Mémoire présenté pour l'obtention du diplôme d'actuaire du CEA / Institut du Risk Management et l'admission à l'Institut des Actuaires

Par : Jérôme Brosseaud et Nicolas Perrin

Titre du mémoire : Tarification des contrats Santé individuelle aux USA dans le cadre de la réforme Obamacare

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

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

Membres présents du jury de l'IA

Membres présents du jury du CEA

Entreprise
Nom : AXA France
Signature :

Directeur du mémoire en entreprise
Nom : Fabienne Cazals
Signature :

Invité
Nom :
Signature :

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

Signature du responsable entreprise

Signature du candidat

Secrétariat :

Bibliothèque :

réinventons / notre métier

Mémoire de fin d'études du Centre d'Études Actuarielles

Tarification des contrats Santé individuelle aux USA dans le cadre de la réforme Obamacare

Jérôme Brosseaud, Nicolas Perrin

2020

Résumé

Dans le contexte d'un système de Santé Américain sans équivalent (3 500 milliards de dépenses annuelles, des coûts moyens par habitant 2,5 fois plus élevés que la moyenne de l'OCDE, une absence de régime général d'assurance maladie), l'administration du Président Barack Obama décide de mettre en place une réforme d'envergure, introduisant notamment une obligation individuelle d'assurance. La problématique de tarification de ces contrats se pose plus spécifiquement à l'aune des différentes contraintes imposées. L'analyse des différentes approches de tarification, la présentation des fondements théoriques et leurs applications opérationnelles font ainsi l'objet de ce mémoire. Au-delà de ces aspects purement actuariels, cette étude vise également à présenter les spécificités d'un marché de la Santé aux Etats-Unis assez peu connu en France, aussi bien sur les parcours de soin (pour développer de la prévention, améliorer les modèles de tarification/suivi de risques, voire mieux suivre sa rentabilité) que sur la connaissance des risques sous-jacents au risque santé.

Mots clés : Système de Santé Américain, Obamacare, Contrats Santé, Tarification, Modèles linéaires généralisés

Abstract

Regarding the unparalleled US Healthcare system (3.5 trillions USD of annual expenses, an average cost per inhabitant 2.5 times higher than the OECD average, no heatlh social security), President Barack Obama's adminsitration decided to introduce an ambitious reform, especillay introducing mandatory individual Health insurance. Pricing problematic is especially raised in the light of the multiple specific constraints imposed. The analysis of the different Health insurance policy pricing, the presentation of the theoretical foundations and their operational applications are also the subject of this thesis. Beyond these purely actuarial aspects, this study also aims to present the characteristics of a Health market in the United States that are little known in France, both on the care pathways (to develop prevention, improve pricing / risk monitoring models, or even better monitor its profitability) as well as on the knowledge of the underlying health risks.

Key words: US Healtcare system, Obamacare, Health contracts, Pricing, GLM

Remerciements

Nous tenons à remercier tout particulièrement Fabienne Cazals, manager de l'équipe Modèles, Rentabilité et Big Data au sein de la Direction Technique et Innovation, ainsi qu'Ange Michel Lago et Déborah Hulot, pour leur soutien, leur apport et leur bienveillance.
Nous remercions également Jean Denis Zafar et Kaoutar Lazrak qui nous ont également accompagnés pour la rédaction de ce mémoire.
Enfin, nous remercions nos collègues, nos amis et nos familles qui nous ont soutenus et accompagnés durant cette période.

Table des matières

Introduction 7
Le contexte de mise en place de contrats « Obamacare » 7
Les contraintes de tarification de ces contrats 7
Objectif de ce mémoire d'actuariat 8
1. Les spécificités du marché de la Santé aux Etats-Unis 9
1.1. L'assurance Santé aux US : Un marché très atypique 9
1.2 Les différents acteurs du système de soins Américain 11
1.2.1 Les prestataires de soin 11
1.2.2 Le financement des dépenses de Santé 14
1.2.3 Les instances législatives et exécutives des Etats-Unis 17
1.2.4 Les lobbies 18
1.2.5 La couverture assurantielle des américains 18
1.3 Brève perspective historique de la protection sociale aux Etats-Unis 19
1.4 Présentation de l'Affordable Care Act (ACA) 21
1.4.1 Points clés de la réforme 21
1.4.2 Mise en œuvre de la réforme et chiffres clés 23
2. Caractéristiques des contrats Santé Obamacare 25
2.1 Les garanties proposées 25
2.2 Les différentes modalités de financement / remboursement 27
2.3. Contraintes réglementaires et impact sur la tarification 28
2.2.1 L'Actuarial Value 28
2.2.2 Autres éléments : risk adjustment et « filling » 30
2.2.3 Critères utilisables pour la tarification 33
3. Portefeuille considéré pour l'étude 34
3.1 Présentation des données 34
3.1.1 Hypothèses et traitements effectués 34
3.1.2 Les bases exploitables 35
3.2. Statistiques descriptives 36
3.2.1 Populations assurées 36
3.2.2 Base de sinistres 44
4. Choix et élaboration du modèle de tarification 48
4.1 Les variables d'intérêt et les variables explicatives 48
4.2 Analyse des modèles envisageables 49

