Tarification et mesure de l'antisélection en assurance santé collective

Ozlem Karatekin

Mémoire présenté devant

l'UFR de Mathématique et d'Informatique

pour l'obtention du Diplôme Universitaire d'Actuaire de Strasbourg

et l'admission à l'Institut des Actuaires

le 02/10/2014

Par : Ozlem KARATEKIN

Titre: Tarification et mesure de l'antisélection en assurance santé collective

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

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

Membres du jury de l'Institut des Actuaires signature

M. MODRY et M. YOU

Membres du jury de l'UdS :

M. BERARD
M. EISELE
M. FRANCHI
M. NETZER

Invités :

M. DUBOIS
M. FITOUCHI
Mme FOATA
M. GADENNE
M. HESS
Mme KELLE-VIGON
Mme MAUMY-BERTRAND
M. VIGON

Secrétariat : Mme Maire-Lantz

Bibliothèque : Mme Christine Disdier

Entreprise :

Nom : Assurances du Crédit Mutuel

Signature :

Directeur de mémoire en entreprise :

Nom : Noémie DREYFUS

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels

(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise

Signature du candidat

Résumé

Mots clés : Assurance santé collective, Tarification, Modèle linéaire généralisé, Modèle fréquence-coût, Risque d'antisélection.

Le contexte très concurrentiel du marché de l'assurance santé collective incite les assureurs à établir des tarifs compétitifs. Ce besoin s'est accru suite à l'accord national interprofessionnel du 11 janvier 2013. L'obligation pour les entreprises du secteur privé de souscrire un contrat complémentaire santé pour l'ensemble de leurs salariés, a pour conséquence une dynamisation du secteur collectif de l'assurance santé. Dans ce cadre, ce mémoire propose de tester une méthode alternative à la méthode traditionnelle de tarification « fréquence - coût moyen », basée sur les modèles linéaires généralisés.

Après une analyse préliminaire du portefeuille, les facteurs expliquant le comportement de consommation de frais de soins de santé ont été sélectionnés et une classification des départements a été effectuée en fonction de la sinistralité observée. La fréquence de consommation et le remboursement moyen ont été modélisés séparément pour l'ensemble des actes de soins médicaux étudiés. Étant donné que le comportement de consommation est différent selon les actes considérés, les résultats obtenus ont été présentés selon les deux cas suivants : une consommation fréquente telle que les analyses et actes de laboratoire et une consommation plus rare telle que les prothèses dentaires. Ainsi, différents GLM ont été testés pour la modélisation de la fréquence de consommation afin de trouver le modèle le plus approprié aux données étudiées, notamment les modèles « modifiés en zéro » et binomial négatif. Enfin, les résultats obtenus des différents modèles ont été confrontés à la méthode actuelle de tarification directe de la fréquence et du coût moyen.

Cette tarification a permis d'établir la prime pure des contrats santé collectifs ayant un caractère obligatoire. Par ailleurs, ce mémoire propose également une analyse et une méthode descriptive pour la prise en compte du risque d'antisélection causé par la commercialisation de contrats collectifs facultatifs. Défini comme l'impossibilité pour l'assureur de distinguer les profils de risque, le risque d'antisélection est un phénomène économique qui a été appréhendé par une approche statistique. L'écart de fréquence de consommation a été observé entre les contrats collectifs obligatoires et contrats individuels, compte tenu de données insuffisantes relatives aux contrats collectifs facultatifs. Pour finir, une analyse par poste de garantie et par âge a permis d'acquérir une meilleure connaissance de ce risque.

Abstract

Keywords : Group health insurance, Pricing, Generalized linear model, Frequency - average cost model, Adverse selection risk.

In the extremely competitive environment of the market of group health insurance, following the French Inter-professional national agreement (Accord National Interprofessionnel) of January 11 in 2013, insurers seek to establish competitive prices. Companies in the private sector have to purchase a health care insurance policy for their salaries and this has the effect of stimulating the group health insurance market. In this context, this report suggests to test an alternative method to the «frequency - average cost» usually used, named generalized linear model.

After a primary analysis of the portfolio, the factors which can influence the medical consumption behaviour were selected and a classification of the geographical location were done according to the medical consumption. The frequency of consumption and the average cost were modeled separately for all acts of medical care studied, the results have been presented according to this two cases : a frequent consumption as laboratory tests and infrequent as dental prosthesis. Thus, different GLM have been tested to model the frequency of consumption in order to find the most appropriate model to the data used, as negative binomial and « zero inflated » models. Finally, the results of the different models used, have been compared to the method of pricing directly the frequency and the average cost.

We used these models in order to estimate the insurance premiums of group health policies which are compulsory. This report also presents an analysis and a descriptive method in order to take into account the adverse selection risk, caused by the marketing of voluntary group policies. Defined as the impossibility for the insurer to distinguish the risk profiles, the adverse selection risk is an economic phenomenon which we try to analyse in a statistical approach. The difference of the frequency of consumption has been observed between the compulsory group policies and the individual policies, because of an insufficient number of voluntary group policies. Finally, an analysis according to category of medical acts and age has permitted to have a better knowledge of this risk.

Remerciements

Tout d'abord, je tiens à exprimer toute ma gratitude envers ma maitre de stage Noémie Dreyfus, responsable du service Actuariat, pour la confiance qu'elle a su m'accorder et pour son suivi et son aide tout au long de ce stage.

Un grand merci à Mario Gugumus pour m'avoir fait bénéficier de ses pertinentes remarques et suggestions. Il a suivi avec beaucoup d'intérêt les travaux effectués et a su être disponible tout au long de ce stage.

Je remercie Patrick Garcia, pour son accueil, ses encouragements et ses précieuses aides concernant la rédaction du mémoire en Latex.

Je suis également reconnaissante envers Mme Muriel Marron et Mme Sabine Klein de m'avoir permis de réaliser ce stage de fin d'études au sein des Assurances du Crédit Mutuel et je remercie l'ensemble du service collectif pour leur accueil.

Mes remerciements s'adressent également à ma tutrice universitaire, Mme Myriam Maumy-Bertrand, pour sa relecture du mémoire et ses conseils avisés.

Enfin, je souhaiterais exprimer ma reconnaissance envers les membres de ma famille, pour leur soutien moral et leur patience durant toute ma formation universitaire.

Table des matières

Introduction générale

L'accord national interprofessionnel du 11 janvier 2013 [1] bouleverse le marché français de l'assurance santé. Cet accord, exigeant la mise en place d'une complémentaire santé obligatoire pour tous les salariés du secteur privé d'ici 2016, devrait entraîner un transfert des contrats individuels des salariés vers les contrats collectifs obligatoires. Ce contrat obligatoire doit respecter les garanties minimales fixés par un projet de décret très attendu par les assureurs. Compte tenu du faible niveau de remboursement fixé par ce projet de décret, les assureurs prévoient également une hausse du recours aux contrats collectif facultatifs par les assurés pour augmenter leur niveau de remboursement et compléter leur panier de soins.

Dans ce contexte très concurrentiel, les assureurs se doivent de proposer des tarifs compétitifs tout en couvrant leurs engagements dans le remboursement des frais de santé des assurés et leurs frais de fonctionnement. Cela rend nécessaire le développement de modèles de tarification permettant d'appréhender au mieux les risques sous-jacents.

C'est dans ce cadre que s'inscrit ce mémoire ayant pour premier objectif la tarification de contrats collectifs en santé dans le cadre d'un modèle fréquence - coût moyen. L'objectif de cette première étude est de mettre à jour le tarificateur actuel et de tester l'adéquation d'un modèle paramétrique aux données étudiées : le modèle linéaire généralisé. Ce type de modèle, utilisé majoritairement en assurance non vie, est une généralisation du modèle linéaire simple. Il permet notamment de modéliser les fréquences et coûts moyens des actes en fonction de variables ayant une influence sur ces deux grandeurs.
Le deuxième axe de ce mémoire vise à évaluer le risque d'antisélection, nécessaire à la tarification de contrats collectifs facultatifs. Pour ce faire, un ou plusieurs coefficients de majoration seront appliqués au tarif d'un contrat collectif obligatoire. Afin d'estimer ces coefficients indépendamment du phénomène d'aléa moral, très présent en santé, une méthode d'analyse de ces coefficients par niveau de garanties du contrat sera préférée.

La première partie de ce mémoire est consacrée à la présentation du secteur de l'assurance santé en France, comprenant notamment une brève description du contexte actuel.

Une deuxième partie est dédiée à l'analyse des données et aux statistiques descriptives. Une analyse plus détaillée est présentée concernant l'influence du lieu d'habitation de l'assuré sur sa consommation en frais de soins de santé par l'utilisation de techniques statistiques multivariées.

1. Accord transcrit dans la loi relative à la sécurisation de l'emploi votée le 15 juin 2013 (Loi n°2013-504).

Dans la troisième partie, la théorie des modèles linéaires généralisée est développée afin de l'appliquer à la modélisation de la fréquence et du coût moyen. Une attention particulière est accordée à la modélisation de la fréquence, pour laquelle divers modèles sont testés et comparés afin de choisir le meilleur modèle ajusté aux données étudiées. Les résultats sont ensuite comparés à la méthode de tarification actuelle.

Enfin, dans la dernière partie de ce mémoire, une analyse à la fois micro-économique et statistique est présentée pour expliquer et mesurer le risque d'antisélection. Une vérification préalable de la présence de ce phénomène est effectuée avant de proposer une analyse par poste de garantie et par âge de l'assuré.

Première partie

L'assurance santé

Dans cette première partie, il convient de décrire le secteur de l'assurance santé en France. Tout d'abord, les deux acteurs de l'assurance santé, c'est à dire la Sécurité sociale et les complémentaires santé seront présentés afin de comprendre leur fonctionnement et leur rôle dans le remboursement des frais de soins de santé. Et ensuite, une brève présentation du contexte actuel sera évoquée notamment avec l'Accord national interprofessionnel dont l'impact sur le marché de l'assurance santé est majeur.

Chapitre 1

Le régime de la Sécurité Sociale

1.1 Le fonctionnement général

Créée après la seconde guerre mondiale, la Sécurité sociale a été mise en place pour protéger les individus face aux conséquences financières des risques sociaux. Elle est composée actuellement de cinq branches : la maladie, les accidents du travail et les maladies professionnelles, la retraite, la famille et le recouvrement. Nous nous intéresserons à la branche maladie, et plus particulièrement au domaine de la santé qui couvre les dépenses de santé des assurés et les dépenses relatives à la maternité. La Sécurité sociale rembourse également les frais de santé des ayants droit, c'est-à-dire des personnes à charge de l'assuré et qui ne peuvent bénéficier d'une protection sociale à titre personnel. Les ayant droits peuvent être les enfants, le conjoint ou les ascendants à charge de l'assuré.

1.2 Les principaux régimes

L'affiliation au régime de la Sécurité sociale est obligatoire pour toutes les personnes qui travaillent et résident en France. Plusieurs régimes de Sécurité sociale existent en France, dont les principaux sont :

- régime général / régime local Alsace-Moselle ;
- régime social des indépendants (RSI) ;
- régime agricole ;
- régimes spéciaux.

Le **régime général** couvre environ 85% [2] de la population française. Il s'agit de la plupart des salariés, mais également d'autres catégories telles que les étudiants qui, au fil du temps, ont été rattachées au régime général.

Le **régime local Alsace-Moselle** est un régime particulier [3] qui concerne les salariés exerçant une activité dans les départements du Bas-Rhin (67), du Haut-Rhin (68) et de la Moselle (57).

2. Source : www.ameli.fr, site de l'assurance maladie
3. Régime mis en place depuis 1946 suite à l'annexion de l'Alsace-Moselle en 1871 par l'Allemagne

Le **régime social des indépendants** (non agricole) regroupe les travailleurs non-salariés tels que les professions libérales.

Le **régime agricole** permet de couvrir les exploitants et salariés agricoles.

Les **régimes spéciaux** des salariés concernent notamment les salariés de la SNCF, de la RATP, d'EDF/GDF etc.

Aujourd'hui, la Sécurité sociale permet de couvrir une très grande majorité des personnes résidant en France [4] grâce à des dispositifs tels que la Couverture Maladie Universelle (CMU) qui permet à ces personnes qui ne sont pas affiliées à un régime obligatoire, l'accès aux soins et le remboursement de ces derniers. D'autres dispositifs d'aides existent afin de permettre à tous les résidents français de bénéficier d'une couverture santé. Cependant, notre étude vise à proposer une couverture complémentaire aux salariés des entreprises, qui ne sont pas concernés par ce type de dispositifs, que nous ne développons pas.

1.3 Le principe de remboursement

1.3.1 La distinction des actes et le conventionnement

En assurance santé, les remboursements sont fonction d'actes (consultations, soins et prothèses dentaires, optique,etc.) codifiés par la Sécurité sociale selon différentes nomenclatures.

La Classification Commune des Actes Médicaux (CCAM) regroupe les actes techniques réalisés par les médecins. Pour ce qui concerne les actes cliniques médicaux, les actes des chirurgiens-dentistes et des auxiliaires médicaux, il convient de consulter la Nomenclature Générales des Actes Professionnels (NGAP).

Le remboursement de la Sécurité sociale va également dépendre d'un autre paramètre : le conventionnement du médecin. Nous distinguons trois types de conventionnement :

- Médecins conventionnés du secteur 1 : ces médecins appliquent le tarif conventionnel et bénéficient, en contrepartie, d'une prise en charge partielle des charges sociales par l'Etat.
- Médecins conventionnés du secteur 2 : ces médecins peuvent fixer des tarifs supérieurs au tarif conventionnel, mais ils sont contraints de payer en totalité leurs charges sociales.
- Médecins non conventionnés : ces médecins n'ont passé aucune convention avec la Sécurité sociale et, sont ainsi libres d'appliquer le tarif souhaité. En contrepartie, la prise en charge par la Sécurité sociale des frais de consultations de leurs patients sera très faible.

Un nouveau type de contrat, appelé « contrat d'accès aux soins » et destiné aux médecins conventionnées du secteur 2 et à certains médecins du secteur 1, est entré en

4. D'après l'INSEE, 99.6% des résidents français en 2003

vigueur depuis le 1er décembre 2013. Les médecins signataires de ce contrat s'engagent à respecter des tarifs de consultations, et bénéficient en contrepartie d'un allégement de leurs cotisations sociales. L'objectif est de diminuer les frais restant à charge de l'assuré à la suite du remboursement de la Sécurité sociale pour les médecins pratiquant des tarifs supérieurs au tarif conventionnel.

1.3.2 Le remboursement

La décomposition des frais médicaux en santé peut être illustrée par le schéma ci-dessous :

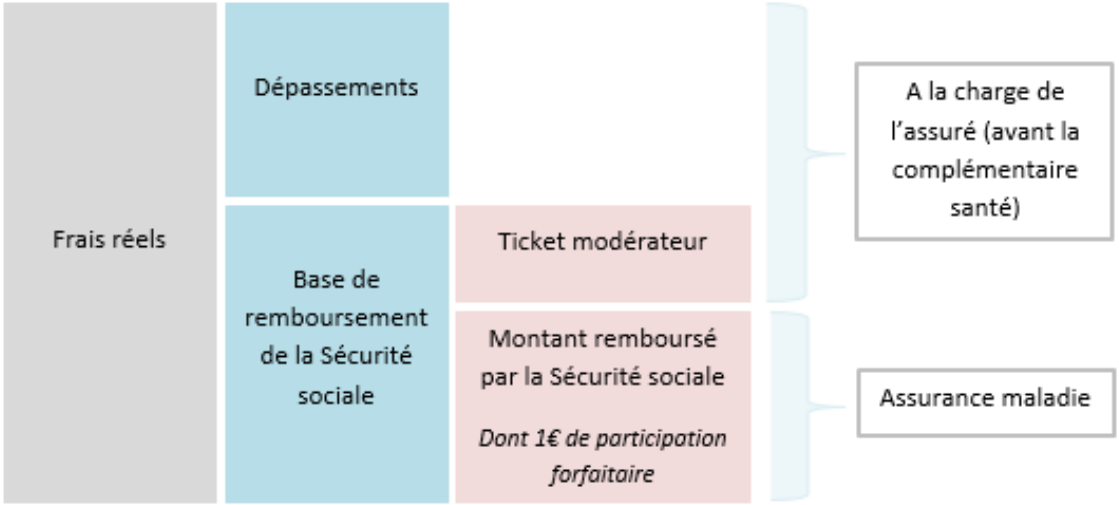


FIGURE 1 – Décomposition des frais de santé

Les **frais réels** correspondent au tarif appliqué par le praticien pour tout acte médical.

L'assurance maladie obligatoire établit une **base de remboursement** pour chaque acte. Il s'agit de tarifs établis par convention ou par arrêté ministériel. La base de remboursement correspond au tarif de convention lorsque l'acte est effectué par un médecin conventionné. Elle correspond au tarif d'autorité, montant très faible, lorsqu'il s'agit d'un médecin non conventionné.

Le montant remboursé par la Sécurité sociale, avant l'application éventuelle de la participation forfaitaire, correspond à la base de remboursement multiplié par un taux de remboursement. Ce taux dépend du type d'acte et du régime auquel adhère le salarié.

La participation forfaitaire, due pour certains actes, correspond à un montant de 1 € à la charge du patient afin d'alléger le déficit de l'assurance maladie.

Le schéma ci-dessous permet d'illustrer le montant à la charge de l'assurance maladie obligatoire d'un salarié au régime général dans le cas d'une consultation de médecin généraliste (n'appliquant pas de dépassements).

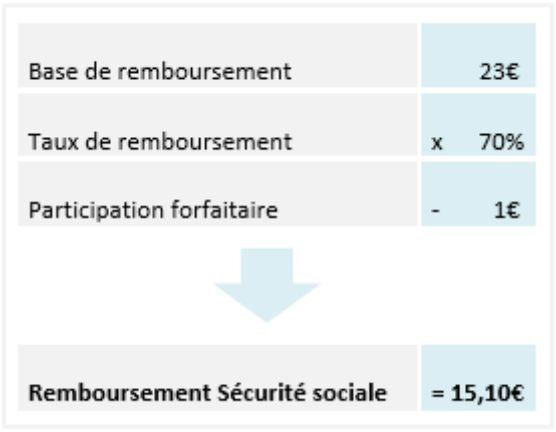


FIGURE 2 – Remboursement de la Sécurité Sociale dans le cas d'une consultation chez le généraliste

1.3.3 Le parcours de soins coordonnés et les franchises médicales

Afin de faire face au déficit financier de l'assurance maladie, la réforme Douste-Blazy a été mise en œuvre entre 2004 et 2007. L'objectif était de responsabiliser les bénéficiaires de l'assurance maladie (assurés et ayants droit) pour ainsi économiser plusieurs milliards d'euros par an. Cette réforme a notamment instauré un parcours de soins coordonnés.

Chaque bénéficiaire d'une couverture maladie est incité à désigner un médecin traitant, qui est le pivot du système. Ce médecin traitant pourra éventuellement orienter le patient vers un médecin spécialiste. En respectant le parcours de soins coordonnés, le bénéficiaire a une meilleure prise en charge des dépenses par la Sécurité sociale. Ainsi, le montant pris en charge par le régime de la Sécurité sociale dépend également de ce dispositif.

D'autres mesures ont été mises en place à partir du 1er janvier 2008 afin de responsabiliser les bénéficiaires d'une couverture maladie, telles que les franchises médicales.

Les montants de ces franchises sont :
- 0,5 € par boîte de médicaments ;
- 0,5 € par acte paramédical ;
- 2 € pour chaque recours au transport sanitaire.

Ces franchises sont plafonnées annuellement à hauteur de 50 € pour l'ensemble des actes et prestations concernées. Un plafond journalier est également appliqué, soit 2 € pour les actes paramédicaux et 4 € pour les transports sanitaires. Elles ne sont pas déduites des remboursements de la Sécurité sociale pour les personnes âgées de moins de 18 ans, les bénéficiaires de la CMU et les femmes enceintes.

Chapitre 2

Les complémentaires santé

2.1 Le remboursement de la complémentaire santé

2.1.1 Le principe de remboursement

La complémentaire santé (mutuelle, institution de prévoyance ou société d'assurance) permet aux bénéficiaires d'une couverture d'assurance maladie, de bénéficier d'un remboursement complémentaire à celui de la Sécurité sociale. Elle rembourse tout ou une partie des frais réels. En effet, après le remboursement de la Sécurité sociale, il reste généralement un montant à charge de l'assuré. Il s'agit d'un montant composé de la participation forfaitaire, du ticket modérateur et des éventuels dépassements. Le ticket modérateur correspond à la différence entre la base de remboursement et ce que rembourse l'assurance maladie obligatoire. En général, les frais réels ne sont pas équivalents au ticket modérateur puisque des dépassements d'honoraires peuvent être pratiqués par les médecins conventionnés et non conventionnés. Il s'agit de la partie des honoraires qui excède la base de remboursement. Ainsi, les bénéficiaires d'une couverture santé peuvent avoir recours à une complémentaire santé afin de réduire les coûts de leur consommation.

Par exemple, dans le cas d'une consultation au régime général de médecin généraliste pratiquant des dépassements d'honoraires :

- Coût de la consultation : 25 €
- Ticket modérateur : 23 € - remboursement théorique de la Sécurité sociale (participation forfaitaire inclus) = 6,90 €
- Dépassements : 25 € - 23 €= 2 €

Dans cet exemple, la complémentaire santé pourra rembourser partiellement ou totalement le montant de 8,90 € résultant du ticket modérateur et des dépassements.

Les organismes d'assurance peuvent également prendre en charge des dépenses qui ne bénéficient d'aucun remboursement par l'assurance maladie obligatoire tels que certains médicaments, des lentilles de contact, etc.

2.1.2 Les expressions de garanties

Sur le marché de l'assurance, il existe différentes formes d'expressions de remboursement utilisées par les complémentaires santé. Il convient de distinguer le pourcentage exprimé, si le remboursement inclut ou non le remboursement de la Sécurité sociale, l'existence d'un plafond, et la base de calcul.

Par exemple, le remboursement par la complémentaire santé des frais de consultation d'un généraliste peut être exprimé par les deux expressions ci-dessous :

- 80% des frais réels ;
- 100% de la base de remboursement en plus du remboursement de la Sécurité sociale

Les différentes expressions de garanties sont décrites ci-dessous :

- Remboursement en fonction de la base de remboursement (BR)
  Cette expression est souvent utilisée dans le cas des actes de soins courants. Le remboursement en fonction de la base de remboursement peut donner des résultats différents si nous considérons que le remboursement de la Sécurité sociale est inclus (BR-RSS) ou non (BR en sus) dans le remboursement de la complémentaire santé

- Remboursement en fonction du remboursement du régime obligatoire (RSS)
  Il s'agit d'un pourcentage exprimé en fonction du montant remboursé par la Sécurité sociale.

- Remboursement en fonction d'un forfait
  Il s'agit d'un montant en euros. Ce montant peut être en fonction du plafond mensuel de la Sécurité sociale [5]. Ce forfait peut être également accompagné d'un remboursement en fonction de la base de remboursement ou d'autres expressions définies ci-dessus. Par exemple, pour une prothèse auditive, la garantie peut être un remboursement de 200% de la base de remboursement et un forfait de 150 €.

- Remboursement en fonction des frais réels (FR)
  Il s'agit d'un pourcentage des dépenses totales. Autrement dit, la complémentaire santé remboursera un pourcentage du prix de l'acte. Ainsi, ce pourcentage ne pourra pas dépasser 100%.

Il est à noter que cette liste n'est pas exhaustive et que les remboursements se font dans la limite des frais réels restant à la charge de l'assuré suite au remboursement de la Sécurité sociale.

5. Le plafond mensuel (3 129€ en 2014) de la Sécurité sociale est utilisé dans le calcul de certaines cotisations sociales et de certaines prestations de la Sécurité sociale.

2.2 Les différents types de contrats santé

En santé, les assureurs proposent plusieurs types de régimes complémentaires. Nous distinguons les régimes individuels des régimes collectifs qui peuvent avoir un caractère obligatoire ou facultatif selon le type d'adhésion. Contrairement aux contrats individuels où l'adhérent souscrit directement chez l'assureur, un contrat collectif est conclu entre une personne morale et l'assureur et vise à couvrir des adhérents. Une description plus précise des différents contrats est effectuée ci-dessous.

Contrat collectif à adhésion obligatoire : Généralement dans le cadre d'une entreprise. Le caractère obligatoire impose à tous les salariés de l'entreprise ou à tous les membres de la catégorie de personnel concernée par le contrat, d'adhérer au régime [6] et impose également à l'assureur d'accepter tous les adhérents.
Ce type de contrat sera obligatoire pour toutes les entreprises dans le cadre de l'ANI.

Contrat collectif à adhésion facultative : Les adhérents ne sont pas contraints, mais ont la possibilité d'adhérer au régime. L'assureur peut alors être confronté à des problèmes d'antisélection, puisque les salariés anticipant de fortes dépenses de santé choisiront de souscrire alors que ceux anticipant de plus faibles dépenses ne souhaiteront pas adhérer au régime ou se limiteront à de faibles garanties (cf. partie 4).

Contrat collectif à adhésion obligatoire et facultative : Il s'agit d'un contrat à adhésion obligatoire tel que défini précédemment qui comporte en complément des garanties ou des options facultatives au choix du salarié.

2.3 Les différents types de cotisation

Généralement, les assureurs proposent plusieurs types de cotisations à l'employeur dans le cadre de contrats collectifs. La prime finale du contrat dépend du type de cotisation choisi par l'employeur. A titre d'exemple, les Assurances du Crédit Mutuel proposent les cotisations suivantes :

- cotisation de type « adulte/enfant » : chaque salarié est assuré en tarif adulte et peut assurer son conjoint en tarif adulte et son enfant en tarif enfant.
- cotisation de type « famille » : tous les salariés sont assurés en tarif famille indépendamment de la composition familiale (un salarié souhaitant affilier conjoint et enfants paiera ainsi la même prime qu'un salarié adhérent seul).
- cotisation de type « isolé/famille » : si le salarié s'assure seul, il bénéficiera d'un tarif isolé, et dans le cas où il ne s'assure pas seul, d'un tarif famille.
- cotisation de type « 1 assuré / 2 assurés / 3 assurés et plus » : chaque salarié peut s'assurer seul, ou assurer une autre personne, ou assurer deux autres personnes et plus.

6. Cette condition n'est pas vérifiée dans le cas des contrats mis en place par décision unilatérale (écrit accordant un avantage supplémentaire par rapport aux contrats de travail) pour lesquels les salariés présents au moment de la mise en place ont le choix d'adhérer ou non.

Chapitre 3

Le contexte actuel de l'assurance santé

3.1 Quelques chiffres

Avant d'étudier les données et de présenter une méthode de tarification des contrats complémentaires santé, il peut être intéressant d'étudier les enjeux actuels.

La santé fait partie d'un des principaux postes de consommation des français. En 2012, si nous ajoutons à la consommation en santé des ménages, les dépenses de consommation en santé des administrations publiques en biens et services individualisables, la santé constitue le deuxième poste de consommation en France après le logement [7].

Postes de consommation	**Répartition de la consommation totale**
Logement, chauffage, éclairage	19.5 %
Santé	12.6 %
Produits alimentaires et boissons non alcoolisées	10.4%
Articles d'habillement et chaussures	3.2%

Table 1 – Les grands postes de consommation des ménages

D'après les comptes de la santé publiés par la Drees, les dépenses courantes de santé ont augmenté de 60% [8] entre 2000 et 2012, soit une évolution de 151 à 243 milliards d'euros par an. Cependant les dépenses totales de santé regroupent différentes catégories de dépenses telles que les indemnités journalières, la dépendance, la formation, la gestion, etc. C'est pourquoi, dans le cadre de ce mémoire, il est plus adapté de suivre l'évolution de la CSBM (consommation de soins et biens médicaux).

7. Source : Insee, Structure des dépenses de consommation des ménages, données 2012.
8. Données recueillies dans le magazine l'actuariel n°11, Janvier 2014, page 21.

Celle-ci regroupe les catégories suivantes :
– les soins hospitaliers ;
– les soins courants : médecins, dentistes et auxiliaires médicaux ;
– les médicaments et autres bien médicaux (optique, prothèses etc.) ;
– le transport de malades.

La CSBM augmente chaque année et s'élève en 2012 à 10,5 [9] milliards d'euros. Son taux de croissance en valeur atteint 2,2% par rapport à 2011. Le graphique ci-dessous illustre l'évolution de ce taux depuis l'année 2000, un ralentissement est observé après 2002, encore plus marqué depuis 2010.

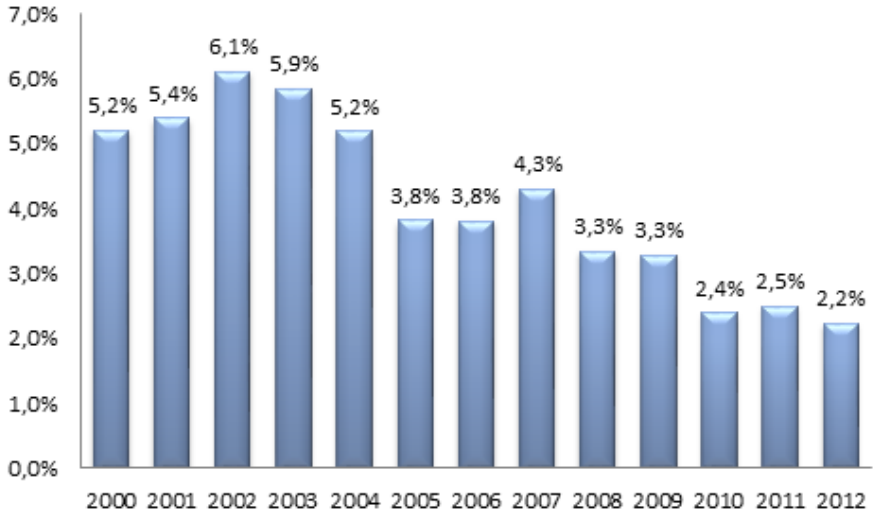


FIGURE 3 – Taux de croissance de la CSBM

Outre l'augmentation de la CSBM, la prise en charge de ces dépenses par la Sécurité sociale est en légère baisse. Ainsi, la part prise en charge par les couvertures complémentaires s'accroît année après année. Entre 2000 et 2012, d'après la DREES, elle est passée de 12,4% à 13,7%. Cette évolution s'explique également par la hausse du coût des dépenses de santé (notamment liés aux progrès de la médecine), l'allongement de l'espérance de vie et l'évolution des pratiques de consommation médicale.

3.2 L'accord national interprofessionnel

Le secteur de l'assurance santé est aujourd'hui marqué par des changements importants. L'Accord National Interprofessionnel (ANI) du 11 janvier 2013 oblige les entreprises à souscrire un contrat frais de santé pour l'ensemble de leurs salariés. Établi entre les syndicats et les organisations patronales, il devra être appliqué avant le 1er janvier 2016.

9. Source : DREES, Comptes nationaux de la santé, données 2012.

L'accord concerne toutes les entreprises du secteur privé avec au moins un salarié et prévoit une participation partielle de l'employeur, a minima à hauteur de 50%.

Il s'agit ici d'un enjeu important pour les assureurs puisqu'une partie non négligeable des salariés ayant une complémentaire santé individuelle va résilier son contrat. Un transfert des salariés du marché de l'assurance santé individuelle vers le marché de l'assurance santé collective devrait s'opérer

3.2.1 Les accords de branche

Avant le 1er juillet 2014, cet accord pouvait relever d'un accord de branche. En effet, au sein d'une branche professionnelle (exemple : branche automobile, branche coiffure, etc.), les partenaires sociaux et les syndicaux peuvent négocier afin de déterminer le contenu du contrat proposé par la branche et le niveau des garanties.

Par ailleurs, ils peuvent également recommander un ou plusieurs organismes assureurs. Il est à noter qu'avant le 13 juin 2013, les branches professionnelles avaient la possibilité de désigner un organisme assureur pour l'ensemble des entreprises du secteur. Cette disposition inscrite dans le Code de la Sécurité sociale a été censurée par le Conseil Constitutionnel, car ces clauses de désignations portaient « *à la liberté d'entreprendre et à la liberté contractuelle une atteinte disproportionnée au regard de l'objectif poursuivi de mutualisation des risques* » [10].

Après le 1er juillet 2014, si les négociations au niveau des branches professionnelles ont échoué, les syndicats pourront négocier jusqu'au 31 décembre 2015. Après cette date, toutes les entreprises seront contraintes de proposer un contrat santé obligatoire à l'ensemble de leurs salariés, dont le contenu inclut les garanties minimales prévues par décret (au stade de projet lors de la rédaction de ce mémoire).

3.2.2 Le panier de soins minimum

Les entreprises ou les branches professionnelles pourront librement définir leurs garanties, à condition de respecter les garanties minimales, appelées « panier minimum de soins », fixées par décret, non encore paru.

Le projet de décret prévoit le contenu du panier minimum suivant :
- le remboursement du ticket modérateur pour la majorité des actes ;
- le remboursement total du forfait journalier ;
- la prise en charge supérieure au montant du ticket modérateur, soit 125% (BR-RSS), dans le cadre du remboursement des prothèses dentaires et de l'orthodontie ;
- un forfait optique de 100 euros par an.

10. Décision n°2013-672 DC du 13 juin 2013 du Conseil Constitutionnel, relative à la loi sur la sécurisation de l'emploi

3.2.3 Le contrat responsable

En vigueur depuis le 1er janvier 2006, la notion de contrat responsable de complémentaire santé est utilisée lorsque la complémentaire santé respecte des obligations de remboursement et des interdictions de remboursement prévues par la loi, afin de bénéficier des aides fiscales et sociales.

La notion de contrat responsable est en cours d'évolution. Le projet précise notamment des plafonds de remboursement en soins courants et en optique.

Deuxième partie

L'analyse préliminaire des données

La tarification d'un contrat d'assurance santé requiert une analyse préliminaire des données. Pour cela, il est d'abord nécessaire de présenter précisément le produit étudié et les données à disposition. Pour l'analyse statistique des données, l'étude démographique et l'analyse des variables affectant le tarif d'un contrat santé seront développées. Une étude particulière du lieu d'habitation de l'assuré sera approfondie par l'utilisation de deux méthodes statistiques : l'analyse en composantes principales et la classification ascendante hiérarchique.

Chapitre 1

Le produit étudié

1.1 Les contrats collectifs sur mesure

Dans le cadre de ce mémoire, un tarif sur mesure sera déterminé pour les contrats collectifs, à adhésion facultative ou obligatoire. Le caractère facultatif des contrats sera pris en compte à travers un coefficient d'antisélection analysé et déterminé dans la dernière partie de ce mémoire.

Ces contrats ne font pas partie de la gamme « Standard » des Assurances du Crédit Mutuel, avec des garanties prédéfinies. L'objectif est de proposer un tarif pour des garanties modulables par l'entreprise pour chaque poste de garanties, en tenant compte de l'existence de différents types d'expressions de garantie.

Le montant total de la cotisation due par l'entreprise doit permettre de faire face aux engagements de l'assureur et tenir compte des frais de fonctionnement. Les engagements de l'assureur sont déterminés en agrégeant les primes pures de chaque assuré, c'est-à-dire le coût de l'assurance au niveau de chaque assuré. Par conséquent, l'étude portera sur la détermination de la prime pure notée P qui peut être décomposée, en deux éléments : le nombre de sinistres N et le coût du remboursement moyen S, sous l'hypothèse d'indépendance de ces deux grandeurs :

$$P = E(N) \times E(S)$$

où, $E(N)$ est le rapport entre le nombre d'actes observé durant la période d'observation et le nombre d'assuré (fréquence de consommation ou de sinistres), et $E(S)$ se déduit du rapport entre le coût total observé sur la période d'observation et le nombre de sinistres.

1.2 Les différents postes de garanties étudiées

Un contrat de complémentaire santé prévoit la couverture de nombreux postes de garanties. Les garanties et sous-garanties traitées dans le cadre de ce mémoire seront étudiées et analysées sous la structure suivante :

Catégories d'actes	**Sous catégories d'actes**
Soins courants	Consultations et visites chez les généralistes Consultations et visites chez les spécialistes Actes de petite chirurgie effectués chez les médecins Transports Auxiliaires médicaux Pharmacie Analyses et actes en laboratoire Radiologie et imagerie médicale
Hospitalisation	Honoraires médicaux ou chirurgicaux Forfait hospitalier Frais de séjour Chambre particulière Maternité Télévision Lit accompagnant
Optique	Montures Verres Lentilles Chirurgie réfractive
Dentaire	Consultations et soins dentaires Prothèses dentaires prises en charge par la Sécurité sociale Prothèses dentaires non prises en charge par la Sécurité sociale Orthodontie
Autres	Orthopédie Prothèses auditives Grand appareillage Cure thermale

TABLE 2 – Les catégories et sous catégories de garantie étudiées

Chapitre 2

La composition du portefeuille

2.1 La description du portefeuille

Les données ont été extraites sur les trois dernières années de la base de données santé des Assurances du Crédit Mutuel, c'est-à-dire 2011, 2012 et 2013 afin d'éviter d'accorder de l'importance à des événements relatifs à une année particulière. Afin d'augmenter le volume de données, les données relatives aux contrats collectifs et aux contrats individuels ont été retenues. Par ailleurs, l'utilisation de données individuelles permet d'effectuer une tarification par niveaux de garanties. En effet, la majorité des contrats collectifs concerne des contrats avec des garanties sur-mesure, pour lesquels l'exploitation des niveaux de garanties est complexe.

Deux fichiers de données ont été créés à l'aide du logiciel SAS 9.3 [11] :

- fichier contenant les effectifs :
  pour chaque bénéficiaire de la couverture complémentaire, le fichier dispose d'informations relatives au contrat, relatives à l'assuré et aux autres bénéficiaires de la couverture, soit environ 913 000 données.

- fichier contenant les sinistres :
  pour chaque sinistre, le fichier indique le remboursement du régime obligatoire, le remboursement des Assurances du Crédit Mutuel, le montant des frais réels, le reste à charge, la catégorie et sous-catégorie du poste de garantie et ainsi que l'identifiant de l'assuré. Nous dénombrons au total 70 millions de sinistres pour l'ensemble des catégories d'actes sinistres.

Un fichier final regroupant les deux fichiers ci-dessus a également été créé pour disposer à la fois des informations sur les sinistres et des informations sur les bénéficiaires. Chaque ligne de ce fichier correspond à l'agrégation du nombre de sinistres et du coût des sinistres par bénéficiaire selon la catégorie d'acte.

