



المندوبية السامية للتخطيط
HAUT-COMMISSARIAT AU PLAN

ROYAUME DU MAROC

..*.*.*

HAUT COMMISSARIAT AU PLAN

..*.*.*

INSTITUT NATIONAL
DE STATISTIQUE ET D'ECONOMIE APPLIQUEE



Projet de Fin d'Etudes

Tarification en assurance maladie de base

Préparé par : M. Othman EL JAMYLY

Sous la direction de : M. Mustapha BERROUYNE (INSEA)

M. Achraf FTOUHI (SANAD)

M. Mohammed ET-TABII (SANAD)

Soutenu publiquement comme exigence partielle en vue de l'obtention du

Diplôme d'Ingénieur d'Etat

Option : ACTUARIAT-FINANCE

Devant le jury composé de :

- M. Jelloul EL MABROUK (INSEA)
- M. Mustapha BERROUYNE (INSEA)
- M. Achraf FTOUHI (SANAD)

RESUME

L'actuaire est amené, dans le cadre de ses fonctions, à mettre en place les techniques nécessaires à la tarification des contrats d'assurance.

Le présent projet de fin d'études fournit les éléments visant à la tarification de la branche d'assurance maladie de base d'une compagnie d'assurance privée.

Sont présentés, de façon synthétique, les fondements de la branche de l'assurance maladie de base ainsi que les particularités de l'environnement de tarification rencontré par l'actuaire. L'essence de ce projet réside dans la description d'une méthodologie destinée à la tarification des assurés de la branche assurance maladie groupe.

Articulé en quatre chapitres, ce présent travail met l'accent, dans son premier tiers, sur les enjeux de l'assurance maladie au Maroc, le cadre économique du marché assurantiel, ainsi qu'une présentation de l'organisme d'accueil.

Le deuxième chapitre, est consacré à une présentation du portefeuille maladie, ses caractéristiques, ses spécificités, ainsi qu'une analyse descriptive de données selon plusieurs facteurs de sinistralité.

Le troisième chapitre présente le cadre conceptuel et théorique ainsi que l'approche adoptée pour modéliser la prime de l'assurance maladie.

Le dernier chapitre, quant à lui, s'est focalisé sur la modélisation de la fréquence et des frais moyens pour aboutir à un calcul du tarif de l'assurance maladie. Les résultats y sont ensuite analysés.

MOTS CLES

- Assurance maladie, tarification, CHAID, GLM, fréquence, sévérité, prime pure.

DEDICACE

Je dédie ce travail aux êtres qui nous en les plus chers

A mes parents avec tous mes sentiments de reconnaissances pour tous les sacrifices déployés pour m'élever dignement et assurer mon éducation dans les meilleures conditions et qui nous ont soutenus par leurs prières et leur amour

A mon frère ISSAM, à mes sœurs NASSIMA et NAJWA, et à toute ma famille

Je dédie aussi ce travail à tous mes amis

REMERCIEMENTS

Au terme de ce travail, je tiens à exprimer ma profonde gratitude envers M. FTOUHI Achraf, directeur du pôle support technique à SANAD Assurances, pour m'avoir accueilli au sein de leur département et ses enseignements qui témoignent d'une grande expérience.

J'adresse mes profonds remerciements à mon encadrant interne, M. BERROUYNE Mustapha enseignant à l'INSEA, pour ses conseils et ses encouragements.

J'exprime également ma reconnaissance à M. ET-TABII Mohamed pour leur encadrement et le soutien chaleureux dont il a toujours fait preuve.

Je tiens à remercier toutes les personnes qui ont contribué de près ou de loin au bon déroulement de mon stage de fin d'études, et toutes les personnes qui m'ont formé tout au long de cette expérience professionnelle.

J'exprime également ma gratitude et mes vifs sentiments à l'égard de l'ensemble du personnel de la compagnie SANAD Assurances qui ont facilité mon incursion, et n'ont pas hésité à me transmettre leurs connaissances, leurs expériences et leurs conseils.

Je profite de cette occasion pour remercier l'ensemble du corps professoral de l'INSEA, pour leurs efforts en vue d'assurer une formation de haut niveau pour l'ensemble des étudiants.

TABLES DES MATIERES

RESUME.....	3
MOTS CLES.....	3
DEDICACE.....	4
REMERCIEMENTS.....	5
TABLES DES MATIERES	6
LISTES DES TABLEAUX.....	10
TABLE DES GRAPHIQUES.....	11
TABLE DES FIGURES.....	12
INTRODUCTION GENERALE	13
CHAPITRE 1. ASSURANCE MALADIE AU MAROC : DES GENERALITES.....	16
INTRODUCTION.....	16
I. PRESENTATION DE L'ORGANISME D'ACCUEIL	16
1. SANAD ASSURANCE	16
2. Branches d'activités	16
3. Organisation	18
4. SANAD ASSURANCE en chiffres	19
5. l'Assurance Marocaine.....	19
5.1. Part marché des assureurs marocains	19
5.2. Densité de l'assurance	20
II. L'assurance maladie au Maroc	21
1. Présentation de l'assurance maladie.....	21
2. L'assurance maladie obligatoire.....	22
2.1. Types d'assurance maladie.....	22
2.2. Organismes de gestion	22
3. La CNOPS et l'assurance maladie	23
3.1. La population couverte.....	23
3.2. Taux de cotisation	23
3.3. Panier de soins.....	23
4. La CNSS et l'assurance maladie	24
4.1. La population éligible.....	24
4.2. Taux de cotisation	25
4.3. Panier de soins.....	25
5. Les organismes privés et l'assurance maladie.....	26
CONCLUSION	26

CHAPITRE 2. SOURCE ET ANALYSE DES DONNEES.....	28
INTRODUCTION.....	28
I. Présentation du portefeuille	28
1. Fichiers reçus.....	28
2. Traitement des données	29
3. Agrégation des données	31
3.1 Attribution de la masse salariale	31
3.2 Attribution de l'exposition	31
3.3 Agrégation de la sinistralité.....	32
II. Analyse des données.....	32
1. Description générale de la base de données	32
2. Distribution de la population.....	33
2.1 Répartition du portefeuille selon le sexe	33
2.2 Répartition du portefeuille selon le lien	33
2.3 Sinistralité du portefeuille	34
3. Analyse de la sinistralité	36
3.1. Selon l'âge et le sexe de l'assuré.....	36
3.2 Selon la masse salariale de l'entreprise assurée	38
3.3 Selon le taux de remboursement	39
CONCLUSION	39
CHAPITRE 3. CADRE CONCEPTUEL ET THEORIQUE.....	41
INTRODUCTION.....	41
I. Choix de modèle par sélection de variables.....	42
1. Pourquoi la sélection de variables	42
2. Critères de sélection de variables	42
3. La procédure stepwise	43
II. Segmentation et codification des variables	43
1. Arbre de régression	43
2. L'algorithme de CHAID	43
3. La méthode de codification	44
4. Test de khi-deux d'indépendance.....	45
III . Etude théorique selon l'approche GLM	45
1. Généralités.....	45
1.1. Des modèles linéaires aux MLG	46
1.2. La famille des distributions exponentielles	46
1.3. les modèles linéaires généralisés	47

2. Estimation des paramètres.....	48
2.1. Maximum de vraisemblance	48
2.2. Paramètres du modèle	48
3. Choix des facteurs	49
3.1. Généralités.....	49
3.2. Inférence sur les paramètres	50
3.3. Test de Wald sur un paramètre.....	50
3.4. Prédiction de la variable réponse	51
4. Validation du modèle	51
4.1. La déviance	51
4.2. Analyse des résidus	53
4.3. Choix entre différents modèles	53
5. Le tarif de l'assurance maladie.....	54
5.1 l'approche fréquence sévérité.....	54
5.2 La prime commerciale.....	55
CONCLUSION	55
CHAPITRE 4. TARIFICATION EN ASSURANCE MALADIE DE BASE	58
INTRODUCTION.....	58
Section 1 : Segmentation et écrêtement des données en assurance maladie	58
I. Classification des assurés.....	58
1. Sélection de variables tarifaires.....	58
2. Segmentation des variables tarifaires	59
II. Ecrêtement des données.....	63
1. Les sinistres graves en assurance maladie.....	63
2. Sélection de seuil.....	65
2.1. Estimateur de Hill.....	65
2.2. La fonction moyenne des excès	66
3. Ecrêtement des valeurs aberrantes	67
Section 2 : Détermination de la prime pure.....	67
I. Modélisation de la fréquence des sinistres.....	67
1. Ajustement et choix du modèle	67
2. Cellule de référence.....	70
3. Analyse de Type III.....	70
4. Analyse des résultats	71
4.1 Estimation des paramètres du modèle établi	71

4.2 Validation du modèle	72
4.3 Prédiction de la fréquence	73
II. Modélisation de la charge des sinistres	74
1. Choix de la distribution (graphique qq-plot).....	74
2. Le modèle tarifaire	76
3. Analyse des résultats	77
3.1 Estimation des paramètres du modèle	77
3.2 Analyse des résidus	77
4. Prédiction	79
CONCLUSION	79
CONCLUSION GENERALE	80
BIBLIOGRAPHIE	82
ANNEXE I. MODELISATION DE LA FREQUENCE DES SINISTRES.....	83
ANNEXE II. MODELISATION DE LA CHARGE DES SINISTRES.....	90
ANNEXE III. EXTRAITS DES PROCEDURES SAS	95

LISTES DES TABLEAUX

Tableau 1. Organigramme de la compagnie d'assurance SANAD	18
Tableau 2. Description des champs de la table de tarification	32
Tableau 3. Sélection stepwise basé sur le F partiel de Fisher	59
Tableau 4. Test de Khi-deux d'indépendance entre l'âge moyen et le nombre de sinistres	60
Tableau 5. Variation de la sinistralité suivant les classes de l'âge moyen.....	61
Tableau 6. Variation de la sinistralité suivant les classes du taux de remboursement.....	61
Tableau 7. Variation de la sinistralité suivant les classes du salaire moyen	61
Tableau 8. Variation de la sinistralité suivant les classes du plafond général.....	62
Tableau 9. Variation de la sinistralité suivant la localisation de l'entreprise.....	62
Tableau 10. Variation de la sinistralité suivant le sexe de l'assuré.....	63
Tableau 11. La classe de référence.....	70
Tableau 12. Analyse de type 3 des facteurs de sinistralité.....	71
Tableau 13. Estimation des paramètres de la fréquence des sinistres	72
Tableau 14. Test de déviance du modèle assurés maladie	73
Tableau 15. Prédiction de la fréquence des sinistres du poste pharmacie.....	74
Tableau 16. Critères AIC et BIC des différents modèles	76
Tableau 17. Analyse de type 3 de la régression log normale	76
Tableau 18. Estimation des paramètres du coût moyen	77
Tableau 19. Prédiction du cout moyen des sinistres du poste pharmacie	79

TABLE DES GRAPHIQUES

Graphique 1. Primes émises par la compagnie SANAD.....	19
Graphique 2. la part du marché assurantiel marocain par entreprise.....	20
Graphique 3. Densité de l'assurance 2006-2011 en dollar (US).....	21
Graphique 4. Répartition des assurés selon le sexe.....	33
Graphique 5. Répartition des assurés selon le lien.....	34
Graphique 6. Evolution de la sinistralité sur la période 2010 – 2014	34
Graphique 7. Variation de la sinistralité suivant le type de bénéficiaire.....	36
Graphique 8. Distribution du nombre de sinistre selon l'âge.....	37
Graphique 9. Evolution de la masse salariale assurée.....	38
Graphique 10. Variation de la sinistralité suivant le salaire moyen déclaré	38
Graphique 11. Distribution de la sinistralité selon le taux de remboursement.....	39
Graphique 12. Courbe de concentration de Lorenz.....	64

TABLE DES FIGURES

Figure 1. Distribution de la fréquence des sinistres	35
Figure 2. qq-plot des adhérents de la prestation pharmacie	64
Figure 3. Hill-plot de la charge des sinistres graves	65
Figure 4. La fonction moyenne des excès	66
Figure 5. tcplot de la charge des sinistres (en 1000 Dh)	67
Figure 6. Ajustement de la fréquence des sinistres par une loi de Poisson.....	68
Figure 7. Ajustement de la fréquence des sinistres par une loi binomiale négative.....	69
Figure 8. qq-plot gamma de la charge des sinistres	74
Figure 9. qq-plot log normal de la charge des sinistres.....	75
Figure 10. qq-plot exponentielle de la charge des sinistres.....	75
Figure 11. Représentation de la déviance pour le poste pharmacie	78

INTRODUCTION GENERALE

Quotidiennement, des milliers de personnes dans le Maroc seront victimes d'une maladie dont de centaines seront mortelles. Ces maladies font également payer un lourd tribut à l'économie marocaine puisque l'estimation de leur coût a atteint 6% du PIB national en 2014. Ce phénomène est dangereusement aggravé par le contexte de crise économique que nous vivons actuellement. Des compromis motivés par des raisons démographiques sont susceptibles d'accroître le nombre de personnes qui tombent malades et d'en amplifier les conséquences.

Cependant, l'instauration d'un tel système de protection sociale est qualifiée d'un devoir de l'Etat selon la déclaration universelle des droits de l'homme de 1948. Face à cette vérité, les États ne tardent pas à mettre en place des systèmes de sécurité sociale destinés à couvrir les frais de soins de santé.

Actuellement, la branche de l'assurance maladie est encore majoritairement gérée par des organismes publics bien que certains États consentent à en libérer partiellement la gestion aux assureurs privés. Pourtant, les spécificités d'exposition de la branche et la forte concurrence qui caractérise le marché assurantiel marocain, font régulièrement subir de larges déconvenues financières aux compagnies s'y engageant.

L'un des outils permettant de s'assurer un équilibre financier stable réside dans l'élaboration du tarif des contrats d'assurance. En effet, un processus de tarification adéquat, doit permettre de contrôler les risques encourus et d'assurer les objectifs définis par la compagnie.

L'objectif assigné à ce projet de fin d'études, est de proposer une méthode de tarification adaptée à la complexité absolue de la branche assurance maladie, en s'attachant à ajuster le tarif en fonction de la sinistralité observée de l'assuré.

Pour atteindre l'objectif fixé, notre projet s'articule autour de quatre chapitres. Le premier chapitre est consacré à la présentation de l'organisme d'accueil, de la branche d'assurance maladie, de ses spécificités et de ses enjeux réglementaires. Il offre une vue globale de la branche étudiée et s'efforce de donner une image fidèle du marché actuel.

Le deuxième chapitre, quant à lui, présente les données utilisées lors de l'étude, le traitement de ces données, ainsi que l'analyse descriptive du portefeuille maladie. Il introduit également les méthodes envisagées qui servent à segmenter les assurés de la branche maladie, en groupes homogènes.

Le troisième chapitre met l'accent sur le cadre conceptuel et théorique, dans lequel nous allons définir les différentes approches adoptées pour modéliser la prime pure de l'assurance maladie. Il définit les étapes de la mise en place du modèle.

Enfin, le chapitre quatre décrit le cadre pratique des éléments constitutifs de la tarification en assurance maladie de base. Il décrit les résultats des différents modèles et présente les limites et la possibilité d'amélioration de la méthode envisagée.

CHAPITRE 1. ASSURANCE MALADIE AU MAROC : DES GENERALITES

CHAPITRE 1. ASSURANCE MALADIE AU MAROC : DES GENERALITES

INTRODUCTION

Le présent chapitre décrit le cadre général dans lequel s'est déroulé notre projet de fin d'études, ainsi qu'une introduction de l'assurance maladie au Maroc. En effet, la première partie est consacrée à la présentation de l'organisme d'accueil SANAD ASSURANCE, la situation financière de la compagnie en question, ainsi qu'une vue globale de l'assurance au Maroc.

La deuxième partie introduit la branche d'assurance maladie au Maroc, ses spécificités et ses enjeux réglementaires. Nous allons découvrir également les organismes qui sont dotés de la gestion assurantielle de la branche maladie de base.

I. PRESENTATION DE L'ORGANISME D'ACCUEIL

1. SANAD ASSURANCE

Créés en 1946, SANAD Assurance filiale de deux géants de la place économique marocaine, HOLMARCOM-groupe marocain multisectoriel d'envergure et la caisse de dépôts et de gestion CDG premier groupe financier du Royaume. SANAD est une société anonyme au capital de 250 Millions de Dirhams avec un chiffre d'affaire plus de 1,5 milliard de dirhams en 2013. Aujourd'hui SANAD fait partie des cinq principaux leaders du marché national de l'assurance, en 5 ans elle a vu son chiffre d'affaire croître de 36,6%.

Forte de son expérience de 65 ans et de la richesse de son capital humain, SANAD a su se démarquer et à s'imposer dans un secteur aussi compétitif que celui de l'assurance et de la réassurance. SANAD aspire à souffler un vent de liberté en proposant à ses clients des produits en permanence diversifiés et innovants.

Assureur par excellence des risques industriels et des risques de pointe, SANAD renforce son positionnement en tant qu'assureur maritime, et nourrit de grande ambition quant au développement des segments particuliers et professionnels en passant par celui des PME.

2. Branches d'activités

Sur le plan technique, la compagnie d'assurance SANAD exerce plusieurs activités que ce soit en assurance non vie ou en assurance vie, ajoutant à cela les opérations de la réassurance. Les différentes opérations traitées par la compagnie SANAD ASSURANCE se présentent comme suit¹ :

¹ Source : <http://www.guide-assurance.ma/sanad-assurance-maroc.php>

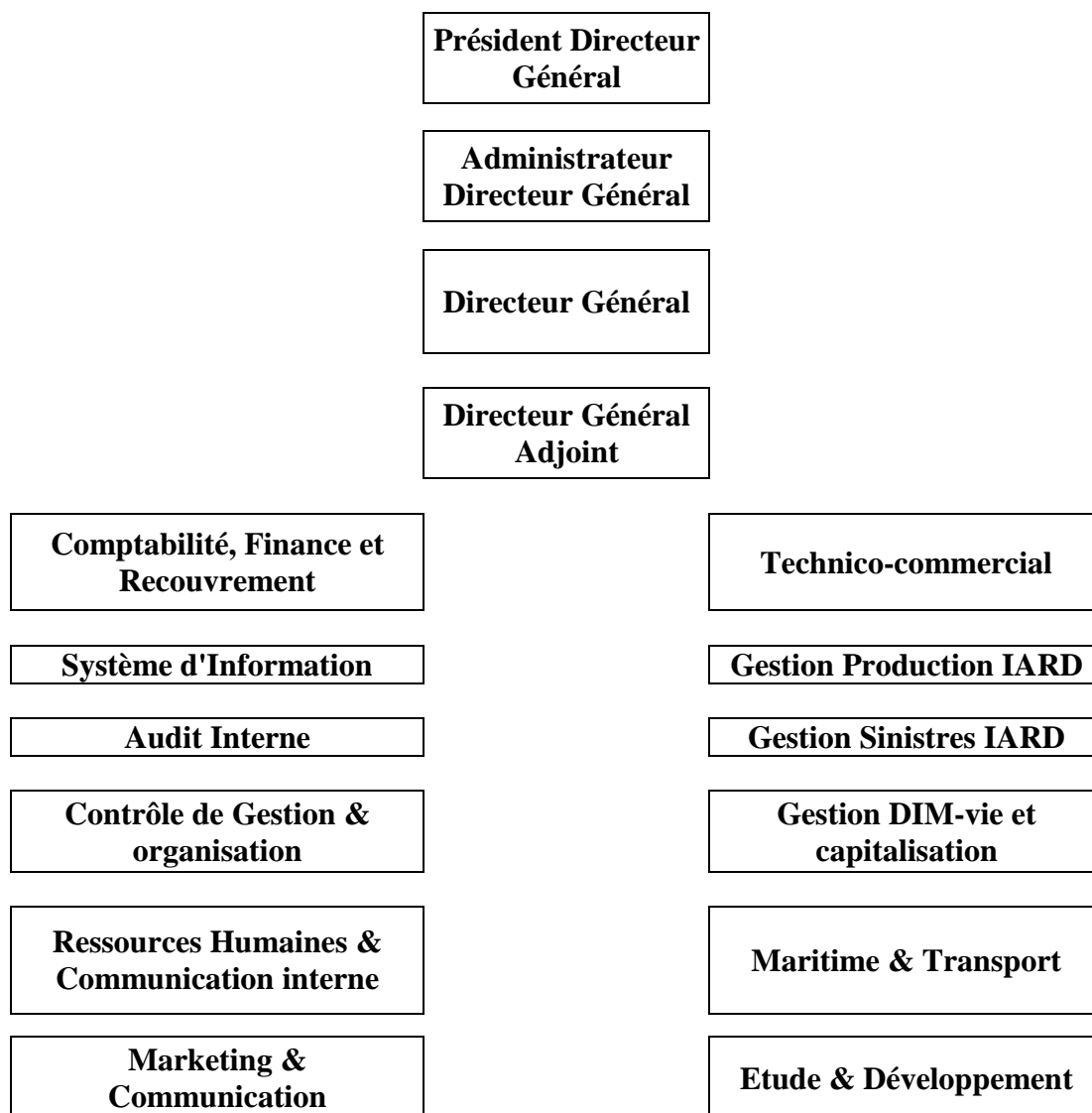
- Vie et décès : toute opération d'assurances comportant des engagements dont l'exécution dépend de la durée de la vie humaine ;
- Capitalisation : toute opération d'appel à l'épargne en vue de la capitalisation et comportant, en échange de versements uniques ou périodiques directs ou indirects, des engagements déterminés ;
- Assurances liées à des fonds d'investissement : toutes opérations comportant des engagements dont l'exécution dépend de la durée de la vie humaine ou faisant appel à l'épargne et liées à un ou plusieurs fonds d'investissement ;
- Opérations faisant appel à l'épargne dans le but de réunir les sommes versées par les assurés en vue de la capitalisation en commun, en les faisant participer aux bénéfices des sociétés gérées ou administrées directement ou indirectement par l'entreprise d'assurances et de réassurance ;
- Opérations d'assurances contre les risques d'accidents corporels.
- Maladie – maternité ;
- Opérations d'assurances contre les risques résultant d'accidents ou de maladies survenus par le fait ou à l'occasion du travail ;
- Opérations d'assurances des corps des véhicules terrestres ;
- Opérations d'assurances contre les risques de responsabilité civile résultant de l'emploi de véhicules terrestres à moteur y compris la responsabilité du transporteur ;
- Opérations d'assurances des corps de navires ;
- Opérations d'assurances contre les risques de responsabilité civile résultant de l'emploi de véhicules fluviaux et maritimes y compris la responsabilité du transporteur ;
- Opérations d'assurances des marchandises transportées ;
- Opérations d'assurances des corps d'aéronefs ;
- Opérations d'assurances contre les risques de responsabilité civile résultant de l'emploi d'aéronefs y compris la responsabilité du transporteur ;
- Opérations d'assurances contre l'incendie et éléments naturels : toute assurance couvrant tout dommage subi par les biens, autres que les biens compris dans les catégories 10°, 12°, 14° et 15°, lorsque ce dommage est causé par : incendie, explosion, éléments et événements naturels autres que la grêle et la gelée, énergie nucléaire et affaissement de terrain ;
- Opérations d'assurances des risques techniques : toute assurance couvrant les risques et engins de chantiers, les risques de montage, le bris de machines, les risques informatiques et la responsabilité civile décennale ;
- Opérations d'assurances contre les risques de responsabilité civile non visés aux paragraphes 9°, 11°, 13°, 16° et 18° ci-dessus ;
- Opérations d'assurances contre le vol ;
- Opérations d'assurances contre les dégâts causés par la grêle ou la gelée ;
- Opérations d'assurances contre les risques de pertes pécuniaires ;

- Protection juridique : toute opération d'assurances consistant à prendre en charge des frais de procédures ou à fournir des services en cas de différends ou de litiges opposant l'assuré à un tiers ;
- Opérations d'assurances contre les risques bris de glaces et dégâts des eaux ;
- Opérations de réassurance.

3. Organisation

Sous la supervision de la présidence direction générale, la compagnie SANAD est composée de plusieurs départements et services qui ont pour objectif la continuité de l'activité assurantielle de la compagnie. Ci-dessous, nous trouverons l'organigramme de la compagnie SANAD.

Tableau 1. Organigramme de la compagnie d'assurance SANAD²

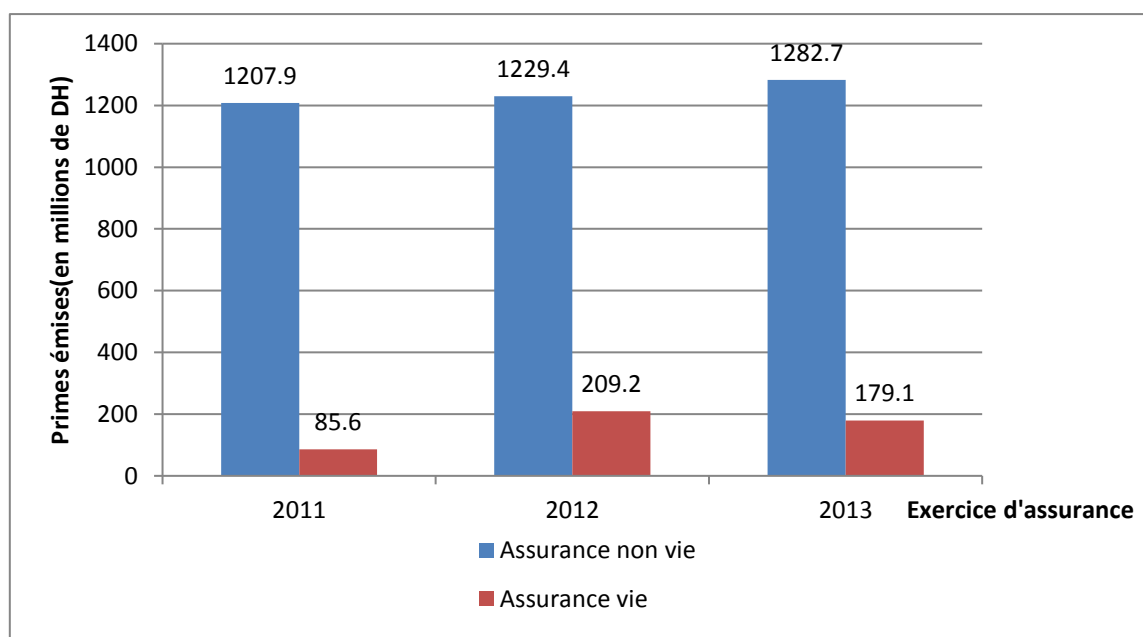


² Source : Rapport annuel de la compagnie SANAD

4. SANAD ASSURANCE en chiffres

Selon la fédération marocaine des sociétés d'assurances et de réassurance, la compagnie SANAD a marqué en 2013 une part de marché de l'ordre de 5,5 %. C'était le résultat d'une évolution régulière en matière des primes émises par la société d'assurances. Le graphique 1 illustre l'évolution des chiffres d'affaires de la compagnie SANAD en 2013.

