



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

Modélisation de la consommation médicale de la population des bénéficiaires de l'AMO du secteur public au Maroc

Préparé par : Mlle Mechrafi Yousra

Mlle Benkhalla Mouna

Sous la direction de : M. Abdelaziz CHAOUBI (INSEA)

Mlle. Nabila BOUZZEGAOUI (ANAM)

**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. Mustapha BERROUYNE (INSEA)
- ❑ M. Abdelaziz CHAOUBI (INSEA)
- ❑ Mlle. Nabila BOUZZEGAOUI (ANAM)

Vu le caractère confidentiel des données fournies, les chiffres figurant dans le présent rapport sont fictifs et sont présentés à titre indicatif.

Résumé

De nombreuses études empiriques ont montré que les personnes qui bénéficient d'une couverture maladie ont des dépenses de santé plus élevées que celles des personnes non assurées. En première analyse, on peut penser que ce phénomène est la manifestation la plus naturelle de la présence d'une assurance santé qui permet à l'individu de solvabiliser une consommation de soins en cas de maladie. Ce phénomène pousse les organismes assureurs à gérer efficacement et continuellement les risques auxquels ils s'exposent en remboursant annuellement des montants très élevés à leur population assurée.

L'objectif de ce mémoire est donc de définir une approche permettant de mesurer l'engagement de la Caisse Nationale des Organismes de Prévoyance Sociale vis-à-vis de ses assurés à travers la modélisation des remboursements annuels probables, dont elle pourra faire face, afin d'apprécier les risques encourus et de juger de sa viabilité financière.

L'étude porte sur la consommation médicale en 2014 de la population assurée de la CNOPS. La variable modélisée étant la consommation annuelle en soins médicaux d'un segment de bénéficiaires ayant les mêmes modalités de variables exogènes. Nous nous penchons dans l'étude sur l'élaboration d'un modèle du type « fréquence * coût * population » où les fréquences de survenance des sinistres par segment de bénéficiaire et le coût moyen par sinistre seront étudiés séparément en se basant sur des modèles linéaires généralisés. Cette approche est la plus couramment utilisée aujourd'hui par les protagonistes du secteur car elle est relativement aisée à mettre en œuvre et permet une estimation cohérente des risques considérés.

La modélisation impose de bien connaître la population à couvrir. En outre, les dépenses d'un régime d'assurance sont étroitement liées à la taille et à la structure de sa population qui peut connaître des changements dans le futur. Nous proposons alors de mettre en œuvre un modèle démographique basé sur la loi « entrée-sortie » pour modéliser les différents flux démographiques que peut connaître la population cible de notre étude. Cette étape nous permettra d'étudier les projections de ladite population, qui serviront par la suite à calculer les dépenses futures du régime.

Mots clés :

Consommation médicale, remboursements, approche Fréquence/ Coût / Population, les modèles linéaires généralisés, loi « entrée-sortie », modèle démographique.



Dédicace

*À mes très chers parents qui ont toujours fait en sorte de m'offrir les
meilleures conditions pour assurer mon bien être et faire de moi ce que
je suis aujourd'hui*

*Je ne saurais trouver les mots pour vous exprimer ma profonde
gratitude pour tout l'amour que vous m'avez porté et pour tous les
sacrifices que vous avez faits pour moi.*

*À mon petit frère Anas à qui je souhaite un avenir plein de bonheur,
de réussite et de sévérité,*

À toute ma grande famille,

À la mémoire de mes grands-parents,

*À ma chère binôme Mouna qui est très studieuse, enthousiasmée et
dynamique, et à qui je souhaite plein de succès et de réussite,*

*À tous mes amis de l'INSEA avec qui j'ai partagé le meilleur et le pire
durant ces merveilleuses années de formations,*

*À tous ceux qui m'ont aidé de proche ou de loin pour réussir mon
cursus,*

À tous ceux que je n'ai pas cités, mais à qui l'affection

Que je porte n'a pas besoin de citation

À tous ceux que j'aime....

Yousra





Dédicace

À mes chers parents

*Que ce travail soit le fruit de toutes vos peines et vos sacrifices,
Acceptez ce travail comme témoignage de l'estime, du respect et du
grand amour que j'éprouve pour vous.*

À mon cher frère Abdou

Avec toutes mes affections, et mes souhaits de bonheur et de réussite.

Je vous remercie pour tout ce que vous avez fait pour moi.

À toute ma famille,

*À ma chère binôme Yousra pour son sérieux, sa motivation et son
soutien et à toute sa famille,*

*À tous mes amis de l'INSEA avec qui j'ai partagé des moments
précieux,*

*À tous ceux dont les bienfaits et l'amitié resteront à jamais dans ma
mémoire,*

*Qu'ils trouvent dans ce mémoire toute ma reconnaissance et ma
considération.*

Mouna



Remerciements

Au terme de ce travail, nous remercions tout d'abord M.CHAOUBI Abdelaziz, pour nous avoir fait l'honneur de nous encadrer, pour son implication, pour ses conseils avisés tout au long de notre stage.

Nous exprimons notre gratitude et nos sentiments de reconnaissance à Mr.HAZIM Jilali, Directeur Général de l'Agence Nationale d'Assurance Maladie, de nous avoir acceptées en tant que stagiaires, ainsi qu'à Mr.MOUHDI Hicham, Chef du département Etudes Economiques et Actuarielles, de nous avoir accueillies au sein de son équipe.

Nos remerciements s'adressent également à notre encadrante externe, Mlle Nabila BOUZEGGAOUI, chef du service Monitoring des Paramètres Financiers, pour sa constante disponibilité, ses efforts et son encouragement qui ont fait de notre stage une expérience enrichissante.

Nos remerciements vont aussi à l'ensemble du personnel du département Etudes Economiques et Actuarielles pour leur coopération et l'accueil chaleureux qu'ils nous ont réservé durant notre projet de fin d'étude.

Nous tenons à remercier M. Mustapha BERROUYNE, professeur à l'INSEA pour avoir accepté d'évaluer ce travail.

Nous adressons nos remerciements à l'ensemble du corps professoral de l'INSEA et à notre Directeur Abdesselam FAZOUANE pour leurs efforts continus afin de nous offrir la meilleure formation, ainsi que toute autre personne ayant participé, à la réussite de notre projet de fin d'études.

Table des matières

Résumé	4
Remerciements	7
Liste des abréviations	11
Liste des tableaux	12
Liste des Figures.....	13
Introduction	14
Chapitre 1 : Contexte de l'étude.....	16
I. Présentation de l'organisme d'accueil.....	16
II. L'Assurance Maladie Obligatoire au Maroc	18
1. Présentation de l'AMO	18
2. Cadre général de l'AMO	18
3. Principes de l'AMO	19
4. La population couverte	19
5. Ouverture des droits.....	19
6. Affection de longue durée (ALD)	20
7. Le taux de cotisation de l'AMO	20
8. Panier de soins et taux de couverture de l'AMO.....	21
Chapitre 2 : Etude descriptive et analyse des données	24
I. Présentation des données utilisées.....	24
II. Structure des données	24
1. La table de la population assurée « fichier effectif ».....	24
2. La table de la population consommatrice « fichier prestation ».....	25
III. La constitution des bases de données	26
IV. Analyse descriptive.....	27
1. Etude de la population sous risque	27
2. Etude de la population consommatrice.....	31
3. Analyse de la fréquence des sinistres	34
4. Méthode de CHAID appliquée au taux de sinistralité.....	36
V. Mesures d'association	39
1. Le test de chi-deux et le V de Cramer	39
2. Analyse des résultats obtenus	40

VI.	Analyse multidimensionnelle	41
1.	L'analyse factorielle des correspondances (AFC)	42
2.	Analyse factorielle des correspondances multiples « ACM »	46
VII.	Analyse bivariée du taux de sinistralité et du montant remboursé moyen	52
Chapitre 3 : Modélisation de la consommation médicale		55
I.	Introduction	55
1.	Les données utilisées	55
2.	L'approche retenue: l'approche « Fréquence * Coût moyen * population »	56
II.	Méthodes de modélisation appliquées en assurance maladie.....	56
1.	La Régression linéaire multiple	56
2.	Mise en œuvre de la régression linéaire multiple	59
3.	Les modèle linéaires généralisés GLM	64
4.	Les modèles additifs généralisés « GAM »	68
5.	Mise en œuvre des modèles linéaires généralisés	69
6.	La modélisation du montant remboursé moyen	74
7.	La modélisation de la fréquence moyenne des sinistres.....	80
Chapitre 4 : Modélisation de la population par la loi « entrée-sortie »		87
1.	Paramètres et notations.....	87
2.	Mise en équation des flux démographiques	90
2.1.	Assurés actifs	90
2.2.	Retraités	90
2.3.	Conjoints	91
2.4.	Les veufs (ves)	93
2.5.	Les enfants	93
3.	Interprétation des résultats	94
Chapitre 5 : Les résultats obtenus		96
1.	Utilisation des résultats	97
2.	Backtesting	99
3.	Conclusion.....	101
Conclusion générale		102
Bibliographie et Webographie :		103
Annexes :.....		104
Annexe 1 : Organigramme de l'ANAM		104

Annexe 2 : Liste des ALD	105
Annexe 3 : Sortie de L'AFC	106
Annexe 4 : CODE SAS POUR LA PROC GENMOD (MODELES LINEAIRES GENERALISES)	108
Annexe 5 : CODE SAS POUR LA PROC REG (REGRESSION LINEAIRE MUPLTIPLE)	112
Annexe 6 : CODE SAS POUR LA DUPLICATION DE LA TABLE « POPULATION »	114

Liste des abréviations

ANAM : Agence Nationale de l'Assurance Maladie

AMO : Assurance Maladie Obligatoire

CNOPS : Caisse Nationale des Organismes de Prévoyance Sociale

CNSS : Caisse Nationale de Sécurité Sociale

ALD : Affection de Longue Durée

AFC : Analyse Factorielle des Correspondances

ACM : Analyse des Correspondances Multiples

CHAID : Chi squared Automatic Interaction Detector

GLM : Generalized Linear Model

GAM : Generalized Additive Model

AIC : Akaike Information Criterion

Liste des tableaux

Tableau 1 : Prestations couvertes et niveaux de couverture pour la CNOPS.....	22
Tableau 2 : Prestations couvertes et niveaux de couverture pour la CNSS.....	23
Tableau 3 : Structure de la base de données de la population de l'AMO- CNOPS de 2014.....	24
Tableau 4 : Structure de la base de données de la consommation de l'AMO - CNOPS de l'année 2014.	25
Tableau 5 : Mesure du test de Khi-Deux de Pearson	40
Tableau 6 : Mesure de V de cramer.....	41
Tableau 7 : Tableau des correspondances des variables type_ass et type_benef	43
Tableau 8 : Tableau des correspondances des variables type_benef et ALD.....	43
Tableau 9 : Tableau des correspondances des variables type_benef et Sexe	43
Tableau 10 : Tableau des correspondances des variables Sexe et ALD.....	43
Tableau 11 : Inertie par dimension dans le cas des variables type_ass et type_benef.....	44
Tableau 12 : Tableau des caractéristiques des profils lignes.....	45
Tableau 13 : Tableau des caractéristiques des profils colonnes	45
Tableau 14 : Récapitulatif des modèles.....	47
Tableau 15 : Coordonnées des modalités sur les 4 premiers facteurs identifiés par l'ACM.....	49
Tableau 16 : Résultat de l'ANOVA à un facteur « type_ass ».....	52
Tableau 17 : Résultat du test de Duncan pour le cas de la variable « type_ass ».....	53
Tableau 18 : Résultat du test de Duncan pour le cas de la variable « type_benef ».....	53
Tableau 19 : Résultat de l'ANOVA à un facteur « type_ass ».....	54
Tableau 20 : Valeurs estimés des paramètres.....	60
Tableau 21 : Statistiques d'ajustement.....	61
Tableau 22 : Fonctions de lien pour les lois de la famille exponentielle.....	67
Tableau 23 : Résultats des calculs SAS à l'issue de la 1ère étape de la méthode de segmentation de la région.....	70
Tableau 24 : Résultats des calculs SAS à l'issue de la dernière étape de la méthode segmentation de la région.....	71
Tableau 25 : Résultats des calculs SAS à l'issue de la 1ère étape de la méthode segmentation de la tranche âge.....	72
Tableau 26 : Résultats des calculs SAS à l'issue de la dernière étape de la méthode segmentation de la tranche âge.....	72
Tableau 27 : Résultat des tests du rapport de vraisemblance	76
Tableau 28 : Les paramètres estimés de loi Log-Normale	77
Tableau 29 : Les paramètres estimés de la loi Poisson.....	82
Tableau 30 : Critère d'évaluation de l'adéquation à la loi de poisson	82
Tableau 31 : Critère d'évaluation de l'adéquation à la loi binomiale-négative.....	83
Tableau 32 : Les paramètres estimés de la loi Binomiale-Négative.....	84
Tableau 33 : Comparaisons 2 à 2 des modalités en termes de moyenne de la fréquence des sinistres .	85
Tableau 34 : Les paramètres estimés de la loi Binomiale-Négative.....	86
Tableau 35 : Les paramètres estimés pour les lois Log normale et Binomiale négative.....	98
Tableau 36 : Résultats du GLM au titre de l'année 2014.....	99

Liste des Figures

Figure 1 : Répartition de la population AMO-CNOPS selon le type d'assuré.....	28	
Figure 2 : Répartition de la population AMO-CNOPS selon le type du bénéficiaire	28	
Figure 3 : Répartition de la population AMO-CNOPS selon le sexe.....	29	
Figure 4 : Répartition de la population AMO-CNOPS selon l'âge.....	29	
Figure 5 : Pyramide des âges de la population AMO-CNOPS.....	30	
Figure 6 : Répartition de la population AMO-CNOPS selon l'atteinte d'une affection de longue durée	30	
Figure 7 : Répartition des montants remboursés moyens selon le type d'assuré	31	
Figure 8 : Répartition des montants remboursés moyens selon le type du bénéficiaire.....	32	
Figure 9 : Répartition des montants remboursés moyens selon le sexe	32	
Figure 10 : Répartition des montants remboursés moyens selon l'âge	33	
Figure 11 : Répartition des montants remboursés selon l'atteinte d'une affection de longue durée	33	
Figure 12 : Taux de sinistralité de la population AMO-CNOPS selon le sexe	34	
Figure 13 : Taux de sinistralité de la population AMO-CNOPS selon l'âge	35	
Figure 14 : Taux de sinistralité de la population AMO-CNOPS selon l'état de santé du bénéficiaire ..	35	
Figure 15 : Importance des variables indépendantes sur la variable cible taux de sinistralité	36	
Figure 16 : Représentation de l'arbre de décision sous SPSS MODELER.....	37	
Figure 17 : Le chemin « si ald=N ET tranche_age=majo ET type_benef=A ET type_ass=A »	38	
Figure 18 : Représentation des modalités dans le premier plan factoriel	51	
Figure 19 : Histogramme ajusté des résidus du modèle simple de la régression linéaire.....	61	
Tableau 20 : Coefficients d'asymétrie et d'aplatissement sur les résidus de la régression linéaire	62	
Tableau 21: résultat du test de normalité des résidus	62	
Figure 22 : Résidus du modèle simple de régression linéaire pondérée.....	63	
Figure 23 : PP-plot de l'ajustement à la loi Log-Normale	Figure 24 : PP-plot de l'ajustement à la loi Gamma	75
Figure 25 : Représentation des résidus de la déviance en fonction des valeurs prédites.....	79	
Figure 26 : L'Ajustement à la loi de Poisson	81	
Figure 27 : Effectifs globaux projetés de la population AMO-CNOPS	95	
Figure 28 : Evolution de l'effectif de la population atteinte d'au moins une ALD entre 2014 et 2019	95	
Figure 29 : L'estimation des montants remboursés pour une période de 5 ans.....	100	
Figure 30 : L'évolution du montant global remboursé par la CNOPS entre 2014 et 2019	101	

Introduction

Le Maroc, comme la plupart des pays en voie de développement, est confronté à une forte consommation de soins médicaux. Malgré les efforts des instances dirigeantes, plusieurs facteurs, dont le plus significatif est l'allongement de la durée de vie, font penser que l'augmentation va se poursuivre dans l'avenir. Le problème de financement du système de santé est donc plus que jamais un sujet important et au cœur des débats publics.

Le chantier de l'assurance maladie obligatoire au Maroc a certes réalisé des avancées notables. Cependant, il a toujours du mal à assurer un accès financier aux soins appropriés et équitables. Le premier grand défi en matière de santé étant sans aucun doute lié à l'accès aux soins, dans un contexte marqué par des besoins de plus en plus croissants d'une population de plus en plus exigeante.

Pour prévenir les situations de déséquilibre, la loi a institué auprès du Premier Ministre, l'Agence Nationale de l'Assurance Maladie, dotée de la personnalité morale et de l'autonomie financière qui a pour objectif d'évaluer d'une manière continue, les dépenses et les données de consommation médicale des différents régimes d'assurance maladie obligatoire.

La diversité des acteurs présents dans le domaine des remboursements des frais de soins de santé, à savoir, les mutuelles et les institutions de prévoyance, les sociétés d'assurance, contribue à créer une situation concurrentielle tendue. Par conséquent, ces intervenants se doivent d'être en mesure de proposer en permanence un tarif adapté au marché, mais tout d'abord ils doivent évaluer le coût du risque le plus finement possible, c'est dans ce contexte que se situe notre étude.

Le présent mémoire a pour objectif d'évaluer l'engagement financier probable de la Caisse Nationale des Organismes de Prévoyance Sociale en termes de dépenses de santé. Ces dépenses ne sont autres que les charges des sinistres à payer. Notre étude se focalisera sur la population des salariés du secteur public, dont la couverture médicale obligatoire est assurée par ladite caisse, et se répartira en cinq parties.

Nous présenterons, dans une première partie, le contexte général actuel de l'assurance maladie obligatoire au Maroc. Nous commencerons en premier lieu par une description brève de l'Agence Nationale d'Assurance Maladie « ANAM », ensuite nous parlerons de l'Assurance Maladie Obligatoire « l'AMO », de son rôle, des différents régimes en faisant partie intégrante et des différents types de financements liés à chaque régime.

La seconde partie de l'étude sera consacrée à l'analyse de la population assurée et de son comportement face aux soins de santé. Pour ce faire nous commencerons par de simples statistiques descriptives afin de caractériser la répartition des bénéficiaires du portefeuille et leur consommation selon les variables discriminantes. Ensuite nous effectuerons une analyse multidimensionnelle de ces variables, ce qui nous permettra

entre autres d'identifier, parmi les données disponibles, les variables qui influent le plus sur la consommation en soins médicaux desdits bénéficiaires et de quantifier cette influence.

La troisième partie traitera la modélisation de la consommation médicale par segment de bénéficiaires, qui sera exprimée à partir de la fréquence moyenne des sinistres et du coût moyen par sinistre, en choisissant un modèle du type « fréquence * coût » basé sur des modèles linéaires généralisés.

La quatrième partie sera consacrée à la modélisation de la population assurée par la Caisse Nationale des Organismes de Prévoyance Sociale. Le modèle choisi est un modèle empirique basé sur le bilan des entrées et sorties que connaît la population étudiée.

La partie finale portera sur la mise en œuvre de l'approche « fréquence * coût * population » afin d'obtenir les dépenses de santé probables que la CNOPS est censée rembourser aux salariés du secteur public immatriculés à l'AMO, ainsi que sur l'analyse des résultats obtenus.

Chapitre 1 : Contexte de l'étude

Le but de ce chapitre est de présenter le secteur d'Assurance Maladie Obligatoire (AMO) au Maroc. Tout d'abord on commence par introduire l'Agence Nationale de l'Assurance Maladie qui est l'un des principaux acteurs de ce secteur et qui représente notre organisme d'accueil.

I. Présentation de l'organisme d'accueil

L'Agence Nationale de l'Assurance Maladie (ANAM) a été créée le 26 mai 2005 en tant qu'établissement public à caractère administratif, doté de la personnalité morale et de l'autonomie financière, et placé sous la tutelle de l'État. La création de l'ANAM vient en vertu de l'article 54 de la loi 65-00, dont un extrait est le suivant :

« Les organismes gestionnaires sont soumis au contrôle technique de l'Etat, qui a pour objet de veiller au respect par ces organismes des dispositions de la présente loi et des textes pris pour son application. »

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. Bulletin officiel n° 5058 du 16 ramadan 1423 (21 novembre 2002).

L'agence est administrée par un conseil présidé par le Premier ministre ou l'autorité gouvernementale déléguée par lui à cet effet.

Il comprend en outre :

- des représentants de l'administration ;
- des représentants des employeurs ;
- des représentants des assurés des secteurs publics et privés désignés par les centrales syndicales les plus représentatives ;
- des représentants des organismes gestionnaires de l'assurance maladie obligatoire de base.

Elle est l'une des grandes réalisations qu'a connues notre pays durant cette dernière décennie pour la concrétisation de la couverture médicale de base (CMB). Tel que prévu par la législation en vigueur, l'ANAM a pour missions principales l'encadrement et la régulation du régime de l'Assurance Maladie Obligatoire (AMO), ainsi que la gestion des ressources du Régime d'Assistance Médicale (RAMED). Elle doit veiller au respect des dispositions de la loi régissant la couverture médicale de base. C'est donc l'institution qui est responsable de la déclinaison effective des droits fondamentaux du citoyen marocain en matière de couverture médicale de base, cités dans l'article 31 de la Constitution de 2011.

Les missions de l'ANAM :

En vertu de l'article 59 de la loi n°65-00 portant code de la couverture médicale de base, l'Agence Nationale de l'Assurance Maladie (ANAM) est chargée de :

- S'assurer de concert avec l'administration, de l'adéquation entre le fonctionnement de l'assurance maladie obligatoire de base et les objectifs de l'Etat en matière de santé;
- Conduire, dans les conditions fixées par voie réglementaire, les négociations relatives à l'établissement des conventions nationales entre les Organismes Gestionnaires d'une part, les prestataires de soins et les fournisseurs de biens et de services médicaux d'autre part ;
- Proposer à l'administration les mesures nécessaires à la régulation du système d'assurance maladie obligatoire de base et en particulier, les mécanismes appropriés de maîtrise des coûts de l'assurance maladie obligatoire de base et veiller à leur respect ;
- Emettre son avis sur les projets de textes législatifs et réglementaires relatifs à l'assurance maladie obligatoire de base dont elle est saisie par l'administration, ainsi que sur toutes autres questions relatives au même objet ;
- Veiller à l'équilibre global entre les ressources et les dépenses pour chaque régime d'assurance maladie obligatoire de base ;
- Apporter l'appui technique aux Organismes Gestionnaires pour la mise en place d'un dispositif permanent d'évaluation des soins dispensés aux bénéficiaires de l'assurance maladie obligatoire de base dans les conditions et selon les formes édictées par l'administration ;
- Assurer l'arbitrage en cas de litiges entre les différents intervenants dans l'assurance maladie ;
- Assurer la normalisation des outils de gestion et documents relatifs à l'assurance maladie obligatoire de base ;
- Tenir les informations statistiques consolidées de l'assurance maladie obligatoire de base sur la base des rapports annuels qui lui sont adressés par chacun des Organismes Gestionnaires ;
- Elaborer et diffuser annuellement un rapport global relatant les ressources, les dépenses et les données relatives à la consommation médicale des différents régimes d'assurance maladie obligatoire de base.
- En vertu de l'article 60 de la même loi n°65-00, l'ANAM s'est vue, en outre, confier la mission de gestionnaire du régime d'assistance médicale (RAMED).

L'organigramme de l'ANAM est cité en annexe.

II. L'Assurance Maladie Obligatoire au Maroc

Il semble important de rappeler dans un premier temps l'ensemble des notions et des mécanismes qui régissent l'assurance maladie obligatoire (AMO) et plus précisément l'assurance santé au Maroc.

1. Présentation de l'AMO

L'Assurance Maladie Obligatoire de base est un système d'assurance sociale instauré par la loi n°65-00 portant code de la couverture médicale de base promulguée par le dahir n° 1-02-296 du 3 octobre 2002 – 25 rejev 1423 pour la couverture des risques et frais de soins de santé inhérents à la maladie ou l'accident, à la maternité et à la réhabilitation physique et fonctionnelle. Elle est fondée sur le principe contributif et la mutualisation des risques.

Le Maroc a connu l'entrée en vigueur de l'AMO jusqu'à Septembre 2005, après la publication des décrets de mise en application de la loi 65-00 (Août 2005).

Elle est plus qu'un simple service d'assurance, c'est tout un engagement étatique qui garantit aux citoyens l'un des droits les plus élémentaires : l'accès aux soins de santé.

Elle représente une convention nationale entre les organismes gestionnaires (CNOPS, CNSS) et les établissements publics de soins et d'hospitalisation parmi eux le centre hospitalier IBN SINA de Rabat, le Centre Hospitalier IBN ROCHD de Casablanca, le Centre Hospitalier HASSAN 2 de Fès, le Centre Hospitalier MOHAMMED 6 de Marrakech et l'Institut Pasteur du Maroc.

2. Cadre général de l'AMO

L'AMO a été institué pour offrir une égalité et une équité dans l'accès aux soins à toute la population avec le principe de prise en charge collective et solidaire des dépenses de la santé.

La garantie de l'AMO est illimitée et sans plafond, elle couvre :

- Des soins pour adultes (51 groupes de pathologies déclinées en 172 maladies)
- Tous les soins pour les enfants couverts y compris les soins préventifs (vaccins)
- Les maladies antérieures
- Les Retraités et Handicapés bénéficiaires à vie.

La garantie de l'AMO est maintenue :

- Pendant 6 mois en cas de cessation d'activité de l'assuré;

- Pendant 12 mois en cas de dissolution du lien de mariage du conjoint ;
- Pendant 24 mois pour le conjoint survivant et les enfants en cas de décès de l'assuré.

3. Principes de l'AMO

- Égalité et équité dans l'accès aux soins à toute la population;
- Prise en charge collective et solidaire des dépenses de la santé;
- Solidarité nationale au profit de la population démunie;
- Progressivité de mise en place;
- Implication des acteurs économiques et sociaux et des professionnels de santé;
- Maintien des acquis;
- Régulation du système.

4. La population couverte

L'Assurance Maladie Obligatoire de base s'applique dans un premier temps aux fonctionnaires et agents de l'Etat, des collectivités locales, des établissements publics, à toutes personnes morales de droit public, aux personnes assujetties au régime de sécurité sociale en vigueur dans le secteur privé, aux titulaires de pension des deux secteurs public et privé, aux travailleurs indépendants, aux personnes exerçant une profession libérale et à toutes autres personnes exerçant une activité non salariée et aux anciens résistants et membres de l'armée de libération et aux étudiants de l'enseignement supérieur public et privé.

Elle couvre outre le salarié assujetti les membres de sa famille qui sont à sa charge, à condition qu'ils ne soient pas bénéficiaires à titre personnel d'une assurance de même nature. Il s'agit :

- Le (s) conjoint (s) de l'assuré,
- Les enfants à charge jusqu'à 21 ans ou en cas de poursuite des études jusqu'à 26,
- Les enfants pris en charge conformément à la législation en vigueur,
- Les enfants handicapés sans limite d'âge lorsque leur handicap physique ou mental les rend dans l'impossibilité de se livrer à une activité.

5. Ouverture des droits

Un bénéficiaire de la couverture AMO sinistré n'a le droit au remboursement (ouverture de droit) que si l'assuré qui le prend en charge a passé une période de

cotisation, dite période de stage, de 54 jours ouvrables successifs ou non pendant les 6 mois précédents sa déclaration.

En cas d'interruption du travail, l'assuré ou les ayants droit (conjoint, enfant à charge de moins de 21 ans ou 26 ans si étudiant et sans limite d'âge si handicapé) bénéficient du maintien de leur droit aux prestations pendant une période maximum de six mois.

En cas de dissolution du mariage, l'ex-conjoint d'un assuré qui ne bénéficie pas d'un régime d'assurance maladie obligatoire de base, continue à bénéficier des prestations de l'AMO pendant un an.

Les ayants droit de l'assuré décédé qui n'ont aucun régime d'assurance maladie obligatoire de base continuent de bénéficier des prestations de l'AMO pendant une période de deux années.

6. Affection de longue durée (ALD)

Les affections de longue durée (ALD) sont définies comme des maladies chroniques, dont la thérapie est coûteuse et pour laquelle l'Assurance Maladie Obligatoire assure une prise en charge pour tous les traitements nécessaires. En vertu de l'article 9 de la loi 65-00, les frais des soins relatifs à ces maladies sont remboursés par la CNSS à la hauteur de 90% si ces soins sont conduits dans un hôpital public et 70% sinon.

La liste des ALD donnant droit à l'exonération a été fixée par arrêté du ministre de la santé n°2518-05. Elle répertorie plus de 140 maladies dont les critères de choix sont :

- la fréquence de la maladie (prévalence ou incidence)
- la gravité de la maladie surtout en termes d'incapacité et d'invalidité
- la chronicité de la maladie
- la charge de morbidité
- le coût de la prise en charge

7. Le taux de cotisation de l'AMO

❖ Pour la caisse nationale des organismes de prévoyance sociale :

Selon le décret n° 2-05-735 du 11 jourmada II 1426 (18 juillet 2005), le taux de cotisation due à la caisse nationale des organismes de prévoyance sociale au titre du régime de l'assurance maladie obligatoire de base est fixé comme suit :

✓ Pour les salariés :

5% de l'ensemble des rémunérations visées à l'article 106 de la loi n°65-00 réparties à parts égales entre l'employeur et le salarié.

L'article premier du décret n°2-05-735 précise que « Chacune des parts de la cotisation est perçue dans la limite d'un montant mensuel minimum de 70 dirhams et d'un plafond mensuel de 400 dirhams ».

✓ Pour les pensionnés :

2,5% du montant global des pensions de base dans la limite d'un montant mensuel minimum de 70 dirhams et d'un plafond mensuel de 400 dirhams.

❖ Pour la caisse nationale de sécurité sociale :

Selon le décret n° 2-11-464 du 7 chaoual 1432 (6 septembre 2011) modifiant et complétant le décret n° 2-05-734 du 11 joumada II 1426 (18 juillet 2005) le taux de cotisation due à la caisse nationale de sécurité sociale au titre du régime de l'assurance maladie obligatoire de base, est fixé comme suit :

✓ Salariés :

4% de l'ensemble des rémunérations visées à l'article 19 du dahir relatif au régime de la sécurité sociale dont 50% à la charge de l'employeur et 50% à la charge du salarié.

1,5% de la rémunération brute mensuelle à la charge exclusive de tous les employeurs affiliés à la CNSS.

✓ Pensionnés :

Le taux de la cotisation due par les titulaires de pensions est fixé à 4 % du montant global des pensions de base servies.

✓ Marins pêcheurs à la part :

Le taux de cotisation des marins pêcheurs à la part est un pourcentage du produit brut de la vente de poisson (1,2% du montant du produit brut de la vente du poisson pêché par les chalutiers ; 1,5% du montant du produit brut de la vente du poisson pêché par les sardiniers et les palangriers.