11. Les traitements statistiques et les graphiques de ce mémoire ont été réalisés à l'aide du logiciel SAS version française 9.3.

2.2 Le traitement des données

Avant d'entamer le traitement des données, il est nécessaire de prendre en compte la durée de présence d'un assuré au sein de la base de données, puisqu'un assuré présent deux mois a une plus forte probabilité de consommer une faible quantité d'actes qu'un assuré présent durant toute la période d'observation. Ainsi, la variable « année risque » a été créée pour tenir compte de la fraction de temps de présence dans la base de données.

Ensuite, en vue d'obtenir un tarif précis et juste, il est important d'étudier et d'analyser les données à disposition afin de traiter les données aberrantes (valeurs non cohérentes) et manquantes. Dans un premier temps, les années risques à valeurs négatives causées par l'inversion de la date de début et de fin de garantie ont été corrigées en sélectionnant la valeur absolue des années risques. La variable recensant la catégorie socio-professionnelle des assurés n'a pas été retenue, par manque de fiabilité. Des valeurs manquantes ont été observées concernant notamment le sexe des assurés et des bénéficiaires, ces contrats n'ont pas été supprimés pour ne pas perdre de l'information. D'autres vérifications ont été effectuées sur les données et aucune valeur aberrante n'a été observée.

Enfin, afin de constituer les bases finales à utiliser pour la tarification, il a été nécessaire de vérifier les comportements de consommation par année d'observation (pour visualiser d'éventuelles tendances) et par type de contrat.

Pour chaque année, les fréquences de consommation et le coût moyen des différents actes ont été calculés en fonction de l'âge des bénéficiaires. Les représentations graphiques de ces grandeurs en fonction de l'âge et par année ont permis de visualiser les écarts de fréquence et de coût moyen entre chacune des trois années étudiées. Les courbes étant pratiquement superposées, le choix de retenir les trois dernières années pour la tarification a été confirmé.

2.3 Les variables tarifaires

L'idée en tarification est la segmentation des risques, c'est à dire de créer des catégories de risques dans lesquelles se trouvent les assurés présentant des caractéristiques assez similaires en matière de risque santé. Autrement dit, les assurés disposant d'un comportement de consommation semblable s'agissant des dépenses de santé seraient regroupés dans une même catégorie de risque, permettant à l'assureur de proposer un tarif adéquat à ce profil. Afin de former ces catégories, nous disposons d'informations décrites dans la partie description du portefeuille. Ainsi parmi ces informations, les variables ci-dessous, qui pourraient avoir une influence sur les dépenses de santé, ont été préalablement sélectionnées :

- l'âge de l'assuré ;
- le sexe de l'assuré ;
- le régime d'affiliation à la Sécurité sociale (régime général ou régime local) ;
- la région et le département ;
- le niveau de garantie des contrats, pour les contrats standards (niveaux 1, 2, 3, 4 et 5) ;
- le type de bénéficiaire (adhérent, conjoint, enfant ou autres) ;
- le nombre d'enfants.

S'agissant de la tarification d'un contrat collectif, les régimes d'adhésion autres que le régime général et le régime Alsace-Moselle ne sont pas concernés par notre étude.

Chapitre 3

Statistiques descriptives et analyse des données

Avant d'entamer la tarification, des statistiques descriptives et des analyses statistiques sont effectuées afin de dresser un profil des données. Les variables tarifaires traditionnellement utilisées en assurance santé sont l'âge, le sexe, le régime d'adhésion et le niveau de garantie du contrat. Nous proposons dans ce chapitre d'analyser l'influence du lieu d'habitation, du type de bénéficiaire et du nombre d'enfants sur la sinistralité.

3.1 L'étude démographique

L'objet de cette partie est d'étudier la répartition des assurés en fonction des critères de segmentation sélectionnés afin de prendre connaissance du profil de risque du portefeuille étudié.

Le graphique ci-dessous illustre la composition du portefeuille en distinguant pour chaque âge, l'effectif en nombre d'années risques des femmes et des hommes. Il comporte également une distinction entre les assurés affiliés au régime général et ceux assurés au régime local.

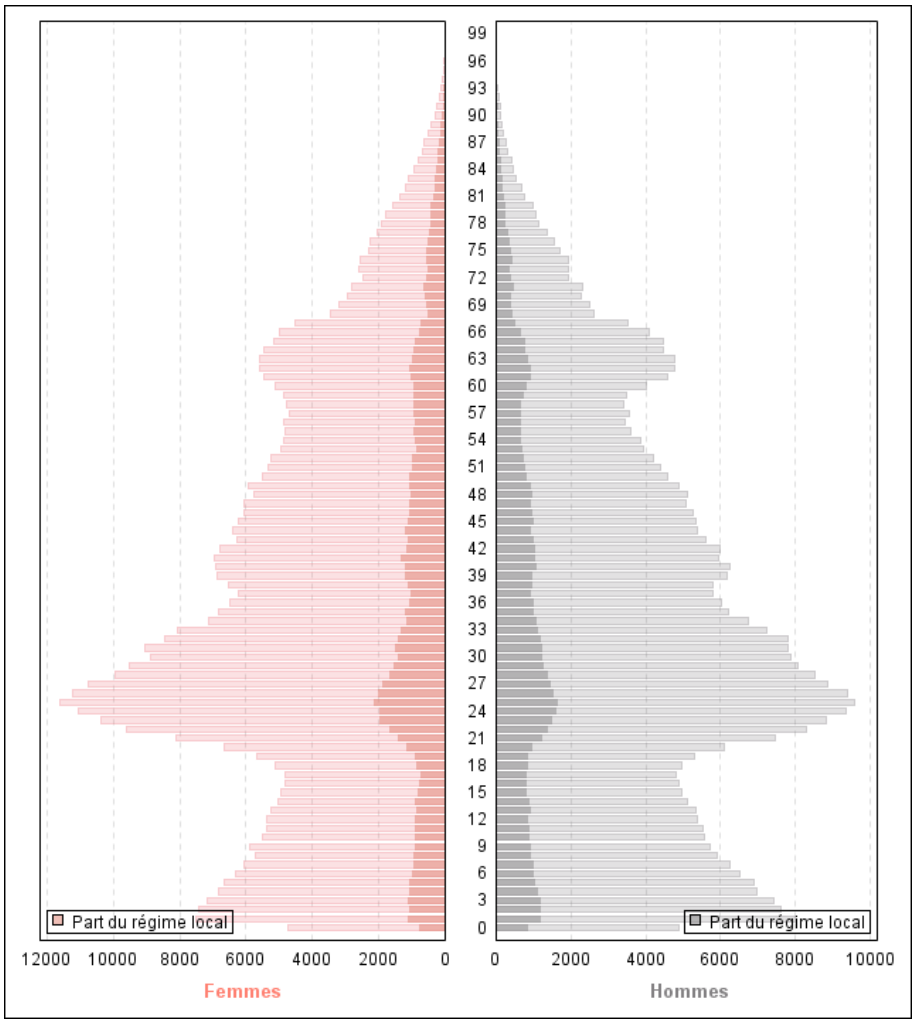


FIGURE 4 – Répartition du portefeuille par âge et par régime d'adhésion

L'effectif semble assez bien réparti entre les hommes et les femmes, bien que le nombre de données diminue lorsque l'âge de l'assuré augmente. Malgré une proportion faible de l'effectif au régime Alsace-Moselle, les données du régime Alsace-Moselle semblent être exploitables. En effet, le groupe Crédit Mutuel étant bien implanté en Alsace-Moselle, nous

disposons de suffisamment de données pour décomposer le portefeuille en fonction du régime d'adhésion de l'assuré. Néanmoins, il faut être vigilant sur le degré de segmentation.

La répartition des contrats par niveau de garanties a également été étudiée. Le groupe commercialise des contrats individuels avec cinq niveaux de garantie : 1, 2, 3, 4 et 5. Comme nous ne disposons pas d'informations suffisantes sur les niveaux de garantie concernant les contrats collectifs, il a été convenu d'utiliser uniquement les contrats individuels dans le cas d'une segmentation par niveau de garantie. La différence de consommation entre les contrats groupes et individuels sera prise en compte à travers un coefficient d'antisélection déterminé dans la partie 4 du mémoire.

	1	**2**	**3**	**4**	**5**	**Total**
General	19,78%	37,90%	33,59%	7,81%	0,92%	100%
Local	0,00%	19,44%	44,68%	30,29%	5,59%	100%
Global	16,58%	34,91%	35,38%	11,45%	1,68%	100%

TABLE 3 – Répartition du portefeuille par niveau de garantie et par régime d'adhésion

Les niveaux de garantie les plus présents dans le portefeuille sont le deuxième et troisième niveau. Il faut porter une attention particulière au cinquième niveau de garantie qui est souscrit par uniquement 1,68% des assurés. La répartition des niveaux de garantie en fonction des deux régimes est différente de la répartition globale. Pour le régime local, environ un tiers des assurés ont un niveau de garantie égal à 4, alors que les assurés du régime général ayant souscrit une garantie de niveau 4 représentent uniquement 7,81% de l'ensemble des assurés du portefeuille. Ainsi, les assurés du régime local ont tendance à souscrire des garanties plus élevées que les assurés du régime général. Cela s'explique par un reste à charge plus faible suite au remboursement de la Sécurité sociale pour les assurés du régime local, qui induit un plus faible remboursement de la complémentaire santé.

3.2 La consommation en fonction de l'âge

L'objectif de cette sous-partie est d'analyser l'effet de l'âge sur la fréquence de consommation et le coût moyen des actes, ce qui nous permettra de réaliser des regroupements au niveau de la variable tarifaire « âge ». Ces regroupements seront utilisés dans la tarification des contrats, puisqu'ils permettront d'augmenter le volume des données pour chaque catégorie d'assuré. Les classes d'âge pourraient être attribuées de façon arbitraire, par exemple par intervalle de cinq ans. Cependant, en santé, l'âge de l'assuré influence très significativement sa consommation. Nous observons en général une forme « en W » avec un pic de consommation à la naissance, à l'adolescence puis une phase de croissance dont le rythme s'accélère avec l'âge. Cette consommation varie également selon la famille d'actes considérée. Par exemple, en dentaire, les enfants consomment beaucoup plus que les adultes (dû à l'orthodontie), alors qu'en pharmacie ce n'est pas le cas.

C'est pourquoi, afin d'obtenir un tarif plus précis par la suite, la formation des classes d'âge s'est effectuée en analysant les courbes de consommation en fonction de l'âge, pour

chaque garantie étudiée. Il s'agit de repérer les âges où le comportement de consommation est similaire et de les regrouper.

3.2.1 Les consultations généralistes

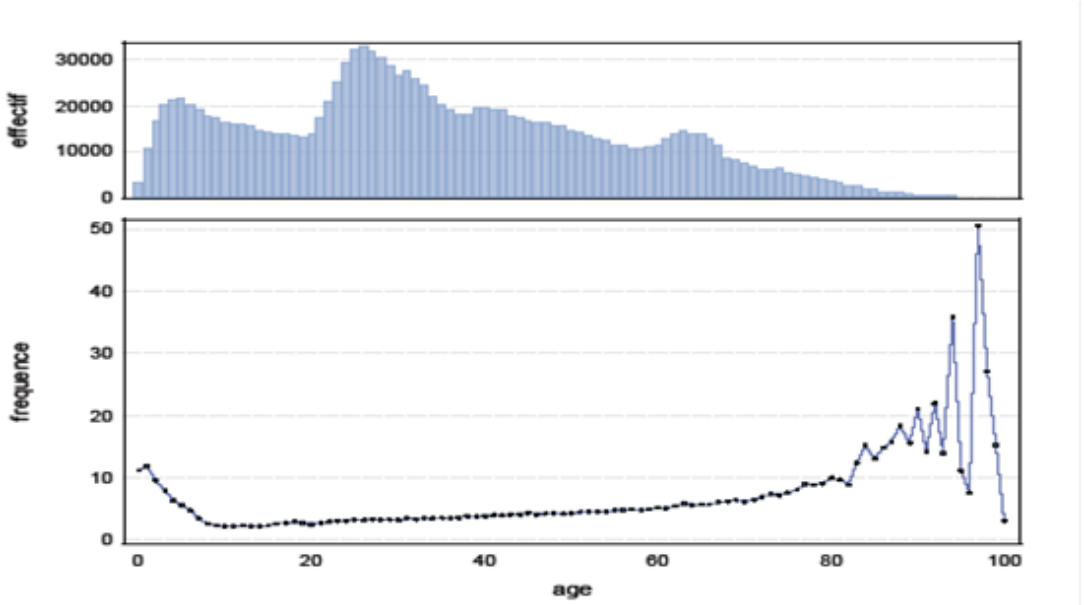


FIGURE 5 – Fréquence de consommation en actes de consultations généralistes en fonction de l'âge

En ce qui concerne la fréquence de consommation pour les consultations généralistes, nous observons une influence de l'âge entre 0 et 9 ans où il convient de créer des classes plus petites. Pour les âges supérieurs à 75 ans, le nombre de données étant faibles, il est nécessaire de créer une seule classe. Par conséquent, nous avons convenu de créer les classes d'âge suivantes :

- entre 0 et 1 an ;
- entre 2 et 3 ans ;
- entre 4 et 5 ans ;
- entre 6 et 7 ans ;
- entre 8 et 20 ans ;
- entre 21 et 40 ans ;
- entre 41 et 60 ans ;
- entre 61 et 70 ans ;
- entre 71 et 75 ans ;
- supérieur ou égal à 76 ans.

3.2.2 Les prothèses dentaires

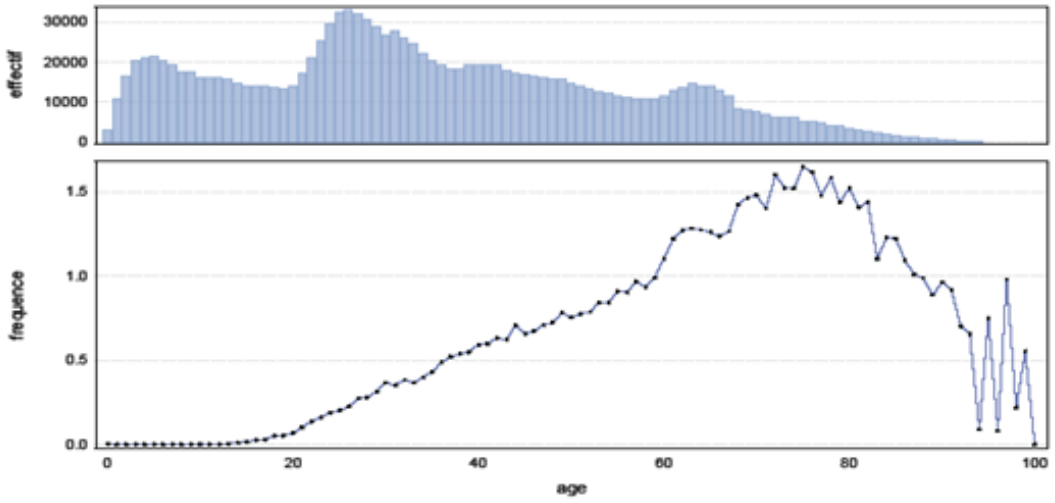


FIGURE 6 – Fréquence de consommation de prothèses dentaires en fonction de l'âge

Pour ce poste de garantie, l'effet de l'âge est différent. La consommation de prothèses dentaires commence uniquement à partir de l'âge de 18 ans, augmente et atteint la fréquence maximale vers 75 ans. Les classes d'âges suivantes ont été créées en tenant compte de la quantité de données :

- strictement inférieur à 20 ans ;
- entre 21 et 25 ans ;
- entre 26 et 30 ans ;
- entre 31 et 35 ans ;
- entre 36 et 40 ans ;
- entre 41 et 45 ans ;
- entre 46 et 50 ans ;
- entre 51 et 55 ans ;
- entre 56 et 60 ans ;
- entre 61 et 65 ans ;
- entre 66 et 70 ans ;
- entre 71 et 75 ans ;
- supérieur ou égal à 76 ans.

3.3 L'analyse des nouvelles variables tarifaires

Parmi les variables présentes dans la base de données, certaines telles que l'âge ou le sexe influent la consommation de façon évidente, contrairement à d'autres variables. Dans cette partie, nous étudierons brièvement l'influence de la variable recensant le type de bénéficiaire et de la variable comptant le nombre d'enfants par adhérent.

Les personnes couvertes par un contrat santé peuvent demander l'adhésion de leurs ayants droits, c'est-à-dire les conjoints ou concubins, enfants et ascendants. La variable type de bénéficiaire permet ainsi de différencier l'adhérent de ses ayants droits. Afin d'examiner l'effet de cette variable sur la consommation en santé, la fréquence de consommation et le coût moyen ont été calculés :

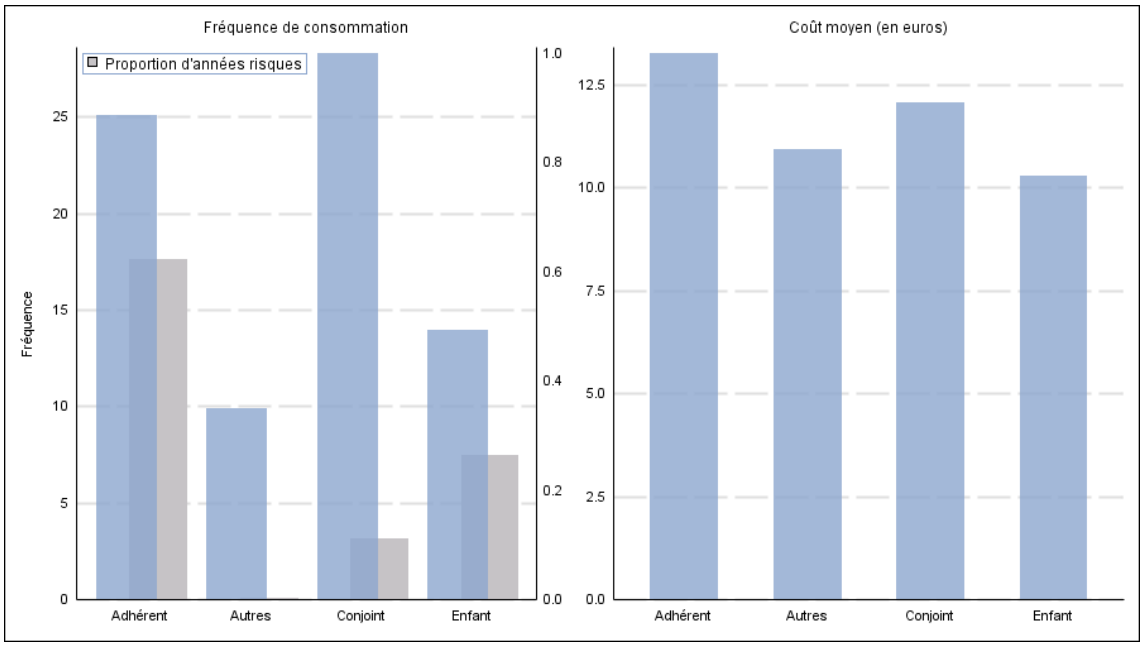


FIGURE 7 – Fréquence de consommation et coût moyen par type de bénéficiaire

La modalité « autres » contenue dans notre portefeuille correspond à un ascendant ou autre type d'ayant droit de l'adhérent. La fréquence de consommation de cette catégorie n'est pas fiable, puisque l'effectif est très faible. Les conjoints et adhérents semblent avoir des comportements de consommation proches que ce soit pour la fréquence ou le coût moyen, contrairement aux enfants et autres bénéficiaires. Toutefois, nous constatons graphiquement que la fréquence de consommation des conjoints est plus élevée que celle des adhérents et que le coût moyen des conjoints est plus faible que celui des adhérents. Ainsi, le type de bénéficiaire peut influer la consommation de frais de santé. Il s'agit ici d'une analyse globale, tous les postes de garanties confondus. Par conséquent, cet effet sera plus ou moins marqué selon les postes de garanties.

Les mêmes analyses ont été effectuées pour la variable comptant le nombre d'enfants par adhérent.

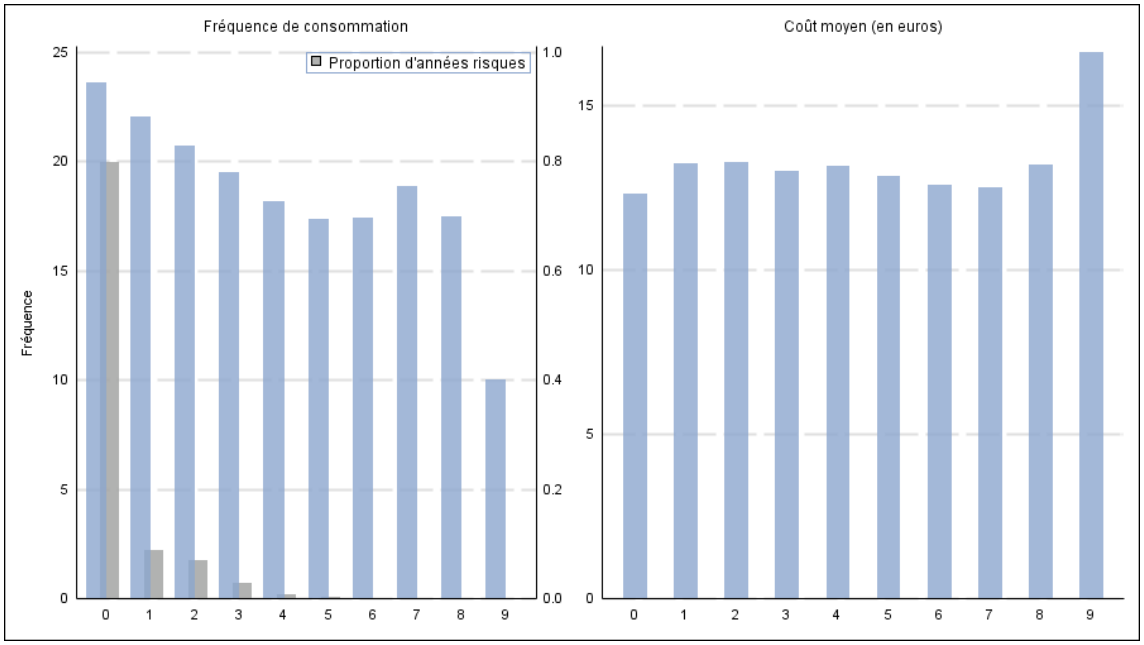


FIGURE 8 – Fréquence de consommation et coût moyen en fonction du nombre d'enfant

Cette variable a peu d'effet sur le coût moyen par acte contrairement à la fréquence de consommation de l'assuré. Néanmoins nous choisissons de retenir cette variable pour la tarification des coûts moyens puisqu'elle peut avoir plus ou moins d'effet selon les différents postes de garanties.

Globalement, la fréquence de consommation de l'assuré diminue en fonction du nombre de ses enfants. Les différences de coût et de fréquence pour les dernières modalités (nombre d'enfants très élevé) ne sont pas fiables puisqu'il y a très peu de données. Afin d'exploiter cette variable, il convient de créer une nouvelle modalité regroupant le nombre d'enfant supérieur ou égal à 3. Ainsi, la variable sera composée des modalités 0, 1, 2 et supérieur à 3.

3.4 L'analyse en composantes principales (ACP) sur le lieu d'habitation

3.4.1 Le principe de l'ACP

L'analyse en composantes principales est une technique de statistique multidimensionnelle permettant d'analyser les liaisons entre plusieurs variables quantitatives simultanément (cf. [6]). L'objectif est de passer d'un espace à p dimensions à un espace de dimension inférieur à p en perdant le moins possible d'informations du tableau de données initial.

Cette technique consiste à projeter orthogonalement le nuage des individus sur un plan factoriel, plan passant au plus près des individus du nuage. Cela permet de créer de nouvelles variables artificielles, nommées les « axes factoriels ». Mathématiquement, il s'agit de diagonaliser une matrice de variance-covariance où les vecteurs propres (axes

factoriels) et les valeurs propres (variances associées aux axes) sont extraits. De la même façon, les variables sont projetées orthogonalement sur un plan factoriel s'inscrivant dans un cercle de rayon unitaire appelé cercle de corrélation, sur lequel nous pouvons observer les corrélations entre les différentes variables.

Pour cela, nous disposons d'un tableau de données contenant pour chaque région, les fréquences de consommation et le coût moyen de chaque catégorie d'acte. Cette analyse permettra ainsi d'analyser l'effet du lieu d'habitation de l'assuré par poste de garantie. Par soucis de clarté et de lisibilité, le choix de l'ACP s'est porté sur les régions plutôt que sur les départements.

3.4.2 Les résultats

L'ACP fournit plusieurs résultats [12], et permet ainsi :

- d'établir un bilan des ressemblances entre les régions ;
- de réaliser un bilan des corrélations linéaires entre les différentes variables initiales ;
- de mettre en liaison l'étude des régions et des variables pour observer les variables caractéristiques d'un groupe d'individus donné.

Sélection du nombre d'axes

Avant toute interprétation des représentations graphiques, il est nécessaire de choisir le nombre optimal d'axes factoriels, afin d'avoir un résumé précis de l'information du tableau de données initial. Nous utilisons le critère du coude (cf. [6]) qui consiste à repérer graphiquement une cassure suivie d'une décroissance régulière sur le graphique représentant le pourcentage de variance expliquée par chacun des axes factoriels.

Le graphique ci-dessous permet de choisir le nombre d'axes pour l'étude des régions :

12. L'ACP a été réalisée avec une macro SAS téléchargée sur le site de l'INSEE.

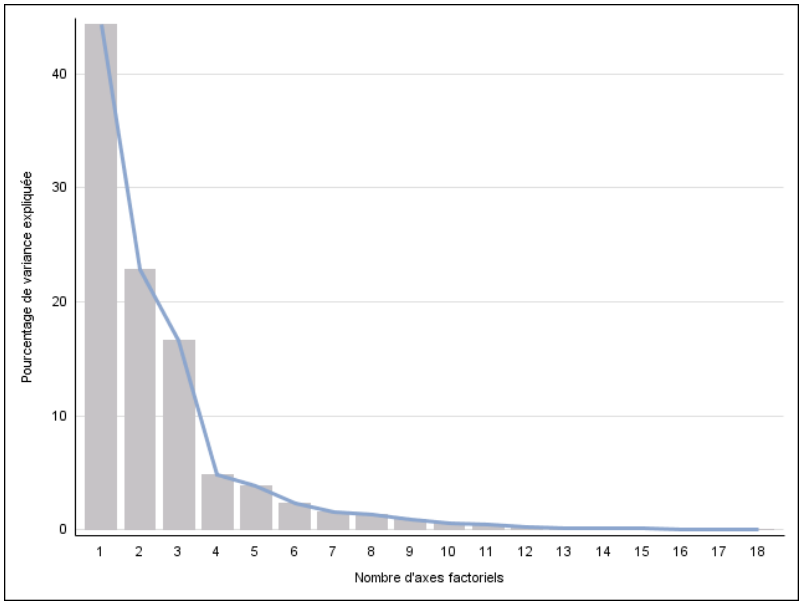


FIGURE 9 – Choix du nombre d'axe factoriel

Dans ce cas, nous observons un décrochement au niveau du quatrième axe, ce qui reviendrait à sélectionner les trois premiers axes. L'ensemble des trois axes représente 83,22% de la variance, ce qui est suffisant pour l'analyse, et confirme le choix de sélection des trois premiers axes factoriels. Cependant, dans la suite nous interpréterons uniquement les deux premiers axes factoriels, puisqu'après analyse du troisième axe factoriel, nous avons constaté qu'il n'apporte pas assez d'informations complémentaires.

Graphique des variables

L'interprétation des représentations graphiques doit se faire de façon prudente. Ainsi, nous interprèterons uniquement les variables bien représentées, c'est à dire celles proches du cercle de corrélation.

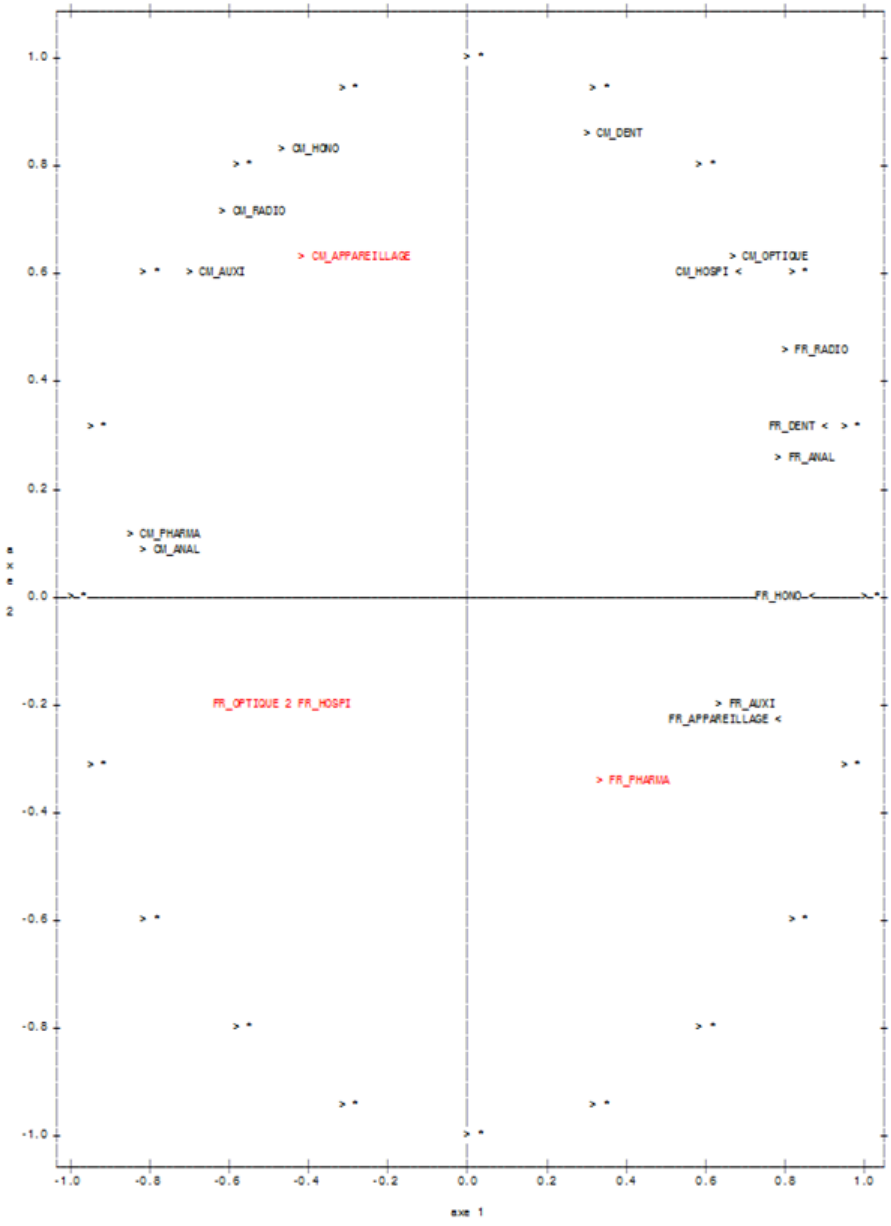


FIGURE 10 – Graphique des variables sur le premier plan factoriel

Nous rappelons ci-dessous les pourcentages de variance expliquée par les deux premiers axes factoriels :

Pourcentage de variance expliquée	
Axe 1	44,34%
Axe 2	22,78%

Dans ce premier plan factoriel composé des axes 1 et 2, l'ensemble des variables est bien représenté, à l'exception de la fréquence pharmacie, de la fréquence hospitalisation, de la fréquence optique et du coût moyen de l'appareillage.

Ce graphique permet d'observer les corrélations linéaires [13] entre les variables initiales. Nous observons notamment une forte corrélation entre la fréquence de consultations chez un auxiliaire médical [14] et la fréquence de consommation en appareillage.

Suite à l'étude des contributions des variables pour chacun des axes, nous pouvons en conclure que les fréquences ont contribué à la formation du premier axe et que les coûts moyens ont contribué à la formation du deuxième axe.

Nous observons également un effet taille sur le premier axe, avec une augmentation du coût moyen de gauche à droite. Les coûts moyens élevés (hospitalisation, dentaire et optique) se trouvant à droite du nuage, et les plus faibles (pharmacie, analyses, honoraires, radiologie, auxiliaire) à gauche du nuage.

13. Tableau des corrélations présenté en annexe A.

14. Il s'agit de professionnels de santé tels que les infirmiers, les orthoptistes, les masseurs kinésithérapeutes, etc.

Graphique des individus

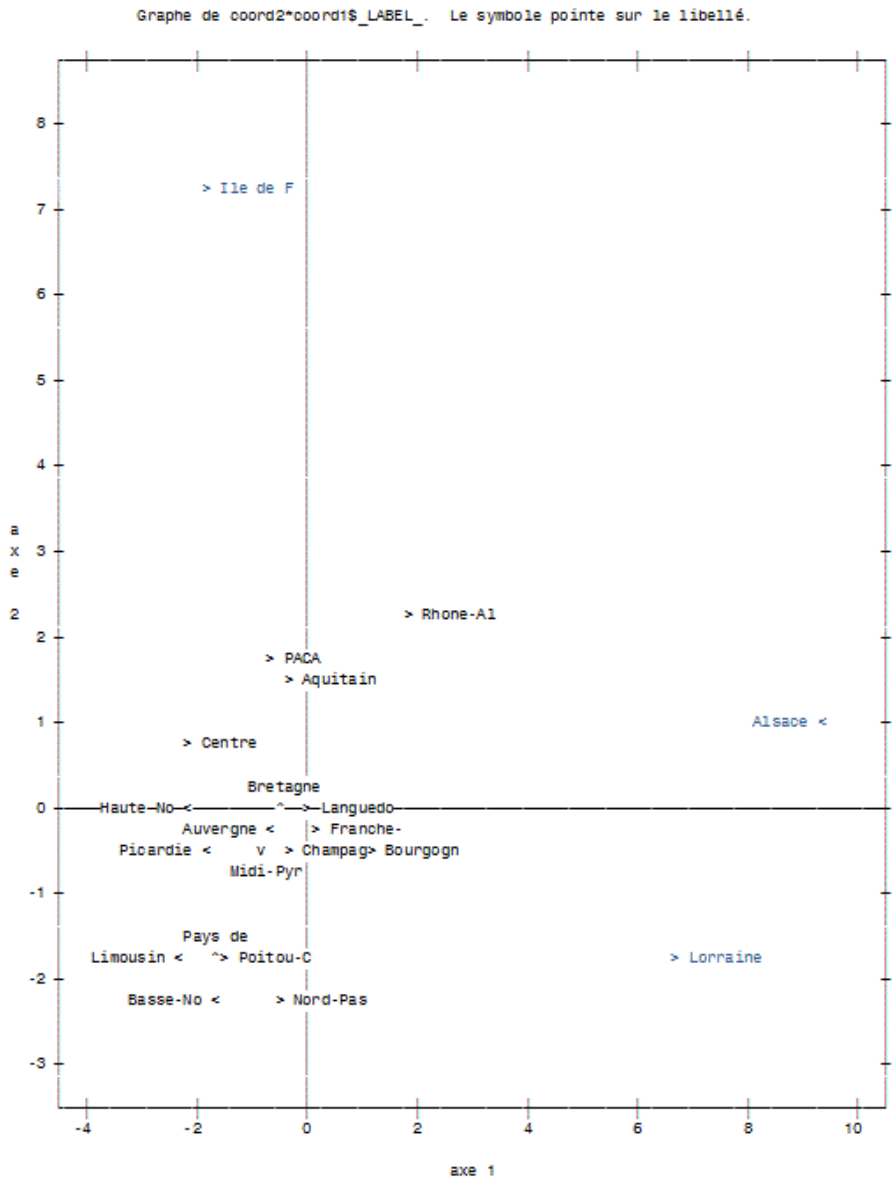


FIGURE 11 – Graphique des individus sur le premier plan factoriel

Pourcentage de variance expliquée	
Axe 1	44,34%
Axe 2	22,78%

De la même façon que les variables, les individus doivent être bien représentés sur ce plan factoriel. Pour cela, il est nécessaire que les sommes des cosinus carrés des angles formés par le vecteur initial et chacun des deux axes soient proche de 1. Cela correspond à la colonne « CO2 » (qualité de représentation) du tableau figurant dans l'annexe A.

Trois régions se détachent de façon très marquée du nuage des individus : l'Ile de France, l'Alsace et la Lorraine. La région Ile de France a les coûts moyens les plus forts pour l'ensemble des postes de garantie. Ces coûts sont d'autant plus élevés pour les honoraires et la radiologie puisque cette région est placée dans la même direction que les variables décrivant le coût moyen des honoraires et le coût moyen de la radiologie. Concernant l'Alsace et la Lorraine, les coûts moyens sont très faibles pour les soins courants dus aux plus forts remboursements du régime local : il est ici évident que ces régions sont atypiques. Cependant, pour l'Alsace les coûts moyens sont plus élevés que pour les autres régions concernant les catégories d'actes optique, dentaire et hospitalisation. Compte tenu des bases de remboursement très faible pour ces catégories d'actes, la différence de taux de remboursement entre le régime général et local n'a pas d'impact sur le coût moyen de l'acte. Une quatrième région, Rhône-Alpes, s'éloigne également des autres régions, en raison de ses coûts moyens élevés. Il est à noter que d'après une récente étude de la DREES, l'Ile de France, le Rhône-Alpes et l'Alsace sont les trois régions où les médecins pratiquent le plus de dépassements d'honoraires (cf. [12]).

Synthèse

Cette ACP semble montrer l'existence d'un effet sur la consommation en frais de santé du lieu d'habitation de l'assuré. Cette variable doit être ainsi considérée dans la tarification, bien que le régime pourrait déjà expliquer le comportement des assurés des régions Alsace et Lorraine. L'ACP nous permet également de voir que l'effet de cette variable dépend des postes de garantie. Par exemple, la région Ile de France prend des valeurs plus fortes que la moyenne pour le coût moyen en honoraires et radiologie, alors que la région Rhône-Alpes prend des valeurs plus fortes que la moyenne pour l'optique et l'hospitalisation.