Graphique 1. Primes émises par la compagnie SANAD



Source : http://www.fmsar.org.ma/docs/Situation_liminaire_2013.pdf

5. l'Assurance Marocaine

5.1. Part marché des assureurs marocains

Le marché marocain d'assurance et de réassurance est constitué, en 2011, par dix-sept entreprises dont quatorze sont commerciales et 3 mutuelles. Sur ce total, sept pratiquent aussi bien les opérations d'assurances non vie que les assurances vie et capitalisation, quatre se limitent aux opérations d'assurances non vie, une pratique exclusivement les opérations d'assurance vie et capitalisation. Trois pratiquent les opérations d'assistance, une pratique exclusivement les opérations d'assurance-crédit, et une entreprise est spécialisée dans la réassurance.

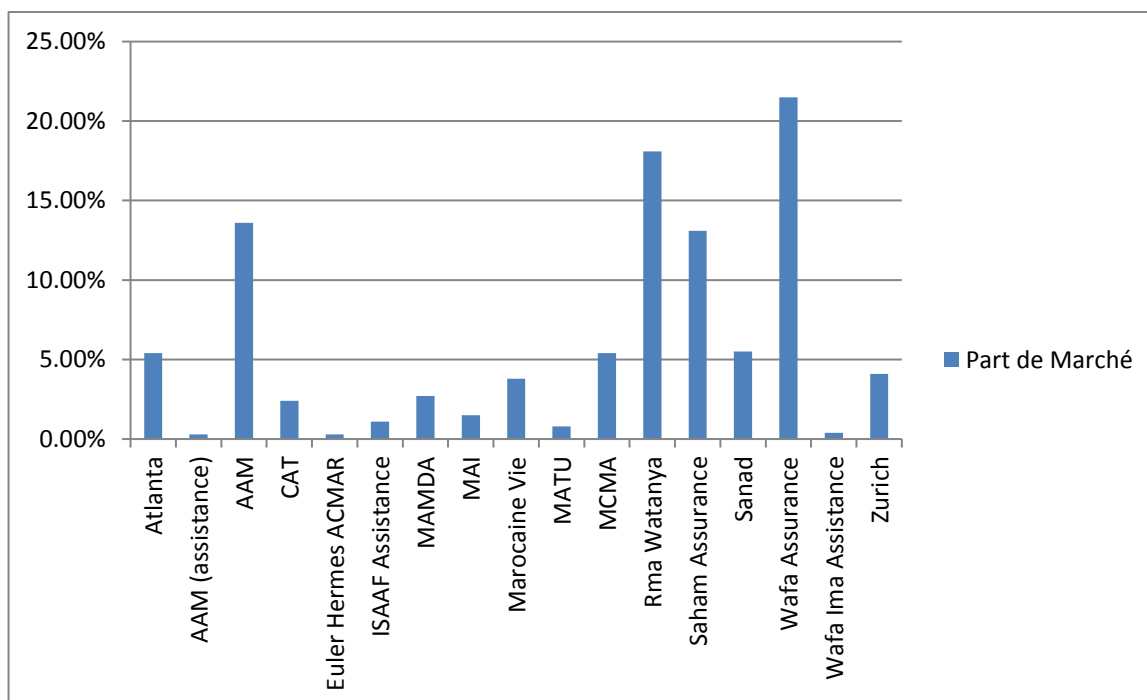
Le nombre des intermédiaires d'assurance a connu une progression significative dans les dernières années. A la fin de 2010, ce nombre a atteint 1364 intermédiaires : 1034 agents d'assurances et 330 courtiers d'assurances.

En 2013, le marché assurantiel marocain était dominé par trois assureurs, à savoir : Wafa Assurance, Rma Watanya et Axa Assurance Maroc. Ces trois premiers assureurs détiennent 53,2% du marché. La compagnie SANAD vient en cinquième rang dans le marché des

assurances avec une part de marché de l'ordre de 5,5%, et ceci suite à l'augmentation des primes émises dans le secteur d'assurance non vie.

Le graphique 2 donne la répartition des parts du marché assurantiel en 2014 détenues par chaque compagnie pratiquant les opérations d'assurance et de réassurance. Il est à signaler que la part du marché dans le secteur d'assurance reflète d'une manière significative la santé financière des assureurs.

Graphique 2. La part du marché assurantiel marocain par entreprise



Source : http://www.fmsar.org.ma/docs/Situation_liminaire_2013.pdf

5.2. Densité de l'assurance

La densité de l'assurance est la somme des dépenses d'assurance effectuées annuellement par habitant. C'est une moyenne qui donne une idée sur la part du revenu consacré à la consommation du service assurance

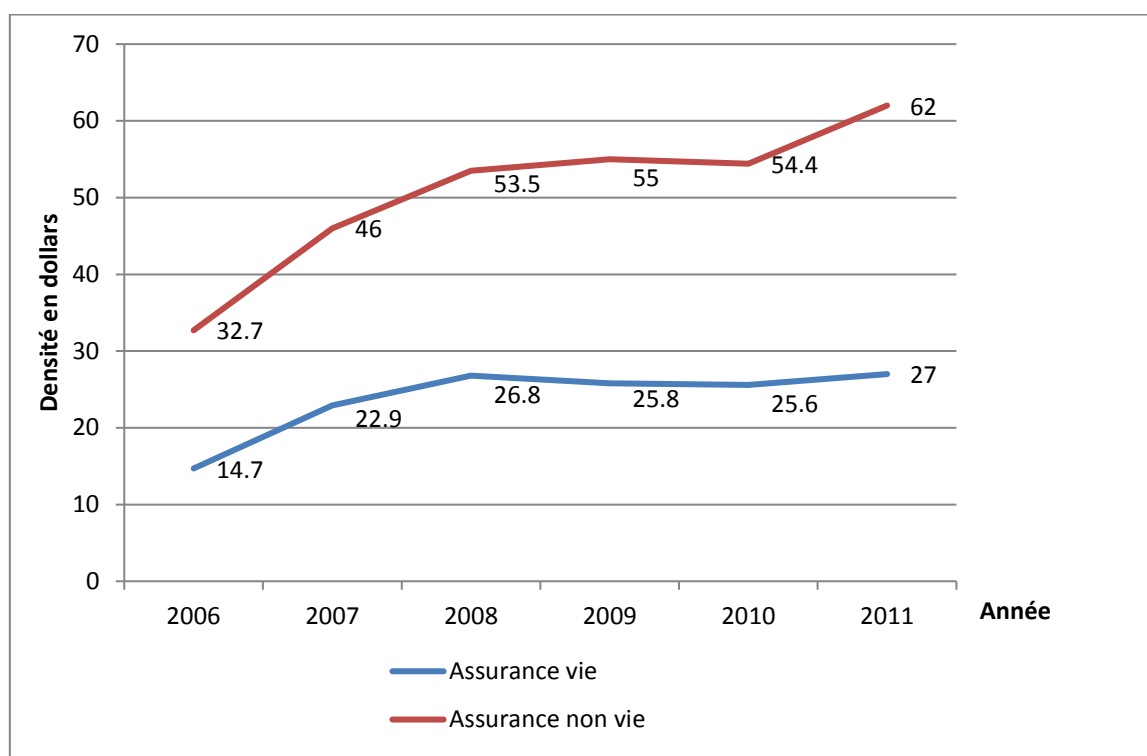
En 2011, le Maroc occupe le 71^{ème} rang mondial et 9^{ème} rang au monde arabe. Dans la même année, les dépenses moyennes annuelles par habitant consacrées à l'achat de produit d'assurances sont estimées à 89 dollars.

L'évolution de la densité d'assurance entre 2006 et 2011, montre que le Maroc a fait une importante avancée, passant de 47,4 \$US en 2006 à 89 \$US en 2011.

Par ailleurs, il convient de signaler que la densité d'assurance n'est pas recommandée pour les pays ayant un nombre d'habitants assez conséquent, comme la Chine et l'Inde. Ces derniers sont classés respectivement au 73^{ème} et 78^{ème} rang en 2000, alors qu'au niveau des primes émises, ils en occupent le 16^{ème} et le 23^{ème}.

Le graphique 3 décrit l'évolution de la densité d'assurance au Maroc durant la période 2006 – 2011 :

Graphique 3. Densité de l'assurance 2006-2011 en dollar (US)



Source : La fédération marocaine des sociétés d'assurance et de réassurance

II. L'assurance maladie au Maroc

1. Présentation de l'assurance maladie

L'amélioration du niveau de santé constitue une des composantes essentielles de la politique de développement social. Une telle politique vise à garantir la pleine participation des citoyens au développement durable du pays.

A cet effet, l'une des priorités de l'Etat en matière de santé consiste à assurer à toute la population l'égalité et l'équité dans l'accès aux soins. Cette priorité fait l'objet d'un consensus national qui s'inscrit dans la mouvance internationale, car elle représente un instrument efficace de justice sociale et de lutte contre les inégalités.³

L'assurance maladie comme étant une assurance contre les dépenses de soin de santé, a pour objet la couverture de l'assuré en cas de survenance d'une maladie, d'un accident ou d'un accouchement entraînant des frais d'hospitalisation. En Pratique, on distingue plusieurs types de garantie de l'assurance maladie, à savoir :

³ Dahir n° 1-02-296 du 25 rejev 1423 (3 octobre 2002) portant promulgation de la loi n° 65-00 portant code de la couverture médicale de base.

- La consultation et la visite ;
- Les médicaments (frais pharmaceutiques) ;
- Le laboratoire (analyses médicales) et la radiologie ;
- L'hospitalisation ;
- L'optique (Verre et Monture) ;
- Les soins et prothèses dentaires ;
- La maternité.

2. L'assurance maladie obligatoire

2.1. Types d'assurance maladie

Le régime marocain de l'assurance maladie couvre à la fois les salariés du secteur public et ceux du secteur privé. Il assure également les intéressés d'une protection contre les risques de nature maladie. Par ailleurs, nous distinguons deux types d'assurance maladie, à savoir :

* **L'assurance maladie obligatoire(AMO)** : Ce régime a été institué en 2002 par la loi 65-00 portant le code de la couverture médicale, et il est entré en vigueur le 18 août 2005. Cette assurance garantit les prestations qui comprennent généralement les actes cités ci-dessus.

* **L'assurance maladie complémentaire(AMC)**: Elle garantit la couverture des soins exclus de l'AMO, du fait qu'elle est qualifiée d'un complément du taux de remboursement des soins de l'AMO.

2.2. Organismes de gestion

Au Maroc, le régime d'assurance maladie obligatoire est géré par les organismes suivants :

* **La Caisse Nationale de Sécurité Sociale** : dénommée ci-après CNSS, instituée par le dahir portant loi n° 1-72-184 du 15 jomada II 1392 (27 juillet 1972), pour les personnes assujetties au régime de sécurité sociale et leurs ayants droits ainsi que pour les titulaires de pensions du secteur privé.

* **La caisse nationale des organismes de prévoyance sociale(CNOPS)** : dénommée ci-après CNOPS, portant statut de la mutualité, pour les fonctionnaires et agents de l'Etat, des collectivités locales, des établissements publics et des personnes morales de droit public et leurs ayants droit ainsi que pour les titulaires de pensions du secteur public.

* **Les compagnies d'assurance** : Ces organismes gèrent l'assurance maladie obligatoire (selon l'article 114 de la loi 65-00 portant le code de la couverture médicale) et l'assurance maladie complémentaire en faveur des salariés des deux secteurs public et privé.

* **Le régime d'assistance médicale** : Bénéficient des prestations du régime d'assistance médicale dans les conditions fixées par voie réglementaire :

- les personnes qui ne sont assujetties à aucun régime d'assurance maladie obligatoire de base et ne disposant pas de ressources suffisantes pour faire face aux dépenses inhérentes aux prestations médicales visées à l'article 121 ci-dessous ;
- leur(s) conjoint(s) ;
- leurs enfants à charge, non-salariés, âgés de 21 ans au plus et non couverts par une assurance maladie obligatoire de base. Cette limite d'âge peut être prorogée jusqu'à 26 ans en cas de poursuite des études dûment justifiée ;
- leurs enfants handicapés quel que soit leur âge, qui sont dans l'impossibilité totale et permanente de se livrer à une activité rémunérée par suite d'incapacité physique ou mentale⁴.

3. La CNOPS et l'assurance maladie

3.1. La population couverte

Pour la caisse nationale des organismes de prévoyance sociale, la population couverte par l'assurance maladie obligatoire est répartie comme suit :

* **Les actifs** : les fonctionnaires et les agents de l'Etat, des collectivités locales, des établissements publics et des personnes morales de droit public.

* **Les retraités** : les titulaires de pensions du secteur public (y compris les pensions de reversions).

* **Les Ayants droits** : les Handicapés à vie, les enfants à l'âge de 21 ans, les étudiants jusqu'à l'âge de 26 ans et les conjoints.

3.2. Taux de cotisation

Le taux de cotisation pour les affiliés de la CNOPS dépend du statut de ces derniers :

* **Pour les actifs** : Le taux de cotisation est de 5% de l'ensemble des rémunérations, 2.5 % à la charge de l'employeur et 2.5% à la charge du salarié.

* **Pour les retraités** : Le taux de cotisation est de 2,5% du montant global des pensions de base.

Il est à noter que toutes les cotisations doivent atteindre un seuil minimal mensuel de 70 DH et plafonné à 400 DH.

3.3. Panier de soins

La caisse nationale des organismes de prévoyance sociale met à la disposition de leurs affiliés un panier de soins bien diversifié, un panier qui répond aux besoins de la population couverte

⁴ Dahir n° 1-02-296 du 25 rejev 1423 (3 octobre 2002) portant promulgation de la loi n° 65-00 portant code de la couverture médicale de base – Article 116.

en matière de l'assurance maladie. Les types de soins et les prestations couverts par la CNOPS se présentent comme suit :

*** Les types de soins**

- Actes de médecine générale et de spécialités médicales et chirurgicales, actes paramédicaux, de rééducation fonctionnelle et de kinésithérapie délivrés à titre ambulatoire hors médicaments ;
- Soins liés à l'hospitalisation et aux interventions chirurgicales y compris les actes de chirurgie réparatrice et le sang et ses dérivés labiles ;
- Médicaments admis au remboursement ;
- Lunetterie médicale, dispositifs médicaux et implants nécessaires aux actes médicaux et chirurgicaux ;
- Appareils de prothèse et d'orthèse médicales admis au remboursement ;
- Soins bucco ;
- Dentaires ;
- Orthodontie médicalement requise pour les enfants ;

*** Les prestations correspondantes**

- ALD et ALC : 100 % sur la base de la tarification nationale de référence (TNR).
- Hospitalisation dans les hôpitaux publics : 100 % de la TNR ;
- Hospitalisation dans les cliniques privées : 90 % de la TNR ;
- Médicaments : 70 % (Le remboursement des dépenses de médicaments s'effectue sur la base du PPM du médicament générique, lorsqu'il existe) ;
- Soins ambulatoires : 80 % de la TNR ;
- Appareillage, lunetterie, prothèses, orthèses, implants et dispositifs médicaux : sous forme de forfaits prévus dans la TNR.

4. La CNSS et l'assurance maladie

4.1. La population éligible

La caisse nationale de la sécurité sociale couvre les salariés du secteur privé contre les risques de maladie obligatoire par le biais d'un panier de soins diversifié. Les bénéficiaires de cette branche d'assurance maladie sont⁵ :

*** Les actifs** : les salariés du secteur privé.

⁵ Source : Dahir n° 1-02-296 du 25 rejev 1423 (3 octobre 2002) portant promulgation de la loi n° 65-00 portant code de la couverture médicale de base – Article 116.

* **Les retraités** : les titulaires de pensions du secteur privé.

* **Les ayants droits** : les handicapés à vie, l'enfant à l'âge de 21 ans, les étudiants jusqu'à l'âge de 26 ans et les conjointes.

4.2. Taux de cotisation

Le taux de cotisation pour les affiliés de la CNSS dépend du statut de ces derniers :

* **Pour les actifs** : le taux de cotisation est de 4% de l'ensemble des rémunérations réparties en deux, 2% à la charge de l'employeur et 2% à la charge du salarié qui est majoré de 1.5% de l'ensemble de la rémunération brute mensuelle.

* **Pour les retraités** : Le taux de cotisation due par les titulaires de pensions est fixé à 4% sur le montant global des pensions de base.

Contrairement à la CNOPS, les cotisations pour la maladie à la CNSS ne sont pas plafonnées.

4.3. Panier de soins

La caisse nationale de la sécurité sociale met à la disposition à leurs affiliés un panier de soins diversifié. Les types de soins couverts par la CNSS ainsi que les prestations correspondantes sont comme suit :

* Les types de soins

- IALD et ALC ;
- Suivi de la grossesse, l'accouchement et ses suites, les actes médicaux et chirurgicaux ;
- Les médicaments admis au remboursement, le sang et ses dérivés labiles, les actes paramédicaux et les actes de rééducation fonctionnelle et de kinésithérapie ;
- Hospitalisation : l'ensemble des prestations et soins rendus dans ce cadre y compris les actes de chirurgie réparatrice ;
- Soins Ambulatoires.

* Les prestations

En ce qui concerne les prestations qui sont à la charge de la CNSS, ce sont des remboursements qui dépendent de l'établissement de soins choisi par l'assuré. En général, ces prestations sont en fonction d'un taux de remboursement qui oscille autour de 80%. Ci-dessous, nous trouverons les différents taux de remboursement déterminés par la CNSS :

- 90% pour les prestations assurées par les professionnels et établissements de soins du secteur public (TNR) ;
- 70% pour les prestations assurées par les professionnels et établissements de soins du secteur privé (TNR) ;

- 90% pour les ALD et ALC (TNR).

5. Les organismes privés et l'assurance maladie

Les compagnies d'assurance ont le droit d'assurer les salariés du secteur privé, les fonctionnaires de l'état, les titulaires de pensions du secteur privé et public, ainsi que leurs ayants droits qui sont les handicapés à vie, les enfants à l'âge de 21 ans, les étudiants jusqu'à l'âge de 25 ans et leurs conjoints, contre les risques de nature maladie.

Les compagnies d'assurance couvrent tous les types de soins définis ci-dessus, avec des prestations qui varient généralement entre 70% et 90% de la charge réelle déclarée par l'assuré, selon le panier choisi par ce dernier. En contrepartie, l'assuré devra payer une prime dite d'assurance au profit de son assureur, afin que ce dernier couvre l'assuré contre les risques de nature maladie.

CONCLUSION

Le présent chapitre a introduit l'organisme d'accueil dans lequel s'est déroulé notre projet de fin d'étude, ainsi que les enjeux économiques du secteur d'assurance au Maroc. Il a décrit également le cadre général de l'assurance maladie, la réglementation de la branche en question, ainsi que ses spécificités juridiques.

Le chapitre suivant sera dédié à la présentation et l'analyse des données de l'étude, afin de tirer le maximum d'informations au niveau de la population étudiée.

Dans le même sens, le prochain chapitre décrira la répartition de la population en matière de la sinistralité, ainsi que la structure du portefeuille maladie de base.

CHAPITRE 2. SOURCE ET ANALYSE DES DONNEES

CHAPITRE 2. SOURCE ET ANALYSE DES DONNEES

INTRODUCTION

Toute étude actuarielle passe nécessairement par la fiabilisation des données de base et, par conséquent, le traitement des données a été perçu comme un paramètre indispensable de notre présent projet. Il constitue un outil permettant l'efficacité, lorsque les objectifs tracés sont atteints et les résultats obtenus sont qualifiés de cohérents.

L'objectif principal de ce chapitre est de présenter les données de base et les différents processus de détection et de correction des anomalies, ainsi que l'analyse du portefeuille maladie.

Ce chapitre vise aussi à proposer quelques statistiques globales relatives à la répartition de la population étudiée, que ce soit au niveau de la sinistralité ou au niveau de la composition du portefeuille. En effet, les études à vocation actuarielle demande une connaissance primaire de la composition du portefeuille étudié, ainsi que la distribution des individus qui composent la base de données.

I. Présentation du portefeuille

1. Fichiers reçus

Dans le but d'établir une tarification en assurance maladie, nous devons préparer et analyser la qualité des données. La population étudiée est celle des contrats d'assurance maladie de base pour des groupes d'assurés de la compagnie SANAD, ces contrats concernent la période 2010 – 2014.

Compte tenu de la spécificité du sujet abordé par notre PFE, il a fallu disposer d'informations assez complètes sur la population étudiée. Cette façon de faire nous a conduit à décomposer le portefeuille en plusieurs fichiers, à savoir :

- **Fichier sinistre** : Ce fichier décrit la sinistralité de la population étudiée en matière du règlement effectué par l'assureur envers chaque assuré.
- **Fichier Sexe** : Ce fichier illustre le sexe de l'assuré.
- **Fichier Effectif** : Ce fichier décrit l'effectif assuré de chaque entreprise qui vient de signer un contrat d'assurance maladie de base "groupe".
- **Fichier Masse salariale** : Ce fichier contient la masse salariale déclarée par l'entreprise assurée.
- **Fichier prestations** : Ce fichier chiffré décrit le panier de soins proposé par la compagnie SANAD.

Il convient de signaler que la modélisation de la prime pure de l'assurance maladie de base "groupe", fait intervenir plusieurs acteurs qui sont considérés comme objet de l'assurance. La distribution de ces acteurs en matière de la sinistralité est déterministe pour l'assureur. Soit les acteurs suivants :

- Les adhérents : les assurés qui cotisent normalement à l'assureur ;
- Les ayants droits : les conjoints et les enfants.

2. Traitement des données

Avant de commencer le traitement des données, il faut détecter les anomalies et redresser les données manquantes qui peuvent exister dans le portefeuille. A cet effet, nous avons utilisé la procédure *means* du logiciel SAS qui permet d'avoir une vue globale sur le portefeuille étudié. Dans toute la suite, nous allons citer les anomalies rencontrées ainsi que les corrections établies.

* La variable AGE

Il s'agit pour cette variable de détecter les observations manquantes et les observations non conformes. Afin de corriger ces erreurs, nous avons affecté par sexe et par lien (adhérent, conjoint et enfant) l'âge moyen des individus ayant les mêmes caractéristiques, c'est-à-dire les individus appartenant aux mêmes groupes et ayant un âge non vide.

Pour les enfants, nous avons rencontré certaines erreurs, à savoir : les enfants ayant des âges supérieurs à 25 ans, les assurés de sexe masculin et de lien "enfant" et les enfants ayant un âge négatif. Afin de rétablir ce problème, nous avons commencé par affecter des points à ces observations pour les qualifier comme des données manquantes. Ensuite, nous avons procédé par la méthode "imputation par moyenne" pour trouver l'âge correspondant à ces points particuliers.

* La variable masse salariale

Il s'agit de la masse salariale totale déclarée par l'entreprise assurée, cette variable est discriminante en matière de l'assurance des groupes. Au sein de notre portefeuille, nous avons rencontré un nombre énorme d'observations manquantes pour la variable MS_T. C'est la raison pour laquelle, nous étions obligé de procéder par un traitement spécial de cette variable. La méthodologie adoptée pour corriger les valeurs manquantes se présente comme suit :

Tout d'abord, la masse salariale totale d'une telle entreprise est la somme des salaires individuels des employés, soit la formule suivante :

$$\sum_{i=1}^N S_{ij} = MS\ totale_j$$

Avec $S_{i,j}$ est le salaire de l'employé i de l'entreprise j assurée, N est le nombre des salariés de l'entreprise en question.

Par ailleurs, le salaire moyen de l'entreprise assurée est le rapport de la somme des salaires individuels par la taille de l'entreprise.

Autrement dit, la masse salariale totale d'une telle entreprise est le produit de l'effectif assuré et le salaire moyen de la même entreprise. Cette formule, nous a permis de trouver la masse salariale correspondante à chaque entreprise marquée d'une valeur manquante au niveau de la variable MS-T. L'affectation de la masse salariale était basée alors sur l'effectif assuré et le salaire moyen annuel. Cette méthodologie était réalisée grâce à l'étape *DATA* du logiciel SAS.(Voir Annexe)

* La variable Exposition

Les contrats d'assurance maladie sont annuels et, par conséquent, l'exposition au risque n'est que la somme des jours d'assurance de l'adhérent divisée par 365 jours. Ce rapport est généralement proche de 1, et il est considéré comme un paramètre indispensable pour le calcul de la prime pure des assurés.

Le calcul de la variable Exposition, nous a demandé d'établir un programme sous *Excel* assez compliqué, du fait qu'il s'agit de déterminer plusieurs paramètres avant de procéder au calcul direct de cette variable. La démarche suivie pour déterminer l'exposition au risque de chaque assuré de la branche maladie est comme suit :

Pour chaque assuré i , l'exposition au risque est définie par la relation suivante :

$$Exposition_i = Nbr\ jours\ d'assurance_i / 365 \quad (1)$$

Où i est l'indice de l'assuré dans le portefeuille d'assurance maladie.