8. Panier de soins et taux de couverture de l'AMO

❖ Pour la caisse nationale des organismes de prévoyance sociale :

Selon le décret n° 2-05-736 du 11 joumada II 1426 (18 juillet 2005) fixant les taux de couverture des prestations médicales à la charge de la caisse nationale des organismes de prévoyance sociale au titre du régime de l'assurance maladie obligatoire de base, les prestations couvertes ainsi que les taux de remboursement y afférents se présentent comme suit :

Prestations couvertes		Taux de remboursement
Groupe I	Actes de médecine générale et de spécialités médicales et chirurgicales, actes paramédicaux, de rééducation fonctionnelle et de kinésithérapie délivrés à titre ambulatoire hors médicaments.	80% de la tarification nationale de référence.
Groupe II	Soins liés à l'hospitalisation et aux interventions chirurgicales y compris les actes de chirurgie réparatrice et le sang et ses dérivés labiles.	90% de la tarification nationale de référence. Ce taux est porté à 100% lorsque les prestations sont rendues dans les hôpitaux publics, les établissements publics de santé et les services sanitaires relevant de l'Etat.
Groupe III	Médicaments admis au remboursement.	70% du prix public Maroc.
Groupe IV	Lunetterie médicale, dispositifs médicaux et implants nécessaires aux actes médicaux et chirurgicaux.	Forfaits fixés dans la tarification nationale de référence.
Groupe V	Appareils de prothèse et d'orthèse médicales admis au remboursement.	Forfaits fixés dans la tarification nationale de référence.
Groupe VI	Soins bucco-dentaires.	80% de la tarification nationale de référence.
Groupe VII	Orthodontie médicalement requise pour les enfants.	Forfait fixé dans la tarification nationale de référence.

Tableau 1 : Prestations couvertes et niveaux de couverture pour la CNOPS.

En cas de maladie grave ou invalidante nécessitant des soins de longue durée ou particulièrement coûteux, l'assuré est exonéré totalement ou partiellement de sa charge selon le type de maladies telles que prévues dans la liste arrêtée par le ministère de la santé.

La part restant à la charge de l'assuré ne peut être supérieure à 10% de la tarification nationale de référence pour ces maladies.

❖ Pour la caisse nationale de sécurité sociale :

Selon le décret n° 2-05-737 du 11 jourmada ii 1426 (18 juillet 2005) fixant les taux de couverture des prestations médicales à la charge de la Caisse Nationale de Sécurité Sociale au titre du régime de l'assurance maladie obligatoire de base et le décret n° 2-09-299 du 23 hijra1430 (11 décembre 2009) complétant le décret n° 2-05-737 du 11 jourmada ii 1426 (18 juillet 2005) fixant les taux de couverture des prestations médicales à la charge de la Caisse Nationale de Sécurité Sociale au titre du régime de l'assurance maladie obligatoire de base, les prestations couvertes ainsi que les taux de remboursement y afférents se présentent comme suit :

	Prestations couvertes	Taux de remboursement
Groupe I	Actes de médecine générale et de spécialités médicales et chirurgicales, actes paramédicaux, de rééducation fonctionnelle et de kinésithérapie délivrés à titre ambulatoire hors médicaments.	70% de la tarification nationale de référence.
Groupe II	Soins liés à l'hospitalisation et aux interventions chirurgicales y compris les actes de chirurgie réparatrice et le sang et ses dérivés labiles.	70% de la tarification nationale de référence. Ce taux est porté à 90 % lorsque les prestations sont rendues dans les hôpitaux publics, les établissements publics de santé et les services sanitaires relevant de l'Etat.
Groupe III	Médicaments admis au remboursement.	70% du prix public Maroc.
Groupe IV	Lunetterie médicale, dispositifs médicaux et implants nécessaires aux actes médicaux et chirurgicaux.	70% de la tarification nationale de référence.
Groupe V	Appareils de prothèse et d'orthèse médicales admis au remboursement.	70% de la tarification nationale de référence.
Groupe VI	Soins bucco-dentaires.	70% de la tarification nationale de référence mais uniquement pour les enfants dont l'âge est inférieur à 12 ans
Groupe VII	Orthodontie médicalement requise pour les enfants.	70% de la tarification nationale de référence mais uniquement pour les enfants dont l'âge est inférieur à 12 ans

Tableau 2 : Prestations couvertes et niveaux de couverture pour la CNSS

Conformément aux dispositions de l'article 9 de la loi n°65-00, la part restant à la charge de l'assuré, fait l'objet de l'exonération partielle ou totale en cas de maladie grave ou invalidante nécessitant des soins de longue durée, ou en cas de soins particulièrement coûteux. A cet effet, le conseil d'administration de l'agence nationale de l'assurance maladie a fixé à l'issue de sa 5ème session (résolution n°38), les modalités de cette exonération et a décidé que le taux de remboursement de chaque affection soit déterminé de façon à ce que la part restante à la charge de l'assuré ne doit pas dépasser 3000Dhs par an.

En guise de conclusion, ce chapitre nous a permis de comprendre les enjeux de l'Assurance Maladie Obligatoire au Maroc et de saisir ses soubassements.

Chapitre 2 : Etude descriptive et analyse des données

La modélisation impose de bien connaître la population à couvrir, et notamment de bien cibler la population potentiellement consommatrice.

La phase d'analyse des données est une phase déterminante, ainsi une part très importante du temps consacrée à l'étude a été réservée à la compréhension et l'analyse des bases fournies par l'organisme d'accueil.

I. Présentation des données utilisées

Les données qui nous ont été assignées sont relatives à la Caisse Nationale des Organismes de Prévoyance Sociale « CNOPS » et se composent de 2 tables ACCESS, une relative à la population sous risque qui comprend tous les individus couverts par l'organisme gestionnaire, il s'agit des assurés et leurs ayants droits, tous susceptibles d'être sinistrés et l'autre relative à la population consommatrice qui comporte tous les individus effectivement sinistrés.

Dans le cadre de notre étude, nous allons nous focaliser sur la population couverte de l'année 2014 et des remboursements des frais de soins effectués au cours de cette année.

II. Structure des données

1. La table de la population assurée « fichier effectif »

La base des données de la population de l'AMO-CNOPS mise à notre disposition comporte les variables décrites dans le tableau suivant :

variables	Etiquettes	Modalités
type_ass	le type de l'assuré	A: Actif P1: Pensionné (Invalidité+vieillesse) P2: Conjoint survivant P3: Enfant survivant
type_benef	le type du bénéficiaire	A: Assuré CA: Conjoint de l'assuré EA: Enfant de l'assuré
Sexe	le sexe du bénéficiaire	F: Femme M: Homme
tranche_age	Tranche d'âge du bénéficiaire du soin médicale	1 : [0,1[, 2 : [1,5[, 3 : [5,10[, . . . , 15 : [65,70[, 16 : 70+ , 17 : non renseigné
ald	Affections Longue Durée	N: Non-ALD O: ALD
ouvert_droit	L'ouverture de droit aux prestations de l'AMO est subordonnée au respect d'une période de cotisation, dite période de stage	F: Fermé O: Ouvert
region	region du bénéficiaire	17 régions
effectif	L'effectif des bénéficiaires selon la classe de population	-

Tableau 3 : Structure de la base de données de la population de l'AMO- CNOPS de 2014

La table de la population assurée comporte une information qui concerne les individus susceptibles d'être sinistrés.

Les données mises à notre disposition sont des données groupées (sous forme de cube).

Chaque ligne de cette table fait référence à un ensemble de personnes ayant les mêmes modalités de nos variables descriptives. Dans ce qui suit, chaque ligne sera dite « classe de population ».

2. La table de la population consommatrice « fichier prestation »

La structure des données disponible sur la consommation en assurance maladie obligatoire est comme suit :

variables	Etiquettes	Modalités
exercice_surv	L'exercice de survenance du sinistre	de 2007 à 2014
code_actes	Un code qui concerne les actes médicaux	-
type_doss	Le type de dossier	T : Tiers payant D : Deboursement direct
secteur_etab	Le secteur d'établissement où le soin a été effectué	PR : secteur privé PU : secteur public NI : non renseigné
type_soins	Le type de soin	A : Ambilatoire H : Hospitalisation
sexe	Le sexe du bénéficiaire	F : Femme M : Homme
tranche_age	La tranche d'âge du bénéficiaire du soin médicale	1 : [0,1[, 2 : [1,5[, 3 : [5,10[, . . . , 15 : [65,70[, 16 : 70 et plus, 17 : non renseigné
type_benef	le type du bénéficiaire	A : Assuré CA : Conjoint de l'assuré EA : Enfant de l'assuré
type_ass	type de l'assuré	A : Actif P1 : Pensionné (Invalidité+Vieillesse) P2 : Conjoint survivant P3 : Enfant survivant
ALD	Affection longue durée	N : Non O : Oui
code_ald	un code qui indique la nature de l'affection de longue durée	41 types, y ALD2 jusqu'à ALD6
région	Ville de résidence du bénéficiaire	17 région
Montant_remb	Montant global liquidé de la classe de population	-
frais_engagés	frais engagés par assuré	-
nb_actes	nombre des actes	-
effectif_sinistré	L'effectif sinistré selon la classe de consommation	-

Tableau 4 : Structure de la base de données de la consommation de l'AMO - CNOPS de l'année 2014.

Ce fichier concerne les prestations réglées par l'organisme gestionnaire « CNOPS » pour la population présente dans le fichier des effectifs.

Elle est aussi organisée sous forme de cube, et le terme « classe de consommation » fera référence à une ligne de cette base de données.

III. La constitution des bases de données

Les données fournies ont nécessité en premier lieu un traitement de manière à ne conserver que les données nécessaires à notre étude mais aussi à fiabiliser ces données. Cette partie sera consacrée aux différents traitements qui y sont apportés.

Les variables inputs:

- Pour la variable exercice de survenance, nous allons filtrer notre base de données sur l'année 2014 comme année de référence pour ne conserver que les personnes dont la période d'affiliation est incluse dans cette année.
- Pour la variable code_actes qui concerne les actes médicaux assujettis à l'assurance maladie, nous allons exploiter la totalité des données, donc toutes les familles d'actes seront confondues. Pour cela, on va filtrer sur la modalité « TOUS ».
- la variable type de dossier indique si le dossier est de nature **déboursement direct** (c'est-à-dire, l'assuré règle lui-même ses frais médicaux et se fait ensuite rembourser par sa caisse d'assurance maladie et sa mutuelle), ou bien **tiers payant**, dans ce cas, les frais médicaux sont pris directement en charge par l'assurance maladie, le tiers payant pourra être intégral (aucun frais à payer) ou partiel (le patient reste alors redevable du ticket modérateur ou des diverses participations forfaitaires).
Dans ce cas aussi, on va filtrer sur la modalité « TOUS », pour ne pas distinguer entre les deux types de dossier.
- Pour les variables type de soins et secteur d'établissement, nous n'allons pas distinguer entre leurs modalités. Pour cela, on va filtrer sur la modalité « TOUS » relative à chacune des deux variables.
- la variable code_ald indique la nature de l'affection de longue durée, nous avons 41 types d'ald, cependant les modalités ALD2 jusqu'à ALD6 désignent les personnes qui sont atteintes de plus qu'une ALD (deux ALD jusqu'à six ALD). Nous n'allons pas distinguer entre ces types dans ce qui suit.

- La variable ouvert_droit du fichier effectif, désigne l'ouverture de droit aux prestations de l'AMO, dans le cadre de la population de CNOPS, nous n'allons pas différencier entre ceux qui ont un droit ouvert et ceux qui ont un droit fermé.

Les variables descriptives qui vont nous permettre de réaliser par la suite une jointure des deux tables décrites précédemment sont : le type de l'assuré, le type du bénéficiaire, le sexe, la tranche d'âge, la région et l'affection longue durée.

Les variables outputs :

La variable Montant_remb représente les dépenses du régime en question, cette dernière dépend de l'effectif sinistré, qui est lui-même un output donc, il faut tenir compte de ces deux variables pour ne pas altérer l'analyse. Pour cela nous allons calculer le montant remboursé moyen que nous allons chercher à modéliser par la suite.

Le montant remboursé moyen, noté MRM, est le montant remboursé global de chaque classe de population divisé par l'effectif sinistré de cette même classe:

$$\text{MRM} = \text{Montant_remb} / \text{Effectif sinistré}$$

IV. Analyse descriptive

Cette partie fait l'objet d'une description de la répartition de la population et de la consommation durant l'année 2014 selon les différentes caractéristiques du bénéficiaire de la couverture médicale obligatoire du régime en question. Il s'agit de modéliser des effets que l'on constate par de simples statistiques descriptives.

1. Etude de la population sous risque

Nous allons présenter la répartition de l'effectif de la population de la CNOPS selon les inputs d'intérêt :

1.1. Répartition de la population selon le type d'assuré

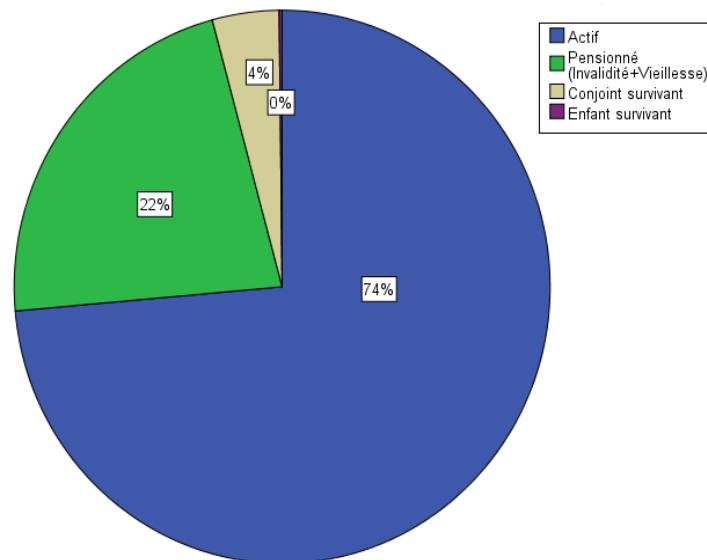


Figure 1 : Répartition de la population AMO-CNOPS selon le type d'assuré

On constate que la plus grande part de la population sous risque est représentée par les actifs (74%), la part des pensionnés pour motif de vieillesse ou d'invalidité est de 22%, cependant l'effectif des conjoints et enfants survivants reste minime par rapport à la population globale.

1.2. Répartition de la population selon le type du bénéficiaire

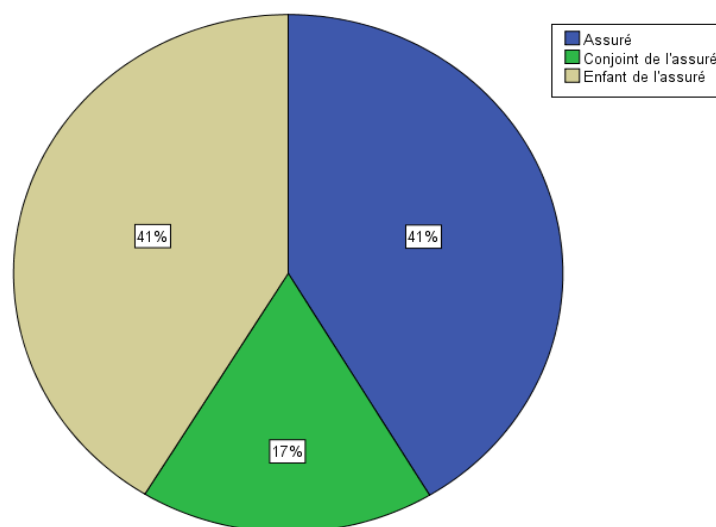


Figure 2 : Répartition de la population AMO-CNOPS selon le type du bénéficiaire

On remarque que la population des bénéficiaires est répartie à part égal entre les assurés et leurs enfants, tandis que la part des conjoints des assurés n'est que de 17% de l'effectif total.

1.3. Répartition de la population selon le sexe

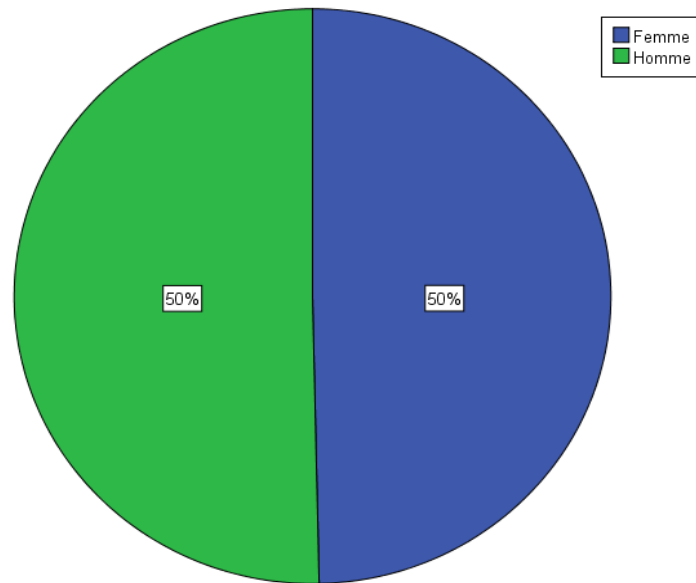


Figure 3 : Répartition de la population AMO-CNOPS selon le sexe

La population cible est uniformément répartie selon le sexe du bénéficiaire.

1.4. Répartition de la population selon la tranche d'âge

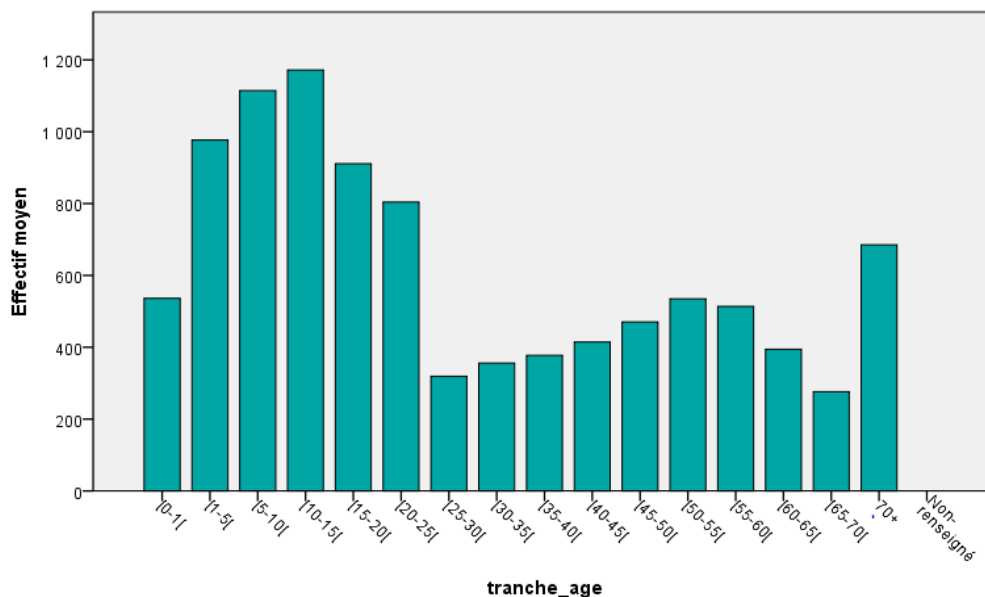


Figure 4 : Répartition de la population AMO-CNOPS selon l'âge

D'après le graphique ci-dessus, nous constatons que la population sous risque de l'année 2014 est majoritairement représentée par les enfants et les jeunes moins de 25 ans. À partir de cet âge, la proportion des bénéficiaires diminue pour croître par la suite chez les vieux ayant plus de 70 ans. On note que cette tranche [70+] représente

une exception, qui peut être expliquée par le fait que l’amplitude de cet intervalle reste imprécise.

Intéressons-nous maintenant à la répartition des bénéficiaires selon l’âge. La pyramide des âges suivante illustre cette répartition :

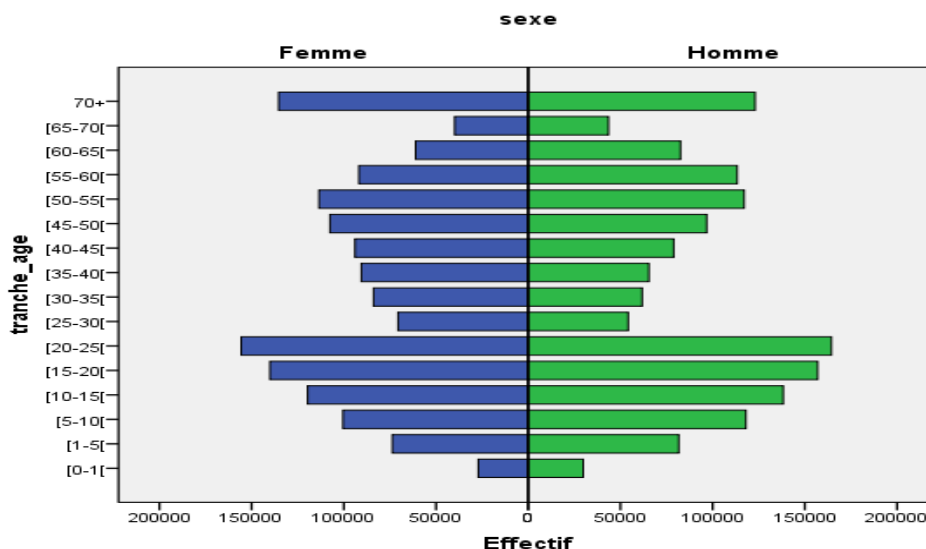


Figure 5 : Pyramide des âges de la population AMO-CNOPS

On constate dans un premier temps que même si cette répartition est plutôt inégale, les effectifs sont en nombre suffisant dans chaque tranche d’âge, ce qui ne pénalisera pas la suite de l’étude.

Cette pyramide représente la répartition de la population couverte selon la tranche d’âge et le sexe. Elle est marquée par la prépondérance de la population jeune âgée entre 20 et 25 ans, ainsi que la population âgée de plus de 70 ans. De plus, on constate que les effectifs hommes et femmes sont du même ordre de grandeur.

1.5. Répartition de la population par ALD

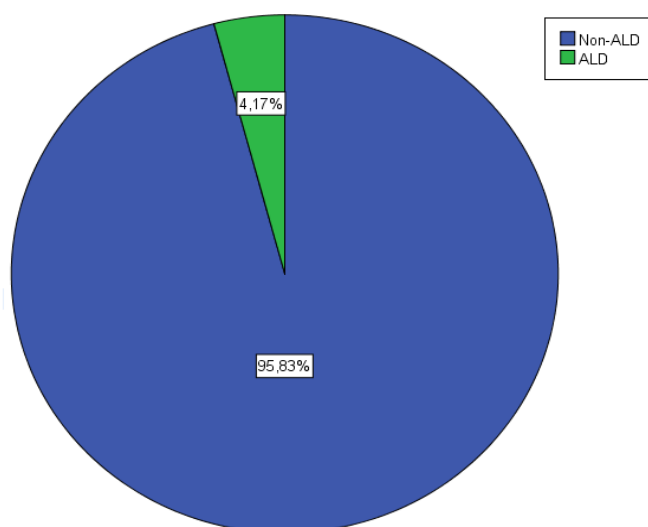


Figure 6 : Répartition de la population AMO-CNOPS selon l'atteinte d'une affection de longue durée

On remarque que la population atteinte d'ALD reste minime par rapport à la population non atteinte. En effet, elle ne représente que 4.17% de l'effectif total, cependant nous allons voir par la suite qu'elle affecte plus la consommation médicale.

2. Etude de la population consommatrice

Dans cette partie nous allons étudier la répartition des prestations versées par l'organisme assureur durant l'année 2014 en fonction des différentes variables d'intérêt mise à notre disposition.

2.1. Répartition des remboursements selon le type d'assuré

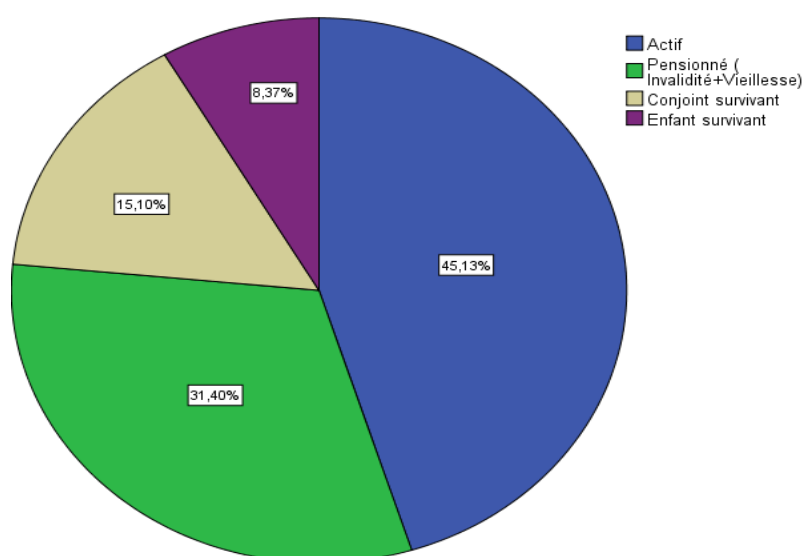


Figure 7 : Répartition des montants remboursés moyens selon le type d'assuré

La consommation médicale des actifs est majoritaire dans le portefeuille, elle constitue 45% du total des remboursements, de plus, les pensionnés s'accaparent de 31% des prestations sanitaires, tandis que la part des conjoints et des enfants survivants reste négligeable.

2.2. Répartition des remboursements selon le type du bénéficiaire

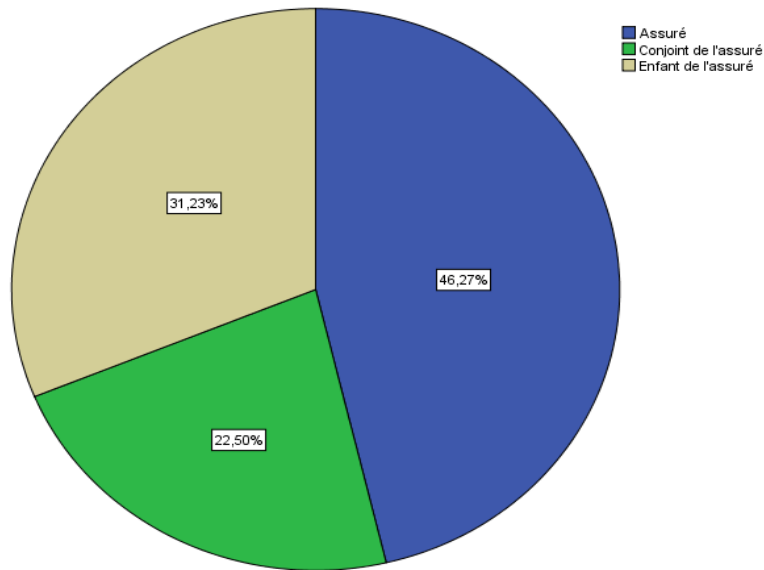


Figure 8 : Répartition des montants remboursés moyens selon le type du bénéficiaire

Pour chaque bénéficiaire, trois modalités de lien vis-à-vis de l'assuré sont possibles : soit il est lui-même l'assuré, soit il est le conjoint de l'assuré, soit il est l'enfant de l'assuré.

Les données ont révélé que les remboursements destinés aux assurés sont significativement plus importants que ceux des autres bénéficiaires. En outre, la part des conjoints des assurés est faible par rapport à la part des assurés, cela peut être dû au fait que les conjoints peuvent bénéficier d'une couverture médicale autre que celle de leurs époux (ses).

2.3. Répartition des remboursements selon le sexe

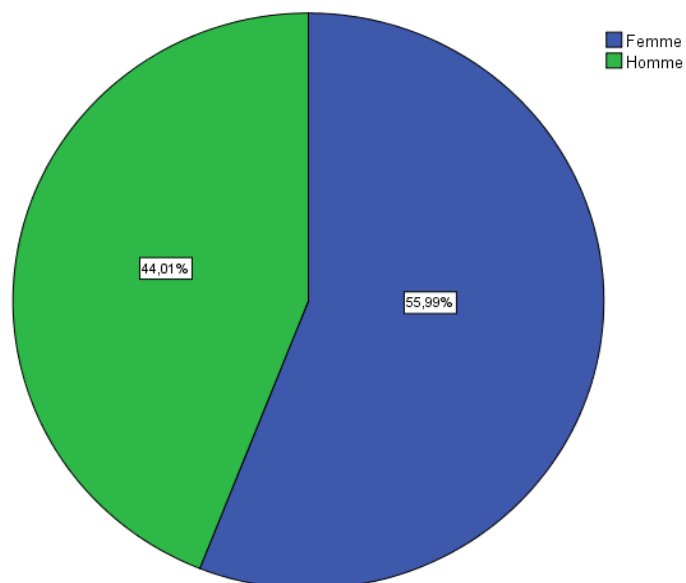


Figure 9 : Répartition des montants remboursés moyens selon le sexe

En calculant le ratio femme par homme on trouve 1.27%, ceci dit que les remboursements destinés aux femmes dépassent ceux destinés aux hommes de 27%.

2.4. Répartition des remboursements selon la tranche d'âge

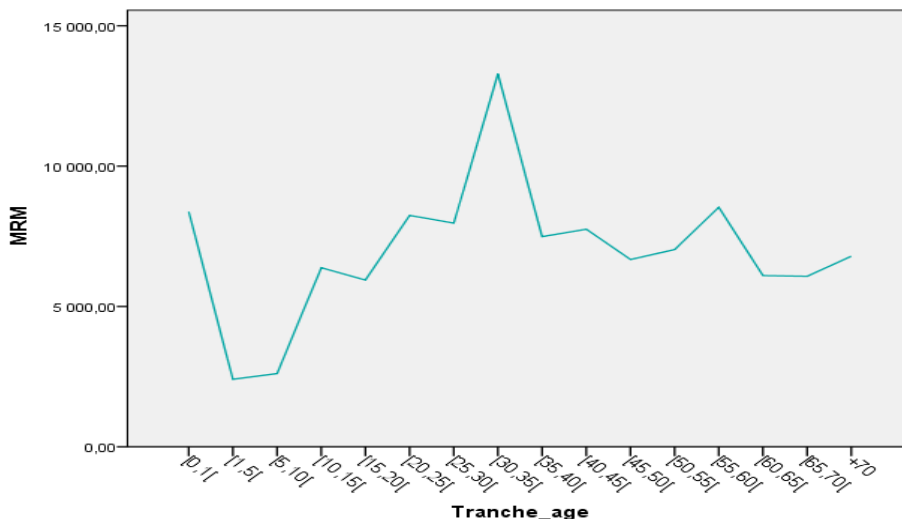


Figure 10 : Répartition des montants remboursés moyens selon l'âge

Au vu de cette courbe, qui trace les remboursements moyens selon l'âge du bénéficiaire, on remarque une consommation élevée aux très jeunes âges due aux maladies infantiles. À l'adolescence, les dépenses commencent à croître due à l'orthodontie. Ainsi un écart sur la consommation moyenne se creuse également entre 20 et 40 ans, provoqué par une fréquence accrue des maternités chez les femmes, l'écart se réduit ensuite à 40 ans et commence encore une fois à se creuser à partir de 50 ans due au vieillissement.

Ainsi on peut conclure que les remboursements varient significativement selon l'âge.