Des ressemblances entre les individus ont pu être observées sur le nuage des individus. Le cercle de corrélation a permis de recenser les corrélations linéaires entre les variables, qui ne sont pas toutes évidentes à interpréter. Pour certains postes de garanties comme l'optique ou les auxiliaires médicaux, des dépendances entre la fréquence et le coût moyen ont été observées (variables anti-corrélées). Même si la fréquence en optique n'est pas bien représentée sur le cercle de corrélation, nous savons qu'en réalité le comportement de consommation en monture ou verres dépend fortement du remboursement de la complémentaire santé. Or, étant donné que le calcul de la prime pure repose sur l'indépendance de ces deux grandeurs, ceci peut biaiser le résultat de la tarification étudié dans la partie suivante.

3.5 La classification ascendante hiérarchique (CAH) sur le lieu d'habitation

La variable « région » doit ainsi être prise en compte dans la tarification, mais cette segmentation en complément de l'âge, du sexe, du régime, du niveau de garantie et d'autres variables tarifaires pourrait réduire le nombre de données par catégories de risque et ainsi générer des estimations non fiables. Par ailleurs, l'utilisation de cette variable composée

d'un grand nombre de modalités pourrait poser des problèmes de lisibilité dans le modèle linéaire généralisé. C'est pourquoi, il convient de regrouper les régions ayant un comportement de consommation proche. Ce regroupement pourrait être plus précis en utilisant les données sur les départements. Ainsi, une méthode de classification des départements a été retenue : la classification ascendante hiérarchique (cf. [6]). Il existe plusieurs méthodes de classification, l'avantage de la CAH réside dans la représentation sous forme d'arbre permettant de choisir facilement le nombre de classes optimal. En effet, il n'est pas nécessaire de fixer le nombre de classes au préalable contrairement à d'autres méthodes de classification.

3.5.1 Le principe de la CAH

La classification permet d'établir des regroupements d'individus en considérant les proximités entre individus sur plusieurs dimensions. La CAH est une méthode de classification se basant sur l'agrégation des individus entre eux de proche en proche, ensuite des classes d'individus entre elles, pour parvenir à une classe recensant l'ensemble des individus. Plusieurs méthodes existent pour agréger les individus entre eux telles que l'indice du lien minimum, l'indice du lien maximum, la distance moyenne, la distance entre centres de gravité et l'indice de Ward. Nous utiliserons dans notre cas l'indice de Ward qui est une méthode couramment utilisée. L'objectif de cette méthode est de regrouper les classes où la perte de variance intra-classe est la plus faible. Cela permet d'obtenir les classes les plus homogènes. Par équivalence, elle consiste à maximiser la variance inter-classe, pour obtenir des classes bien séparées. La variation de la variance inter-classe entre les classes C_1 et C_2 est donnée par la formule suivante :

$$\delta_{WARD}\left(C_1, C_2\right) = \frac{m_1 m_2}{m_1 + m_2} \times d^2\left(g_1, g_2\right)$$

où :
- m_1 et m_2 sont les poids respectifs des classes C_1 et C_2 ;
- $d(g_1, g_2)$ représente la distance entre les centres de gravité g_1 et g_2 des classes respectives C_1 et C_2.

3.5.2 Les résultats

Dendrogramme

La CAH a été réalisée [15] sur le même tableau de données et avec les mêmes variables que précédemment, à l'exception des régions qui sont remplacées par les départements. L'arbre hiérarchique suivant illustre le processus d'agrégation des classes :

15. Macro SAS téléchargée sur le site de l'INSEE.

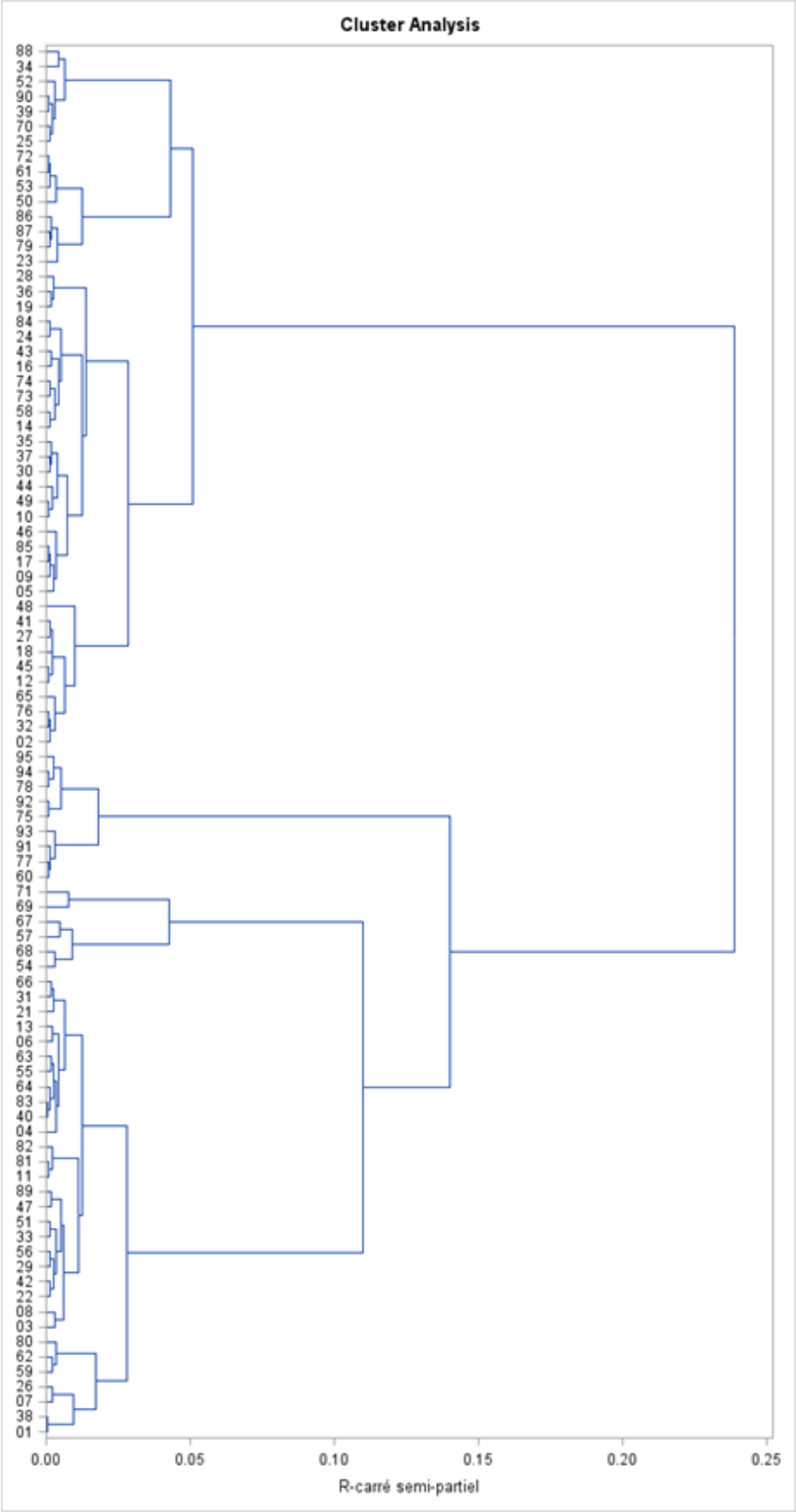


FIGURE 12 – Dendrogramme des départements

Le choix du nombre de classes optimal peut être déterminé grâce au graphique du R^2 partiel (cf. Annexe B) qui représente la décroissance de la variance inter-groupes en fonction du nombre de classes. La méthode consiste à lire le graphique de droite à gauche et de sélectionner le nombre de classe se situant avant un saut. Dans notre cas, nous observons un premier saut entre six et sept classes. Il convient de retenir sept classes puisque cela permet de classer dans une classe à part les départements 54, 57, 67 et 68, départements de la Lorraine et de l'Alsace. En effet, les habitants des départements 57, 67 et 68 sont affiliés au régime local et bénéficient ainsi d'un remboursement plus élevé. La CAH regroupe également le département 54 dans cette classe, puisque le portefeuille étudié contient une grande majorité d'assurés affiliée au régime local dans ce département. Cela est lié notamment au fait que les salariés exerçant une activité en Alsace-Moselle bénéficient également du régime local indépendamment de leur lieu de résidence.

Composition des classes

La sortie SAS ci-dessous fournit la composition des classes :

```
VARIABLE DE CLASSE :


 Classe numéro  1

01   03   04   06   07   08   11   13   21   22   26   29   31   33   38   40   42
47   51
55   56   59   62   63   64   66   80   81   82   83   89


 Classe numéro  2

25   34   39   52   70   88   90


 Classe numéro  3

02   05   09   10   12   14   16   17   18   19   24   27   28   30   32   35   36
37   41
43   44   45   46   48   49   58   65   73   74   76   84   85


 Classe numéro  4

23   50   53   61   72   79   86   87


 Classe numéro  5

60   75   77   78   91   92   93   94   95


 Classe numéro  6

54   57   67   68


 Classe numéro  7

69   71
```

FIGURE 13 – La composition des classes de départements

Nous pouvons également observer la moyenne des valeurs prises par les différentes variables en fonction de la classe créée. Voici les résultats pour les analyses et pour le dentaire :

Fréquence	Statistique	1	2	3	4	5	6	7	Ensemble
Analyses	Moyenne	0,83	0,96	0,77	0,69	0,87	1,06	1,09	0,83
Dentaire	Moyenne	0,60	0,59	0,52	0,45	0,61	0,81	0,63	0,57

TABLE 4 – Fréquence moyenne par classes de département

Coût moyen	Stat.	1	2	3	4	5	6	7	Ensemble
Analyses	Moy.	29,17	20,94	25,30	26,00	30,23	24,41	22,59	26,71
Dentaire	Moy.	62,40	53,44	57,83	53,78	72,76	62,78	64,67	60,48

TABLE 5 – Coût moyen par classes de département

Les fréquences de consommation les plus élevées concernent les assurés habitant dans les départements de la classe 6 (apparentée au régime local). Cela nous parait cohérent, puisque la Sécurité sociale rembourse à des taux plus élevés dans ces départements et par conséquent les assurés sont incités à consommer plus que ceux du régime général.
En ce qui concerne les coûts moyens, nous nous attendons à avoir des coûts élevés pour la catégorie 5 regroupant les départements d'Ile de France et des coûts plus faibles pour la catégorie 6 regroupant les départements du régime local. À la lecture des tableaux, les coûts moyens de la classe 5 sont supérieurs aux coûts des autres classes. En revanche, la classe 6 ne contient pas le coût le plus faible. Cette différence peut être liée à la répartition des assurés au sein des contrats de différents niveaux de garanties qui a été étudiée précédemment. En effet, en moyenne, les assurés du régime local ont des contrats de niveaux de garantie plus élevés que les assurés du régime général, ce qui entraîne des remboursements plus élevés par la complémentaire santé et explique le chiffre obtenu pour la classe n°6.
Les différentes études descriptives réalisées sur notre portefeuille ont permis d'obtenir une meilleure connaissance des différentes variables tarifaires utilisées en assurance santé et de leurs impacts sur la fréquence et le coût moyen.

Troisième partie

La tarification

L'objet de cette partie est de proposer une méthode alternative à la méthode de tarification directe fréquence-coût actuelle [a] : le modèle linéaire généralisé. Avant d'appliquer le modèle à la modélisation de la fréquence et du coût moyen, il est nécessaire d'en présenter les aspects théoriques. Pour finir, les résultats obtenus pourront faire l'objet d'une comparaison avec les résultats de la méthode directe.

Pour des raisons de lisibilité, nous présenterons uniquement deux sous-catégories d'acte : les analyses et actes de laboratoire et les prothèses dentaires. Ces deux exemples permettront d'illustrer l'adéquation des modèles dans le cas où nous observons une forte fréquence de consommation et dans le cas d'une faible fréquence de consommation.

a. Cette méthode est présentée dans le chapitre 5.

Chapitre 1

La théorie des modèles linéaires généralisés (GLM)

Les modèles linéaires gaussiens (cf. [7]) ont longtemps été utilisés pour modéliser la fréquence et le coût moyen. Cependant, ils ne sont pas adaptés à la réalité, puisque la variable à modéliser, c'est-à-dire la variable réponse, n'est pas nécessairement gaussienne. Ainsi, le modèle linéaire généralisé a été créé afin d'étendre le modèle linéaire aux variables non gaussiennes et plus précisément aux variables dont la loi fait partie de la famille exponentielle.

1.1 La présentation générale

L'objectif du modèle linéaire généralisé [16] est de modéliser la relation existante entre une variable réponse et une ou plusieurs variables explicatives. Avant d'exposer le modèle linéaire généralisé, il est primordial de comprendre le modèle linéaire gaussien.

1.1.1 Le modèle linéaire gaussien

Nous considérons n observations indépendantes $y_1, y_2, \ldots, y_n$ correspondant à des réalisations de la variable réponse Y_i. L'équation s'écrit sous la forme suivante :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot X_{ij} + \epsilon_i \quad i = 1, \ldots, n$$

où :

- $X_{i1}, \ldots, X_{ip}$: variables explicatives associées à l'individu i ;
- $\beta_0, \beta_1, \ldots, \beta_p$: paramètres inconnus à estimer ;
- ϵ_i : terme d'erreur provenant de la différence entre l'observation et l'estimation de la variable réponse. ϵ_i est supposé de moyenne nulle et de variance constante.

Le modèle est qualifié de «gaussien» dès lors que nous supposons que les erreurs sont distribuées selon une loi normale d'espérance nulle et de variance constante inconnue σ^2.

16. Modèle introduit initialement par John Nelder et Robert Wedderburn en 1972, et présenté d'une façon plus détaillée et complète par Mc Cullagh et John Nelder en 1989.

Dans ce cas, il s'agit de modèle linéaire gaussien. L'hypothèse d'espérance nulle permet d'écrire la relation suivante :

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

1.1.2 Le modèle linéaire généralisé

Le modèle linéaire généralisé se distingue du modèle linéaire gaussien par les trois composantes suivantes :

- **la composante aléatoire :**
  Nous supposons que les observations y_i sont indépendantes et associées à une loi de probabilité issue d'une structure exponentielle. Cette notion de structure exponentielle sera détaillée dans la section 1.2.

- **la composante systématique :**
  La composante systématique η_i, nommée prédicteur linéaire, correspond à une combinaison linéaire des variables explicatives. Soit x_{ij} les observations de la variable explicative X_{ij}, nous avons
  $$\eta_i = x_i^t \beta$$
  où,
  $$x_i^t = \begin{pmatrix} 1 & x_{i1} & \dots & x_{ip} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

- **la fonction de lien :**
  La relation entre la composante aléatoire et le prédicteur linéaire est exprimée par la troisième composante appelée fonction de lien g, strictement monotone et différentiable. Notons $\mu_i = E(Y_i)$, alors
  $$g(\mu_i) = \eta_i \ ou \ \mu_i = g^{-1}(\eta_i) = g^{-1}\left(x_i^t \beta\right)$$
  Ainsi, l'espérance de Y correspond à une transformation du prédicteur linéaire.

  Contrairement aux modèles linéaires simples et multiples, il s'agit ici de modéliser une transformation de l'espérance de la variable réponse.

  Le tableau ci-dessous nous renseigne sur les fonctions de lien classiques :

Identité	$g(x) = x$
Log	$g(x) = log(x)$
Logit	$g(x) = log\left(\frac{x}{1-x}\right)$
Inverse	$g(x) = \frac{1}{x}$
Probit	$g(x) = \varphi(x)$ [17]

TABLE 6 – Les fonctions de lien classiques

1.2 Distribution d'une famille exponentielle

La famille exponentielle contient toutes les lois disposant d'une fonction de densité pouvant s'écrire sous la forme suivante :

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

L'espérance et la variance sont données par les formules suivantes :

$$E(Y) = b'(\theta) = \frac{db(\theta)}{d\theta}$$

$$Var(Y) = a(\phi)\, b''(\theta)$$

où le paramètre θ est appelé le paramètre de position et ϕ le paramètre de dispersion ou d'échelle, les fonctions $a(.)$, $b(.)$ et $c(.,.)$ sont des fonctions réelles.

Exemple

Afin d'illustrer l'utilisation de cette formule, nous détaillerons les étapes permettant de passer de l'expression générale de la fonction de densité d'une loi normale de paramètres μ et σ^2 à la forme de l'expression d'une famille exponentielle. Dans ce cas, nous supposons que la variable aléatoire réponse Y suit une loi normale d'espérance μ et de variance σ^2. Sa fonction de densité est donnée par :

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

La première étape consiste à intégrer tous les termes dans la fonction exponentielle :

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

17. φ : fonction de densité d'une variable aléatoire qui suit une loi $N(0, 1)$

Ensuite, l'objectif est de distinguer les différents paramètres de la formule caractérisant la famille exponentielle :

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

$$f\left(y;\mu,\sigma^2\right) = exp\{\frac{y\mu-\frac{\mu^2}{2}}{\sigma^2} - \frac{\frac{y^2}{\sigma^2}+log\left(2\pi\sigma^2\right)}{2}\}$$

La loi normale de paramètres μ et σ appartient ainsi à la famille exponentielle avec les paramètres et les fonctions suivants :

$$\begin{aligned}
\theta &= \mu; \\
\phi &= \sigma^2; \\
a\left(\phi\right) &= \sigma^2; \\
b\left(\theta\right) &= \frac{\theta^2}{2}; \\
c\left(y,\phi\right) &= -\frac{1}{2}\left(\frac{y^2}{\phi} + log\left(2\pi\phi^2\right)\right).
\end{aligned}$$

Ainsi, $E\left(Y\right) = \theta = \mu$ et $Var\left(Y\right) = 1 \times \sigma^2 = \sigma^2$, confirmant le résultat obtenu.

Les composantes de la famille exponentielle d'autres distributions et de la loi normale sont données dans les tableaux ci-dessous :

Distribution	Notation	Densité	θ	ϕ
Binomiale	$B(n,p)$	$\binom{n}{ny}p^{ny}\left((1-p)^{n(1-y)}\right)$	$log\left(\frac{p}{1-p}\right)$	$\frac{1}{n}$
Binomiale négative	$BN(r,p)$	$\binom{y+r-1}{y}(1-p)^y p^r$	$log(1-p)$	1
Poisson	$P(\mu)$	$\frac{1}{y!}e^{-\mu}\mu^y$	$log(\mu)$	1
Gamma	$GA(\mu,\nu)$	$\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}exp\left(-\frac{\nu}{\mu}y\right)$	$-\frac{1}{\mu}$	ν^{-1}
Normale	$N(\mu,\sigma^2)$	$f(y)=\frac{1}{\sigma\sqrt{2\pi}}exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	μ	σ^2

Distribution	$a(\phi)$	$b(\theta)$	$c(y,\phi)$
Binomiale	$\frac{1}{n}$	$\frac{\theta^2}{2}$	$log\binom{n}{y}$
Binomiale négative	1	$-rlog\left(1-e^{\theta}\right)$	$log\binom{y+r-1}{y}$
Poisson	1	e^{θ}	$-log(y!)$
Gamma	ν^{-1}	$-log(-\theta)$	$\nu log(\nu y)-log(y)-log(\Gamma(\nu))$
Normale	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+log\left(2\pi\phi^2\right)\right)$

TABLE 7 – Les composantes de la famille exponentielle

1.3 L'estimation des paramètres

Un des objectifs du GLM est d'estimer les coefficients de régression $\beta_0, \beta_1, \ldots, \beta_p$. La méthode d'estimation couramment utilisée dans le cadre du GLM est le maximum de vraisemblance que nous détaillerons dans cette section.

Log-vraisemblance

Tout d'abord, nous considérons la variable réponse Y_i indépendante et issue d'une famille exponentielle. L'expression de la vraisemblance s'écrit :

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

Notons $L = L(y_1,\ldots,y_n;\theta_i,\phi)$, nous obtenons l'expression de la log-vraisemblance suivante :

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

Il faut ainsi maximiser cette dernière expression, ce qui consiste à calculer tout d'abord la dérivée en fonction des paramètres β_j :

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

Exemple

Considérons une variable aléatoire réponse Y qui suit une loi de Poisson de paramètre μ :

$$L = \prod_{i=1}^{n} e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}$$

$$log\left(L\right) = \sum_{i=1}^{n} -\mu_i + y_i log\left(\mu_i\right) - log\left(y_i!\right)$$

$$\frac{\partial}{\partial \beta_j} log\left(L\right) = \sum_{i=1}^{n} \frac{\partial}{\partial \mu_i} \{-\mu_i + y_i log\left(\mu_i\right) - log\left(y_i!\right)\} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Dans le cas où nous utilisons la fonction de lien log, nous avons $g(\mu_i) = \eta_i = log(\mu_i)$. Sachant que, $\eta_i = x_i^t$, nous obtenons :

$$\frac{\partial}{\partial \beta_j} log\left(L\right) = \sum_{i=1}^{n} \left(y_i - \mu_i\right) x_{ij}.$$

Ainsi les équations de vraisemblance sont :

$$\sum_{i=1}^{n} \left(y_i - \mu_i\right) x_{ij} = 0 \ \forall \ j = 1, \ldots, p$$

La résolution de ces équations requiert une méthode itérative telle que la méthode de Newton-Raphson que nous ne développerons pas (cf. [1]).

1.4 Synthèse

La régression par modèles linéaires généralisés sera réalisée par l'utilisation du logiciel SAS 9.3, notamment avec les procédures « Genmod » et « Countreg ».

Pour ce faire, il faut auparavant choisir les éléments ci-dessous :

- la distribution de la variable à expliquer, sachant que celle-ci doit faire partie de la famille exponentielle ;
- la fonction de lien ;
- le poids de la variable à expliquer ;
- les variables explicatives pouvant avoir une influence sur la valeur de la variable réponse.

Le tableau suivant indique la fonction de lien associée à quelques lois de probabilité usuelles. Si la fonction de lien est la même que celle qui lit le paramètre μ au paramètre θ alors il s'agit de fonction de lien canonique. Par exemple, pour la loi de Poisson $\theta = log(\mu)$ (cf. section 1.2 de cette partie), par conséquent la fonction de lien canonique est la fonction « log ».

Binomiale	Logit
Poisson	Log
Binomiale négative	Logit
Normale	Identité
Gamma	Inverse

TABLE 8 – Les fonctions de lien associées aux lois de probabilité usuelles

Pour des raisons pratiques, il est préférable d'utiliser la fonction de lien log pour les lois binomiale négative et gamma. En effet, l'utilisation de cette fonction de lien permet :

- d'obtenir des coefficients positifs des paramètres estimés ;
- d'avoir un modèle multiplicatif qui permet de connaître facilement l'effet de chaque paramètre sur la variable réponse.

Chapitre 2

Les critères de choix de modèle

Après avoir présenté le cadre théorique des GLM, nous proposons dans ce chapitre d'étudier les différents critères utilisés dans la sélection et la comparaison de modèles. Ce chapitre présente également la méthode de sélection de variables retenue.

2.1 La validation et la comparaison de modèles

2.1.1 La déviance

Afin de vérifier l'ajustement du modèle aux données utilisées, nous pouvons calculer la déviance du modèle (cf. [1]). Elle consiste à comparer le modèle étudié à un modèle saturé, *i.e.* un modèle avec une distribution identique, une même fonction de lien et où les termes μ_i sont remplacés par les termes y_i, observations de la variable réponse. Autrement dit, elle permet de comparer un modèle où la variable réponse est supposée suivre une certaine loi à un modèle avec les valeurs observées de la variable réponse.

Définition

La déviance D est définie comme deux fois la différence entre la log-vraisemblance du modèle saturé et la log-vraisemblance du modèle étudié multipliée par le paramètre de dispersion :

$$D = 2\phi\{logL\,(y, y, \phi) - logL\,(y, \mu, \phi)\}.$$

Elle peut également être définie comme la déviance standardisée multipliée par le paramètre de dispersion :

$$D = \phi D^* \text{ où } D^* = 2\{logL\,(y, y, \phi) - logL\,(y, \mu, \phi)\}.$$

Le modèle avec la déviance la plus faible sera préféré aux autres modèles puisque ce critère indique un écart plus faible entre les log-vraisemblances, et ainsi une distance plus faible entre les valeurs modélisées et les valeurs observées.

Test

La déviance standardisée D^* suit asymptotiquement une loi du χ^2_{n-p}, où n représente le nombre de paramètres du modèle saturé (équivalent au nombre d'observations de la variable réponse) et p celui du modèle étudié. Par conséquent, nous pouvons construire un test permettant de rejeter ou d'accepter le modèle étudié. Ce test sera détaillé et illustré dans le cas de la sélection des variables (cf. section 2.2.2).

2.1.2 Les critères AIC et BIC

Outre la déviance, les critères AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion) (cf. [10]) permettent également de comparer les modèles entre eux. L'utilisation de ces critères semble plus appropriée dans la comparaison de modèles construits avec des distributions de variables réponses différentes, puisque la déviance permet uniquement de comparer les modèles emboités.

L'AIC est défini par la formule suivante :

$$AIC = -2log(L) + 2k$$

où $log(L)$ constitue la log-vraisemblance maximisée et k le nombre de paramètres. Ainsi, le critère d'Akaike permet d'effectuer un compromis entre la réduction du biais (avec l'augmentation du nombre de paramètres) et le besoin de modéliser les données avec le plus petit nombre de paramètres.

Le critère BIC (également nommé SBC) doit être privilégié lorsqu'il s'agit de modèles disposant d'un grand nombre d'observations. En effet, dans la littérature (cf. [13]) il est précisé que le critère AIC a tendance à choisir les modèles avec de nombreuses variables explicatives dans le cas de grands échantillons. Afin d'écarter ce problème, le nombre de paramètres dans la formule du BIC est multiplié par le logarithme du nombre d'observations $log(n)$ et permet ainsi d'appliquer une pénalité plus sévère afin de privilégier l'utilisation de modèles avec moins de variables explicatives :

$$BIC = -2log(L) + klog(n).$$

2.1.3 Les résidus

L'analyse des différents critères d'ajustement présentés ci-dessus n'est pas suffisante pour valider un modèle linéaire généralisé. Une analyse plus précise à l'aide des résidus permet d'analyser individuellement les écarts entre les valeurs observées et les valeurs prédites $\widehat{y}_i$ par le modèle. Les résidus de déviance ou de Pearson sont couramment utilisés dans le cadre de GLM.

Les résidus de déviance sont définis à partir d'un terme d_i représentant la contribution de l'$i^{\text{ème}}$ observation y_i à la déviance D,

$$r_{D_i} = signe\,(y_i - \hat{y}_i)\,\sqrt{d_i} \;et\; d_i = 2\{logL(y_i, y_i, \phi) - logL(y_i, \hat{y}_i, \phi)\}.$$

Les résidus de Pearson sont définis comme le rapport entre la distance entre la valeur observée et la valeur prédite $\hat{y}_i$ et la racine carrée de la variance $\widehat{V(\hat{y}_i)}$ estimée du modèle :

$$r_{P_i} = \frac{(y_i - \hat{y}_i)}{\sqrt{\widehat{V(\hat{y}_i)}}}.$$

Il est préférable de les normaliser (résidus standardisés) :

$$r^*_{P_i} = \frac{(y_i - \hat{y}_i)}{\sqrt{\widehat{V(\hat{y}_i)} h_{ii}}}$$

$$r^*_{D_i} = \frac{signe\,(y_i - \hat{y}_i)\sqrt{d_i}}{\sqrt{h_{ii}}}$$

où h_{ii} correspond aux termes sur la diagonal de la matrice H défini par
$H = W^{\frac{1}{2}} X (X^t W X)^{-1} X^t W^{\frac{1}{2}}$, W correspondant à la matrice diagonale de « pondération » (cf. [11]).
La matrice H, telle que $\hat{y} = Hy$ permet d'évaluer la variation des valeurs prédites en fonction des autres observations.

2.2 La sélection des variables

2.2.1 Présentation des méthodes

Une sélection initiale des variables tarifaires a été réalisée par une analyse descriptive dans la partie 2. Cependant, dans le cadre du GLM, nous souhaitons sélectionner uniquement celles qui ont une réelle influence sur la variable réponse. Nous retrouvons plusieurs types d'algorithme de sélection de variables dans la littérature, dont les plus couramment utilisés sont les méthodes backward, forward et stepwise.

La méthode backward (descendante) consiste à intégrer toutes les variables explicatives dans le modèle et à éliminer à chaque étape la variable la moins significative, contrairement à la méthode forward (ascendante), qui débute l'algorithme par la variable la plus significative et intègre à chaque étape la variable qui contribue le plus au modèle.

Dans le cadre de ce mémoire, nous nous intéresserons à la méthode stepwise qui est un mélange des méthodes forward et backward. Elle est semblable à la méthode forward et diffère par la possibilité d'éliminer après chaque insertion de variable, une variable qui ne serait plus significative.

L'insertion et l'élimination des variables se basent sur l'analyse de la déviance. Il s'agit de comparer la valeur de la déviance standardisée du modèle avant l'ajout et après l'ajout d'une ou plusieurs variables. Cela revient à analyser la différence de déviance standardisé qui suit asymptotiquement une statistique du χ^2 à $q - p$ degrés de liberté, avec

– p le nombre de variables dans le modèle avant l'ajout d'une ou plusieurs variables ;
– q le nombre de variables dans le modèle suite à l'ajout d'une ou plusieurs variables supplémentaires.

Considérons deux modèles simplifiés sans le terme d'erreur :

$$M_1 : Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$$M_2 : Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$$

$$avec\ q > p$$

et les déviances standardisées respectives D_{M_1} et D_{M_2}.

La différence de déviances standardisées des deux modèles est :

$$\Delta D = D_{M_1} - D_{M_2} = 2\{logL_{M_2} - logL_{M_1}\}$$

où L_{M_2} et L_{M_1} correspondent respectivement à la vraisemblance du deuxième et du premier modèle.
Sous l'hypothèse $H0 : \beta_{p+1} = \cdots = \beta_q = 0$, ΔD suit asymptotiquement une loi χ^2_{q-p}. Ainsi, nous pouvons comparer la statistique de test ΔD au quantile $1-\alpha$ d'une loi χ^2_{q-p}. L'hypothèse $H0$ est rejetée si la quantité $P(\chi^2_{q-p} > \Delta D)$ appelée p-value, est supérieure à α. Or, ce test effectué par rapport à l'ajout de plusieurs variables permet uniquement d'indiquer que le modèle M_1 peut être complété par d'autres variables, et pas nécessairement de l'ensemble des variables considérées dans le modèle M_2. Ainsi, il est préférable d'effectuer ce test en considérant l'intégration d'une seule variable.

2.2.2 L'application

Nous détaillons les étapes de l'algorithme d'un exemple ci-dessous, avec un seuil de significativité fixé à 5% :

– Étape 0 : choix de la première variable à intégrer dans le modèle

$$M_0 : Y = \beta_0,\ avec\ la\ déviance\ standardisée\ D_0$$

$$M_1 : Y = \beta_0 + \beta_1 x_1,\ avec\ la\ déviance\ standardisée\ D_1$$

Le tableau ci-dessous fournit pour l'ensemble des variables la valeur de la déviance,

la statistique de test dans la colonne « Chi-Square » qui correspond à ΔD, et la probabilité $P(\chi^2_{2-1} > \Delta D)$ dans la dernière colonne.

Variable x_1	Déviance D_1	Chi-Square	Pr > ChiSq
Age	780591	194408	<.0001
Sexe	964860	9688	<.0001
Régime	950894	24104	<.0001
Département	929929	37006	<.0001
Garantie	848082	126916	<.0001
Bénéficiaire	855019	119979	<.0001
Nombre d'enfants	966825	563	<.0001

TABLE 9 – Exemple : Choix de la première variable à intégrer

La variable qui apporte la plus grande information au modèle est l'âge, puisque la statistique de test est de 194408, avec une p-value inférieur au seuil α.

– Étape 1 : introduction de la deuxième variable

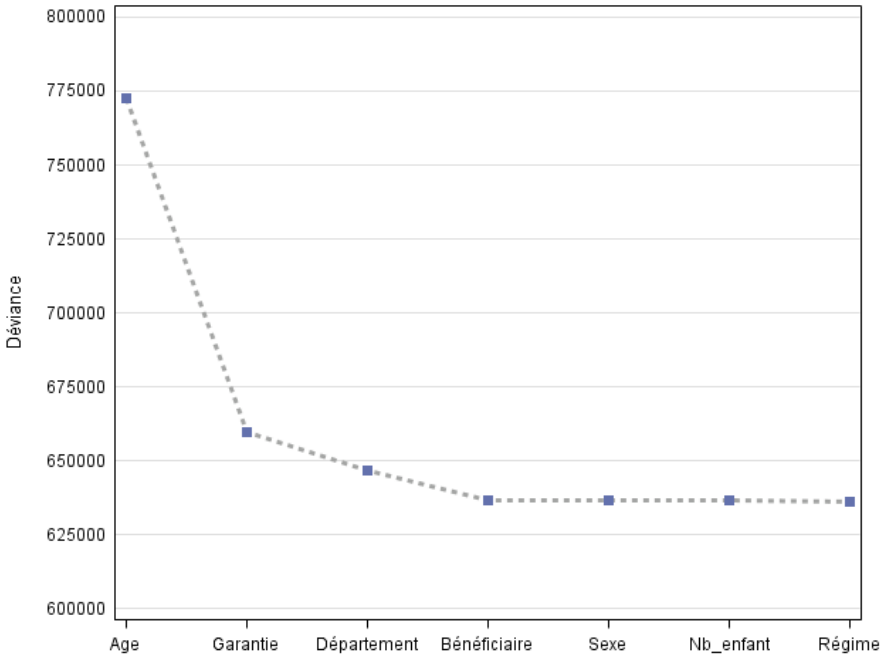
$$M_1 \ : \ Y = \beta_0 + \beta_1 (\hat{a}ge) \, , avec \ la \ déviance \ D_1$$
$$M_2 \ : \ Y = \beta_0 + \beta_1 (\hat{a}ge) + \beta_2 x_2, \ avec \ la \ déviance \ D_2$$

Variable x_2	Déviance D_2	Chi-Square	Pr > ChiSq
Sexe	774089	6177	<.0001
Régime	760584	20006	<.0001
Département	747217	26658	<.0001
Garantie	665380	115211	<.0001
Bénéficiaire	776189	4401	<.0001
Nombre d'enfants	771885	1166	<.0001

TABLE 10 – Exemple : Choix de la seconde variable à intégrer

Dans cette étape, il convient de sélectionner la variable « garantie » puisqu'elle contribue le plus au modèle. Ensuite, conformément à la méthode stepwise, il y a lieu de vérifier que les variables sélectionnées dans l'étape précédente ont une p-value supérieur à α. Dans notre cas, il suffit de vérifier la p-value de la variable âge. Si elle est supérieure à 5% alors elle sera retirée du modèle, sinon il s'agit de continuer la sélection des variables en appliquant la même méthodologie que dans cette étape.

L'introduction des variables peut être arrêtée lorsqu'aucune des p-valeur des variables non sélectionnées n'est inférieure à α, ou lorsque l'introduction d'une nouvelle variable apporte peu d'information au modèle (très faible baisse de la déviance). Nous représentons ci-dessous la différence des déviances de toutes les variables introduites une à une avec la méthode stepwise :



var.png

FIGURE 14 – Exemple : la contribution des variables au modèle

A partir de la variable « bénéficiaire », nous considérons que la contribution des variables à l'explication de la variable réponse est négligeable. Ainsi, nous décidons de retenir les quatre premières variables représentées.

Chapitre 3

La prise en compte de la dispersion

Dans ce chapitre, nous proposons d'étudier les différents modèles permettant de prendre en compte la sur-dispersion des données.

3.1 La présentation du phénomène

La modélisation d'une variable discrète et positive est souvent réalisée à partir d'une loi de Poisson. Or, en théorie celle-ci est construite sur une hypothèse forte qui est l'équidispersion des données, *i.e.* l'espérance est égale à la variance. Dans cette partie, il convient d'analyser ce phénomène et de donner les différentes solutions possibles pour palier à un éventuel problème de sur-dispersion ou de sous-dispersion de nos données.

La variance, dans le cas d'une sur-dispersion, est définie ci-dessous :

$$Var(Y) = \phi E(Y),\ \phi \geqslant 1,\ \textit{paramètre de dispersion}$$

La sous-dispersion est plus rare et représente le cas où le paramètre de dispersion est inférieur à 1.

Plusieurs causes peuvent être à l'origine de cette sur-dispersion dont la présence importante de zéro pour la variable réponse et l'hétérogénéité du portefeuille étudié : l'unique paramètre de la loi de Poisson ne serait ainsi pas suffisant à expliquer les données. Il est nécessaire d'utiliser un autre modèle puisque la présence de sur-dispersion peut affecter les estimations de la statistique du Khi-deux, qui intervient dans le choix de sélection de variables explicatives.

Nous retrouvons dans la littérature plusieurs alternatives permettant de prendre en compte ce phénomène de sur-dispersion, dont les principales sont :

- le modèle quasi-Poisson ;
- le modèle binomial négatif ;
- les modèles modifiés en zéro (« zero inflated »).

3.2 Le modèle quasi-Poisson

Afin de rester dans le cadre d'un modèle de Poisson qui constitue un modèle simple, il convient de corriger la sur-dispersion par un coefficient. Cela consiste à multiplier les écarts-types des paramètres β estimés par un estimateur du coefficient de dispersion :

$$E(Y_i) = \mu_i \qquad Var(Y_i) = \phi\mu_i.$$

Le coefficient de dispersion ϕ peut être estimé par le coefficient de Pearson généralisé :

$$\hat{\phi} = \frac{\chi^2}{n-p} \qquad \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(\mu_i)}$$

avec :

- n : le nombre d'observations ;
- p : le nombre de variables ;
- $Var(.)$: la fonction de variance de la distribution.

Nous constatons que les estimations des termes μ_i sont identique aux estimations du modèle de Poisson. Par conséquent, ce modèle permet d'obtenir les mêmes estimations des paramètres β_j que le modèle de Poisson. La seule différence réside dans l'expression de la variance.

3.3 Le modèle binomial négatif

La loi binomiale négative est la loi habituellement utilisée pour prendre en compte la sur-dispersion. Nous précisons que les notations utilisées dans cette partie font référence aux notations fournies par la documentation du logiciel SAS (cf. [23]).