A partir de la formule (1), nous remarquons que l'aléa réside dans le paramètre nombre de jours d'assurance de chaque individu. C'est la raison pour laquelle, nous avons distingué tous les cas possibles.

Cas 1



Si la date de sortie est incluse dans l'exercice d'assurance, le nombre de jours d'assurance se présente comme suit : $Nbr\ jours\ d'assurance_i = (date\ de\ sortie - Exercice)_{jours}$

Dans le cas contraire, c'est-à-dire si la date de sortie n'est pas incluse dans l'exercice, le nombre de jours d'assurance est égale à 365, et l'exposition vaut 1.

Cas 2

Le deuxième cas concerne les adhérents qui s'assurent au cours de l'exercice d'étude.



Dans ce cas, le *Nombre jours d'assurance*_{*i*} = (*date d'adhésion* – *Exercice*)_{jours} si la date de sortie est incluse dans l'exercice d'assurance.

Si non, c'est-à-dire si la date de sortie n'est pas incluse dans l'exercice, le *Nombre jours d'assurance*_{*i*} = 0 et l'exposition vaut 0.

Cas 3



Dans ce dernier cas, nous remarquons que la date de sortie est strictement inférieure à l'exercice d'assurance, et par conséquent l'exposition vaut 0 puisque l'adhérent n'est assuré pendant l'exercice d'étude.

3. Agrégation des données

Afin de rendre nos fichiers exploitables pour la modélisation de la prime pure, nous avons acheminé quelques traitements sur tous les fichiers précédemment cités. Ces traitements consistent d'abord à agréger les données pour optimiser le temps de calcul. En suite, effectuer des études actuarielles afin d'évaluer la pertinence du traitement établi. Il est à noter que le fichier de base auquel nous avons attribué les variables manquantes est le fichier *sinistre*.

3.1 Attribution de la masse salariale

La clé d'attribution de la masse salariale est basée sur :

- La police d'assurance ;
- Le code succursale ;
- L'exercice d'assurance.

En utilisant ainsi ces références, nous avons affecté la masse salariale à chaque entreprise assurée.

3.2 Attribution de l'exposition

La clé d'agrégation pour ce fichier est basée sur : la police d'assurance, le code succursale, l'exercice d'assurance et le code de l'individu assuré.

La prise en compte de l'exposition au risque lors de la modélisation de la prime pure est indispensable, du fait que la précision de l'estimation du tarif repose sur la durée d'assurance.

3.3 Agrégation de la sinistralité

Finalement, nous avons agrégé pour chaque assuré le nombre de sinistre enregistré, le règlement et la charge réelle selon les critères suivants⁶ : l'exercice d'assurance, la police, le code succursale, le code de l'individu et le poste.

II. Analyse des données

1. Description générale de la base de données

Les données utilisées proviennent de la branche maladie de base de l'entité marocaine SANAD du groupe HOLMARCOM. La base de données à l'origine du modèle de tarification, est obtenue par la fusion des tables auparavant citées, et comportant l'ensemble des polices présentes dans le portefeuille durant la période 2010 – 2014. La table de tarification ainsi obtenue comporte les champs suivants :

Tableau 2. Description des champs de la table de tarification

Champs	Description
Exercice	Année d'assurance
POLICE	Numéro de police
CodeSuc	Numéro de la filiale
REGION	Localisation de l'entreprise
INDIVIDU	Code de l'assuré
POSTE	Prestation maladie
NB_SIN	Nombre de sinistres
AGE	Âge de l'assuré
CHARGE	Charge des sinistres
LIEN	Lien entre l'assuré et l'adhérent
SEXE	Sexe de l'assuré
EFFECTIF	Taille de l'entreprise
Exposition	Exposition temporelle
MS_T	Masse salariale assurée
Tauxremb	Taux de remboursement
PLAFGEN	Plafond général
PLAFPROT	Plafond prothèse
PLAFODF	Plafond ODF
PLAFMONT	Plafond monture
VALEDENT	Valeur de la dent

⁶ L'agrégation était réalisée grâce à la procédure SQL du logiciel SAS (Voir Annexe)

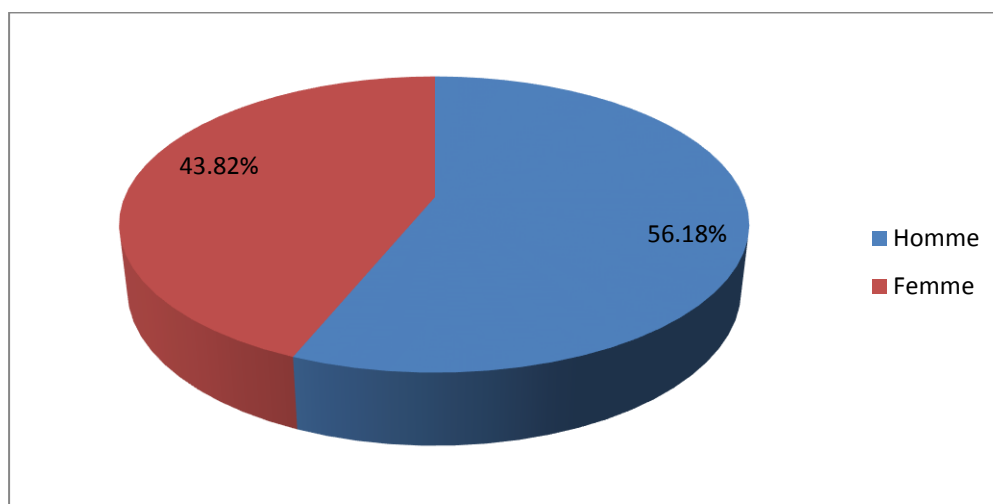
Le portefeuille est composé de 3650 polices au 31 décembre 2014 pour un montant de sinistres survenus en 2014 atteint plus de 121 millions dirhams. L'occurrence des sinistres⁷ est de 1 pour 9 282 DH de salaire assuré.

2. Distribution de la population

2.1 Répartition du portefeuille selon le sexe

La variable sexe désigne le genre du bénéficiaire (homme ou femme). Cette variable permet de mesurer la proportion des hommes dans le portefeuille étudié, ainsi que la proportion des femmes. Elle permet également d'étudier l'évolution de la sinistralité des assurés selon leurs sexes, afin d'examiner l'impact du genre de l'assuré sur la sinistralité enregistrée durant l'exercice d'assurance. Ci-dessous, nous trouverons la proportion des hommes et des femmes dans le portefeuille étudié :

Graphique 4. Répartition des assurés selon le sexe



Source : Elaboré à partir du portefeuille maladie

Nous remarquons que les hommes sont assez présents que les femmes. Il s'agit de 56.18% des assurés de sexe masculin contre 43.82% de sexe féminin.

2.2 Répartition du portefeuille selon le lien

En assurance maladie, la variable lien signifie le statut du bénéficiaire au sein du portefeuille. En effet, on distingue trois types d'assurés :

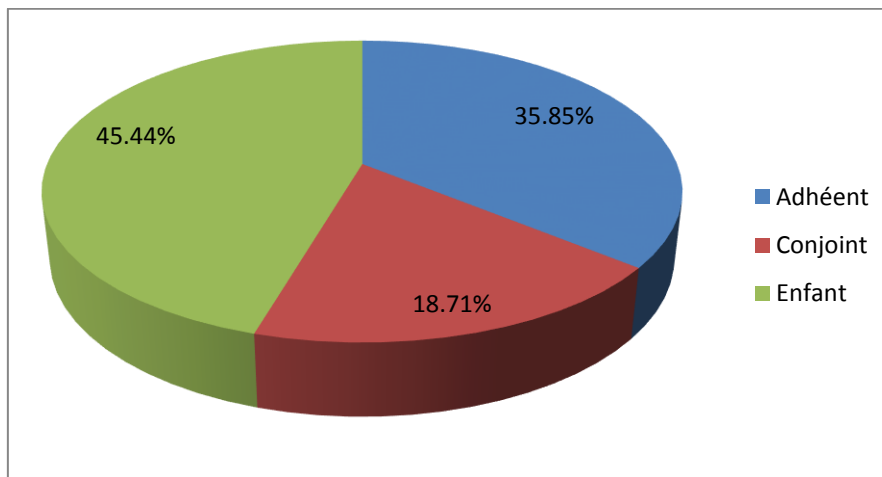
- Adhérent : c'est l'assuré lui-même
- Conjoint : c'est le conjoint de l'adhérent
- Enfants : ce sont les enfants de l'adhérent

⁷ Rapport du nombre de sinistres à la masse salariale déclarée par les assurés

Il est à noter que la variable LIEN est une variable discriminante en assurance maladie de base, c'est une variable qui permet de segmenter la population étudiée par tête assurée. Ainsi, l'étude de la sinistralité des assurés sera établie par type de bénéficiaire (adhérent, conjoint et enfants). Autrement dit, il faut estimer les paramètres de chaque tête bénéficiaire du produit assurance maladie.

Le graphique 5 représente la proportion des bénéficiaires du produit assurance maladie de la population étudiée.

Graphique 5. Répartition des assurés selon le lien



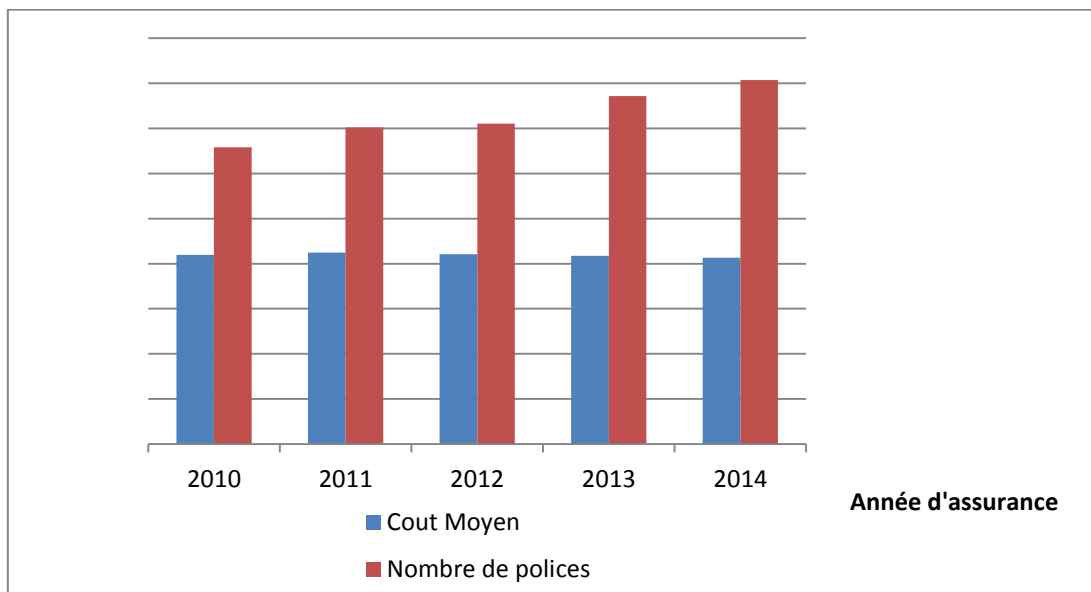
Source : Elaboré à partir du portefeuille maladie

2.3 Sinistralité du portefeuille

Les axes des graphiques figurant dans cette section ont été modifiés afin de protéger la confidentialité des données.

Le graphique 6 représente l'évolution de la sinistralité durant la période 2010 – 2014 :

Graphique 6. Evolution de la sinistralité sur la période 2010 – 2014

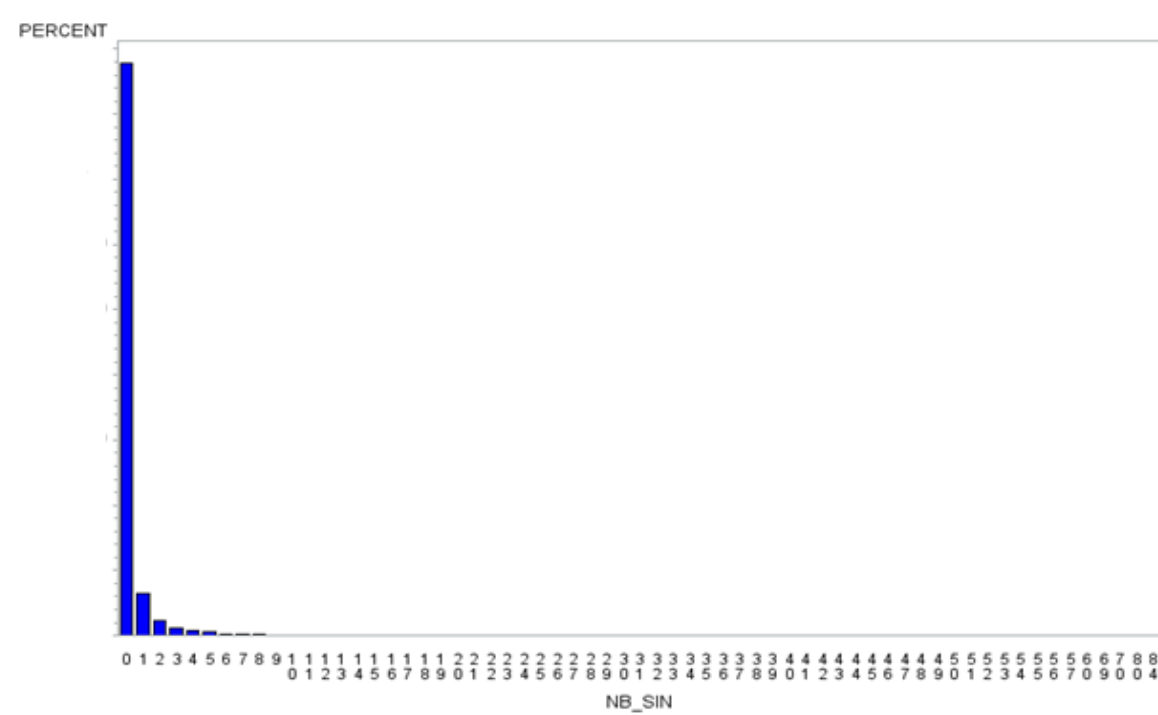


Les données de ce graphique correspondent aux sinistres vus sur la période 2010 – 2014. Cependant, ces chiffres n'intègrent pas les sinistres tardifs. L'occurrence des sinistres augmente de 15% et le coût moyen des sinistres diminue de 1,62% sur cette période. Ces chiffres témoignent d'une tendance marquée par la diminution des coûts moyens et l'augmentation des nombres de sinistres. Ce décalage entre l'occurrence des sinistres et le coût moyen, s'explique par l'apparition d'un nombre important de sinistres ayant des coûts faibles. Le coût des sinistres est composé majoritairement des frais de soins.

Concernant la distribution du nombre de sinistres, nous trouvons qu'il s'agit d'un excès en zéro sinistre, du fait que 80% des assurés de la population n'ont enregistré aucun sinistre. Ce phénomène (l'excès en zéro sinistre) est qualifié d'un événement classique en assurance maladie, et ceci s'explique par la nature de la branche étudiée. En effet, l'assurance maladie de base au Maroc est une assurance obligatoire et, par conséquent, si l'adhérent et son conjoint sont employés chez deux entreprises différentes, ils seront affiliés à deux régimes d'assurance différents les uns des autres.

Dans ce cas particulier, les enfants sont couverts par le régime d'assurance du père, alors que le conjoint (ici la mère) est couvert tout seul par l'assureur adverse. D'où l'absence de la déclaration d'un nombre important de sinistres enregistrés par les enfants et le père en faveur de l'assureur de la mère, et vise vers ça. Ci-dessous, nous trouverons la distribution du nombre de sinistres de la population étudiée :

Figure 1. Distribution de la fréquence des sinistres



Source : Elaboré à partir du portefeuille maladie

La figure 1 montre qu'environ 80% des assurés n'ont enregistré aucun sinistre durant la période d'étude. Cette masse de zéros sinistres est expliquée par la nature de la branche d'assurance étudiée.

3. Analyse de la sinistralité

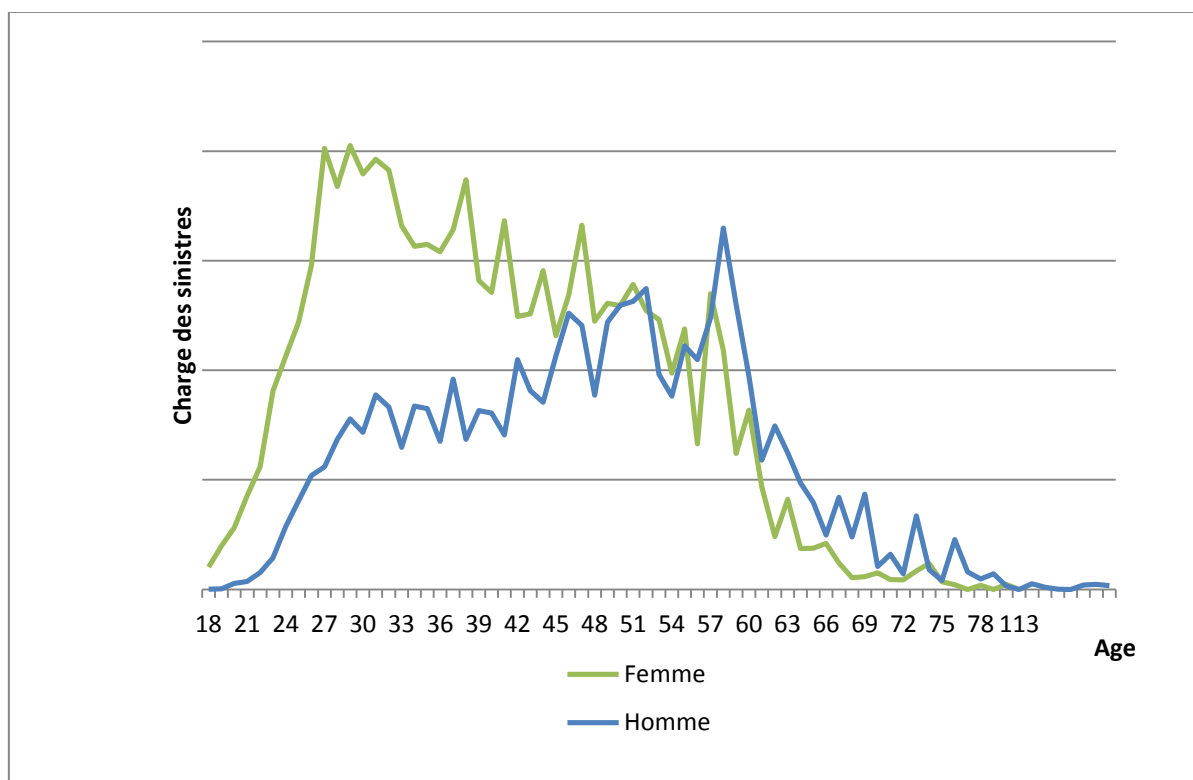
3.1. Selon l'âge et le sexe de l'assuré

* Nombre de sinistres

La charge des sinistres sera calculée par sexe du bénéficiaire et par âge. Dans ce cas, la variable âge sert à jouer le rôle de l'indicateur temporel.

Les bénéficiaires de l'assurance maladie de base ont des profils de risque variés selon le sexe. Nous effectuons une analyse univariée de ce facteur afin de distinguer les différences entre les types de bénéficiaires.

Graphique 7. Variation de la sinistralité suivant le type de bénéficiaire



Source : Elaboré à partir du portefeuille maladie

L'axe vertical correspond à la charge des sinistres, alors que l'axe horizontal correspond à l'âge de l'assuré. La courbe bleue représente la charge des sinistres chez les hommes, et la courbe verte représente la charge chez les femmes. D'après le graphique 7, nous remarquons que les femmes consomment beaucoup plus le produit assurance maladie que les hommes avant l'âge de 52 ans. Mais à partir de l'âge de 52 ans, nous trouvons que la consommation

médicale devient importante pour les hommes que pour les femmes. Cependant, il sera très utile d’avoir une idée à propos de la distribution du nombre de sinistres chez les assurés, ainsi que l’évolution de la quantité des sinistres de la population étudiée.

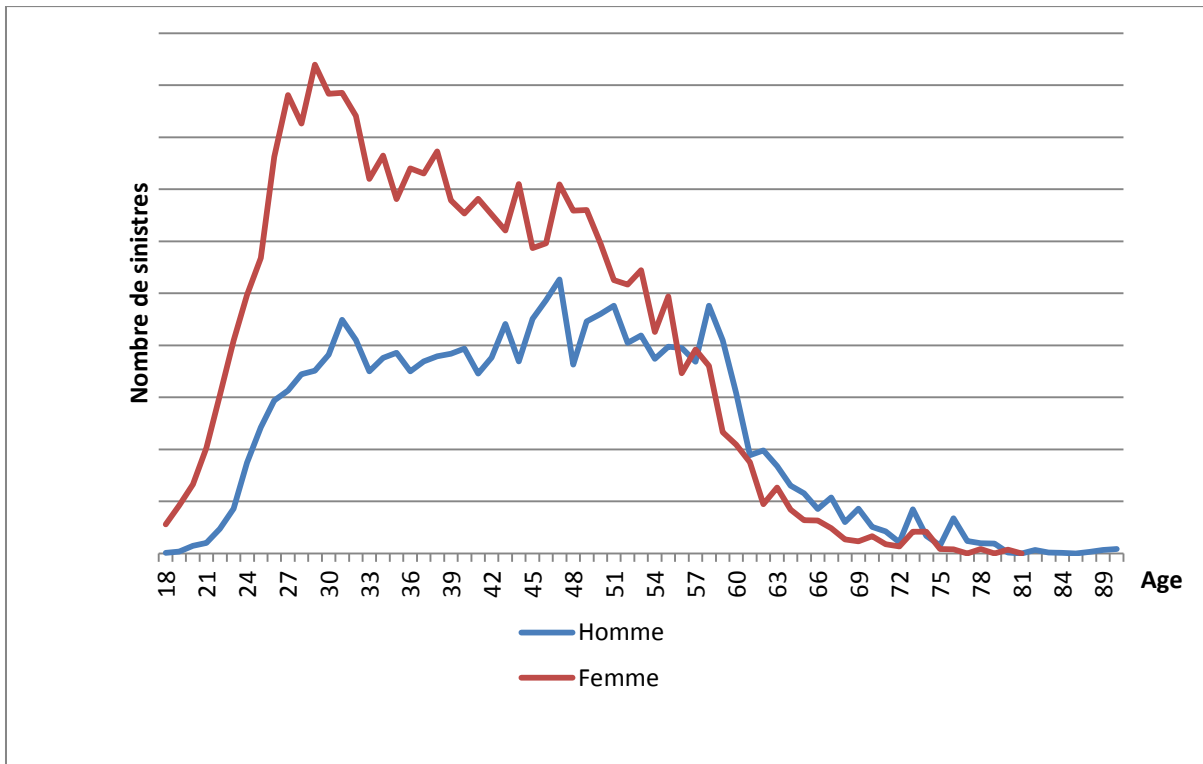
*** Nombre de sinistres**

En ce qui concerne le nombre de sinistres enregistré par les deux types de bénéficiaires (hommes et femmes), nous trouvons qu’il s’agit d’une confirmation de l’hypothèse précédemment citée. Les femmes sont plus sinistrées que les hommes.

Ce dernier constat peut être perçu comme un argument qui nous permet de procéder par type de bénéficiaire dans le processus de tarification en assurance maladie de base. En effet, nous allons remarquer dans le graphique suivant, que la variation de la sinistralité entre les hommes et les femmes est très significative. Il s’agit également d’une consommation importante du produit assurance maladie par les femmes, et ceci nous amène à prendre en considération le sexe de l’assuré comme un paramètre de la tarification maladie.