2.5. Répartition des remboursements par ALD

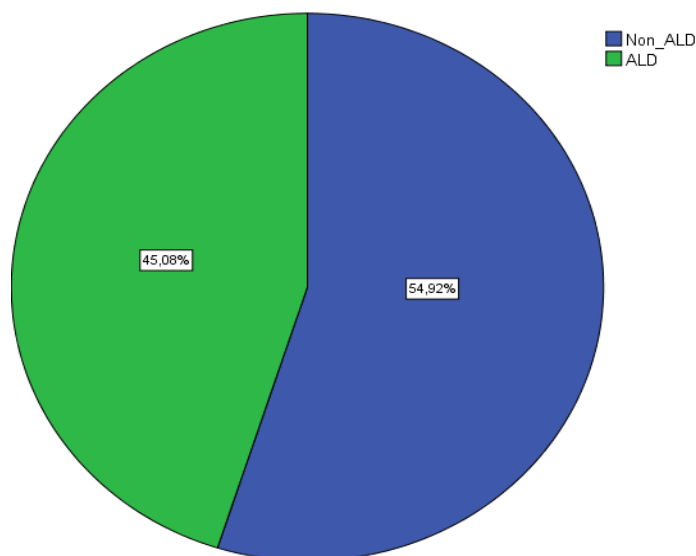


Figure 11 : Répartition des montants remboursés selon l'atteinte d'une affection de longue durée

Les remboursements destinés aux bénéficiaires atteints d'ALD représentent 45 % des remboursements globaux, tandis que ceux destinés aux bénéficiaires non atteints est de 54.9%.

D'après ce qui précède, la population des bénéficiaires atteints d'ALD n'est que de 4% de la population globale, mais elle s'accapare de plus de 45% des montants remboursés, ceci peut être expliquée par le fait que les soins médicaux relatifs aux maladies de longue durée sont plus onéreux.

3. Analyse de la fréquence des sinistres

Dans cette section, nous allons nous intéresser au taux de sinistralité afin d'avoir un premier aperçu des risques considérés. Pour cela, nous devons dans un premier temps réaliser une jointure à partir des deux fichiers « effectif » et « prestation » à l'aide d'ACCESS en liant les variables communes entre elles. Ensuite, nous calculons le taux de sinistralité, tel que :

$$\text{Taux de sinistralité} = \text{Effectif sinistré/Effectif}$$

La sinistralité à la CNOPS s'élève à 42.3%. Le taux de sinistralité a déjà atteint sa stationnarité. La fréquence de sinistralité quant à elle, est de l'ordre de 3.6, elle représente le nombre moyen de dossier par an et par bénéficiaire de soins.

3.1. Taux de sinistralité selon le sexe

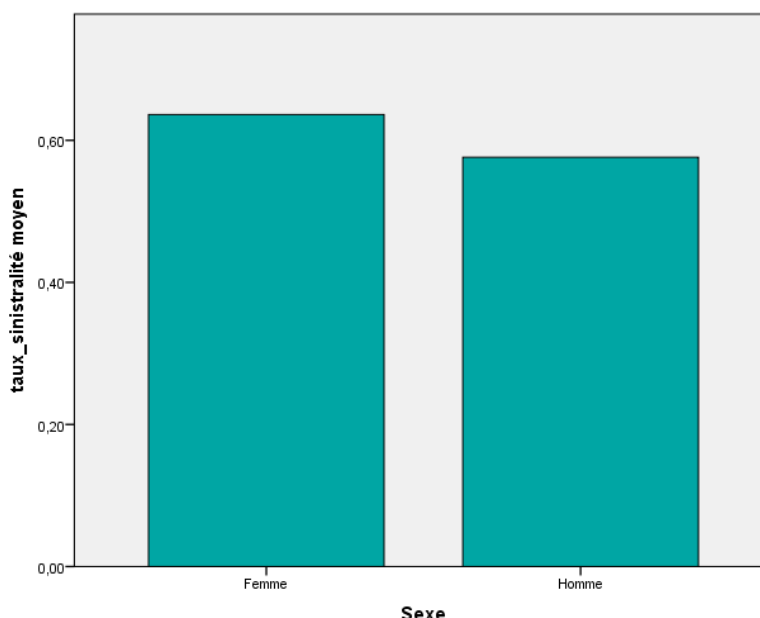


Figure 12 : Taux de sinistralité de la population AMO-CNOPS selon le sexe

On remarque que le taux de sinistralité des femmes est un peu plus important que celui des hommes.

3.2. Taux de sinistralité selon la tranche d'âge



Figure 13 : Taux de sinistralité de la population AMO-CNOPS selon l'âge

La répartition de la sinistralité selon l'âge du bénéficiaire, révèle qu'elle se concentre en premier lieu chez les bénéficiaires moins de 20 ans. À partir de 30 ans les bénéficiaires sont sujets à des sinistres de plus en plus fréquents et coûteux, c'est-à-dire qu'ils deviennent de plus en plus risqués avec le vieillissement. A partir de 70 ans, le taux de sinistralité connaît une certaine chute, cela est généralement dû au non déclaration des sinistres survenus.

3.3. Taux de sinistralité selon la variable ALD

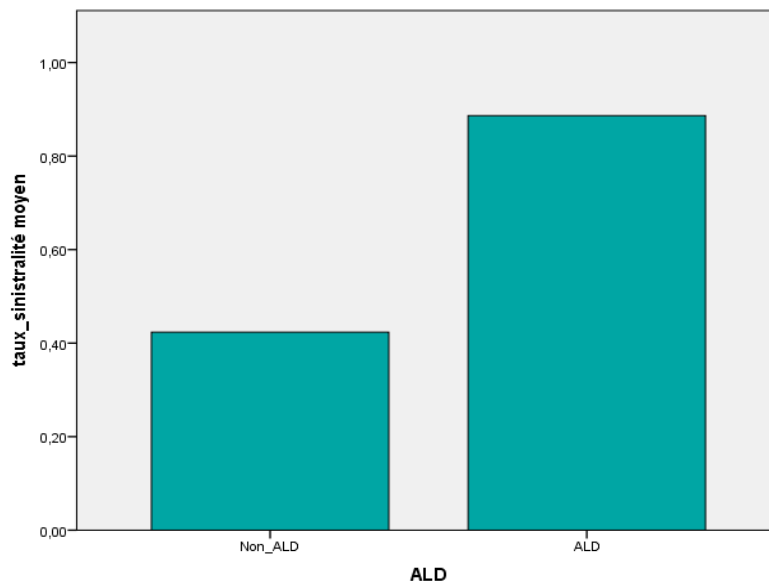


Figure 14 : Taux de sinistralité de la population AMO-CNOPS selon l'état de santé du bénéficiaire

On constate que le taux de sinistralité est plus important chez les personnes atteintes d'une affection de longue durée (90%) que chez celles non atteintes (40%).

4. Méthode de CHAID appliquée au taux de sinistralité

Afin de décrire la répartition de notre population des bénéficiaires en fonction de leur taux de sinistralité, nous faisons recours à la méthode de CHAID : « **Chi-squared Automatic Interaction Detector** ».

CHAID est une technique de type arbre de décision. Elle a été publiée en 1980 par Gordon V.Kass. Elle peut être utilisée pour la prédiction ou pour la détection d'interaction entre variables. La popularité de la méthode repose en grande partie sur sa simplicité. Il s'agit de trouver un partitionnement des individus que l'on représente sous la forme d'un arbre de décision.

L'objectif est de produire des groupes d'individus les plus homogènes possibles du point de vue de la variable à prédire. Il est d'usage de représenter la distribution empirique de l'attribut à prédire sur chaque sommet (nœud) de l'arbre.

Pour mieux appréhender la démarche, nous faisons recours au nœud de modélisation CHAID du logiciel SPSS MODELER qui représente un ensemble d'outils de data mining permettant de développer des modèles prédictifs et de les déployer dans des applications professionnelles afin de faciliter la prise de décision.

Analysons dans un premier temps le tableau d'importance des variables exogènes fournit par cet outil d'exploration de données :

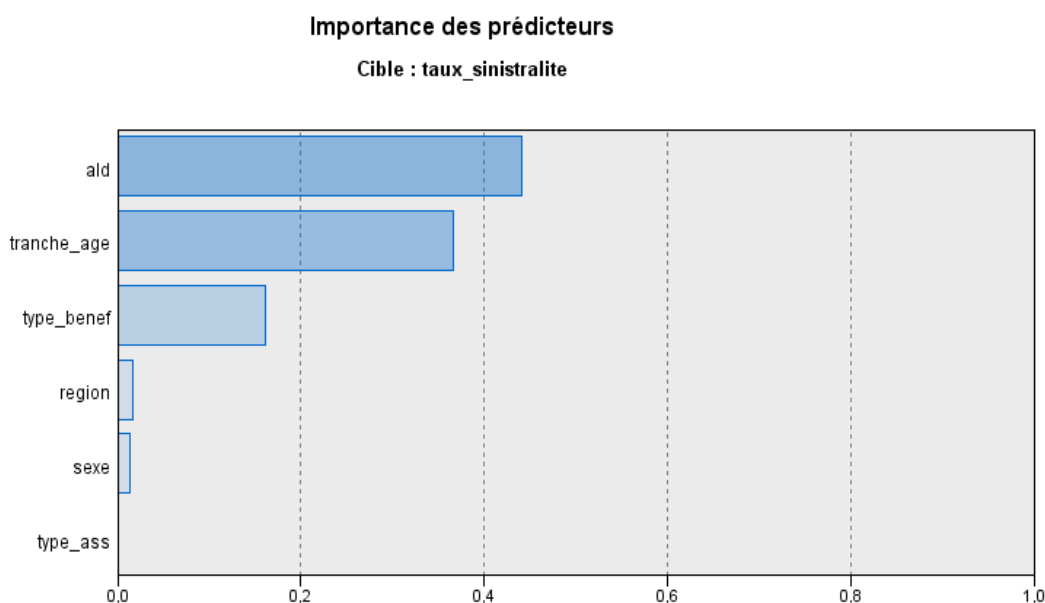


Figure 15 : Importance des variables indépendantes sur la variable cible taux de sinistralité

Ce graphique nous montre l'importance relative de chaque variable indépendante dans l'estimation du taux de sinistralité. Nous pouvons observer que la variable ald est le critère le plus important dans ce cas et que les seuls autres facteurs intéressants sont la variable tranche_age et la variable type_benef.

En voici notre modèle obtenu grâce à la technique CHAID sous forme d'un arbre :

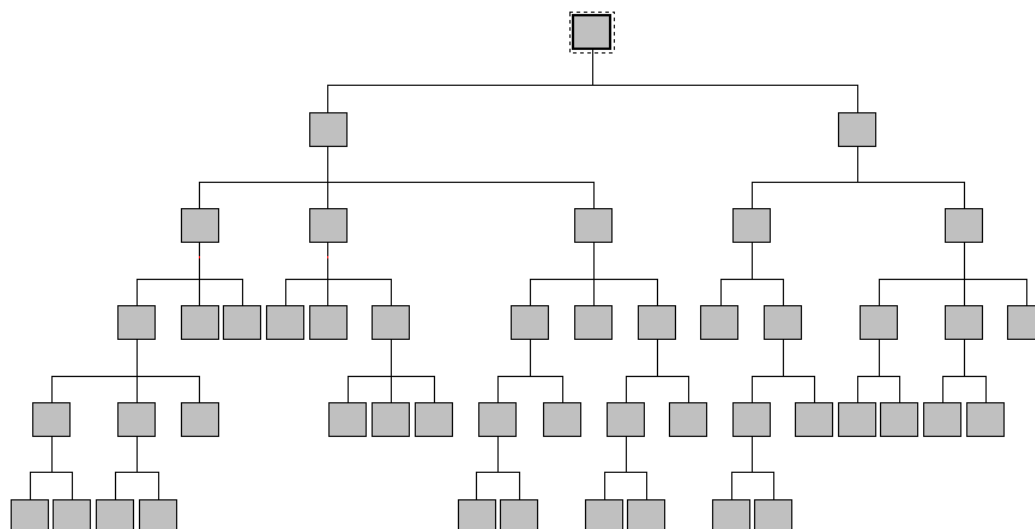


Figure 16 : Représentation de l'arbre de décision sous SPSS MODELER

Il est à noter que les critères permettant de passer d'un nœud à ses descendants n'ont pas été mentionnés sur cette figure car cela rendrait la lecture de l'arbre très difficile.

Si l'on regarde la partie supérieure de l'arbre élaborée grâce à SPSS MODELER, le premier sommet appelé la « racine » de l'arbre propose un récapitulatif de tous les enregistrements dans l'ensemble de données. En moyenne le taux de sinistralité des bénéficiaires prédit est de 61%. Il s'agit d'une proportion élevée. Voyons si l'arbre peut nous donner des informations sur les facteurs responsables.

La première division se situe au niveau de la variable « ald », on parle de variable de segmentation. Comme elle est composée de 2 modalités {N : non atteint d'une affection de longue durée, O : atteint d'une affection de longue durée}, elle produit donc 2 nœuds.

La première branche à gauche, sur le deuxième niveau, est produite à partir de la modalité « N » de la variable « ald ». Le nœud qui en résulte couvre 2693 observations, le taux de sinistralité moyen prédit correspondant à ce groupe est de 42,4%.

La seconde branche, à droite, correspond à la modalité « O » de la variable de segmentation « ald », le sommet correspondant couvre 1821 observations, leur taux de sinistralité moyen prédit est de 40.34%.

Ce processus est réitéré sur chaque nœud de l'arbre jusqu'à l'obtention de feuilles pures. Il s'agit bien d'un arbre de partitionnement, un segment représentant un

ensemble de bénéficiaires ayant les mêmes critères ne peut être situé dans deux feuilles différentes de l'arbre.

Le modèle de prédiction peut être lu très facilement. On peut traduire un arbre en une base de règles sans altération de l'information. Le chemin menant d'un sommet vers la racine de l'arbre peut être traduit en une partie prémisse d'une règle de prédiction de type attribut-valeur « SI variable 1 = valeur 1 ET variable 2 = valeur 2 ... ».

Par exemple le chemin menant à la première feuille pure située à gauche est défini comme suit : « si ald= N ET tranche_age=majo¹ ET type_benef=A ET type_ass=A », cette feuille correspond aux femmes assurées actives d'âge supérieure à 45 ans et non atteintes d'une affection de longue durée. Elles constituent (96/4514 = 2.127%) de l'ensemble des observations. En effet, cette feuille est composée de 96 observations, cependant le total des observations est de 4514. Pour rappel chaque observation correspond à un segment comme expliqué précédemment.

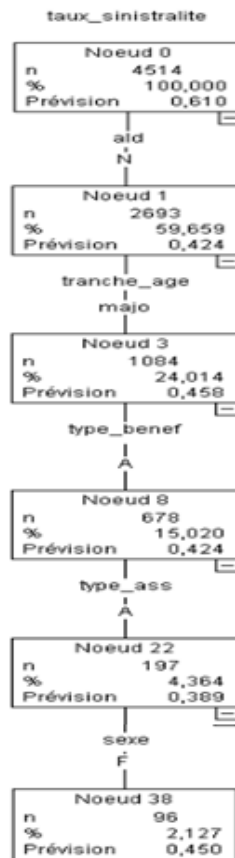


Figure 17 : Le chemin « si ald=N ET tranche_age=majo ET type_benef=A ET type_ass=A »

Pour classer un nouveau segment, il suffit de l'injecter dans l'arbre et de lui associer la conclusion attachée à la feuille dans laquelle il aboutit.

¹ La modalité « majo » de la variable tranche d'âge sera expliquée dans le chapitre 3

V. Mesures d'association

Les tableaux croisés à deux ou plusieurs variables qualitatives ne permettent pas de démontrer l'existence d'une association du point de vue statistique. Afin de mesurer véritablement la relation entre les variables, il est nécessaire de mettre en place des tests de signification statistique de l'association.

Dans cette section, nous allons utiliser le test de khi-deux et le V de Cramer pour mesurer et ordonner les différentes associations entre les variables qualitatives dont nous disposons.

1. Le test de chi-deux et le V de Cramer

Le test de khi 2 sert à apprécier l'existence ou non d'une relation entre deux caractères au sein d'une population, en d'autres termes, il permet de tester l'indépendance entre deux variables qualitatives. Le principe du test consiste à comparer les effectifs observés aux effectifs attendus à partir du tableau de contingence comme suit :

Soit une table de contingence bidimensionnelle à i lignes et j colonnes. On considère les notations suivantes :

- n_{ij} : L'effectif figurant dans la cellule (i, j)
- $n_{i.}$: L'effectif total de la ligne i
- $n_{.j}$: L'effectif total de la ligne j
- n : Effectif globale de la table de contingence
- $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$: L'effectif espéré de la cellule (i, j) en cas d'indépendance

La formule de la statistique de Chi-deux est la suivante :

$$\chi^2 = \sum_{(i,j)} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi_{(i-1)(j-1)}^2$$

Cette statistique suit une loi de chi-deux à $(i - 1)(j - 1)$ degrés de libertés.

Il s'agit d'un test non paramétrique des hypothèses :

$$\begin{cases} H_0 : \text{Les deux variables sont indépendantes} \\ H_1 : \text{Les deux variables sont dépendantes} \end{cases}$$

On peut alors lire sur la table du χ^2 la valeur attendue pour la valeur du degré de liberté obtenue, et au seuil de signification voulu. Si la statistique de χ^2 est supérieure à la valeur obtenue dans la table pour un seuil suffisamment élevé, nous pouvons conclure qu'il existe une liaison statistique significative entre les deux variables.

On s'intéresse particulièrement à la p-value calculée par les logiciels statistiques et qui fournit le niveau de signification pour lequel l'hypothèse nulle est rejetée, c'est-à-dire pour lequel nous rejetons l'indépendance entre les variables.

Ainsi, on rejette H_0 quand la statistique dépasse le quantile de la loi de chi-deux au seuil fixé (5% généralement).

Il existe plusieurs mesures d'association qui permettent d'ordonner les liaisons entre les différents variables.

Le test V de Cramer permet de comparer l'intensité du lien entre les deux variables étudiées.

Le V de Cramer est la racine carrée du χ^2 divisé par le χ^2_{max} . Ce χ^2_{max} théorique est égal à l'effectif multiplié par le plus petit côté du tableau (nombre de lignes ou de colonnes) moins 1.

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n * [\min(l, c) - 1]}}$$

Plus V est proche de zéro moins les variables étudiées sont dépendantes. Plus V est proche de 1 plus la liaison entre les deux variables étudiées est forte.

2. Analyse des résultats obtenus

Sous le logiciel SPSS, on effectue le test de Chi-deux sur toutes les combinaisons de variables catégorielles, on trouve alors le résultat suivant :

Variable 1	Variable 2	Khi –deux de Pearson		Décision
		Valeur	Signification	
Type_ass	Sexe	3,062	0,382	On accepte H_0
Type_ass	Région	139,682	0	On rejette H_0
Type_ass	ALD	90,406	0	On rejette H_0
Type_benef	Sexe	196,164	0	On rejette H_0
Type_benef	Région	45,892	0,053	On accepte H_0
Type_benef	ALD	0,531	0,767	On accepte H_0
ALD	Sexe	0,805	0,369	On accepte H_0

Tableau 5 : Mesure du test de Khi-Deux de Pearson

Avec H_0 : Les deux variables catégorielles sont indépendantes. On rejette H_0 quand la p-value du test est inférieur à 5%.

D'après le tableau ci-dessus, on note qu'il y a une indépendance entre le type d'assuré et son sexe. Ainsi entre le type du bénéficiaire et sa région, et le type du bénéficiaire et son état de santé.

Pour les variables corrélés entre elles deux à deux, le V de cramer permet de mesurer le degré d'association entre les variables. Plus V est proche de zéro, moins les variables étudiées sont dépendantes. Plus V est proche de 1 plus la liaison entre les deux variables étudiées est forte.

Le tableau suivant établie un classement croissant des liaisons entre les variables associées deux à deux :

Variable 1	Variable2	V de cramer
Type_ass	Région	0,102
Type_ass	ALD	0,142
Type_benef	Sexe	0,208

Tableau 6 : Mesure de V de cramer

Les V de cramer calculés sont tous proche de 0, on peut donc conclure que les variables sont peu associées entre elles.

VI. Analyse multidimensionnelle

Les méthodes d'Analyse de Données sont rangées en deux grandes familles : les méthodes d'analyse factorielle et les méthodes de classification. Ces deux familles de méthodes ont pour objet de résumer l'information contenue dans les données. Elles sont plus complémentaires que concurrentes, et peuvent avec profit être conjointement utilisées.

Les méthodes factorielles ont pour objet de résumer l'information apportée par un ensemble de variables, par un nombre plus restreint de variables nouvelles appelées "facteurs". Les méthodes de classification consistent à regrouper en classes ou catégories l'ensemble des individus jugées les plus homogènes possibles et cela au regard d'un critère.

Dans le cadre de notre travail, les variables à analyser sont toutes qualitatives, alors les méthodes d'analyses de données qui seront utilisée sont l'analyse factorielle des correspondances (AFC) et l'analyse des correspondances multiples (ACM).

1. L'analyse factorielle des correspondances (AFC)

L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été développée essentiellement par Jean Paul Benzécri durant la période 1970-1990.

L'AFC a pour but de décrire le maximum de l'information contenue dans un tableau de contingence, Elle a pour objectif d'étudier les similarités, les associations ou les interactions qui peuvent exister entre deux variables qualitatives.

Le terme correspondance provient du fait que l'on cherche à mettre ces deux variables en correspondance.

1.1. Principe général de l'AFC

L'objectif de l'AFC est la visualisation de données de fréquences issues du croisement de deux variables qualitatives observées sur une population.

En AFC, la comparaison de deux lignes (ou de deux colonnes) n'est pas facile car les effectifs sont inégaux. Cette comparaison peut se faire à l'aide des profils-lignes qu'on obtient en divisant chaque ligne par son effectif total et de la même façon on compare deux colonnes par les profils-colonnes qu'on obtient en divisant chaque colonne par son effectif total.

L'AFC est une double ACP faite sur les profils des lignes et des colonnes. De plus, la métrique utilisée pour faire cette comparaison est la distance du Khi deux.

1.2. Mise en œuvre de l'AFC

Dans le cadre de notre étude, Nous effectuons une AFC sur les variables explicatives qui nous intéressent, à savoir :

- Type de l'assuré et type du bénéficiaire
- Type du bénéficiaire et ALD
- Type du bénéficiaire et Sexe
- Sexe et ALD.

Comme l'AFC est une méthode exploratoire d'analyse des tableaux de contingence. On commence tout d'abord par croiser nos variables d'intérêt deux à deux comme suit:

type_benef	type_ass				
	Actif	Pensionné	Enfant survivant	Conjoint survivant	Marge active
Assuré	704	509	452	201	1866
Conjoint de l'assuré	542	392	16	26	976
Enfant de l'assuré	476	575	424	197	1672
Marge active	1722	1476	892	424	4514

Tableau 7 : Tableau des correspondances des variables type_ass et type_benef

type_benef	ald		
	Non-ald	ald	Marge active
Assuré	1110	756	1866
Conjoint de l'assuré	592	384	976
Enfant de l'assuré	991	681	1672
Marge active	2693	1821	4514

Tableau 8 : Tableau des correspondances des variables type_benef et ALD

type_benef	sexe		
	Femme	Homme	Marge active
Assuré	983	883	1866
Conjoint de l'assuré	745	231	976
Enfant de l'assuré	836	836	1672
Marge active	2564	1950	4514

Tableau 9 : Tableau des correspondances des variables type_benef et Sexe

ald	sexe		
	Femme	Homme	Marge active
Non-ald	1515	1178	2693
ald	1049	772	1821
Marge active	2564	1950	4514

Tableau 10 : Tableau des correspondances des variables Sexe et ALD

Ces tableaux croisés permettent de constater quelles sont les catégories les plus représentées, à savoir les actifs assurés, les assurés non atteints d'une affection de longue durée, les assurés de sexe féminin, et les femmes non atteintes d'ALD.

1.2.1. Le choix du nombre de dimensions :

L'inertie totale correspond à la somme des carrés des distances à l'origine de toutes les dimensions. Les valeurs propres (ou valeurs singulières) peuvent être interprétées comme étant des coefficients de corrélation entre les coordonnées des lignes et des colonnes.

Concernant le choix de la dimension, le nombre maximal de facteurs à conserver doit être égal à $\inf(p, q) - 1$, où p est le nombre de ligne et q est le nombre de colonne.

En effectuant une analyse des correspondances simple entre les variables type_ass et type_benef par exemple, on trouve le récapitulatif suivant qui indique la décomposition de l'inertie totale le long de chaque dimension.

Dimension	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie		Valeur singulière de confiance	
					Expliqué	Cumulé	Ecart-type	Corrélation
1	,295	,087			,931	,931	,009	
2	,080	,006			,069	1,000	,014	
Total		,093	421,839	,000 ^a	1,000	1,000		-.094

a. 6 degrés de liberté

Tableau 11 : Inertie par dimension dans le cas des variables type_ass et type_benef

Ce tableau récapitulatif affiche également un coefficient de corrélation entre les estimations des valeurs propres. Cette valeur nous donne une idée de la stabilité des résultats obtenus.

Dans cet exemple, on a $\inf(3,4)-1=2$, cependant on remarque que la première composante factoriel explique 93% de l'inertie totale du nuage de points (proportion de l'inertie totale : $0,087/0,093= 93\%$). ceci dit que le premier axe nous assure une très bonne représentation.

L'inertie totale(0,093) correspond à la somme des carrés des distances à l'origine dans toutes les dimensions.

Le carré de chaque valeur propre nous donne l'inertie de la dimension à laquelle elle appartient ($0,295^2=0,087$).

Par raisonnement analogue, il s'avère en analysant les tableaux récapitulatifs des couplets suivants (type_benef et ALD), (type_benef et Sexe) et (Sexe et ALD) que la première dimension explique 100 % de l'inertie totale.

Les tableaux récapitulatifs seront exposés en annexe.

1.2.2. Contributions

Dans cette partie, on cherche à savoir dans quelle mesure les lignes et les colonnes contribuent à l'inertie du nuage de points, en d'autres termes on cherche à vérifier la pertinence de ces axes en regardant dans quelles mesures les variables expliquent les axes sélectionnés.

À titre d'exemple, les contributions sont données dans le tableau suivant pour la variable type_benef :

1.2.2.1. Contributions : points lignes

Caractéristiques des points lignes^a

type_benef	Masse	Score dans la dimension		Inertie	Contribution				
		1	2		De point à inertie de dimension		De dimension à inertie de point		Total
					1	2	1	2	
Assuré	,413	,209	-,320	,009	,061	,525	,612	,388	1,000
Conjoint de l'assuré	,216	-1,026	,068	,067	,771	,013	,999	,001	1,000
Enfant de l'assuré	,370	,365	,317	,018	,167	,462	,830	,170	1,000
Total actif	1,000			,093	1,000	1,000			

a. Normalisation principale symétrique

Tableau 12 : Tableau des caractéristiques des profils lignes

Ce tableau affiche les contributions des modalités de la variable en ligne «Type_benef » à la construction des axes factoriels.

Les scores de la dimension permettent de placer les modalités sur les axes retenus. Il s'agit en quelques sortes, des coordonnées des modalités.

Le premier axe oppose les conjoints dont les coordonnées sont négatives, aux Assurés et leurs enfants dont les coordonnées sont positives.

On remarque également que la modalité conjoint de l'assuré contribue plus à la construction du premier axe, tandis que le deuxième axe est représenté par l'assuré et ses enfants.

Ainsi, les contributions « De point à inertie de dimension » permettent de savoir comment ces variables expliquent les axes (dimensions) afin de pouvoir dire si les positions des modalités les unes par rapport aux autres sur les axes sont significatives. Ces contributions peuvent être interprétées comme des corrélations.

Les contributions « De dimension à inertie de point » nous donnent la façon dont les modalités sont expliquées par les axes.

1.2.2.2. Contributions : points colonnes

Caractéristiques des points colonnes^a

type_ass	Masse	Score dans la dimension		Inertie	Contribution				
		1	2		De point à inertie de dimension		De dimension à inertie de point		Total
					1	2	1	2	
Actif	,381	-,462	-,268	,026	,276	,342	,916	,084	1,000
Pensionné	,327	-,197	,390	,008	,043	,619	,483	,517	1,000
Enfant survivant	,198	,886	-,126	,046	,526	,039	,994	,006	1,000
Conjoint survivant	,094	,698	-,002	,014	,155	,000	1,000	,000	1,000
Total actif	1,000			,093	1,000	1,000			

a. Normalisation principale symétrique

Tableau 13 : Tableau des caractéristiques des profils colonnes

Ce tableau affiche les contributions sur chaque dimension de la variable en colonne : « type_ass ».

Le premier axe oppose les conjoints dont les coordonnées sont négatives, aux Assurés et leurs enfants dont les coordonnées sont positives.

On constate de même que le premier axe factoriel est représenté par les Enfants survivants, tandis que le deuxième est représenté par les Pensionnés.

Les tableaux des contributions des autres couplets seront exposés en annexe.

2. Analyse factorielle des correspondances multiples « ACM »

L'ACM est une technique d'analyse de données qui s'applique à des données catégorielles, elle constitue une généralisation de l'AFC. C'est l'une des méthodes les plus utilisées en analyse des données, ses principaux domaines d'applications sont le traitement des questionnaires et l'exploitation des enquêtes par sondages.

L'objectif de la méthode consiste à décrire les liaisons entre deux ou plusieurs variables qualitatives, et à observer les relations existantes entre les modalités de ces variables.

2.1. La sélection des variables

Pour identifier ces dimensions, une ACM a été réalisée sur 8 variables actives issues de la base de données jointe, soit 2 variables déterminant le type de l'assuré et du bénéficiaire de la couverture maladie, 3 variables sociodémographiques, 1 variable déterminant la pesanteur de la maladie et 2 variables déterminant respectivement le taux de sinistralité et le montant à rembourser par l'organisme gestionnaire:

- Le « Type de l'assuré » type_ass : 4 modalités
- Le « Type du bénéficiaire » type_benef : 3 modalités
- Le « Sexe » (sexe) : 2 modalités
- La « Tranche d'âge » (tranche_age) : 16 modalités
- Le « Région » (Région) : 17 modalités
- L' « Affection longue durée » (ALD) : 2 modalités
- Le « Taux de sinistralité » (taux_sinistralité) : 10 modalités
- Le « Montant moyen remboursé » (MRM) : 6 modalités

Remarque : le taux de sinistralité et le montant remboursé moyen sont à la base des variables quantitatives que nous avons convertis en des variables qualitatives afin de les joindre dans notre analyse factorielle des correspondances multiples.

On note qu'il existe des variables prenant un grand nombre de modalités, ce qui peut engendrer une grande disparité dans les résultats, cependant vue l'intérêt de ces dernières on va les garder dans l'analyse.

2.2. Le choix du nombre de facteurs

Le nombre maximal de facteurs que peut extraire une ACM est égal au nombre de modalités M – le nombre de variables V : soit dans ce cas-ci $66-8 = 58$. Ce nombre est bien entendu trop important et le souci de synthétiser cet espace multidimensionnel de façon optimal conduit à retenir les facteurs les plus performants - statistiquement parlant- en termes de synthèse de l'information et ceux qui ont du sens par rapport à l'espace des variables initial.

L'inertie totale du nuage de points dépend du nombre total de modalités M et du nombre total de variables V , soit $I = (M-V)/V$, soit dans ce cas-ci $I = [(66-8)/8] = 7,25$.

On retient en général les facteurs ayant une valeur propre supérieure à 1 ou encore une inertie (variation expliquée) supérieure à $1/V$, soit une inertie supérieure à l'inertie moyenne d'une variable active. Dans ce cas-ci, cette valeur-seuil s'établit à $0,125 (= 1/8)$.