Afin de prendre en compte la sur-dispersion, nous pouvons introduire une hétérogénéité dans l'espérance conditionnelle de la loi de Poisson à travers un terme ϵ_i :

$$E(Y_i|X_i, \tau_i) = exp\{x_i^t\beta + \epsilon_i\} = \mu_i\tau_i.$$

Ainsi, la fonction de densité correspondante est définie par :

$$f(y_i|x_i, \tau_i) = exp\{-\mu_i\tau_i\}\frac{(\mu_i\tau_i)^{y_i}}{y_i!}.$$

Ce terme τ_i est supposé suivre une loi gamma d'espérance 1 et de variance $1/\theta$. Cela nous permet de calculer la densité de la variable Y_i conditionnellement à X_i :

$$f(y_i|x_i) = \frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}(\frac{\theta}{\theta+\mu_i})^{\theta}(\frac{\mu_i}{\theta+\mu_i})^{y_i},\ y_i \in \mathbb{N}$$

où l'espérance et la variance conditionnelle sont définies par :

$$E(Y_i|X_i) = \mu_i \qquad Var(Y_i|X_i) = \mu_i(1+\alpha)$$

où $\alpha = {}^{1}/_{\theta}$ correspond au paramètre à estimer.

L'expression de l'espérance et de la variance de la loi binomiale négative prouve que l'effet de sur-dispersion peut être pris en compte à travers ce modèle puisque $\mu_i(1+\alpha) > \mu_i$. Nous remarquons ici que nous sommes dans le cas de la loi quasi-Poisson où la variance d'une distribution de Poisson est multipliée par un paramètre estimé. Cameron et Trivedi (1986) ont proposé des modèles binomiaux négatifs où l'expression de la variance serait de la forme $\mu_i + \alpha\mu_i^p$. En pratique, les modèles classiques utilisent $p = 1$ (cas présenté ci-dessus) ou $p = 2$. Par exemple, la loi binomiale négative par défaut de la procédure « Countreg » sur SAS, correspond au cas où p est égal à 2.

3.4 Les modèles modifiés en zéro

Lorsque les modèles quasi-Poisson et binomial négatif ne permettent pas de prendre en compte la sur-dispersion, cela peut être lié à un excès de zéros prise par la variable réponse. Cet excès de zéros existe lorsque le nombre de zéros observé par la variable réponse est supérieur au nombre de zéros estimé par un ajustement avec la loi de Poisson. En santé, ce phénomène est présent dans le cas d'actes à utilisation rare tels que la consommation en prothèses dentaires. Les modèles modifiés en zéro [18], ou zero-inflated ont été développés afin de prendre en compte cet excès de zéros. Dans la littérature, nous retrouvons deux modèles modifiés en zéro : les modèles zero inflated Poisson et zero inflated binomial negative.

Dans le cadre de la modélisation de la fréquence, l'idée de ces modèles est de distinguer

- la présence ou non de sinistres, c'est à dire le fait qu'un assuré consomme ou ne consomme pas l'acte considéré ;
- la quantité d'actes médicaux consommés.

Ainsi un modèle modifié en zéro consiste à considérer deux processus. L'utilisation d'une loi de Bernoulli permet de connaître la probabilité attribuée à chaque processus. Le premier processus permet de déterminer la probabilité π_i de non sinistralité fournie par une loi de Bernoulli et le deuxième processus génère les valeurs estimées par une loi définie (Poisson ou binomiale négative).

18. Modèles développés initialement par Lambert (1992) et Greene (1994)

Soit Y_i une variable de comptage positive, le modèle modifié en zéro est définie ci-dessous :

$$\begin{cases} \pi_i + (1-\pi_i)f(0) & si\ y_i = 0 \\ (1-\pi_i)f(y_i) & si\ y_i > 0 \end{cases}$$

où :

$$Y_i \sim 0\ avec\ la\ probabilité\ \pi_i$$
$$Y_i \sim f(y_i)\ avec\ la\ probabilité\ 1-\pi_i$$

où la fonction $f(.)$ suit une loi de Poisson ou une loi binomiale négative.

3.4.1 Le modèle Zero Inflated Poisson (ZIP)

Le modèle ZIP (cf. [25]) constitue un mélange entre une loi de Poisson de paramètre μ_i et une masse de Dirac en 0.

Le modèle s'écrit sous la forme suivante :

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)exp\{-\mu_i\} & si\ y_i = 0 \\ (1-\pi_i)exp\{-\mu_i\}\frac{\mu_i^{y_i}}{y_i!} & si\ y_i > 0 \end{cases}$$

L'espérance et la variance du modèle sont données par :

$$E(Y_i) = (1-\pi_i)\mu_i$$
$$Var(Y_i) = (1-\pi_i)(\mu_i + \pi_i\mu_i^2) = E(Y_i)(1+\pi_i\mu_i).$$

Nous retrouvons ici un modèle de Poisson lorsque la probabilité d'avoir la valeur 0 est nulle. Nous constatons également que la variance est strictement supérieure à la valeur de l'espérance et par conséquent le modèle permet de prendre en compte la sur-dispersion.

3.4.2 Le modèle Zero Inflated Negative Binomial (ZINB)

Similairement au modèle ZIP, le modèle ZINB (cf. [25]) correspond à un mélange entre une loi binomiale négative (de paramètres μ_i et ν) et une masse de Dirac en 0.

Nous considérons le modèle suivant :

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)(1+\nu\mu_i)^{-1/\nu} & si\ y_i = 0 \\ (1-\pi_i)\frac{\Gamma(y_i+1/\nu)}{\Gamma(y_i+1)\Gamma(1/\nu)}(\frac{1/\nu}{1/\nu+\mu_i})^{1/\nu}(\frac{\mu_i}{1/\nu+\mu_i})^{y_i} & si\ y_i > 0 \end{cases}$$

L'espérance et la variance du modèle sont données par :

$$E(Y_i) = (1 - \pi_i)\mu_i$$
$$Var(Y_i) = (1 - \pi_i)(\mu_i + (\pi_i + \nu)\mu_i^2) = E(Y_i)(1 + (\pi_i + \nu)\mu_i).$$

Chapitre 4

L'application à la modélisation de la fréquence

L'objet de cette sous-partie est de modéliser le nombre d'actes consommés (nombre de sinistres) par an par individu. Il s'agit ainsi d'une variable de comptage qui nécessite une modélisation par une loi discrète.

Dans la littérature, la modélisation d'un événement de comptage est souvent réalisée par une loi de Poisson supposant une équidispersion des données. Nous montrerons que cette loi de probabilité est inadaptée au portefeuille étudié et nous proposerons d'autres modèles.

4.1 L'analyse de la variable expliquée

Avant d'entamer la modélisation, il convient d'analyser la variable comptant le nombre de sinistres pour les deux actes sélectionnés.

Les analyses et actes de laboratoire

Le tableau ci-dessous contient les fréquences empiriques pour la variable comptant le nombre de sinistre (uniquement de 0 à 15) sur la période observée.

Nombre de sinistres Y	Fréquence empirique	Pourcentage (%)
0	573413	66,53
1	31231	3,62
2	59476	6,9
3	40948	4,75
4	28787	3,34
5	18249	2,11
6	19796	2,29
7	11327	1,31
8	11092	1,28
9	9139	1,06
10	7546	0,87
11	6044	0,7
12	5772	0,66
13	4370	0,5
14	4044	0,46
15	3546	0,41

TABLE 11 – Nombre de sinistres pour les analyses et actes de laboratoire

Moyenne	Ecart-type	Valeur max de Y
2,24	5,73	269

TABLE 12 – Statistiques descriptives du nombre de sinistres (les analyses et actes de laboratoire)

La fréquence de consommation moyenne en analyses et actes de laboratoire est de l'ordre de deux actes avec un écart-type d'environ 5,73. Il s'agit ici d'une variable décrivant des valeurs très dispersées, puisque l'écart-type est important et le nombre maximum d'acte par personne sur la période observée atteint le nombre de 269 actes.

Les prothèses dentaires

Nous analysons également la distribution du nombre de sinistres (uniquement de 0 à 15) dans le cas des prothèses dentaires prises en charge par la Sécurité sociale.

Nombre de sinistres Y	Fréquence empirique	Pourcentage (%)
0	845113	92,56
1	17554	1,92
2	20104	2,2
3	8781	0,96
4	6505	0,71
5	3524	0,38
6	3356	0,36
7	1849	0,2
8	1575	0,17
9	1110	0,12
10	821	0,08
11	554	0,06
12	561	0,06
13	326	0,03
14	275	0,03
15	195	0,02

TABLE 13 – Nombre de sinistres pour les prothèses dentaires

Moyenne	Ecart-type	Valeur max deY
0,25	1,23	29

TABLE 14 – Statistiques descriptives du nombre de sinistres (prothèses dentaires)

Dans le cas des prothèses dentaires, l'étendue du nombre d'acte consommé est plus faible. La non-sinistralité correspond à 92% des valeurs prises par cette variable. Par conséquent, la fréquence de consommation moyenne est faible et les données sont moins dispersées que dans le cas des analyses et actes de laboratoire, ce qui est cohérent.

4.2 Application de la loi de Poisson

Afin de modéliser des données discrètes telles que le nombre de sinistres, nous utilisons une loi de Poisson. La régression de Poisson sera effectuée en intégrant un terme offset prenant en compte le nombre d'années de présence de l'assuré sur la période 2011-2013.

$$log\left(\frac{E(Y/X}{années\ risques}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

ce qui est équivalent à :

$$E(Y|X) = exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + log(années\ risques)\}.$$

Dans SAS, la modélisation sera réalisée à l'aide de la procédure « GENMOD », procédure spécifique aux modèles linéaires généralisés, à laquelle sera appliquée une méthode de sélection pas à pas des variables.

Les analyses et actes de laboratoire

Les sorties SAS permettent d'analyser l'ajustement de la loi de Poisson aux données utilisées. Le tableau ci-dessous fournit la valeur du coefficient de dispersion ϕ qui est d'environ 8,98 pour les deux critères. Par conséquent, le modèle de Poisson ne s'ajuste pas à nos données, étant donné la présence d'une forte sur-dispersion.

Criterion	DF	Value	Value/DF
Pearson Chi-Square	8,50E+05	7604233.1850	8.9798
Scaled Pearson X2	8,50E+05	7604233.1850	8.9798

TABLE 15 – Critères d'ajustement à une loi de Poisson (analyses et actes de laboratoire)

Le mauvais ajustement de la loi de Poisson est confirmé par le graphique ci-dessous, permettant de comparer la probabilité moyenne d'avoir y_i sinistres avec la loi de Poisson et la probabilité d'observer la valeur y_i sur nos données. Il s'agit ici d'une méthode permettant d'analyser l'ajustement du modèle discret aux données étudiées. La probabilité moyenne prédite par le modèle peut être comparée à la probabilité observée, *i.e.* la fréquence empirique.

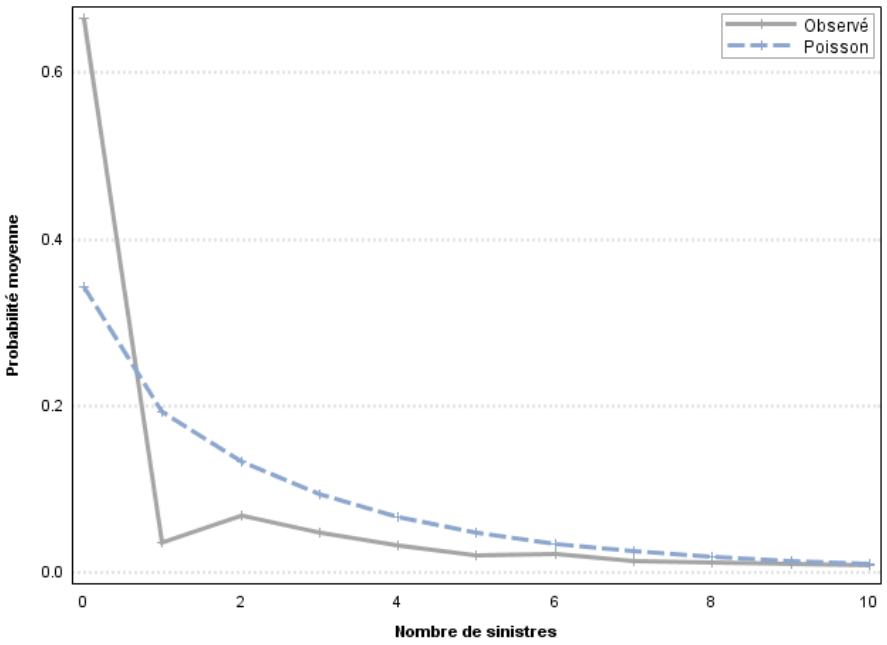


FIGURE 15 – Ajustement des données à une loi de Poisson (analyses et actes de laboratoire)

La loi de Poisson sous-estime fortement la probabilité d'avoir aucun sinistre, et surestime les autres valeurs.

Les prothèses dentaires

De façon similaire, nous analysons l'ajustement d'une loi de Poisson à la fréquence des sinistres pour les prothèses dentaires.

Criterion	DF	Value	Value/DF
Pearson Chi-Square	4,40E+05	2174081.7507	4.9255
Scaled Pearson X2	4,40E+05	2174081.7507	4.9255

TABLE 16 – Critères d'ajustement à une loi de Poisson (prothèses dentaires)

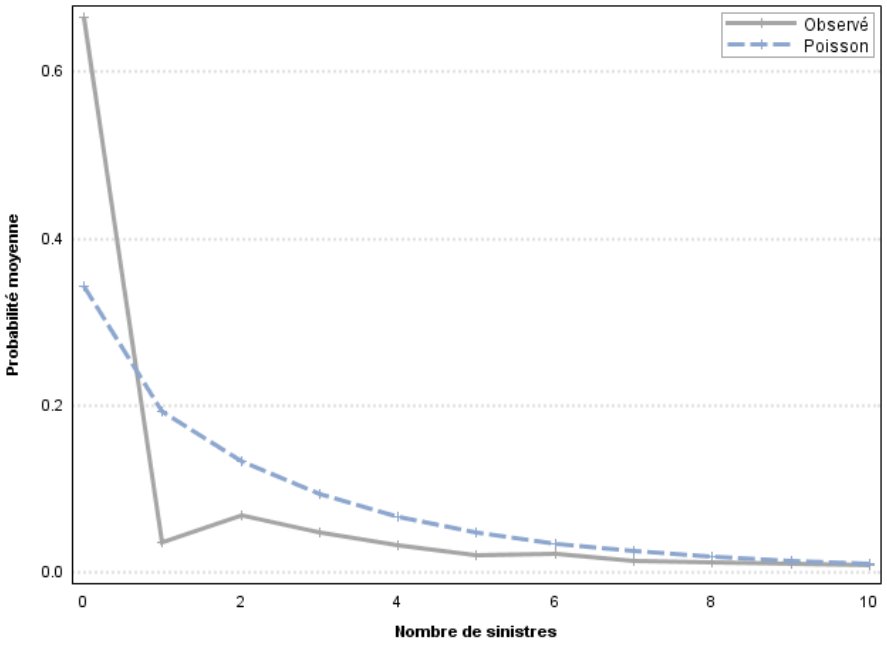


FIGURE 16 – Ajustement des données à une loi de Poisson (prothèses dentaires)

Nous constatons à nouveau une sur-dispersion des données utilisées. D'après le graphique, nous pouvons en déduire que la loi de Poisson s'ajuste mal pour un nombre de sinistres inférieur à trois. Au contraire, l'estimation du nombre de sinistres supérieur ou égal à quatre semble être plutôt bonne.

4.3 Les modèles alternatifs

La modélisation de la fréquence des sinistres ne peut pas être réalisée en appliquant une régression de Poisson puisque les données utilisées sont sur-dispersées. La prise en compte de cette sur-dispersion peut être effectuée en appliquant d'autres modèles présentés précédemment.

Dans le cadre des prothèses dentaires, nous développerons précisément l'utilisation d'un modèle ZINB compte tenu du nombre important de "zéro" observé, que nous pourrons comparer au modèle de Poisson, au modèle ZIP et binomial négatif. Comme il a été précisé précédemment, le modèle ZINB fait référence à deux lois : la loi de Bernoulli et la loi binomiale négative. Dans ce type de modèle, la sélection de variables est plus complexe puisque les variables ayant un impact sur le fait de consommer ou de ne pas consommer ne sont pas nécessairement les mêmes variables qui influent le nombre d'actes consommés. Ainsi, il convient de sélectionner deux groupes de variables pour chacune des distributions : les variables pour la modélisation avec une loi de Bernoulli (régression logistique) et les variables pour la modélisation avec une loi binomiale négative.

Une première sélection des variables avec la méthode stepwise a été effectuée à l'aide d'un modèle linéaire généralisé utilisant la loi binomiale négative. Cette régression est presque semblable au modèle de Poisson, puisque nous utilisons le même terme offset et la même fonction de lien :

$$E(Y|X) = exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + log(années\ risques)\}$$
$$Y|X \sim BN.$$

Les variables sélectionnées sont les suivantes (cf. Annexe C) :

Age
Garantie

La régression par un modèle ZINB a été réalisée avec la procédure « Countreg » du logiciel SAS. Nous avons tout d'abord défini les deux groupes de variables à intégrer dans les deux parties du modèle. Pour la partie modélisée par une loi binomiale négative, les variables sélectionnées ci-dessus dans le cadre d'un modèle binomiale négative simple ont été intégrées. Pour la partie « zéro », toutes les variables ont été dichotomisées et introduites une à une en fonction des critères de significativité des variables avec un seuil identique pour la partie binomiale négative de 5%. Les variables finales retenues sont les suivantes :

Partie BN de ZINB	Partie zéro de ZINB
Age	Age
Garantie	Garantie
	Sexe

Table 17 – Variables sélectionnées pour le modèle ZINB

Les variables expliquant les deux parties du modèle sont identiques à l'exception de la variable sexe qui n'influe pas le nombre de sinistres. Elle explique uniquement la partie modélisant la présence ou non de sinistralité. L'intégration des variables dans la partie zéro a requis un regroupement des modalités de la variable âge, qui n'étaient plus toutes significatives. Les classes de la variable âge retenues pour les deux parties du modèle ZINB sont définies ci-dessous :

Partie BN de ZINB	Partie zéro de ZINB
[0,20[	[0,20[
[20,25[	[20,25[
[25,30[	[25,30[
[30,35[	[30,35[
[35,40[	[35,40[
[40,45[	[40,45[
[45,50[	[45,50[
[50,75[	50 et plus
[75,80[	
80 et plus	

Table 18 – Classes d'âge sélectionnées pour le modèle ZINB

Nous constatons que la modélisation par la loi de Bernoulli permettant de modéliser la présence ou non de consommation par l'assuré ne requiert pas une explication plus précise de la variable âge pour les âges supérieurs à 50 ans. Les individus ayant un âge supérieur à 50 ans ont la même probabilité de consommer. Dans la partie BN, qui modélise le nombre de sinistres supérieur ou égal à 0, la variable âge peut être décomposée en intervalles plus petits pour les individus âgés de plus de 50 ans. Cela paraît cohérent avec la réalité, puisque pour les «jeunes assurés» la consommation en prothèses dentaires est rare. De ce fait, une modélisation analysant le fait de consommer ou non est adaptée pour ces classes d'âges. Cependant pour les « grands » âges, la consommation en prothèses dentaires est moins rare, il n'est plus question de consommer ou de ne pas consommer, mais de la quantité d'actes qui a été consommé. Ainsi, la modélisation d'une probabilité de sinistralité/non sinistralité n'est pas adaptée pour ces classes d'âge.

Cette application permet de comprendre l'intérêt d'un modèle modifié en zéro. En effet, dans le comptage du nombre de sinistres deux effets peuvent être distingués : le fait de consommer ou ne pas consommer et le fait de consommer 0, 1 ou plusieurs actes. En distinguant les variables et les modalités des variables pour ces deux processus, nous pouvons obtenir un tarif plus précis.

Le tableau en annexe (cf. Annexe D) contient les estimations des coefficients et des écart-types des paramètres β_j pour les deux parties du modèle ZINB. Nous nous intéressons ici au calcul des valeurs prédites puisque le modèle fournit simultanément les estimations des paramètres en distinguant les deux modèles. La partie Bernoulli du modèle étant réalisée à partir d'une fonction de lien logit, nous estimons la probabilité π :

$$\hat{\pi} = \frac{exp\{\hat{\gamma_0} + \hat{\gamma_1} z_1 + \cdots + \hat{\gamma_q} z_q\}}{1 + exp\{\hat{\gamma}_0 + \hat{\gamma}_1 z_1 + \cdots + \hat{\gamma}_q z_q\}}.$$

Or, nous avons précédemment fourni l'expression de l'espérance de la variable réponse dans le cas d'un modèle ZINB :

$$\widehat{E(Y/X)} = (1 - \hat{\pi})\hat{\mu} \quad o\text{ù} \quad \hat{\mu} = exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p + log(ann\acute{e}es\ risques)\}.$$

En remplaçant les termes $\hat{\mu}$ et $\hat{\pi}$ par leurs expressions respectives, nous obtenons :

$$\widehat{E(Y|X,Z)} = \frac{exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p + log(ann\acute{e}esrisques)\}}{1 + exp\{\hat{\gamma}_0 + \hat{\gamma}_1 z_1 + \cdots + \hat{\gamma}_q z_q\}}.$$

Les variables X_i et Z_j peuvent être identiques.

Le logiciel SAS nous fournit les probabilités moyennes suivantes de la variable expliquée, que nous comparons avec les fréquences empiriques :

Nombre de sinistres	ZINB	Observé
0	92,57%	92,48%
1	1,92%	3,06%
2	2,20%	1,44%
3	0,96%	0,85%
4	0,71%	0,55%
5	0,39%	0,38%
6	0,37%	0,27%
7	0,20%	0,20%
8	0,17%	0,15%
9	0,12%	0,12%
10	0,09%	0,09%
11	0,06%	0,07%
12	0,06%	0,06%
13	0,04%	0,05%
14	0,03%	0,04%
15	0,02%	0,03%
16	0,02%	0,03%
17	0,01%	0,02%
18	0,01%	0,02%
19	0,01%	0,02%
20	0,01%	0,01%
21	0,01%	0,01%
22	0,00%	0,01%
23	0,00%	0,01%
24	0,00%	0,01%
25	0,00%	0,01%
26	0,00%	0,01%
27	0,00%	0,00%
28	0,00%	0,00%
29	0,00%	0,00%

TABLE 19 – Probabilités moyennes observées et prédites par le modèle ZINB

Globalement, ce modèle fournit une meilleure adéquation aux données que le modèle de Poisson. Il estime mieux la probabilité d'avoir 0 et 1 sinistre malgré une probabilité plus élevée d'avoir 0 sinistre dans le cas d'un modèle ZINB. Toutefois, il convient de comparer ce modèle avec d'autres modèles avant de valider notre choix.

4.4 La comparaison des modèles

Les résultats obtenus en appliquant un modèle ZINB sont globalement satisfaisants. Cependant, le modèle ZINB est un modèle complexe à mettre en place, il serait inutile de choisir ce type de modèle alors qu'un modèle binomial négatif permettrait aussi bien de prendre en compte la sur-dispersion. Nous avons réalisé quatre procédures « countreg » avec les modèles Poisson, binomiale négative, ZIP et ZINB. La régression avec le modèle

ZIP a été réalisée de façon similaire au modèle ZINB.

Pour chaque modèle, nous avons calculé les probabilités moyennes afin de les comparer entre elles et avec la fréquence empirique. Le graphique ci-dessous permet de comparer les probabilités estimées pour les 10 premiers sinistres.

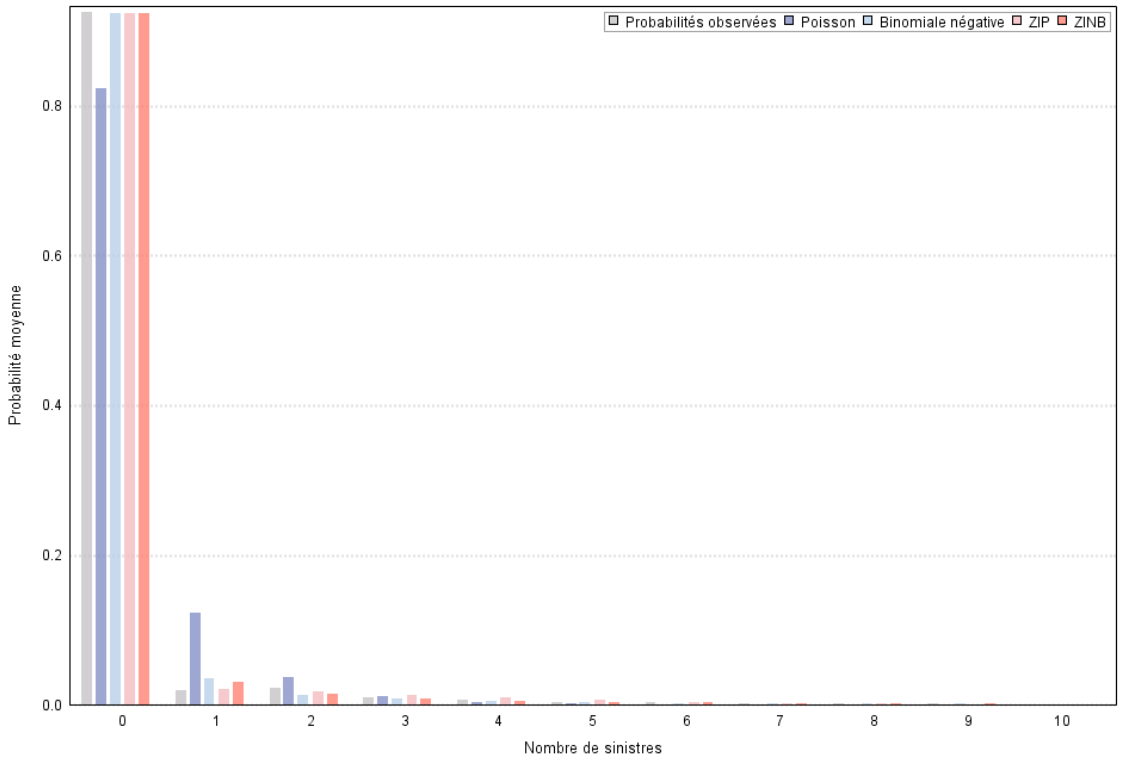


FIGURE 17 – Probabilités moyennes observées et prédites (modèles Poisson, binomial négatif, ZIP et ZINB)

Nous constatons ici que les trois modèles, le modèle binomial négatif, ZIP et ZINB permettent de prendre en compte l'excès de zéros observé sur le portefeuille. La comparaison peut s'effectuer à partir du premier sinistre, où les modèles binomial négatif et ZINB surestiment la probabilité, contrairement au modèle ZIP qui parait proposer une meilleure estimation. La probabilité d'avoir deux sinistres est quant à elle mal ajustée par l'ensemble des modèles. Au-delà de deux sinistres, les modèles binomial négatif et ZINB semblent donner des résultats très proches. Analysons plus précisément les différences entre les probabilités observées sur les données et les différents modèles étudiés à l'aide du graphique et du tableau suivant :

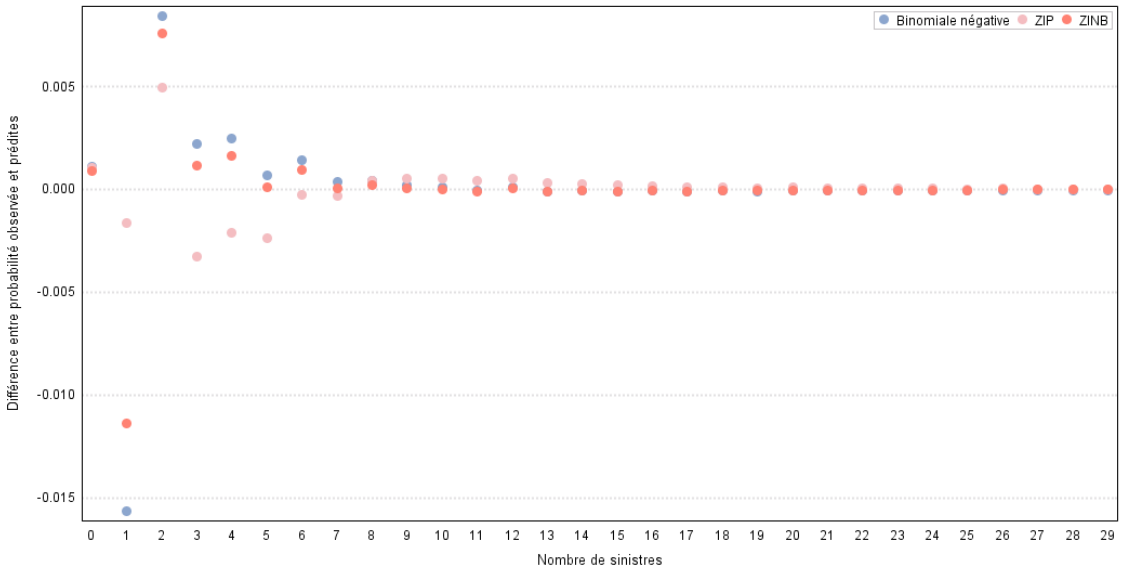


FIGURE 18 – Différence entre la probabilité observée et prédites (modèles Poisson, binomial négatif, ZIP et ZINB)

Nombre de sinistres	Poisson	Binomiale négative	ZIP	ZINB	Observé
0	82,40%	92,46%	92,46%	92,48%	**92,57%**
1	12,34%	3,49%	2,08%	3,06%	**1,92%**
2	3,61%	1,36%	1,71%	1,44%	**2,20%**
3	1,11%	0,74%	1,29%	0,85%	**0,96%**
4	0,36%	0,47%	0,92%	0,55%	**0,71%**
5	0,12%	0,32%	0,63%	0,38%	**0,39%**
6	0,04%	0,23%	0,40%	0,27%	**0,37%**
7	0,02%	0,17%	0,24%	0,20%	**0,20%**
8	0,01%	0,13%	0,13%	0,15%	**0,17%**
9	0,00%	0,10%	0,07%	0,12%	**0,12%**
10	0,00%	0,08%	0,04%	0,09%	**0,09%**
11	0,00%	0,07%	0,02%	0,07%	**0,06%**
12	0,00%	0,05%	0,01%	0,06%	**0,06%**
13	0,00%	0,04%	0,01%	0,05%	**0,04%**
14	0,00%	0,04%	0,00%	0,04%	**0,03%**
15	0,00%	0,03%	0,00%	0,03%	**0,02%**
16	0,00%	0,03%	0,00%	0,03%	**0,02%**
17	0,00%	0,02%	0,00%	0,02%	**0,01%**
18	0,00%	0,02%	0,00%	0,02%	**0,01%**
19	0,00%	0,02%	0,00%	0,02%	**0,01%**
20	0,00%	0,02%	0,00%	0,01%	**0,01%**
21	0,00%	0,01%	0,00%	0,01%	**0,01%**
22	0,00%	0,01%	0,00%	0,01%	**0,00%**
23	0,00%	0,01%	0,00%	0,01%	**0,00%**
24	0,00%	0,01%	0,00%	0,01%	**0,00%**
25	0,00%	0,01%	0,00%	0,01%	**0,00%**
26	0,00%	0,01%	0,00%	0,01%	**0,00%**
27	0,00%	0,01%	0,00%	0,00%	**0,00%**
28	0,00%	0,01%	0,00%	0,00%	**0,00%**
29	0,00%	0,01%	0,00%	0,00%	**0,00%**

TABLE 20 – Probabilités moyennes observées et prédites (modèles Poisson, binomial négatif, ZIP et ZINB)

Globalement les modèles ZINB et ZIP fournissent des résultats proches de la fréquence observée sur le portefeuille. Cette analyse permet également de mettre en évidence le mauvais ajustement du modèle ZIP pour un nombre de sinistres supérieur à huit puisque la probabilité estimée reste inférieure à la probabilité observée.

L'analyse des probabilités n'est pas suffisante pour comparer des modèles entre eux, il convient d'analyser les critères AIC et BIC :

Modèle	AIC	BIC
Poisson	1096820	1097078
Binomial négatif	659563	659739
ZIP	713511	713921
ZINB	**647352**	**647727**

TABLE 21 – Critères AIC et BIC des différents modèles

Les critères AIC et BIC fournis par les sorties SAS des procédures « countreg » et « genmod » permettent de confirmer le choix du modèle ZINB. En effet, la ZINB dispose de la plus faible valeur d'AIC et de BIC parmi les quatre modèles étudiés.

Il est également possible d'utiliser le test de Vuong afin de choisir, entre le modèle binomial négatif et le modèle ZINB, celui qui propose le meilleur ajustement aux données observées. Ce test peut aussi s'appliquer dans le cas du modèle de Poisson et du modèle ZIP.

Le test de Vuong met en place les hypothèses suivantes :

$$H_0 \ : \ \textit{les deux modèles s'ajustent aux données}$$

$$H_1 \ : \ \textit{un des deux modèles est plus adapté aux données.}$$

Si le test est significatif, alors l'hypothèse H_0 est rejetée par conséquent l'hypothèse H_1 est acceptée et nous réalisons un risque de premier espèce α. Le choix entre le premier et le deuxième modèle est déterminé en fonction du signe et de la valeur de la statistique de ce test.

Nous définissons la statistique de test ci-dessous, où f_1 et f_2 correspondent aux fonctions de densité des deux distributions testées :

$$V = \frac{\sqrt{n}\bar{m}}{\sigma_m^2}$$

$$m_i = log\frac{f_1(y_i)}{f_2(y_i)}$$

où $\bar{m}$ et σ_m^2 sont la moyenne et la variance du rapport de vraisemblance des deux distributions testées.

Sous l'hypothèse H_0, nous supposons que la statistique de Vuong peut être approximée par une loi normale centrée réduite. Ainsi, pour un niveau de significativité de 5%, nous avons :

- $V < -1,96$: choix du deuxième modèle ;
- $V \in [-1,96; 1,96]$: choix des deux modèles ;
- $V > 1,96$: choix du premier modèle.

Il est préférable d'utiliser les ajustements d'Akaike et de Schwarz dans le cas où les deux modèles n'ont pas le même nombre de coefficients. Une macro fournie par le logiciel SAS permet de réaliser ce test. Nous souhaitons ainsi connaître le meilleur modèle entre le modèle binomial négatif et le modèle modifié en zéro qui lui est associé.

Vuong Statistic	**Z**	**Pr>|Z|**	**Preferred Model**
Unadjusted	65.6582	<.0001	Zinb
Akaike Adjusted	65.5606	<.0001	Zinb
Schwarz Adjusted	65.0232	<.0001	Zinb

TABLE 22 – Test de Vuong : modèle binomial négatif - ZINB

La statistique de test est très largement supérieure à 1,96 avec une probabilité très significative pour les trois types de statistique de test. Ainsi, le test confirme à nouveau le choix de retenir le modèle binomial négatif modifié en zéro.
Ce test peut également être utilisé pour d'autres distributions. Le tableau ci-dessous fournit les résultats du test pour le modèle de Poisson et ZIP :

Vuong Statistic	**Z**	**Pr>|Z|**	**Preferred Model**
Unadjusted	148.2018	<.0001	Zip
Akaike Adjusted	148.1879	<.0001	Zip
Schwarz Adjusted	148.1065	<.0001	Zip

TABLE 23 – Test de Vuong : modèle Poisson - ZIP

Concernant les actes et analyses de laboratoire, étant donné la forte sur-dispersion des données, nous avons également eu recours à des modèles modifiés en zéro. Pour des raisons de lisibilité, les résultats obtenus pour les analyses et actes de laboratoire sont présentés en annexe (Annexe E).

4.5 Conclusion

Les lois classiques utilisées pour modéliser une variable de comptage ne sont pas adaptées à nos données. La surdispersion observée peut avoir plusieurs causes conjointes :

- non prise en compte de variables importantes pour l'explication de la consommation en frais de soins de santé, car elles ne sont pas disponibles dans notre portefeuille ;
- non fiabilité des données étudiées ;
- structure de la loi de Poisson inadaptée à la modélisation de la fréquence de consommation ;
- la présence d'une importante masse en zéro.

Concernant nos données, il s'agit principalement de la présence d'une importante masse en zéro pour les postes de garantie où le nombre d'actes consommés constitue un évènement rare. Cependant, la sur-dispersion a également été observée parmi les postes de

garantie où le nombre d'actes consommé est fréquent (masse en zéro plus faible), cela peut être lié à plusieurs raisons citées ci-dessus. Par conséquent, nous avons ajusté un modèle binomial négatif modifiée en zéro pour les postes de garanties où la non sinistralité est très représentative et un modèle binomial négatif pour les autres postes de garanties tels que la pharmacie ou les consultations de généralistes.

Chapitre 5

L'application à la modélisation du coût moyen

5.1 L'analyse de la variable expliquée

L'objet de cette partie est de modéliser le remboursement moyen d'un acte par la complémentaire santé. Pour cela, nous disposons pour chaque assuré du coût total de l'ensemble de sa consommation sur la période étudiée qu'il convient de diviser par le nombre d'actes consommés.