Ci-dessous, nous trouverons le nombre de sinistres total déclaré par les assurés selon le sexe et l’âge du bénéficiaire :

Graphique 8. Distribution du nombre de sinistre selon l’âge

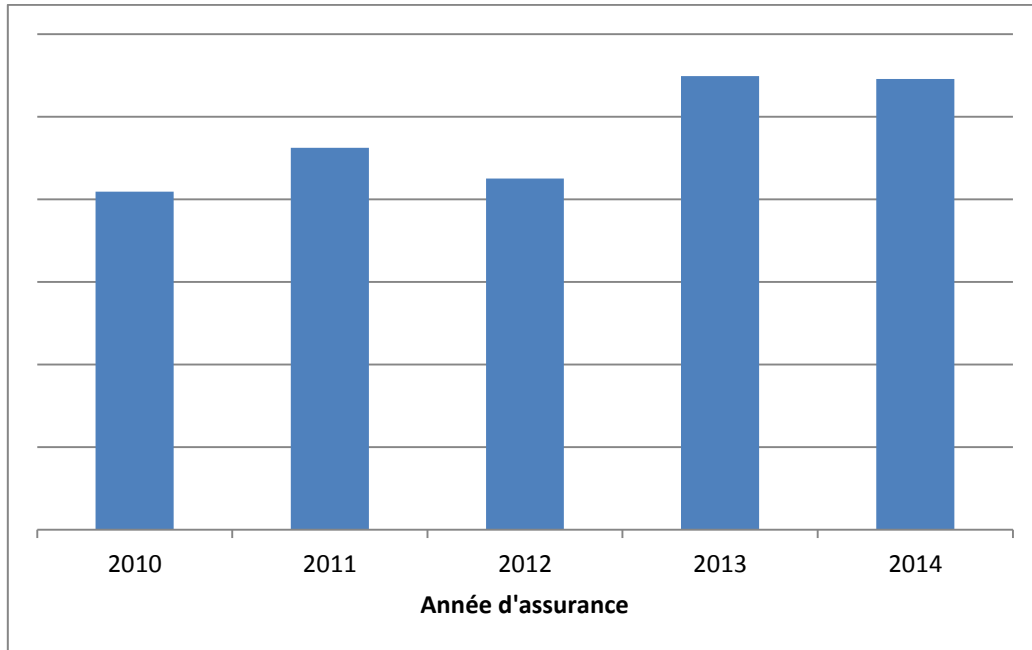


Source : Elaboré à partir du portefeuille maladie

3.2 Selon la masse salariale de l'entreprise assurée

D'une part, la masse salariale déclarée est fortement corrélée au nombre d'employés de l'entreprise et donc à l'exposition au risque de l'assurance maladie. D'autre part, pour mesurer la variation de la sinistralité des entreprises assurées selon la masse salariale, nous aurons besoins d'étudier l'évolution du coût moyen des sinistres selon le salaire moyen déclaré par les entreprises assurées. D'où l'intérêt des graphiques suivants :

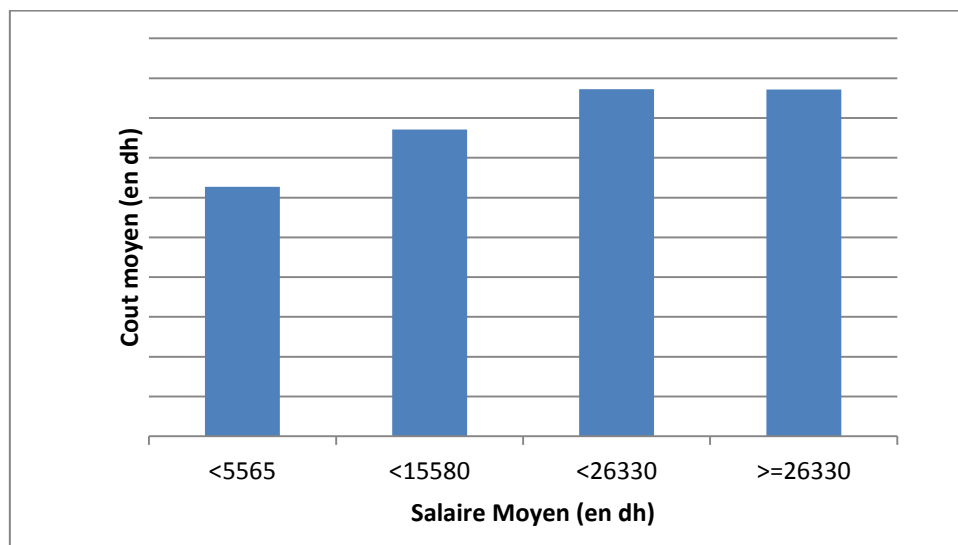
Graphique 9. Evolution de la masse salariale assurée



Source : Elaboré à partir du portefeuille maladie

Observons désormais l'analyse de la masse salariale sur le portefeuille.

Graphique 10. Variation de la sinistralité suivant le salaire moyen déclaré



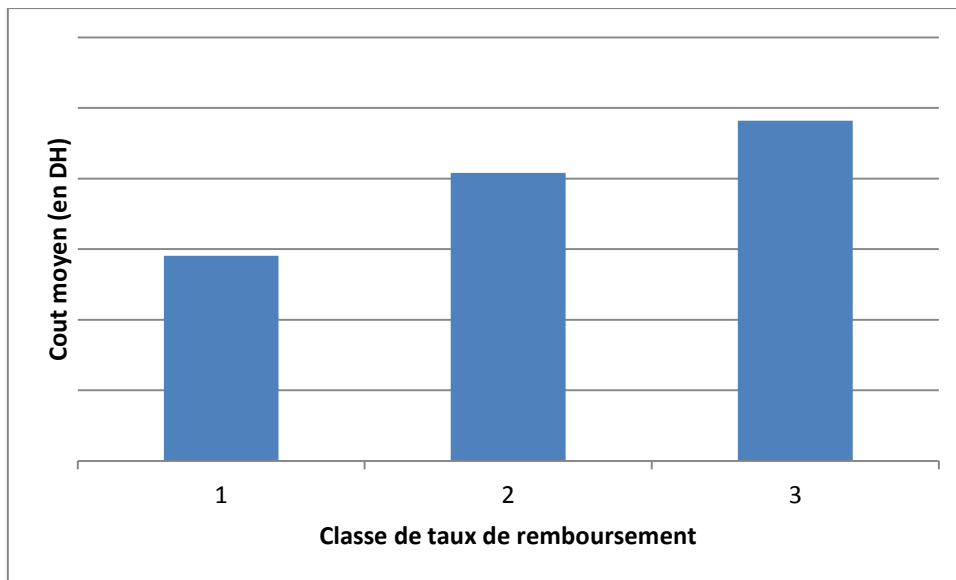
Source : Elaboré à partir du portefeuille maladie

Le graphique 10, représente l'évolution du coût moyen suivant la classe du salaire moyen de l'entreprise assurée. Il paraît que les deux variables en question sont corrélées positivement. En effet, plus qu'une entreprise déclare un salaire moyen grand, plus que le coût moyen de la même entreprise est important.

3.3 Selon le taux de remboursement

Le taux de remboursement est un paramètre contractuel et annuel, qui oscille généralement autour de 80%. Ce paramètre détermine le règlement de chaque sinistre déclaré par un tel assuré. Cependant, le coût moyen des sinistres et le taux de remboursement varient dans le même sens.

Graphique 11. Distribution de la sinistralité selon le taux de remboursement



Source : *Elaboré à partir du portefeuille maladie*

CONCLUSION

Ce chapitre a été consacré à la présentation des données de base et les différents processus de détection et de correction des anomalies, ainsi que l'analyse du portefeuille maladie. Il propose également quelques statistiques globales sur la répartition de la population étudiée, que ce soit au niveau de la sinistralité ou au niveau de la composition du portefeuille.

Le chapitre suivant sera dédié au cadre conceptuel et théorique, dans lequel nous allons définir les différentes approches adoptées pour modéliser la prime pure de la branche assurance maladie. En effet, l'élaboration du tarif maladie nécessite un processus de modélisation approprié, qui s'adapte avec la complexité de la branche étudiée.

CHAPITRE 3. CADRE CONCEPTUEL ET THEORIQUE

CHAPITRE 3. CADRE CONCEPTUEL ET THEORIQUE

INTRODUCTION

L'activité d'assurance est caractérisée par un cycle de production inversé : en contrepartie d'une prime d'un montant connu à la souscription du contrat, l'assureur s'engage à couvrir un risque de montant inconnu dont il ignore la date de réalisation. La tarification de l'offre d'assurance consiste à évaluer la prime nécessaire pour couvrir les engagements et les frais de fonctionnement de l'assureur. Le prix de l'assurance est appelée prime commerciale, qui est en fonction de la prime pure et des frais de fonctionnement.

La prime pure est la part de la prime commerciale qui couvre les engagements de l'assureur vis à vis les assurés. Ce paramètre représente également le coût futur des risques et, par conséquent, l'assureur est invité à calculer la prime pure correspondante à chaque groupe d'assurés. C'est la raison pour laquelle, l'assureur est amené à segmenter son portefeuille afin de répartir ses assurés en groupes homogènes. En effet, dans un marché fluide, si deux compagnies identiques ont les mêmes offres, les même frais et la même distribution, la compagnie la moins segmentée court le risque d'anti sélection.

Cependant, l'actuaire cherche à définir des classes de risques homogènes, c'est-à-dire ayant le même cout du risque (prime pure). Il dispose de deux grandes classes de variables : les variables exogènes qui représentent les informations relatives au risque à l'exclusion de toute donnée relative aux réalisations du risque, et les variables endogènes qui représentent les informations permettant la réalisation du risque.

En assurance maladie groupe, les variables exogènes sont liées aux critères de l'entreprise assurée (taille de l'entreprise, localisation de l'entreprise, branche d'activité, salaire moyen, plafond général...) et aux critères des employés et leurs ayant droit (âge individuel, sexe, lien, taux de remboursement, plafond contractuel...). Par contre, les variables endogènes sont liées à la réalisation du risque assuré : le nombre de sinistres par garantie, le coût des sinistres...

Le choix de la méthode de segmentation et des variables tarifaires, repose à la fois sur des objectifs commerciaux et sur des méthodes statistiques. Dans ce présent chapitre, nous allons commencer par sélectionner les variables de tarification en utilisant un algorithme de sélection par optimisation. Ensuite, nous allons exposer une approche statistique de segmentation appelé algorithme de CHAID. C'est un outil permettant la classification hiérarchique des assurés de la branche assurance maladie.

Il est à signaler que ce chapitre vise à proposer une méthode de tarification pour les assurés de la branche maladie. Cette méthode intègre plusieurs facteurs de tarification, et introduit les modèles linéaires généralisés comme étant une approche statistique qui consiste à modéliser

la prime pure de l'assurance maladie. Ensuite, nous allons exposer la théorie des valeurs extrêmes comme étant une approche statistique qui répond au problème de la survenance des événements rares. En effet, la branche d'assurance maladie est caractérisée par la présence de sinistres graves (rares) dans certaines classes, ce qui entraîne une perturbation de l'hypothèse d'homogénéité des classes et de stabilité des indicateurs de risque comme la prime pure. C'est la raison pour laquelle, l'introduction de cette partie dans le processus de tarification est indispensable, afin que les résultats soient cohérents.

I. Choix de modèle par sélection de variables

1. Pourquoi la sélection de variables

La sélection de variables est une étape clé de la modélisation. Dans les études réelles, nous sommes confrontés à des bases de données avec un nombre considérable de descripteurs. Ce sont autant de variables explicatives potentielles. Certaines d'entre elles sont redondantes, d'autres n'ont aucun rapport avec la variable dépendante. La méthode statistique doit nous donner des indications sur le sous-ensemble des bonnes variables à inclure dans le modèle. Dans l'idéal, elles devraient être fortement liées avec la variable dépendante.

Certains auteurs encensent la sélection automatique de variables parce qu'elle constitue un outil fort utile pour une première approche sur des données que l'on ne connaît pas très bien. D'autres par contre, la critiquent vertement car elle nous rend dépendante des fluctuations aléatoires dans les données, d'un échantillon à l'autre nous sommes susceptibles d'obtenir des solutions. Il reste qu'elle est précieuse lorsque la qualité de prédiction est l'objectif principal ou lorsque nous sommes dans un contexte exploratoire. Même si l'expert du domaine a une certaine idée des explicatives à retenir, une sélection automatique peut l'aiguiller sur les pistes à étudier.

2. Critères de sélection de variables

Une variable quantitative Y dite à expliquer (ou encore réponse) est mise en relation avec p variables explicatives X_1, \dots, X_p (ou encore indépendantes). Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n .

Cependant, la détermination des variables explicatives nécessite un processus de sélection de variables approprié. Dans la littérature des modèles linéaires, plusieurs critères de choix de modèle sont présentés à savoir : les critères AIC et BIC, le coefficient de détermination, la statistique de Fisher...

Dans notre cas, nous allons se baser sur le F partiel de Fisher comme étant un critère de sélection de variables dites de tarification. Le paragraphe suivant présente la procédure de sélection ainsi que le critère utilisé pour détecter les variables explicatives.

3. La procédure stepwise

Cette procédure est un mix des approches backward et forward. A la première étape, nous commençons par construire le meilleur modèle à une exogène, c'est-à-dire le descripteur le plus significatif (ayant la statistique du Fisher la plus élevée). Par la suite, et à chaque étape, nous regardons si l'ajout d'une variable ne provoque pas le retrait d'une autre. Cela est possible lorsqu'une variable exogène expulse une autre variable qui lui est corrélée, et qui semblait pourtant plus significative dans les étapes précédentes.

Généralement, nous fixons un risque plus exigeant pour la sélection (ex. 5%, nous ne faisons entrer la meilleure variable que si elle est significative à 5%) que pour la suppression (ex. 10%, nous supprimons la variable la moins pertinente si elle est non significative à 10%).

II. Segmentation et codification des variables

1. Arbre de régression

Les arbres de régression sont des outils non paramétriques de segmentation. Dans un arbre de décision, on cherche à détecter des critères permettant de répartir les individus en 2 classes, caractérisées par $Y = 0$ et $Y = 1$. On commence par choisir la variable, qui, par ses modalités, sépare le mieux les individus de chacune des classes. On constitue alors un premier nœud. On réintère alors la procédure sur chaque nouveau nœud. Dans la méthode CART (Classification And Regression Tree), on regarde toutes les possibilités. On continue soit jusqu'à ce qu'il ne reste plus qu'un seul individu dans chaque nœud, soit suivant un critère d'arrêt. Les critères de discrimination et de constitution des nœuds sont généralement les suivants :

- lorsque les variables explicatives X_j sont qualitatives, ou discrètes, on utilise la distance du Chi-deux (on parle d'arbre CHAID) ;
- en présence de variables de tous types, on peut utiliser l'indice de Gini (méthode CART) ;

Pour une variable qui prend des valeurs continues, on distinguera les deux ensembles $\{X_i \leq s\}$ et $\{X_i > s\}$. Par contre, pour une variable qualitative, on distinguera les ensembles $\{X_i = s\}$ et $\{X_i \neq s\}$. Pour chacune des variables, on regarde l'ensemble des classifications possibles.

2. L'algorithme de CHAID

L'acronyme CHAID signifie "Chi-squared Automatic Interaction Detector". Il s'agit de l'une des méthodes d'arbres de classification les plus anciennes. Cet algorithme était proposé par Kass (1980), mais la version actuelle n'est qu'une modification de l'algorithme THAID qui a été développé par Morgan et Messenger en 1973. Cette approche statistique est très utilisée en assurance pour classer les assurés en groupes homogènes en matière de la sinistralité. En effet, l'assureur cherche à détecter des groupes d'assurés dans lesquels la sinistralité est

considérée à la fois intra homogène (à l'intérieur de chaque groupe) et inter-hétérogène (entre les groupes).

3. La méthode de codification

L'algorithme de segmentation CHAID mesure la liaison qui peut exister entre la variable dépendante Y et la variable explicative considérée. Son principe réside dans le fait de chercher les classes qui engendrent la meilleure partition possible, vis-à-vis de la variable expliquée. Cette partition est le résultat à la fois d'une réduction de la variance intragroupes et d'une maximisation de la variance intergroupes.

Le test statistique utilisé pour mesurer le degré de liaison entre la variable dépendante Y et les descripteurs X_i , se présente comme suit :

Sur une population P de taille n, sont définies K variables explicatives X_1, \dots, X_k . La moyenne générale de Y sur P est notée μ et sa variance est notée σ^2 , soit $P(\mu, \sigma^2)$. Notre population sera divisée en deux sous populations $P_1(\mu_1, \sigma_1^2)$ et $P_2(\mu_2, \sigma_2^2)$. On note par X^j la j^{ème} variable explicative, soit $X^j = [X_1^j, \dots, X_n^j]$.

$$\begin{aligned} \text{Nous avons alors} \quad \mu &= \frac{1}{n} \sum_i \sum_j X_i^j \\ &= \frac{n_1}{n} \mu_1 + \frac{n_2}{n} \mu_2 \end{aligned}$$

Où n_1 et n_2 sont les effectifs des deux sous populations.

$$\text{D'autre part } \sigma^2 = \left[\frac{n_1}{n} \sigma_1^2 + \frac{n_2}{n} \sigma_2^2 \right] + \left[\frac{n_1}{n} (\mu_1 - \mu) + \frac{n_2}{n} (\mu_2 - \mu) \right]$$

Et $\sigma^2 = \sigma^2(D) + \sigma^2(E)$ avec $\sigma^2(D)$ est la variance intragroupes et $\sigma^2(E)$ désigne la variance intergroupes.

Finalement, pour séparer la population en deux groupes homogènes en terme de la variable réponse, l'algorithme de segmentation CHAID cherche à construire des segments en réalisant les deux conditions suivantes : une variance intragroupes $\sigma^2(D)$ minimale une variance intergroupes $\sigma^2(E)$ maximale.

Il est à noter qu'à chaque étape de l'algorithme, on doit vérifier les hypothèses suivantes :

H_0 : Les moyennes de Y sur chaque segment sont égales ;

H_1 : Les moyennes de Y sur chaque segment sont différentes.

Le test se base sur la statistique suivante :

$$R = \frac{\sigma_{\text{inter-groupes}}^2}{\sigma_{\text{intra-groupes}}^2}$$

La statistique R suit une loi de Fisher de paramètres (1,n-1), n est la taille de la population. Il est à préciser que cette analyse appliquée sur les 2 groupes, peut être généralisée à k groupes.

4. Test de khi-deux d'indépendance

Il arrive en pratique que l'on étudie plusieurs variables simultanément. Dans le cas particulier de deux variables, on peut être amené à vérifier s'il existe un lien entre les deux. Ce dernier postulat est indispensable pour l'algorithme de CHAID, qui nécessite de mesurer la dépendance entre la variable réponse et la variable explicative avant de commencer la segmentation.

La méthode du Khi-deux permet d'effectuer le test d'indépendance entre deux variables, et ceci grâce test suivant :

Soient X et Y deux variables aléatoires.

H_0 : X et Y sont indépendantes ;

H_1 : X et Y sont dépendantes ;

Afin d'effectuer un tel test, on prélève un échantillon de taille n de la population que l'on classe conjointement selon les r modalités de X et les c modalités de Y. On obtient alors un tableau de contingence.

Le principe du test du Khi-deux consiste à comparer les effectifs observés O_{ij} aux effectifs attendus E_{ij} si H_0 est vraie. Si les deux variables sont indépendantes, les effectifs attendus E_{ij} (avec $i = 1, \dots, r$ et $j = 1, \dots, c$) sont calculés à partir du tableau de contingence :

$$E_{ij} = \frac{1}{n} \left(\sum_{k=1}^c O_{ik} \right) \left(\sum_{l=1}^r O_{lj} \right)$$

La statistique du test est

$$\chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ik} - E_{ij})^2}{E_{ij}}$$

Lorsque H_0 est vraie, la statistique χ_0^2 suit une loi de khi-deux à $v = (r - 1) \times (c - 1)$ degrés de liberté. Pour un niveau critique α donné, le test consiste à rejeter H_0 si $\chi_0^2 > \chi_{0,v}^2$.

III. Etude théorique selon l'approche GLM

1. Généralités

Aujourd'hui les modèles linéaires généralisés sont couramment utilisés par les professionnels de l'assurance pour tarifer les branches personnelles. La plupart des assureurs marocains utilisent les MLG pour analyser leur portefeuille. Dans le passé, les actuaires avaient recours aux analyses à un facteur afin de tarifer et contrôler la rentabilité des produits d'assurance.

Une analyse à un facteur regroupe des statistiques telles que la fréquence de sinistres ou encore le ratio de sinistres à primes pour chaque valeur de chaque variable explicative, sans prendre en compte l'effet des autres variables.

Les analyses à un facteur peuvent être déformées par les corrélations entre les facteurs de tarification. Par exemple, une analyse à un facteur de la masse salariale montrera une grande sinistralité pour les grandes entreprises. Cependant cela résulte principalement du fait que les grandes entreprises sont majoritairement caractérisées par une masse salariale importante, et par conséquent elles ont un profil de risque élevé. Les mesures de l'analyse à un facteur sur la masse salariale et la branche d'activité vont alors comptabiliser deux fois l'effet de la masse salariale sur le niveau de risque.

En outre, l'analyse à un facteur ne considère pas les interactions entre facteurs. Ces interdépendances existent lorsque l'effet d'un facteur varie selon le niveau pris par un autre facteur.

Les méthodes multivariées, telles que les modèles linéaires généralisés, ajustent les effets induits par la corrélation et permettent de représenter les phénomènes d'interaction. Dans le même sens, les MLG sont qualifiés d'un outil permettant l'efficacité en assurance, lorsque les hypothèses sont bien respectées. Dans quelle mesure alors les MLG répondent aussi bien aux besoins des assureurs que les modèles linéaires simples ? Et si c'est le cas, doit-on considérer que la transition d'une régression linéaire simple pour modéliser une telle variable aux modèles linéaires généralisés est indispensable en assurance ?

1.1. Des modèles linéaires aux MLG

Les modèles linéaires de la forme $y = X\beta + e$ dans lesquels les éléments de e sont supposés indépendants, identiquement distribués et de loi $N(0; \sigma^2)$ ont longtemps formé la base de l'analyse de données. Cependant, l'évolution de la théorie statistique nous permet désormais d'utiliser des méthodes analogues aux modèles linéaires dans les cas suivants :

1. La variable réponse suit une loi différente de la loi Normale ;
2. La relation entre la réponse et les variables explicatives n'est pas nécessairement linéaire.

Ces modèles faisant appel à la famille des distributions exponentielles sont appelés modèles linéaires généralisés (MLG).

1.2. La famille des distributions exponentielles

Soit Y une variable aléatoire dont la densité de probabilité (Loi de probabilité si la variable est discrète) dépend d'un unique paramètre θ . La distribution appartient à la famille exponentielle si elle peut être écrite sous la forme :

$$f(y; \theta) = \exp(yb(\theta) + c(\theta) + d(y))$$

Où b , c et d sont des fonctions connues.

Nombre de distributions connues appartiennent à la famille exponentielle, telles que les distributions Poisson, Normale, et Binomiale.

Prenons l'exemple d'une loi de Poisson. La fonction de probabilité pour une variable qui suit une loi de Poisson s'écrit :

$$f(y; \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

Celle-ci peut s'écrire sous la forme suivante :

$$f(y; \lambda) = \exp(y \log(\lambda) - \lambda - \log(y!))$$

1.3. Les modèles linéaires généralisés

Soit un échantillon de variables aléatoires indépendantes Y_1, \dots, Y_N dont les distributions appartiennent à la famille exponentielle et prennent toutes la même forme. Nous pouvons écrire la fonction de densité conjointe de Y_1, \dots, Y_N :

$$f(y_1, \dots, y_N; \theta_1, \dots, \theta_N) = \exp \left[\sum_{i=1}^N y_i b(\theta_i) + \sum_{i=1}^N c(\theta_i) + \sum_{i=1}^N d(y_i) \right]$$

Dans le cadre des modèles linéaires généralisés, nous considérons les paramètres β_1, \dots, β_p ($p < N$) tels que la combinaison linéaire des β est fonction de l'espérance de Y_i notée μ_i :

$$g(\mu_i) = x_i^T \beta$$

Où g est une fonction monotone et différentiable appelée fonction de lien, x_i est un vecteur des variables explicatives (p éléments) et β est le vecteur des paramètres (p éléments).

Par conséquent, un modèle linéaire généralisé a trois composants :

1. Des variables de réponse Y_1, \dots, Y_N
2. Un ensemble de paramètres β et de variables explicatives X

$$X = \begin{bmatrix} x_1^T \\ \cdot \\ \cdot \\ \cdot \\ x_p^T \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}$$

3. Une fonction de lien monotone g telle que :

$$g(\mu_i) = x_i^T \beta \quad \text{où} \quad \mu_i = E(Y_i)$$

La fonction de lien logarithmique possède la propriété de donner un effet multiplicatif aux variables. En écrivant $g(x) = \ln(x)$ on obtient :

$$\mu_i = g^{-1}(\beta_1 x_{i_1} + \dots + \beta_p x_{i_p}) = \exp(\beta_1 x_{i_1}) \dots \exp(\beta_p x_{i_p})$$

En d'autres termes, lorsqu'une fonction de lien logarithmique est utilisée, le MLG estime les logarithmes des effets multiplicatifs.

2. Estimation des paramètres

2.1. Maximum de vraisemblance

Soient N variables aléatoires indépendantes $Y_1 \dots Y_N$ avec la fonction de densité conjointe $f(y_1, \dots, y_N; \theta_1, \dots, \theta_p)$ qui dépend des paramètres $\theta_1, \dots, \theta_p$. Dans toute la suite, nous notons les deux composantes y et θ par :

$$y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix} \quad ; \quad \theta = \begin{bmatrix} \theta_1 \\ \cdot \\ \cdot \\ \cdot \\ \theta_p \end{bmatrix}$$

La fonction de vraisemblance correspond à la fonction de densité conjointe. Nous la notons $L(\theta; y)$. Notons Ω l'ensemble des valeurs possibles du vecteur θ . L'estimateur du maximum de vraisemblance de θ est la valeur $\hat{\theta}$ qui maximise la fonction de vraisemblance.

$\hat{\theta}$ est également la valeur qui maximise le logarithme de la fonction de vraisemblance (La fonction logarithmique étant monotone). Il est souvent plus simple de travailler avec le logarithme de la fonction de vraisemblance qu'avec la fonction de vraisemblance elle-même.

Nous obtenons alors l'estimateur $\hat{\theta}$ en dérivant le logarithme de la fonction de vraisemblance par rapport par à chaque élément de θ et en résolvant l'équation suivante :

$$\frac{\partial l(\theta, y)}{\partial \theta_j} = 0 \quad \text{pour } j = 1, \dots, p \quad \text{avec } l(\theta, y) = \log L(\theta, y) .$$

2.2. Paramètres du modèle

Nous souhaitons obtenir les estimateurs des paramètres β qui maximisent la fonction de vraisemblance du MLG. La fonction logarithmique de vraisemblance pour des réponses indépendantes Y_1, \dots, Y_N s'écrit :

$$l(\theta, y) = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)$$

où $E(Y_i) = \mu_i = -c'(\theta_i)/b'(\theta_i)$

et $g(\mu_i) = x_i^T \beta = \omega_i$

La famille des distributions exponentielles satisfait les conditions de régularité qui permettent d'assurer que la solution du système d'équations est l'unique maximum global du logarithme de la fonction de vraisemblance.

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^N \frac{y_i - \mu_i x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \omega_i}$$

Où x_{ij} est le $j^{\text{ème}}$ élément de x_i^T . les équations sont résolues grâce à un algorithme numérique (méthode de Newton Raphson).

Cependant, la convergence de l'algorithme Newton Raphson est une question classique lors de l'estimation des paramètres du modèle. C'est la raison pour laquelle qu'il faut envisager de modifier le modèle en :

- En choisissant une autre distribution de la famille exponentielle (ZIP, ZINB, ...) pour la composante aléatoire du modèle ;
- En ajoutant ou en enlevant des variables explicatives ou des interactions entre ces variables ;
- En changeant la fonction de lien.

