dans le même cadre, à savoir le Gradient boosting.

L'approche de la modélisation du coût a été identique à celle de la fréquence, sauf que les lois utilisées dans le GLM sont la loi Gamma et la loi Log-Normale, l'utilisation de ces lois a été causée par l'asymétrie positive existante dans la structure du coût. Les méthodes d'apprentissage automatique utilisées sont similaires à celles utilisées dans la fréquence, cependant nous avons réalisé le modèle de forêt aléatoire en plus et qui a dominé tous les autres modèles en termes de performance sur les données de test.

Enfin, nous nous sommes intéressés à la partie déploiement, dans la quelle nous avons montré le travail réalisé sur le développement d'un API Web dans lequel des fonctions ont été développé pour gérer le processus de la tarification de l'importation brutes à la modélisation avec les méthodes déjà vues, la prédiction et l'analyse des résultats.

En réponse aux questions posées dans l'introduction, les modèles GLM parviennent dans notre cas à surmonter certains problèmes liés à la structure des données et donc à expliquer la fréquence et le coût des variables et ce grâce à sa capacité à ajuster plusieurs distributions et à prouver les résultats avec des tests statistiques. Tous ces avantages expliquent la concentration récente des compagnies d'assurance à développer ces modèles. Cependant, avec l'augmentation des données et l'orientation du développement vers l'intelligence artificielle, les modèles GLM peuvent être complètement remplacés si les techniques d'apprentissage s'améliorent et se débarrassent de l'effet "boîte noire".

Annexes

Résultats du modèle ZINB:

```
Call:
zeroinfl(formula = Nombre_actes_poste ~ Sexe + Classes_age + Statut + Situation_matrimonia
    Classe, data = train_freq, dist = "negbin", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.4784 -0.3629 -0.3271 -0.2982 12.0687

Count model coefficients (negbin with log link):
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -0.856682   1.256776  -0.682  0.49546
SexeHomme                      -0.130076   0.068279  -1.905  0.05677 .
Classes_age[9-18[              -0.500282   0.153565  -3.258  0.00112 **
Classes_age[18-27[             -0.195711   0.188015  -1.041  0.29791
Classes_age[27-36[             -0.303084   0.256299  -1.183  0.23699
Classes_age[36-45[             -0.445726   0.259480  -1.718  0.08584 .
Classes_age[45-54[             -0.504228   0.262047  -1.924  0.05433 .
Classes_age[54-63[             -0.180871   0.275548  -0.656  0.51156
Classes_age63+                  0.055836   0.407932   0.137  0.89113
StatutPrincipal                 0.609581   0.255058   2.390  0.01685 *
StatutConjoint                  0.577324   0.261972   2.204  0.02754 *
Situation_matrimonialeMarié(e)  0.139313   1.250991   0.111  0.91133
Situation_matrimonialeCelibataire -0.009563   1.252969  -0.008  0.99391
Situation_matrimonialeVeuf(ve)  0.775089   1.474527   0.526  0.59913
ClasseCLASSE A                 -0.025233   0.063298  -0.399  0.69015
Log(theta)                      0.002482   0.233553   0.011  0.99152

Zero-inflation model coefficients (binomial with logit link):
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      0.23307    1.55754   0.150   0.8810
SexeHomme                       -0.15617    0.10819  -1.443   0.1489
Classes_age[9-18[               -0.84812    0.49703  -1.706   0.0879 .
Classes_age[18-27[              -0.28593    0.39618  -0.722   0.4705
Classes_age[27-36[              -0.35047    0.45676  -0.767   0.4429
Classes_age[36-45[              -0.36229    0.45982  -0.788   0.4308
Classes_age[45-54[              -0.42488    0.46544  -0.913   0.3613
Classes_age[54-63[              -0.29504    0.47807  -0.617   0.5371
Classes_age63+                  -0.86967    0.70410  -1.235   0.2168
StatutPrincipal                  1.06278    0.45817   2.320   0.0204 *
StatutConjoint                   0.88041    0.46383   1.898   0.0577 .
Situation_matrimonialeMarié(e)  -0.50463    1.54163  -0.327   0.7434
Situation_matrimonialeCelibataire -0.38047    1.54459  -0.246   0.8054
Situation_matrimonialeVeuf(ve) -13.69369  351.15022  -0.039   0.9689
ClasseCLASSE A                   0.45765    0.09975   4.588 4.48e-06 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.0025
Number of iterations in BFGS optimization: 43
Log-likelihood: -1.306e+04 on 31 Df
```

Estimation des paramètres de la loi Gamma

```
Fitting of the distribution ' gamma ' by maximum likelihood
Parameters :
       estimate  Std. Error
shape 5.2431079 0.105746679
rate  0.1482126 0.003136985
Loglikelihood:  -18903.86   AIC:  37811.72   BIC:  37824.59
```