Les analyses et actes de laboratoire

Le graphique ci-dessous illustre la distribution empirique du coût moyen pour les analyses et les actes de laboratoire :

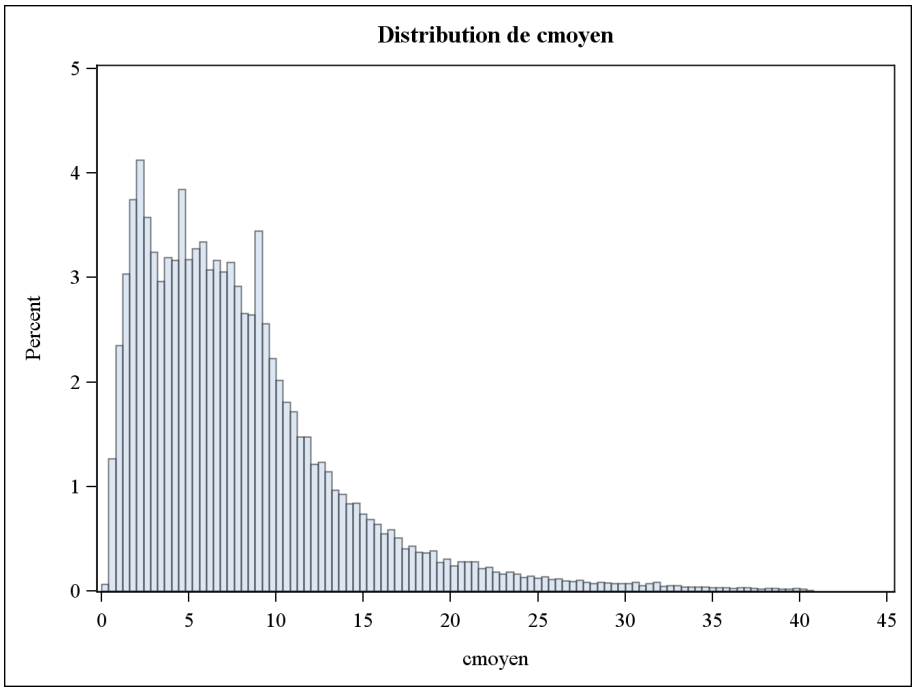


FIGURE 19 – Distribution du coût moyen pour les analyses et les actes de laboratoire

Moyenne	Ecart-type [19]	Coût maximum
7,96	5,99	40,63

TABLE 24 – Statistiques descriptives pour le coût moyen (analyses et actes de laboratoire)

Le remboursement en analyses et en actes de laboratoire repose sur des montants faibles. Nous observons une grande concentration des données sur l'intervalle [0-10 €], et une baisse progressive du nombre de données lorsque le coût augmente pour atteindre un coût maximum de 40,63 €.

Les prothèses dentaires

Le graphique ci-dessous illustre la distribution du coût moyen empirique pour les prothèses dentaires :

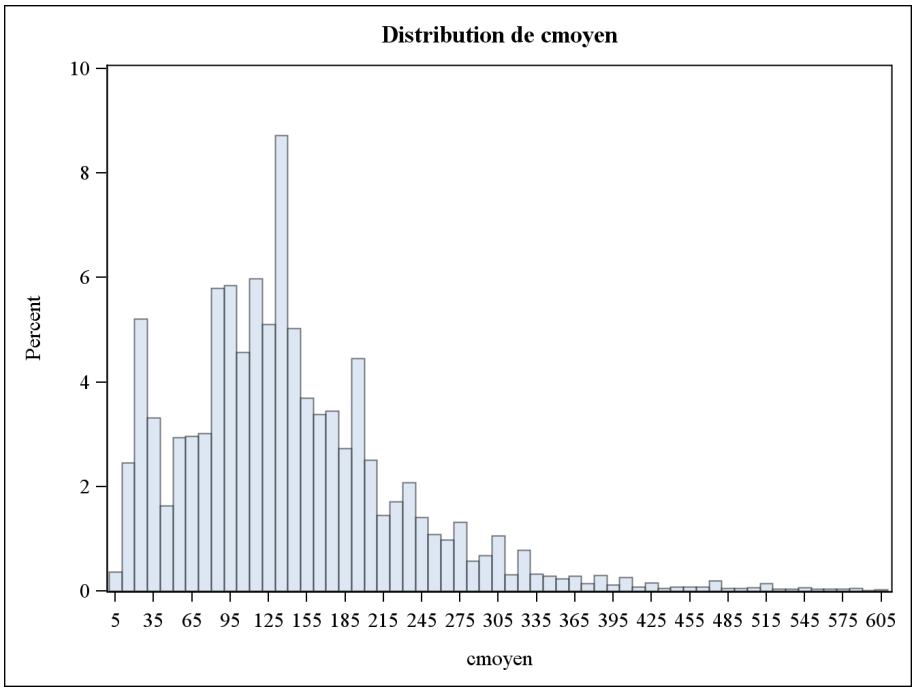


FIGURE 20 – Distribution du coût moyen (prothèses dentaires)

Moyenne	Écart-type [20]	Coût maximum
142,46	87,88	603,07

TABLE 25 – Statistiques descriptives du coût moyen(prothèses dentaires)

Compte tenu du coût des prothèses dentaires, il est évident que le remboursement moyen (142,46 €) est plus élevé que dans le cas des analyses et actes de laboratoire. Les

19. L'écart-type fourni par SAS est de la forme : $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

valeurs étant plus étendues avec la présence de valeurs très faibles et de valeurs très élevées, l'écart-type est égal à 87,88.

5.2 Le choix de la loi de probabilité

Nous appliquons dans cette partie les modèles linéaires généralisés à la modélisation du coût moyen d'un acte. Cette variable étant continue et positive, un modèle suivant une distribution classique telle que la loi normale n'est pas très approprié. Les distributions les plus utilisées dans ce cas sont les lois exponentielles, gamma et log-normale. Il s'agit ainsi de choisir le modèle proposant le meilleur ajustement aux données étudiées.

Les analyses et actes de laboratoire

La procédure « Univariate » permet de comparer la distribution empirique avec la fonction de densité des trois lois : exponentielle, gamma et log-normale. Les paramètres des différentes distributions sont estimés par maximum de vraisemblance et indiqués sous le graphique ci-dessous :

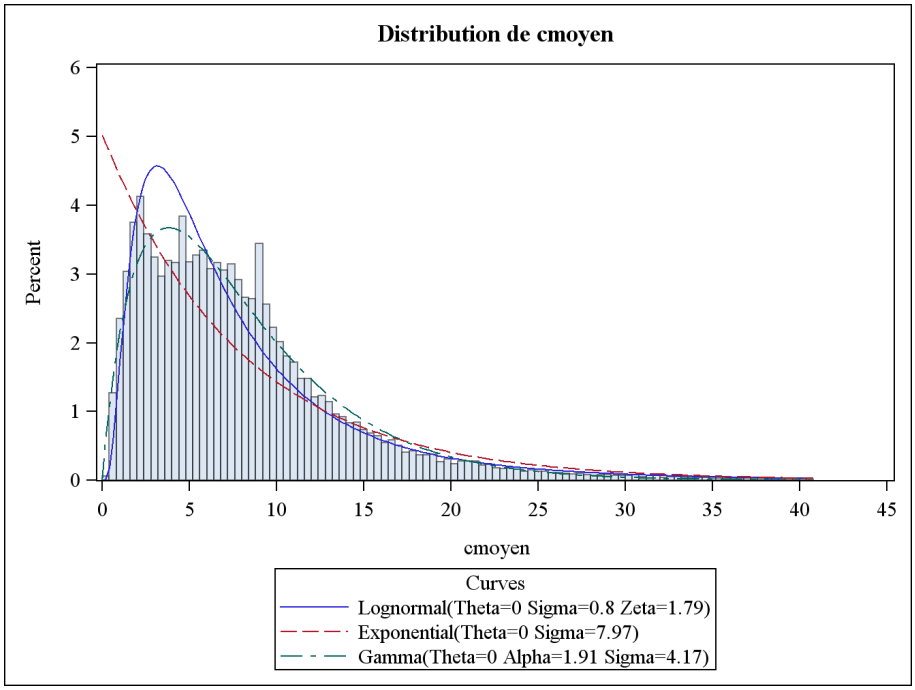


FIGURE 21 – Ajustement du coût moyen (les analyses et actes de laboratoire)

La forme de la distribution empirique correspond globalement à la forme de la loi gamma et de la loi log-normale. Néanmoins, la loi gamma semble mieux s'ajuster, puisque la loi log-normale accorde un poids important aux valeurs proches de 5. La loi exponentielle peut être écartée puisqu'elle ne permet pas de prendre en compte la forme en cloche de la distribution des données.

L'analyse graphique des histogrammes n'est pas suffisante pour choisir un modèle permettant le meilleur ajustement aux données. Elle permet uniquement d'avoir un avis sur

le type de distribution qui pourrait être utilisé. Cette analyse doit être complétée par une analyse graphique des Q-Q plot.

Le Q-Q plot (diagramme quantile-quantile) est une technique graphique employée pour vérifier la pertinence de l'ajustement d'une loi à des données empiriques. Le principe est de vérifier que les quantiles de la loi théorique correspondent aux quantiles des données étudiées. Le graphique représente en abscisses les quantiles de la loi théorique et en ordonnées les quantiles observés. Ainsi les points représentés sur le graphique doivent être alignés sur une droite.

Les graphiques ci-dessous représentent le Q-Q plot de la distribution des coûts moyens en analyse et actes de laboratoire par rapport aux lois gamma et log-normale :

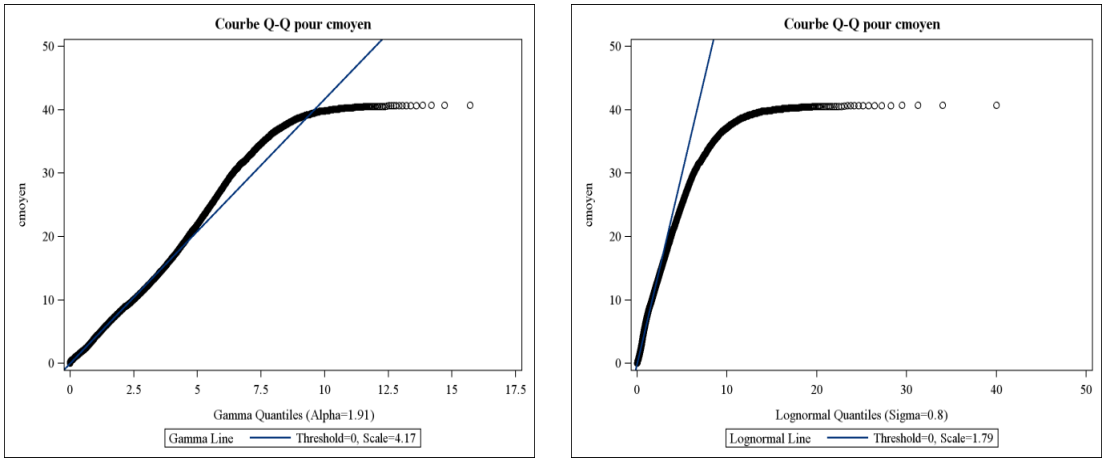


FIGURE 22 – Q-Q plot de la distribution du coût moyen (analyses et actes de laboratoire)

Le mauvais ajustement de la loi log-normale observé sur l'histogramme précédent est confirmé avec l'analyse des Q-Q plot. Les points sont alignés pour les premiers quantiles et se détachent progressivement de la bissectrice. Concernant la loi gamma, malgré un mauvais ajustement pour les quantiles extrêmes, nous observons que les points sont globalement alignés. Par ailleurs, le graphique ci-dessous, fourni par la procédure « Capability » sous SAS, conforte le choix d'utiliser la loi gamma. La fonction de répartition de la loi gamma est superposée à la fonction de répartition de nos données.

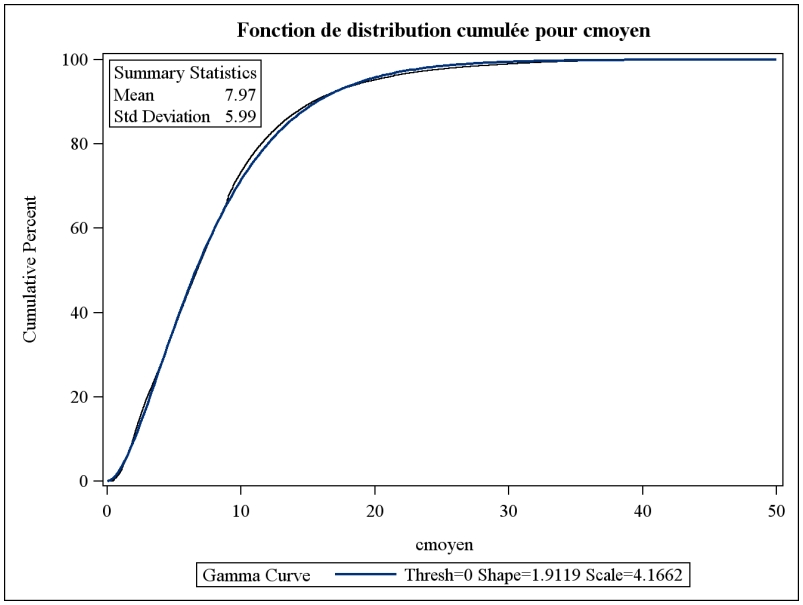


FIGURE 23 – Distribution cumulée du coût moyen (les analyses et actes de laboratoire)

Les prothèses dentaires

Similairement au poste de garantie « analyses et actes de laboratoire », il est nécessaire de choisir le modèle adéquat aux données étudiées pour les prothèses dentaires. Nous analysons ici directement les Q-Q plot de la distribution empirique par rapport aux lois gamma et log-normale :

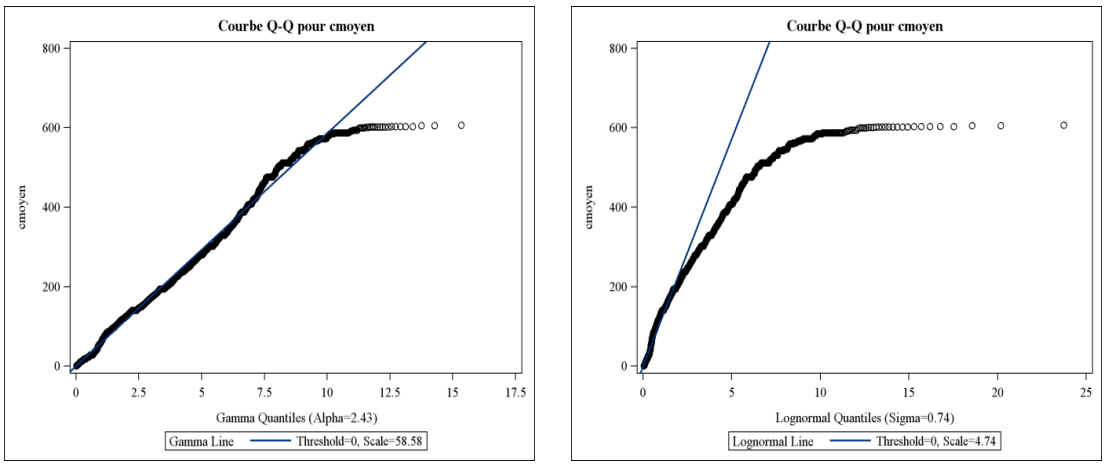


FIGURE 24 – Q-Q plot de la distribution du coût moyen (prothèses dentaires)

Dans le cas de la loi gamma, les points sont alignés sur la droite. Le Q-Q plot permet de valider l'hypothèse d'une meilleure adéquation de la loi gamma avec les données étudiées. Le Q-Q plot de la loi log-normale permet de rejeter l'hypothèse d'un ajustement pertinent des données au modèle théorique.

Nous confirmons notre choix avec l'étude de la fonction de répartition des données théoriques et de la loi gamma. Même si les deux courbes ne sont pas parfaitement superposées, le choix de valider l'ajustement d'une loi gamma est maintenu puisque les distances entre les points des deux courbes paraissent négligeables.

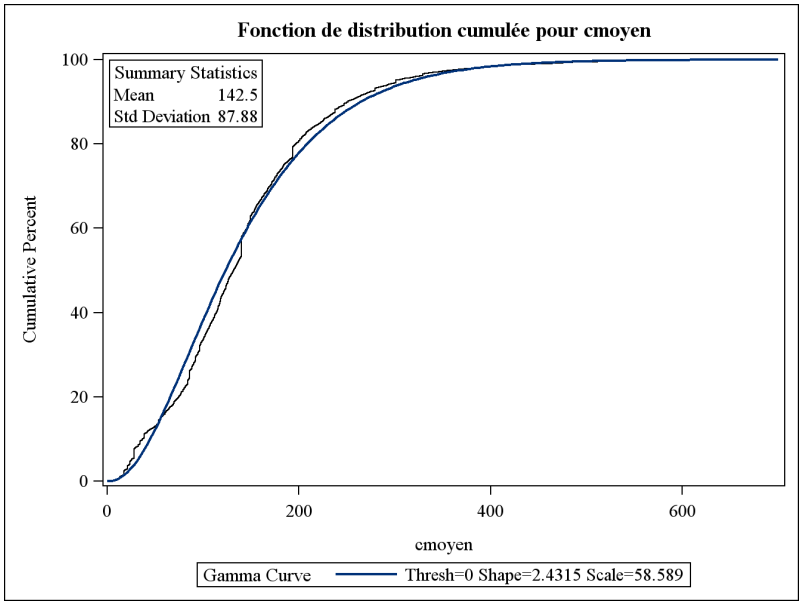


FIGURE 25 – Distribution cumulée du coût moyen (les prothèses dentaires)

5.3 L'estimation des paramètres

La régression en utilisant les modèles linéaires généralisés du coût moyen s'est effectuée soit par une loi gamma soit par une loi log-normale avec la procédure « Genmod ». En effet, pour chaque poste de garantie, la même analyse que ci-dessus a été effectuée et la loi la plus adéquate à nos données parmi ces deux distributions a été retenue.

Comme la loi log-normale ne fait pas partie de la famille exponentielle, l'idée est de modéliser le logarithme du coût moyen par une loi normale et une fonction de lien identité, puisque la loi normale appartient à la famille exponentielle. Étant donné que la loi théorique sélectionnée pour les deux familles d'actes est identique, nous présenterons uniquement les résultats des prothèses dentaires.

Les prothèses dentaires

Pour ce poste de garantie, les variables sélectionnées avec la méthode stepwise sont uniquement le niveau de garantie du contrat (niveau 1 et 2 regroupé), l'âge et le régime d'adhésion. Les valeurs estimées des paramètres sont données ci-dessous :

Variable	Modalité	Valeur estimée
Intercept		5,322
garantie	2	-0,930
garantie	3	-0,603
garantie	4	-0,210
garantie	5	0,000
age	0-20	0,091
age	21-30	0,107
age	31-50	0,106
age	51-60	0,057
age	61-70	0,036
age	71et plus	0,000
regime	General	0,0680
regime	Local	0,000

TABLE 26 – Paramètres estimés pour les prothèses dentaires

Les graphiques ci-dessous fournissent les valeurs prédites pour chaque modalité, ce qui correspond à l'exponentielle des valeurs estimées du tableau ci-dessus :

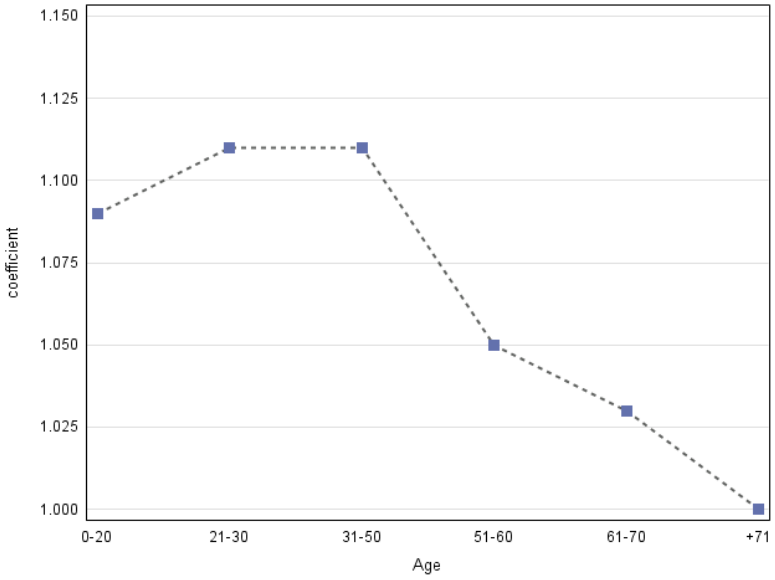


FIGURE 26 – Coefficients du GLM relatifs à l'âge

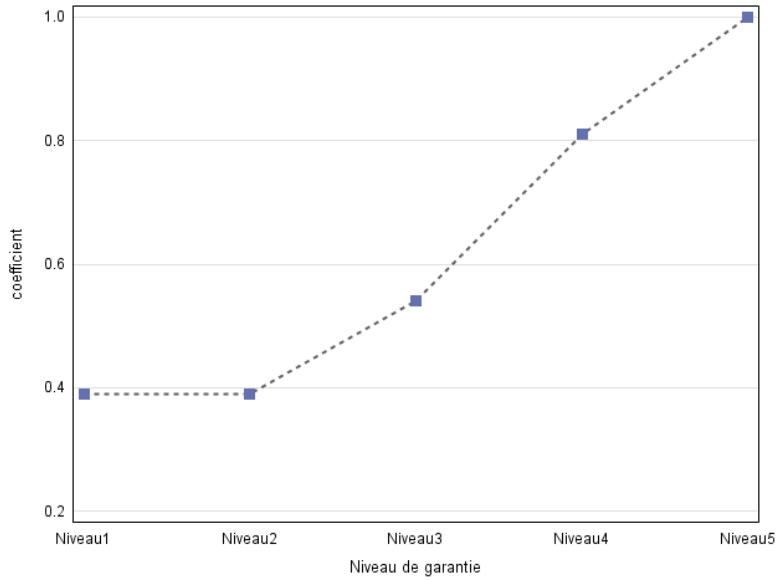


FIGURE 27 – Coefficients du GLM relatifs au niveau de garantie

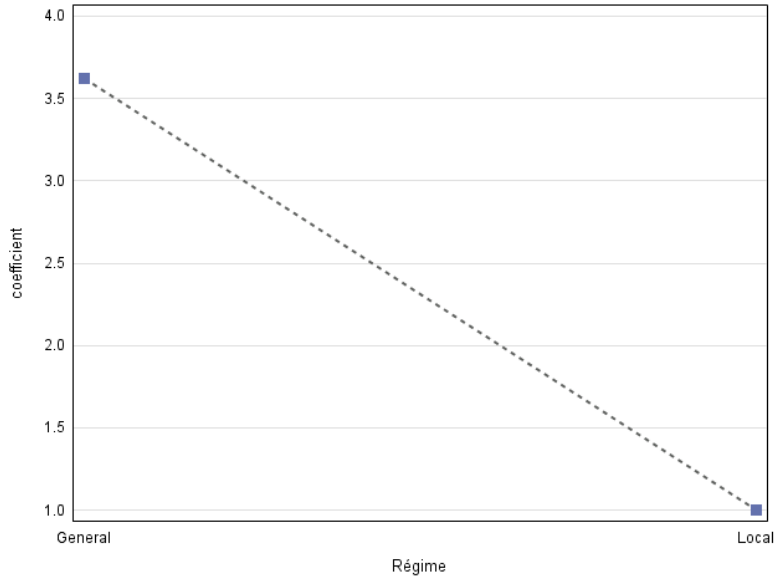


FIGURE 28 – Coefficients du GLM relatifs au régime d'adhésion

Les différents coefficients représentés sur ces graphiques semblent être cohérents. En effet, le remboursement moyen des frais de soins de santé augmente en fonction du niveau de garantie du contrat, et est plus élevé au régime général qu'au régime local. Concernant l'âge, il est plus difficile d'interpréter les coefficients. Il semblerait que les personnes âgées ont recours à des soins ayant un reste à charge plus faible suite au remboursement de la Sécurité sociale.

5.4 L'analyse des résidus du modèle sélectionné

Afin de valider la régression effectuée, il est nécessaire d'analyser les résidus du modèle. Pour cela, nous vérifierons :

- la répartition des résidus en fonction des valeurs prédites pour détecter d'éventuels points aberrants ;
- la répartition des résidus autour de la valeur zéro de façon symétrique pour valider l'hypothèse d'homoscédasticité et d'espérance nulle ;

Nous représentons les résidus pour les prothèses dentaires :

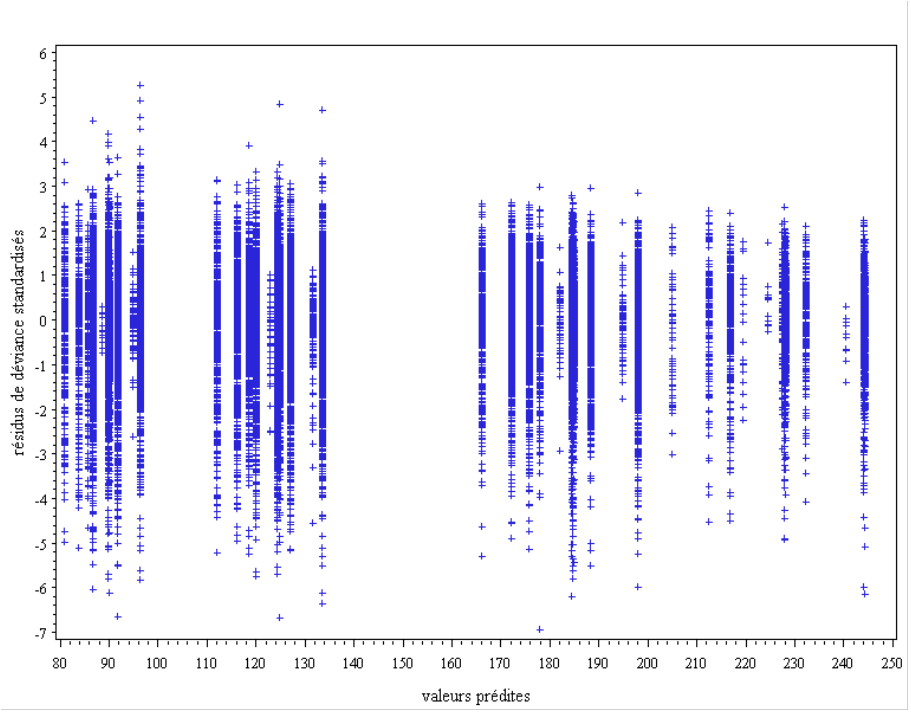


FIGURE 29 – Résidus de déviance standardisés en fonction des valeurs prédites

La répartition des résidus pour chaque valeur prédite ne permet pas de détecter de valeurs aberrantes ou de tendance particulière. Les résidus semblent être plus ou moins centrés en zéro, ce qui vérifie l'hypothèse d'espérance nulle. Cependant, l'hypothèse de variance constante des erreurs ne semble pas vérifiée, puisque les résidus ne sont pas répartis de façon symétrique. Nous pouvons observer des variances plus élevées pour les valeurs prédites comprises entre 80 €et 135 €, contrairement aux valeurs prédites comprises entre 165 €et 250 €.

Analysons à présent les résidus en fonction des variables explicatives sélectionnées par le modèle :

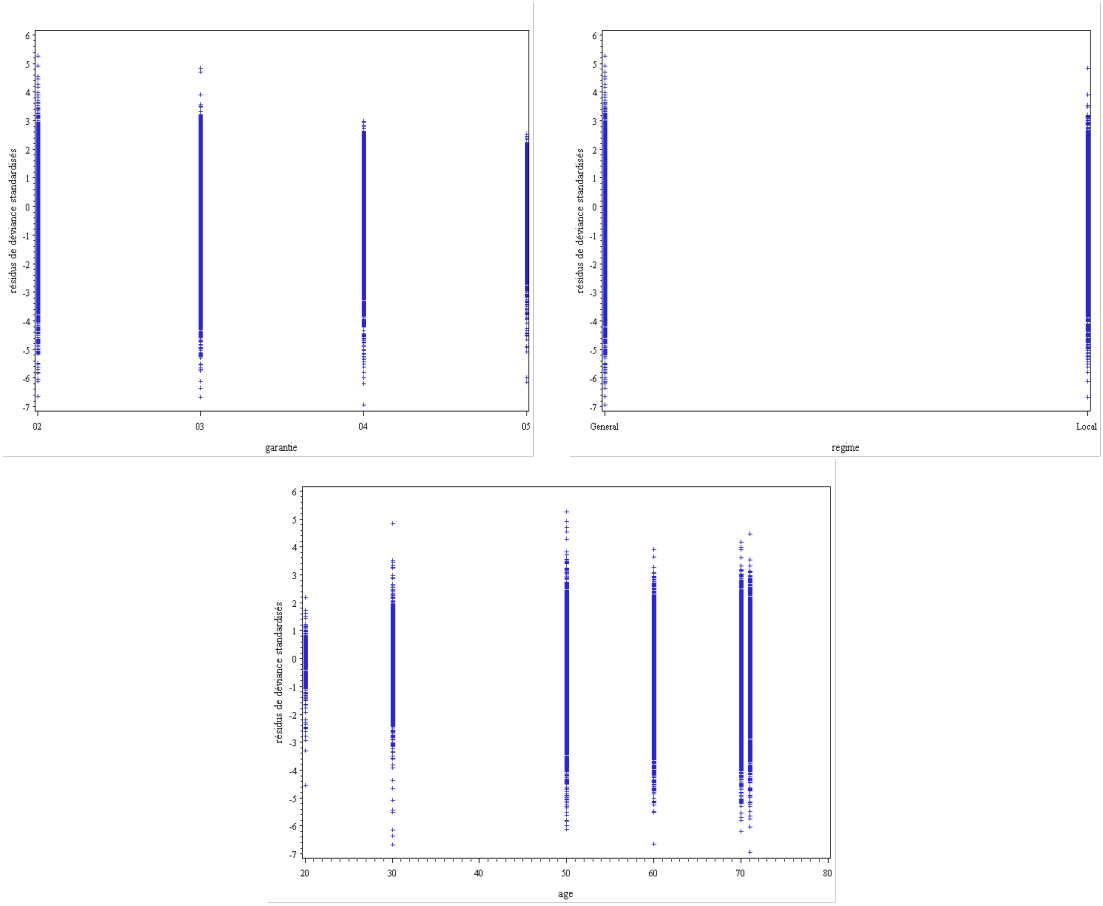


FIGURE 30 – Résidus de déviance standardisés en fonction des variables explicatives

Les résidus représentés en fonction de la garantie et du régime semblent être globalement de variance constante. Contrairement à l'âge, où la variance de la première classe d'âge semble être significativement inférieure à celle des autres classes d'âge.

5.5 Conclusion

Nous avons effectué un GLM sur les coûts moyens pour pratiquement chaque sous-poste de garanties. Les sous-postes de garantie tels que le remboursement du forfait télévision ou du lit d'accompagnant n'ont pas été modélisé, compte tenu du faible nombre de données et du faible impact de ces garanties sur la prime.

La loi de probabilité adaptée pour chaque modélisation a été sélectionnée en fonction des Q-Q plot. Dans l'ensemble, il semblerait que la loi gamma soit plus adaptée que la loi log normale, hormis pour trois sous-postes de garantie modélisés par une loi log normale : les honoraires en hospitalisation, la maternité et le transport.

Similairement aux prothèses dentaires, nous avons analysé la cohérence des coefficients estimés par le GLM. Nous n'avons détecté aucune incohérence pour chacun des postes de garanties.

Enfin, afin de vérifier les hypothèses du GLM, une analyse graphique des résidus de chaque modèle a été effectuée. Globalement, nous avons obtenu les mêmes résultats que les résidus des prothèses dentaires prises en charge par la Sécurité sociale. L'hypothèse d'espérance nulle semblerait être vérifiée contrairement à l'hypothèse d'homogénéité des variances. Ces résultats ne nous permettent pas de valider les GLM du coût moyen.

Chapitre 6

La comparaison avec la méthode directe

Dans ce chapitre, nous proposons dans un premier temps une analyse de cohérence des coefficients estimés avec le GLM, et ensuite une comparaison de la prime estimée avec le GLM et la méthode directe.

6.1 La cohérence de la prime estimée avec le GLM

Dans le cadre des prothèses dentaires, les variables retenues pour les GLM relatifs à la fréquence et au coût moyen sont l'âge, le niveau de garantie et le sexe. Ainsi, analysons l'évolution de la prime pure en fonction de l'âge pour chaque niveau de garantie, sachant que la fréquence a été estimée par un modèle ZINB.

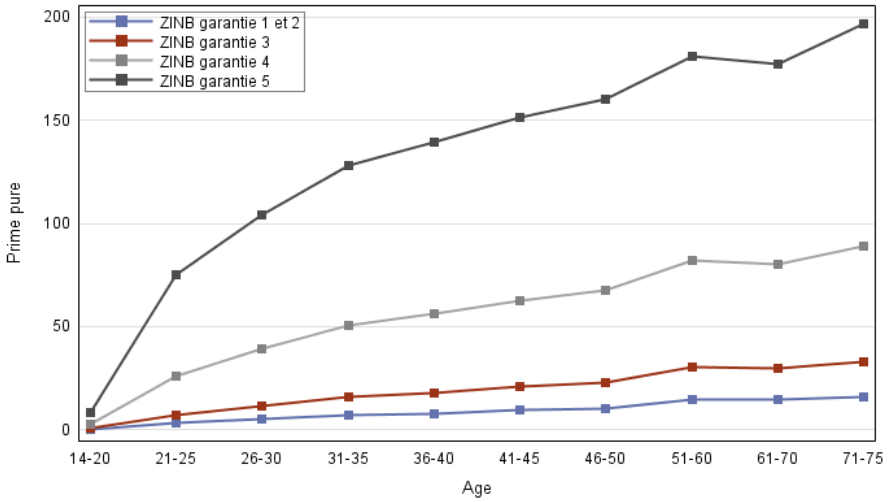


FIGURE 31 – Prime pure en fonction du niveau de garantie

Les coefficients estimés avec le GLM paraissent cohérents, puisque la prime augmente en fonction de l'âge. Par ailleurs, il est également évident que nous observons une translation vers le haut des courbes de consommation par niveaux de garantie, puisque d'une part la fréquence de consommation augmente avec le niveau de garantie et d'autre part le

remboursement d'un acte dépend du niveau de la garantie. Cependant, nous remarquons que l'écart de prime entre deux niveaux de garantie augmente en fonction du niveau de garantie.

Il est à noter que la forme des courbes est due à la tarification par classes d'âge établie dans le cadre du GLM.

Analysons à présent l'évolution de la prime par sexe :

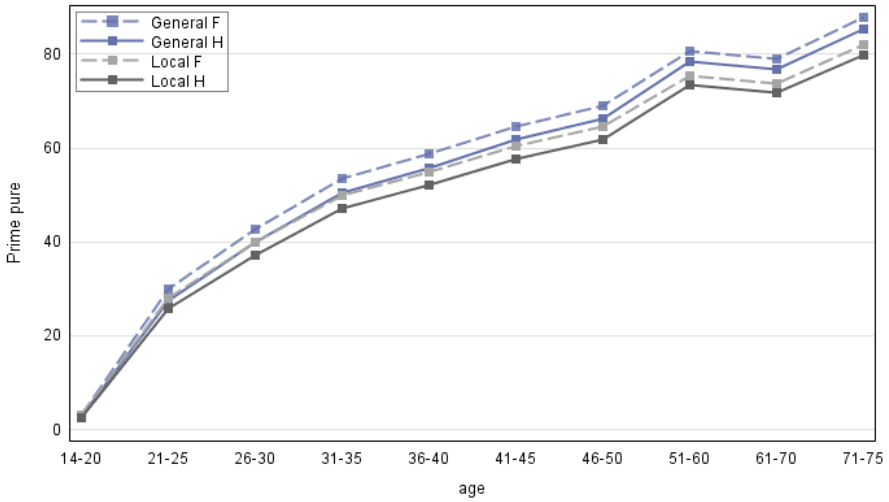


FIGURE 32 – Prime pure en fonction du sexe et du régime d'adhésion (prothèses dentaires

Les primes sont à nouveau cohérentes puisque la prime est légèrement plus élevée pour les femmes, considérées comme ayant une fréquence de consommation plus élevée.

Dans le cadre des analyses et actes de laboratoire :

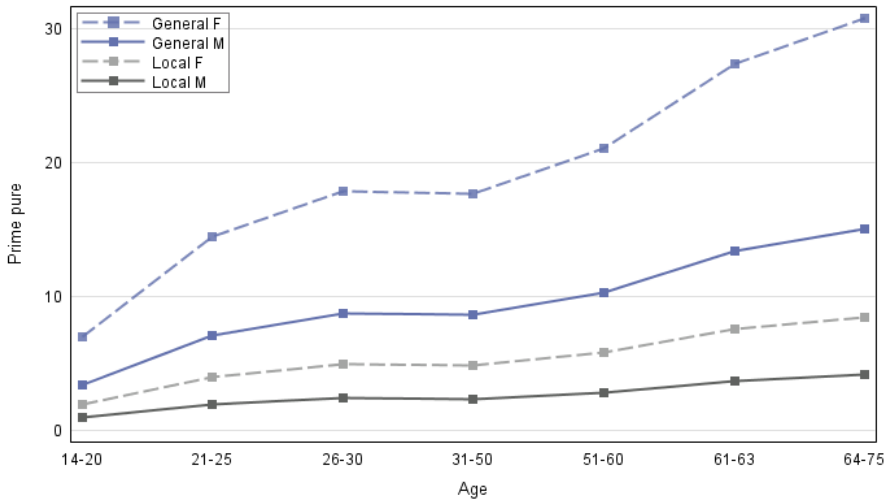


FIGURE 33 – Prime pure en fonction du sexe et du régime d'adhésion (analyses en actes de laboratoire)

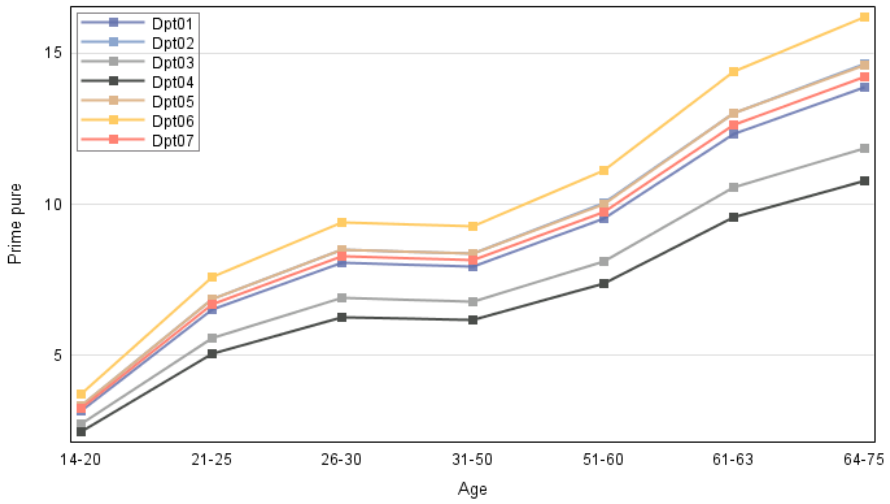


FIGURE 34 – Prime pure en fonction du département

6.2 La comparaison avec la méthode directe

6.2.1 La présentation de la méthode directe

La tarification actuelle utilisée afin d'estimer la prime pure pour un contrat collectif santé sur-mesure est basée majoritairement sur une méthode de détermination directe de la fréquence et du coût moyen. Le tarif tient compte de différentes variables qui ont été utilisées dans le cadre du modèle linéaire généralisé à l'exception du lieu d'habitation, du type de bénéficiaire et du nombre d'enfants. Il est à noter que la variable « âge » n'est pas regroupée sous forme de classes d'âge. Les fréquences et coûts moyens sont dans un

second temps lissés par la méthode la plus adaptée à chaque poste de garantie (ajustement paramétrique, lissage Whittaker-Henderson,etc.).