3. Choix des facteurs

Dans cette section nous décrivons le processus de choix des facteurs intervenants dans la modélisation. Nous étudions préalablement les distributions des dérivées partielles de la fonction de log-vraisemblance, des estimateurs du maximum de vraisemblance et de la log-vraisemblance afin de déterminer les propriétés de la déviance. Des tests statistiques sur la déviance nous permettent alors de sélectionner les facteurs pertinents.

3.1. Généralités

Si $\hat{\theta}$ est un estimateur consistant du paramètre θ et $\text{var}(\hat{\theta})$ la variance de l'estimateur, alors pour un grand échantillon de données, la statistique

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{var}(\hat{\theta})}}$$

suit une loi $N(0,1)$ et son carré suit une loi du χ^2 à un degré de liberté. La généralisation de ce résultat à p paramètres donne :

$$(\hat{\theta} - \theta)^T V^{-1}(\hat{\theta} - \theta) \sim \chi_p^2$$

où $\hat{\theta}$ est un estimateur convergent de θ , V la matrice de variance covariance de $\hat{\theta}$. $\hat{\theta}$ est asymptotiquement un estimateur sans biais de θ , et V est une matrice inversible.

3.2. Inférence sur les paramètres

Faire de l'inférence statistique signifie émettre des conclusions concernant une population à partir de résultats obtenus sur un échantillon de cette population. Nous avons vu comment ajuster un GLM à des données provenant d'un échantillon. Nous allons maintenant chercher à établir des conclusions sur la population cible à l'étude à partir des estimations calculées pour les coefficients.

En plus d'estimer le vecteur β de paramètres, on calcule une matrice de variance-covariance asymptotique pour le vecteur des estimations obtenues.

Cette matrice est $\hat{\sigma}^2(\hat{\beta}) = \text{var}(\hat{\beta}) = I^{-1}(\hat{\beta})$ où I est la matrice d'information observée dans l'algorithme de Newton-Raphson.

Les erreurs-types des paramètres sont la racine carrée des éléments sur la diagonale de cette matrice.

Ainsi, $\hat{\sigma}(\hat{\beta}_j)$ est la racine carrée de l'élément correspondant à β_j la matrice $\text{var}(\hat{\beta})$.

On peut se servir de ces erreurs-types pour construire des intervalles de confiance de Wald de niveau $(1-\alpha)\%$ pour tous les paramètres du vecteur comme suit :

$$\beta_j \in [\hat{\beta}_j - z_{\alpha/2} \hat{\sigma}(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2} \hat{\sigma}(\hat{\beta}_j)]$$

Où $j = 0, 1, \dots, p$

3.3. Test de Wald sur un paramètre

Tout d'abord, un paramètre j nul signifie que la variable explicative dont il est le coefficient, x_j , n'a pas de lien avec la variable réponse N . Il est donc d'intérêt de tester si la valeur d'un paramètre est nulle dans la population étudiée. Nous pouvons faire ce type d'évaluation grâce au test de Wald pour confronter les hypothèses suivantes :

$H_0 : \beta_j = 0$ (il n'y a pas de lien entre x_j et N)

$H_1 : \beta_j \neq 0$ (il y a un lien entre x_j et N)

Nous utilisons le test de Wald, basé sur la distribution asymptotique de $\hat{\beta}_j$. La statistique du test est $Z = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$.

Sous H_0 , la statistique du test suit approximativement une loi normale $N(0,1)$. Il est à noter que le carré de la statistique du test de Wald suit approximativement une loi de khi-deux à un degré de liberté.

Soit $Z_W^2 = \left(\frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}\right)^2 \sim \chi_{(1;1-\alpha)}^2$ où α est le seuil de signification du test.

En général, nous prenons $\alpha=0.05$, c'est-à-dire si le p-value observé est inférieur de ce seuil, nous rejetons l'hypothèse nulle du test. Dans ce cas, et seulement dans ce cas, la variable explicative x_j est significative (il y a un lien entre x_j et la variable réponse).

3.4. Prédiction de la variable réponse

Une valeur prédite par le modèle est définie par

$$\hat{\lambda}_j = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_p x_{i,p})$$

Donc, pour une observation n_i associée au vecteur de variables explicatives x_i , sa valeur prédite par le modèle ayant le logarithme de fonction de lien est

$$\hat{n}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_p x_{i,p})$$

Ainsi, nous prédisons la variable réponse N par son espérance, qui est elle fonction des valeurs prises par les variables explicatives.

Nous pouvons aussi prédire la valeur moyenne de N pour les valeurs de x non observées dans l'échantillon étudié. D'où l'importance des modèles linéaires généralisés dans ce type de traitement.

4. Validation du modèle

L'étape de la validation du modèle est qualifiée de l'une des parties les plus importantes des modèles linéaires généralisés. En effet, il est indispensable de s'assurer que le modèle considéré s'ajuste bien aux données avant de pratiquer l'inférence statistique. A cet effet, nous utilisons la déviance comme statistique de test qui permet de valider un modèle linéaire généralisé.

4.1. La déviance

La déviance D est l'écart en terme de log-vraisemblance entre le modèle saturé d'ajustement maximum et le modèle considéré.

$$D = 2 \sum_{i=1}^m \left(I_{\text{saturé}}(n_i) - I(\hat{\lambda}_i) \right)$$

Si le modèle était parfait (modèle saturé), $\hat{\lambda}_j$ serait égale à n_i ($E(N_i|x_i) = \lambda_i = n_i$).

Un modèle présente un bon ajustement aux données (Goodness-of-fit), si les valeurs prédites par le modèle sont similaires aux valeurs observées. Le modèle saturé présente donc un ajustement parfait puisque ses valeurs prédites sont toutes égales aux valeurs observées.

Un tel modèle n'est pas désirable puisqu'il comporte trop de paramètres, il est trop complexe. De plus, trop coller aux données n'est pas toujours une bonne chose puisque celles-ci proviennent d'un échantillon aléatoire et non de la population complète. On désire un modèle parcimonieux qui présente des relations théoriques interprétables.

La valeur de la déviance obtenue en se basant sur le modèle considéré, permet de tirer la qualité de l'adéquation de ce modèle aux données de l'étude. En effet, si la déviance est petite, le modèle considéré sera adéquat. Dans le cas contraire, c'est-à-dire si la déviance est grande, alors le modèle considéré est loin du modèle saturé et donc il n'est pas très adéquat.

* La statistique du test

La statistique de déviance permet de tirer le modèle qui présente le bon ajustement aux données de base. Pour cela, nous confrontons les deux hypothèses suivantes :

H_0 : Le modèle considéré à p paramètres est adéquat

H_1 : Le modèle considéré à p paramètres n'est pas adéquat

La statistique du test est notée D , et sa formule est la suivante :

$$D = 2 \sum_{i=1}^m \left(n_i \log \left(\frac{n_i}{\hat{\lambda}_i} \right) - (n_i - \hat{\lambda}_j) \right) \sim \chi_{m-p}^2$$

Le modèle est adéquat si la valeur de D s'approche de $m - p$. Il est à préciser que m représente le nombre d'observation et p le nombre de paramètres estimés.

Pratiquement, si $D_{\text{observé}} > \chi_{m-p, 1-\alpha}^2$ alors H_0 est rejetée. Autrement dit, le modèle considéré n'est pas adéquat.

Il existe d'autres statistiques de déviance, c'est le cas de la statistique de Pearson qui permet de mesurer l'adéquation du modèle considéré aux données de l'étude. Cette statistique se base sur les mêmes hypothèses précédemment citées, mais la différence réside dans la formule de la statistique du test :

$$Q^P = \sum_{i=1}^m \left(\frac{n_i - \hat{\lambda}_j}{\sqrt{\hat{\lambda}_j}} \right) \sim \chi_{m-p}^2$$

Le modèle est adéquat si la valeur de Q^P s'approche de $m - p$.

4.2. Analyse des résidus

Considérons un modèle Normal dont les variables de réponse Y_i sont modélisées par :

$$Y_i = \mu_i + e_i$$

Où les termes d'erreur e_i sont supposés indépendants et identiquement distribués $N(0, \sigma^2)$ et les espérance μ_i sont des fonctions de combinaisons linéaires des paramètres du vecteur β .

Pour ce modèle $\frac{(y_i - \mu_i)}{\sigma} \sim N(0,1)$

Le résidu associé à Y_i est défini par $(y_i - \hat{\mu}_i)$ où $\hat{\mu}_i$ est déterminé par le modèle. Il est calculé à partir de l'estimation du maximum de vraisemblance. Le résidu standardisé est défini comme $r_i = (y_i - \hat{\mu}_i)/\hat{\sigma}$ où $\hat{\sigma}$ est un estimateur de σ . Par conséquent, les résidus standardisés suivent approximativement la loi $(0,1)$.

Pour les autres modèles linéaires généralisés, les résidus sont définis par analogie avec le cas Normal $r_i = (y_i - \hat{\mu}_i)/s_i$ où s_i est la valeur estimée de l'écart type de la valeur modélisée $\hat{\mu}_i$.

Les résidus standardisés peuvent être comparés à la distribution Normale pour évaluer la qualité du choix de la distribution dans le modèle.

4.3. Choix entre différents modèles

Plusieurs modèles concurrents sont en compétition, le but est de choisir le plus adéquat aux données de l'étude. Pour cela, et lorsque les modèles sont emboîtés, il est possible d'utiliser le test de déviance entre modèles ou bien des critères de choix de modèles (AIC et BIC).

H_0 : le modèle simple à p_1 paramètres est adéquat

H_1 : le modèle simple à p_2 paramètres est adéquat ($p_2 > p_1$)

Pour comparer deux modèles emboîtés, nous allons utiliser l'écart de déviance qui mesure la différence entre les deux modèles en question : $\Delta D = D_{simple} - D_{grand}$

Si ΔD est grande, le fait de passer d'un modèle simple à un modèle plus grand a donc apporté un écart de déviance significatif et donc le modèle général est acceptable.

Si l'écart est faible, le modèle simple et celui plus général sont voisins et par parcimonie le modèle simple est conservé. Sous l'hypothèse nulle, la statistique du test d'hypothèse (H_0 contre H_1) précédemment citée ΔD , suit une loi de khi-deux à $p_2 - p_1$ degré de liberté.

Pratiquement, on rejette H_0 si $\Delta D_{obs} > \chi_{p_2-p_1, 1-\alpha}^2$. Dans ce cas, le modèle à p_2 est le plus adéquat aux données de l'étude.

Il est à préciser qu'un mauvais ajustement d'un modèle peut s'expliquer de différentes façons :

- Des variables explicatives importantes ont été oubliées lors de l'écriture de la composante systématique du modèle ;
- La distribution théorique qui permet d'ajuster les observations est mal choisie ;
- La fonction de lien est inappropriée.

Finalement, il reste à noter que pour le choix entre différents modèles non emboîtés, nous comparons les valeurs d'un indice d'ajustement tel le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC). En effet, le modèle qui présente le plus faible critère est sélectionné.

5. Le tarif de l'assurance maladie

5.1 L'approche fréquence sévérité

Pour chaque police d'assurance, la prime est fonction de variables dites de tarification. Généralement, on considère :

- Des informations sur l'assuré, comme l'âge moyen et le nombre de salariés pour une entreprise ;
- Des informations sur le bien assuré, comme le chiffre d'affaire de l'entreprise en perte d'exploitation ;
- Des informations géographiques comme la localisation de l'entreprise.

La fréquence est le nombre de sinistres divisé par l'exposition (correspondant au nombre d'années police) pour une police d'assurance, ou un groupe de polices d'assurance. La plupart des contrats étant annuels, nous ramènerons toujours le nombre de sinistres à une exposition annuelle lors du calcul de la prime, et nous noterons N la variable aléatoire associée. Durant la période d'exposition, nous noterons Y_i les coûts des sinistres, c'est à dire les indemnités versées par l'assureur à l'assuré (ou une tierce personne). La charge totale par police est alors $S = 0$ s'il n'y a pas eu de sinistres, ou sinon : $S = \sum_{i=1}^N Y_i$

Classiquement (et ce point sera important pour constituer la base de données) $Y_i > 0$ et N est alors le nombre de sinistres en excluant les sinistres classés sans suite (i.e. de coût nul).

La prime pure est $E(S) = E(N)E(Y_i)$ dès lors que les coûts individuels sont i.i.d, indépendants du nombre de sinistres.

5.2 La prime commerciale

Considérons une entreprise i ($i = 1, \dots, n$) qui vient d'assurer leurs collaborateurs contre le risque maladie. Cette entreprise est caractérisée par un certain nombre de facteurs de risque comme étant des variables explicatives de la sinistralité. Dans le même sens, nous supposons que cette entreprise compte :

- N_h salariés hommes ;
- N_f salariés femmes ;
- N_a salariés assurés ($N_a = N_h + N_f$) ;
- N_c salariés conjoints ;
- N_e salariés enfants.

Dans ce cas, et seulement dans ce cas, la prime pure est définie par la formule suivante :

$$Prime = N_h Prime_h + N_f Prime_f + N_c Prime_c + N_e Prime_e$$

Avec :

- $Prime_h$: La prime relative à chacun des hommes salariés de l'entreprise ;
- $Prime_f$: La prime relative à chacune des femmes salariées de l'entreprise ;
- $Prime_c$: La prime relative aux bénéficiaires conjoints ;
- $Prime_e$: La prime relative aux bénéficiaires enfants.

Pratiquement, il est jugé difficile de proposer un tarif qui prend en considération le nombre de femmes et des hommes comme un paramètre fondamental de la prime commerciale. En effet, nous trouverons qu'un nombre important d'entreprises ne peuvent pas communiquer à leurs assureurs le nombre des hommes et des femmes qui constituent l'entreprise en question. C'est la raison pour laquelle, nous n'allons pas prendre en considération la composition de l'entreprise assurée en matière du sexe de leurs collaborateurs. Dans ce cas, la prime pure sera calculée de la manière suivante :

$$Prime = N_a Prime_a + N_c Prime_c + N_e Prime_e$$

Où $Prime_a$ est la prime relative à l'ensemble des collaborateurs (hommes et femmes) de l'entreprise assurée.

CONCLUSION

Comme nous l'avons exposé, ce chapitre a introduit le cadre théorique et conceptuel qui servira à mettre en place le tarif de l'assurance maladie. En effet, ce chapitre a défini l'algorithme du CHAID comme étant une méthodologie qui permet de segmenter les assurés de la branche maladie. Ensuite, nous avons mis l'accent sur les modèles linéaires généralisés comme une approche justifiée pour élaborer le tarif de l'assurance maladie.

Le chapitre suivant sera consacré à la mise en pratique des approches théoriques adoptées. En effet, nous allons commencer par sélectionner les variables tarifaires de la prestation pharmacie. Ensuite, nous allons segmenter la population assurée en groupes homogènes, en nous basant sur les descripteurs de la sinistralité que nous allons déterminer par la procédure de sélection *stepwise*.

Enfin, nous allons modéliser la prime pure de l'assurance maladie, en nous basant sur la sinistralité individuelle des assurés. Le tarif sera ainsi en fonction du profil de risque de chaque assuré de la branche maladie de base.

CHAPITRE 4. TARIFICATION EN ASSURANCE MALADIE DE BASE

CHAPITRE 4. TARIFICATION EN ASSURANCE MALADIE DE BASE

INTRODUCTION

Après avoir introduit les modèles linéaires généralisés et expliqué la structure de cette approche statistique, ce chapitre sera consacré à la modélisation de la prime pure de la branche d'assurance maladie de base. En effet, nous allons commencer par sélectionner les variables dites de tarification, ensuite nous allons segmenter le portefeuille étudié en nous basant sur les variables tarifaires sélectionnées. Cette segmentation de la population nous aura permis de proposer des tarifs différents pour les assurés de la branche maladie, et ceci selon la structure de la population en matière de la sinistralité. A cet effet, nous modélisons séparément le coût moyen et le nombre de sinistres afin d'isoler les effets des facteurs sur la fréquence et la sévérité. Cette méthodologie permet d'obtenir une meilleure compréhension de l'influence des facteurs sur le risque.

Certains facteurs ont un grand impact sur la fréquence de sinistres sans pour autant être significatifs quant à l'explication de la sévérité des sinistres. Cette distinction offre de nombreuses possibilités dans le processus de tarification.

Section 1 : Segmentation et écrêtement des données en assurance maladie

I. Classification des assurés

1. Sélection de variables tarifaires

Dans le but de déterminer les descripteurs potentiels de la sinistralité des adhérents du poste pharmacie, nous allons appliquer le processus de sélection *stepwise* aux données des adhérents de la prestation pharmacie avec 201288 observations (durant les cinq derniers exercices d'assurance 2010 – 2014).

Les variables susceptibles d'expliquer la sinistralité des adhérents de la prestation pharmacie sont :

- Taux de remboursement ;
- Age moyen de l'entreprise assurée ;
- Salaire moyen ;
- Localisation de l'entreprise (la région) ;
- Plafond général d'indemnisation ;
- Sexe de l'assuré.

Après avoir appliqué le processus de sélection *stepwise* par le logiciel SAS, plus précisément la requête *selection=stepwise* de la procédure *reg*, nous avons obtenu les résultats suivants :

Tableau 3. Sélection stepwise basée sur le F partiel de Fisher

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AGE_MOY		1	0.0387	0.0387	14723.7	8090.46	<.0001
2	Salaire_Moy		2	0.0277	0.0664	8503.57	5969.86	<.0001
3	SEXE		3	0.0264	0.0928	2567.26	5863.59	<.0001
4	Tauxremb		4	0.0095	0.1023	441.062	2123.59	<.0001
5	REGION		5	0.0017	0.1040	59.1254	383.84	<.0001
6	Plafgen1		6	0.0002	0.1042	7.0000	54.13	<.0001

Source : *Elaboré à partir de la table tarification*

Le tableau 3 montre que toutes les variables introduites pour expliquer la sinistralité des assurés de la prestation pharmacie sont significatives au seuil 5%. Et par conséquent, les variables que nous allons utiliser pour modéliser la fréquence et la sévérité de la poste pharmacie et qui se rapportent aux entreprise assurées contre le risque maladie sont présentées comme suit :

*** Variables caractérisant l'entreprise assurée :**

- Localisation de l'entreprise.
- Age moyen ;
- Salaire moyen ;

*** Variables caractérisant l'assuré ⁸ :**

- Plafond général ;
- Sexe de l'assuré ;
- Taux de remboursement ;

2. Segmentation des variables tarifaires

Après l'étape "sélection des variables tarifaires", il vient une étape très importante dans le processus de tarification, c'est la segmentation de la population assurée en groupes homogènes. En nous basant sur les variables précédemment sélectionnées, nous allons déterminer les groupes d'assurés ayant un comportement analogue en matière de la sinistralité.

Dans cette section, nous souhaitons donc appréhender le comportement des variables candidates à la tarification de notre portefeuille face aux critères de la fréquence et du coût

⁸ L'affilié au régime d'assurance maladie, prestation pharmacie

moyen des sinistres, ainsi que de la prime moyenne et du rapport S/P. Nous créons ainsi, des classes intra-homogènes et inter-hétérogènes pour les variables quantitatives et des classes de modalités pour les variables qualitatives. Nous allons tenter de calculer, pour chacune des variables dites de tarification, et pour chacun de ses segments, l'ensemble des quantités statistiques qui mesurent l'intensité de la sinistralité.

Il est à noter que les segments qui seront exposés, concernent la prestation la plus sinistrée des assurés de l'assurance maladie de base.

*** L'âge moyen**

La variable âge moyen est une variable quantitative continue, qui varie d'une entreprise à une autre. La segmentation de cette variable explicative était élaborée grâce à l'algorithme de CHAID. Il est à noter que nous avons pris comme variable dépendante le nombre de sinistres déclaré par les entreprises assurées. En effet, nous avons trouvé qu'il s'agit d'une forte dépendance entre les deux variables en question (âge moyen et nombre de sinistres). Le test suivant mesure la dépendance entre l'âge moyen et le nombre de sinistres :

Tableau 4. Test de Khi-deux d'indépendance entre l'âge moyen et le nombre de sinistres

Statistic	DF	Value	Prob
Chi-Square	101080	304638	<.0001
Likelihood Ratio Chi-Square	101080	82345	1
Mantel-Haenszel Chi-Square	1	7751	<.0001
Phi Coefficient		1.23194	
Contingency Coefficient		0.77641	
Cramer's V		0.19985	

Source : Elaboré à partir de la table tarification

Le tableau 4 montre que la p-value est largement inférieure à 5% (seuil de signification). Il s'agit alors d'une dépendance entre la variable âge moyen et la variable nombre de sinistres des assurés.

Dans ce cas, nous pouvons utiliser l'algorithme du CHAID pour segmenter la variable âge moyen des assurés de la prestation pharmacie, en prenant en considération le nombre de sinistres comme variable réponse. Il est à signaler que nous avons forcé l'algorithme du CHAID en terme de nombre de classes souhaitées. En effet, la politique de tarification nous a imposé de segmenter les assurés en quatre groupes en matière de l'âge moyen. Le tableau suivant représente les quantités statistiques de la sinistralité des assurés de la prestation pharmacie :

Tableau 5. Variation de la sinistralité suivant les classes de l'âge moyen

Classe d'âge	Poids	Fréquence	Cout moyen	Prime pure
1	9.99	0.52286	179.045	93.616
2	29.99	0.88576	245.857	217.771
3	29.86	1.19743	309.06	370.079
4	30.16	1.38537	412.918	572.045

Source : Elaboré à partir de la table tarification

*** Le taux de remboursement**

En ce qui concerne la variable "taux de remboursement", nous avons limité le nombre de classes à cinq segments : le moins risqué, le plus risqué et les intermédiaires de ces deux derniers segments. Le tableau suivant représente les différents chiffres relatifs à la sinistralité des assurés selon le taux de remboursement :

Tableau 6. Variation de la sinistralité suivant les classes du taux de remboursement

Classe de remboursement	Poids	Fréquence	Cout moyen	Prime pure
1	14.26	0.35302	192.366	67.909
2	0.49	1.39741	203.743	284.711
3	51.92	1.23283	329.623	406.371
4	16.01	1.13106	322.003	364.203
5	17.33	1.34989	371.792	501.877

Source : Elaboré à partir de la table tarification

De ce dernier tableau, on en déduit que le taux de remboursement et la sinistralité varient dans le même sens. Autrement dit, plus le taux de remboursement augmente, plus la sinistralité est importante et réciproquement. Ces résultats sont homogènes avec la réalité économique, du fait que les assurés de la branche maladie ont une tendance de consommer beaucoup plus le produit "assurance maladie", lorsqu'ils avaient un taux de remboursement assez important.

*** Le salaire moyen**

Le salaire moyen des employés est aussi une variable relative à l'entreprise assurée. Le tableau montre l'évolution de la sinistralité selon le salaire moyen.

Tableau 7. Variation de la sinistralité suivant les classes du salaire moyen

Classe de salaire moyen	Poids	Fréquence	Cout moyen	Prime pure
1	10.06	0.28939	234.077	67.738
2	9.93	0.73529	268.333	197.302
3	24.97	0.91188	306.763	279.73
4	35.11	1.19245	369.267	440.331
5	19.95	1.82685	320.113	584.798

Source : Elaboré à partir de la table tarification

A partir de ces résultats, nous constatons que les entreprises ayant un salaire moyen important, consomment beaucoup plus le produit assurance maladie. C'est la raison pour laquelle, ce type d'entreprises est pénalisé par des primes d'assurance énormes.

*** Le plafond général**

Le plafond général d'indemnisation est une limite contractuelle qui repose sur les caractéristiques des assurés. Cette variable pénalise toute déclaration d'une charge des sinistres énorme, et qui dépasse le montant maximum d'indemnisation des actes de la branche maladie. L'assureur peut aussi procéder par un sous plafond, afin de ne pas prendre en considération les sinistres ayant un montant faible. En réalité, cette politique de souscription est compliquée dans un marché marqué par la concurrence.

Le tableau 8 met l'accent sur la sinistralité des assurés de la prestation pharmacie, selon les modalités de la variable plafond général :

Tableau 8. Variation de la sinistralité suivant les classes du plafond général

Classe de plafond	Poids	Fréquence	Cout moyen	Prime pure
1	27.7	1.05388	279.169	294.212
2	72.3	1.12943	348.261	393.338

Source : Elaboré à partir de la table tarification

Il apparaît que les assurés appartenant à la classe deux du descripteur "plafond général" ont la prime pure la plus importante. Ce résultat est adéquat, du fait que les assurés du groupe deux ont le poids le plus important, et ils ont un plafond assez grand.

*** La localisation de l'entreprise**

La variable localisation de l'entreprise est une variable discriminante en assurance de groupes. En effet, la sinistralité des assurés varie d'une région à une autre, d'où l'intérêt de prendre en considération la localisation de l'entreprise comme variable explicative de la sinistralité.

Tableau 9. Variation de la sinistralité suivant la localisation de l'entreprise

Classe de région	Poids	Fréquence	Cout moyen	Prime pure
1	73.54	1.24406	342.157	425.665
2	10.49	0.8546	275.859	235.75
3	4.09	0.7787	237.208	184.713
4	11.87	0.58611	285.57	167.375

Source : Elaboré à partir de la table tarification

Le tableau 9 montre que la sinistralité des assurés dépend aussi de la localisation de l'entreprise. En effet, la fréquence des sinistres varie d'une région à une autre, ce qui impacte le montant de la prime selon la région de l'entreprise.

*** Le sexe de l'assuré**

Le tableau 10 représente la variation de la sinistralité des assurés de la prestation pharmacie selon la variable sexe. Il décrit également la variabilité de la prime pure entre les hommes et les femmes.