Le tableau **Récapitulatif des modèles** produit par SPSS a retenu les 4 premiers facteurs sur base de ces critères : ensemble ils rendent compte de 12,78% de l'inertie totale du nuage de points ($0,927/7,25=0,1278$), ce qui est une proportion acceptable dans le cadre de l'ACM. En effet, la variance « expliquée » par les facteurs d'une ACM est sous-estimée et les chercheurs tiennent rarement compte de cette statistique : l'important est de pouvoir donner du sens aux facteurs retenus et de visualiser l'espace complexe des variables initiales en écartant ce qui est considéré comme du bruit.

Model Summary

Dimension	Cronbach's Alpha	Variance Accounted For		
		Total (Eigenvalue)	Inertia	% of Variance
1	,710	2,640	,330	33,004
2	,571	1,998	,250	24,978
3	,322	1,391	,174	17,394
4	,318	1,386	,173	17,325
Total		7,416	,927	
Mean	,526 ^a	1,854	,232	23,175

a. Mean Cronbach's Alpha is based on the mean Eigenvalue.

Tableau 14 : Récapitulatif des modèles

2.3.L'interprétation du premier plan factoriel

- Analyse du tableau des coordonnées des modalités sur les 4 dimensions retenues :

Le tableau ci-dessous détaille les coordonnées des modalités des 8 variables actives par rapport aux 4 dimensions retenues : il a été construit à partir des 8 tableaux partiels (un par variable active) que produit SPSS. Ce tableau est une aide à l'interprétation des plans factoriels qu'on peut constituer à partir des coordonnées des modalités sur chaque couple de facteurs.

Remarque : Ce sont les modalités les plus distantes du « barycentre » (qui correspondant aux coordonnées 0,0) qui contribuent le plus à la construction des axes. Pour rappel, le barycentre correspond au comportement « moyen » de toutes les variables, ainsi plus une modalité s'écarte de ce comportement « moyen », plus elle caractérise des personnes qui diffèrent de ce comportement moyen.

Variables	Modalités	Dimension			
		1	2	3	4
Type_ass	Actif	-.243	.085	.099	-.448
	Pensionné	-.097	.054	-.168	-.220
	Conjoint survivant	.285	-.122	.041	.815
	Enfant survivant	.728	-.277	.097	.870
Type_benef	Assuré	-.312	.393	.117	.797
	Conjoint de l'assuré	-.491	.626	-.267	-.964
	Enfant de l'assuré	.635	-.804	.026	-.327
Sexe	Femme	-.125	.191	-.100	-.315
	Homme	.165	-.252	.132	.415
Tranche_age	[0,1[1.127	-.377	-.450	-.946
	[1,5[1.173	-1.138	-.684	-.358
	[5,10[1.108	-1.262	-.669	-.317
	[10,15[.936	-1.378	-.545	-.380
	[15,20[.708	-.874	-.197	-.305
	[20,25[.428	-.103	2.626	-.815
	[25,30[-.088	-.116	-.201	-.538
	[30,35[-.099	.015	-.348	-.327
	[35,40[-.094	.206	-.225	-.174
	[40,45[-.123	.281	-.327	.447
	[45,50[-.270	.260	-.365	.400
	[50,55[-.458	.367	-.227	.351
	[55,60[-.521	.438	-.189	.373
	[60,65[-.548	.606	-.328	.361
	[65,70[-.500	.648	.383	.271
70+	-.382	.362	.852	.712	
Région	reg1	.109	.399	-.054	.075
	reg2	.093	.113	-.250	-.003
	reg3	.150	.039	.133	.069
	reg4	.001	-.094	-.136	.094
	reg5	-.001	.041	.035	-.095
	reg6	-.057	.054	.015	-.270
	reg7	-.022	.000	.050	-.054
	reg8	.028	-.100	-.017	.105
	reg9	-.079	.008	.059	.189
	reg10	-.070	-.122	-.011	.158
	reg11	-.061	-.025	.041	-.160
	reg12	.042	.043	-.148	-.167
	reg13	.083	-.107	-.017	.001
	reg14	-.069	-.014	.018	.048
	reg15	.025	.007	-.113	-.323
	reg16	-.024	-.118	.148	.223
	reg17	.053	.396	.287	-.168
ALD	ALD-N	.636	.415	-.029	.010
	ALD-O	-.940	-.614	.043	-.014
taux_sinistralité	[0 ; 0,1[1.015	.088	3.548	-.244
	[0,1 ; 0,2[1.142	-.174	.017	.553
	[0,2 ; 0,3[1.218	-.453	-.556	-.040
	[0,3 ; 0,4[.856	.108	-.681	.052
	[0,4 ; 0,5[.327	.647	-.189	.652
	[0,5 ; 0,6[-.040	1.455	-.168	-.966
	[0,6 ; 0,7[-.160	.808	.135	-.558
	[0,7 ; 0,8[-.490	-.924	.213	-.405
	[0,8 ; 0,9[-1.253	-.882	-.045	.344
	[0,9 ; 1[-.852	-.506	-.035	.019
MRM (DH)	(0,1000)	1.125	-.523	-.087	-.024
	(1000,5000)	.072	.755	.001	-.008
	(5000,10000)	-.851	-.794	.089	.403
	(10000,100000)	-1.272	-.663	.059	-.163
	(100000,190000)	-.748	-1.066	.441	1.345
	(190000,1755139)	-.572	-1.802	1.280	.723

Tableau 15 : Coordonnées des modalités sur les 4 premiers facteurs identifiés par l'ACM

L'objectif d'une ACM étant d'offrir une visualisation interprétable d'un espace-variables complexe, le sens donné aux axes et l'analyse des proximités entre variables et modalités sont généralement élaborés à partir des plans factoriels. On se limitera ici au **premier plan factoriel**, composé par les deux premiers facteurs qui représentent ensemble près de 8% de la variance initiale ($0,58/7,25 = 0,08$ avec $0,566 = 0,330 + 0,250$).

- La première dimension est surtout déterminée par l'âge du bénéficiaire, son taux de sinistralité, l'ALD et le montant remboursé moyen : elle oppose clairement ceux ayant un taux de sinistralité plus que 50 % et un montant remboursé plus que 5000 DH à ceux ayant moins. De même, l'âge contribue à cette dimension : plus les bénéficiaires sont âgées, plus elles s'éloignent négativement de la valeur moyenne de cet axe. En fait, le premier axe factoriel oppose les bénéficiaires ayant plus de 25 ans aux ceux ayant moins.
- De même, Ce sont : l'âge du bénéficiaire, le taux de sinistralité et le montant remboursé moyen qui se positionne le plus clairement le long de la deuxième dimension : elle oppose les bénéficiaires ayant plus de 35 ans et ceux ayant moins.
- Les modalités « Homme » et « Femme » se situent à proximité du barycentre du graphique, ce qui veut dire que même si les caractéristiques des hommes et des femmes diffèrent (car leurs modalités sont légèrement distantes l'une de l'autre), elles ne contribuent que faiblement à la construction des axes factoriels. Il s'agit cependant d'un effet de perspective : le sexe contribue un peu plus à la construction de la quatrième dimension.
- Les modalités de la variable région se condensent autour du barycentre, donc elles ne contribuent que faiblement à la construction des axes factoriels.
- Les modalités « avoir une ALD » et « ne pas avoir une ALD » s'opposent. On peut remarquer que la variable ALD contribue plus à la construction de la première dimension.

■ Analyse du diagramme joint des points de modalités :

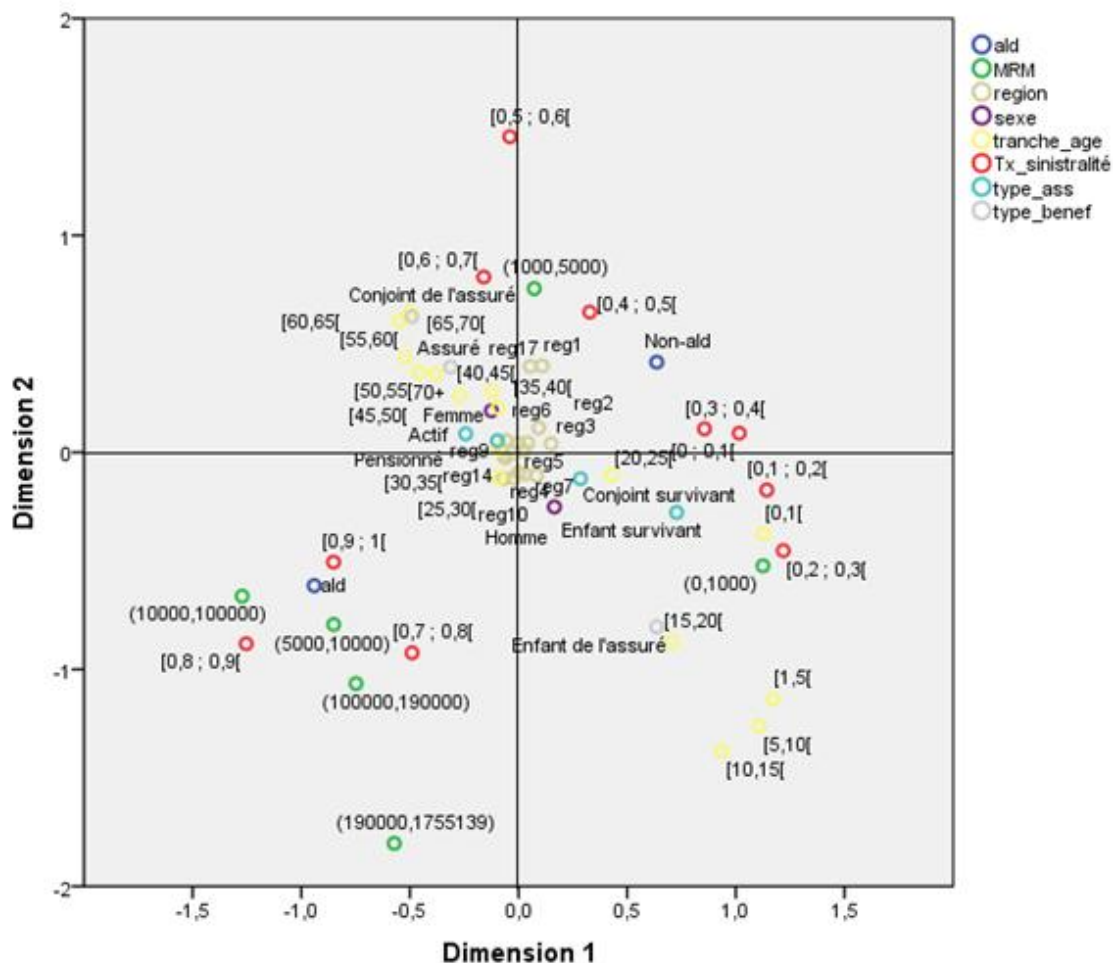


Figure 18 : Représentation des modalités dans le premier plan factoriel

- La proximité entre la modalité « enfant de l'assuré » et les autres modalités « [0,1[» à « [20, 25[» approuve le fait que les enfants bénéficiaires de la couverture médicale sont âgés entre 1 an et 25 ans. Ainsi, on constate un rapprochement entre les modalités « Assuré » et « conjoint de l'assuré » avec les modalités « [25,30[» à « 70+ », cela veut dire que les assurés et leurs conjoints sont âgés de 25 ans à 70 ans et plus, ce qui paraît logique.
- On constate que la modalité « avoir une ALD » est proche des montants remboursés moyens qui dépassent les 5000 DH et des taux de sinistralité dépassant 80%, ce qui prouve que les affections longues durées sont très onéreuses et présentent une sinistralité grave.
- on remarque aussi que les taux de sinistralité moins de 40% ne sont pas loin des montants remboursés moyens moins de 1000 DH, alors que les taux de sinistralité variant entre 40% et 70% sont proche de la tranche 1000 DH à 5000 DH.

VII. Analyse bivariée du taux de sinistralité et du montant remboursé moyen

- L'analyse de la variance à un facteur

L'analyse de la variance entre dans le cadre général du modèle linéaire où une variable quantitative est expliquée par une variable qualitative ou plusieurs. L'objectif essentiel est de comparer les moyennes empiriques de la variable quantitative observées pour les variables qualitatives (facteurs) découpées en classes (niveaux).

Dans cette section, on étudiera l'impact de chaque facteur sur le taux de sinistralité, ainsi que sur le montant remboursé moyen. Mais avant d'entamer une ANOVA à un facteur, il convient de vérifier les conditions de **normalité** de la distribution de la variable quantitative dans les sous-populations définies par les niveaux de chaque facteur, et les conditions d'**homoscédasticité**, qui renvoie à l'égalité des variances intra-groupes de la variable endogène continue.

Le test de Kolmogorov-Smirnov montre que la variable « taux de sinistralité » ainsi que la variable « montant remboursé moyen » ne suivent pas une loi normale dans aucune des sous-populations relatives aux niveaux des variables qualitatives. De même la condition d'homoscédasticité n'est pas satisfaite dans le cas de ces deux variables endogènes.

On pense alors à utiliser des tests non paramétriques qui ne supposent ni homogénéité de la variance, ni une distribution normale, par exemple le test de **Kruskall-Wallis**, cependant ces types de tests sont valables seulement pour les échantillons de tailles petites, ce qui n'est pas notre cas.

On décide alors de continuer avec le test de l'ANOVA, d'ailleurs, les statisticiens ont vu que ce test est « robuste ». Autrement dit, même avec des variables pas tout à fait gaussiennes, on pourra utiliser ce test avec une bonne confiance. C'est aussi le cas lorsque les effectifs sont très importants. En outre, en présence des échantillons de grande taille, la condition d'homoscédasticité peut être négligée.

- Application de l'ANOVA à un facteur pour le cas de la variable « taux de sinistralité »

ANOVA

taux_sinistralité

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5,253	3	1,751	18,400	,000
Within Groups	429,217	4510	,095		
Total	434,470	4513			

Tableau 16 : Résultat de l'ANOVA à un facteur « type_ass »

Plus le p-value est petit, plus la preuve est forte contre l'hypothèse nulle. Ici, les moyennes sont très différentes (car $p\text{-value} = 0,000 < 5\%$). L'hypothèse nulle est rejetée, la variable type de l'assuré a donc un effet sur le taux de sinistralité des bénéficiaires de la CNOPS mais à ce stade, nous ne savons pas quels sont les niveaux de la variable qui sont significativement différents des autres. Pour cela, il faut réaliser un test de comparaisons multiples, aussi appelé test post hoc.

taux_sinistralité

Duncan^{a,b}

type ass	N	Subset for alpha = 0.05		
		1	2	3
Conjoint survivant	424	,5447		
Enfant survivant	892		,5839	
Pensionné	1476		,5996	
Actif	1722			,6494
Sig.		1,000	,294	1,000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 844,275.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Tableau 17 : Résultat du test de Duncan pour le cas de la variable « type_ass »

Le test de Duncan, souvent employé pour des tests de comparaisons de plusieurs moyennes, montre que la modalité « actif » est supérieure aux autres. Le taux de sinistralité moyen à son égard est de 0.64, significativement plus élevé que celui des autres modalités pensionné (0.59), enfant survivant (0.58) ou conjoint survivant (0.54).

On constate aussi que le taux de sinistralité est significativement identique pour les pensionnés et les enfants survivants puisque ces deux modalités appartiennent au même sous-ensemble homogène numéro 2.

De façon analogue, on trouve que les variables type du bénéficiaire, sexe, ald, tranche d'âge et région ont bien un effet sur le taux de sinistralité.

Pour savoir quelles sont les modalités de la variable « type_benef » qui sont significativement différentes en termes de la moyenne du taux de sinistralité, on effectue des tests post hoc.

taux_sinistralité

Duncan^{a,b}

type benef	N	Subset for alpha = 0.05		
		1	2	3
Enfant de l'assuré	1672	,5446		
Assuré	1866		,6336	
Conjoint de l'assuré	976			,6785
Sig.		1,000	1,000	1,000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 1389,801.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Tableau 18 : Résultat du test de Duncan pour le cas de la variable « type_benef »

On constate que la distribution du taux de sinistralité n'est pas identique à l'intérieur des sous-populations définies par les niveaux de la variable « type_benef » et cela pour les 3 niveaux. En effet, le test nous fournit 3 sous-ensembles homogènes chacun comporte une et une seule modalité.

- Application de l'ANOVA à un facteur pour le cas de la variable « MRM »

ANOVA

MRM

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6060274593	3	2020091531	1,904	,127
Within Groups	4,785E+12	4510	1061007875		
Total	4,791E+12	4513			

Tableau 19 : Résultat de l'ANOVA à un facteur « type_ass »

D'après ce tableau, le niveau de signification du test de Fisher est de $0.127 > 5\%$ donc l'hypothèse nulle est acceptée. Ceci dit, que la distribution du montant remboursé moyen est statistiquement identique pour tous les types de l'assuré.

De façon analogue, on trouve que le montant remboursé moyen est identique pour tous les niveaux des facteurs type du bénéficiaire, sexe, tranche d'âge et région, à l'exception de la variable « ald ».

Conclusion

On peut conclure que l'étude descriptive nous a permis de mieux connaître notre portefeuille et d'identifier les variables qui affectent le plus la consommation médicale. Aussi, les techniques de l'analyse de données et les mesures d'association nous ont permis d'obtenir une première vision sur les liens pouvant exister entre les différentes variables. A ce stade les résultats obtenus sont logiques et acceptables.

Chapitre 3 : Modélisation de la consommation médicale

I. Introduction

Dans le contexte de la croissance sans cesse des coûts en assurance maladie, il est évident que des stratégies de gestion de risques de plus en plus sophistiquées s'imposent, et poussent les acteurs à faire recours aux méthodes de modélisation qui contribueront à l'atteinte de quelques objectifs essentiels tel que :

- ✓ Evaluer le coût du risque le plus finement possible,
- ✓ optimiser les bénéfices par rapport aux coûts,
- ✓ identifier les sources des économies potentielles,
- ✓ augmenter le degré de précision des estimations des coûts futurs,
- ✓ assurer la stabilité et la viabilité du régime à long terme.

L'objectif de cette partie est de présenter une méthode d'estimation de la consommation moyenne d'un assuré.

L'orientation générale de l'étude est la recherche d'un modèle global et explicatif qui lie les différentes catégories de dépenses d'assurance maladie entre elles, en tenant compte des spécificités de la population consommatrice et celle sous risque.

1. Les données utilisées

Les données dont nous disposons initialement se décomposent en deux fichiers : le fichier des « Effectifs » concernant les informations sur la population assurée et le fichier des « Prestations » portant sur l'ensemble des prestations médicales dont ont bénéficié ces mêmes assurés.

En vue d'obtenir une base sur laquelle portera l'étude de la modélisation, nous réalisons une jointure entre les 2 fichiers afin d'identifier nos variables d'intérêts à savoir le montant remboursé moyen, noté MRM, qui représente le coût du risque et que nous obtenons en divisant le montant remboursé global par l'effectif sinistré. Et le taux de sinistralité qui représente la fréquence des sinistres que nous obtenons également en divisant l'effectif sinistré de chaque classe par l'effectif.

Nous ne disposons pas des données tête-par-tête, nous suggérons de dupliquer chaque ligne de la base de données selon l'effectif des individus effectivement sinistrés, cependant, vu que le total des effectifs sinistrés est très grand, nous allons garder la structure agrégée de nos données.

2. L'approche retenue: l'approche « Fréquence * Coût moyen * population »

L'objet de cette section est de définir une méthode d'estimation de la consommation médicale. On se base sur une approche de type « Fréquence * Coût moyen * effectif », méthode classiquement utilisée dans le monde de l'assurance car elle est relativement aisée à mettre en œuvre et permet une estimation cohérente des risques considérés.

Elle peut être abordée de manière empirique en se basant sur le principe de décomposition du risque entre la fréquence et le coût moyen, ainsi le montant remboursé par le régime peut être déterminé comme étant le produit de la fréquence par le remboursement moyen par l'effectif.

On a:

$$\begin{aligned} \text{Fréquence} * \text{Coût moyen} * \text{effectif} &= \frac{\text{Effectifs sinistré}}{\text{Effectif}} \times \frac{\text{remboursement}}{\text{Effectifs sinistré}} \times \text{Effectif} \\ &= \text{remboursement} \end{aligned}$$

II. Méthodes de modélisation appliquées en assurance maladie

Il existe de multiples méthodes statistiques qui permettent de quantifier le lien entre la variable à expliquer Y et les variables explicatives X_i , $i = 1, \dots, p$. Elles peuvent être classées dans deux grandes familles :

- ✚ les méthodes paramétriques pour lesquelles la relation recherchée entre Y et X est spécifiée de manière paramétrique a priori telle que les méthodes de la régression linéaire dans le cadre gaussien ou linéaires généralisées à différentes lois de distributions des erreurs (GLM).
- ✚ Les méthodes non paramétriques pour lesquelles la relation recherchée entre Y et X n'est pas spécifiée a priori telles que les modèles additifs généralisés (GAM).

Notre première intuition conduit à l'utilisation de la régression linéaire multiple qui est considérée comme un modèle de référence.

1. La Régression linéaire multiple

1.1.Principe

La régression linéaire multiple est une méthode statistique qui cherche à expliquer les valeurs prises par une variable Y, dite endogène à l'aide d'un nombre p de variables X_j ($j = 1, \dots, p$), dites exogènes.

L'équation de régression s'écrit :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i$$

Avec :

- $i = 1, \dots, n$: indice représentant la $i^{\text{ème}}$ observation, avec $n \geq p$
- y_i : la $i^{\text{ème}}$ observation de la variable Y
- $x_{i,j}$: la $i^{\text{ème}}$ observation de la variable X_j
- ε_i : composante aléatoire représentant l'information omise par le modèle linéaire.

La régression linéaire repose sur les hypothèses suivantes :

H1 : Les X_j sont non aléatoires ;

H2 : $E[\varepsilon_i]$ l'espérance de l'erreur est nulle ;

H3 : $E[\varepsilon_i^2] = \sigma_\varepsilon^2$ la variance de l'erreur est constante ;

H4 : $COV(\varepsilon_i, x_{i,j}) = 0$, l'hypothèse de l'homoscédasticité ;

H5 : $COV(\varepsilon_i, \varepsilon_{i'}) = 0$ pour $i \neq i'$, l'hypothèse de la non-autocorrélation des résidus ;

H6 : $\varepsilon_i \sim N(0, \sigma_\varepsilon)$, l'hypothèse de la normalité des erreurs ;

H7 : La matrice $X'X$ est régulière;

H8 : $\frac{X'X}{n}$ tend vers une matrice finie non singulière lorsque n tends vers l'infini ;

H9 : $n > p + 1$, le nombre d'observations est strictement supérieur au nombre des paramètres à estimer;

La méthode d'estimation des paramètres la plus utilisée est celle des Moindres Carrée Ordinaire (MCO). Le principe est de trouver les paramètres $(\beta_0, \beta_1, \dots, \beta_p)$ qui minimisent la quantité suivante :

$$S = \sum_{i=1}^n (y_i - \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})^2 = \sum_{i=1}^n \varepsilon_i^2$$

Écriture matricielle :

En adoptant une écriture matricielle, l'équation de régression devient :

$$Y = X\beta + \varepsilon$$

Avec :

$$\bullet \quad X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\bullet \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Donc nous avons :

$$S = \varepsilon' \varepsilon = (Y - X\beta)'(Y - X\beta)$$

En annulant la dérivée matricielle de S par rapport à β , on retrouve :

$$\frac{\partial S}{\partial x} = -2(X'Y) + 2(X'X)\beta = 0$$

D'où :

$$(X'X)\beta = (X'Y)$$

L'estimateur des moindres carrés ordinaires est donc :

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

1.2. Régression linéaire pondérée

La régression linéaire multiple considère que toutes les observations ont la même importance, ce qui n'est pas toujours le cas notamment quand :

- Chaque observation est retrouvée plusieurs fois.
- Les observations des variables exogènes sont des moyennes sur des populations de tailles différentes.

Pour remédier à ces situations, on affecte une pondération (ω_i) à chaque observation. On parle alors de Régression Linéaire Pondérée.

Dans ce cas on cherche à minimiser la grandeur suivante :

$$S = \sum_{i=1}^n \omega_i (y_i - \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})^2 = (Y - X\beta)'W(Y - X\beta)$$

Avec :

$$W = \begin{pmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & \omega_n \end{pmatrix}$$

On peut remarquer que :

$$S = (W^{1/2}Y - W^{1/2}X\beta)'(W^{1/2}Y - W^{1/2}X\beta)$$

En posant $W^{1/2}Y = Y_*$ et $W^{1/2}X = X_*$ nous avons :

$$S = (Y_* - X_*\beta)'(Y_* - X_*\beta)$$

Comme établi plus tard, l'estimateur des moindres carrés est :

$$\hat{\beta} = (X_*'X_*)^{-1}X_*'Y_* = (X'WX)^{-1}X'WY$$

2. Mise en œuvre de la régression linéaire multiple

A présent, appliquons un modèle de régression linéaire pondérée pour expliquer le montant remboursé moyen « MRM » à l'aide des inputs de la base de données jointe.

On utilisera le logiciel SAS qui met à notre disposition la procédure « PROC REG » qui permet de réaliser la régression linéaire de la variable numérique MRM sur le sous-espace engendré par les variables numériques $x_1 \dots x_p$.

Dans notre cas, les variables exogènes sont toutes catégorielles. Il s'avère donc nécessaire de recoder nos variables à l'aide de variables binaires.

L'instruction « WEIGHT » de la procédure « PROC REG » nous permet de pondérer nos observations par la variable « effectif_sinistré ».

Le modèle à ajuster sera représenté par l'équation symbolique suivante :

$$\text{MRM} = \text{Type_ass} + \text{Type_benef} + \text{Ald} + \text{Region} + \text{Tranche_age} + \text{Sexe}$$

On évitera d'introduire les variables d'interaction afin d'obtenir un modèle moins complexe.

Par défaut, le modèle intègre une constante parmi les régresseurs.

La sortie fournie par SAS est la suivante :

Valeurs estimées des paramètres						
Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > t
Intercept	Intercept	1	1781,34792	1761,71137	1,01	0,312
actif		1	-1166,63871	1416,84474	-0,82	0,4103
pensionne		1	-861,80544	1421,01693	-0,61	0,5442
CS		1	-1270,85674	1433,72201	-0,89	0,3754
assuré		1	887,6502	384,59982	2,31	0,021
CA		1	612,12475	379,83068	1,61	0,1071
homme		1	-251,9489	101,67274	-2,48	0,0132
ald_oui		1	10658	153,35678	69,5	<.0001
region1		1	1912,96783	1211,43302	1,58	0,1144
region2		1	628,07455	1021,98513	0,61	0,5389
region3		1	502,50061	1018,24802	0,49	0,6217
region4		1	604,48125	981,83971	0,62	0,5381
region5		1	728,29634	989,62276	0,74	0,4618
region6		1	771,55489	989,1799	0,78	0,4354
region7		1	714,05234	983,21416	0,73	0,4677
region8		1	531,70773	983,22626	0,54	0,5887
region9		1	1052,88356	975,379	1,08	0,2804
region10		1	931,91401	973,88569	0,96	0,3387
region11		1	905,23438	989,17471	0,92	0,3602
region12		1	936,4324	997,48684	0,94	0,3479
region13		1	688,46293	981,98598	0,7	0,4833
region14		1	827,26749	985,02515	0,84	0,401
region15		1	700,74651	997,48274	0,7	0,4824
region16		1	660,38088	984,17573	0,67	0,5023
age1		1	-2,04582	544,4276	0	0,997
age2		1	-596,05139	488,15203	-1,22	0,2221
age3		1	-669,42078	478,21874	-1,4	0,1616
age4		1	-529,51957	474,07145	-1,12	0,2641
age5		1	-400,71889	453,94521	-0,88	0,3774
age6		1	-132,49082	361,12676	-0,37	0,7137
age7		1	-172,97592	281,68273	-0,61	0,5392
age8		1	-153,01527	268,35439	-0,57	0,5686
age9		1	-332,24585	264,11408	-1,26	0,2085
age10		1	-332,68866	257,31617	-1,29	0,1961
age11		1	-346,18804	242,59006	-1,43	0,1536
age12		1	-325,93998	229,60075	-1,42	0,1558
age13		1	-406,13755	223,41577	-1,82	0,0692
age14		1	-355,22983	229,36739	-1,55	0,1215
age15		1	62,65969	265,03282	0,24	0,8131

Tableau 20 : Valeurs estimées des paramètres

Le tableau ci-dessus liste les régresseurs avec une estimation du paramètre associé, l'écart-type estimé pour le dit paramètre, ainsi que la statistique et la p-value relative à un test de Student de nullité du paramètre (l'hypothèse nulle étant justement la nullité du paramètre).

Ici on voit que les variables ald_oui, assuré et homme sont les seules à être significatives car leur p-value est inférieur à 5%. Cependant presque la totalité de nos régresseurs sont non significatives, ceci dit que le modèle de régression linéaire s'ajuste mal à nos données.

Racine MSE	47597	R carré	0,5723
Moyenne dépendante	2639,09943	R car. ajust.	0,5687
Coeff Var	1803,53523		

Tableau 21 : Statistiques d'ajustement

Le coefficient de détermination R^2 est un indicateur qui permet de juger la qualité d'une régression linéaire, simple ou multiple. Il mesure l'adéquation entre le modèle et les données observées.

Dans notre cas, il est de 57.23%, ce qui prouve que l'ajustement est faible.

Nous allons maintenant examiner les résidus du modèle ajusté pour s'assurer s'ils vérifient les hypothèses de la régression linéaire citées précédemment.

On commence tout d'abord par dessiner l'histogramme des résidus pour avoir une idée de leur distribution :

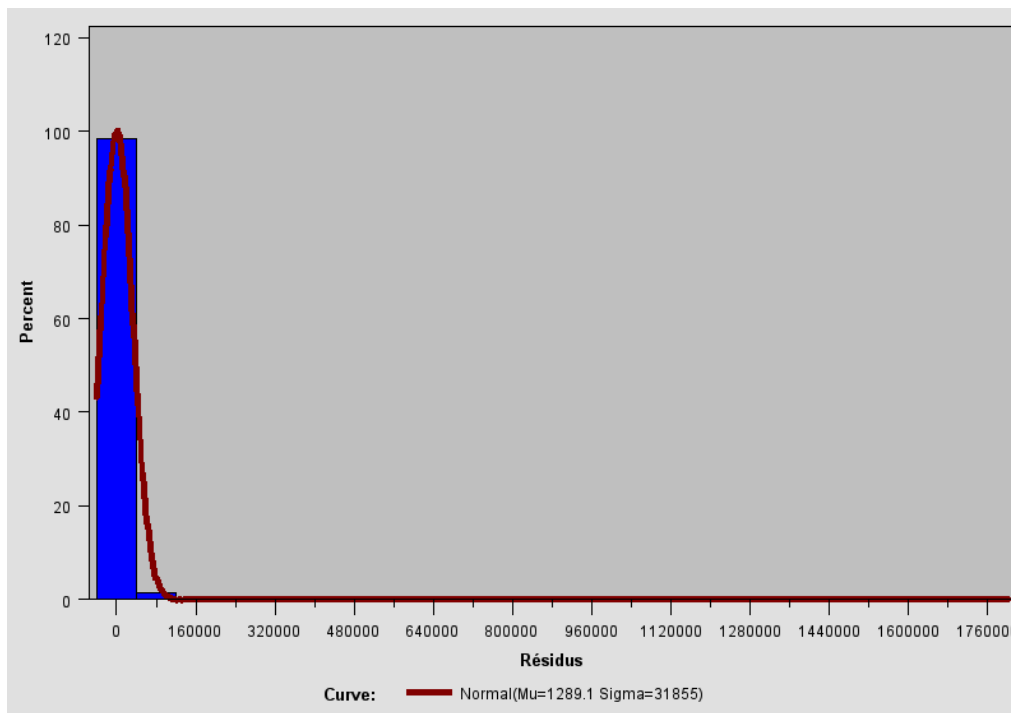


Figure 19 : Histogramme ajusté des résidus du modèle simple de la régression linéaire

D'après ce graphique, on constate que la distribution est étalée vers la droite, ce qui ne présente pas une allure s'apparentant à une loi normale.