Pour chaque ensemble d'actes (garantie lentilles, consultations généralistes, etc.), un tarif annuel est déterminé en fonction des variables impactant significativement la fréquence et le coût moyen. L'agrégation des primes des différents postes de garantie fournit la prime finale proposée à l'entreprise avant la prise en compte des différents frais et taxes.

Cette méthode de tarification, basée sur les données 2009 à 2011, a été actualisée avec les données 2011 à 2013.

6.2.2 Comparaison de la prime pure

L'analyse des résidus du GLM appliqué aux coûts moyens ne nous conduit pas à valider l'utilisation d'un GLM pour l'estimation de la prime. Toutefois, nous souhaitons comparer les résultats de la méthode directe avec ceux du GLM. Il serait également intéressant ici de comparer les modèles ZINB, ZIP et binomial négatif avec la fréquence déterminée avec la méthode directe afin d'analyser le modèle s'en rapprochant le plus (pour un même modèle de coût gamma). Nous nous intéressons à nouveau à la tarification des prothèses dentaires acceptées par la Sécurité sociale et à la tarification des actes et des analyses de laboratoire, par soucis de lisibilité.

La méthode directe est comparée à trois modèles de GLM :

	Loi de probabilité pour la fréquence	Loi de probabilité pour le coût moyen
Modèle 1	Binomiale négative	Gamma
Modèle 2	ZIP	Gamma
Modèle 3	ZINB	Gamma

Table 27 – Les différents modèles de GLM

Pour commencer, analysons l'évolution de la prime pure moyenne en fonction de l'âge pour les prothèses dentaires :

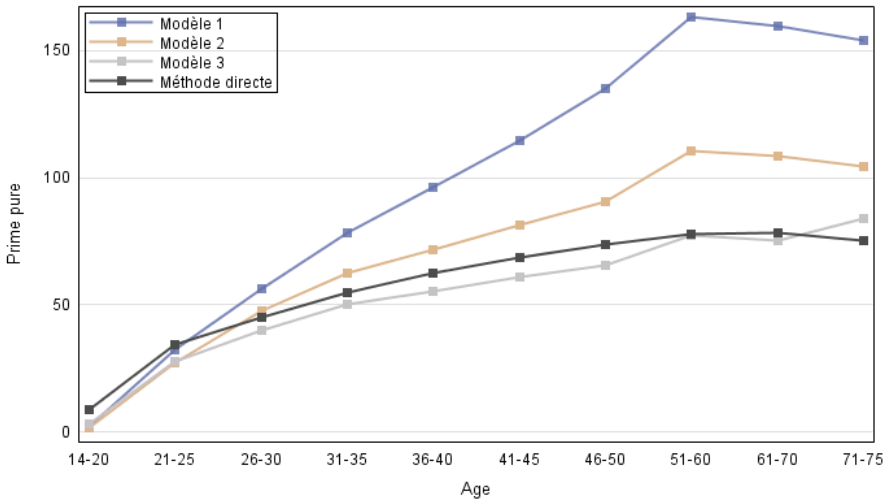


FIGURE 35 – Comparaison de la prime pure (prothèses dentaires)

Age	Méthode directe	Modèle 1	Modèle 2	Modèle 3
14-20	8,65	2,00	1,78	2,91
21-25	34,45	32,40	27,39	27,88
26-30	45,24	56,18	47,66	39,94
31-35	54,53	78,39	62,56	50,24
36-40	62,34	96,23	71,41	55,33
41-45	68,65	114,69	81,50	61,12
46-50	73,48	134,96	90,34	65,40
51-60	77,73	163,21	110,53	77,04
61-70	78,44	159,82	108,24	75,44
71-75	75,24	153,54	104,10	83,57

TABLE 28 – Prime pure par âge et par type de modèle (prothèses dentaires)

Les différences par rapport au modèle directe :

Age	Modèle 1	Modèle 2	Modèle 3
14-25	-76,84%	-79,47%	-66,35%
26-30	24,18%	5,35%	-11,72%
31-35	43,75%	14,72%	-7,87%
36-40	54,37%	14,56%	-11,25%
41-45	67,06%	18,72%	-10,97%
46-50	83,67%	22,95%	-11,00%
51-60	109,97%	42,20%	-0,88%
61-70	103,75%	37,99%	-3,82%
71-75	104,87%	38,75%	11,38%

TABLE 29 – Différence entre la prime pure calculée par la méthode directe et par les différents modèles GLM (prothèses dentaires)

Le graphique précédent montre bien que le modèle binomial négatif surestime fortement la fréquence et ainsi la prime pure, qui est dû au mauvais ajustement de ce modèle à nos données. Le modèle ZINB, quant à lui, se rapproche beaucoup de la méthode directe. Le tarif estimé est globalement légèrement inférieur à la méthode directe.

Analysons de façon similaire l'évolution de la prime pure moyenne en fonction de l'âge pour les analyses et actes de laboratoire :

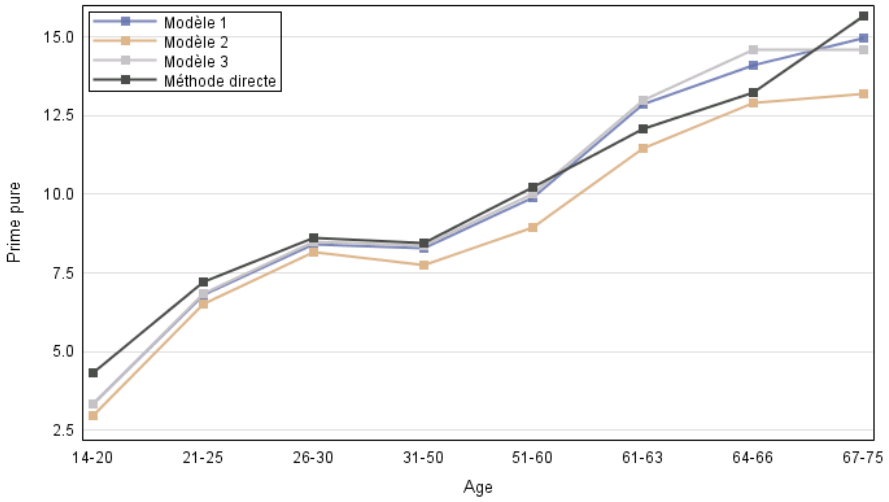


FIGURE 36 – Comparaison de la prime pure (analyses et actes de laboratoire)

Age	Méthode directe	Modèle 1	Modèle 2	Modèle 3
14-20	4,34	3,31	2,95	3,35
21-25	7,23	6,79	6,53	6,86
26-30	8,61	8,4	8,18	8,49
31-50	8,46	8,28	7,77	8,37
51-60	10,21	9,91	8,94	10,02
61-63	12,09	12,87	11,48	12,99
64-66	13,25	14,1	12,9	14,62
67-75	15,68	14,96	13,19	14,62

TABLE 30 – Prime pure par âge et par type de modèle (prothèses dentaires)

Les différences par rapport au modèle directe :

Age	Modèle 1	Modèle 2	Modèle 3
14-20	-23,74%	-32,04%	-22,91%
21-25	-6,16%	-9,73%	-5,12%
26-30	-2,43%	-5,01%	-1,38%
31-50	-2,04%	-8,12%	-0,96%
51-60	-2,98%	-12,44%	-1,92%
61-63	6,42%	-5,05%	7,45%
64-66	6,41%	-2,67%	10,29%
67-75	-4,58%	-15,83%	-6,75%

TABLE 31 – Différence entre la prime pure calculée par la méthode directe et par les différents modèles GLM (prothèses dentaires)

6.3 La conclusion et limites du GLM

Dans le cadre de la modélisation de la fréquence, l'application du GLM aux données étudiées a mis en évidence un problème de dispersion des données qui n'a pas pu être totalement pris en compte par un modèle binomial négatif. Ainsi, des modèles plus complexes ont dû être utilisés. L'utilisation de ces modèles peut être contraignante puisqu'il s'agit de sélectionner simultanément des variables explicatives pour deux processus distincts, sollicitant beaucoup de temps. Par ailleurs, contrairement aux coefficients estimés avec la loi de Poisson, ceux qui sont estimés par ces modèles ne sont pas faciles à interpréter comme dans le cas de la loi de Poisson.

Dans le cadre du coût moyen, l'analyse des résidus n'a pas permis de valider le modèle GLM malgré le choix préalable de la loi la plus adaptée aux données étudiées. Le rejet du modèle par l'analyse des résidus a été observé sur la majorité des postes et sous postes de garanties étudiées.
Par conséquent, nous avons décidé de ne pas retenir le GLM.

Dans la deuxième partie du mémoire, l'analyse des résultats de l'ACP a permis de mettre en évidence la présence d'une certaine corrélation linéaire entre la fréquence de consommation et le coût moyen notamment pour les actes de consultations et de visites chez les auxiliaires. En général, ce phénomène est très présent dans le cadre de l'optique, même si elle n'a pas pu être confirmée par l'ACP (mauvaise représentation des variables sur les axes factoriels). La valeur de la prime peut être ainsi biaisée, puisqu'elle repose sur l'hypothèse d'indépendance entre la fréquence et le coût moyen. Une solution serait d'estimer directement la prime sans distinguer la fréquence et le coût moyen. Mais, ce choix ne permettrait pas de réaliser un tarif précis, puisque les variables influant la fréquence de consommation ne sont pas forcément les mêmes que les variables ayant une influence sur le coût moyen.

Quatrième partie

L'analyse et la mesure du risque d'antisélection

La commercialisation de contrats collectifs facultatifs en assurance santé suscite des interrogations sur la prise en compte du risque d'antisélection. L'objet de cette partie est de définir ce phénomène, de l'analyser dans une optique micro-économique et statistique et de proposer un ou plusieurs coefficients de majoration à retenir pour la tarification de tels contrats.

Chapitre 1

La présentation du phénomène d'antisélection

1.1 Définition générale

Avant de définir la notion d'antisélection, prédéfinissons de façon plus générale la notion d'asymétrie d'information. En effet, l'asymétrie d'information peut être définie comme le fait que l'information n'est pas partagée ou connue par tous.

Par exemple :

- dans le domaine de l'assurance, lorsque l'assureur n'a pas la même connaissance du risque de l'assuré que l'assuré lui-même ;
- au niveau d'une société quelconque, lorsque les dirigeants et les investisseurs ne disposent pas des mêmes renseignements concernant la société.

Deux phénomènes peuvent être distingués en présence d'asymétrie d'information : l'aléa moral (ou risque moral) et l'antisélection (ou sélection adverse).

Dans le cas d'un contrat d'assurance, l'aléa moral peut être défini comme ci-dessous selon l'économètre français P. A. Chiappori :

> « L'on parle d'aléa moral lorsqu'une spécificité du contrat induit chez l'assuré un comportement non observable par l'assureur contraire à l'intérêt commun. »

Ainsi ce sont les caractéristiques du contrat qui auront une influence sur la consommation ou le nombre de sinistres de l'assuré. Dans le cas de l'assurance santé, est-ce que l'assuré consommera plus ou moins d'actes médicaux en fonction des garanties de son contrat ?

L'antisélection se distingue de l'aléa moral car elle ne porte pas sur les actions des individus. Il est question d'antisélection, dans le cadre d'un contrat d'assurance, lorsque les assurés ayant la possibilité ou non de souscrire le contrat, détiennent une information sur leur risque non connue par l'assureur. Dans ce cas, les personnes ayant une forte probabilité d'avoir un sinistre seront plus intéressées par le contrat d'assurance que les personnes ayant une faible probabilité. La notion d'antisélection est liée à l'incapacité de l'assureur de distinguer les « bons » risques des « mauvais » risques.

Les conséquences de la présence d'aléa moral et d'antisélection sont significatives pour un assureur. Nous montrerons dans une analyse micro-économique qu'elle constitue une limite au bon fonctionnement des marchés d'assurance.

1.2 L'approche économique

L'antisélection est considérée comme un phénomène économique que nous proposons d'appréhender de manière statistique. Avant de présenter une approche statistique dans le chapitre suivant, il semble légitime ici de développer une approche économique dans le domaine de l'assurance.

Trois principaux auteurs analysant le phénomène de sélection adverse peuvent être cités : Arrow, Rothschild et Stiglitz. Dans le cas de l'antisélection, nous présenterons le modèle et brièvement les résultats de l'article de Rothschild et Stiglitz (1976), « Equilibrium in Complete Insurance Markets : an Essay on the Economics of Imperfect Information ». Il est à noter que l'intention n'est pas d'exposer la théorie économique relative à l'antisélection, mais uniquement de mettre en évidence la problématique par le biais d'outils économiques.

1.2.1 Le cadre général du modèle

Les auteurs considèrent un marché d'assurance de concurrence parfaite où ils distinguent deux types d'assurés :

- les assurés avec un niveau de risque élevé (les hauts risques) d'une proportion λ (connue par l'assureur) disposant d'une probabilité de sinistre p^H ;
- les assurés avec un niveau de risque faible (les bas risques) d'une proportion $1-\lambda$ disposant d'une probabilité de sinistre p^L.

D'où la probabilité moyenne de survenance d'un sinistre $\hat{p}$ dans la population :

$$\hat{p} = \lambda p^H + (1-\lambda)p^L$$

$$o\text{ù } p^H > p^L.$$

Les auteurs considèrent deux états de la nature, W_1 et W_2 représentant respectivement la richesse de l'assuré lorsque l'assuré ne subit pas de sinistres et lorsqu'il subit un sinistre. Les assurés ont tous la même richesse initiale indépendamment du type de risque. Les préférences $V(.)$ [21] des deux types d'assurés sont caractérisées par la même fonction d'utilité $U(.)$ où U est strictement croissante et concave [22] :

$$V_H(p_H, W_1, W_2) = (1-p_H)U(W_1) + p_H U(W_2)$$

21. Les choix des assurés sont caractérisés par la fonction $V(.)$ représentant une fonction de l'utilité espérée

22. Une fonction d'utilité concave indique l'aversion au risque des assurés.

$$V_L(p_L, W_1, W_2) = (1 - p_L)U(W_1) + p_L U(W_2)$$

Nous caractérisons le contrat d'assurance par $\alpha = (\pi, Q)$ où π représente le profit de l'assureur pour une couverture d'assurance Q (quantité). Si Q est égale au montant du sinistre, il s'agit d'une couverture complète et si Q est inférieure au montant du sinistre, il s'agit ainsi d'une couverture partielle ou d'un contrat avec franchise.

L'équilibre selon Rothschild et Stiglitz a été défini par les deux conditions suivantes :

- le contrat réalise un profit non négatif ;
- il n'existe pas d'autres contrats qui, s'il était proposé, réaliserait un profit positif.

Par ailleurs, l'hypothèse de concurrence parfaite suppose que le profit espéré doit être nul à l'équilibre.

1.2.2 L'équilibre en information parfaite

Afin de déterminer l'équilibre en information parfaite, nous procédons par une analyse graphique [23].

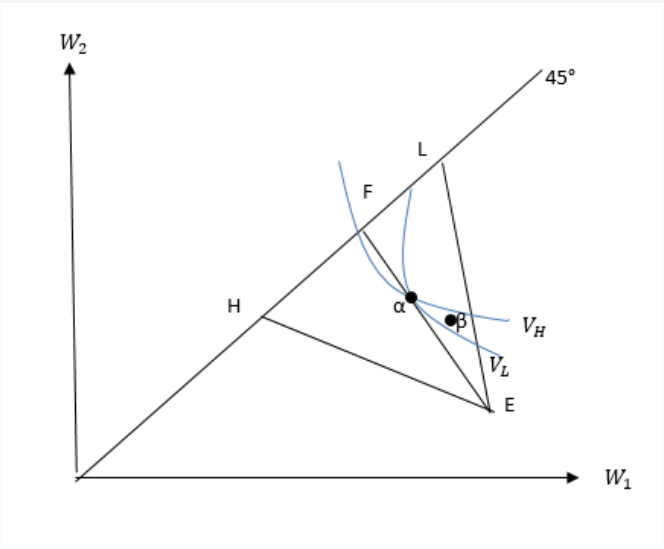


FIGURE 37 – Équilibre en information parfaite

Le graphique ci-dessus permet de visualiser l'équilibre lorsque l'assureur connait parfaitement le type de risque de l'assuré.

Définissons à présent les différents éléments du graphique :

23. Graphiques inspirés de l'article de Rothschild et Stiglitz, « Equilibrium in Complete Insurance Markets : an Essay on the Economics of Imperfect Information »

– La droite de 45 degrés représente toutes les situations qui permettent d'avoir un même niveau de richesse, indépendamment de la présence ou non de sinistres (couverture totale). Ainsi, tout contrat situé sur cette droite assure un niveau de couverture complet et tous points en-dessous de cette courbe représentent un contrat avec franchise ;
– Les préférences des deux types d'individus sont représentées par les courbes d'indifférence des fonctions d'utilité V_H et V_L, c'est-à-dire toutes les situations de richesse ($W1$,$W2$) qui conduisent l'assuré au même niveau de bien-être. Le niveau de bien-être de l'individu augmente lorsque la courbe d'indifférence s'éloigne de l'origine.
– Les contrats situés sur les droites (EH) et (EL) sont au prix actuariel (droites actuarielles). En effet, tous contrats situés sur (EH) et souscrit par les hauts risques assurent un niveau de profit espéré nul. De la même façon pour (EL), puisque l'assureur obtient un profit espéré nul uniquement si le contrat est souscrit par un bas risque.

Dans ce cas, l'équilibre est atteint en maximisant les fonctions d'utilité des assurés de chaque type avec la contrainte de profit espéré nul pour l'assureur. Le contrat doit donc se situer sur la courbe d'indifférence respective la plus élevée tangente à la droite actuarielle respective. Par conséquent, la solution est un niveau de couverture complet pour chaque type d'assuré en échange d'une prime d'assurance actuarielle.

1.2.3 Le problème en présence d'antisélection

En cas de présence d'antisélection, l'assureur est informé de la présence de deux types de risque, connaît la proportion de personnes présentes pour chaque type de risque, mais n'a pas connaissance du profil de risque de chaque assuré observé individuellement.
Analysons le graphique suivant :

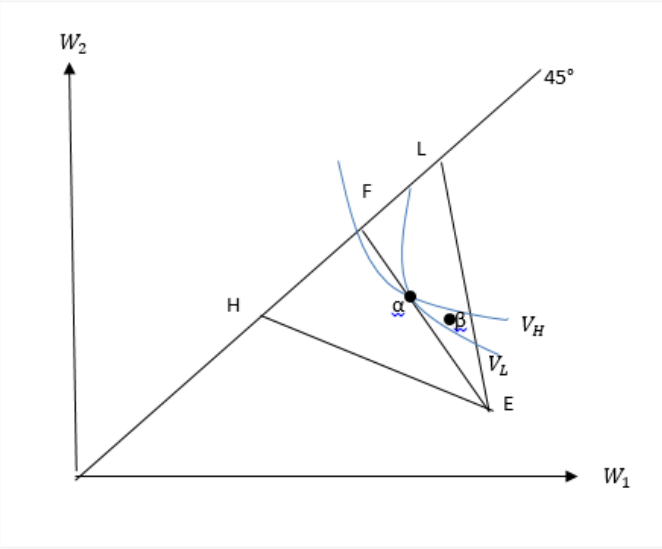


FIGURE 38 – Équilibre en information imparfaite

L'assureur connait uniquement le risque moyen $\bar{p}$, il propose ainsi une prime actuarielle moyenne située sur la droite (EF), notée $\pi = \bar{p}Q$.

Or, nous savons que :

$$p_H > \bar{p} \quad et \quad p_L < \bar{p}$$

Ainsi, la prime que les bas risques sont prêts à payer est inférieure à la prime moyenne : $p_L Q < \bar{p}Q$. La prime proposée par l'assureur est trop élevée pour ces assurés. Par ailleurs, si le contrat proposé par l'assureur est α et qu'un autre assureur propose le contrat β, le niveau de bien être des bas risques augmenterait en choisissant le contrat β.

Concernant les hauts risques, la prime qu'ils sont prêts à payer est supérieure à la prime moyenne : $p_H Q > \bar{p}Q$. Le contrat α proposé par l'assureur est intéressant pour ces assurés, contrairement au contrat β qui réduirait leur niveau de satisfaction.

Par conséquent, un contrat tarifé à la prime moyenne aurait pour conséquence d'attirer uniquement les « mauvais » risques.

Rothschild et Stiglitz propose un équilibre séparateur, c'est-à-dire un contrat pour chaque type de risque de façon à ce que les assurés se différencient d'eux-mêmes. Un contrat d'équilibre incitant les hauts risques à ne pas choisir le contrat approprié aux bas risques. La solution serait un contrat avec un niveau de couverture complet pour les hauts risques et un contrat avec un niveau de couverture partiel pour les bas risques, tous deux générant un profit moyen nul pour l'assureur.

Nous ne développerons pas ici les explications aux solutions proposées par les deux auteurs, puisque cela ne constitue pas l'objet du mémoire (pour plus d'informations, cf. [17]).

1.3 Les solutions possibles

Le modèle élémentaire d'antisélection de Rothschild et Stiglitz nous a permis de comprendre l'importance de la prise en compte de ce risque dans la tarification de contrats d'assurance. A présent, il convient d'énumérer les principales solutions proposées dans la littérature :

- **Augmenter le niveau de discrimination dans la tarification**
  La présence d'antisélection est majoritairement liée à l'impossibilité d'obtenir certaines informations sur l'assuré permettant de distinguer son profil de risque. Par exemple, l'assureur ne dispose d'aucune information sur l'état de santé des salariés souhaitant un contrat facultatif. Par ailleurs, les informations sont très limitées dans le cadre d'un contrat collectif contrairement à la santé individuelle. En tout état de cause, la prise en compte d'un grand nombre de variables explicatives dans le modèle de tarification pourrait poser des problèmes de robustesse de données, générant des tarifs non fiables.
- **Différencier les niveaux de garantie des contrats**
  Nous avons vu dans le modèle de Rothschild et Stiglitz qu'en cas de présence d'anti-

sélection, une solution serait de proposer des contrats différenciés pour chaque type de profil. L'individu pourrait révéler son risque indirectement dans le choix de son type de contrat. Nous avons indiqué que dans le cas de présence de deux types de risques, un individu avec un risque élevé serait intéressé rationnellement par un niveau de couverture total d'assurance contrairement à un individu avec un risque faible, qui opterait plutôt pour un contrat avec une couverture partielle. Il s'agit, ainsi de l'autosélection, puisque les individus choisiront chacun les contrats qui leur sont réservé.

Chapitre 2

L'analyse statistique

Après avoir analysé par une approche économique le problème de l'antisélection et évoqué l'importance de la mise en place d'une solution, nous proposons dans cette partie une solution pour prendre en compte ce phénomène dans le cadre de notre tarification.

2.1 La présentation de la méthode d'analyse retenue

2.1.1 La présentation des contrats facultatifs

Dans le cadre de la commercialisation de contrats collectifs santé, les Assurances du Crédit Mutuel proposent à la fois des contrats obligatoires et facultatifs. Le caractère facultatif de l'offre se retrouve à deux niveaux :

– Étant donné que, d'ici 2016, chaque entreprise est contrainte de souscrire un contrat obligatoire santé respectant le panier de soins minimum imposé par le futur décret suite à l'ANI, la souscription d'options peut être proposée aux assurés. En effet, les assurés peuvent augmenter leur niveau de garantie de façon facultative en souscrivant une option supérieure. L'exemple fictif suivant illustre les contrats à options :

	Contrat socle obligatoire	Option 1	Option 2
Soins courants	100% BR-RSS	200% BR-RSS	300% BR-RSS
Optique	100 €par an	300 €par an	500 €par an
Prothèses dentaires	125% BR-RSS	150% BR-RSS	300% BR-RSS
Soins dentaires	125% BR-RSS	150% BR-RSS	300% BR-RSS
Forfait hospitalier	100% BR-RSS	100% BR-RSS	100% BR-RSS
Hospitalisation honoraire	100% BR-RSS	150% BR-RSS	200% BR-RSS

Table 32 – Exemple de contrats à options

– Des contrats collectifs facultatifs peuvent également être proposés sans que l'entreprise ne doive souscrire au préalable un contrat obligatoire. Il s'agit notamment de contrats à destination des membres d'associations, non concernés par l'ANI. Par ailleurs, certains salariés disposant uniquement du panier de soins prévu par le décret (socle minimal) auprès d'un autre assureur vont vouloir améliorer leur couverture ou bénéficier de nouveaux services en ayant recours à une surcomplémentaire.

2.1.2 La description de la méthodologie

Afin de prendre en compte l'antisélection pour les contrats facultatifs, l'objectif est d'appliquer un ou plusieurs coefficient(s) de majoration à la tarification établie pour les contrats obligatoires. Ce(s) coefficient(s) de majoration permettra ainsi de prendre en compte le profil de risque des individus ayant opté pour un contrat facultatif.

La mesure de l'antisélection est une tâche complexe, étant donné qu'il s'agit de modéliser le comportement d'individus et de le différencier de l'aléa moral. Étant donné que la quantité de données relative aux contrats obligatoires facultatifs est très faible et compte tenu de la difficulté de leur différenciation dans nos données, nous avons opté pour une autre solution. Nous avons choisi d'évaluer les écarts de consommation entre un contrat collectif obligatoire et un contrat individuel. En effet, le salarié ne choisit pas volontairement d'adhérer à un contrat collectif obligatoire mis en place par son entreprise, contrairement aux contrats de complémentaire santé individuels où l'individu effectue la démarche d'adhérer. Pour cela, nous disposons de données relatives aux contrats obligatoires et de données relatives aux contrats individuels. L'évaluation des écarts se réalise sur les fréquences de consommation.

Cependant, cette étude ne permet pas de mettre à l'écart la présence d'aléa moral qui se manifeste par un changement de comportement de l'assuré en fonction du niveau des garanties de son contrat. Le fait d'avoir un contrat avec de meilleures garanties incitera l'assuré à consommer davantage d'actes médicaux, contrairement à l'antisélection qui se manifeste avant la souscription du contrat, lorsque l'individu choisit de souscrire volontairement. Par ailleurs, la présence d'antisélection peut être plus ou moins marquée en fonction des niveaux de garanties des contrats. Enfin, l'observation d'un écart de consommation entre contrats collectifs et individuels peut être biaisée, compte tenu des différents niveaux de garanties. Par conséquent, il est plus prudent d'estimer un coefficient de majoration dépendant également du niveau de garantie.

Cette étude a été ainsi réalisée sur un type de produit collectif détenant cinq niveaux de garantie. Ces niveaux de garanties sont également présents pour un type de produit proposé par l'assurance santé individuelle.

Les schémas ci-dessous permettent d'illustrer la distinction de l'antisélection et de l'aléa moral avec la méthode que nous avons sélectionné. Nous avons fait les hypothèses suivantes :

- la fréquence augmente en fonction du niveau de garantie ;
- pour un niveau de garantie donné, la fréquence de consommation d'un contrat collectif obligatoire sera plus faible que la fréquence de consommation d'un contrat individuel.

Ainsi, la longueur des différentes flèches a été choisie de façon arbitraire.

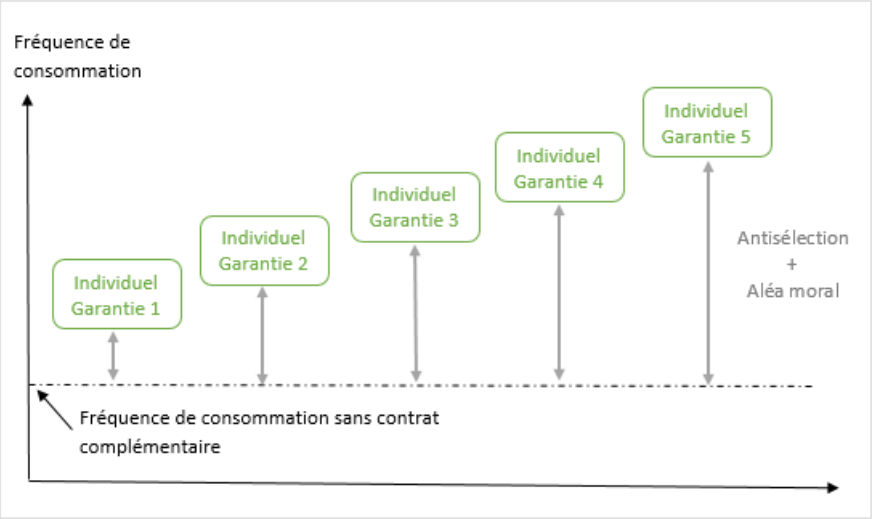


FIGURE 39 – Fréquence de consommation des contrats individuels

Ce schéma permet de montrer la difficulté de différencier l'antisélection de l'aléa moral. En considérant uniquement les contrats individuels, il n'est pas possible de connaître les raisons du choix de consommation de l'assuré.
Est-ce que l'assuré consomme car il considère que son niveau de garantie est avantageux ou est-ce qu'il s'agit uniquement d'un comportement prévu lors de la signature du contrat d'assurance ? Ces phénomènes sont présents uniquement à partir d'une fréquence minimale qui correspond à la fréquence de consommation pour un individu n'ayant pas de contrat de couverture santé.

Le schéma ci-dessous illustre le cas des contrats collectifs :

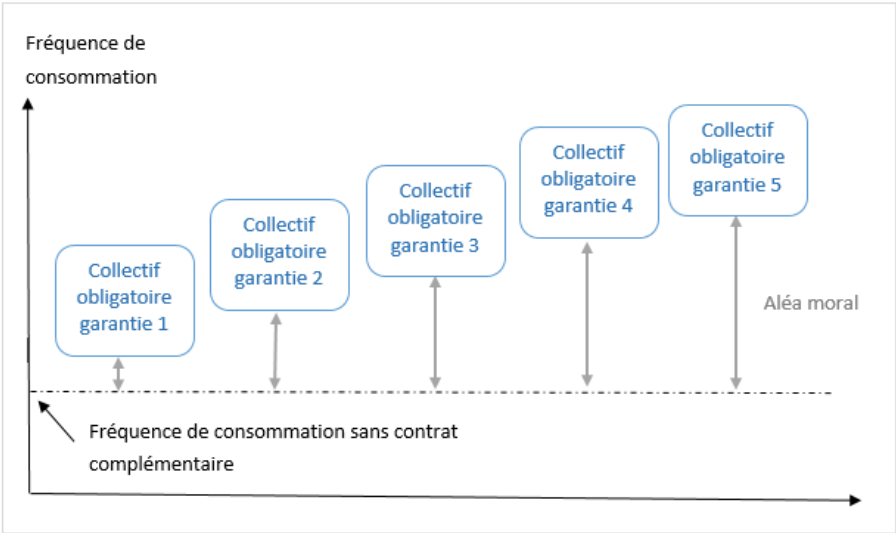


FIGURE 40 – Fréquence de consommation des contrats collectifs

Dans le cas des contrats collectifs, l'écart de consommation entre un individu n'ayant pas de couverture santé et un individu ayant un contrat collectif obligatoire est considéré uniquement comme de l'aléa moral. En effet, puisque les salariés n'ont pas le choix entre souscrire ou non, il ne peut pas y avoir de l'antisélection. Analysons à présent le dernier schéma :

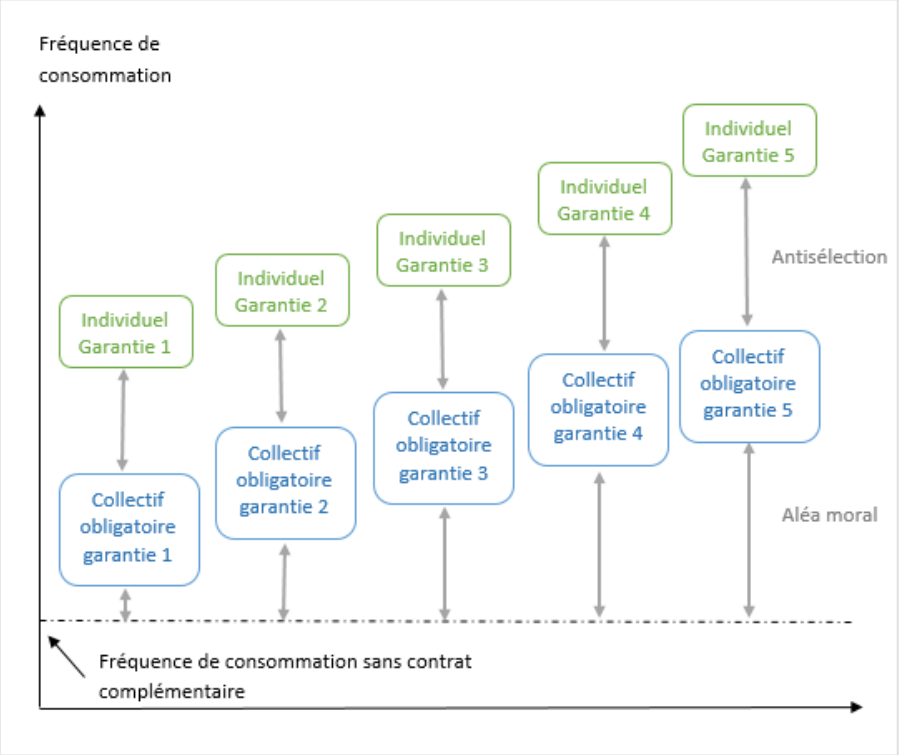


FIGURE 41 – Fréquence de consommation des contrats individuels et collectifs

Le contrat collectif obligatoire permet ainsi de différencier l'aléa moral de l'antisélection, qui sera mesurée dans notre étude comme la différence entre la fréquence de consommation pour un contrat collectif obligatoire et un contrat individuel.

2.1.3 Le traitement des données

La comparaison entre un contrat collectif obligatoire et un contrat individuel doit s'effectuer avec vigilance. Il s'agit de deux populations différentes en termes de démographie.

Dans un premier temps, il convient de vérifier la quantité de nos données. En effet, les données ont été réduites suite au choix d'un unique type de contrat, proposant des niveaux de garanties (1 à 5) identiques en collectif et en individuel. Concernant les données collectives, la quantité de contrat de niveau de garantie 1 est très faible. Ainsi, nous décidons de ne pas tenir compte des contrats collectifs de niveau de garantie 1. Par conséquent, il est également nécessaire de supprimer les données individuelles de niveau de garantie 1. Pour les autres niveaux de garanties, il semblerait que nous disposons de données suffisantes, hormis le niveau de garantie 2 en collectif pour lequel il faut être prudent dans l'interpré-

tation des résultats.

Dans un deuxième temps, il est nécessaire de tenir compte de la répartition par âge au sein des contrats collectifs et individuels. En effet, pour les contrats collectifs où les assurés sont des salariés, nous disposons de données jusqu'à environ 70 ans, contrairement aux données individuelles, où nous observons des assurés excédant l'âge de 100 ans. Dans les parties précédentes, nous avons observé une forte hausse de la fréquence de consommation sur les « grands » âges. Ainsi, la présence de ces « grands » âges expliquerait partiellement le niveau de fréquence élevé en individuel, et biaiserait fortement le coefficient d'antisélection. Par conséquent, il est nécessaire de sélectionner uniquement les assurés ayant un âge inférieur à 70 ans.

2.2 L'étude de la démographie par type de contrat

L'objectif étant de déterminer un coefficient d'antisélection par niveau de garantie, il est primordial d'effectuer au préalable une analyse statistique sur la répartition des données en fonction des différentes variables explicatives dont nous disposons.

2.2.1 L'ACP sur la démographie

Dans les parties précédentes, une présentation des différentes variables disponibles pour analyser le risque de consommation de frais de santé a été effectuée. Nous avons vu, par l'intermédiaire du modèle linéaire généralisé que ces variables peuvent impacter très significativement la fréquence de consommation. Or, si la répartition des contrats n'est pas proportionnelle pour les contrats individuels et collectifs par rapport aux différentes variables, cela peut biaiser la valeur de la fréquence.

Par conséquent pour analyser la composition des contrats collectifs et individuels avec différents niveaux de garantie, une analyse en composantes principales peut être intéressante.

L'ACP a été réalisée sur huit individus qui représentent la nature du contrat associé à son niveau de garantie :

- contrat individuel de niveau de garantie 2 ;
- contrat individuel de niveau de garantie 3 ;
- contrat individuel de niveau de garantie 4 ;
- contrat individuel de niveau de garantie 5 ;
- contrat collectif obligatoire de niveau de garantie 2 ;
- contrat collectif obligatoire de niveau de garantie 3 ;
- contrat collectif obligatoire de niveau de garantie 4 ;
- contrat collectif obligatoire de niveau de garantie 5.

Nous utiliserons par la suite, la notion de "type de contrat" qui regroupe la nature du contrat et son niveau de garantie.

Les variables et modalités considérées dans l'étude sont les suivantes :

- proportion d'hommes et de femmes ;

– proportion d'adhérent au régime général (noté gen) et local (noté loc) ;
– proportion de personnes avec un âge compris entre 0 et 10, 11 et 20, etc (noté age10, age20, etc.) ;
– proportion de personnes habitant dans les classes de départements créées dans la deuxième partie du mémoire (noté dpt1, dpt2, etc.).

L'objectif recherché avec cette ACP est ainsi de distinguer les variables et modalités représentatives des différents types de contrat.