Tableau 10. Variation de la sinistralité suivant le sexe de l'assuré

Classe de sexe	Poids	Fréquence	Coût moyen	Prime pure
1	73.1	0.88771	360.865	320.345
2	26.9	1.72409	286.581	494.09

Source : Elaboré à partir de la table tarification

II. Ecrêtement des données

1. Les sinistres graves en assurance maladie

Tout d'abord, la stabilité temporelle des indicateurs de risque est nécessaire pour avoir une bonne adéquation entre la sinistralité et la tarification. Dans ce sens, la survenance de sinistres graves dans une classe, vient perturber l'hypothèse d'homogénéité des classes et de stabilité des indicateurs.

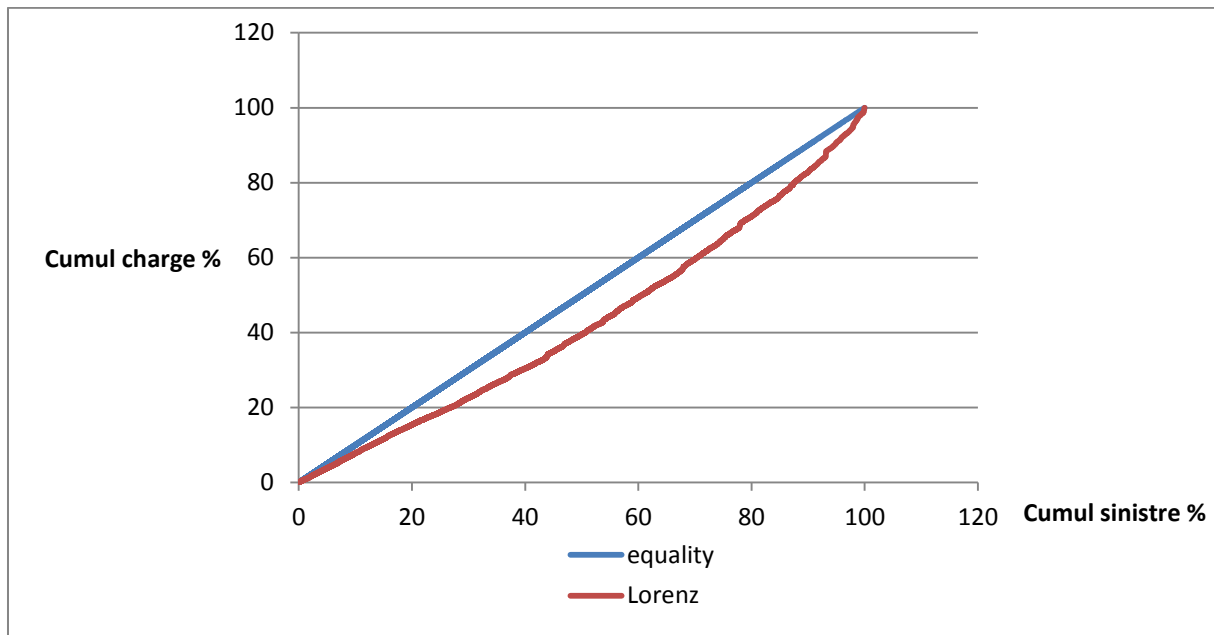
Cependant, l'ecrêtement des données a pour but de corriger l'hétérogénéité des risques dans une classe donnée. Il s'agit d'étudier la loi des extrêmes ou des maximums d'une suite de variables aléatoires réelles même si, et spécialement si, la loi du phénomène n'est pas connue.

Dans cette partie du rapport, nous allons donc écrêter les données de la prestation pharmacie, et ceci en nous basant sur la théorie des valeurs extrêmes. C'est une approche qui permet de déterminer un seuil suffisamment élevé, à partir duquel les sinistres sont jugés graves.

Un des moyens qui permet de prouver l'existence des sinistres graves au sein d'une telle population étudiée, est la courbe de Lorenz. C'est un outil graphique qui décrit la répartition de la sinistralité entre les assurés. Ajoutant à cela, le graphique qq-plot qui permet aussi de visualiser certaines données dites aberrantes, et qui ne sont pas homogènes avec une masse importante d'observations.

Dans notre cas particulier, nous allons utiliser la courbe de Lorenz pour représenter la répartition de la charge des sinistres au sein de la population étudiée. Le but est de mesurer les inégalités qui peuvent exister au sein de la population des assurés de la branche maladie (prestation pharmacie) en terme de la distribution de la charge des sinistres. Il est à préciser que la courbe de Lorenz offre aussi un autre paramètre qui mesure l'inégalité de répartition appelé indice de Gini.

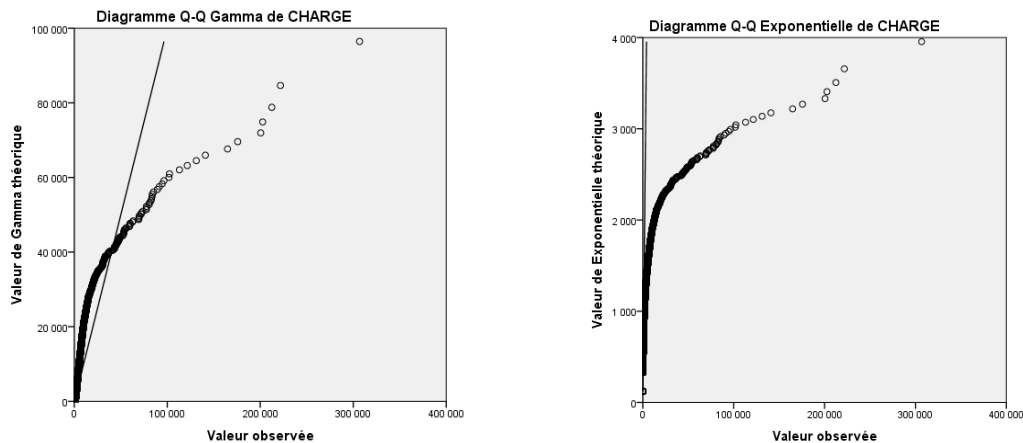
Graphique 12. Courbe de concentration de Lorenz



A partir du graphique 12, nous remarquons que la courbe de Lorenz est une fonction convexe. En plus, nous avons trouvé que 7% des sinistres ont généré plus de 13% de la charge. On en déduit qu'il s'agit de certains sinistres graves au sein du portefeuille étudié, et par conséquent, nous sommes obligé de traiter ce type de sinistres en nous basant sur une méthode appropriée.

Dans la figure 2, nous trouverons le qq-plot de la charge des sinistres des assurés de la prestation pharmacie :

Figure 2. qq-plot des adhérents de la prestation pharmacie



D'après les deux graphiques du QQ-plot ci-dessus, nous pouvons facilement constater l'existence des sinistres graves qui se présentent par des points très éloignés de la droite de Henry. Dans ce cas, il est indispensable de séparer les grands sinistres des petits sinistres afin que le tarif souhaité soit cohérent avec la prime commerciale réelle.

2. Sélection de seuil

Avant de pouvoir estimer le modèle, il nous faut trouver un seuil u de sélection des données extrêmes suffisamment élevé pour que le tarif de l'assurance maladie soit cohérent. Si nous choisissons un seuil trop bas, les estimations seront biaisées. Il est à signaler qu'au-dessus de ce seuil, nous conservons assez de données pour des estimations précises (si le seuil est trop élevé, les écarts-types des estimateurs seront très importants).

Un des outils de choix du seuil est le graphe de la fonction moyenne des excès (FME) en u (ME-plot) que nous allons définir ci-dessus, et l'autre outil est l'estimateur de Hill

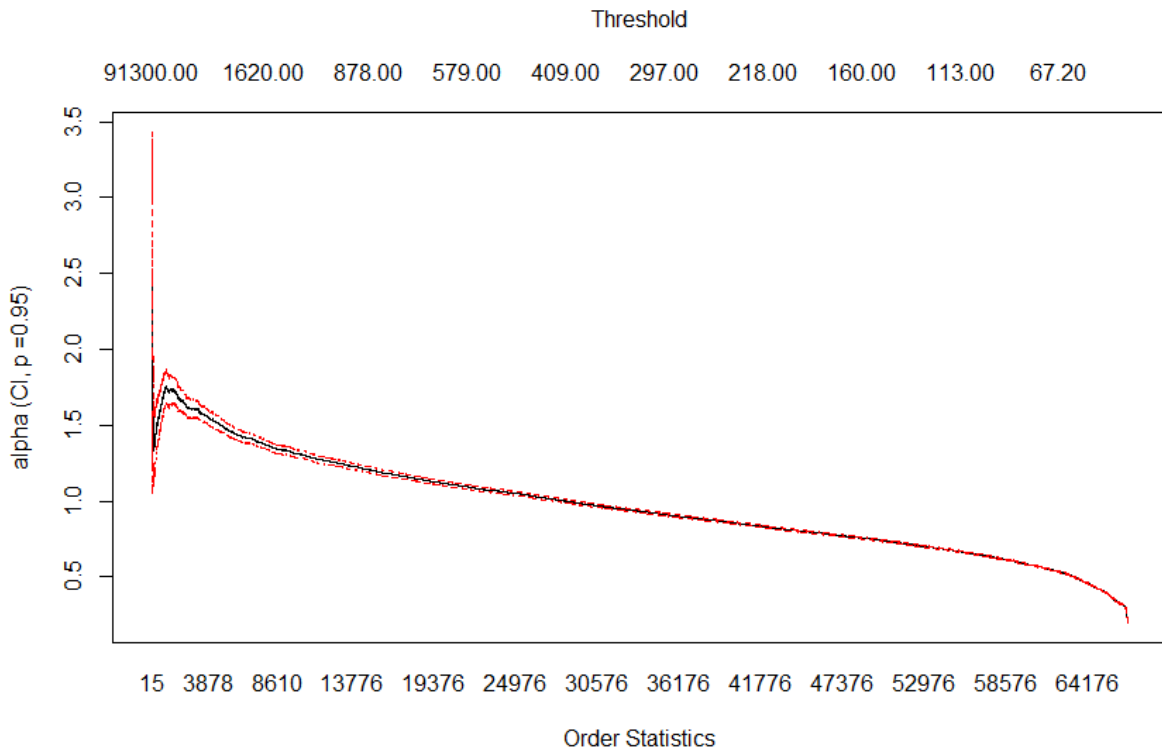
2.1. Estimateur de Hill

Hill (1975) a proposé l'estimateur de la distribution GPD suivant :

$$\hat{\xi} = \frac{1}{k-1} (\ln(X_{i,n}) - \ln(X_{k,n})) \quad \text{pour } k \geq 2$$

Avec k , l'ordre statistique le plus élevé (le nombre des excès), n est la taille de l'échantillon et $\alpha = \frac{1}{\xi}$ est l'indice de la queue de distribution.

Figure 3. Hill-plot de la charge des sinistres graves



Ce graphique Hill-plot, nous permet d'avoir des estimations du paramètre α en fonction de l'ordre statistique le plus élevé (nombre des excès), nous choisissons ainsi l'indice le plus stable. Le cas de la figure ci-dessus n'est pas assez informatif, car nous n'observons pas de stabilité. Nous pouvons encore utiliser une représentation graphique de la moyenne des excès.

2.2. La fonction moyenne des excès

La fonction moyenne des excès est par définition la somme des excès dépassant un certain seuil élevé, noté u , divisé par le nombre de points des données qui dépassent ce seuil.

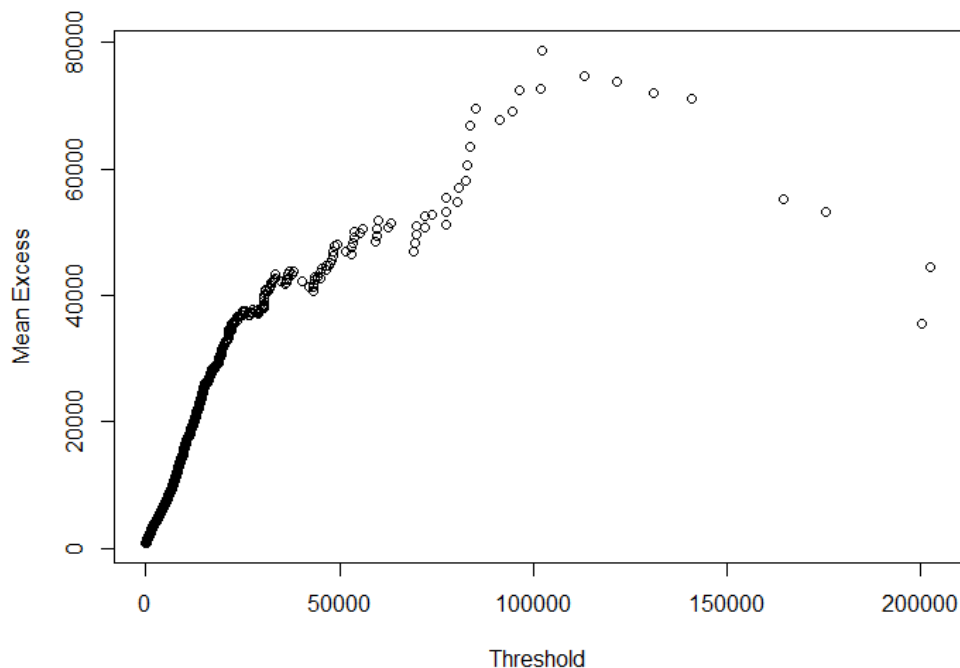
$$e_n(u) = \frac{\sum_{i=1}^n (X_i - u)}{\sum_{i=1}^n I_{\{X_i > u\}}} \quad \text{avec} \quad I_{\{X_i > u\}} = \begin{cases} 1, & \text{si } X_i > u \\ 0, & \text{si non} \end{cases}$$

Autrement dit, il s'agit d'une estimation de la fonction moyenne des excès (notée FME) qui permet de décrire la prédiction du dépassement du seuil lorsqu'un excès se produit.

Trois cas peuvent se présenter :

- Si à un certain seuil, la FME empirique est marquée par une pente positive, les données suivent la distribution GPD avec un paramètre ξ positif ;
- Si la fonction moyenne des excès est horizontale, les données suivent une distribution exponentielle ;
- Si la FME empirique est marquée par une pente négative, les données suivent une distribution à queue légère.

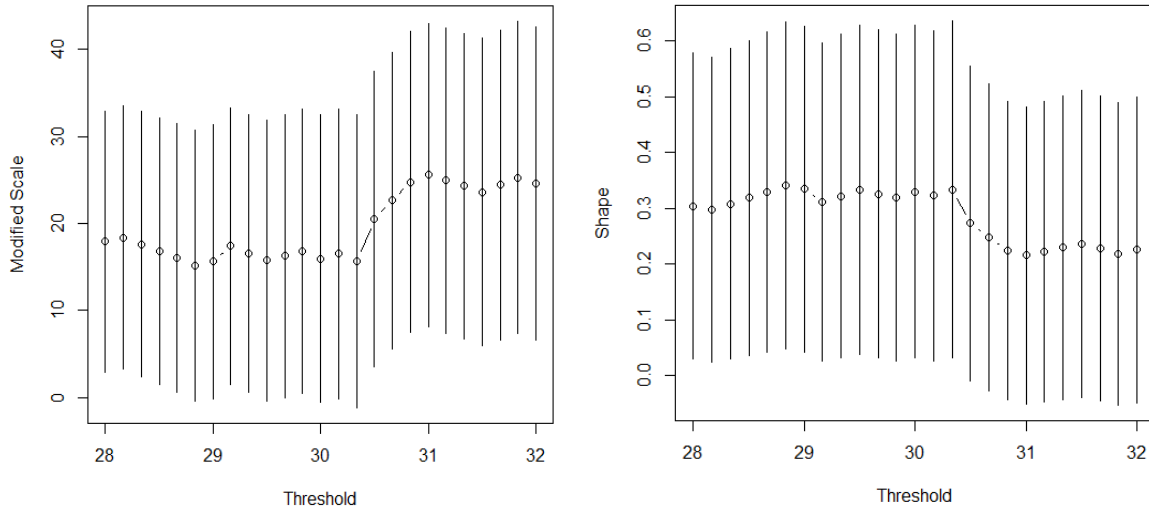
Figure 4. La fonction moyenne des excès



Nous observons une stabilité linéaire jusqu'au seuil 29000. Pour compléter notre décision du seuil, nous allons utiliser un autre outil graphique qui permet de visualiser la limite de la linéarité sur la figure de la moyenne des excès. A l'aide de la fonction `teplot` (Threshold

Choice Plot), dans notre exemple nous avons représenté entre les seuils 28000 et 32000 pour mieux observer.

Figure 5. tcplot de la charge des sinistres (en 1000 Dh)



La stabilité linéaire combinée de ces deux représentations nous permet de prendre un seuil égal 30333 pour la modélisation.

3. Ecrêtement des valeurs aberrantes

Pour écrêter les valeurs aberrantes, nous avons retenu 30333 comme seuil d'écrêtement pour les assurés de la prestation pharmacie. Ensuite, nous avons réparti l'excès de la charge des sinistres qui dépasse le seuil retenu sur l'ensemble des assurés sinistrés ayant une charge inférieure à 30333 DH.

Section 2 : Détermination de la prime pure

I. Modélisation de la fréquence des sinistres

1. Ajustement et choix du modèle

Tout d'abord, les modèles linéaires généralisés nécessitent un ajustement des données par une loi théorique qui appartient à la famille exponentielle. Cependant, les principales caractéristiques des lois de probabilité les plus souvent utilisées pour le dénombrement des sinistres en assurance sont :

- La loi de Poisson ;
- La loi binomiale négative ;
- La loi binomiale.

Pour vérifier ce constat, nous avons effectué un ajustement graphique de la fréquence de consommation pour chaque poste sous le logiciel R. Il en résulte que pour la totalité des

postes, l'ajustement par une binomiale négative est bien meilleur que celui de Poisson. Ceci peut être expliqué par le phénomène de sur-dispersion constaté pour l'ensemble des postes. D'ailleurs l'ajustement par une loi de Poisson est meilleur pour les portefeuilles qui ne présentent pas de sur-dispersion, chose non vérifiée pour le portefeuille santé étudié. Ainsi, nous nous contenterons d'exposer dans ce chapitre la modélisation de la fréquence pour le poste pharmacie. Les résultats des autres postes seront exposés en annexes.

*** Loi de Poisson**

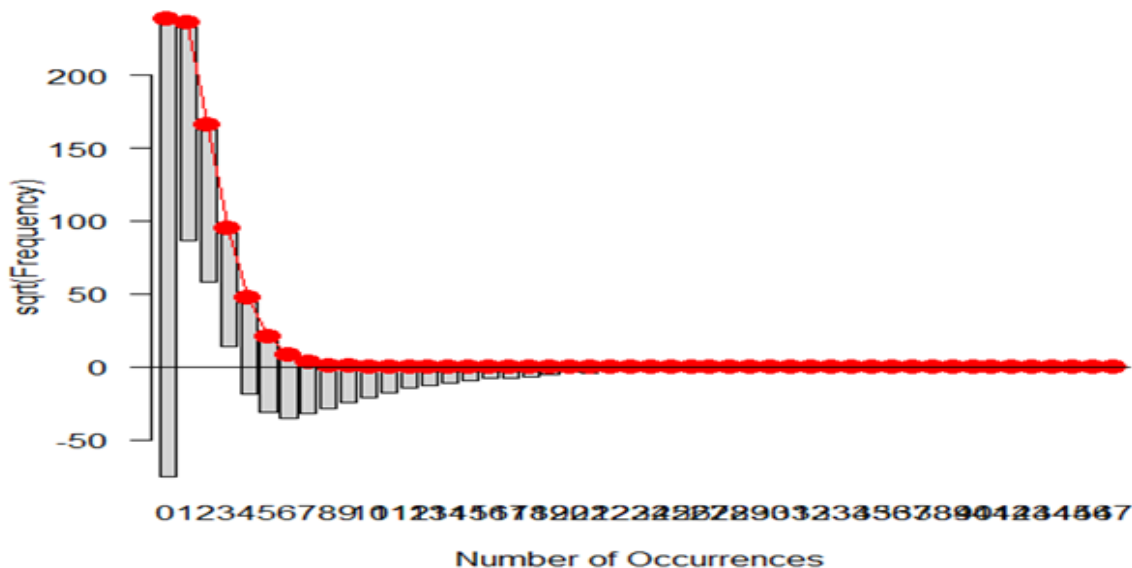
Soit N une variable aléatoire qui suit une loi de Poisson de paramètre λ , sa fonction de probabilité s'écrit :

$$P(N = n) = \exp(-\lambda) \frac{\lambda^n}{n!}$$

Avec $E(N) = \lambda$ et $V(N) = \lambda$

Comme indiqué précédemment, l'étude serait exposée uniquement pour le poste pharmacie qui occupe le premier rang au niveau de la fréquence de consommation et des frais engagés totaux. L'ajustement a été effectué par le logiciel R et a fourni un mauvais ajustement pour la loi Poisson.

Figure 6. Ajustement de la fréquence des sinistres par une loi de Poisson



Les points rouges représentent la loi théorique et les histogrammes les fréquences observées, qui sont collés par le sommet à la loi théorique. Tout écart de la base d'un histogramme avec l'axe des abscisses indique donc un mauvais ajustement des observations par la loi théorique.

Nous remarquons ainsi que l'ajustement par la loi de Poisson est peu satisfaisant. En effet les écarts observés sont conséquents et cela quelque soit le nombre d'occurrences.

La loi Binomiale-Négative est en effet une bonne alternative à la loi de Poisson, en particulier en cas de sur-dispersion des données. En effet, l'utilisation du modèle de Poisson revient à supposer l'égalité entre le nombre moyen de sinistres et la variabilité de ce nombre. Bien souvent, et c'est le cas sur notre jeu d'observation, cette observation n'est pas satisfaite.

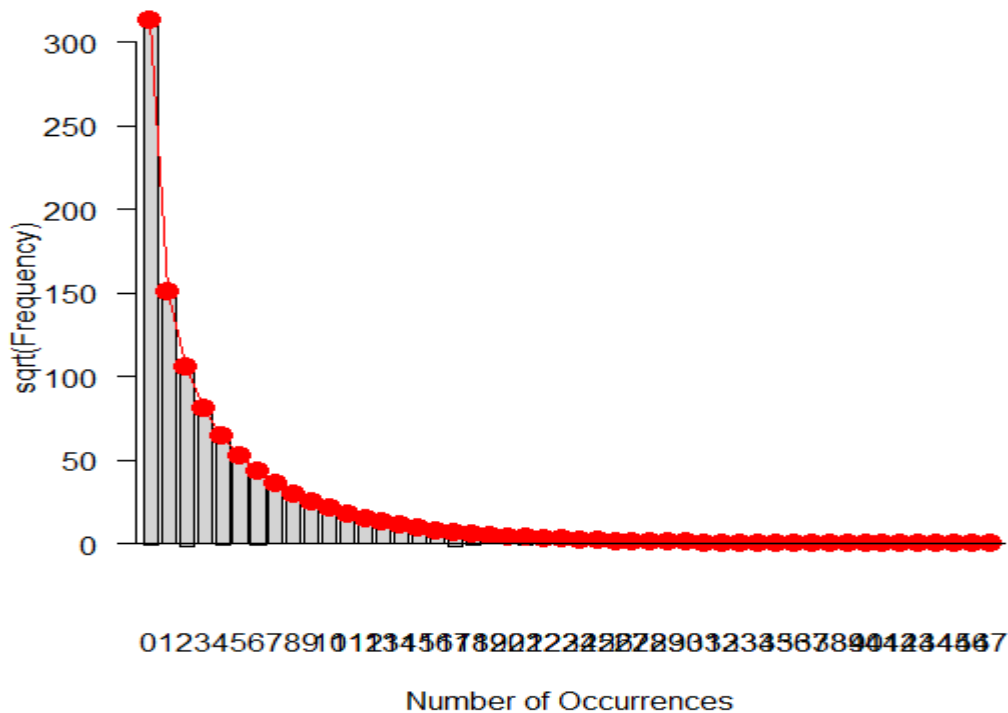
*** Loi binomiale négative**

Soit N une variable aléatoire qui suit une loi binomiale négative de paramètres (n,p) , sa fonction de probabilité s'écrit :

$$P(N = k) = C_{n-1}^{k-1} p^n (1 - p)^{k-n}$$

Avec $E(N) = n/p$ et $V(N) = \frac{n(1-p)}{p^2}$

Figure 7. Ajustement de la fréquence des sinistres par une loi binomiale négative



La figure 7 montre que l'ajustement est bien meilleur pour la loi binomiale négative que pour la loi de Poisson, les écarts avec l'axe des abscisses sont en effet très faibles. A la vue de ces résultats, nous sommes donc fortement tentés de modéliser la fréquence des sinistres par une loi Binomiale-Négative. Pour le cas général, nous observons que la loi de Poisson ne peut s'ajuster que sur très peu de frais, alors que la loi binomiale négative convient pour tous les types de soins. C'est donc cette dernière que l'on retiendra pour la modélisation par les modèles linéaires généralisés.

2. Cellule de référence

En pratique, il est plus aisé de paramétrer le modèle linéaire généralisé en considérant une cellule de référence, qui est un paramètre qui s'applique à toutes les observations. Un niveau de chaque facteur ne doit pas avoir de paramètre afin que le modèle demeure identifiable.

Pour chaque variable, nous choisissons comme niveau de référence (i.e. celui pour lequel toutes les variables binaires utilisées pour la coder valent simultanément 0) la modalité la plus représentée dans le portefeuille. Les résultats s'interpréteront ensuite comme une sur ou sous-sinistralité par rapport à cette classe de référence.

Prenons l'exemple des facteurs explicatifs de la sinistralité de la prestation pharmacie. Le tableau 11 représente les classes de référence adoptées pour ces descripteurs potentiels :

Tableau 11. La classe de référence

Variable tarifaire	classe de référence
Taux de remboursement	Classe 3
Salaire moyen	Classe 4
Région de l'entreprise	Classe 1
Age moyen	Classe 4
Sexe de l'assuré	Classe 1

Source : Elaboré par l'auteur

3. Analyse de Type III

Pour étudier la significativité de l'impact d'une variable introduite dans le modèle, une analyse de type III s'impose. Elle peut être effectuée suivant le test de Wald ou du rapport de vraisemblance. Pour notre part, nous avons effectué cette analyse grâce au logiciel SAS avec le critère du rapport de vraisemblance, en testant la significativité des variables explicatives une à une. Ce test peut être aussi effectué par le critère de l'AIC (le meilleur modèle étant celui qui possède la valeur de l'AIC la plus faible).