4.3 L'approche « Fréquence x Coût moyen » 49
4.3.1 Calcul de la Fréquence....49
4.3.2 Calcul du Coût moyen....49
4.3.3 Calcul de la prime pure 50
4.3.4 Limites liées à l'utilisation d'une approche Fréquence x Coût moyen 50
4.4 L'approche Probabilité de consommer x Charge de consommation 51
4.4.1 Calcul de la Probabilité de consommer 51
4.4.2 Calcul de la charge de consommation....51
4.4.3 Calcul de la prime pure 52
4.4.4 Rationnel du choix de l'approche....52
4.5 La prise en compte de franchises et plafonds contractuels dans les tarifs 53
5. Les modèles linéaires généralisés pour la tarification 55
5.1 La formalisation du modèle linéaire gaussien 55
5.1.1 Estimation des paramètres....55
5.1.2 Validation du modèle....57
5.1.3 Limites du modèle linéaire gaussien et généralisation 58
5.2 La 1ère généralisation du modèle linéaire classique : les lois de la famille exponentielle comme loi pour la variable réponse....59
5.3 La deuxième généralisation du modèle linéaire classique : la fonction de lien ...60
5.3.1 La formalisation du modèle GLM 61
5.3.2 La validation du modèle....62
5.4 La modélisation de la probabilité de consommer dans l'année 64
5.4.1 Intuition....64
5.4.2 Introduction d'une variable latente pour la modélisation....64
5.5 La modélisation de la charge annuelle de consommation....66
5.5.1 Le choix de la loi gamma 66
5.5.2 Le choix du lien log 66
5.5.3 La régression log-gamma....67
5.6 La modélisation de la fréquence annuelle de consommation 69
5.7 La modélisation du coût moyen d'un acte....69
6. Application....70
6.1 Choix et structuration des variables tarifaires....70
6.1.1 Définition des garanties à tarifer 70
6.1.2 Choix des variables explicatives pour la tarification 72
6.2 La tarification d'un acte classique : la consultation d'un généraliste....74
6.2.1 Modélisation du coût moyen 74

6.2.2 Modélisation du taux d'incidence.....77
6.2.3 Modélisation de la prime pure 80
6.3 La tarification de la garantie hospitalisation.....80
6.3.1 Modélisation de la charge annuelle de consommation 81
6.3.2 Modélisation de la probabilité de consommer 84
6.3.3 Modélisation de la prime pure 87
Pistes de réflexion pour aller plus loin.....87
Conclusion.....89
Références et bibliographie.....90

Introduction

Le contexte de mise en place de contrats « Obamacare »

D'après l'Organisation Mondiale de la Santé (OMS), les Etats-Unis ont dépensé en 2016 plus que n'importe quel autre pays en soins médicaux par habitant (près de 9 000 dollars, contre 3 400 en moyenne pour les pays de l'OCDE). Plus de 17 % du PIB Américain était consacré en 2016 aux dépenses de Santé, contre 11% en France. Le système de Santé américain est en effet complexe : il fait intervenir différents acteurs des secteurs public et privé, il est de plus centré sur le concept de « service » : des entités des secteurs public ou privé « achètent » des services de santé à des « fournisseurs » ; ceux-ci étant soumis à des réglementations imposées par le gouvernement fédéral mais également à des réglementations spécifiques à chaque Etat. Contrairement au système français, il n'existe pas aux Etats-Unis de régime général d'assurance-maladie : les Etats-Unis restent la seule nation industrialisée à ne pas posséder de système de Santé universel.

En 2010, 17% de la population, soit environ 50 M de citoyens américains, ne disposait d'aucune assurance en Santé. C'est dans ce contexte que le gouvernement de Barack Obama décide d'introduire une réforme structurelle de la santé aux Etats-Unis, avec " l'Affordable Care Act" plus connue sous le nom de « Obamacare ». Cette réforme permet d'encadrer, réguler et corriger les déficiences du système de Santé. Mais surtout, la stratégie centrale vise à instaurer une obligation individuelle d'assurance ("Individual Mandate").