Le choix du nombre d'axes

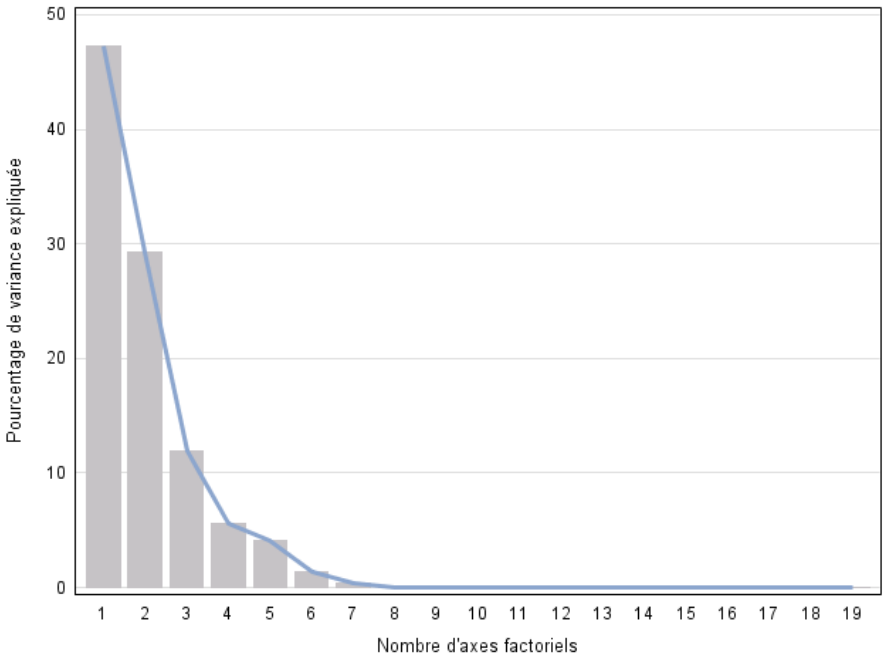


FIGURE 42 – Choix du nombre d'axes factoriels

Ce graphique représente le pourcentage de variance expliquée en fonction du nombre d'axes. Le nombre d'axes optimal correspond au nombre avant la « cassure ». Dans ce cas, il convient ainsi de retenir deux axes, qui représentent une grande partie de l'information (76,57%).

Le graphique des individus et des variables

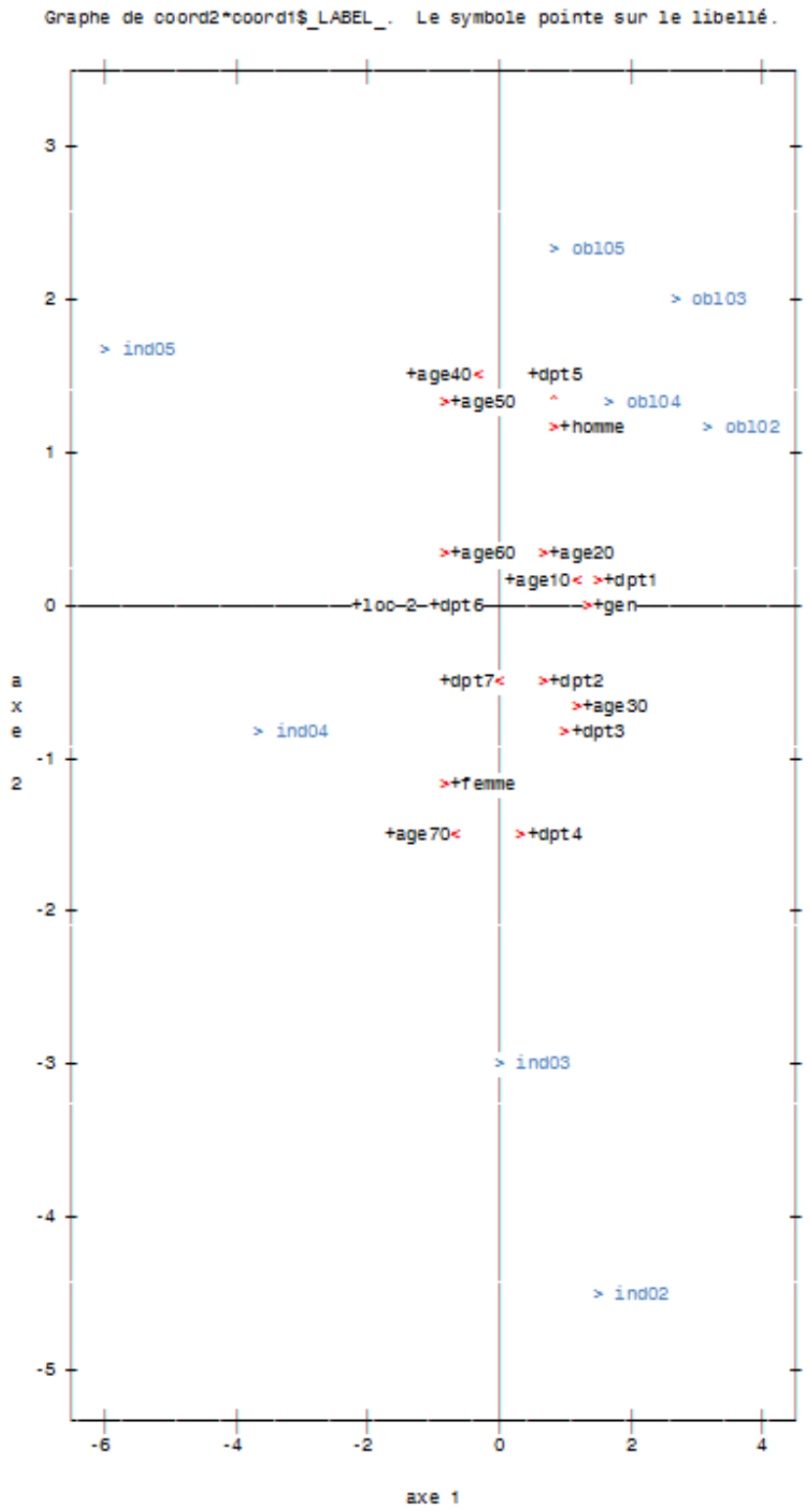


FIGURE 43 – Graphique des individus et des variables

L'analyse du graphique des individus et des variables permet de fournir les résultats suivants :

- un individu du graphique prend des valeurs plus élevées que la moyenne pour les variables allant dans sa direction ;
- un individu du graphique prend des valeurs moins élevées que la moyenne pour les variables allant en direction opposée.

Or, pour pouvoir utiliser cette règle de lecture, les variables doivent être bien représentées sur le graphique, c'est-à-dire proche des bords du cercle de corrélation (figure 43). Les variables qui ne pourront pas être interprétées sont : le département n°7, n°2, et les personnes dont l'âge est compris entre 10 et 20 ans et 50 et 60 ans.

Les résultats suivants ont été relevés du graphique :

- les hommes sont en proportion plus élevée parmi les différents contrats obligatoires que les femmes et les femmes sont plus présentes que les hommes en individuel.
- les adhérents du régime local sont plus représentatifs des contrats individuel ayant un haut niveau de garantie. Ce constat avait déjà été fait dans l'analyse de la démographie (partie 2).
- la variable « âge » ne semble pas donner des résultats pertinents. Nous pouvons uniquement constater que les personnes âgées de 60 ans ou plus sont plus présentes dans les contrats individuels, ce qui est lié aux départs en retraite des salariés ayant des contrats collectifs.
- les classes de départements 3 et 4 sont plus représentatives des contrats individuels de niveau de garantie 2 et 3.

La répartition des contrats par sexe et par âge nécessite une analyse plus précise puisque ces variables ont beaucoup d'influence sur la fréquence de consommation.

Analysons à présent la variable « sexe », qui n'est clairement pas distribuée de façon similaire, d'après l'ACP, entre les contrats obligatoires et individuels.

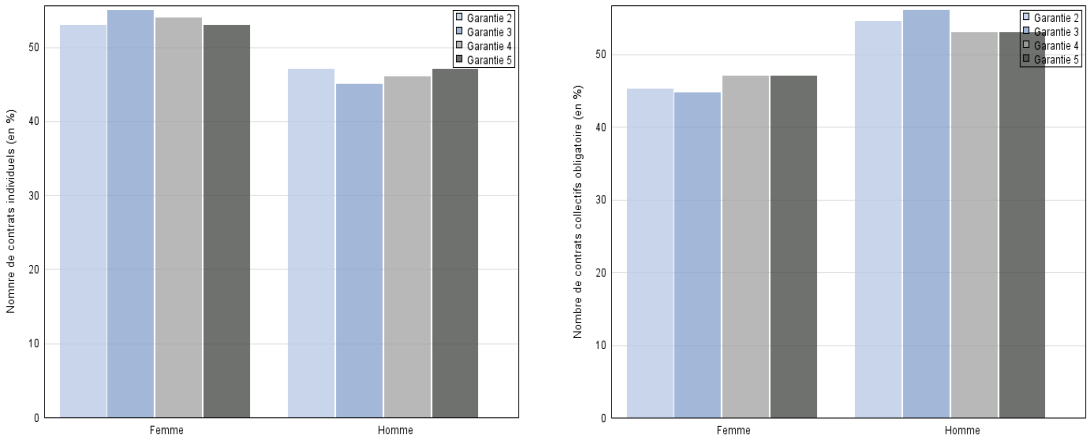


FIGURE 44 – Répartition des assurés par type de contrat

Le graphique de gauche représente la répartition des assurés d'un contrat individuel par niveau de garantie en fonction du sexe. Par exemple, parmi les contrats individuels de niveau de garantie 2, 53% sont des femmes et 47% sont des hommes.

La répartition par sexe semble être semblable entre les différents niveaux de garantie d'un contrat de même nature. Cependant, lorsque nous comparons les deux types de contrat : les femmes sont plus présentes que les hommes en individuel, et moins présentes en obligatoire. Les écarts ne semblent toutefois pas être majeurs.

La répartition des contrats par classes de département pour chaque type de contrat est représentée par les graphiques ci-dessous :

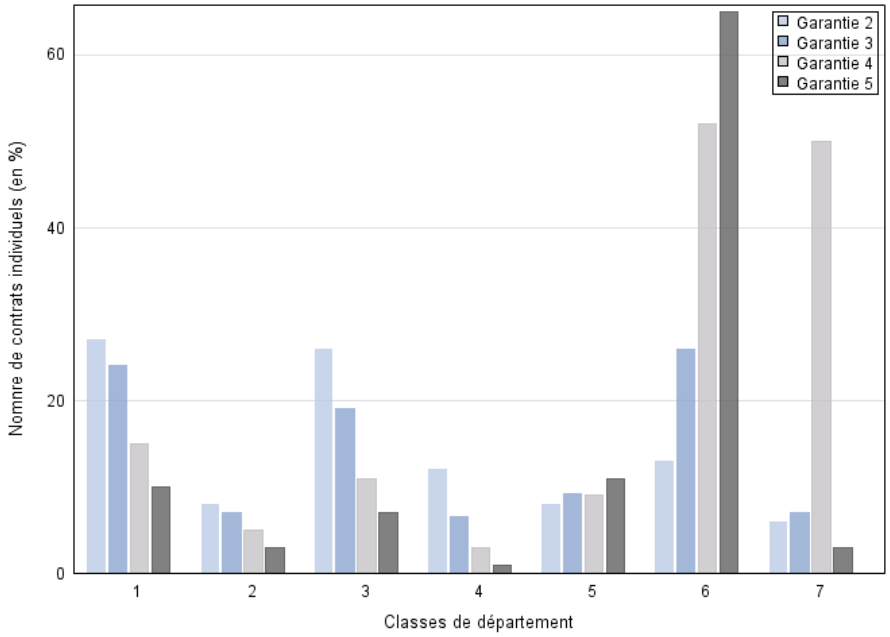


FIGURE 45 – Répartition des contrats individuels en fonction des classes de département

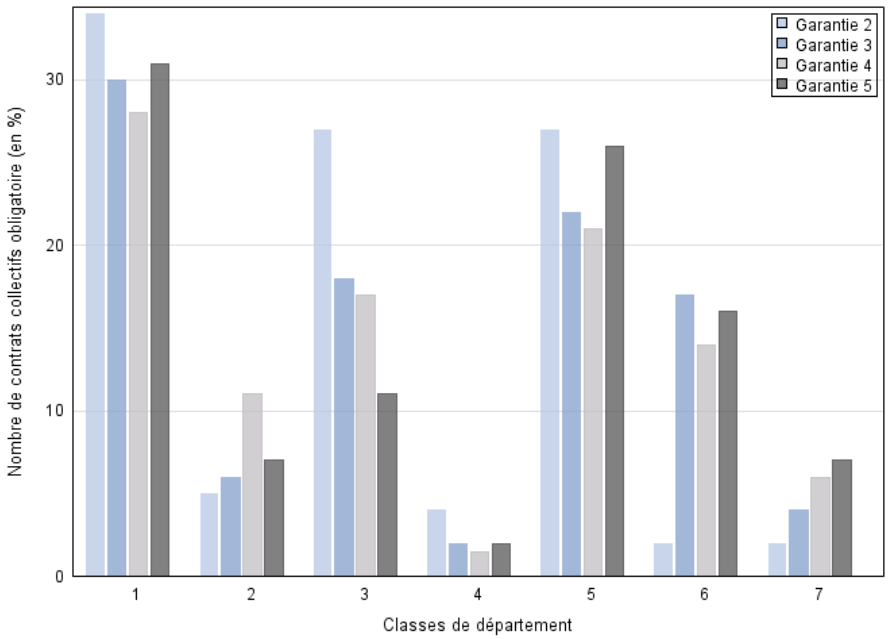


FIGURE 46 – Répartition des contrats collectifs obligatoire en fonction des classes de département

Concernant les départements, les différences sont plus visibles. Nous observons surtout une différence de répartition entre les contrats individuels et collectifs pour les classes de département 1, 5 et 6.

2.2.2 L'analyse univariée plus précise sur l'âge

Le facteur le plus influent sur la fréquence de consommation en assurance santé est l'âge. Ainsi, il est nécessaire d'analyser la répartition par âge des individus des différents contrats. En effet, il faudrait que cette répartition soit identique pour les différents contrats étudiés afin de calculer une fréquence par type de contrat non perturbée par la structure des âges.

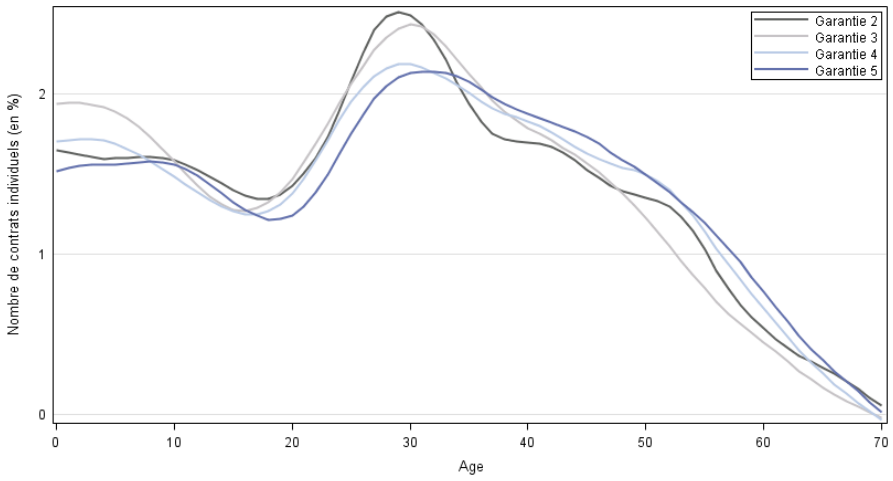


FIGURE 47 – Répartition des contrats individuels en fonction de l'âge

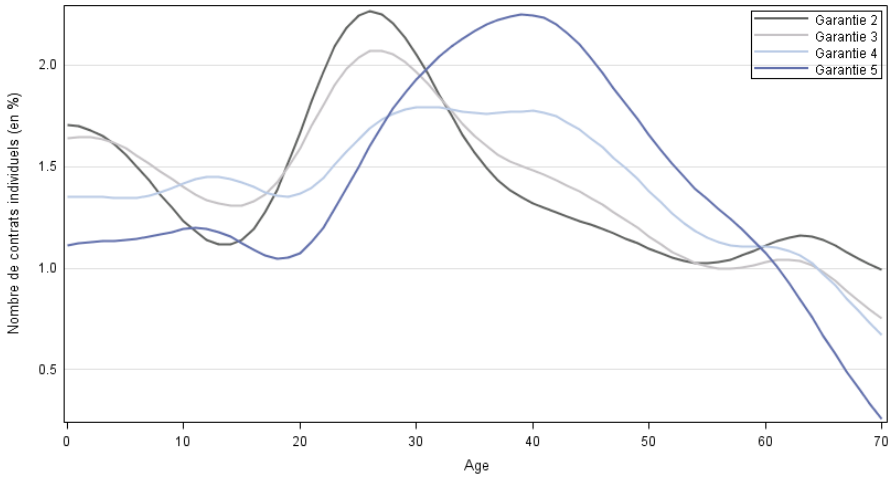


FIGURE 48 – Répartition des contrats collectifs obligatoire en fonction de l'âge

Nous déduisons les résultats suivants des graphiques ci-dessus :

- Pour les contrats individuels facultatifs, nous observons globalement une répartition plus forte de contrats avec de faibles niveaux de garantie avant l'âge de 35 ans (40 ans en obligatoire). Après cet âge, l'ordre des courbes est inversé, la part de contrats avec des hauts niveaux de garantie est plus élevée que les contrats avec de faibles niveaux de garantie.
- En collectif, la proportion d'enfants est plus élevée qu'en individuel.
- En individuel, la proportion de « grands » âges est plus élevée qu'en collectif. Sachant que la fréquence de consommation est relativement forte à ces âges pour l'ensemble des postes de garantie, il est nécessaire de prendre en compte cette différence de répartition dans l'estimation du coefficient d'antisélection global.

2.3 La vérification de l'existence du phénomène d'antisélection

Dans cette section, nous souhaitons mettre en évidence la présence d'antisélection entre les contrats facultatifs et obligatoires avec un test d'analyse de la variance.

2.3.1 La présentation du modèle ANOVA

La définition du modèle

Le modèle ANOVA (Analysis of variance) permet de tester l'effet d'une ou plusieurs variables qualitatives appelés facteurs sur une variable quantitative observée appelée réponse. Ce test est fondé sur la comparaison des moyennes de plusieurs groupes constitués par les modalités des différents facteurs étudiés.

Le modèle ANOVA où nous considérons un unique facteur α_j avec Y_ij la réponse est le suivant :

$$Y_ij = \mu + \alpha_j + \epsilon_{ij} \quad i = 1, \dots, n_j \; j = 1, \dots, J$$

où j est l'indice du niveau du facteur étudié et n_j le nombre de répétitions pour une modalité j. Le paramètre μ correspond à l'effet moyen, α_j l'effet du facteur étudié et ϵ_{ij} représente les erreurs du modèle.

L'analyse de variance est fondée sur l'équation de décomposition de la variation totale des données :

$$Variation\ totale = Variation\ intergroupe + Variation\ intragroupe.$$

La variation intergroupe traduit l'effet du facteur. En effet, elle correspond aux écarts entre les moyennes de chaque groupe et la moyenne générale. Les écarts entre chaque observation et la moyenne du groupe est détenue dans la variation intragroupe, que nous nommons également variation résiduelle. Cette décomposition sera utilisée dans la construction de test pour analyser l'influence du facteur.

Ces variations sont calculées à partir de la relation suivante de la somme des carrés des écarts :

$$SCT = SCE + SCR$$
$$\sum_{i,j}(y_{ij} - \bar{y})^2 = \sum_i n_j(\bar{y}_j - \bar{y})^2 + \sum_{i,j}(y_{ij} - \bar{y}_j)^2.$$

Ainsi, avec la prise en compte des degrés de liberté de chaque composante, nous obtenons les carrés moyens CM associés suivants :

$$CMtotale = \frac{SCT}{n-1}$$

$$CMexpliquée = \frac{SCE}{J-1}$$
$$CMrésiduelle = \frac{SCR}{n-J}$$

où $n = \sum_{j=1}^{J} n_j$.

Le test de Fisher

L'analyse de la variance repose sur le test de Fisher où nous testons les hypothèses suivantes :

H_0 : l'effet de tous les niveaux du facteur est identique

$$\alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$$

H_1 : au moins un niveau de facteur a un effet différent des autres

$$\exists i \in \{1, \ldots, n_j\} \text{ où } \alpha_i \neq 0.$$

La statistique de test est la suivante :

$$F = \frac{CMexpliquée}{CMrésiduelle} = \frac{SCE}{SCR}\frac{n-J}{J-1}$$

Il convient de la comparer au quantile d'ordre α d'une loi de Fisher à $(J-1)$ et $(n-J)$ degrés de liberté. Ainsi, nous rejetons l'hypothèse H_0 dès lors que $F > f_{J-1,n-J}^{1-\alpha}$ ou dès lors que $P(F_{J-1,n-J} > F) < \alpha$.

Cependant, l'utilisation de ce test requiert au préalable la vérification des conditions suivantes :

- les erreurs ϵ_{ij} doivent être indépendantes ;
- ϵ_{ij} doit suivre une loi normale. Cette hypothèse ne doit pas être nécessairement vérifiée lorsque la population étudiée est assez grande ;
- les variances des différents erreurs doivent être égales (hypothèse d'homoscédasticité).

Plusieurs tests permettent de vérifier l'hypothèse d'homoscédasticité tel que les tests paramétriques de Bartlett, de Hartley ou de Levene. Les tests de Bartlett et Hartley, contrairement au test de Levene, sont très sensibles à l'hypothèse de normalité de la variable observée. Ainsi, le test de Levene est plus adapté dans le cas où les données ne seraient pas distribuées selon une loi normale.

Les tests d'homogénéités consistent à vérifier l'égalité des variances pour chaque niveau du facteur :

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_J^2$$
$$H_1 : \exists k \neq j,\ \sigma_j^2 \neq \sigma_k^2$$

Le rejet de H_0 signifie que l'hypothèse d'homoscédasticité des erreurs n'est pas vérifiée et que le test ANOVA ne serait pas applicable dans ce cas. La solution serait de recourir à une transformation des données, à un test non paramétrique ou à un test paramétrique hétéroscédastique. Compte tenu de la puissance des tests paramétriques par rapport aux tests non paramétriques, nous choisissons de réaliser un test paramétrique hétéroscédastique.

Le test de Welch

Pour contourner le problème de normalité et d'égalité des variances des données étudiées, nous pouvons utiliser un test ne supposant pas l'égalité des variances afin d'analyser l'effet d'un facteur. Parmi ces tests, nous retrouvons le test de Welch qui est considéré comme une alternative au test de Fisher.

Le test de Welch consiste à tester les mêmes hypothèses que le test de Fisher sur deux groupes :

H_0 : l'effet des deux niveaux du facteur est identique ;
H_1 : l'effet des deux niveaux du facteur est différent.

Soit μ_1 et μ_2 les moyennes respectives des groupes 1 et 2 constitués par les deux niveaux du facteur étudié, σ_1 et σ_2 leurs variances respectives, et n_1 et n_2 les tailles respectives. La statistique de test est la suivante :

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Sous H_0, T va suivre une loi de Student avec ν degré de liberté :

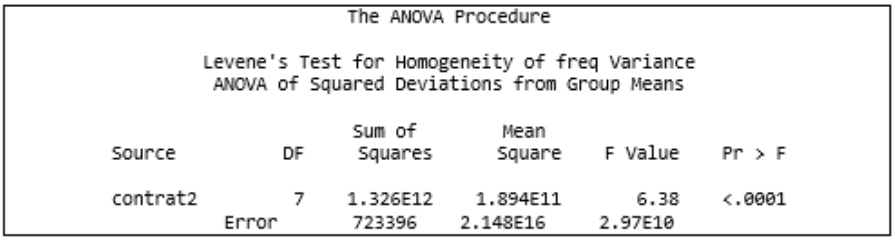
$$\nu = \frac{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^2}{\frac{{\sigma_1^2}^2}{n_1^2(n_1-1)} + \frac{{\sigma_2^2}^2}{n_2^2(n_2-1)}}.$$

2.3.2 L'application à l'étude de l'antisélection

Dans cette sous-section, nous souhaitons mettre en évidence la présence d'antisélection dans nos données. Pour cela, nous décidons au préalable d'appliquer un test ANOVA sur la fréquence de consommation avec comme unique facteur, le type de contrat (c'est-à-dire individuel de niveau 1, individuel de niveau 2, etc).

Le test ANOVA peut être utilisé sans que l'hypothèse de normalité soit vérifiée si l'échantillon de données est suffisamment grand, ce qui est le cas dans nos données. Nous privilégions l'utilisation de tests paramétriques étant donné qu'ils sont plus puissants et plus efficaces.

Par ailleurs, nous supposons l'hypothèse d'indépendance vérifiée et analysons l'hypothèse d'égalité des variances par le test de Levene fourni par la procédure « GLM » de SAS. Rappelons que la statistique de test est identique à la statistique de test d'une analyse de la variance classique sur la variable transformée suivante : $|y_{ij} - y_j|$.



The ANOVA Procedure

Levene's Test for Homogeneity of freq Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
contrat2	7	1.326E12	1.894E11	6.38	<.0001
Error		723396	2.148E16	2.97E10	

FIGURE 49 – Résultats du test de Levene

Compte tenu de la significativité du test (p-valeur <5%), nous rejetons l'hypothèse H_0 d'égalité des variances des différents modèles, et par conséquent nous réalisons une erreur de premier espèce. Le modèle ANOVA n'est donc pas applicable à nos données. Il est préférable d'utiliser un test hétéroscédastique. Ainsi, nous utilisons plusieurs tests de Welch sous SAS, avec la procédure « GLM » :

- test 1 : fichier de données contenant uniquement les contrats de niveau de garantie 2 ;
- test 2 : fichier de données contenant uniquement les contrats de niveau de garantie 3 ;
- test 3 : fichier de données contenant uniquement les contrats de niveau de garantie 4 ;
- test 4 : fichier de données contenant uniquement les contrats de niveau de garantie 5.

Nous obtenons les résultats suivants :

Welch's ANOVA (garantie 2)			
Source	**DDL**	**Valeur F**	**Pr > F**
contrat2	1.0000	60.50	<.0001
Error	311.8		

TABLE 33 – Test de Welch : niveau de garantie 2

Welch's ANOVA (garantie 3)			
Source	**DDL**	**Valeur F**	**Pr > F**
contrat2	1.0000	108.53	<.0001
Error	1941.3		

TABLE 34 – Test de Welch : niveau de garantie 3

Welch's ANOVA (garantie 4)			
Source	**DDL**	**Valeur F**	**Pr > F**
contrat2	1.0000	155.75	<.0001
Error	2171.9		

Table 35 – Test de Welch : niveau de garantie 4

Welch's ANOVA (garantie 5)			
Source	**DDL**	**Valeur F**	**Pr > F**
contrat2	1.0000	74.20	<.0001
Error	6671.5		

Table 36 – Test de Welch : niveau de garantie 5

La p-value de ces tests pour chaque niveau de garantie étant inférieure au seuil de 5%, nous pouvons rejeter l'hypothèse d'absence d'antisélection indépendamment de l'aléa moral. Ainsi nous réalisons également une erreur de première espèce.

2.4 La mesure de l'antisélection

Nous avons montré dans la section précédente qu'indépendamment de l'aléa moral, il existe un écart de fréquence entre les contrats individuels et collectifs, considéré comme de l'antisélection. Nous proposons à présent, dans cette section, une mesure de ce risque par poste de garantie et tous postes de garantie confondus.

2.4.1 La normalisation des données

Nous souhaitons estimer la fréquence de consommation pour chaque niveau de garantie pour les contrats individuels et collectifs. Or, le calcul de la fréquence sur ces populations peut être fortement biaisé en raison de la structure par âge, par sexe et par département de ces différentes populations. En effet, ces fréquences pourront être comparables uniquement si la structure des échantillons étudiés est similaire pour des facteurs ayant une influence sur la fréquence de consommation.

La solution serait d'estimer la fréquence en fonction des différents facteurs influençant le comportement de consommation en santé, en calculant par exemple un coefficient d'antisélection fixé en fonction de l'âge, du sexe et du département de l'assuré. Une autre solution serait de corriger l'effet des facteurs sur la fréquence de consommation. Il s'agit d'une technique de standardisation des données. Les méthodes les plus utilisées sont : la standardisation directe et la standardisation indirecte.

Afin de déterminer un coefficient d'antisélection indépendamment de l'âge de l'assuré, nous appliquons la méthode de standardisation indirecte à nos données. Il est à noter qu'il serait nécessaire a priori de corriger également l'effet du sexe et du département, mais cela poserait un problème de fiabilité du coefficient, suite à un manque de données.

Pour cela, il est d'abord nécessaire de calculer la fréquence moyenne en fonction du type de contrat et de l'âge, que nous notons f_{ij} avec i la $i^{\text{ème}}$ modalité de la variable âge et j la $j^{\text{ième}}$ modalité de la variable type de contrat.

Rappelons que le type de contrat est une variable constitué de la nature du contrat et du niveau de garantie.

Nous notons :

- n_{ij} : le nombre d'assurés (en années risques) ayant un âge i et un type de contrat de j ;
- n_i : le nombre d'assurés (en années risques) ayant un âge i pour l'ensemble des contrats ;
- n_j : le nombre d'assurés (en années risques) ayant un type de contrat j dans l'ensemble de la population ;

Pour appliquer la technique de standardisation indirecte [24] des données, calculons d'abord les éléments suivants :

- La fréquence moyenne par âge pondérée par le nombre de contrats :

$$\bar{f_i} = \frac{1}{n_i} \sum j n_{ij} f_{ij}.$$

- La fréquence moyenne par type de contrat pondérée par le nombre de contrats :

$$\bar{f_j} = \frac{1}{n_j} \sum_i n_{ij} f_{ij}.$$

- La fréquence moyenne théorique par type de contrat : calculée à partir de la fréquence moyenne par âge :

$$\bar{f_j^*} = \frac{1}{n_j} \sum_i n_{ij} \bar{f_i}.$$

Dans ce cas, au lieu de calculer une fréquence moyenne en fonction de la fréquence de chaque case, nous réalisons une moyenne pondérée sur une fréquence moyenne par âge calculée sur l'ensemble des contrats. Cette moyenne pondérée prend en compte la structure par âge du type de contrat j.

A partir des fréquences moyennes théoriques et réelles, nous calculons un indice par type de contrat j :

$$ind_j = \frac{\bar{f_j}}{\bar{f_j^*}}.$$

24. Méthode reprise de Lemel et Villeneuve « Les consommations médicales des Français », Les Collections de l'INSEE, 1977.

Cet indice permet d'apprécier l'effet de la fréquence d'un type de contrat par rapport à la fréquence moyenne de la population. Ainsi, un indice de 0,8 signifierait que la fréquence de ce type de contrat représente 80% de la fréquence moyenne de la population.

2.4.2 Les résultats

Pour des raisons de confidentialité, les valeurs exactes des coefficients obtenus ne seront pas fournies.

Après avoir corrigé les données par la structure par âge de la population, nous obtenons les fréquences suivantes pour les contrats obligatoires et individuels :

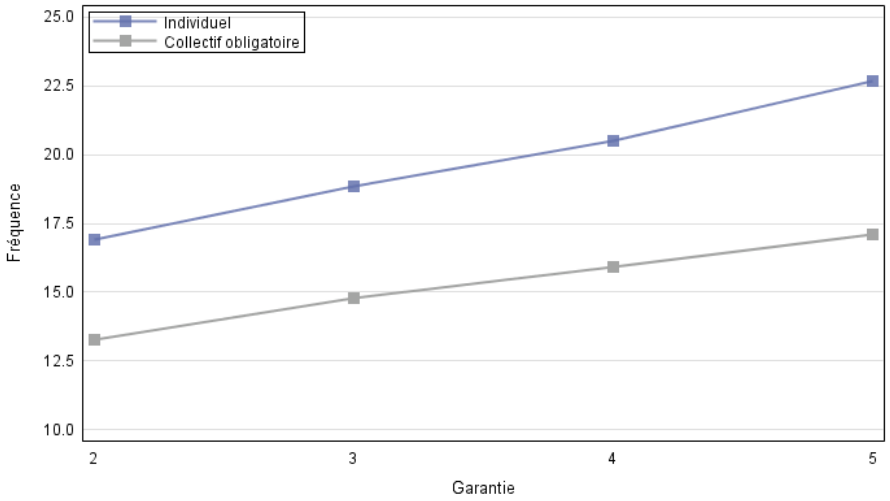


FIGURE 50 – Variation du niveau d'antisélection en fonction du niveau de garantie

Ce graphique met en évidence la présence d'antisélection, compte tenu de l'écart de fréquence de consommation entre les contrats individuels et collectifs. Par ailleurs, le phénomène d'aléa moral peut également être observé entre les différents niveaux de garantie des contrats collectifs.

Le coefficient d'antisélection α correspond ainsi à :

$$\alpha = \frac{Fréquence\ moyenne\ individuelle}{Fréquence\ moyenne\ collective} - 1$$

Le graphique précédent permet de montrer que le coefficient d'antisélection augmente en fonction du niveau de garantie. Cette augmentation est plus marquée entre les niveaux de garantie 4 et 5. Ces différents coefficients semblent toutefois assez proches pour les niveaux de garantie 2, 3 et 4.

L'impact de la normalisation des données est représenté sur le graphique ci-dessous.

Niveaux de garantie	2 à 3	3 à 4	4 à 5
Variation du coefficient de majoration	2%	4%	12%

TABLE 37 – Variation du coefficient de majoration en fonction du niveau de garantie

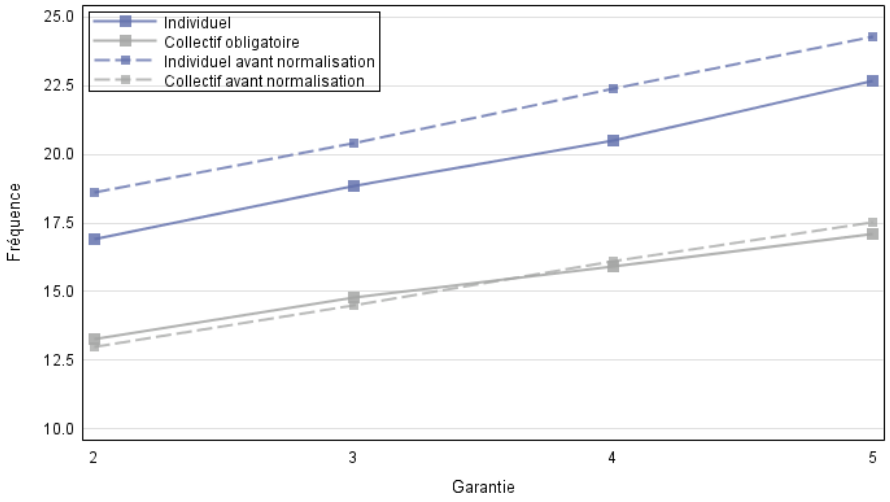


FIGURE 51 – Impact de la standardisation des données

Le graphique ci-dessus montre que la standardisation des données par âge rapproche les deux courbes et réduit ainsi le coefficient d'antisélection. La structure par âge des contrats individuels accordant un poids important aux « grands » âges considérés comme de « grands » consommateurs avait pour effet de surestimer les fréquences.

2.5 La mesure de l'antisélection par postes de garantie

Nous nous intéressons également à l'analyse de l'antisélection de certains grands postes de garantie.

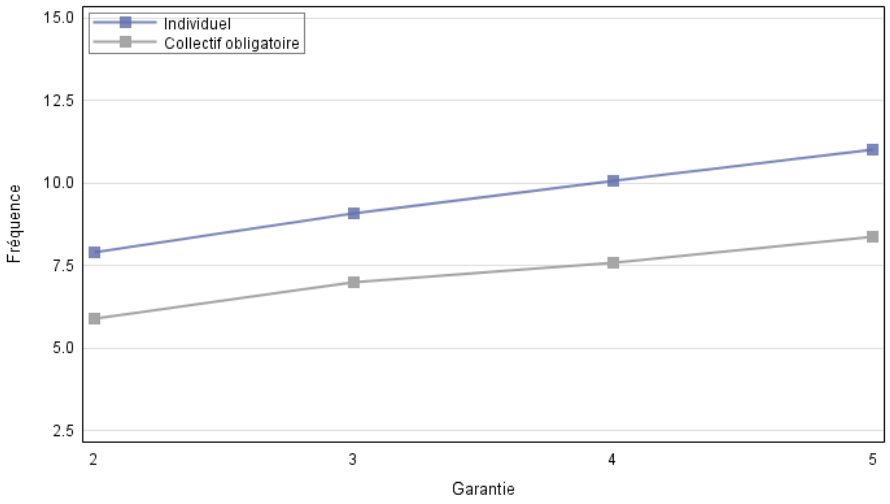


FIGURE 52 – Analyse de l'antisélection pour les soins courants

Le graphique ci-dessus représente la fréquence de consommation moyenne pour les contrats individuels et collectifs obligatoires concernant les actes de soins courants hors actes de pharmacie. Par conséquent, nous constatons par analyse graphique que le phénomène d'antisélection est significativement présent pour les actes de soins courants. Cet écart est majoritairement lié aux actes de transport et aux visites et consultations de spécialistes.

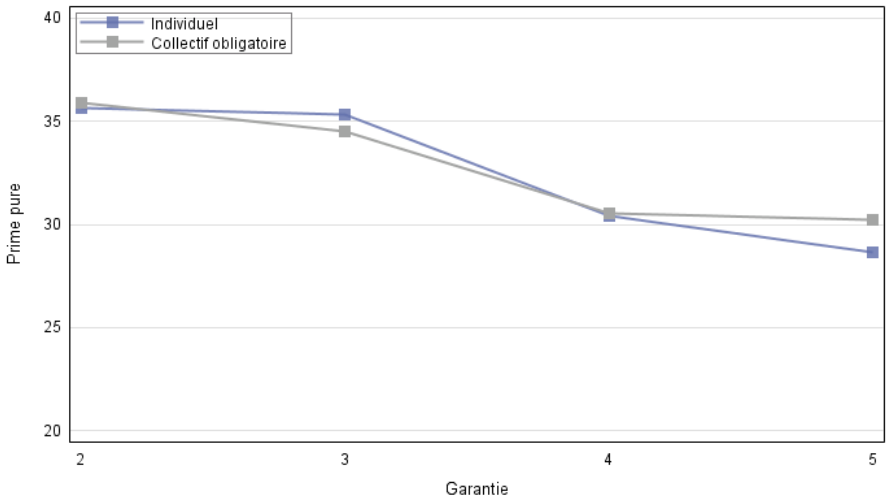


FIGURE 53 – Analyse de l'antisélection pour la pharmacie

Concernant la pharmacie, l'analyse de l'antisélection est réalisée sur la prime pure. Nous observons l'écart entre la prime pure d'un contrat collectif obligatoire et un contrat individuel puisque dans le cas des dépenses de médicaments, la notion de fréquence ne

parait pas fiable. En effet, sur une même ordonnance, plusieurs médicaments peuvent être prescrits, et peuvent être comptabilisés pour un unique acte ou plusieurs actes.