* Le principe du test

Le test du rapport de vraisemblance permet de comparer un modèle à un modèle réduit, dans le sens où il comportera moins de variables. La valeur renvoyée par le test indique aussi à quel point la variable possède le pouvoir explicatif. Comme son nom l'indique, le test s'appuie sur le rapport de vraisemblance et ainsi sur l'effet que peut avoir l'omission d'une variable sur la vraisemblance du modèle.

* La statistique du test

Tout d'abord, les hypothèses à confronter se présentent comme suit :

H_0 : La variable X introduite n'est pas influente dans le modèle ;

H_1 : La variable X introduite est influente dans le modèle.

Sous H_0 la statistique $V = -2\ln(R)$ suit asymptotiquement une loi de Khi-deux à n degrés de liberté, où :

$$R = \frac{\text{Vraisemblance du modèle sans } X}{\text{Vraisemblance du modèle avec } X}$$

$n =$ nombre de paramètres du modèle complet (avec la variable X)
 – nombre de paramètres du modèle réduit (sans la variable X)

Les résultats du test sont exposés comme suit :

Tableau 12. Analyse de type 3 des facteurs de sinistralité

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
CLASSE_AGE	3	4766.34	<.0001
CLASSE_Remb	4	2960.47	<.0001
CLASSE_SM	4	4957.91	<.0001
CLASSE_SEXE	1	5014.56	<.0001
CLASSE_REG	3	726.47	<.0001

Source : Elaboré à partir du portefeuille maladie

Le tableau 12 montre que la p-value de chaque variable tarifaire introduite dans le modèle est largement inférieure à 5% (seuil de signification). On en déduit ainsi que toutes les variables ont un impact significatif sur la fréquence de consommation des actes du poste pharmacie. Nous remarquons aussi que les variables sexe de l'assuré, le taux de remboursement, le salaire moyen et l'âge moyen de l'entreprise assurée sont les facteurs les plus significatifs au sens de l'analyse de type 3.

4. Analyse des résultats

4.1 Estimation des paramètres du modèle établi

Dans cette partie du rapport, nous allons estimer les paramètres du modèle par la méthode de maximum de vraisemblance. Il est à signaler que la qualité de l'estimation des paramètres, repose sur la convergence de l'algorithme de Newton Raphson, comme étant la méthode numérique qui permet de prédire les valeurs des paramètres introduits dans le modèle. Le tableau suivant, illustre l'estimation de tous les paramètres du modèle adopté pour modéliser la fréquence des sinistres :

Tableau 13. Estimation des paramètres de la fréquence des sinistres

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4307	0.0119	0.4073	0.454	1307.83	<.0001
CLASSE_AGE	AGE1	1	-1.1854	0.0214	-1.2273	-1.1436	3080.87	<.0001
CLASSE_AGE	AGE2	1	-0.74	0.0137	-0.7668	-0.7132	2937.51	<.0001
CLASSE_AGE	AGE3	1	-0.2935	0.0127	-0.3184	-0.2685	532.13	<.0001
CLASSE_AGE	AGE4	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_Remb	Remb1	1	-1.0428	0.0196	-1.0811	-1.0044	2837.08	<.0001
CLASSE_Remb	Remb2	1	0.2256	0.0676	0.0931	0.3582	11.13	0.0008
CLASSE_Remb	Remb3	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_Remb	Remb4	1	-0.1957	0.0143	-0.2237	-0.1676	186.82	<.0001
CLASSE_Remb	Remb5	1	-0.1897	0.0141	-0.2173	-0.1621	181.79	<.0001
CLASSE_SM	SM1	1	-1.1754	0.0247	-1.2237	-1.127	2270.43	<.0001
CLASSE_SM	SM2	1	-0.4649	0.0195	-0.5031	-0.4267	568.93	<.0001
CLASSE_SM	SM3	1	-0.1125	0.0133	-0.1386	-0.0865	71.7	<.0001
CLASSE_SM	SM4	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_SM	SM5	1	0.5215	0.0138	0.4944	0.5486	1422.93	<.0001
CLASSE_SEXE	A	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_SEXE	B	1	0.7709	0.011	0.7494	0.7924	4923.79	<.0001
CLASSE_REG	REG1	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_REG	REG2	1	-0.347	0.0172	-0.3808	-0.3132	405.22	<.0001
CLASSE_REG	REG3	1	0.1401	0.0282	0.0847	0.1954	24.61	<.0001
CLASSE_REG	REG4	1	-0.3453	0.0181	-0.3809	-0.3098	362.52	<.0001
Dispersion		1	2.1743	0.017	2.1413	2.2078		

Source : Elaboré à partir du portefeuille maladie

Le tableau 13 montre que la p-value de chaque paramètre introduit dans le modèle est largement inférieure à 5%. On en déduit alors que les variables retenues pour modéliser la fréquence de la sinistralité des assurés de la prestation pharmacie, sont toutes significatives.

4.2 Validation du modèle

Tout d'abord, il est indispensable de s'assurer que le modèle considéré s'ajuste bien aux données avant de pratiquer l'inférence statistique. A cet effet, nous utilisons la déviance comme statistique de test qui permet de valider un modèle linéaire généralisé. Elle permet ainsi de tirer le modèle qui présente le bon ajustement aux données de base. Pour cela, nous confrontons les deux hypothèses suivantes :

H_0 : Le modèle considéré à p paramètres est adéquat contre H_1 : le modèle considéré à p paramètres n'est pas adéquat.

Afin de valider le modèle considéré, nous avons mis en pratique le test de déviance grâce à la procédure *genmod* du logiciel SAS.

Tableau 14. Test de déviance du modèle assurés maladie

Obs	Criterion	DF	Value	ValueDF	pvalue
1	Deviance	1.50E+05	117759.84	0.7756	1
2	Scaled Deviance	1.50E+05	117759.84	0.7756	1
3	Pearson Chi-Square	1.50E+05	190675.41	1.2558	0
4	Scaled Pearson X2	1.50E+05	190675.41	1.2558	0
5	Log Likelihood	–	-59183.07	–	.
6	Full Log Likelihood	–	-184270.2	–	.
7	AIC (smaller is better)	–	368574.48	–	.
8	AICC (smaller is better)	–	368574.48	–	.
9	BIC (smaller is better)	–	368743.3	–	.

Source : Elaboré à partir du portefeuille maladie

Le tableau montre que la p-value du test de déviance est largement supérieure à 5%, et par conséquent, nous acceptons l’hypothèse nulle du test considéré. Autrement dit, le modèle mis en place pour modéliser la fréquence des sinistres des assurés de la prestation pharmacie est validé.

4.3 Prédiction de la fréquence

Pour une observation n_i associée au vecteur de variables explicatives x_i , sa valeur prédite par le modèle ayant le logarithme comme fonction de lien est

$$\hat{n}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_p x_{i,p})$$

Cependant, les résultats obtenus par le modèle précédemment établi, nous permettent de prédire les fréquences de la consommation du poste pharmacie suivant les caractéristiques de l’individu assuré et de la police souscriptrice du contrat. Or, la fréquence obtenue change de manière multiplicative en passant de la classe de référence à une autre classe. Ainsi, un coefficient d’une telle classe inférieur à 0, correspond à une réduction de la fréquence de consommation. Par contre, un coefficient supérieur à 0 indique qu’il s’agit d’une majoration de la fréquence des sinistres.

Prenons maintenant l’exemple d’un assuré homme qui vient de s’assurer contre le risque maladie et employé chez une entreprise ayant les caractéristiques suivantes : un âge moyen, un taux de remboursement et une région de la classe 3 – un salaire moyen de la classe 2.

Cet assuré souhaite enregistrer une fréquence de sinistres d’ordre de 83%. Le tableau 15 représente l’estimation de chaque paramètre de cet assuré ayant les critères cités précédemment.

Tableau 15. Prédiction de la fréquence des sinistres du poste pharmacie

	Fréquence				
	AGE3	Remb3	SM2	A	REG3
Estimate/classe	-0.2935	0	-0.4649	0	0.1401
Fréquence	83%				

Source : Elaboré à partir de la table estimation des paramètres de la fréquence des sinistres

II. Modélisation de la charge des sinistres

1. Choix de la distribution (graphique qq-plot)

Principe

Il ne s'agit pas d'un test au sens statistique du terme. Le graphique Q-Q plot (quantile-quantile plot) est un graphique "nuage de points" qui vise à confronter les quantiles de la distribution empirique et les quantiles d'une distribution théorique normale, de moyenne et d'écart type estimés sur les valeurs observées. Si la distribution est compatible avec la loi normale, les points forment une droite. Dans la littérature francophone, ce dispositif est appelé Droite de Henry.

Application sur les données

Tout d'abord, la charge des sinistres est une variable continue, c'est pour cela que les choix élus pour représenter la distribution de cette variable sont :

1. Gamma
2. Exponentiel
3. Log normal

Pour accepter ou rejeter une telle loi en matière de sa distribution vis-à-vis la variable coût des sinistres, nous allons construire un graphique Q-Q plot pour les différentes lois grâce au logiciel SPSS. Ci-dessous, nous trouverons les trois diagrammes qui illustrent la distribution de la charge des sinistres.

Figure 8. qq-plot gamma de la charge des sinistres

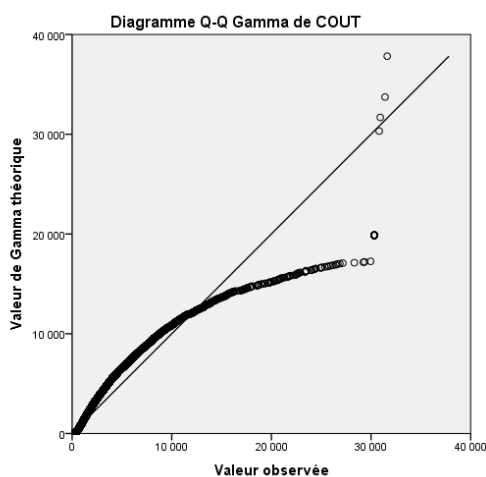


Figure 9. qq-plot log normal de la charge des sinistres

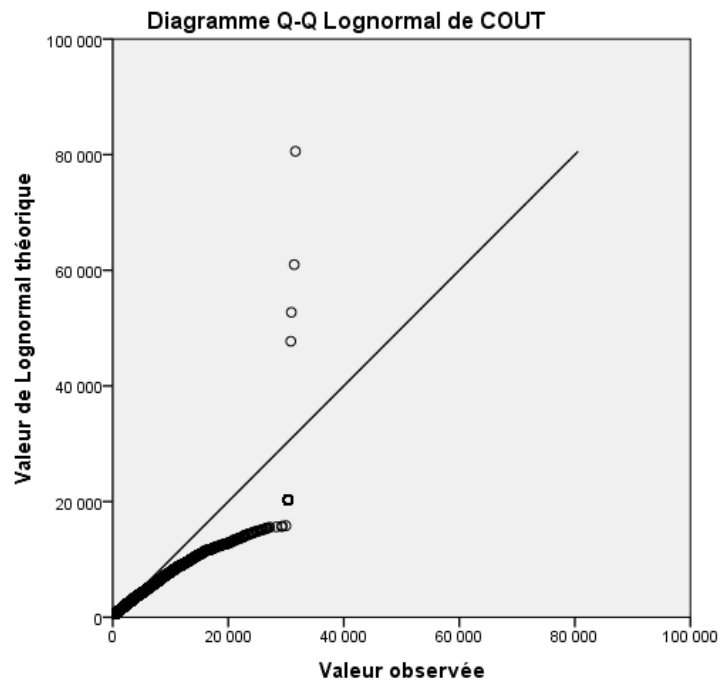
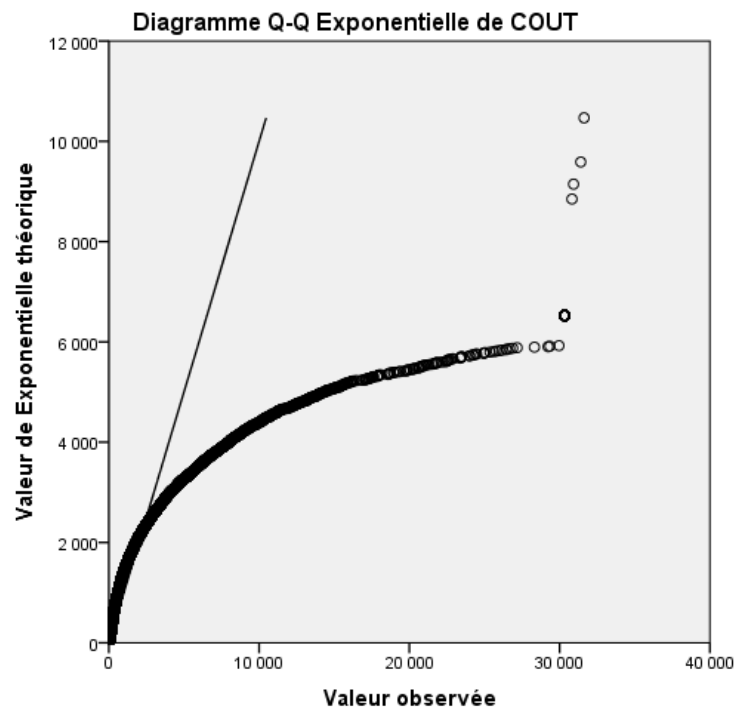


Figure 10. qq-plot exponentielle de la charge des sinistres



La figure 10 montre que la loi exponentielle ne s'ajuste pas aux données observées de la charge des sinistres. Par contre, les qq-plots de gamma et log normale montre qu'il s'agit d'un ajustement relativement bon de la distribution empirique des coûts de sinistres.

Cependant, pour accepter ou rejeter une telle loi théorique candidate à l'ajustement des données, nous allons se base sur les critères AIC et BIC. En effet, la distribution ayant un AIC ou BIC minimal, est considérée la meilleure distribution possible pour modéliser la charge des sinistres.

Tableau 16. Critères AIC et BIC des différents modèles

Loi théorique	Log Normal	Gamma
Criterion	Value	Value
AIC (smaller is better)	174715.475	833022.077
AICC (smaller is better)	174715.4864	833022.0911
BIC (smaller is better)	174866.7057	833191.0996

Source : Elaboré à partir du portefeuille maladie

Le tableau 16 montre que la loi log normale a le critère AIC le plus faible. Par conséquent, la distribution Log Normal est la meilleure loi que nous devons utiliser pour modéliser la charge des sinistres.

2. Le modèle tarifaire

Concernant la modélisation de la charge des sinistres, nous avons retenu six facteurs qui se présentent comme suit : l'âge moyen, la localisation de l'entreprise, le salaire moyen, le taux de remboursement et le sexe de l'assuré.

Afin de justifier un tel choix, l'analyse de type III permet d'évaluer le pouvoir explicatif des descripteurs potentiels de la sinistralité des assurés.

Tableau 17. Analyse de type 3 de la régression log normale

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
CLASSE_AGE	3	1470.14	<.0001
CLASSE_Remb	5	762.98	<.0001
CLASSE_SM	4	263.99	<.0001
CLASSE_SEXE	1	722.15	<.0001
CLASSE_REG	2	272.75	<.0001

Source : Elaboré à partir du portefeuille maladie

Il paraît que toutes les variables introduites pour modéliser la charge des sinistres de la prestation pharmacie sont significatives au seuil 5% (les p-value sont toutes inférieures à 5%).

Il est à préciser que le facteur âge moyen, est le facteur le plus significatif en terme de l'explication de la variation de la sinistralité. Par contre, la localisation de l'entreprise (CLASSE_REG) est le descripteur le moins significatif.

3. Analyse des résultats

3.1 Estimation des paramètres du modèle

Après la construction du modèle tarifaire, nous allons estimer les paramètres de notre modèle grâce à la méthode du maximum de vraisemblance. Les résultats se présentent comme suit :

Tableau 18. Estimation des paramètres du coût moyen

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	6.2509	0.0122	6.227	6.2748	262915	<.0001
CLASSE_AGE	AGE1	1	-0.6539	0.024	-0.7009	-0.6069	742.71	<.0001
CLASSE_AGE	AGE2	1	-0.491	0.016	-0.5224	-0.4596	939.73	<.0001
CLASSE_AGE	AGE3	1	-0.3009	0.0132	-0.3267	-0.275	519.61	<.0001
CLASSE_AGE	AGE4	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_Remb	Remb1	1	-0.6135	0.024	-0.6605	-0.5665	654.27	<.0001
CLASSE_Remb	Remb2	1	-0.2121	0.0612	-0.3319	-0.0922	12.03	0.0005
CLASSE_Remb	Remb3	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_Remb	Remb4	1	-0.0374	0.0151	-0.067	-0.0078	6.14	0.0132
CLASSE_Remb	Remb5	1	0.0955	0.0156	0.065	0.126	37.7	<.0001
CLASSE_Remb	Remb6	1	0.1978	0.0306	0.1378	0.2577	41.78	<.0001
CLASSE_SM	SM1	1	-0.3416	0.0324	-0.4052	-0.2781	110.91	<.0001
CLASSE_SM	SM2	1	-0.2457	0.0224	-0.2896	-0.2017	120.01	<.0001
CLASSE_SM	SM3	1	-0.1067	0.0146	-0.1353	-0.078	53.12	<.0001
CLASSE_SM	SM4	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_SM	SM5	1	0.0452	0.0143	0.0172	0.0732	10.04	0.0015
CLASSE_SEXE	A	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_SEXE	B	1	0.3066	0.0114	0.2843	0.3289	727	<.0001
CLASSE_REG	REG1	0	0.0000	0.0000	0.0000	0.0000	.	.
CLASSE_REG	REG2	1	-0.3003	0.0183	-0.3362	-0.2645	270.2	<.0001
CLASSE_REG	REG3	1	-0.0742	0.0182	-0.11	-0.0384	16.53	<.0001
Scale		1	1.2212	0.0037	1.2139	1.2285		

Source : Elaboré à partir du portefeuille maladie

Le tableau 18 montre que tous les paramètres du modèle considéré sont significatifs au seuil de 5%. En effet, il paraît que les p-value correspondantes à ces paramètres sont largement inférieures à 0,05 (test de wald).

3.2 Analyse des résidus

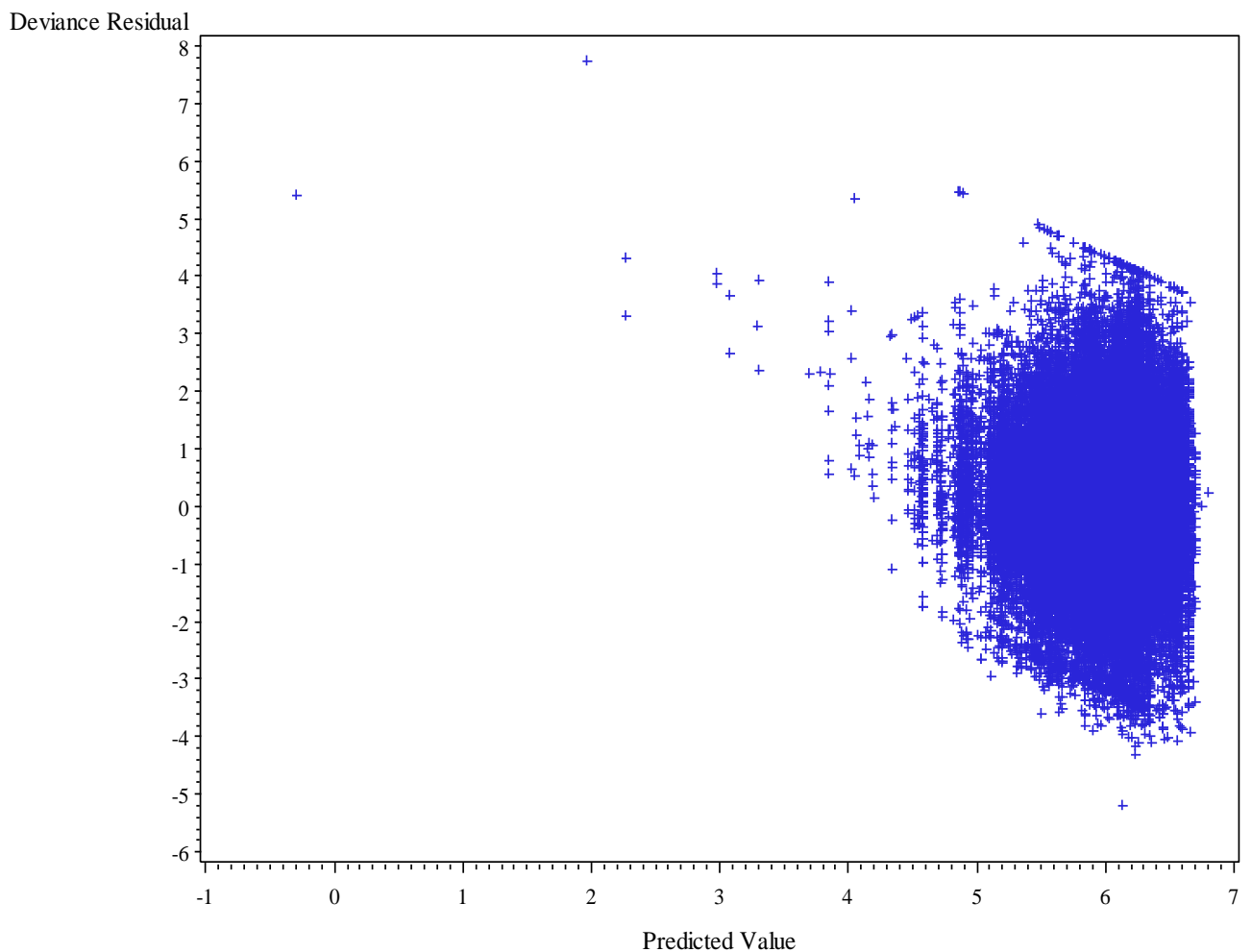
L'analyse des résidus permet une analyse plus poussée que les statistiques vues dans le paragraphe précédent. Elle permet en effet de comprendre d'où proviennent les éventuels écarts entre les valeurs prédite et les données en détectant les observations particulières.

En effet l'existence de certaines valeurs aux caractéristiques très atypiques peut biaiser fortement les coefficients calculés dans le modèle. Il conviendra donc éventuellement d'ôter ces valeurs atypiques et de relancer le modèle afin d'obtenir des résultats plus stables.

Deux types de résidus sont classiquement utilisés pour les modèles linéaires généralisés : les résidus de Pearson et les résidus de la déviance (voir le chapitre cadre conceptuel et théorique). C'est cette dernière approche que nous utilisons dans les calculs.

Pour que nous obtenions un modèle valide, il faut que les résidus soient assez proches de 0 et soient répartis de manière assez uniforme autour de l'axe des abscisses. Nous prenons ici l'exemple de la prestation pharmacie.

Figure 11. Représentation de la déviance pour le poste pharmacie



Les résidus sont correctement répartis sur l'axe des abscisses et sont d'une manière générale relativement proches de 0. Nous n'observons pas de valeurs éloignées de 0 pour les résidus de la déviance. Nous concluons alors que le modèle adopté pour modéliser la charge des sinistres est adéquat.

4. Prédiction

D'après les résultats obtenus par le modèle appliqué au poste pharmacie, nous pouvons prédire les frais moyens de la consommation suivant les caractéristiques de l'individu et de la police souscriptrice du contrat. Les frais obtenus changent de manière multiplicative en passant de la classe de référence à une autre classe (les paramètres nuls correspondent à la classe de référence). En effet, un coefficient inférieur à 0 correspond à un effet minorant des frais de consommation, alors qu'un coefficient supérieur à 0 indique un effet majorant des frais.

Prenons maintenant l'exemple d'un assuré homme qui vient de s'assurer contre le risque maladie et employé chez une entreprise ayant les caractéristiques suivantes: un taux de remboursement et un salaire moyen de la classe 1, un âge moyen et une région de la classe 3.

Cet assuré souhaite enregistrer un coût moyen de sinistres d'ordre de 288.98 DH

Tableau 19. Prédiction du cout moyen des sinistres du poste pharmacie

	COUT				
	AGE3	Remb1	SM1	A	REG3
Estimate	-0.3009	-0.6135	-0.3416	0	-0.0742
Cout Moyen	288.982092				

Source : Elaboré à partir de la table estimation des paramètres du coût moyen

CONCLUSION

Ce dernier chapitre, a mis l'accent sur la mise en place des différentes approches théoriques adoptées pour modéliser la prime de l'assurance maladie. Cette modélisation a fait appel aux techniques des modèles linéaires généralisés, comme étant une approche justifiée pour élaborer le tarif de la branche assurance maladie.

Cependant, la sélection des variables tarifaires était une étape indispensable pour aboutir à une segmentation assez précise de la population assurée. En effet, nous avons remarqué que la construction des classes de risque homogènes, repose sur le pouvoir explicatif des descripteurs de la sinistralité.

Par ailleurs, l'approche fréquence – sévérité était la méthodologie adoptée pour estimer la prime de l'assurance maladie. A cet effet, nous avons modélisé séparément le coût moyen et le nombre de sinistres afin d'isoler les effets des facteurs sur la fréquence et la charge des sinistres.

CONCLUSION GENERALE

Le présent projet de fin d'études propose une méthodologie de tarification intégrant la sinistralité des assurés comme facteur déterminant de la prime. Une analyse préliminaire de la sinistralité du portefeuille justifie l'utilisation d'une segmentation, afin de séparer les assurés en groupes intra-homogènes. Cette segmentation était élaborée grâce à l'algorithme du CHAID, comme étant un outil puissant de séparation. La segmentation était basée sur la sinistralité des assurés.