Ceci est confirmé par les résultats obtenus des coefficients d'asymétrie et d'aplatissement suivants :

Moments			
N	4514	Somme des poids	4514
Moyenne	1289,12687	Somme des observations	5819118,67
Ecart-type	31855,1148	Variance	1014748342
Skewness	40,0266926	Kurtosis	2038,16484
Somme des carrés non corrigée	4,58706E+12	Somme des carrés corrigée	4,57956E+12
Coeff. variation	2471,06128	Moy. erreur std	474,13105

Tableau 20 : Coefficients d'asymétrie et d'aplatissement sur les résidus de la régression linéaire

Le coefficient d'asymétrie «**Skewness**» de la distribution des résidus est de l'ordre de 40.02 >> 0 donc la distribution penche à droite.

Le coefficient d'aplatissement «**kurtosis**» donne une information sur les queues de distribution. Ce coefficient est de 2038,16 >> 0 ce qui indique que les queues comptent plus d'observations que dans une distribution normale. Sachant que le kurtosis d'une distribution normale est égale à 3.

Tests for Normality			
Test	Statistique		p Value
Kolmogorov-Smirnov	D	0,36	Pr > D <0.0100
Cramer-von Mises	W-Sq	234,032	Pr > W-Sq <0.0050
Anderson-Darling	A-Sq	1137,389	Pr > A-Sq <0.0050

Tableau 21: résultat du test de normalité des résidus

Le test de Kolmogorov-Smirnov, permet de comparer une distribution observée avec une autre, ou avec une distribution connue de type loi de probabilité. Notamment, ce test donne une bonne indication d'ajustement à une loi normale. On constate que le nombre des observations est égale 4514, ce qui est largement supérieur à 30, donc le test de Kolmogorov-Smirnov est le plus approprié pour tester la normalité des résidus.

Comme la p-value de ce test est inférieur à 5%, notre distribution est donc différente de la normale.

Dressons le diagramme des résidus en fonction des valeurs prédites comme suit :

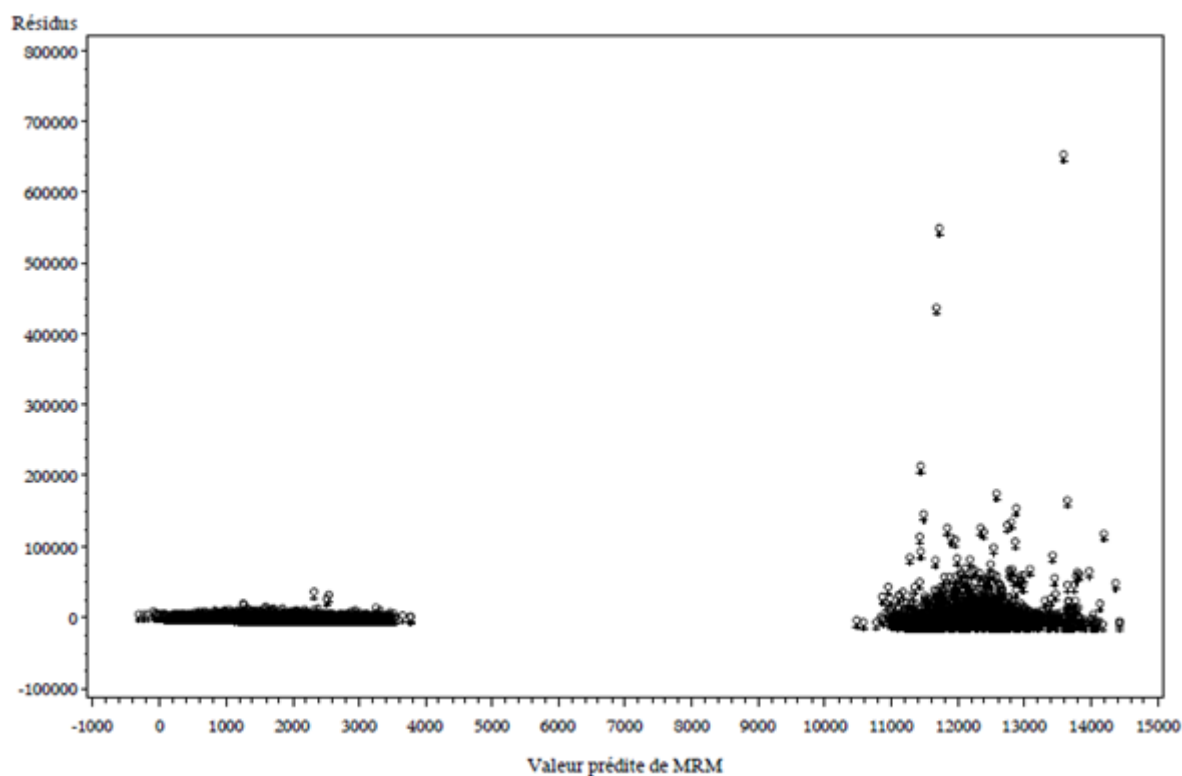


Figure 22 : Résidus du modèle simple de régression linéaire pondérée

On remarque que les résidus présentent une certaine discontinuité dans leur distribution. Ceci peut être expliqué par la non-homogénéité des observations. Chose qui montre que le modèle de la régression linéaire est inapproprié à la distribution du montant moyen remboursé.

En guise de conclusion, la distribution de la variable d'intérêt MRM est non gaussienne, donc le modèle de régression multiple impose de sérieuses limitations.

Face aux problèmes que posent le modèle de régression multiple à savoir la non-linéarité de la variable endogène par rapport aux variables exogènes, la non-homogénéité de la population (discontinuité des résidus) et la non-normalité des résidus du modèle de régression linéaire, on se tourne vers les modèles linéaires généralisés afin de quantifier l'impact des variables explicatives sur une notre variable d'intérêt.

3. Les modèles linéaires généralisés GLM

3.1. Introduction

Les modèles linéaires généralisés ne sont utilisés que depuis récemment dans les problèmes rencontrés en assurance.

Les actuaires se sont en effet longtemps limités aux modèles linéaires gaussiens afin de quantifier l'impact des variables explicatives sur une variable d'intérêt. La complexité des phénomènes à expliquer s'étant considérablement accrue ces dernières années, les actuaires ont dû petit à petit se tourner vers des modèles prenant mieux en compte cette complication croissante.

Les modèles linéaires généralisés ont été utilisés pour la première fois en assurance par les actuaires de la City University de Londres à la fin du 20^{ème} siècle. Ils sont depuis couramment utilisés pour la modélisation de phénomènes complexes, aussi bien en assurance non-vie qu'en assurance vie.

3.2. Principe

L'objectif du GLM est de modéliser la moyenne d'une variable Y endogène notée $\mu = E(Y/X)$ à l'aide d'une combinaison de variables exogènes.

Ces modèles permettent de tenir compte des distributions de réponses autres que la distribution normale.

Un modèle GLM a la structure suivante :

$$G(\mu) = X'\beta$$

Où :

- $\mu = E(Y/X)$;
- G : fonction de lien monotone ;
- X_i : la $i^{\text{ème}}$ ligne de la matrice X relative à l'individu i ;
- β : le vecteur de paramètres (inconnu).

L'estimation des modèles GLM est basée sur la théorie du maximum de vraisemblance qui requiert une approche d'estimation par moindres carrés itératifs.

Ainsi, pour chacune des distributions de la famille exponentielle il existe au moins une fonction de lien dite canonique qui simplifie la procédure d'estimation.

3.3. La famille exponentielle

On dit qu'une loi de probabilité appartient à la famille exponentielle si sa distribution s'écrit sous la forme :

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

Les fonctions a, b et c varient d'une distribution à une autre.

ϕ est un paramètre d'échelle arbitraire et θ est connu comme le paramètre canonique de la distribution.

La famille exponentielle regroupe plusieurs lois usuelles parmi eux :

- **Loi de Poisson :**

Soit Y une variable aléatoire suivant une loi Poisson de paramètre λ , sa fonction de densité s'écrit sous la forme :

$$f_{\lambda}(y) = P_{\lambda}(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!} = \exp(y \ln \lambda - \lambda - \ln y!)$$

Par identification:

$$\theta = \ln \lambda ; \quad \phi = 1; \quad a(\phi) = \phi \quad ; \quad b(\theta) = \exp(\theta) = \lambda \quad ; \quad c(y, \phi) = -\ln y!$$

- **Loi Binomiale-Négative :**

Soit Y une variable aléatoire suivant une loi Binomiale-Négative de paramètre r et p, sa fonction de densité s'écrit sous la forme :

$$f_{r,p}(y) = \binom{y+r-1}{y} (1-p)^r p^y \quad y \in \mathbb{N}$$

Que l'on peut écrire :

$$f_{r,p}(y) = \exp\left(y \log p + r \log(1-p) + \log\left(\binom{y+r-1}{y}\right)\right)$$

Par identification :

$$\theta = \log p ; \quad \phi = 1; \quad a(\phi) = \phi \quad ; \quad b(\theta) = -r \log p \quad ;$$

$$c(y, \phi) = \log\left(\binom{y+r-1}{y}\right)$$

▪ **Loi normale :**

Soit Y une variable aléatoire suivant une loi normale d'espérance μ et de variance σ^2 , pour toute $y \in \mathbb{R}$ la fonction de densité est comme suit :

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

qui peut être mise sous la forme :

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

Par identification :

$$\theta = \mu \quad ; \quad \emptyset = \sigma^2 \quad ; \quad a(\emptyset) = \emptyset \quad ; \quad b(\theta) = \frac{\theta^2}{2} \quad ; \quad c(y, \emptyset) = -\frac{1}{2}\left[\frac{y^2}{\emptyset} + \ln(2\pi\emptyset)\right]$$

Donc la loi normale appartient à la famille exponentielle.

▪ **Loi Gamma :**

Soit Y une variable aléatoire qui suit une loi Gamma de paramètre r et α . La densité associée à la loi Gamma peut s'écrire sous la forme :

$$f_{r,\alpha}(y) = \frac{\alpha^r}{\sqrt{\Gamma(r)}} y^{r-1} \exp(-\alpha y)$$

Avec : $\Gamma(x) = \int_0^\infty e^{-u} u^{x-1} du$

Cette densité appartient bien à la famille exponentielle, elle peut en effet se mettre sous la forme :

$$f_{r,\alpha}(y) = \exp(-r \ln \alpha - \alpha \ln y - \alpha y + (r-1) \ln y - \ln(\Gamma(\alpha)))$$

Par identification :

$$\theta = -\alpha \quad ; \quad \emptyset = 1 \quad ; \quad a(\emptyset) = 1 \quad ; \quad b(\theta) = -r \ln(-\theta) \quad ; \\ c(y, \emptyset) = (r-1) \ln y - \ln(\Gamma(\alpha))$$

3.4. Fonction de lien canonique

Pour une loi de probabilité appartenant à la famille exponentielle, la fonction de lien canonique G est celle qui permet d'écrire :

$$G(\mu) = \theta$$

Cette fonction de lien permet de simplifier la procédure d'estimation.

Loi de probabilité	Fonction de lien canonique
Normale	$g = \mu$
Poisson	$g = \ln \mu$
Gamma	$g = 1/\mu$
Binomiale	$g = \ln \mu - \ln (1-\mu)$

Tableau 22 : Fonctions de lien pour les lois de la famille exponentielle

3.5. Estimation des paramètres

L'appartenance de la distribution de Y à la famille exponentielle permet de calculer les estimateurs du maximum de vraisemblance des paramètres des modèles linéaires généralisés.

On admet les deux résultats suivants relatifs aux lois de la famille exponentielle

$$E(Y) = \mu = b'(\theta)$$

$$var(Y) = V(\mu)a(\varnothing) = b''(\theta)a(\varnothing)$$

Comme leur nom l'indique, les estimateurs du maximum de vraisemblance reposent sur la maximisation de la fonction de vraisemblance ou celle du log-vraisemblance de l'échantillon étudié.

Pour des observations (Y_i, X_i) , où $i = 1, \dots, n$, la fonction de log-vraisemblance s'écrit :

$$l(Y, \mu, \varnothing) = \sum_{i=1}^n l(Y_i, \mu_i, \varnothing) = \sum_{i=1}^n \left\{ \frac{Y_i \theta_i - b(\theta_i)}{a(\varnothing)} + c(Y_i, \varnothing) \right\}$$

Maximiser cette grandeur revient à maximiser :

$$\sum_{i=1}^n \{Y_i \theta_i - b(\theta_i)\}$$

On définit la déviance comme étant :

$$D = -2 \sum_{i=1}^n \{Y_i \theta_i - b(\theta_i)\}$$

La littérature propose généralement de minimiser la déviance, ce qui est équivalent à la maximisation de la vraisemblance.

Pour minimiser D nous introduisons son gradient comme suit :

$$\nabla(\beta) = \frac{\partial}{\partial \beta} \left[-2 \sum_{i=1}^n \{Y_i \theta_i - b(\theta_i)\} \right] = -2 \sum_{i=1}^n \{Y_i - b'(\theta_i)\} \frac{\partial}{\partial \beta} \theta_i$$

Notre objectif est de résoudre :

$$\nabla(\beta) = 0$$

Généralement, résoudre ce problème revient à résoudre un système d'équations non-linéaires en β . Une approche itérative est donc nécessaire. La méthode généralement utilisée est la méthode IRWLS (Iteratively reweighted least squares) qui repose sur l'algorithme de Newton-Raphson. Il est à noter que l'IRWLS est implémentée dans la plus part des logiciels statistique, et donc il n'est pas nécessaire de la programmer étape-par-étape.

4. Les modèles additifs généralisés « GAM »

On note que les GLM se basent sur un modèle linéaire, cependant on ne peut pas être sûr que les variables explicatives interviennent toutes linéairement.

Pour remédier à ce problème, on introduit les modèles Additifs Généralisés (GAM) qui consistent à permettre à Y d'être une fonction additive non nécessairement linéaire des variables explicatives.

Les modèles additifs généralisés ont connu une popularité croissante durant l'année 1995.

Les GAM cherchent à estimer un vecteur f de fonctions non paramétriques modélisant l'impact de chaque variable explicative dans le modèle.

Ils se présentent comme suit :

$$E(Y) = \mu = \beta_0 + \sum_{j=1}^P f_j(X_j)$$

Comme pour les GLM,

- $E(\varepsilon_i) = 0$ et $var(\varepsilon_i) = \sigma_i^2$; $\forall i \in [1, n]$
- La fonction de distribution de Y appartient à la famille exponentielle.
- La fonction g (fonction de lien) est monotone et différentiable.

Les fonctions f_j sont des fonctions quelconques d'une ou plusieurs variables. Elles peuvent être paramétriques (polynomiales, trigonométriques) ou non paramétriques comme les fonctions splines.

Les fonctions qui caractérisent le GAM sont estimées seulement pour les variables quantitatives.

Dans le cas de notre portefeuille, les variables sont toutes qualitatives, donc elles continuent à intervenir sous forme linéaire comme dans le GLM.

On peut conclure alors que le modèle de GAM n'apportera pas un plus à la qualité de prédiction par rapport au GLM.

5. Mise en œuvre des modèles linéaires généralisés

On étudie dans cette partie la façon dont les modèles linéaires généralisés sont appliqués à notre étude. Notamment on se penche sur les paramétrages nécessaires à la prise en compte des variables explicatives, ainsi que les modélisations qui sont retenues pour nos deux variables expliquées qui sont la fréquence moyenne des sinistres et leur coût moyen annuel.

5.1. La segmentation des variables

Pour les variables explicatives retenues pour la modélisation, la détermination d'un niveau de segmentation s'impose.

Si cette segmentation est implicite pour certaines variables comme le sexe (Homme ; Femme), le type du bénéficiaire (Assuré ; Conjoint ; Enfant) et l'Affections Longue Durée (oui ; Non), elle l'est moins pour des variables prenant un grand nombre de modalités, comme l'âge ou la région du bénéficiaire.

Un traitement spécifique à ces 2 variables doit donc être effectué.

5.1.1. Le niveau de segmentation de la région

On effectue ici une analyse sur l'effet de la région dans le but d'en réduire le nombre de classes. Pour cela, une première modélisation est réalisée sur l'effet de la région sur le montant remboursé moyen. Les procédures d'ajustement nous conduisent à utiliser l'ajustement à la loi Log-Normale, avec le lien logarithme.

La région « reg9 » est choisie comme région de référence par le paramétrage SAS.

On obtient les résultats suivants :

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de Wald à		Khi-2 de Wald	Pr > Khi-2
Intercept		1	7.6938	0.0292	7.6366	7.751	69430.1	<.0001
region	reg1	1	-0.396	0.1961	-0.7803	-0.0116	4.08	0.0435
region	reg10	1	-0.0649	0.0385	-0.1404	0.0106	2.84	0.0918
region	reg11	1	-0.2148	0.0602	-0.3327	-0.0969	12.74	0.0004
region	reg12	1	-0.3574	0.0693	-0.4933	-0.2216	26.58	<.0001
region	reg13	1	-0.3907	0.0511	-0.4909	-0.2904	58.37	<.0001
region	reg14	1	-0.1924	0.0552	-0.3006	-0.0842	12.16	0.0005
region	reg15	1	-0.4447	0.0691	-0.5802	-0.3092	41.39	<.0001
region	reg16	1	-0.2765	0.0541	-0.3825	-0.1705	26.12	<.0001
region	reg17	1	-0.435	0.2605	-0.9456	0.0756	2.79	0.0949
region	reg2	1	-0.5323	0.0909	-0.7105	-0.3542	34.3	<.0001
region	reg3	1	-0.6771	0.0879	-0.8493	-0.5049	59.4	<.0001
region	reg4	1	-0.4297	0.0504	-0.5285	-0.3309	72.63	<.0001
region	reg5	1	-0.1972	0.0608	-0.3163	-0.0781	10.53	0.0012
region	reg6	1	-0.3111	0.0601	-0.4289	-0.1933	26.78	<.0001
region	reg7	1	-0.2602	0.0527	-0.3635	-0.1568	24.36	<.0001
region	reg8	1	-0.3545	0.0527	-0.4578	-0.2512	45.21	<.0001
region	reg9	0	0	0	0	0	.	.
Scale		1	12.7164	0.1339	12.4568	12.9815		

Tableau 23 : Résultats des calculs SAS à l'issue de la 1^{ère} étape de la méthode de segmentation de la région

Certaines modalités ne sont pas significatives, on décide donc d'adopter une méthode de pas à pas pour segmenter cette variable.

Une méthode de pas à pas classique de type « **backward** » consiste à éliminer une à une les variables les moins significatives du modèle jusqu'à obtenir un modèle où toutes les variables sont significatives.

Ici, nous ne raisonnons pas avec des variables mais avec des modalités de la variable Région, modalités que nous voulons au final conserver en totalité ou, au moins que nous voulons conserver sous forme de modalités agrégées.

Le modèle tel qu'il est construit utilise une variable de référence, pour laquelle le coefficient obtenu sera nul (et au final égal à 100 % en passant à l'exponentiel).

On agrège ainsi tour à tour les modalités les moins significatives à la modalité de référence qui est ici est la région « reg9 ». En effet, à chaque tour, le fait qu'une modalité ne soit pas significative indique qu'elle n'est pas significativement différente du coefficient de référence. La première modalité agrégée est la région « reg17 ».

L'ensemble des modalités du modèle sont significatives lorsque l'on a agrégé avec la région « reg9 » les régions « reg17 », « reg10 » et « reg1 ».

On obtient un modèle avec les résultats suivants :

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de		Khi-2 de Wald	Pr > Khi-2
Intercept		1	7.6509	0.0189	7.6139	7.688	163618	<.0001
region	reg11	1	-0.172	0.056	-0.2816	-0.0623	9.44	0.0021
region	reg12	1	-0.3146	0.0657	-0.4434	-0.1858	22.92	<.0001
region	reg13	1	-0.3478	0.0461	-0.4381	-0.2575	56.98	<.0001
region	reg14	1	-0.1496	0.0505	-0.2486	-0.0505	8.76	0.0031
region	reg15	1	-0.4019	0.0655	-0.5302	-0.2735	37.64	<.0001
region	reg16	1	-0.2337	0.0494	-0.3304	-0.1369	22.41	<.0001
region	reg2	1	-0.4895	0.0882	-0.6624	-0.3166	30.79	<.0001
region	reg3	1	-0.6343	0.0851	-0.801	-0.4676	55.59	<.0001
region	reg4	1	-0.3868	0.0453	-0.4756	-0.2981	72.98	<.0001
region	reg5	1	-0.1544	0.0566	-0.2653	-0.0434	7.44	0.0064
region	reg6	1	-0.2683	0.0559	-0.3778	-0.1587	23.04	<.0001
region	reg7	1	-0.2173	0.0478	-0.3111	-0.1236	20.65	<.0001
region	reg8	1	-0.3117	0.0478	-0.4054	-0.2179	42.45	<.0001
region	reg_n	0	0	0	0	0	.	.
Scale		1	12.7285	0.134	12.4685	12.9938		

Tableau 24 : Résultats des calculs SAS à l'issue de la dernière étape de la méthode segmentation de la région

On décide au final de conserver 2 modalités agrégées pour les régions : celles ayant un effet neutre sur la variable de référence, et celles ayant un effet minorant sur la variable de référence (c'est-à-dire, celles dont le coefficient est négatif).

A la vue de ces résultats, les 2 groupes se composent respectivement des régions suivantes :

- Un effet minorant des montants remboursés moyens pour « reg2 », « reg3 », « reg4 », « reg5 », « reg6 », « reg 7 », « reg8 », « reg11 », « reg12 », « reg13 », « reg14 », « reg15 » et « reg16 », toutes agrégés dans la modalité « reg_m ».
- Un effet neutre pour les autres, soit « reg9 », les régions « reg17 », « reg10 » et « reg1 », toutes agrégés dans la modalité « reg_n ».

5.1.2. Le niveau de segmentation de la variable tranche d'âge

Selon la même démarche, on effectue ici une analyse sur l'effet de la variable tranche d'âge dans le but d'en réduire le nombre de classes.

La modalité « AG9 » est choisie comme tranche d'âge de référence par le paramétrage SAS.

On obtient les résultats suivants :

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de		Khi-2 de Wald	Pr > Khi-2
Intercept		1	7.4534	0.0334	7.3879	7.5189	49730.6	<.0001
tranche_age	AG1	1	-0.3918	0.0698	-0.5286	-0.255	31.5	<.0001
tranche_age	AG10	1	0.0046	0.0463	-0.0862	0.0954	0.01	0.921
tranche_age	AG11	1	0.1147	0.0441	0.0284	0.2011	6.78	0.0092
tranche_age	AG12	1	0.2819	0.0426	0.1984	0.3655	43.73	<.0001
tranche_age	AG13	1	0.4342	0.0433	0.3493	0.5191	100.44	<.0001
tranche_age	AG14	1	0.6397	0.0465	0.5486	0.7308	189.31	<.0001
tranche_age	AG15	1	0.8378	0.0537	0.7324	0.9431	243.05	<.0001
tranche_age	AG16	1	0.7524	0.0458	0.6626	0.8422	269.82	<.0001
tranche_age	AG2	1	-0.8963	0.0509	-0.9961	-0.7964	309.46	<.0001
tranche_age	AG3	1	-0.938	0.0478	-1.0317	-0.8443	384.92	<.0001
tranche_age	AG4	1	-0.7604	0.0476	-0.8537	-0.6671	255.01	<.0001
tranche_age	AG5	1	-0.6424	0.0462	-0.7331	-0.5518	192.94	<.0001
tranche_age	AG6	1	-0.125	0.0564	-0.2356	-0.0145	4.91	0.0267
tranche_age	AG7	1	0.0549	0.0506	-0.0443	0.1541	1.18	0.2779
tranche_age	AG8	1	0.0522	0.0479	-0.0416	0.1461	1.19	0.2753
tranche_age	AG9	0	0	0	0	0	.	.
Scale		1	9.4789	0.0998	9.2854	9.6765		

Tableau 25 : Résultats des calculs SAS à l'issue de la 1^{ère} étape de la méthode segmentation de la tranche âge

On agrège tour à tour les modalités les moins significatives à la modalité de référence qui est ici « AG9 ». En effet, à chaque tour, le fait qu'une modalité ne soit pas significative indique qu'elle n'est pas significativement différente du coefficient de référence. La première modalité agrégée est la tranche d'âge « AG10 ». L'ensemble des modalités du modèle sont significatives lorsque l'on a agrégé avec la modalité « AG9 » les tranches d'âges « AG10 », « AG7 » et « AG8 ».

On obtient un modèle avec les résultats suivants :

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de		Khi-2 de Wald	Pr > Khi-2
Intercept		1	7.4789	0.0171	7.4453	7.5125	190618	<.0001
tranche_age	AG1	1	-0.4173	0.0636	-0.542	-0.2925	42.99	<.0001
tranche_age	AG11	1	0.0893	0.0334	0.0237	0.1548	7.13	0.0076
tranche_age	AG12	1	0.2565	0.0315	0.1947	0.3183	66.14	<.0001
tranche_age	AG13	1	0.4087	0.0325	0.3451	0.4724	158.51	<.0001
tranche_age	AG14	1	0.6142	0.0366	0.5425	0.6859	281.87	<.0001
tranche_age	AG15	1	0.8123	0.0454	0.7232	0.9014	319.53	<.0001
tranche_age	AG16	1	0.7269	0.0357	0.6569	0.7969	414.48	<.0001
tranche_age	AG2	1	-0.9217	0.0421	-1.0043	-0.8392	479.24	<.0001
tranche_age	AG3	1	-0.9635	0.0382	-1.0384	-0.8885	634.68	<.0001
tranche_age	AG4	1	-0.7859	0.038	-0.8604	-0.7114	427.6	<.0001
tranche_age	AG5	1	-0.6679	0.0363	-0.739	-0.5968	339.01	<.0001
tranche_age	AG6	1	-0.1505	0.0486	-0.2457	-0.0553	9.6	0.0019
tranche_age	neut	0	0	0	0	0	.	.
Scale		1	9.4813	0.0998	9.2876	9.6789		

Tableau 26 : Résultats des calculs SAS à l'issue de la dernière étape de la méthode segmentation de la tranche âge

Finalement, on décide de conserver 3 modalités agrégées pour la variable tranche d'âge : celles ayant un effet majorant sur la variable de référence, celles ayant un effet neutre sur la variable de référence, et celles ayant un effet minorant sur la variable de référence.

A la vue de ces résultats, les 3 groupes se composent respectivement des régions suivantes :

- Un effet majorant des montants remboursés moyens pour les tranches d'âges « AG11 », « AG12 », « AG13 », « AG14 », « AG15 », et « AG16 » agrégés dans la modalité « majo », elle concerne les personnes âgées de plus de 45 ans.
- Un effet neutre pour les autres, soit les tranches d'âges « AG9 », « AG10 », « AG7 » et « AG8 » agrégés dans la modalité « neut », elle concerne les personnes âgées de 25 ans à 45 ans.
- Un effet minorant des montants remboursés moyens pour « AG1 », « AG2 », « AG3 », « AG4 », « AG5 » et « AG6 » agrégés dans la modalité « mino », elle concerne les personnes âgées de moins de 25 ans.

5.2. Le choix du modèle

Il est couramment constaté dans ce type d'étude la présence d'interactions entre les variables explicatives. On entend par interaction le fait que la modalité d'une variable influence la manière dont est adoptée la modalité d'une autre variable.

Dans note étude, on va considérer le modèle simple sans interaction, c'est-à-dire, il comporte seulement les facteurs suivants : « type_ass », « type_benef », « sexe », « ald », « tranche_age » et « région ».

Symboliquement, on écrit :

$$MRM = type_ass + type_benef + sexe + ald + tranche_age + region$$

On procédera par la suite à une étude de l'effet de chaque facteur introduit dans l'équation ci-dessus sur la variable endogène afin de savoir s'il est indispensable à notre modèle ou on peut l'éliminer.

5.3. Le choix de la fonction lien logarithme

Le choix de la fonction lien est une étape importante de la modélisation, car ce choix ne sera pas neutre sur le résultat. On note toutefois qu'il est courant dans ce type d'étude de retenir la fonction de lien logarithme, qui présente l'avantage de fournir un tarif multiplicatif facilement compréhensible.

Notre choix se porte donc sur un modèle multiplicatif avec une fonction lien logarithme qui exprime la prime pure d'un assuré quelconque comme un pourcentage

de la prime d'un assuré de référence. En pratique, les bases de tarification seront donc facilement interprétables.

Par le choix du lien logarithme on obtient :

$$\ln \mu_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad \text{avec} \quad \mu_i = E(y_i), \quad y_i \text{ est la variable endogène}$$

Et ainsi :

$$\mu_i = \exp \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right) = \exp(\beta_0) \prod_{j=1}^p \exp(\beta_j x_{ij})$$

On distingue ainsi clairement la prime de base à laquelle sont appliqués les coefficients de manière multiplicative.

5.4. L'approche fréquence/coût moyen

La fréquence de survenance et le coût moyen d'un sinistre sont des éléments importants dans la modélisation des remboursements en assurance santé. Il est donc important pour l'actuaire de les connaître au mieux. Une modélisation de ces deux piliers est un excellent moyen d'apprécier le risque engendré par un groupe d'assurés. Nous aborderons dans ce qui suit la modélisation du coût moyen des sinistres d'une part et celle de la fréquence de survenance d'autre part.

L'objectif est de trouver une loi statistique connue la plus proche possible du profil des observations et ensuite estimer ses paramètres en fonction de l'expérience du portefeuille. En général la fréquence des sinistres est modélisée par une loi de Poisson ou Binomiale-Négative, et le coût moyen des sinistres par une des lois Gamma ou bien Log-Normale.

6. La modélisation du montant remboursé moyen

Le coût moyen d'un sinistre est un paramètre actuariel que l'on peut calculer lorsque la loi statistique du coût est parfaitement connue. Il s'agit ici de déterminer le montant que l'organisme gestionnaire aura à déboursier pour faire face à ses sinistres. Ainsi le coût moyen peut être aisément déterminé si l'on dispose de données en quantité suffisamment importante.

Dans cette partie, On va chercher à modéliser le montant remboursé moyen « **MRM** » de chaque classe de la population consommatrice. Classiquement, on utilise la loi Gamma ou la loi Log-Normale.