Coefficients estimés par le modèle GLM-LogNormale

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.62577    0.01854 195.524  < 2e-16 ***
Sexe_Femme            0.07990    0.01612   4.958 7.48e-07 ***
`Classes_age_[0-9[`   0.04574    0.03045   1.502  0.13313
`Classes_age_[36-45[` -0.07357    0.02395  -3.072  0.00215 **
`Classes_age_[45-54[` -0.06261    0.02463  -2.542  0.01107 *
Statut_Enfant         -0.12957    0.02956  -4.383 1.21e-05 ***
`Classe_CLASSE 1`     -0.08232    0.01609  -5.116 3.30e-07 ***
```

Calcul de l'influence relative des variables GBM-Gamma

```
Variable Importances:
                variable relative_importance scaled_importance percentage
1           Classes_age          106.140892          1.000000   0.358949
2                Classe           71.742767          0.675920   0.242621
3                  Sexe           63.515373          0.598406   0.214797
4                Statut           44.377338          0.418098   0.150076
5 Situation_matrimoniale            9.922684          0.093486   0.033557
```

Modélisation de la fréquence par GLM pour la catégorie pharmacie

```
glm.nb(formula = Nombre_actes_poste ~ Classes_age + Classe, data = train_ph,
    init.theta = 0.3113395523, link = log)
```

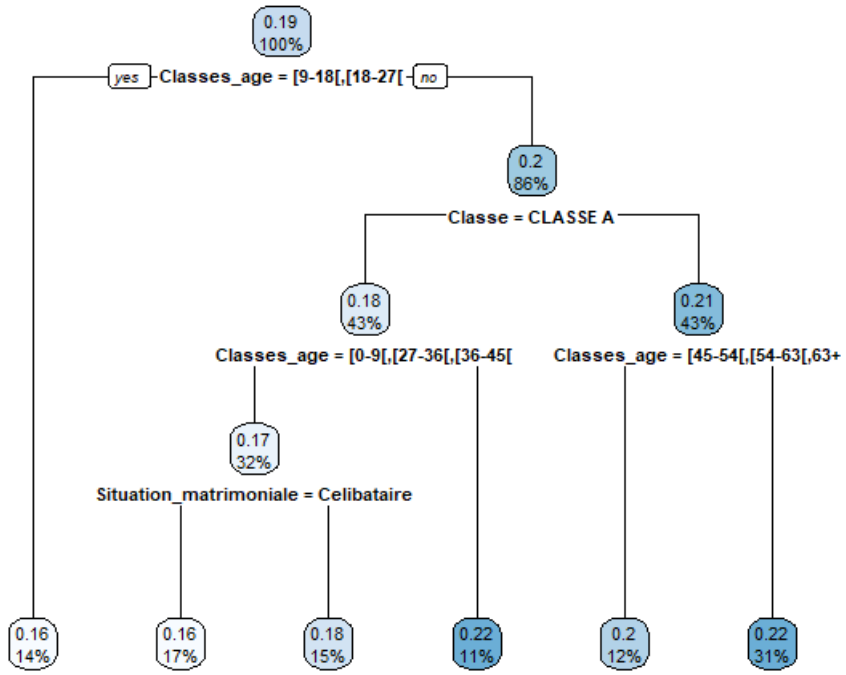
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7404  -0.5672  -0.5347  -0.5292   3.3947

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.57353    0.03973 -39.606  < 2e-16 ***
Classes_age[9-18[  -0.24671    0.07979  -3.092  0.00199 **
Classes_age[18-27[ -0.14551    0.08190  -1.777  0.07564 .
Classes_age[27-36[  0.04613    0.05398   0.855  0.39278
Classes_age[36-45[  0.02547    0.05425   0.470  0.63868
Classes_age[45-54[  0.01575    0.05726   0.275  0.78322
Classes_age[54-63[  0.15952    0.07359   2.168  0.03017 *
Classes_age63+      0.75136    0.24224   3.102  0.00192 **
ClasseCLASSE A     -0.15900    0.03607  -4.409 1.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.3113) family taken to be 1)

    Null deviance: 11935  on 26287  degrees of freedom
Residual deviance: 11887  on 26279  degrees of freedom
AIC: 27255

Number of Fisher Scoring iterations: 1


              Theta:  0.3113
          Std. Err.:  0.0128

 2 x log-likelihood:  -27234.6530
```