Le phénomène d'antisélection semble être inexistant dans le cas de la pharmacie, puisque l'écart entre les primes est très faible. Nous constatons cependant, un petit écart pour les contrats de niveaux de garantie 5.

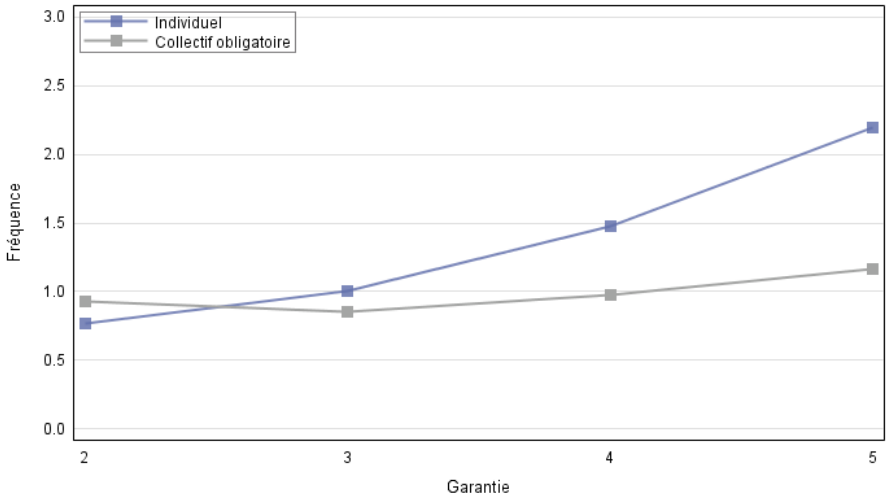


FIGURE 54 – Analyse de l'antisélection pour le dentaire

En dentaire, le phénomène d'antisélection semble être très présent. Cet effet parait cohérent puisque cette catégorie d'acte est notamment composée des prothèses dentaires non remboursées par la Sécurité sociale. Un assuré souhaitant consommer ce type d'acte sera très incité à souscrire un contrat. Nous constatons également que les écarts augmentent en fonction des niveaux de garanties. Compte tenu des garanties très faibles pour le dentaire concernant le niveau de garantie 2, la fréquence moyenne de consommation des assurés d'un contrat obligatoire collectif est plus élevée que celle des assurés d'un contrat facultatif. En effet, l'assuré ne sera pas incité à souscrire un contrat de complémentaire santé individuel uniquement pour le remboursement de ses prothèses dentaires puisque le niveau y est faible.

L'évolution du phénomène d'aléa moral par niveaux de garantie est peu marquée dans le cas du dentaire, puisque la fréquence de consommation pour les contrats obligatoires est presque constante selon les différents niveaux de garantie.

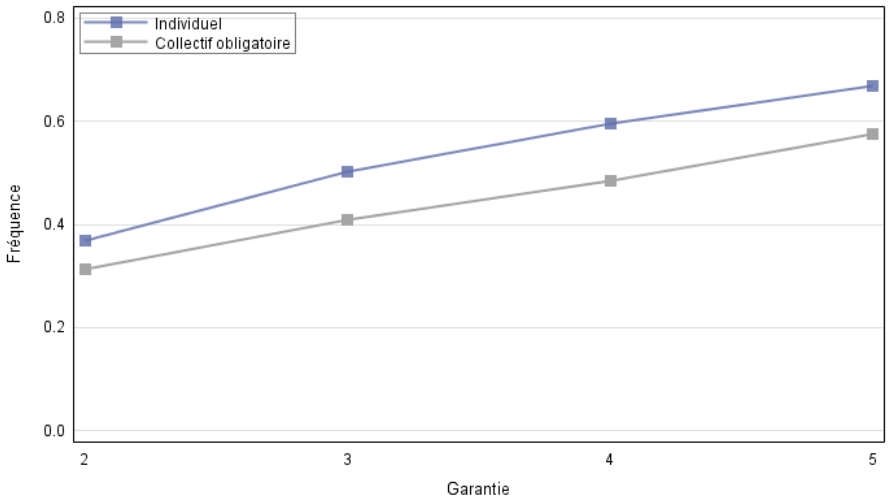


FIGURE 55 – Analyse de l'antisélection pour l'optique

L'optique semble être un poste de garantie où le phénomène d'antisélection est plus faible que le dentaire et les soins courants. Cependant, contrairement au dentaire, l'antisélection ne dépend pas du niveau de garantie. Le coefficient semble être assez stable.

Concernant l'aléa moral, il est significativement présent en optique, puisque la fréquence de consommation augmente en fonction du niveau de garantie dans le cadre d'un contrat collectif.

2.6 La mesure de l'antisélection en fonction de l'âge

Compte tenu de la forte influence de l'âge sur la fréquence de consommation en santé, nous proposons dans cette partie un coefficient d'antisélection en fonction de l'âge, en plus du type du contrat et du niveau de garantie. Les âges ont été regroupés par classes en fonction de la courbe de consommation globale en santé afin de rendre l'analyse plus robuste.

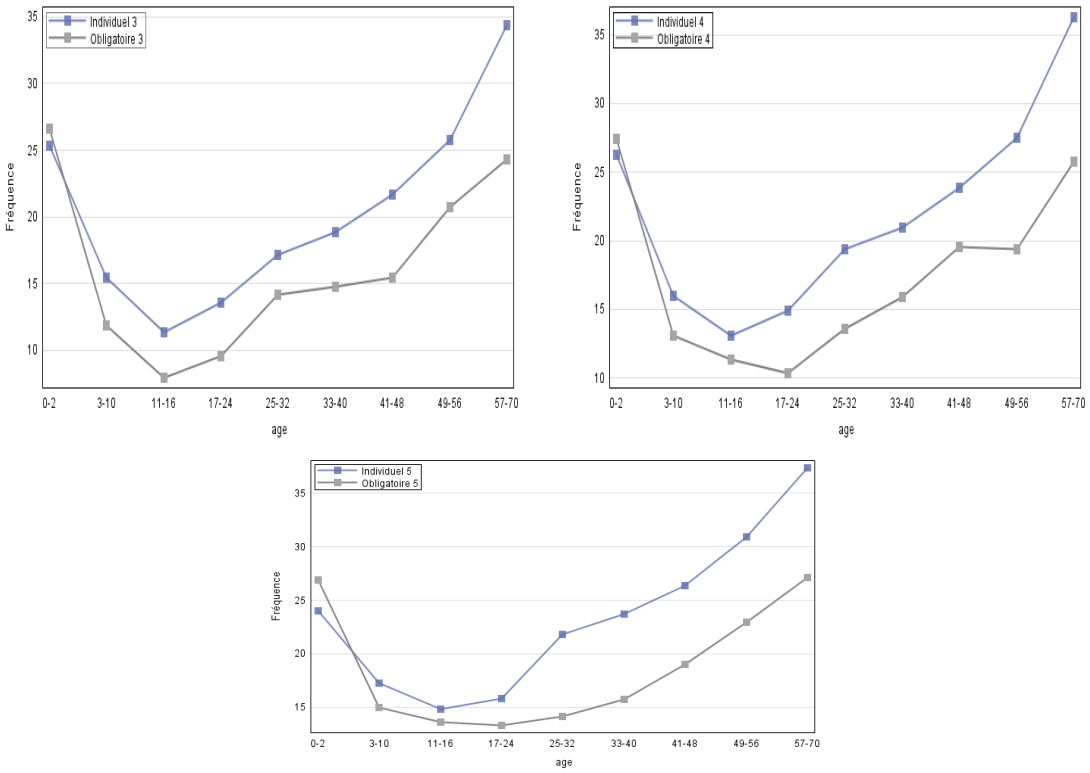


FIGURE 56 – Variation du comportement antisélectif en fonction de l'âge

Globalement, nous constatons en premier lieu que l'effet de l'âge ne dépend pas du niveau de garantie. En effet, l'écart entre les deux courbes semble être assez similaire pour les trois graphiques. Il est évident que pour les nouveau-nés, il n'y a pas d'antisélection. Nous remarquons que le phénomène d'antisélection apparaît de façon significative vers l'âge de 17 ans et augmente en fonction de l'âge pour atteindre son maximum entre 57 et 70 ans.

Conclusion générale

Une étude préalable du portefeuille nous a permis de sélectionner les facteurs impactant la sinistralité en assurance santé : l'âge, le régime d'adhésion, le sexe, le niveau de garantie du contrat, le département, le type de bénéficiaire et le nombre d'enfants par adhérent. Nous avons porté une attention particulière à l'impact du département sur la sinistralité observée. Compte tenu du nombre important de modalités de cette variable, nous avons eu recours à une CAH qui a permis de regrouper les départements en sept classes selon la consommation observée sur les différents postes de garantie.

Ensuite, après avoir rappelé les fondements théoriques des GLM, nous l'avons appliqué à la modélisation de la fréquence de consommation et du coût moyen.
La modélisation de la fréquence avec le GLM, basé sur la modélisation du nombre d'actes consommé, a nécessité l'utilisation de lois de probabilités discrètes. Étant donné la présence d'une hétérogénéité non observable parmi nos données et la forte présence d'assuré n'ayant pas consommé pour certaines familles d'actes, des modèles « modifiées en zéro » ont été utilisées en plus des lois usuelles. Pour les deux sous-catégories d'actes considérées dans cette étude, le modèle binomial négatif modifié en zéro a fourni les meilleurs résultats.
Concernant les coûts moyens, les lois de probabilités classiques modélisant des distributions continues ont été sélectionnées. Globalement, la loi gamma nous a semblé la plus adaptée. Toutefois, l'analyse des résidus des estimations fournies par le GLM a mis en évidence un mauvais ajustement du modèle aux données étudiées.
La prime pure issue de ces modèles a été comparée à la méthode de tarification classique de la fréquence et du coût moyen, et nous avons constaté que le modèle de fréquence binomial négatif modifié en zéro est le plus adapté à nos données.

Compte tenu de la mauvaise représentation des résidus du coût moyen et de la complexité d'utilisation des modèles modifiés en zéro, nous avons choisi de garder le modèle actuel de tarification directe de la fréquence et du coût moyen. Cette méthode de tarification, même si nous ne pouvons pas la valider par des tests statistiques, présente l'avantage d'être plus flexible et d'une utilisation plus simple que le GLM.

Enfin, dans la dernière partie du mémoire, nous nous sommes intéressés au risque d'antisélection dans l'objectif de tarifer les contrats collectifs facultatifs. Une analyse économique a permis de montrer que la non prise en compte de ce phénomène a pour conséquence d'attirer uniquement les individus avec des niveaux de risques élevés. Dans une approche statistique, compte tenu du faible nombre de contrats collectifs facultatifs parmi nos données, nous avons ainsi choisi d'évaluer l'écart de fréquence de consommation entre les contrats collectifs obligatoire et les contrats individuels. Le choix d'analyse par niveaux de garantie du contrat a permis de retirer tout comportement relevant du phénomène d'aléa moral. Ainsi, nous avons calculé des coefficients d'antisélection permettant

de prendre en compte l'excès de consommation par rapport à un contrat obligatoire. Tous postes de garanties confondues, ces coefficients sont presque identiques selon les différents niveaux de garantie considérés. Une analyse par postes de garantie, a montré que le phénomène d'antisélection est presque inexistant dans le cas de la pharmacie, contrairement au poste dentaire où le coefficient atteint son niveau maximal notamment dû aux prothèses dentaires non prises en charge par la Sécurité sociale. Pour finir, une étude en fonction de l'âge de l'assuré a permis de conclure que le comportement antisélectif, inexistant pour les nouveaux nés, est particulièrement marqué à partir de l'âge de 17 ans.

L'étude préalable de la sinistralité a permis également de constater que l'utilisation d'un modèle fréquence - coût dans le cadre de la santé peut biaiser les résultats. En effet, l'hypothèse d'indépendance entre la fréquence de consommation et le coût moyen n'est pas vérifiée pour certains postes de garantie tels que l'optique où le comportement de consommation de l'assuré dépend du reste à charge de l'acte. Par ailleurs, la détermination d'une prime par famille d'actes ne permet pas de prendre en compte les interactions existant entre certaines familles d'actes. Un exemple simple est la corrélation entre la consommation de médicaments et la visite chez un généraliste, ou la corrélation entre l'achat d'une monture de lunette et l'achat de verres. Ainsi, il peut être intéressant de tester des méthodes de tarification alternatives au modèle fréquence - coût afin d'appréhender au mieux le risque santé.
Les modèles non paramétriques sont de plus en plus utilisées afin de s'affranchir des limites d'un modèle fréquence - coût et du GLM. Il s'agit des algorithmes d'apprentissage statistique, qui, contrairement au GLM ne nécessite pas de formuler une hypothèse sur la distribution de la variable modélisée. Plusieurs types de méthodes peuvent être cités : les arbres de décision, les modèles additifs généralisées, les réseaux de neurones, etc. La tarification en santé avec la méthode des réseaux de neurones a notamment été testé dans le cadre d'un mémoire (cf. [5]) et a fourni des meilleurs résultats que le GLM.

Liste des abréviations

ANI : Accord national interprofessionnel

DREES : Direction de la recherche, des études, de l'évaluation et des statistiques

INSEE : Institut national de la statistique et des études économiques

CCAM : Classification commune des actes médicaux

NGAP : Nomenclature générale des actes médicaux

CMU : Couverture maladie universelle

BR : Base de remboursement de la Sécurité sociale

RSS : Remboursement de la Sécurité sociale

FR : Frais réels

CSBM : Consommation de soins et biens médicaux

ACP : Analyse en composantes principales

CAH : Classification ascendante hiérarchique

GLM : Modèle linéaire généralisé

ZIP : Zero inflated Poisson

ZINB : Zero inflated binomial negative

ANOVA : Analyse de la variance

Table des figures

Liste des tableaux

Annexes

Annexe A : Résultats de l'ACP

Cet annexe contient certains résultats de l'ACP fourni par SAS.
Les variables sont codifiées selon quatre lettres : les deux premières lettres «FR» et «CM» permettent d'indiquer la fréquence ou le coût moyen et les deux dernières lettres indiquent le type de famille d'actes. Nous considérons les familles d'actes suivants :

- pharmacie («PH») ;
- honoraires («HO») ;
- auxiliaire («AU») ;
- hospitalisation («HS») ;
- dentaire («DE») ;
- appareillage («AP») ;
- analyses («AN») ;
- radiologie («RA») ;
- optique («OP») ;

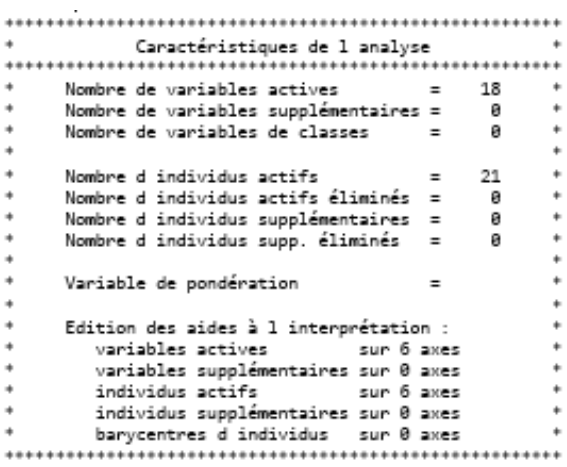


FIGURE 57 – ACP sur les départements

Matrice de corrélation :

Variable	CM PH	CM HO	CM AU	CM DE	CM HS	CM AP
CM PH	1,00	0,49	0,69	-0,21	-0,58	0,53
CM HO	0,49	1,00	0,82	0,56	0,24	0,70
CM AU	0,69	0,82	1,00	0,32	-0,21	0,59
CM DE	-0,21	0,56	0,32	1,00	0,69	0,38
CM HS	-0,58	0,24	-0,21	0,69	1,00	0,06
CM AP	0,53	0,70	0,59	0,38	0,06	1,00
CM AN	0,72	0,49	0,74	-0,10	-0,54	0,29
CM RA	0,63	0,96	0,86	0,44	0,05	0,68
CM OP	-0,47	0,17	-0,08	0,77	0,81	0,17
FR PH	-0,27	-0,30	-0,43	-0,06	0,13	-0,14
FR HO	-0,66	-0,41	-0,51	0,30	0,52	-0,31
FR AU	-0,37	-0,39	-0,36	0,07	0,27	-0,26
FR DE	-0,69	-0,19	-0,37	0,51	0,72	-0,14
FR HS	0,56	0,13	0,36	-0,20	-0,47	0,36
FR AP	-0,52	-0,45	-0,59	0,05	0,46	-0,40
FR OP	-0,64	0,05	-0,20	0,69	0,82	-0,08
FR AN	-0,44	-0,14	-0,39	0,43	0,61	-0,02
FR RA	-0,55	0,02	-0,19	0,61	0,74	-0,07

Variable	CM AN	CM RA	CM OP	FR PH	FR HO	FR AU
CM PH	0,72	0,63	-0,47	-0,27	-0,66	-0,37
CM HO	0,49	0,96	0,17	-0,30	-0,41	-0,39
CM AU	0,74	0,86	-0,08	-0,43	-0,51	-0,36
CM DE	-0,10	0,44	0,77	-0,06	0,30	0,07
CM HS	-0,54	0,05	0,81	0,13	0,52	0,27
CM AP	0,29	0,68	0,17	-0,14	-0,31	-0,26
CM AN	1,00	0,60	-0,55	-0,23	-0,52	-0,32
CM RA	0,60	1,00	0,02	-0,27	-0,52	-0,43
CM OP	-0,55	0,02	1,00	0,05	0,55	0,32
FR PH	-0,23	-0,27	0,05	1,00	0,34	0,62
FR HO	-0,52	-0,52	0,55	0,34	1,00	0,73
FR AU	-0,32	-0,43	0,32	0,62	0,73	1,00
FR DE	-0,65	-0,39	0,78	0,07	0,85	0,55
FR HS	0,49	0,26	-0,33	0,49	-0,25	0,27
FR AP	-0,54	-0,51	0,38	0,67	0,76	0,88
FR OP	-0,48	-0,14	0,81	0,02	0,73	0,51
FR AN	-0,73	-0,26	0,66	0,33	0,67	0,52
FR RA	-0,49	-0,18	0,73	0,07	0,78	0,59

Variable	FR DE	FR HS	FR AP	FR OP	FR AN	FR RA
CM PH	-0,69	0,56	-0,52	-0,64	-0,44	-0,55
CM HO	-0,19	0,13	-0,45	0,05	-0,14	0,02
CM AU	-0,37	0,36	-0,59	-0,20	-0,39	-0,19
CM DE	0,51	-0,20	0,05	0,69	0,43	0,61
CM HS	0,72	-0,47	0,46	0,82	0,61	0,74
CM AP	-0,14	0,36	-0,40	-0,08	-0,02	-0,07
CM AN	-0,65	0,49	-0,54	-0,48	-0,73	-0,49
CM RA	-0,39	0,26	-0,51	-0,14	-0,26	-0,18
CM OP	0,78	-0,33	0,38	0,81	0,66	0,73
FR PH	0,07	0,49	0,67	0,02	0,33	0,07
FR HO	0,85	-0,25	0,76	0,73	0,67	0,78
FR AU	0,55	0,27	0,88	0,51	0,52	0,59
FR DE	1,00	-0,43	0,59	0,87	0,78	0,92
FR HS	-0,43	1,00	0,04	-0,42	-0,16	-0,35
FR AP	0,59	0,04	1,00	0,51	0,65	0,59
FR OP	0,87	-0,42	0,51	1,00	0,60	0,91
FR AN	0,78	-0,16	0,65	0,60	1,00	0,74
FR RA	0,92	-0,35	0,59	0,91	0,74	1,00

Table 38 – Matrice de corrélation fournie par l'ACP

Qualité de représentation des individus sur l'axe 1 :

	COORD axe 1	CO2 axe 1	COORD axe 2	CO2 axe 2
Alsace	9,25	94,4	0,92	0,9
Aquitain	-0,36	2,9	1,38	42,5
Auvergne	-0,67	14,1	-0,32	3,2
Basse-No	-1,7	28,8	-2,17	47,2
Bourgogn	1,14	5,2	-0,48	0,9
Bretagne	-0,45	3,4	0,1	0,2
Centre	-2,2	41,7	0,67	3,8
Champagn	-0,37	6,7	-0,45	9,9
Franche-	0,18	0,4	-0,33	1,3
Haute No	-2,21	49,8	0,01	0
Ile de F	-1,79	5,6	7,21	91,3
Languedo	-0,05	0,1	-0,07	0,1
Limousin	-2,39	43,3	-1,81	24,8
Lorraine	6,73	85,1	-1,81	6,1
Midi-Pyr	-0,85	11	-0,59	5,2
Nord-Pas	-0,57	2,6	-2,23	39,8
PACA	-0,62	4,5	1,71	34,1
Pays de	-1,59	29,5	-1,82	38,9
Picardie	-1,78	38,3	-0,41	2
Poitou-C	-1,53	35	-1,63	40,2
Rhone-Al	1,85	12,4	2,13	16,5

TABLE 39 – Qualité de représentation des individus sur les deux premiers axes

Annexe B : Résultats de la CAH

Critères de sélection du nombre optimal de classes :

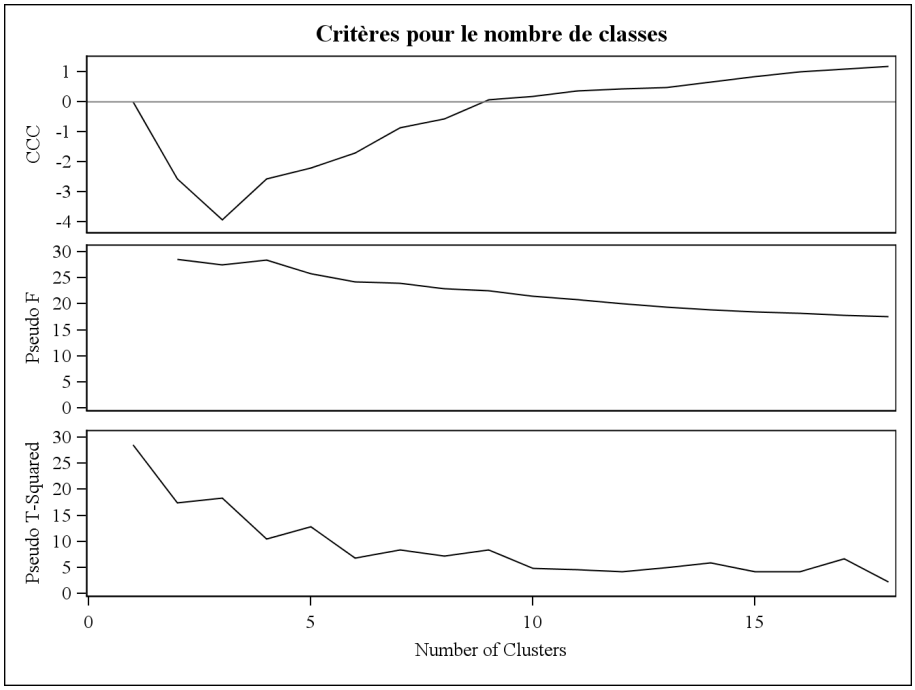


FIGURE 58 – CAH : sélection du nombre de classes optimal

Annexe C : GLM - Modèle binomial négatif (prothèses dentaires)

Les résultats du modèle binomial négatif (sorties SAS) estimant la fréquence relative aux actes de prothèses dentaires sont fournis ci-dessous :

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	91E4	185095.2657	0.2042
Scaled Deviance	91E4	185095.2657	0.2042
Pearson Chi-Square	91E4	2120254.1595	2.3387
Scaled Pearson X2	91E4	2120254.1595	2.3387
Log Likelihood		-110960.4704	
Full Log Likelihood		-329766.8911	
AIC (smaller is better)		659563.7822	
AICC (smaller is better)		659563.7827	
BIC (smaller is better)		659739.5440	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DDL	Valeur estimée	Erreur type	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Khi-2
Intercept		1	0.5414	0.0473	0.4488	0.6340	131.26	<.0001
age	20	1	-4.3384	0.0469	-4.4304	-4.2465	8552.40	<.0001
age	25	1	-1.5704	0.0414	-1.6515	-1.4892	1437.97	<.0001
age	30	1	-1.0201	0.0404	-1.0992	-0.9410	638.86	<.0001
age	35	1	-0.6864	0.0406	-0.7660	-0.6068	285.76	<.0001
age	40	1	-0.4814	0.0410	-0.5617	-0.4010	137.91	<.0001
age	45	1	-0.3059	0.0410	-0.3862	-0.2256	55.78	<.0001
age	50	1	-0.1432	0.0413	-0.2241	-0.0623	12.04	0.0005
age	75	1	0.0966	0.0383	0.0214	0.1718	6.35	0.0118
age	80	1	0.2537	0.0499	0.1558	0.3515	25.80	<.0001
age	81	0	0.0000	0.0000	0.0000	0.0000	.	.
garantie	01	1	-5.5101	0.0475	-5.6032	-5.4169	13441.8	<.0001
garantie	02	1	-2.0217	0.0299	-2.0803	-1.9632	4581.09	<.0001
garantie	03	1	-1.5543	0.0296	-1.6124	-1.4962	2753.63	<.0001
garantie	04	1	-0.7882	0.0312	-0.8493	-0.7271	639.10	<.0001
garantie	05	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	10.0790	0.0577	9.9665	10.1927		

LR Statistics For Type 1 Analysis				
Source	2*Log Likelihood	DF	Khi-2	Pr > Khi-2
Intercept	-295018.73			
age	-253277.67	9	41741.1	<.0001
garantie	-221920.94	4	31356.7	<.0001

FIGURE 59 – Résultats du modèle binomial négatif

La variable âge a été codifié de la manière suivante :
- «20» : âge inférieur à 20 ans ;
- «25» : âge compris entre 21 et 25 ans ;
- «30» : âge compris entre 26 et 30 ans ;

– «35» : âge compris entre 31 et 35 ans ;
– «40» : âge compris entre 36 et 40 ans ;
– «45» : âge compris entre 41 et 45 ans ;
– «50» : âge compris entre 31 et 50 ans ;
– «75» : âge compris entre 51 et 75 ans ;
– «80» : âge compris entre 76 et 80 ans ;
– «81» : âge supérieur ou égal à 81 ans.

Annexe D : GLM - Modèle ZINB (prothèses dentaires)

Les résultats du modèle ZINB (sorties SAS) estimant la fréquence relative aux actes de prothèses dentaires sont fournis ci-dessous :

Model Fit Summary	
Dependent Variable	nsin
Number of Observations	904748
Missing Values	8231
Data Set	ETUDEOZ.TAB_DISJ3
Model	ZINB
Offset Variable	log_expo
ZI Link Function	Logistic
Log Likelihood	-323644
Maximum Absolute Gradient	0.01488
Number of Iterations	103
Optimization Method	Quasi-Newton
AIC	647352
SBC	647727

Résultats estimés des paramètres					
Paramètre	**DDL**	**Valeur estimée**	**Erreur type**	**Valeur du test t**	**Approx. de Pr > |t|**
Intercept	1	0.322056	0.047979	6.71	<.0001
age 20	1	-0.241842	0.076556	-3.16	0.0016
age 25	1	0.113654	0.044912	2.53	0.0114
age 30	1	0.093718	0.039033	2.40	0.0164
age 35	1	0.101668	0.038055	2.67	0.0075
age 40	1	0.159783	0.038010	4.20	<.0001
age 45	1	0.186515	0.037326	5.00	<.0001
age 50	1	0.245686	0.037364	6.58	<.0001
age 75	1	0.141251	0.032026	4.41	<.0001
age 80	1	0.233904	0.041348	5.66	<.0001
age 81	0	0	.	.	.
garantie 01	1	-4.603713	0.047106	-97.73	<.0001
garantie 02	1	-1.116001	0.028988	-38.50	<.0001
garantie 03	1	-0.868335	0.028246	-30.74	<.0001
garantie 04	1	-0.470146	0.029339	-16.02	<.0001
garantie 05	0	0	.	.	.

Inf_Intercept	1	4.785556	0.056654	84.47	<.0001
Inf_age20	0	0	.	.	.
Inf_age25	1	-2.505721	0.059685	-41.98	<.0001
Inf_age30	1	-3.196784	0.057470	-55.63	<.0001
Inf_age35	1	-3.603633	0.058527	-61.57	<.0001
Inf_age40	1	-3.798231	0.059823	-63.49	<.0001
Inf_age45	1	-4.020844	0.060797	-66.14	<.0001
Inf_age50	1	-4.188454	0.062236	-67.30	<.0001
Inf_age80	1	-4.864824	0.067361	-72.22	<.0001
Inf_garantie02	0	0	.	.	.
Inf_garantie03	1	-0.348134	0.018246	-19.08	<.0001
Inf_garantie04	1	-0.921439	0.029311	-31.44	<.0001
Inf_garantie05	1	-1.407589	0.060604	-23.23	<.0001
Inf_sexeF	0	0	.	.	.
Inf_sexeH	1	0.112318	0.012939	8.68	<.0001
_Alpha	1	3.208387	0.119020	26.96	<.0001

FIGURE 60 – Résultats du modèle ZINB

La codification de la variable âge est présentée dans le corps du mémoire.

Annexe E : GLM - Comparaison des modèles de fréquence (analyses et actes de laboratoire)

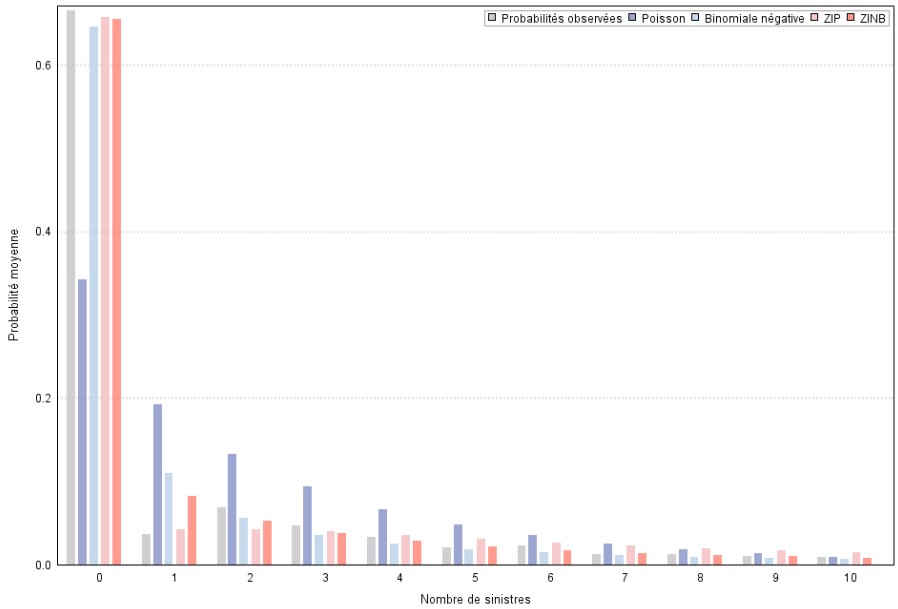


FIGURE 61 – Probabilités moyenne observées et prédites par les différents modèles GLM (analyses et actes de laboratoire)

Modèle	AIC	BIC
Poisson	4998144	4998388
Binomial négatif	2507475	2507684
ZIP	3479600	3479950
ZINB	2461370	2461778

TABLE 40 – Critère AIC et BIC pour les analyses et actes de laboratoire

Annexe F : GLM - Coût moyen

Prothèses dentaires :

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	67E3	52975.5210	0.7941
Scaled Deviance	67E3	70783.7106	1.0611
Pearson Chi-Square	67E3	47363.1838	0.7100
Scaled Pearson X2	67E3	63284.7367	0.9487
Log Likelihood		-374323.0356	
Full Log Likelihood		-374323.0356	
AIC (smaller is better)		748668.0712	
AICC (smaller is better)		748668.0751	
BIC (smaller is better)		748768.2617	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DDL	Valeur estimée	Erreur type	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Khi-2
Intercept		1	5.3220	0.0090	5.3043	5.3396	349264	<.0001
garantie	02	1	-0.9300	0.0078	-0.9453	-0.9147	14157.7	<.0001
garantie	03	1	-0.6034	0.0073	-0.6178	-0.5890	6754.16	<.0001
garantie	04	1	-0.2101	0.0074	-0.2247	-0.1955	795.62	<.0001
garantie	05	0	0.0000	0.0000	0.0000	0.0000	.	.
age	20	1	0.0917	0.0236	0.0453	0.1380	15.02	0.0001
age	30	1	0.1073	0.0079	0.0919	0.1228	184.73	<.0001
age	50	1	0.1068	0.0065	0.0941	0.1195	271.66	<.0001
age	60	1	0.0571	0.0071	0.0431	0.0711	64.11	<.0001
age	70	1	0.0361	0.0069	0.0226	0.0497	27.24	<.0001
age	71	0	0.0000	0.0000	0.0000	0.0000	.	.
regime	General	1	0.0685	0.0043	0.0600	0.0770	250.04	<.0001
regime	Local	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	1.3362	0.0069	1.3227	1.3498		

FIGURE 62 – Résultats GLM du coût moyen des prothèses dentaires

Le niveau de garantie 1 a été regroupé avec le niveau de garantie 2 et la variable âge a été codifié de la manière suivante :

- «20» : âge inférieur à 20 ans ;
- «30» : âge compris entre 21 et 30 ans ;
- «50» : âge compris entre 31 et 50 ans ;
- «60» : âge compris entre 51 et 60 ans ;
- «70» : âge compris entre 61 et 70 ans ;
- «71» : âge supérieur ou égal à 71 ans.

Analyses et actes de laboratoire :

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	31E4	640771.6491	2.0847
Scaled Deviance	31E4	339820.6578	1.0950
Pearson Chi-Square	31E4	726189.1416	2.3400
Scaled Pearson X2	31E4	385120.1471	1.2410
Log Likelihood		-895216.1372	
Full Log Likelihood		-895216.1372	
AIC (smaller is better)		1790450.2744	
AICC (smaller is better)		1790450.2749	
BIC (smaller is better)		1790548.0835	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DDL	Valeur estimée	Erreur type	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Khi-2
Intercept		1	0.6026	0.0056	0.5917	0.6135	11686.6	<.0001
regime	General	1	1.2879	0.0043	1.2795	1.2963	91170.2	<.0001
regime	Local	0	0.0000	0.0000	0.0000	0.0000	.	.
departement	01	1	0.2208	0.0041	0.2126	0.2289	2842.34	<.0001
departement	02	1	0.0293	0.0048	0.0199	0.0387	37.35	<.0001
departement	03	1	0.0371	0.0041	0.0290	0.0452	80.50	<.0001
departement	04	1	0.2165	0.0050	0.2066	0.2264	1841.46	<.0001
departement	05	1	0.1984	0.0049	0.1889	0.2080	1657.51	<.0001
departement	06	1	0.1304	0.0053	0.1201	0.1407	615.98	<.0001
departement	07	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.5303	0.0012	0.5279	0.5328		

FIGURE 63 – Résultats GLM du coût moyen des analyses et actes de laboratoire

Bibliographie

Ouvrages

[1] M. DENUIT, A. CHARPENTIER [2005] Mathématiques de l'assurance non-vie, Tome 2 : tarification et provisionnement, Econometrica, p.70-109.

[2] J. GUIZOUARN, N.MARESCAUX [2004] Assurance santé, segmentation et compétitivité, Economica.

Mémoires

[3] F. LAGADEC [2009] Tarification d'un contrat de complémentaire santé par un modèle linéaire généralisé, EURIA.

[4] M. VAUTRIN [2008/2009] Élaboration d'une méthode de tarification avec indicateurs de risque pour des contrats complémentaires santé collectifs, ISUP.

[5] J. AOUIZERATE [2010] Alternative neuronale en tarification santé, CNAM.

Cours

[6] E. PERINEL [2013] Analyse de données, Université de Strasbourg.

[7] A. GUILLOU [2012] Statistiques, Université de Strasbourg.

[8] A. YOU [2013] Tarification non vie, Université de Strasbourg.

[9] S. SPAETER [2012] Économie et gestion du risque, Université de Strasbourg.

[10] F. BERTRAND, M. MAUMY-BERTRAND, Choix du modèle, notes de cours téléchargés sur www.irma.u-strasbg.fr, Université de Strasbourg.

[11] P. BESSE, Introduction au modèle linéaire général consulté sur le site www.math.univ-toulouse.fr, Université de Toulouse.

Publications

[12] DREES [2014] Recueil d'indicateurs régionaux : offre de soins et état de santé.

[13] INRIA [2004] Le critère BIC : fondements théoriques et interprétation.

[14] E. ALLAIN, T. BRENAC [2001] Modèles linéaires généralisés appliqués à l'étude des nombre d'accidents sur des sites routiers : le modèle de Poisson et ses extensions.

[15] M. PERRONNIN [2013] Effet de l'assurance complémentaire santé sur les consommations médicales : entre risque moral et amélioration de l'accès aux soins, thèse, Université Paris-Dauphine.

[16] S. ETTNER [1995] Adverse selection and the purchase of Medigap insurance by the elderly, Journal of Health Economics (16) p.543-562.

[17] M ROTHSCHILD, J. STIGLITZ [1976] Equilibrium in competitive insurance markets : an essay on the economics of imperfect information.

[18] CREDES [2002] La consommation de médicaments varie-t-elle selon l'assurance complémentaire ?.

Sites Internet

[19] Site de l'assurance maladie : www.ameli.fr.

[20] Site de l'INSEE : www.insee.fr.

[21] Site de la DREES : www.drees.sante.gouv.fr.

[22] Site comportant la loi n°2013-504 du 14 juin 2013 relative à la sécurisation de l'emploi : www.legifrance.gouv.fr.

[23] Site de documentation du logiciel SAS : www.support.sas.com.

Autres

[24] SAS, SAS/STAT 9.3 User's guide, Chapter 39 The GENMOD Procedure p.2607-2801, Chapter 41 The GLM Procedure p.3154-3333.

[25] SAS, SAS/ETS User's guide, Chapter 11 The COUNTREG Procedure, p.419-443.

[26] Magazine L'actuariel n°11, janvier 2014, p.20-24.