6.1. L'adéquation des montants remboursés moyens aux lois théoriques

Deux lois sont à notre disposition pour modéliser notre variable d'intérêt « MRM », il nous reste donc à choisir celle qui s'ajustera le mieux aux données. Pour cela, on procède dans un premier temps à un ajustement graphique de la loi théorique à la loi empirique grâce au PP-PLOT comme suit :

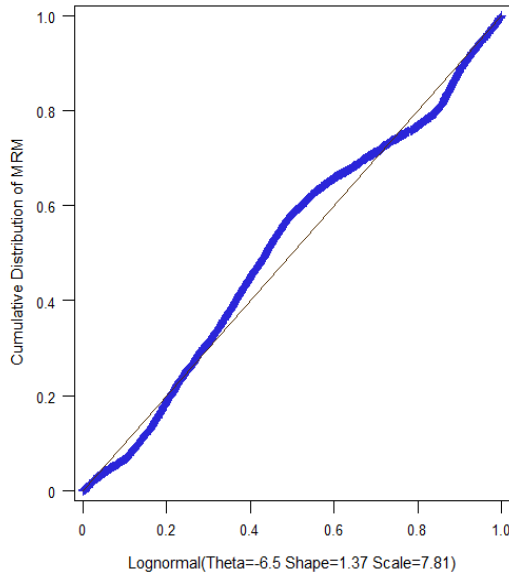


Figure 23 : PP-plot de l'ajustement à la loi Log-Normale

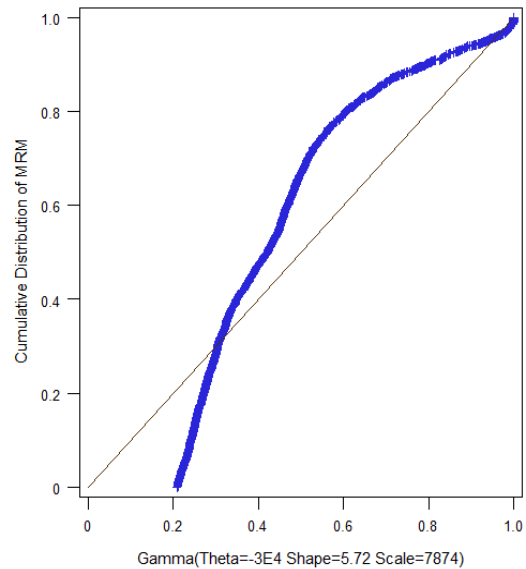


Figure 24 : PP-plot de l'ajustement à la loi Gamma

L'alignement du PP-PLOT du modèle Log-normale est un peu meilleur que celui du modèle Gamma, donc nos données s'ajustent le plus à une loi Log-Normale.

Pour confirmer cette hypothèse, nous allons comparer entre ces deux modèles grâce au critère **AIC** « **Akaike Information Criterion** ». En effet, l'AIC permet de faire un choix entre plusieurs modèles, son principe repose sur le fait que : plus la vraisemblance du modèle est grande, plus la log-vraisemblance est grande et par conséquent le modèle est le meilleur.

On écrit :

$$AIC = -2\ln(L) + 2p$$

Où L désigne la vraisemblance maximisée du modèle et p le nombre de paramètre à estimer.

Le modèle qui minimise l'AIC est considéré comme étant le meilleur.

L'AIC relatif au modèle Gamma est égale à **85438.6559** alors que celui relatif au modèle Log-normale est de **12439.6203**, donc la loi Log-normale permet une meilleure modélisation du MRM que la loi Gamma.

Pour la suite de l'étude, on retient comme loi du MRM, la loi Log-Normale qui propose des résultats meilleurs aux tests d'adéquation.

6.2. L'analyse des effets liés aux variables explicatives retenues:

Dans le but de déterminer maintenant quels sont réellement les facteurs influant sur les montants remboursés moyens, on effectue des **tests du Rapport de Vraisemblance**.

On rappelle brièvement dans un premier temps le principe du test du rapport de vraisemblance.

Le test du rapport de vraisemblance permet de comparer un modèle à un modèle réduit, dans le sens où il comportera moins de variables. Il s'appuie pour cela sur le rapport de vraisemblance, et donc sur l'effet que peut avoir l'omission d'une variable sur la vraisemblance du modèle.

On utilise la statistique du rapport des vraisemblances suivante, pour un test sur la variable V :

$$R = \frac{\text{Vraisemblance du modèle sans la variable } V}{\text{Vraisemblance du modèle avec la variable } V}$$

Avec les hypothèses :

$$\begin{cases} H_0: \text{La variable } V \text{ n'est pas influente dans le modèle} \\ H_1: \text{La variable } V \text{ est influente dans le modèle} \end{cases}$$

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

$$n = \frac{\text{dimension(modalités du modèle avec la variable } V)}{\text{dimension(modalités du modèle sans la variable } V)}$$

Les tests sont effectués à l'aide du logiciel SAS, pour le calcul des vraisemblances. On obtient le tableau suivant :

Statistiques LR pour Analyse de Type 3			
Source	DDL	Khi-2	Pr > Khi-2
type_ass	3	277.54	<.0001
type_benef	2	1044.59	<.0001
sexe	1	244.42	<.0001
ald	1	7284	<.0001
tranche_age	2	54.69	<.0001
region	1	559.75	<.0001

Tableau 27 : Résultat des tests du rapport de vraisemblance

Les résultats de ces tests nous montrent que sur les variables testées (le type de l'assuré, le type du bénéficiaire, le sexe, ALD, la tranche d'âge et la région), nous ne pouvons négliger l'influence d'aucune d'entre elles.

On conserve donc l'ensemble des variables explicatives pour la suite de l'étude.

6.3. L'estimation des paramètres:

On estime dans cette partie les paramètres de la loi Log-Normale. Une variable est dite Log-Normale si son logarithme suit une loi Normale. Elle présente l'avantage d'être positive, donc adaptée à la modélisation de coût, et permet d'ajuster des phénomènes asymétriques.

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de		Khi-2 de Wald	Pr > Khi-2
Intercept		1	8.2294	0.1289	7.9769	8.482	4078.48	<.0001
type_ass	A	1	0.2154	0.1258	-0.0311	0.4619	2.93	0.0868
type_ass	P1	1	0.3873	0.126	0.1403	0.6343	9.44	0.0021
type_ass	P2	1	0.2417	0.1271	-0.0074	0.4908	3.62	0.0572
type_ass	P3	0	0	0	0	0	.	.
type_benef	A	1	0.9011	0.0269	0.8485	0.9538	1124.04	<.0001
type_benef	CA	1	0.7297	0.0266	0.6775	0.7819	750.78	<.0001
type_benef	EA	0	0	0	0	0	.	.
sexe	F	1	0.1424	0.009	0.1248	0.1601	251.16	<.0001
sexe	M	0	0	0	0	0	.	.
ald	N	1	-1.8217	0.0135	-1.8482	-1.7952	18159.4	<.0001
ald	O	0	0	0	0	0	.	.
tranche_age	majo	1	0.0754	0.0107	0.0544	0.0964	49.5	<.0001
tranche_age	mino	1	0.0901	0.0262	0.0387	0.1415	11.81	0.0006
tranche_age	neut	0	0	0	0	0	.	.
region	reg_m	1	-0.1925	0.0079	-0.2079	-0.177	595.95	<.0001
region	reg_n	0	0	0	0	0	.	.
Scale		1	4.2277	0.0445	4.1413	4.3158		

Tableau 28 : Les paramètres estimés de loi Log-Normale

Ce tableau donne les estimations des paramètres du modèle selon le paramétrage SAS et teste la nullité de chaque paramètre à l'aide de la statistique du test de WALD, sachant que les autres variables explicatives sont dans le modèle. Il ne donne pas les résultats des tests sur l'effet de chaque composante du modèle.

Les paramètres estimés correspondent à l'écart entre un niveau et le niveau qui sert de référence (celui dont le paramètre estimé est nul selon le paramétrage SAS).

Par exemple, le paramètre associé à ald=N et qui vaut -1.8217 mesure l'écart entre les bénéficiaire non atteints d'une affection de longue durée (ald=N) et ceux qui sont atteints (ald=O, pris pour référence selon le paramétrage SAS).

6.4. La validation du modèle retenu :

La validation est une étape importante de l'établissement d'un modèle. Elle permet en effet d'évaluer la qualité de l'ajustement qui a été fait entre les moyennes μ_i et les observations faites sur le jeu de données à notre disposition, et d'améliorer le modèle initial.

Pour cela, la théorie des modèles linéaires généralisés nous fournit plusieurs statistiques dont l'analyse peut nous donner des éléments de validation. Même si

l'analyse de ces statistiques ne nous procure pas une assurance de la validité du modèle, elle nous permet, toute au moins, de ne pas le rejeter.

Avant d'étudier les résidus de la déviance obtenus sur le modèle, on définit le concept de déviance.

6.4.1. La déviance

La déviance est un indicateur basé sur la vraisemblance du modèle, qui varie, une fois la densité choisie, selon le nombre de paramètre du modèle.

On considère qu'une description parfaite des données est possible lorsqu'il y a autant de paramètres que de données, les moyennes calculées étant ainsi strictement égales aux données. Cette description n'est pas intéressante puisqu'elle ne résume pas les données, cependant la vraisemblance du modèle ainsi paramétré peut nous servir de référence en vue d'une comparaison avec la vraisemblance réellement obtenue.

On note la vraisemblance du modèle avec autant de paramètres que d'observations

$L(y/y)$ est la vraisemblance du modèle réellement obtenu $L(\hat{\mu}/y)$. L'idée de la déviance est alors de s'intéresser au rapport de vraisemblance :

$$\Lambda = \frac{L(y/y)}{L(\hat{\mu}/y)}$$

Un rapport proche de 1 indique une bonne description des données. Au contraire, plus ce rapport s'éloignera de 1, moins bonne sera la description.

En pratique, on utilisera plutôt l'équation similaire :

$$\ln \Lambda = \ln [L(y/y)] - \ln [L(\hat{\mu}/y)]$$

Et par la statistique suivante, appelée déviance réduite :

$$D = 2 \ln \Lambda$$

Une valeur faible pour la statistique D indique ainsi une bonne description des données, alors qu'une valeur élevée supposera une description de mauvaise qualité.

6.4.2. L'analyse des résidus

L'analyse des résidus permet une analyse plus poussée. Elle permet en effet de comprendre d'où proviennent les éventuels écarts entre les valeurs prédites μ_i 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** ».

En pratique ces deux approches conduisent à des résultats peu différents et, dans le cas contraire, c'est une indication de mauvaise approximation de la loi asymptotique.

Dans notre cas, nous utilisons « les résidus de la déviance » pour juger l'ajustement de notre modèle à la loi asymptotique.

Pour cela, on considère que chaque observation y_i apporte sa contribution c_i à la déviance, de sorte que la somme des c_i nous fournisse la déviance :

$$D = \sum_{i=1}^n c_i$$

On définit alors les résidus de déviance r_i comme la racine carrée de la contribution c_i , auquel on affecte le signe du résidu brut $y_i - \mu_i$:

$$r_i = \text{signe}(y_i - \mu_i)\sqrt{c_i}$$

On a ainsi que :

$$D = \sum_{i=1}^n r_i^2$$

On interprète les résultats sur un graphique en plaçant en ordonnée les observations des résidus de la déviance r_i et en abscisses les valeurs estimées $\widehat{\mu}_i$.

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.

Le graphique des résidus obtenu est alors le suivant :

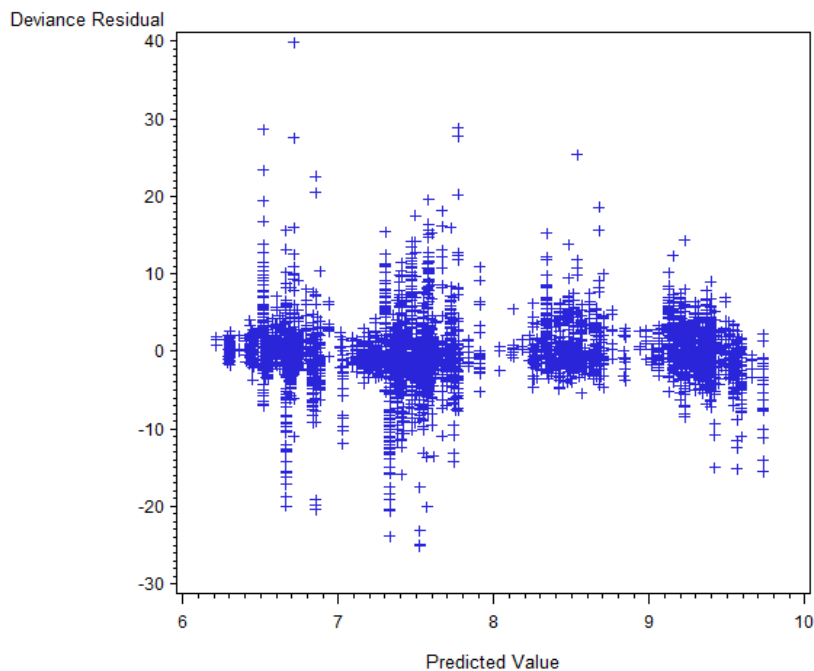


Figure 25 : Représentation des résidus de la déviance en fonction des valeurs prédites

Les résidus sont correctement répartis sur l'axe des abscisses et sont d'une manière générale relativement proches de 0. On n'observe pas de valeurs éloignées de 0 pour les résidus de la déviance ce qui peut traduire une certaine justesse du modèle.

7. La modélisation de la fréquence moyenne des sinistres

La fréquence de survenance d'un sinistre peut être considérée comme étant une variable aléatoire à valeurs entières. Pour rappel, le taux de sinistralité s'obtient en divisant l'effectif sinistré de chaque classe de consommation par l'effectif qui lui est associé.

L'objectif de cette partie est d'étudier la relation entre le nombre de sinistres (variable endogène) et les variables descriptives (variables exogènes)

La variable endogène est une variable de comptage. Les modèles linéaires généralisés permettent l'étude de telles variables. En effet, la loi de probabilité de Poisson et la loi binomiale négative sont des lois adaptées aux phénomènes de comptage et appartiennent à la famille exponentielle.

Nous allons, dans ce qui suit, appliquer ces deux modèles linéaires généralisés à l'effectif sinistré.

Remarque : Les modèles linéaires généralisés adaptés aux données de comptage introduisent la notion de variable d'*offset* ou d'*exposition*. L'introduction d'une telle variable revient à rapporter la variable endogène sur l'offset.

Dans notre cas nous allons considérer l'effectif de chaque segment comme offset. Nous modélisons donc en réalité le nombre de sinistre par personne.

7.1. Ajustement des fréquences par une loi de Poisson

On teste dans un premier temps l'ajustement à la loi de Poisson, loi la plus classiquement utilisée pour modéliser les fréquences de sinistre.

On procède tout d'abord à un ajustement graphique de la loi poisson à l'aide de la fonction **goodfit** du logiciel R comme suit :

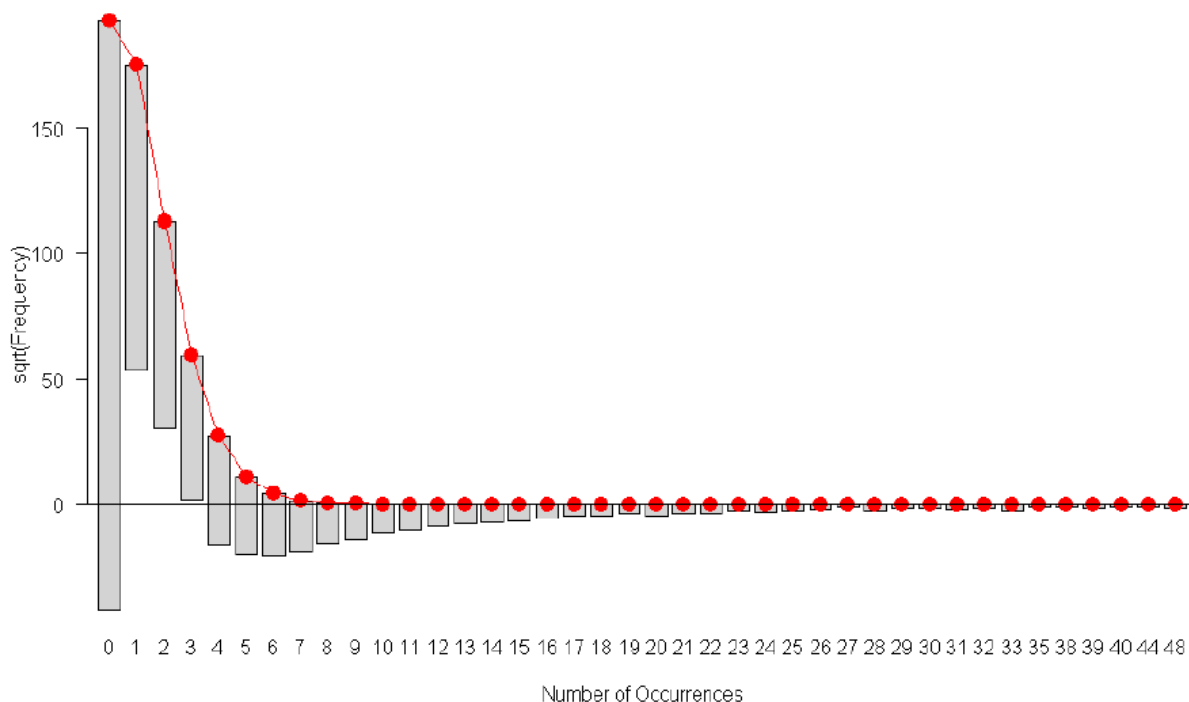


Figure 26 : L’Ajustement à la loi de Poisson

On interprète le graphique de la manière suivante : 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.

On remarque ainsi que l’ajustement par la loi de Poisson est peu satisfaisant. En effet les observés sont conséquents et cela quel que soit le nombre d’occurrences.

En calculant la moyenne empirique $\hat{\mu}$ et la variance empirique $\hat{\sigma}^2$ on trouve les valeurs suivantes :

$$\hat{\mu} = 279,41 \quad ; \quad \hat{\sigma}^2 = 551018,89$$

Empiriquement, on observe une moyenne bien inférieure à la variance. Cela peut-être signe de la sur-dispersion.

Passons maintenant à l’estimation des paramètres de notre modèle, la sortie de l’ajustement est la suivante :

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de		Khi-2 de Wald	Pr > Khi-2
Intercept		1	-0.8662	0.0306	-0.9262	-0.8062	801.6	<.0001
type_ass	A	1	0.4141	0.0298	0.3558	0.4724	193.73	<.0001
type_ass	P1	1	0.3262	0.0298	0.2678	0.3846	119.79	<.0001
type_ass	P2	1	0.2292	0.03	0.1704	0.2881	58.22	<.0001
type_ass	P3	0	0	0	0	0	.	.
type_benef	A	1	0.4082	0.0064	0.3956	0.4207	4043.19	<.0001
type_benef	CA	1	0.34	0.0064	0.3274	0.3526	2795.92	<.0001
type_benef	EA	0	0	0	0	0	.	.
sexe	F	1	0.2453	0.0021	0.2411	0.2495	13202.9	<.0001
sexe	M	0	0	0	0	0	.	.
ald	N	1	-0.7136	0.0032	-0.7199	-0.7073	49304	<.0001
ald	O	0	0	0	0	0	.	.
tranche_age	majo	1	-0.012	0.0025	-0.0169	-0.0072	23.63	<.0001
tranche_age	mino	1	-0.1793	0.0063	-0.1917	-0.167	810.81	<.0001
tranche_age	neut	0	0	0	0	0	.	.
region	reg_m	1	-0.0639	0.0019	-0.0676	-0.0603	1170.55	<.0001
region	reg_n	0	0	0	0	0	.	.
Scale		0	1	0	1	1		

Tableau 29 : Les paramètres estimés de la loi Poisson

On constate très bien que les résultats des tests de WALD sont tous significatifs, nous devons vérifier que notre modèle ne présente pas de sur-dispersion.

La sur-dispersion est un phénomène qui concerne la modélisation de données selon une loi binomiale ou selon une loi de Poisson, il se traduit par le fait d'avoir une variance observée supérieure à celle estimée par le modèle. Ce phénomène provient du fait que la loi de Poisson n'a qu'un seul paramètre, ce qui ne permet pas d'ajuster la variance indépendamment de la moyenne.

On diagnostique une sur-dispersion lorsque la déviance normalisée ou le khi-deux de Pearson normalisé sont nettement supérieurs à 1.

La sur-dispersion n'intervient pas au niveau de l'estimation de β . En revanche, ce phénomène a pour effet de rendre les résultats des tests de Wald et des tests basés sur le rapport de vraisemblance trop significatifs.

Pour savoir si on est effectivement en présence de ce phénomène, on analysera la déviance normalisée ou le khi-deux de Pearson normalisé :

Critère d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Deviance	4503	159412.2012	35.4013
Scaled Deviance	4503	159412.2012	35.4013
Pearson Chi-Square	4503	126914.9395	28.1845
Scaled Pearson X2	4503	126914.9395	28.1845
Log Likelihood		7790742.782	
Full Log Likelihood		-90658.8752	
AIC (smaller is better)		181339.7504	
AICC (smaller is better)		181339.809	
BIC (smaller is better)		181410.3147	

Tableau 30 : Critère d'évaluation de l'adéquation à la loi de poisson

On remarque bien que la déviance normalisée (scaled deviance) est plus de 35 fois plus élevée que le nombre de degrés de libertés. De même, le khi-deux de Pearson normalisé (scaled pearson x2) rapporté sur le nombre de degrés de libertés est largement supérieur à 1 ($28.1845 \gg 1$). **On est alors en présence du phénomène de la sur-dispersion.**

7.2. Ajustement des fréquences des sinistres par une loi Binomiale-négative

L'ajustement n'étant pas satisfaisant par la loi de Poisson, on s'intéresse à la loi Binomiale-Négative.

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 cette observation n'est pas satisfaite.

On obtient le résultat suivant :

Critère d'évaluation de l'adéquation			
Critère	DDL	Valeur	Valeur/DDL
Deviance	4503	4080.0439	0.9061
Scaled Deviance	4503	4080.0439	0.9061
Pearson Chi-Square	4503	3609.3473	0.8015
Scaled Pearson X2	4503	3609.3473	0.8015
Log Likelihood		7864044.551	
Full Log Likelihood		-17357.106	
AIC (smaller is better)		34738.2119	
AICC (smaller is better)		34738.2812	
BIC (smaller is better)		34815.1912	

Tableau 31 : Critère d'évaluation de l'adéquation à la loi binomiale-négative

On constate que la déviance normalisée est légèrement inférieure au nombre de degrés de liberté ($0.9061 \approx 1$) ce qui indique un bien meilleur ajustement qu'avec la régression de Poisson.

C'est donc la loi Binomiale-Négative que l'on retiendra dans la suite pour la modélisation des fréquences des sinistres.

Après avoir identifié la loi adéquate qui modélise le mieux la fréquence des sinistres, nous allons passer à l'estimation des coefficients de notre modèle comme suit :

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de		Khi-2 de Wald	Pr > Khi-2
Intercept		1	-0.3674	0.0516	-0.4686	-0.2662	50.6	<.0001
type_ass	A	1	0.2382	0.0439	0.1522	0.3241	29.5	<.0001
type_ass	P1	1	0.0983	0.0443	0.0115	0.1852	4.92	0.0265
type_ass	P2	1	0.0585	0.0462	-0.0321	0.149	1.6	0.2058
type_ass	P3	0	0	0	0	0	.	.
type_benef	A	1	0.1598	0.0296	0.1017	0.2178	29.05	<.0001
type_benef	CA	1	0.15	0.0321	0.087	0.213	21.76	<.0001
type_benef	EA	0	0	0	0	0	.	.
sexe	F	1	0.1567	0.0176	0.1223	0.1911	79.6	<.0001
sexe	M	0	0	0	0	0	.	.
ald	N	1	-0.8613	0.017	-0.8947	-0.8279	2554.01	<.0001
ald	O	0	0	0	0	0	.	.
tranche_age	majo	1	0.018	0.0211	-0.0234	0.0594	0.73	0.3944
tranche_age	mino	1	-0.2555	0.0286	-0.3115	-0.1995	79.97	<.0001
tranche_age	neut	0	0	0	0	0	.	.
region	reg_m	1	-0.0187	0.0189	-0.0557	0.0182	0.99	0.3207
region	reg_n	0	0	0	0	0	.	.
Dispersion		1	0.175	0.0049	0.1653	0.1846		

Tableau 32 : Les paramètres estimés de la loi Binomiale-Négative

On constate que les coefficients relatifs aux modalités P2 (conjoint survivant) de la variable type_ass, majo (âge > 45 ans) de la variable tranche_age et reg_m de la variable region sont non significatifs, cela veut dire que l'écart de chacune avec la référence, en termes de moyennes de la variable endogène en question, est inexistant. Autrement dit, les moyennes de la fréquence des sinistres pour P2 et P3 sont tellement proches qu'elles sont statistiquement indiscernables. De même pour majo et neut ainsi que pour reg_m et reg_n.

Sachant que les modalités P3 (enfant survivant), EA (enfant de l'assuré), M (homme), O (avoir une ald), neut (âge entre 25 ans et 45 ans) et reg_n sont choisies comme modalités de référence selon le paramétrage SAS.

Dans ce cas, si un tel regroupement a un sens, on peut fusionner les deux modalités P2 et P3 en une seule, de même pour majo et neut, ainsi que pour reg_m et reg_n. Et refaire ensuite le modèle. Le risque c'est de sauter aux conclusions à la seule vue du tableau des coefficients, voir P2 et P3 ne sont pas significativement différents, alors qu'en réalité, l'écart entre P2 et P1, à titre d'exemple, est encore moins significatif, c'est-à-dire ce sont ces deux dernières modalités qui devraient être fusionnées en priorité. Mais puisque c'est P3 qu'est la référence, rien dans le tableau de coefficients n'indique la significativité de l'écart entre P2 et P1. Il faudrait alors choisir une de ces 2 modalités comme référence.

Le plus simple dans ce cas est d'utiliser l'instruction LSMEANS et son option DIFF du logiciel SAS. On aura ainsi toutes les comparaisons 2 à 2 de modalités. On fusionne les deux dont la p-value sera la plus élevée, **à condition que cette fusion ait un sens**. Et on refait le modèle, et ainsi de suite, jusqu'à ce que toutes les p-values soient, soit en-dessous du seuil choisi à l'avance (5%), soit associées à des couples non fusionnables.

On obtient le tableau suivant :

Différences des moyennes des moindres carrés							
Effet			Valeur estimée	Erreur type	DDL	Khi-2	Pr > Khi-2
type_ass	A	P1	0.1398	0.0189	1	54.68	<.0001
type_ass	A	P2	0.1797	0.0246	1	53.39	<.0001
type_ass	A	P3	0.2382	0.0439	1	29.5	<.0001
type_ass	P1	P2	0.0399	0.0248	1	2.57	0.1087
type_ass	P1	P3	0.0983	0.0443	1	4.92	0.0265
type_ass	P2	P3	0.0585	0.0462	1	1.6	0.2058
type_benef	A	CA	0.0098	0.0221	1	0.2	0.657
type_benef	A	EA	0.1598	0.0296	1	29.05	<.0001
type_benef	CA	EA	0.15	0.0321	1	21.76	<.0001
sexe	F	M	0.1567	0.0176	1	79.6	<.0001
ald	N	O	-0.8613	0.017	1	2554	<.0001
tranche_age	majo	mino	0.2735	0.0319	1	73.41	<.0001
tranche_age	majo	neut	0.018	0.0211	1	0.73	0.3944
tranche_age	mino	neut	-0.2555	0.0286	1	79.97	<.0001
region	reg_m	reg_n	-0.0187	0.0189	1	0.99	0.3207

Tableau 33 : Comparaisons 2 à 2 des modalités en termes de moyenne de la fréquence des sinistres

Les résultats obtenus montrent que les moyennes de la fréquence des sinistres sont statistiquement indiscernables pour chacune des combinaisons de modalités suivantes :

- P1 et P2
- P2 et P3
- A et CA
- majo et neut
- reg_m et reg_n

Donc on procède comme suit :

- On fusionne dans un premier temps les deux modalités P2 et P3 puisque leur p-value est plus grande que celle de P1 et P2.
- On fusionne reg_m et reg_n. Ceci est équivalent à exclure la variable région de l'analyse en conservant des coefficients à 100% pour chacune de ces modalités. D'ailleurs bien que cette variable n'a pas un effet influant sur la variable endogène selon le résultat du test du rapport de vraisemblance, on ne peut pas l'éliminer car elle est retenue dans le modèle ajustant le MRM. En effet, par cohérence un facteur devra être conservé à la fois pour les fréquences et les coûts, ou pour aucune de ces deux variables.
- On fusionne les modalités majo et neut. On obtient ainsi de nouvelles tranches d'âges : « plus25 », elle correspond aux bénéficiaires ayant plus que 25 ans, c'est la fusion entre majo et neut. « moins25 », elle correspond aux bénéficiaires ayant moins que 25ans.
- On décide de ne pas fusionner les modalités A (assuré) et CA (conjoint de l'assuré), puisque le nouveau groupe engendré n'a pas de sens.

Une fois qu'on a fusionné les modalités dont les moyennes de la variable endogène sont proches, on estime à nouveau notre modèle retenu dans le cas de la loi binomiale négative :

Analyse des valeurs estimées du paramètre de vraisemblance maximum								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de		Khi-2 de Wald	Pr > Khi-2
Intercept		1	-0.3243	0.0327	-0.3884	-0.2602	98.41	<.0001
type_ass	A	1	0.1866	0.0225	0.1424	0.2308	68.5	<.0001
type_ass	P1	1	0.0503	0.0233	0.0047	0.0959	4.67	0.0307
type_ass	survivants	0	0	0	0	0	.	.
type_benef	A	1	0.1665	0.0279	0.1119	0.2212	35.68	<.0001
type_benef	CA	1	0.155	0.0309	0.0944	0.2156	25.09	<.0001
type_benef	EA	0	0	0	0	0	.	.
sexe	F	1	0.1568	0.0175	0.1225	0.1912	80.21	<.0001
sexe	M	0	0	0	0	0	.	.
ald	N	1	-0.8638	0.017	-0.897	-0.8305	2595.63	<.0001
ald	O	0	0	0	0	0	.	.
tranche_age	moins25	1	-0.2634	0.0279	-0.3182	-0.2087	89	<.0001
tranche_age	plus25	0	0	0	0	0	.	.
region	reg	0	0	0	0	0	.	.
Dispersion		1	0.1753	0.0049	0.1656	0.1849		

Tableau 34 : Les paramètres estimés de la loi Binomiale-Négative

On constate en conséquence que toutes les modalités sont significatives.

Conclusion

L'utilisation des modèles linéaires généralisés appliqués à l'assurance maladie nous a permis de formuler des hypothèses sur la structure paramétrique du modèle et sur la loi de distribution de la variable à modéliser. De plus, ils nous ont permis de mettre en évidence les variables réellement influentes sur la consommation et de quantifier cette influence.

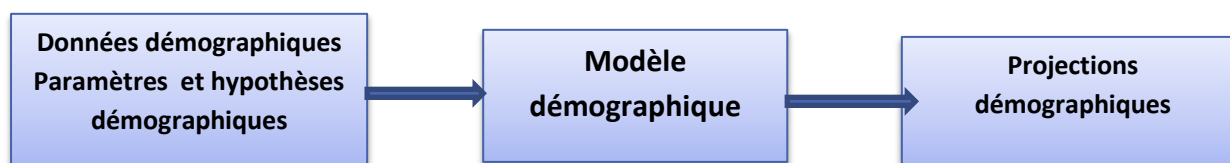
L'approche « fréquence/coût moyen » nous a permis de modéliser les montants remboursés moyens durant une période considérée, cependant pour estimer les montants remboursés futurs par le régime on devra projeter la population. C'est dans ce contexte que se situe le chapitre suivant.