Comme nous l'avons remarqué, la survenance de sinistres graves dans une classe vient perturber l'hypothèse d'homogénéité des classes et de stabilité des indicateurs. A cet effet, l'écrêtement des données était indispensable pour corriger l'hétérogénéité des risques dans une classe donnée. Il s'agit d'étudier la loi des extrêmes ou des maximums d'une suite de variables aléatoires réelles.

Cependant, la méthodologie adoptée pour élaborer le tarif de l'assurance maladie se base sur le modèle de prime pure. C'est le résultat d'un modèle de fréquence et d'un modèle de sévérité. Ces deux modélisations font appel aux techniques des modèles linéaires généralisés.

Toutefois, il convient de signaler que la mise en place de l'approche MLG nécessite un processus de modélisation approprié. En effet, nous avons commencé par identifier les lois théoriques qui ajustent correctement les observations du nombre et de la charge des sinistres. Ensuite, nous avons déterminé la classe de référence des facteurs de risque, et ce afin de trancher quant à la nécessité de majorer ou de réduire le tarif des assurés. Ainsi, nous avons estimé séparément le coût moyen et le nombre de sinistres afin d'isoler les effets des facteurs sur la fréquence et la sévérité. Cette méthodologie d'estimation permet d'obtenir une meilleure compréhension de l'influence des facteurs sur le risque.

Une analyse approfondie des résultats de la modélisation permet de montrer l'impact du système sur la structure tarifaire du portefeuille étudié. Le modèle permet d'obtenir une prime pure qui repose sur la sinistralité des assurés. Autrement dit, le tarif obtenu a pris en considération les critères relatifs à la police souscriptrice du contrat.

Cette étude a été réalisée dans le cadre d'un stage de fin d'études actuariel au sein de la compagnie d'assurance SANAD. Notons que certains points n'ont pu être approfondis et que certaines pistes d'amélioration demeurent inexplorées :

- Les sinistres extrêmes peuvent être traités au moyen de techniques appropriées ;
- L'introduction de mesures et facteurs (mesures de sécurité, mesures de prévention, etc.) dans la récolte des données permettrait d'améliorer la vision du risque.
- L'utilisation de la théorie de la crédibilité constituerait un argument actuariel puissant pour corriger la prime pure à priori.

Enfin, le périmètre du projet de fin d'étude se limite aux primes pures. Les chargements techniques et les coûts de distribution ne sont pas étudiés. Or, les frais et les commissions de distribution sont des sujets importants afin d'agir sur la stabilité financière.

BIBLIOGRAPHIE

- [1].M. Denuit and A. Charpentier. *Mathématiques de l'assurance non-vie: Tarification et provisionnement*. Tome 2. Economica, 2005.
- [2].Ricco Rakotomalala. *Pratique de la régression: Régression logistique binaire et polytomique*. Version 2.0. Université Lumière Lyon2, 2014.
- [3].Christian Yamdjieu Ngadeu, Brehima Mariko. *Statistique des extrêmes : Théorie et application*. Université Claude Bernard, 2013.
- [4].Frédéric Bertrand. *Choix du modèle*. Université de Strasbourg. ESIEA, 2010.
- [5].Gwladys Toulemonde. *Estimation et tests en théorie des valeurs extrêmes. Mathematics*. Université Pierre et Marie Curie - Paris VI, 2008. French.
- [6].Fouad Marri. *Mathématique de l'assurance non-vie : Théorie de risque*. INSEA Rabat, 2013.
- [7].Arthur Charpentier. *Actuariat IARD: Modélisation des couts individuels de sinistres*. Université du Québec à Montréal, 2013.
- [8].Matthieu Vautrin. *Elaboration d'une méthode de tarification des contrats complémentaires santé collectifs*. Institut de Statistique de l'Université de Paris, 2009.
- [9].Olivier Jean Baptiste. *Tarification santé en Allemagne pour un produit soins complets*. Université Paris Dauphine, 2003.

ANNEXE I. MODELISATION DE LA FREQUENCE DES SINISTRES

1. Poste pharmacie

1.1. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	<.0001
CLASSE_AGE	AGE1	1	réduction	<.0001
CLASSE_AGE	AGE2	1	réduction	<.0001
CLASSE_AGE	AGE3	0	0
CLASSE_AGE	AGE4	1	majoration	<.0001
CLASSE_Remb	Remb1	1	réduction	<.0001
CLASSE_Remb	Remb2	1	réduction	<.0001
CLASSE_Remb	Remb3	0	0
CLASSE_Remb	Remb4	1	majoration	0.0117
CLASSE_Remb	Remb5	1	majoration	<.0001
CLASSE_SM	SM1	1	réduction	<.0001
CLASSE_SM	SM2	0	0
CLASSE_SM	SM3	1	majoration	<.0001
CLASSE_SM	SM4	1	majoration	<.0001
Dispersion		1		

1.2. Enfant

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	<.0001
CLASSE_AGE	AGE1	1	majoration	<.0001
CLASSE_AGE	AGE2	1	majoration	<.0001
CLASSE_AGE	AGE3	0	0
CLASSE_AGE	AGE4	1	réduction	<.0001
CLASSE_Remb	Remb1	1	réduction	<.0001
CLASSE_Remb	Remb2	1	réduction	0.018
CLASSE_Remb	Remb3	0	0
CLASSE_Remb	Remb4	1	majoration	<.0001
CLASSE_SM	SM1	1	réduction	<.0001
CLASSE_SM	SM2	0	0
CLASSE_SM	SM3	1	majoration	<.0001
CLASSE_SM	SM4	1	majoration	<.0001
Dispersion		1	

2. Poste consultation

2.1. Adhérent

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	<.0001
CLASSE_AGE	AGE1	1	Réduction	<.0001
CLASSE_AGE	AGE2	1	Réduction	<.0001
CLASSE_AGE	AGE3	1	Réduction	<.0001
CLASSE_AGE	AGE4	0	0	
CLASSE_SM	SM1	1	Réduction	<.0001
CLASSE_SM	SM2	1	Réduction	<.0001
CLASSE_SM	SM3	1	Réduction	<.0001
CLASSE_SM	SM4	0	0	
CLASSE_SM	SM5	1	Majoration	<.0001
Dispersion		1	

2.2. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	38.17	<.0001
CLASSE_AGE	AGE1	1	Réduction	.	.	.	77.53	<.0001
CLASSE_AGE	AGE2	1	Réduction	.	.	.	59.4	<.0001
CLASSE_AGE	AGE3	0	0	.	.	.		
CLASSE_AGE	AGE4	1	Majoration	.	.	.	721.82	<.0001
CLASSE_SM	SM1	1	Réduction	.	.	.	2623.07	<.0001
CLASSE_SM	SM2	1	Réduction	.	.	.	853.22	<.0001
CLASSE_SM	SM3	1	Réduction	.	.	.	121.04	<.0001
CLASSE_SM	SM4	0	0	.	.	.		
CLASSE_SM	SM5	1	Majoration	.	.	.	263.26	<.0001
Dispersion		1		

2.3. Enfant

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	398.85	<.0001
CLASSE_AGE	AGE1	1	majoration	.	.	341.18	<.0001
CLASSE_AGE	AGE2	1	majoration	.	.	24.68	<.0001
CLASSE_AGE	AGE3		0	.	.		
CLASSE_AGE	AGE4	1	réduction	.	.	27.76	<.0001
CLASSE_AGE	AGE5	1	réduction	.	.	227.84	<.0001
CLASSE_Remb	Remb1	1	réduction	.	.	3889.14	<.0001
CLASSE_Remb	Remb2	1	réduction	.	.	363.81	<.0001
CLASSE_Remb	Remb3	1	0	.	.	363.81	<.0001
CLASSE_Remb	Remb4	1	majoration	.	.	91.3	<.0001
CLASSE_Remb	Remb5	1	majoration	.	.	13.18	0.0003
CLASSE_SM	SM1	1	réduction	.	.	1234.1	<.0001
CLASSE_SM	SM2	1	réduction	.	.	1193.3	<.0001
CLASSE_SM	SM3	1	réduction	.	.	143.42	<.0001
CLASSE_SM	SM4		0	.	.		
CLASSE_SM	SM5	1	majoration	.	.	1073.25	<.0001
Dispersion		1	.	.	.		

3. Poste analyse et radiologie

3.1. Adhérent

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	904.58	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	2221.76	<.0001
CLASSE_AGE	AGE2	1	réduction	.	.	1022.79	<.0001
CLASSE_AGE	AGE3	1	réduction	.	.	318.65	<.0001
CLASSE_AGE	AGE4	0	0				
CLASSE_AGE	AGE5	1	majoration	.	.	810.44	<.0001
CLASSE_SM	SM1	1	réduction	.	.	2112.13	<.0001
CLASSE_SM	SM2	1	réduction	.	.	344.77	<.0001
CLASSE_SM	SM3	0	0				
CLASSE_SM	SM4	1	majoration	.	.	124.05	<.0001
CLASSE_SM	SM5	1	majoration	.	.	1909.67	<.0001
Dispersion		1	.	.	.		

3.2. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	267.13	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	97.65	<.0001
CLASSE_AGE	AGE2	0	0	.	.		
CLASSE_AGE	AGE3	1	majoration	.	.	209.43	<.0001
CLASSE_AGE	AGE4	1	majoration	.	.	788.17	<.0001
CLASSE_SM	SM1	1	réduction	.	.	2011.45	<.0001
CLASSE_SM	SM2	1	réduction	.	.	700.29	<.0001
CLASSE_SM	SM3	1	réduction	.	.	288.83	<.0001
CLASSE_SM	SM4	0	0	.	.		
CLASSE_SM	SM5	1	majoration	.	.	202.93	<.0001
Dispersion		1	.	.	.		

3.3. Enfant

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	6003.48	<.0001
CLASSE_AGE	AGE1	1	majoration	.	.	317.97	<.0001
CLASSE_AGE	AGE2	1	majoration	.	.	129.96	<.0001
CLASSE_AGE	AGE3	1	majoration	.	.	157.88	<.0001
CLASSE_AGE	AGE4	0	0	.	.		
CLASSE_SM	SM1	1	réduction	.	.	1889.25	<.0001
CLASSE_SM	SM2	1	réduction	.	.	1098.91	<.0001
CLASSE_SM	SM3	1	réduction	.	.	1095.69	<.0001
CLASSE_SM	SM4	0	0	.	.		
Dispersion		1	.	.	.		

4. Poste optique

4.1. Adhérent

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	9262.03	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	639.81	<.0001
CLASSE_AGE	AGE2	1	réduction	.	.	391.63	<.0001
CLASSE_AGE	AGE3		0	.	.		
CLASSE_AGE	AGE4	1	majoration	.	.	162.26	<.0001
CLASSE_MONT	MONT1	1	réduction	.	.	490.58	<.0001
CLASSE_MONT	MONT2		0	.	.		
CLASSE_MONT	MONT3	1	majoration	.	.	13.4	0.0003
CLASSE_SM	SM1	1	réduction	.	.	612.16	<.0001
CLASSE_SM	SM2	1	réduction	.	.	39.5	<.0001
CLASSE_SM	SM3		0	.	.		
CLASSE_SM	SM4	1	majoration	.	.	1242.57	<.0001
Dispersion		1	0	.	.		

4.2. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	5055.46	<.0001
CLASSE_AGE	AGE1	0	0	.	.		
CLASSE_AGE	AGE2	1	majoration	.	.	116.04	<.0001
CLASSE_AGE	AGE3	1	majoration	.	.	81.27	<.0001
CLASSE_AGE	AGE4	1	majoration	.	.	765.78	<.0001
CLASSE_MONT	MONT1	1	réduction	.	.	256.69	<.0001
CLASSE_MONT	MONT2	0	0	.	.		
CLASSE_MONT	MONT3	1	majoration	.	.	12.89	0.0003
CLASSE_SM	SM1	1	réduction	.	.	272.66	<.0001
CLASSE_SM	SM2	1	réduction	.	.	39.73	<.0001
CLASSE_SM	SM3	0	0	.	.		
CLASSE_SM	SM4	1	majoration	.	.	140.28	<.0001
Dispersion		1	0	.	.		

4.3. Enfant

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	11441.7	<.0001
CLASSE_AGE	AGE1	0	0	.	.	.	
CLASSE_AGE	AGE2	1	réduction	.	.	11.91	0.0006
CLASSE_AGE	AGE3	1	réduction	.	.	12.81	0.0003
CLASSE_MONT	MONT1	1	réduction	.	.	256.96	<.0001
CLASSE_MONT	MONT2	0	0	.	.	.	
CLASSE_MONT	MONT3	1	majoration	.	.	8.12	0.0044
CLASSE_SM	SM1	1	réduction	.	.	239.65	<.0001
CLASSE_SM	SM2	1	réduction	.	.	68.85	<.0001
CLASSE_SM	SM3	0	0	.	.	.	
CLASSE_SM	SM4	1	majoration	.	.	296.87	<.0001
Dispersion		1	0	.	.	.	

5. Poste dentaire

5.1 Adhérent

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	.	.	.	12877.4	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	268.94	<.0001
CLASSE_AGE	AGE2	0	0	.	.	.	
CLASSE_AGE	AGE3	1	majoration	.	.	189.01	<.0001
CLASSE_AGE	AGE4	1	majoration	.	.	262.68	<.0001
CLASSE_SM	SM1	1	réduction	.	.	1009.23	<.0001
CLASSE_SM	SM2	0	0	.	.	.	
CLASSE_SM	SM3	1	majoration	.	.	462.31	<.0001
CLASSE_SM	SM4	1	majoration	.	.	3189.86	<.0001
Scale		0	1	.	.	.	

5.2. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	3936.16	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	.	41.37	<.0001
CLASSE_AGE	AGE2	0	0	.	.	.		
CLASSE_AGE	AGE3	1	majoration	.	.	.	228.16	<.0001
CLASSE_AGE	SM1	1	réduction	.	.	.	433.68	<.0001
CLASSE_SM	SM2	0	0	.	.	.		
CLASSE_SM	SM3	1	majoration	.	.	.	106.19	<.0001
CLASSE_SM	SM4	1	majoration	.	.	.	470.11	<.0001
Dispersion		1		

5.3. Enfant

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	8877.26	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	.	18.73	<.0001
CLASSE_AGE	AGE2	0		
CLASSE_AGE	AGE3	1	majoration	.	.	.	17.96	<.0001
CLASSE_SM	SM1	1	réduction	.	.	.	474.47	<.0001
CLASSE_SM	SM2	0		
CLASSE_SM	SM3	1	majoration	.	.	.	131.23	<.0001
CLASSE_SM	SM4	1	majoration	.	.	.	940.49	<.0001
Dispersion		1		

ANNEXE II. MODELISATION DE LA CHARGE DES SINISTRES

1. Poste Pharmacie

1.1. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	182917	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	.	28.63	<.0001
CLASSE_AGE	AGE2	1	réduction	.	.	.	6.34	0.0118
CLASSE_AGE	AGE3	0	0	.	.	.		
CLASSE_AGE	AGE4	1	majoration	.	.	.	567.9	<.0001
CLASSE_Remb	Remb1	1	réduction	.	.	.	143.42	<.0001
CLASSE_Remb	Remb2	1	réduction	.	.	.	24.86	<.0001
CLASSE_Remb	Remb3	0	0	.	.	.		
CLASSE_Remb	Remb4	1	majoration	.	.	.	30.43	<.0001
CLASSE_Remb	Remb5	1	majoration	.	.	.	180.34	<.0001
Scale		1		

1.2. Enfant

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	215702	<.0001
CLASSE_Remb	Remb1	1	réduction	.	.	.	144.96	<.0001
CLASSE_Remb	Remb2	1	réduction	.	.	.	13.1	0.0003
CLASSE_Remb	Remb3	0	0	.	.	.		
CLASSE_Remb	Remb4	1	majoration	.	.	.	125.24	<.0001
CLASSE_SM	SM1	1	réduction	.	.	.	4.15	0.0416
CLASSE_SM	SM2	0	0	.	.	.		
CLASSE_SM	SM3	1	majoration	.	.	.	114.87	<.0001
CLASSE_SM	SM4	1	majoration	.	.	.	159.62	<.0001
Scale		1		

2. Poste Consultation

2.1. Adhérent

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	665785	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	.	221.97	<.0001
CLASSE_AGE	AGE2	1	réduction	.	.	.	386.56	<.0001
CLASSE_AGE	AGE3	1	réduction	.	.	.	153.52	<.0001
CLASSE_AGE	AGE4	0	0	.	.	.		
CLASSE_SM	SM1	1	réduction	.	.	.	1486.44	<.0001
CLASSE_SM	SM2	1	réduction	.	.	.	562.16	<.0001
CLASSE_SM	SM3	1	réduction	.	.	.	108.09	<.0001
CLASSE_SM	SM4	0	0	.	.	.		
CLASSE_SM	SM5	1	majoration	.	.	.	237.93	<.0001
Scale		1	0.8047	.	.	.		

2.2. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	288664	<.0001
CLASSE_AGE	AGE1	1	réduction	.	.	.	24.77	<.0001
CLASSE_AGE	AGE2	1	réduction	.	.	.	25.6	<.0001
CLASSE_AGE	AGE3	0	0
CLASSE_AGE	AGE4	1	majoration	.	.	.	125.42	<.0001
CLASSE_SM	SM1	1	réduction	.	.	.	167.71	<.0001
CLASSE_SM	SM2	1	réduction	.	.	.	278.86	<.0001
CLASSE_SM	SM3	1	réduction	.	.	.	76.16	<.0001
CLASSE_SM	SM4	0	0
CLASSE_SM	SM5	1	majoration	.	.	.	39.74	<.0001
Scale		1		

2.3. Enfant

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square Pr > ChiSq
Intercept		1	333080 <.0001
CLASSE_AGE	AGE1	1	majoration	.	.	.	141.6 <.0001
CLASSE_AGE	AGE2	1	majoration	.	.	.	23.41 <.0001
CLASSE_AGE	AGE3	0	0
CLASSE_AGE	AGE4	1	réduction	.	.	.	15.28 <.0001
CLASSE_AGE	AGE5	1	réduction	.	.	.	11.14 0.0008
CLASSE_Remb	Remb1	1	réduction	.	.	.	486.45 <.0001
CLASSE_Remb	Remb2	0	0
CLASSE_Remb	Remb3	1	majoration	.	.	.	37.15 <.0001
CLASSE_Remb	Remb4	1	majoration	.	.	.	150.78 <.0001
CLASSE_Remb	Remb5	1	majoration	.	.	.	85.76 <.0001
CLASSE_SM	SM1	1	réduction	.	.	.	97.91 <.0001
CLASSE_SM	SM2	1	réduction	.	.	.	197.84 <.0001
CLASSE_SM	SM3	1	réduction	.	.	.	42.43 <.0001
CLASSE_SM	SM4	0	0
CLASSE_SM	SM5	1	majoration	.	.	.	170.7 <.0001
Scale		1	0.8256

3. Poste optique

3.1 Adhérent

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square Pr > ChiSq
Intercept		1	1276252 <.0001
CLASSE_MONT	MONT1	1	réduction	.	.	.	3225.54 <.0001
CLASSE_MONT	MONT2	0	0
CLASSE_MONT	MONT3	1	majoration	.	.	.	2300.26 <.0001
CLASSE_SM	SM1	1	réduction	.	.	.	96.36 <.0001
CLASSE_SM	SM2	1	réduction	.	.	.	88.25 <.0001
CLASSE_SM	SM3	0	0
CLASSE_SM	SM4	1	majoration	.	.	.	313.59 <.0001
Scale		1

3.2. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	254177	<.0001
CLASSE_MONT	MONT1	1	réduction	.	.	.	628.06	<.0001
CLASSE_MONT	MONT2	0	0	.	.	.		
CLASSE_MONT	MONT3	1	majoration	.	.	.	452.43	<.0001
CLASSE_SM	SM1	1	réduction	.	.	.	3.91	0.048
CLASSE_SM	SM2	1	réduction	.	.	.	11.2	0.0008
CLASSE_SM	SM3	0	0	.	.	.		
CLASSE_SM	SM4	1	majoration	.	.	.	12.27	0.0005
Scale		1		

3.3. Enfant

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	238946	<.0001
CLASSE_MONT	MONT1	1	réduction	.	.	.	664.58	<.0001
CLASSE_MONT	MONT2	0	0	.	.	.		
CLASSE_MONT	MONT3	1	majoration	.	.	.	333.75	<.0001
CLASSE_SM	SM1	1	réduction	.	.	.	15.08	0.0001
CLASSE_SM	SM2	1	réduction	.	.	.	10.05	0.0015
CLASSE_SM	SM3	0	0	.	.	.		
CLASSE_SM	SM4	1	majoration	.	.	.	21.77	<.0001
Scale		1		

4 Poste Dentaire

4.1. Adhérent

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	510076	<.0001
CLASSE_DENT	DENT1	1	réduction	.	.	.	48.39	<.0001
CLASSE_DENT	DENT2	0	0	.	.	.		
CLASSE_DENT	DENT3	1	majoration	.	.	.	42.89	<.0001
CLASSE_DENT	DENT4	1	majoration	.	.	.	148.66	<.0001
CLASSE_DENT	DENT5	1	majoration	.	.	.	27.52	<.0001
CLASSE_DENT	DENT6	1	majoration	.	.	.	86.31	<.0001
Scale		1		

4.2. Conjoint

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	196179	<.0001
CLASSE_DENT	DENT1	1	réduction	.	.	.	10.7	0.0011
CLASSE_DENT	DENT2	0	0	.	.	.		
CLASSE_DENT	DENT3	1	majoration	.	.	.	33.97	<.0001
CLASSE_DENT	DENT4	1	majoration	.	.	.	46.96	<.0001
CLASSE_DENT	DENT5	1	majoration	.	.	.	17.27	<.0001
Scale		1		

4.3. Enfant

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	90612.4	<.0001
CLASSE_SM	SM1	1	réduction	.	.	.	4.6	0.032
CLASSE_SM	SM2	0	0	.	.	.		
CLASSE_SM	SM3	1	majoration	.	.	.	4.1	0.0429
CLASSE_SM	SM4	1	majoration	.	.	.	7.43	0.0064
CLASSE_DENT	DENT1	1	réduction	.	.	.	6.68	0.0097
CLASSE_DENT	DENT2	0	0	.	.	.		
CLASSE_DENT	DENT3	1	majoration	.	.	.	36.44	<.0001
CLASSE_DENT	DENT4	1	majoration	.	.	.	49.43	<.0001
CLASSE_DENT	DENT5	1	majoration	.	.	.	64.75	<.0001
Scale		1		

ANNEXE III. EXTRAITS DES PROCEDURES SAS

1. Modélisation de la fréquence des sinistres : Modèle Binomiale négative

```
/* NOMBRE DE SINISTRE */
```

```
proc genmod data=pharma_a;  
class CLASSE_AGE(param=reference ref="AGE4") CLASSE_Remb(param=reference  
ref="Remb3") CLASSE_Plaf(param=reference ref="Plaf2")  
CLASSE_SM(param=reference ref="SM4") CLASSE_SEXE(param=reference ref="A")  
CLASSE_REG(param=reference ref="REG1") ;  
model NB_SIN=CLASSE_AGE CLASSE_Remb CLASSE_SM CLASSE_SEXE  
CLASSE_REG / dist=NB link=log offset=log_expo type3;  
make 'modelfit' out=negbinmodel;  
run;  
data fitnbinom;  
set negbinmodel;  
pvalue=1-cdf('CHISQ',value,df);  
run;  
proc print data=fitnbinom;  
run;
```

2. Modélisation du coût moyen des sinistres : Modèle log normale

```
/* CHARGE DES SINISTRES */
```

```
proc genmod data=pharma_a;  
class CLASSE_AGE(param=reference ref="AGE4") CLASSE_Remb(param=reference  
ref="Remb3") CLASSE_Plaf(param=reference ref="Plaf2")  
CLASSE_SM(param=reference ref="SM4") CLASSE_SEXE(param=reference ref="A")  
CLASSE_REG(param=reference ref="REG1") ;  
model LOGA=CLASSE_AGE CLASSE_Remb CLASSE_SM CLASSE_SEXE  
CLASSE_REG / dist=normal link=id offset=log_expo type3;  
make 'modelfit' out=NORMALMODEL;  
run;  
data FITNORM;  
set NORMALMODEL;  
pvalue=1-cdf('CHISQ',value,df);  
run;  
proc print data=FITNORM;
```

run;

3. Graphique des résidus

```
proc genmod data=pharma_c
```

```
class CLASSE_AGE(param=reference ref="AGE3") CLASSE_Remb(param=reference  
ref="Remb2") CLASSE_Plaf(param=reference ref="Plaf2")
```

```
CLASSE_SM(param=reference ref="SM4") CLASSE_SEXE(param=reference ref="B")  
CLASSE_REG(param=reference ref="REG1") ;
```

```
model LOGA=CLASSE_AGE CLASSE_Remb CLASSE_SM CLASSE_SEXE  
CLASSE_REG / dist=normal link=id offset=log_expo type3;
```

```
output OUT=RESIDUC RESDEV=RdevianceC RESCHI=RpearsonC PRED=MUchapeauC;
```

run;

```
proc GPLOT data=RESIDUC;
```

```
plot RdevianceC*MUchapeauC;
```

run;

4. Calcul de la prime pure empirique par prestation

```
proc sql;
```

```
create table maladie.Prest_Poste as
```

```
select Poste, LIEN, sum(Exposition) as Expo_T, sum(NB_SIN) as NB_ACT, sum(CHARGE)  
as CHARGE_T
```

```
from Portefeuille
```

```
group by Poste, LIEN
```

```
;
```

```
quit;
```

```
data maladie.Prest_Poste;
```

```
set maladie.Prest_Poste;
```

```
Freq=NB_ACT/Expo_T;
```

```
CM=CHARGE_T/NB_ACT;
```

```
PP=Freq*CM;
```

```
Run;
```