Modélisation du coût par GLM pour la catégorie pharmacie

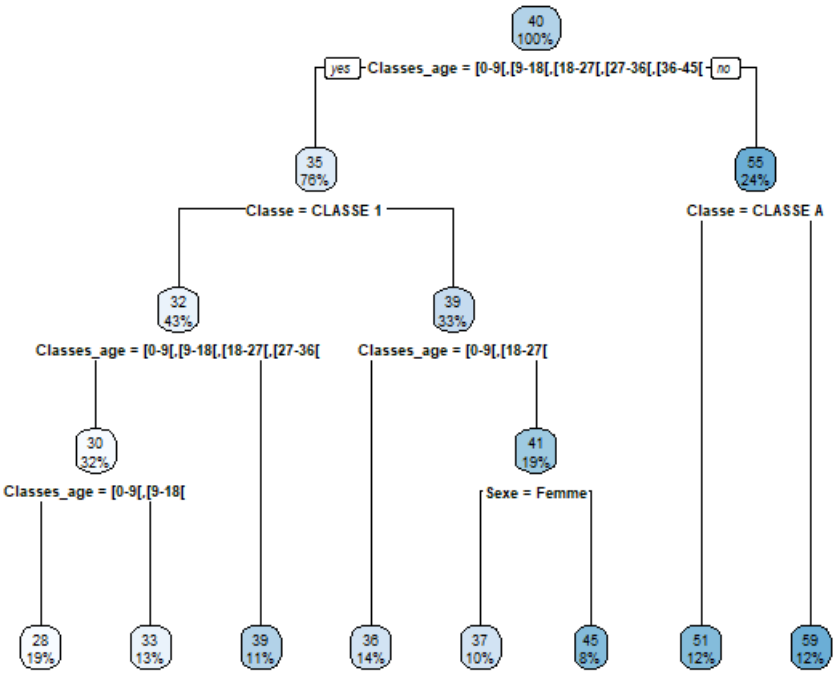
```
glm(formula = Cout_moyen_rembourse ~ Sexe + Situation_matrimoniale +
    Classes_age + Classe, family = Gamma(link = "inverse"), data = train_ph_cout)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9643  -0.7004  -0.2526   0.2163   4.0112

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.0895602  0.0524183   1.709 0.087619 .
SexeHomme                       -0.0015859  0.0008024  -1.977 0.048169 *
```

```
Situation_matrimonialeMarié(e)     -0.0536128  0.0523958  -1.023 0.306271
Situation_matrimonialeCelibataire -0.0558207  0.0524114  -1.065 0.286927
Situation_matrimonialeVeuf(ve)     -0.0473927  0.0550381  -0.861 0.389248
Classes_age[9-18[                  -0.0024066  0.0020776  -1.158 0.246777
Classes_age[18-27[                 -0.0037529  0.0022009  -1.705 0.088247 .
Classes_age[27-36[                 -0.0068705  0.0017516  -3.922 8.93e-05 ***
Classes_age[36-45[                 -0.0088031  0.0018619  -4.728 2.36e-06 ***
Classes_age[45-54[                 -0.0153332  0.0018326  -8.367  < 2e-16 ***
Classes_age[54-63[                 -0.0167140  0.0019572  -8.540  < 2e-16 ***
Classes_age63+                     -0.0157448  0.0042242  -3.727 0.000197 ***
ClasseCLASSE A                     -0.0019137  0.0008037  -2.381 0.017316 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.9189984)

    Null deviance: 2593.7  on 3522  degrees of freedom
Residual deviance: 2416.5  on 3510  degrees of freedom
AIC: 32481
```

Arbre de régression pour la fréquence de la catégorie pharmacie



Arbre de régression pour le coût de la catégorie pharmacie



Bibliography

[Bellina(2014)] R. Bellina. Méthodes d'apprentissage appliquées à la tarification non-vie. *Mémoire ISFA*, 2014.

[Boucher et al.(2008)Boucher, Denuit, and Guillén] J.-P. Boucher, M. Denuit, and M. Guillén. Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions. *Variance*, 2(1):135–162, 2008.

[Charpentier(2010)] A. Charpentier. Statistique de l'assurance. Lecture, Sept. 2010. URL `https://cel.archives-ouvertes.fr/cel-00550583`.

[Ferrario et al.(2020)Ferrario, Noll, and Wuthrich] A. Ferrario, A. Noll, and M. V. Wuthrich. Insights from inside neural networks. *CompSciRN: Industry Practical Application (Topic)*, 2020.

[Guelman(2012)] L. Guelman. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Syst. Appl.*, 39:3659–3667, 2012.

[Haberman and Renshaw(1996)] S. Haberman and A. Renshaw. Generalized linear models and actuarial science. *The Statistician*, 45:407, 01 1996. doi: 10.2307/2988543.

[Henckaerts et al.(2020)Henckaerts, Côté, Antonio, and Verbelen] R. Henckaerts, M.-P. Côté, K. Antonio, and R. Verbelen. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, pages 1–31, 2020.

[Karatekin(2014)] O. Karatekin. *Tarification et mesure de l'antiselection en assurance sante collective*. PhD thesis, Crédit Mutuel, 88-90 rue Cardinet, 75017 Paris, 2014.

[Lebarbier and Mary-Huard(2004)] E. Lebarbier and T. Mary-Huard. Le critère bic : fondements théoriques et interprétation. 147, 01 2004.

[Nelder and Wedderburn(1972)] J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[Paglia and Phelippe-Guinvarc'h(2011)] A. Paglia and M. V. Phelippe-Guinvarc'h. Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. *Bulletin français d'Actuariat*, 11(22):49–81, 2011.

[Ver Hoef and Boveng(2007)] J. M. Ver Hoef and P. L. Boveng. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.

[Wuthrich and Buser(2020)] M. V. Wuthrich and C. Buser. Data analytics for non-life insurance pricing. *Swiss Finance Institute Research Paper*, (16-68), 2020.