Chapitre 4 : Modélisation de la population par la loi « entrée-sortie »

Les projections démographiques permettent de prévoir les tendances futures que peut connaître une population dans le temps.

On désigne par les flux démographiques les entrées ou les sorties que peut connaître le régime à savoir le décès, l'invalidité, les entrées au régime, les départs en retraite, le départ du régime, l'adhésion de nouveaux conjoints, les nouveaux nés...

Le but de cette partie est de décrire et de mettre en œuvre un modèle démographique pour modéliser les flux démographiques qui serviront aux projections de la population CNOPS.



Tout d'abord, on commence par présenter les paramètres et les notations qu'on va adopter.

1. Paramètres et notations

On distingue entre les différentes sous-populations suivantes :

- Actifs (A, A)
- Retraités (vieillesse et invalidité) (R, A)
- Conjoints d'actifs (A, C)
- Conjoints de retraités (R, C)
- Enfants d'actifs (A, E)
- Enfants de retraités (R, E)
- Veufs (ves) (V, A)

Pour chacune de ces sous-populations, nous disposons des effectifs selon les variables suivantes :

- Age, notée x
- Sexe, notée S
- ALD (1 : atteint d'ALD, 0 : non atteint)
- Région de résidence, notée rr

Les effectifs seront notés E_n . En guise d'exemple, l'effectif des assurés, actifs, de sexe masculin, d'âge 40, résidant dans la région 05 et non ALD (ALD = 0) sera noté :

$$E_n(A, A, M, 40, 05, 0)$$

Remarque : Un signe «.» est marqué à la place d'une variable sur laquelle on veut appliquer une sommation. Par exemple, l'effectif des assurés retraités d'âge 68 et non ALD est noté :

$$E_n(A, R, ., 68, ., 0)$$

Après avoir défini les différentes sous-populations, on présente ci-dessous les différents paramètres qui serviront à la modélisation des différents flux démographiques.

Paramètres relatifs aux Actifs :

- $adhes(n)$: nombre de nouvelles adhésions de l'année n, estimé à partir de l'effectif initial moyennant un taux moyen d'entrée TME tel que :

$$adhes(n) = TME * \sum E_n(A, A, *, *, *, *)$$

- $TR(type_ass, x, Sexe, rr)$: Taux de répartition des nouvelles recrues par âge, sexe et région de résidence.
- $TS(x)$: taux de sortie par âge des actifs pour une raison autre que le décès ou l'invalidité.
- r : âge de la retraite.
- min : âge minimum des actifs.

Paramètres relatifs aux ALD :

- $poidsALD(x)$: Poids des ALD par âge, c'est-à-dire la proportion d'ALD pour un âge donné.
- $TAGpoids_ALD$: taux d'aggravation de $poids_ALD$.
- $Tinv$: probabilité de tomber invalide avec un taux d'invalidité ne permettant plus de pouvoir continuer à travailler, mais plutôt de bénéficier d'une pension de retraite pour motif d'invalidité.

Paramètres matrimoniaux :

- $Tac(Type_ASS, S)$: taux d'adhésion des conjoints par sexe, il sert au calcul des nouveaux conjoints issus des mariages des adhérents.
- d : différence d'âge moyenne entre un homme et sa femme.

Paramètres relatifs aux enfants :

- $PSe(x)$: Probabilité de scolarisation des enfants à l'âge $x > 21$.
- $Tfec$: Taux de fécondité.
- $PE(S)$: Poids des enfants par sexe.
- Nme : Nombre moyen d'enfants par assuré.
- e : Différence d'âge moyenne entre l'assuré et un enfant à sa charge.

Autres paramètres :

- $q(x)$: Probabilité de décéder entre l'âge x et $x+1$.
- $D_n(Type_ass, Type_Ben, s, x, rr)$: Décès de l'année n .
- $INV_n(inv, A, S, x, rr)$: Nouveaux invalides de l'année n selon le sexe, l'âge et la région de résidence.
- $N_n(Type_ass, Type_Ben, s, x, rr)$: Nouvelles adhésions des assurés selon le type d'assuré, le type du bénéficiaire, le sexe, l'âge et la région de résidence.
- $S_n(A, A, s, x, rr)$: Nouvelles sorties des actifs pour une raison autre que le décès ou l'invalidité.
- $NV_n(A, C, s, x, rr)$: Nouveaux veufs.
- $ENR_n(A, E, *, x, *)$: représente les enfants des nouveaux retraités au cours de l'année $n - 1$.

Équations préliminaires :

Les équations suivantes sont des équations préliminaires qui définissent les liens entre quelques paramètres énoncés ci-dessus.

Remarque : Le signe « * » sera marqué à la place d'une variable lorsqu'une formule est valable indépendamment de la modalité prise par cette variable.

En ce qui concerne les décès de chaque année on retient la formule suivante :

$$D_n(*,*,*,x,*) = q(x) * E_n(*,*,*,x,*,.)$$

Les probabilités de décès proviendront de la table de mortalité TV 88-90.

Cette table présente une espérance de vie de 106 ans, ce qui permettra de prendre une position de prudence face à la vieillesse de la population.

Les nouvelles adhésions d'assurés par catégorie seront déduites à partir de l'adhésion globale de chaque année moyennant le taux de répartition des nouvelles recrues selon la formule suivante :

$$N_n(*,A,*,x,*) = Adhes(n) * TR(*,x,*,*)$$

Finalement les sorties autres qu'en cas de décès ou d'invalidité seront calculées moyennant le taux de sortie pour autre cause que le décès comme suit :

$$S_n(A,A,*,x,*,*) = TS * E_n(A,A,*,x,*,*)$$

2. Mise en équation des flux démographiques

Dans cette partie, nous allons définir les équations qui permettent d'obtenir les effectifs selon les variables d'intérêt, d'une année n à partir de ceux de l'année $n - 1$. Ces équations sont basées sur un bilan des entrées et sorties et seront catégorisées par type de bénéficiaire.

2.1. Assurés actifs

Pour $min \leq x < r$, les effectifs des actifs de l'année n s'écrivent :

$$\begin{aligned} E_n(A,A,*,x,*,.) &= E_{n-1}(A,*,x-1,*,.) + N_{n-1}(A,A,*,x-1,*) - \\ &N_{inv_{n-1}}(Inv,A,*,x-1,*) - D_{n-1}(A,A,*,x-1,*) - \\ &S_{n-1}(A,A,*,x-1,*) \end{aligned}$$

2.2. Retraités

S'agissant des retraités, il y'a lieu de distinguer entre les anciens retraités $x > r$, et les nouveaux $x = r$.

Pour $x > r$:

$$E_n(R, A, *, x, *, .) = E_{n-1}(R, A, *, x - 1, *, .) + N_{n-1}(R, A, *, x - 1, *) - D_{n-1}(R, A, *, x - 1, *) + N_{inv_{n-1}}(Inv, A, *, x - 1, *)$$

Pour $x = r$:

$$E_n(R, A, *, x, *, .) = E_{n-1}(A, A, *, x - 1, *, .) + E_{n-1}(R, A, *, x - 1, .) + N_{n-1}(A, A, *, x - 1, *) - D_{n-1}(R, A, *, x - 1, *)$$

Dans l'équation ci-dessus, le terme « $E_n(R, A, *, x - 1, *, .)$ » représente les gens ayant bénéficié d'une retraite anticipée.

2.3. Conjoints

On s'intéresse ici aux conjoints des gens étant affilié à l'AMO-CNOPS.

L'effectif des conjoints d'actifs est le suivant :

$$E_n(A, C, *, x, *, .) = E_{n-1}(A, C, *, x - 1, *, .) + N_{n-1}(A, C, *, x - 1, *) - CNR_{n-1}(A, C, *, x - 1, *) - D_{n-1}(A, C, *, x - 1, *) - CNInv_{n-1}(A, C, *, x - 1, *) - S_{n-1}(A, C, *, x - 1, *) - NV_{n-1}(A, C, *, -1, *)$$

En ce qui concerne les conjoints des retraités :

$$E_n(R, C, *, x, *, .) = E_{n-1}(R, C, *, x, *, .) + N_{n-1}(R, C, *, x - 1, *) - D_{n-1}(R, C, *, x - 1, *) + CNInv_{n-1}(A, C, *, x - 1, *) + CNR_{n-1}(A, C, *, x - 1, *) - NV_{n-1}(R, C, *, x - 1, *)$$

Le terme « $N_{n-1}(*, C, *, x - 1, *)$ » représente les nouvelles adhésions des conjoints, d'âge $x - 1$ au cours de l'année $n - 1$. Ce qui englobe les mariages des adhérents au régime.

Cette grandeur peut se calculer comme suit :

- Conjoint de sexe féminin :

$$N_{n-1}(*, C, F, x - 1, *) = Tac(*, F) * E_{n-1}(*, A, H, x + d, *)$$

- Conjoint de sexe Masculin :

$$N_{n-1}(*, C, H, x - 1, *) = Tac(*, H) * E_{n-1}(*, A, F, x - d, *)$$

Le terme « $NV_{n-1}(*, C, *, x - 1, *)$ » représente les nouveaux veufs (ves) d'adhérents de l'année $n - 1$.

Pour calculer cet effectif, nous définissons le poids des conjoints à charge selon le sexe comme suit :

$$P_Conj_n(*, F, x, *) = \frac{E_n(*, C, F, x, *, .)}{E_n(*, A, H, x + d, *, .)}$$

et

$$P_Conj_n(*, H, x, *) = \frac{E_n(*, C, H, x, *, .)}{E_n(*, A, F, x - d, *, .)}$$

Donc nous avons :

$$NV_{n-1}(*, C, F, x, *) = P_Conj_n(*, F, x, *) * D_{n-1}(*, A, H, x + d, *)$$

et

$$NV_{n-1}(*, C, H, x, *) = P_Conj_n(*, H, x, *) * D_{n-1}(*, A, F, x - d, *)$$

Nous allons calculer de la même manière les « $S_{n-1}(A, C, *, x - 1, *)$ », qui représente le départ des conjoints suite au départ de l'adhérent pour autre raison que le décès et l'invalidité, selon les deux équations suivantes :

$$S_n(A, C, F, x, *) = P_Conj_n(*, F, x, *) * S_n(A, A, H, x + d, *)$$

et

$$S_n(A, C, H, x, *) = P_Conj_n(*, H, x, *) * S_n(A, A, F, x - d, *)$$

Le terme « $CNR_{n-1}(*, C, *, x - 1, *)$ » représente les conjoint des adhérents sortant à la retraite à l'année n . On le calcule, selon le sexe comme suit :

$$CNR_{n-1}(*, C, M, x - 1, *) = \begin{cases} E_{n-1}(A, C, M, x - 1, *, .) & \text{si } x - 1 - d + 1 = r \\ 0 & \text{sinon} \end{cases}$$

et

$$CNR_{n-1}(*, C, F, x - 1, *) = \begin{cases} E_{n-1}(A, C, F, x - 1, *, .) & \text{si } x - 1 + d + 1 = r \\ 0 & \text{sinon} \end{cases}$$

Finalement, le terme « $CNInv_{n-1}(*, C, *, x - 1, *)$ » représente les conjoints des nouveaux invalides et se calcule selon le sexe comme suit :

$$CNInv_n(A, C, F, x, *) = P_Conj_n(*, F, x, *) * Ninv_n(A, A, H, x + d, *)$$

et

$$CNInv_n(A, C, H, x, *) = P_Conj_n(*, H, x, *) * Ninv_n(A, A, F, x - d, *)$$

2.4. Les veufs (ves)

Pour une année donnée n , l'effectif des veufs selon les variables d'intérêt est le suivant:

$$E_n(V, A, *, x, *, .) = E_{n-1}(V, A, *, x - 1, *, .) - D_{n-1}(V, A, *, x - 1, *) + \\ NV_{n-1}(A, C, *, x - 1, *) + NV_{n-1}(R, C, *, x - 1, *)$$

2.5. Les enfants

En ce qui concerne les enfants des adhérents, on distingue entre les enfants d'actifs et ceux des retraités. Selon ces deux catégories, on calcule l'effectif des enfants comme suit :

Pour $x > 0$:

$$E_n(A, E, *, x, *, .) = PSe(x) * [E_{n-1}(A, E, *, x - 1, *, .) - D_{n-1}(A, E, *, x - 1, *) - ENR_{n-1}(A, E, *, x - 1, *) - ENInv_{n-1}(A, E, *, x - 1, *) - S_{n-1}(A, E, *, x - 1, *) - NO_{n-1}(A, E, *, x - 1, *)]$$

et

$$E_n(R, E, *, x, *, .) = PSe(x) * [E_{n-1}(R, E, *, x - 1, *, .) - D_{n-1}(R, E, *, x - 1, *) + ENR_{n-1}(A, E, *, x - 1, *) + ENInv_{n-1}(A, E, *, x - 1, *) - NO_{n-1}(R, E, *, x - 1, *)]$$

Si $x = 0$:

$$E_n(*, E, s, x, *, .) = Tfec * PE(s) * \sum_{k=15}^{49} [E_{n-1}(*, A, F, k, *, .) + E_{n-1}(*, C, F, k, *, .)]$$

avec :

➤ $NO_{n-1}(*, E, *, x - 1, *)$: Nombre des nouveaux orphelins suite aux décès d'actifs ou de retraités, il se calcule comme suit :

$$NO_{n-1}(*, E, s, x - 1, *) = Nme * D_{n-1}(*, A, *, x - 1 + e, *) * PE(s)$$

- $S_{n-1}(A, E, *, x - 1, *)$: représente les sorties des enfants suite à la sortie des parents pour autres raisons que le décès, il se calcule comme suit :

$$S_{n-1}(A, E, s, x - 1, *) = Nme * S_{n-1}(A, A, *, x - 1 + e, *) * PE(s)$$

- $ENR_{n-1}(A, E, *, x - 1, *)$: représente les enfants des nouveaux retraités au cours de l'année $n - 1$, il se calcule comme suit :

$$ENR_{n-1}(A, E, *, x - 1, *) = \begin{cases} E_{n-1}(A, E, *, x - 1, *) & \text{si } x - 1 + e + 1 = r \\ 0 & \text{sinon} \end{cases}$$

- $ENInv_{n-1}(A, E, *, x - 1, *)$: représente les enfants des nouveaux invalides au cours de l'année $n-1$, il se calcule comme suit :

$$ENInv_{n-1}(A, E, s, x - 1, *) = Nme * NInv_{n-1}(Inv, A, *, x - 1 + e, *) * PE(s)$$

3. Interprétation des résultats

Dans cette section, nous allons estimer les différents paramètres du modèle démographique définis ci-dessus.

Nous allons utiliser la base de données relative à la population de CNOPS de l'année 2014, que nous serons amenés à dupliquer afin de travailler avec des âges annuels au lieu des tranches d'âges quinquennales.

Pour ce faire, nous allons répartir uniformément l'effectif de chaque classe de population sur les tranches d'âges, par exemple, pour le cas de la tranche d'âge quinquennal $[5 ; 10[$, nous allons diviser l'effectif relatif à cette ligne par l'amplitude 5 et l'affecter à chaque ligne dupliquée respectivement aux âges annuels 5 ans jusqu'à 9 ans. Cependant, pour les tranches d'âges $[0 ; 1[$ et $[1 ; 5[$, nous allons diviser l'effectif respectivement par les amplitudes 2 et 4.

Le code relatif à cette tâche sera exposé en annexe.

Après avoir dupliqué la base de données de la population sous risque, nous allons procéder à des estimations des équations démographiques présentées dans la partie précédente grâce à la commande « rechercheV » sous EXCEL, et cela pour chaque sous-populations des bénéficiaires afin d'obtenir l'effectif global de la population pour l'année 2015. Par raisonnement analogue, nous allons projeter la population sur la période 2016 - 2019.

Pour des soucis de confidentialité, les tables utilisées et les valeurs retenues des paramètres ne figureront pas dans ce rapport. Cependant nous allons présenter l'allure de l'évolution de cette projection de 5 ans.

La figure suivante représente l'évolution de la population globale selon notre modèle :

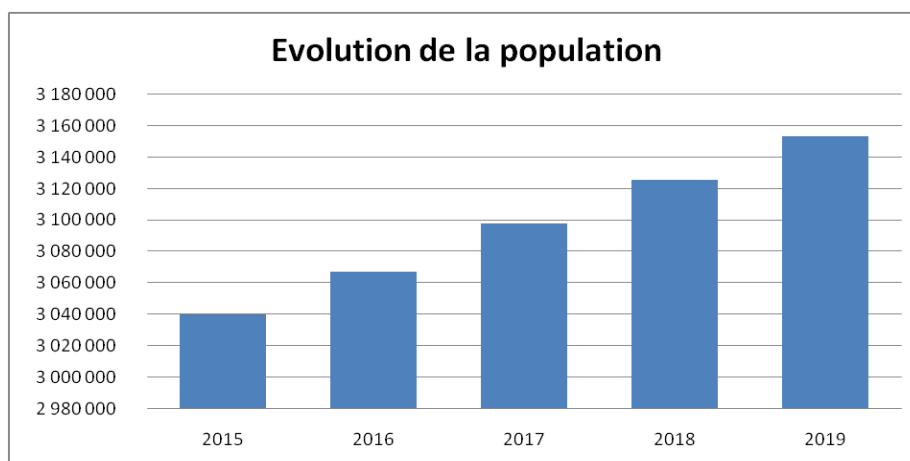


Figure 27 : Effectifs globaux projetés de la population AMO-CNOPS

En général, la population de la CNOPS est en augmentation continue légère, soit de 0,5% à 1% en moyenne.

Pour mieux visualiser cette évolution, nous allons estimer l'effectif des ALD à partir de l'effectif total en utilisant la formule suivante :

$$E_n(*,*,*,x,1,*) = \text{poidsALD}(x) * (1 + \text{TAGpoids}_{\text{ALD}})^n * E_n(*,*,*,x,.,*)$$

Cette équation suppose que le caractère ALD d'un bénéficiaire ne dépend que de son âge.

Le graphique suivant représente l'évolution de la population des personnes atteintes d'une affection longue durée comme suit :

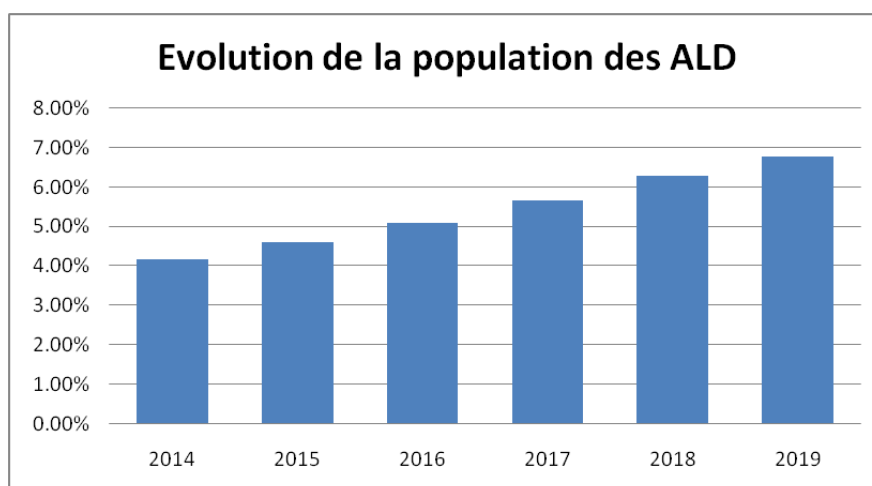


Figure 28 : Evolution de l'effectif de la population atteinte d'au moins une ALD entre 2014 et 2019

L'effectif des personnes déclarées ayant au moins une ALD a évolué de 4,17% de la population globale de 2014, pour atteindre 6,78% de la population globale de 2019.

Sachant que la population des ALD de 2014 s'accapare à elle seule presque de la moitié des dépenses du régime, donc le fait qu'elle évolue dans le temps implique une augmentation des dépenses sanitaires de la CNOPS.

Dans le double but de réduire le poids des maladies chroniques à l'avenir et contribuer par la même occasion à réduire les dépenses engendrées par des pathologies lourdes, coûteuses ou de longue durée, la CNOPS est amenée à mettre en place un plan d'action portant sur la communication et la sensibilisation sur certains comportements nuisibles à la santé tel que le tabagisme, l'alcoolisme, la mal nutrition etc.

Par conséquent, la prévention est un des outils performants pour faire de la régulation car elle est perçue comme une gestion du capital de santé avec une participation individuelle responsable.

Chapitre 5 : Les résultats obtenus

Ce dernier chapitre sera consacré à l'étude de la situation économique de la caisse nationale des organismes de prévoyance sociale en termes de dépenses pour l'année 2014 et le dressement d'un aperçu sur sa situation entre 2015 et 2019.

Pour ce faire, nous allons se baser sur l'approche « fréquence*coût-moyen *population », illustrée comme suit:

$$\begin{aligned} \text{Fréquence} * \text{Coût moyen} * \text{effectif} &= \frac{\text{Effectif sinistré}}{\text{Effectif}} \times \frac{\text{remboursement}}{\text{Effectif sinistré}} \times \text{Effectif} \\ &= \text{remboursement} \end{aligned}$$

Les méthodes linéaires généralisées nous ont permis de modéliser la fréquence des sinistres et le coût moyen, en se basant sur le choix de la loi qui ajuste le mieux les données.

Le modèle démographique nous a permis également de modéliser la population sous risque en se basant sur la loi entrée-sortie et d'obtenir des projections pour une période de 5 ans.

A ce stade, nous allons pencher sur l'exploitation des résultats des deux chapitres précédents afin de pouvoir modéliser la consommation annuelle en soins médicaux des bénéficiaires de la couverture médicale obligatoire en secteur public.

1. Utilisation des résultats

Pour ce type d'étude, on définit un agent de référence qui est le profil auquel nous référons lors du calcul des autres profils, on considère que la consommation qui lui correspond représente une consommation de base ou de référence. Mathématiquement, le choix des modalités de l'agent de référence n'affecte en rien les résultats trouvés car toutes les informations censés être contenues dans ces modèles éliminées, nous les retrouvons dans la constante du modèle β_0 notée « Intercept ».

En appliquant à nos données les modèles linéaires généralisés tels que nous les avons formalisés précédemment, Le logiciel SAS nous renvoie ainsi une série de coefficients estimés applicables aux différentes modalités des variables explicatives retenues, que ce soit pour les fréquences et les coûts moyens.

Ces coefficients du modèle estimé nous traduisent soit une majoration par rapport à la consommation de l'agent de référence s'ils sont positifs, soit une réduction s'ils sont négatifs.

On rappelle que notre choix a porté sur un modèle multiplicatif avec une fonction de lien logarithme tel que :

$$\ln \mu_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad \text{avec} \quad \mu_i = E(y_i), \quad y_i \text{ est la variable endogène}$$

Pour obtenir les fréquences et les coûts à proprement dits, On doit calculer donc l'exponentiel des coefficients renvoyés par le logiciel SAS.

Ainsi :

$$\mu_i = \exp \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right) = \exp(\beta_0) \prod_{j=1}^p \exp(\beta_j x_{ij})$$

Le tableau ci-dessous regroupe les paramètres estimés des deux lois retenues pour la modélisation de la consommation, à savoir la loi Log-Normale et loi Binomiale négative.

Paramètre		Log-Normale		Binomiale-Négative	
		Valeur estimée	EXP(valeur estimée)	Valeur estimée	EXP(valeur estimée)
Intercept		8,23	3749,58	-0,32	0,72
type_ass	A	0,22	1,24	0,19	1,21
type_ass	P1	0,39	1,47	0,05	1,05
type_ass	P2	0,24	1,27	0,00	1,00
type_ass	P3	0,00	1,00	0,00	1,00
type_benef	A	0,90	2,46	0,17	1,18
type_benef	CA	0,73	2,07	0,16	1,17
type_benef	EA	0,00	1,00	0,00	1,00
sexe	F	0,14	1,15	0,16	1,17
sexe	M	0,00	1,00	0,00	1,00
ald	N	-1,82	0,16	-0,86	0,42
ald	O	0,00	1,00	0,00	1,00
tranche_age	majo	0,08	1,08	0,00	1,00
tranche_age	mino	0,09	1,09	-0,26	0,77
tranche_age	neut	0,00	1,00	0,00	1,00
region	reg_m	-0,19	0,82	0,00	1,00
region	reg_n	0,00	1,00	0,00	1,00

Tableau 35 : Les paramètres estimés pour les lois Log normale et Binomiale négative

Calculons par exemple les dépenses annuelles estimées pour le segment suivant : « les Femmes Actives, Assurées, âgées de 46 ans (majo), habitants la région 05 (reg_m) et non atteintes d'une affection de longue durée ».

On a alors pour la fréquence moyenne :

Fréquence moyenne estimée = $0,72 \times 1,21 \times 1,18 \times 1,17 \times 0,42 \times 1 \times 1$

Fréquence moyenne estimée = 50,52%

En moyenne, la fréquence des sinistres pour l'ensemble des femmes actives, assurées, âgées de 46 ans, habitants la région 05 et non atteintes d'une affection de longue durée est de 50,52%

De la même façon, on calcule le coût moyen estimé pour ce segment comme suit :

Coût moyen estimé = $3\,749,58 \times 1,24 \times 2,46 \times 1,15 \times 0,16 \times 1,08 \times 0,82$

Coût moyen estimé = 1 863,78 Dhs

En moyenne, le coût des sinistres pour l'ensemble des femmes actives, assurées, âgées de 46 ans, habitants la région 05 et non atteintes d'une affection de longue durée est de 1 863,78 Dhs

On précise que la fréquence moyenne et le coût moyen s'interprètent comme le produit du coefficient de la classe de référence par les coefficients des modalités du segment i. Les coefficients des autres modalités viennent ainsi majorer ou minorer le coût moyen ou la fréquence moyenne du segment de référence.

L'effectif de ce segment pour l'année 2014 est de 65758.

Donc, en appliquant notre approche déjà définie précédemment, on obtient pour notre segment un montant remboursé estimé de $50,52\% \times 1\,863,78 \times 65\,758 = 61\,916\,526$ Dhs pour l'année 2014.

La même démarche peut être adoptée quel que soit le segment choisi. Il est à noter que nous disposons de 6 variables qualitatives ayant respectivement entre 2 à 4 modalités, ce qui donne un total de 288 possibilités de combinaisons.

2. Backtesting

Pour notre modèle retenu, nous disposons de l'estimation des montants remboursés pour l'année 2014, vérifions si cette estimation permet de couvrir les remboursements réellement effectués :

Total Montants Remboursés réels	3 328 698 746 Dhs
Total Montants Remboursés estimés	3 242 923 155 Dhs
Quotient Réel/ Estimé	102.64 %

Tableau 36 : Résultats du GLM au titre de l'année 2014

D'après le tableau ci-dessus, il s'avère que les remboursements réels dépassent ceux estimés de 2.64%.

On peut conclure donc que notre modèle est bon vu que le biais 2.64% est très acceptable.

En se basant sur les projections démographiques obtenues dans le chapitre de la modélisation de la population, on peut estimer les montants remboursés futurs par le régime comme suit:

$$\begin{aligned} & \text{Le montant remboursé estimé de l'année } i \\ & = \\ & \text{Fréquence moyenne estimée} \times \text{Coût moyen estimé} \\ & \quad \times \text{population estimée de l'année } i \end{aligned}$$

Avec comme hypothèses :

- fréquence moyenne constante
- un taux d'inflation de 1%.

On dresse les résultats obtenus comme suit :

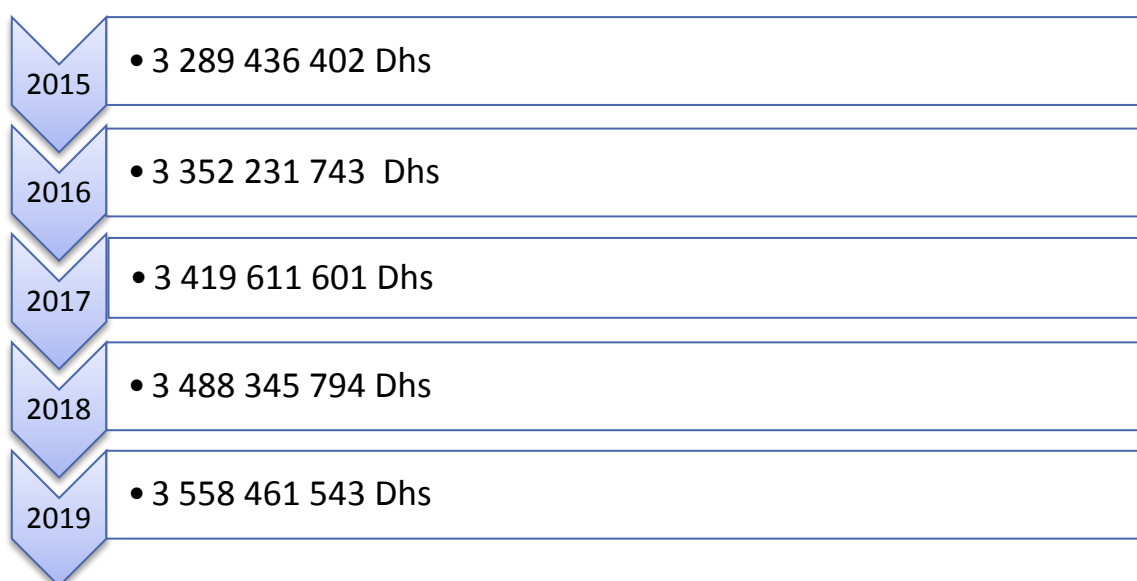


Figure 29 : L'estimation des montants remboursés pour une période de 5 ans.

Il est à noter que ces montants sont donnés à **titre indicatif** par souci de confidentialité.

Pour bien illustrer la performance de notre modèle établi, nous représentons graphiquement l'évolution du montant remboursé par année :

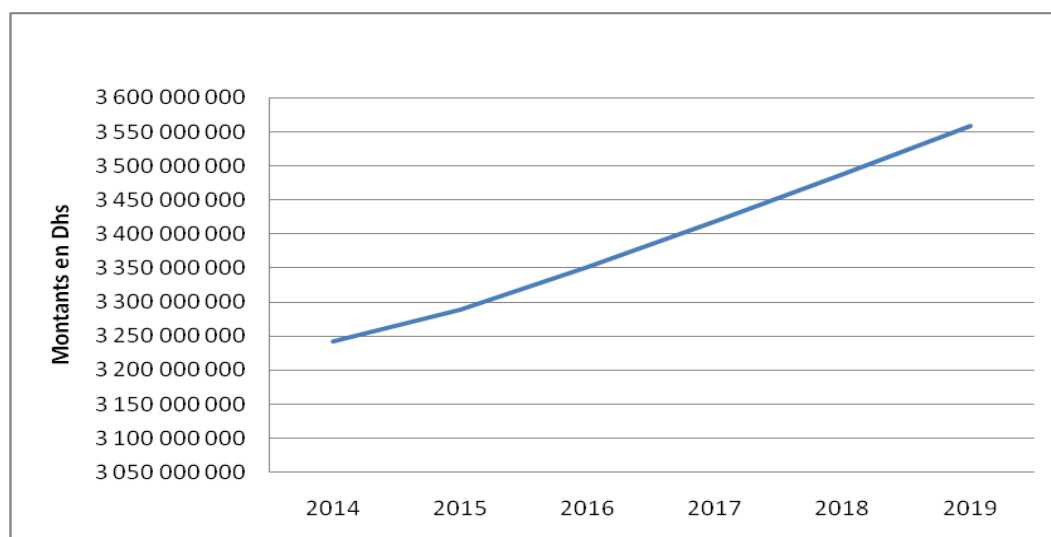


Figure 30 : L'évolution du montant global remboursé par la CNOPS entre 2014 et 2019

Le graphique ci-dessus met en évidence une évolution presque linéaire des montants remboursés par la CNOPS durant la période 2014-2019 qui est de 1.62% en moyenne par an.