Les contraintes de tarification de ces contrats

Les polices d'assurance proposées dans le cadre de l'Affordable Care Act doivent respecter un certain nombre de contraintes, tant en termes de garanties qu'en termes de couverture financière ou partage des dépenses entre assureur et assuré. Au-delà des garanties en elles-mêmes, plusieurs dispositifs, tant de franchise, de coassurance ou co-paiement ainsi que de plafonds annuels coexistent. Tous ces paramètres complexifient d'autant plus la lisibilité des contrats pour des individus souhaitant comparer les offres proposées par les acteurs de l'assurance.

Afin de remédier à cette problématique de comparaison des offres et d'information du consommateur quant à la qualité du produit souscrit, la mesure de la valeur actuarielle (Actuarial Value) a été introduite. Celle-ci mesure le pourcentage des dépenses de Santé couvertes par la police d'assurance souscrite.

In fine, seuls 5 facteurs peuvent être pris en compte pour déterminer le niveau des primes au sein d'Obamacare : le lieu de résidence, l'âge, le nombre de bénéficiaires, le statut fumeur vs. non-fumeur et enfin le « niveau métallique » (notion qui sera présentée ultérieurement) du plan. C'est dans ce contexte et avec ces contraintes que la tarification de ces contrats Santé est envisagée.

Objectif de ce mémoire d'actuariat

Ce mémoire d'actuariat vise dans un 1er temps à présenter la « problématique métier », à savoir appréhender le système de soin américain, le contexte de la réforme et enfin les contraintes liées à la tarification des contrats Obamacare.

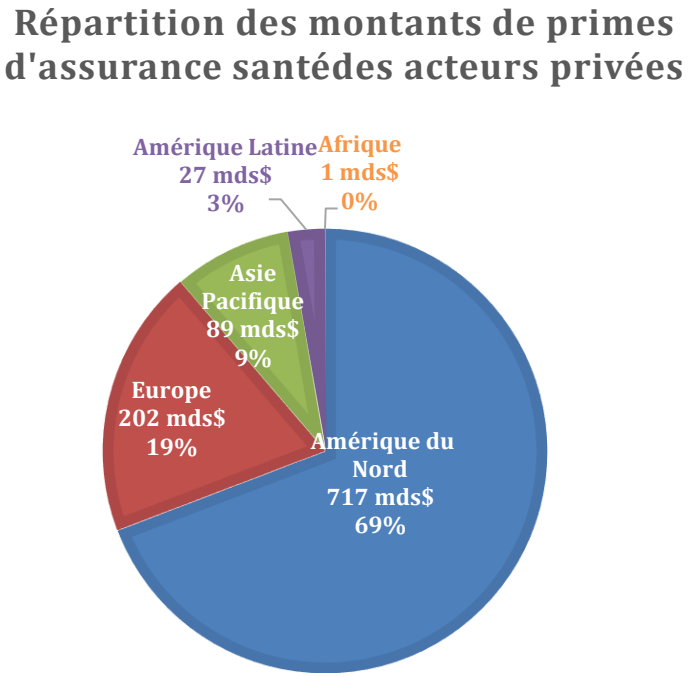
Une fois le contexte et la problématique posés, nous nous intéressons tout d'abord à l'analyse des données disponibles puis aux différentes méthodologies de tarification de contrats Santé (« Fréquence x Coût moyen » d'une part et « Probabilité de consommer x Charge de consommation » d'autre part) et enfin à la présentation des modèles linéaires généralisés (formalisation, estimation des paramètres, méthodes de validation et limites).

Une application de ces éléments théoriques est finalement mise en œuvre à partir des données exploitées. Pour conclure une ouverture sur des pistes pour approfondir le sujet sera présentée à la fin.