3. Conclusion

Les résultats obtenus sur l'étude des remboursements en consommation médicale sont globalement satisfaisants, ainsi, le modèle choisi prend correctement en compte les différents types de comportement des bénéficiaires vis-à-vis des soins médicaux.

La modélisation retenue permet en effet de déterminer si une variation de la consommation annuelle provient d'une évolution de la fréquence moyenne ou d'une évolution du coût moyen ou bien de la population sous risque. Toutefois, le choix de cette modélisation nous pousse à accepter l'hypothèse d'indépendance de la survenance des sinistres.

On remarque de plus que la modélisation retenue pour l'étude présente l'énorme avantage d'être très flexible et ainsi de pouvoir être actualisée très rapidement. Ceci peut notamment être utile dans le cas d'une modification des remboursements estimés d'une année à une autre.

Conclusion générale

L'uniformisation des données a été un travail important en termes de temps passé, de rigueur exigée et de ressources informatiques. Une fois cette uniformisation terminée, nous avons pu réaliser les premières études statistiques sur la population sous risque et la population consommatrice de l'année 2014, ainsi que sur leur jointure. Nous avons ainsi constaté que le portefeuille étudié respectait les grandes tendances observées sur la consommation médicale obligatoire au Maroc.

L'objectif de ce mémoire est de donner lieu à la mise en place de l'approche « fréquence *coût* population » afin d'obtenir les remboursements annuels en soins médicaux des bénéficiaires de la Caisse Nationale des Organismes de Prévoyance Sociale. En effet, la consommation médicale annuelle desdits bénéficiaires a pu être modélisée à travers deux études menées conjointement. La première étude concerne la modélisation de la fréquence moyenne de survenance de sinistres et le coût moyen par sinistre en se basant sur des modèles linéaires généralisés, tandis que la deuxième étude porte sur la modélisation de la population en question.

Afin de permettre l'utilisation des modèles linéaires généralisés, la fréquence moyenne de survenance de sinistres et le coût moyen par sinistre ont été modélisés par des lois de probabilité usuelles. L'application du modèle retenu à ces lois nous a permis d'obtenir des coefficients correctifs qui expriment les remboursements annuels moyens en consommation médicale d'un segment de bénéficiaires en fonction des remboursements annuels moyens d'un segment de référence. Par ailleurs, 6 variables explicatives sont retenues pour l'étude, à savoir le type de l'assuré, le type du bénéficiaire, le sexe, l'âge, la région, et l'affection de longue durée.

L'étude de la population assurée par la Caisse Nationale des Organismes de Prévoyance Sociale repose sur l'élaboration d'un modèle empirique servant à la projection de la population. Ce modèle a été basé sur le bilan des flux, internes et externes, que connaît la population en question. L'estimation des différents paramètres du modèle a été omise dans ce rapport vu le caractère confidentiel des résultats obtenus.

Ainsi, l'approche proposée, nous a permis d'obtenir les dépenses probables en soins de santé sur la période allant de 2014 à 2019. Le premier constat que permette de faire l'analyse des résultats obtenus est que les dépenses en question sont en constante évolution, ce qui met la CNOPS au défi et l'incite à prendre les mesures nécessaires en vue d'assurer sa viabilité.

Bibliographie et Webographie

Référence :

- Rapport Annuel Global de L'assurance Maladie Obligatoire au titre de l'année 2013.
- RAPPORT D'ACTIVITES, Branche Assurance Maladie Obligatoire, Année 2014.
- Feuille de route 2014-2018
- Pratique de l'analyse de données, Rafael Costa et G. Masuy Stroobant Louvain la Neuve 2013
- LANGAGE SAS, Axelle Chauvet-Peyrard 2012

Notes de cours :

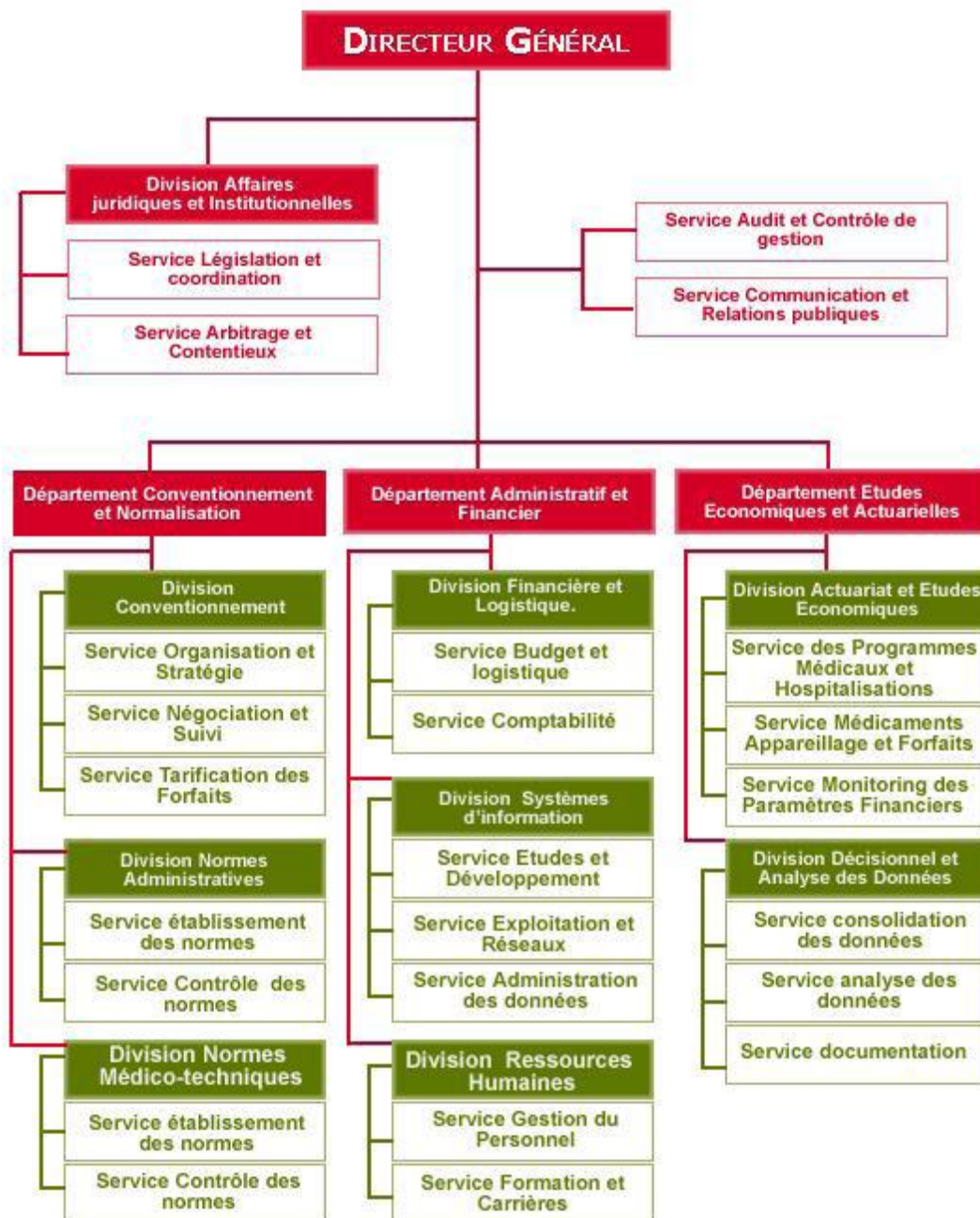
- Pr. CHAOUBI Abdelaziz : cours d'analyse de données
- Pr. TIRARI Mohammed EL Haj : cours de l'analyse de la variance
- Pr. MARRI Fouad : Cours des modèles linéaires généralisés (GLM)

Sites web :

- <http://www.assurancemaladie.ma/>
- <http://www.ressources-actuarielles.net/>
- <http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-mlg.pdf>
- http://perso.math.univ-toulouse.fr/ldinetan/files/2012/08/tp3_mlg_sid1213.pdf
- http://www.coopami.org/fr/countries/countries/marocco/projects/2015/pdf/2015_113007.pdf
- <http://www2.sas.com/proceedings/forum2008/333-2008.pdf>

Annexes :

Annexe 1 : Organigramme de l'ANAM



Annexe 2 : Liste des ALD

- 1- Accident vasculaire cérébral ou médullaire ischémique ou hémorragique
- 2- Affections malignes du tissu lymphatique ou hématopoïétique
- 3- Anémies hémolytiques chroniques sévères
- 4- Aplasies médullaires sévères
- 5- Artériopathies chroniques
- 6- Asthme sévère
- 7- Cardiopathies congénitales
- 8- Cirrhoses du foie
- 9- Diabète insulino-dépendant et diabète non insulino-dépendant
- 10- Epilepsie grave
- 11- Etat de déficit mental
- 12- Formes graves des affections neurologiques et neuromusculaires
- 13- Glaucome chronique
- 14- Hypertension artérielle sévère
- 15- Insuffisance cardiaque
- 16- Insuffisance rénale aiguë
- 17- Insuffisance rénale chronique terminale
- 18- Insuffisance respiratoire chronique grave
- 19- Lupus érythémateux aigu disséminé
- 20- Maladie coronaire
- 21- Maladie de Crohn évolutive
- 22- Maladie de Parkinson
- 23- Maladies chroniques actives du foie (hépatites B et C)
- 24- Myélodysplasies sévères
- 25- Néphropathies graves
- 26- Polyarthrite rhumatoïde évolutive grave
- 27- Psychoses
- 28- Rectocolite hémorragique évolutive
- 29- Rétinopathie diabétique
- 30- Sclérodémie généralisée évolutive
- 31- Sclérose en plaques
- 32- Spondylarthrite ankylosante grave
- 33- Syndrome d'immunodéficience acquise (SIDA)
- 34- Syndromes néphrotiques
- 35- Troubles graves de la personnalité
- 36- Troubles héréditaires de l'hémostase
- 37- Troubles mentaux et/ou de personnalité dus à une lésion, à un dysfonctionnement cérébral ou à une lésion physique
- 38- Troubles permanents du rythme et de la conductivité
- 39- Tumeurs malignes
- 40- Valvulopathies rhumatismales
- 41- Vascularites

Annexe 3 : Sortie de L'AFC

Pour le couplet Type_benef et Sexe :

Récapitulatif

Dimension	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie		Valeur singulière de confiance
					Expliqué	Cumulé	Ecart-type
1	,208	,043			1,000	1,000	,013
Total		,043	196,164	,000 ^a	1,000	1,000	

Tableau 35 : Inertie par dimension des deux variables type_benef et sexe

Caractéristiques des points lignes^a

type_benef	Masse	Score dans la dimension	Inertie	Contribution		
		1		De point à inertie de dimension	De dimension à inertie de point	
				1	1	Total
Assuré	,413	,182	,003	,066	1,000	1,000
Conjoint de l'assuré	,216	-,864	,034	,773	1,000	1,000
Enfant de l'assuré	,370	,301	,007	,161	1,000	1,000
Total actif	1,000		,043	1,000		

a. Normalisation principale symétrique

Tableau 36 : caractéristiques des profils lignes pour la variable type_benef

Caractéristiques des points colonnes^a

sexe	Masse	Score dans la dimension	Inertie	Contribution		
		1		De point à inertie de dimension	De dimension à inertie de point	
				1	1	Total
Femme	,568	-,398	,019	,432	1,000	1,000
Homme	,432	,524	,025	,568	1,000	1,000
Total actif	1,000		,043	1,000		

a. Normalisation principale symétrique

Tableau 37 : Caractéristiques des profils colonnes pour la variable sexe

Pour le couplet Type_benef et ALD :

Récapitulatif

Dimension	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie		Valeur singulière de confiance
					Expliqué	Cumulé	Ecart-type
1	,011	,000			1,000	1,000	,015
Total		,000	,531	,767 ^a	1,000	1,000	

a. 2 degrés de liberté

Tableau 38 : Inertie par dimension des deux variables type_benef et ALD

Caractéristiques des points lignes^a

type_benef	Masse	Score dans la dimension	Inertie	Contribution		
		1		De point à inertie de dimension	De dimension à inertie de point	
				1	1	Total
Assuré	,413	,034	,000	,044	1,000	1,000
Conjoint de l'assuré	,216	-,195	,000	,759	1,000	1,000
Enfant de l'assuré	,370	,076	,000	,197	1,000	1,000
Total actif	1,000		,000	1,000		

a. Normalisation principale symétrique

Tableau 39 : Caractéristiques des profils lignes pour la variable type_benef

Caractéristiques des points colonnes^a

ald	Masse	Score dans la dimension	Inertie	Contribution		
		1		De point à inertie de dimension	De dimension à inertie de point	
				1	1	Total
Non-ald	,597	-,086	,000	,403	1,000	1,000
ald	,403	,127	,000	,597	1,000	1,000
Total actif	1,000		,000	1,000		

a. Normalisation principale symétrique

Tableau 40 : Caractéristiques des profils colonnes pour la variable ald

Pour le couplet Sexe et ALD :

Récapitulatif

Dimension	Valeur singulière	Inertie	Khi-deux	Sig.	Proportion d'inertie		Valeur singulière de confiance
					Expliqué	Cumulé	Ecart-type
1	,013	,000			1,000	1,000	,015
Total		,000	,805	,369 ^a	1,000	1,000	

a. 1 degrés de liberté

Tableau 41 : Inertie par dimension des deux variables sexe et ALD

Caractéristiques des points lignes^a

sexe	Masse	Score dans la dimension	Inertie	Contribution		
		1		De point à inertie de dimension	De dimension à inertie de point	
				1	1	Total
Femme	,568	-,101	,000	,432	1,000	1,000
Homme	,432	-,133	,000	,568	1,000	1,000
Total actif	1,000		,000	1,000		

a. Normalisation principale symétrique

Tableau 42 : Caractéristiques des profils lignes pour la variable sexe

Caractéristiques des points colonnes^a

ald	Masse	Score dans la dimension	Inertie	Contribution		
		1		De point à inertie de dimension	De dimension à inertie de point	
				1	1	Total
Non-ald	,597	-,095	,000	,403	1,000	1,000
ald	,403	,141	,000	,597	1,000	1,000
Total actif	1,000		,000	1,000		

a. Normalisation principale symétrique

Tableau 43 : Caractéristiques des profils lignes pour la variable ALD

Annexe 4 : CODE SAS POUR LA PROC GENMOD (MODELES LINEAIRES GENERALISES)

```
libname CNOPS 'C:\Users\HP 650\Desktop\CNOPS';  
%let biblio = CNOPS;
```

➤ Segmentation des variables tranche d'âge et région

```
/*Région*/
```

```
ods TAGSETS.EXCELXP  
body='C:\Users\HP 650\Desktop\non\agetetsh.xls';  
proc genmod data=cnops.Mrtesth;  
CLASS type_ass type_benef sexe tranche_age region;  
Model loga = region /dist=Normal link=id type3;  
Weight effectif_sinistre;  
run;  
ods TAGSETS.EXCELXP close;  
  
data cnops.Mr2; set cnops.Mr1;  
if region="reg2" or region="reg3" or region="reg4" or region="reg5" or region="reg6" or region="reg7" or  
region="reg8" or region="reg11" or region="reg12" or region="reg13" or region="reg14" or  
region="reg15" or region="reg16" then region="reg_m";  
if region="reg9" or region="reg17" or region="reg10" or region="reg1" then region="reg_n";  
run;
```

```
/* Tranche d'âge */
```

```
ods TAGSETS.EXCELXP  
body='C:\Users\HP 650\Desktop\non\agetetsh.xls';  
proc genmod data=cnops.Mrtesth;  
CLASS type_ass type_benef sexe tranche_age region;  
Model loga = tranche_age /dist=Normal link=id type3;  
Weight effectif_sinistre;  
run;  
ods TAGSETS.EXCELXP close;  
  
data cnops.Mr2; set cnops.Mr1;  
if tranche_age="AG9" or tranche_age="AG10" or tranche_age="AG7" or tranche_age="AG8" then  
tranche_age="neutre";  
if tranche_age="AG11" or tranche_age="AG12" or tranche_age="AG13" or tranche_age="AG14" or  
tranche_age="AG15" or tranche_age="AG16" then tranche_age="majorant";  
if tranche_age="AG1" or tranche_age="AG2" or tranche_age="AG3" or tranche_age="AG4" or  
tranche_age="AG5" or tranche_age="AG6" then tranche_age="minorant";  
run;
```

➤ Modélisation des coûts moyens des sinistres

*/*PP-PLOT d'ajustement de MRM aux lois GAMMA et Log-Normale*/*

```
Ods pdf
body='C:\Users\HP 650\Desktop\tout\qqplot_MRM_lognormal.pdf';
proc CAPABILITY data=cnops.Tout;
VAR MRM;
QQPLOT MRM /Lognormal (SIGMA=EST SLOPE=EST THETA=est) square;
PPLOT MRM /Lognormal (SIGMA=EST THETA=est) square;
histogram MRM /Lognormal(SIGMA=EST THETA=est);
run;
ods pdf close;
```

```
ods pdf
body='C:\Users\HP 650\Desktop\tout\qqplot_MRM_gamma.pdf';
PROC CAPABILITY DATA=cnops.Tout ;
VAR MRM;
QQPLOT MRM /gamma (ALPHA=EST THETA=estSIGMA=EST) square;
PPLOT MRM /gamma (ALPHA=EST THETA=estSIGMA=EST) square ;
histogram MRM /gamma (ALPHA=EST THETA=est SIGMA=EST) ;
RUN;
Ods pdf close;
```

*/*Calcul de log de MRM*/*

```
Data cnops.Tout;
Set cnops.Tout;
loga = log(MRM);
run;
```

*/*Estimation des paramètres du modèle sans interaction avec la loi Gamma*/*

```
ods TAGSETS.EXCELXP
body='C:\Users\HP 650\Desktop\tout\gamma.xls';
proc genmod data=cnops.Tout;
CLASS type_ass type_benef sexe ald tranche_age region;
model MRM = type_asstype_benefsexealdtranche_age region /dist=gamma link=log type3;
weight effectif_sinistre;
run;
ods TAGSETS.EXCELXP close;
```

*/*Estimation des paramètres du modèle sans interaction avec la loi Log-Normale*/*

```
ods TAGSETS.EXCELXP
body='C:\Users\HP 650\Desktop\tout\loga.xls';
proc genmod data=cnops.Tout;
CLASS type_asstype_benefsexealdtranche_age region;
Model loga =type_asstype_benefsexealdtranche_age region /dist=normal link=id type3;
Weight effectif_sinistre;
run;
ods TAGSETS.EXCELXP close;
```

*/*Rédidu de la deviance avec la loi Log-Normale*/*

```
Proc genmod data=cnops.Tout;
CLASS type_ass type_benef sexe ald tranche_age region;
```

```

modelloga = type_asstype_benefsexaldtranche_age region /dist=normal link=id;
weighteffectif_sinistre;
outputout=cnops.residual_mrmPRED=mu_chapeauRESCHI=residual_pearsonRESDEV=residual_deviance;
run;

```

/ Graphique des résidus de la déviance */*

```

procplotdata= cnops.Residual_mrm;
plotresidual_deviance*mu_chapeau;
goptionshsize=15cmvsize=15cm;
run;

```

*/*le MRM estimé*/*

```

datacnops.residual_mrm; setcnops.residual_mrm;
MRM_estime=exp(mu_chapeau);
run;

```

➤ Modélisation de la fréquence moyenne des sinistres

*/*Calcul de la moyenne et de la variance de l'effectif_sinistré*/*

```

procmeansdata = cnops.Toutmeanvar;
vareffectif_sinistre;
run;

```

*/*L'ajustement graphique à la loi de poisson*/*

```

>library(MASS)
>library("vcd")
>effectif_sin=read.csv2("jointure.csv")
>x=effectif_sin$x
>plot(goodfit(x,"pois"),main="ajustement à la loi Poisson")

```

*/*L'ajustement à la loi de poisson*/*

```

ods TAGSETS.EXCELXP
body='C:\Users\HP 650\Desktop\tout\poisson.xls';
procgenmoddata=cnops.Tout;
CLASStype_asstype_benefsexaldtranche_age region;
modeleffectif_sinistre = type_asstype_benefsexaldtranche_age region /dist=poi link=log offset=offset
type3;
run;
ods TAGSETS.EXCELXP close;

```

*/*L'ajustement à la loi binomiale négative*/*

```

ods TAGSETS.EXCELXP
body='C:\Users\HP 650\Desktop\tout\negbin.xls';
procgenmoddata=cnops.Tout;
CLASStype_asstype_benefsexaldtranche_age region;
modeleffectif_sinistre = type_asstype_benefsexaldtranche_age region /dist=negbinlink=log offset=offset
type3;
run;
ods TAGSETS.EXCELXP close;

```

```
/*Rédidu de la deviance avec la loi binomiale négative*/
```

```
Procgenmoddata=cnops.Tout;  
CLASStype_asstype_benefsexealdtranche_age region;  
modeleffectif_sinistre= type_asstype_benefsexealdtranche_age region /dist=negbinlink=log offset=offset;  
outputout=cnops.residual_esPRED=mu_chapeauRESDEV=residual_devianceRESCHI=residual_pearson ;  
run;
```

```
/* Graphique des résidus de la déviance */
```

```
Procplotdata= cnops.residual_es;  
plotresidual_deviance*mu_chapeau;  
run;
```

```
/*Comparaison 2 à 2 des modalités en termes de moyenne de la fréquence des sinistres*/
```

```
ods TAGSETS.EXCELXP  
body='C:\Users\HP 650\Desktop\tout\lsmeans.xls';  
procgenmoddata=cnops.Tout ;  
CLASStype_asstype_benefsexealdtranche_age region;  
modeleffectif_sinistre = type_asstype_benefsexealdtranche_age region /dist=negbinlink=log offset=offset ;  
lsmeantype_asstype_benefsexealdtranche_age region / DIFF;  
run;  
ods TAGSETS.EXCELXP close;
```

```
/*Regroupement des modalités*/
```

```
data cnops.ls; setcnops.tout;  
iftype_ass="P2" or type_ass="P3"thentype_ass="survivants";  
iftranche_age="majo" or tranche_age="neut"thentranche_age="plus25";  
iftranche_age="mino"thentranche_age="moins25";  
if region="reg_m" or region="reg_n"then region="reg";  
run;
```

Annexe 5 : CODE SAS POUR LAPROC REG (REGRESSION LINEAIRE MUPLTIPLE)

*/*Codification des variables*/*

```
data cnops.MRM1; setcnops.Mrm;
actif=0; iftype_ass="A"thenactif=1;
pensionne=0; iftype_ass="P1"thenpensionne=1;
CS=0; iftype_ass="P2"then CS=1;
ES=0; iftype_ass="P3"then ES=1;
```

```
assure=0; iftype_benef="A"then assure=1;
CA=0; iftype_benef="CA"then CA=1;
EA=0; iftype_benef="EA"then EA=1;
```

```
homme=0; ifsexe="M"thenhomme=1;
femme=0; ifsexe="F"then femme=1;
```

```
ald_oui=0; ifald="O"thenald_oui=1;
ald_non=0; ifald="N"thenald_non=1;
```

```
region1=0; if region="reg1"then region1=1;
region2=0; if region="reg2"then region2=1;
region3=0; if region="reg3"then region3=1;
region4=0; if region="reg4"then region4=1;
region5=0; if region="reg5"then region5=1;
region6=0; if region="reg6"then region6=1;
region7=0; if region="reg7"then region7=1;
region8=0; if region="reg8"then region8=1;
region9=0; if region="reg9"then region9=1;
region10=0; if region="reg10"then region10=1;
region11=0; if region="reg11"then region11=1;
region12=0; if region="reg12"then region12=1;
region13=0; if region="reg13"then region13=1;
region14=0; if region="reg14"then region14=1;
region15=0; if region="reg15"then region15=1;
region16=0; if region="reg16"then region16=1;
region17=0; if region="reg17"then region17=1;
```

```
age1=0; iftranche_age="AG1"then age1=1;
age2=0; iftranche_age="AG2"then age2=1;
age3=0; iftranche_age="AG3"then age3=1;
age4=0; iftranche_age="AG4"then age4=1;
age5=0; iftranche_age="AG5"then age5=1;
age6=0; iftranche_age="AG6"then age6=1;
age7=0; iftranche_age="AG7"then age7=1;
age8=0; iftranche_age="AG8"then age8=1;
age9=0; iftranche_age="AG9"then age9=1;
age10=0; iftranche_age="AG10"then age10=1;
age11=0; iftranche_age="AG11"then age11=1;
age12=0; iftranche_age="AG12"then age12=1;
age13=0; iftranche_age="AG13"then age13=1;
age14=0; iftranche_age="AG14"then age14=1;
age15=0; iftranche_age="AG15"then age15=1;
age16=0; iftranche_age="AG16"then age16=1;
```

```
run;
```

```
/*Régression linéaire multiple du MRM*/
```

```
ods TAGSETS.EXCELXP  
body='C:\Users\HP 650\Desktop\CNOPS\reg1.xls';  
PROCREG data=cnops.MRM1;  
MODEL MRM = actifpensionne CS assure CA hommeald_oui region1 region2 region3 region4 region5  
region6 region7 region8 region9 region10 region11 region12 region13 region14 region15 region16 age1  
age2 age3 age4 age5 age6 age7 age8 age9 age10 age11 age12 age13 age14 age15;  
WEIGHT effectif_sinistre ;  
OUTPUT OUT=cnops.resultat1 P=MRM_estime R=residu_reg;  
run;  
ods TAGSETS.EXCELXP close;
```

```
/*Test de normalité des résidus*/
```

```
ods TAGSETS.EXCELXP  
body='C:\Users\HP 650\Desktop\CNOPS\test_normalité_reg1.xls';  
PROCCAPABILITY DATA=cnops.resultat1 NORMAL;  
VARresidu_reg;  
QQPLOTresidu_reg /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);  
HISTOGRAM /NORMAL(COLOR=MAROON W=4) CFILL=BLUE CFRAME=LIGR;  
RUN;  
ods TAGSETS.EXCELXP close;
```

```
/*Détection l'hétérogénéité*/
```

```
Odspdf  
body='C:\Users\HP 650\Desktop\CNOPS\heteroplot.pdf';  
procgplot data=cnops.resultat1;  
plotresidu_reg*MRM_estime="*" ;  
run;  
odspdfclose;
```

Annexe 6 : CODE SAS POUR LA DUPLICATION DE LA TABLE « POPULATION »

```
data dup.pop; set dup.pop;  
do i=1 to agedup;  
output;  
end;  
run;
```

```
data dup.pop; set dup.pop;  
if (tranche_age=1 and i=1 ) then age=0;  
if (tranche_age=1 and i=2 ) then age=1;  
  
if (tranche_age=2 and i=1 ) then age=1;  
if (tranche_age=2 and i=2 ) then age=2;  
if (tranche_age=2 and i=3 ) then age=3;  
if (tranche_age=2 and i=4 ) then age=4;  
  
if (tranche_age=3 and i=1 ) then age=5;  
if (tranche_age=3 and i=2 ) then age=6;  
if (tranche_age=3 and i=3 ) then age=7;  
if (tranche_age=3 and i=4 ) then age=8;  
if (tranche_age=3 and i=5 ) then age=9;  
  
if (tranche_age=4 and i=1 ) then age=10;  
if (tranche_age=4 and i=2 ) then age=11;  
if (tranche_age=4 and i=3 ) then age=12;  
if (tranche_age=4 and i=4 ) then age=13;  
if (tranche_age=4 and i=5 ) then age=14;  
  
if (tranche_age=5 and i=1 ) then age=15;  
if (tranche_age=5 and i=2 ) then age=16;  
if (tranche_age=5 and i=3 ) then age=17;  
if (tranche_age=5 and i=4 ) then age=18;  
if (tranche_age=5 and i=5 ) then age=19;  
  
if (tranche_age=6 and i=1 ) then age=20;  
if (tranche_age=6 and i=2 ) then age=21;  
if (tranche_age=6 and i=3 ) then age=22;  
if (tranche_age=6 and i=4 ) then age=23;  
if (tranche_age=6 and i=5 ) then age=24;  
  
if (tranche_age=7 and i=1 ) then age=25;  
if (tranche_age=7 and i=2 ) then age=26;  
if (tranche_age=7 and i=3 ) then age=27;  
if (tranche_age=7 and i=4 ) then age=28;  
if (tranche_age=7 and i=5 ) then age=29;  
  
if (tranche_age=8 and i=1 ) then age=30;  
if (tranche_age=8 and i=2 ) then age=31;  
if (tranche_age=8 and i=3 ) then age=32;  
if (tranche_age=8 and i=4 ) then age=33;  
if (tranche_age=8 and i=5 ) then age=34;  
  
if (tranche_age=9 and i=1 ) then age=35;  
if (tranche_age=9 and i=2 ) then age=36;
```

```

if (tranche_age=9 and i=3 ) then age=37;
if (tranche_age=9 and i=4 ) then age=38;
if (tranche_age=9 and i=5 ) then age=39;

if (tranche_age=10 and i=1 ) then age=40;
if (tranche_age=10 and i=2 ) then age=41;
if (tranche_age=10 and i=3 ) then age=42;
if (tranche_age=10 and i=4 ) then age=43;
if (tranche_age=10 and i=5 ) then age=44;

if (tranche_age=11 and i=1 ) then age=45;
if (tranche_age=11 and i=2 ) then age=46;
if (tranche_age=11 and i=3 ) then age=47;
if (tranche_age=11 and i=4 ) then age=48;
if (tranche_age=11 and i=5 ) then age=49;

if (tranche_age=12 and i=1 ) then age=50;
if (tranche_age=12 and i=2 ) then age=51;
if (tranche_age=12 and i=3 ) then age=52;
if (tranche_age=12 and i=4 ) then age=53;
if (tranche_age=12 and i=5 ) then age=54;

if (tranche_age=13 and i=1 ) then age=55;
if (tranche_age=13 and i=2 ) then age=56;
if (tranche_age=13 and i=3 ) then age=57;
if (tranche_age=13 and i=4 ) then age=58;
if (tranche_age=13 and i=5 ) then age=59;

if (tranche_age=14 and i=1 ) then age=60;
if (tranche_age=14 and i=2 ) then age=61;
if (tranche_age=14 and i=3 ) then age=62;
if (tranche_age=14 and i=4 ) then age=63;
if (tranche_age=14 and i=5 ) then age=64;

if (tranche_age=15 and i=1 ) then age=65;
if (tranche_age=15 and i=2 ) then age=66;
if (tranche_age=15 and i=3 ) then age=67;
if (tranche_age=15 and i=4 ) then age=68;
if (tranche_age=15 and i=5 ) then age=69;

do j=1to37; if (tranche_age=16 and i=j) then age=69+j;end;
run;

```