1. Les spécificités du marché de la Santé aux Etats-Unis

1.1. L'assurance Santé aux US : Un marché très atypique

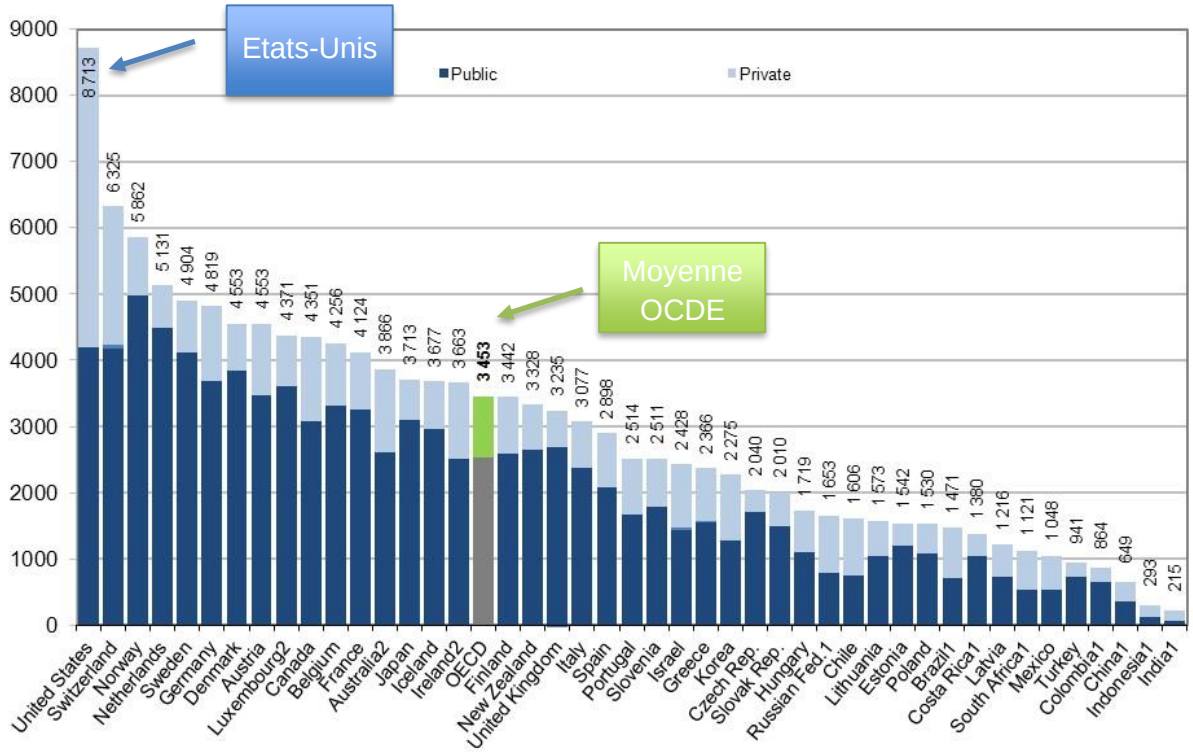
En 2015, l'activité de l'assurance Santé réservée aux acteur privés dans le monde représentait un marché d'environ $ 1 032 Mds de Primes Brutes, dont plus des 2 tiers pour l'Amérique du Nord :



Avec une concentration aussi forte du marché de la santé en Amérique du Nord, principalement aux Etats-Unis, il est intéressant de mieux comprendre quelles sont les particularités expliquant cette situation.

L'analyse de la dépense de santé par habitant au sein des différents pays de l'OCDE permet d'apporter des éléments d'explication.

En 1er lieu, on peut s'intéresser au montant moyen de dépense de Santé par habitant, pays par pays :



Source : Department for Professional Employees, AFL-CIO

On constate ainsi 2 éléments particulièrement notables :

1. Les Etats-Unis sont le pays avec de loin les plus fortes dépenses par habitant
    - Avec un montant de 8,700 dollars par habitant alors que la moyenne des pays de l'OCDE est autour de 3,400 dollars par habitant
2. Par ailleurs, on constate également une part prépondérante de la sphère privée dans les dépenses de santé aux Etats-Unis
    - Alors que les dépenses publiques constituent près de 85 à 90% des dépenses de santé par habitant au sein des pays de l'OCDE, les Etats-Unis se démarquent fortement avec une part de dépenses publiques de l'ordre de 50% seulement des dépenses de santé par habitant.

Bien qu'il n'y ait pas de consensus sur les raisons exactes d'une telle disparité entre les dépenses de santé aux Etats-Unis et la moyenne de l'OCDE, plusieurs leviers peuvent contribuer à en expliquer les causes :

- Le développement important des nouvelles technologies et médicaments soutenant une demande pour des traitements toujours plus coûteux
- La montée en puissance des maladies chroniques, notamment l'obésité, le diabète, ... Ces maladies conduisant à des hospitalisations et visites de praticiens répétées
- L'absence de système de Santé national, hormis pour certaines catégories de populations (Medicare, ~15% de la population, pour les plus de 65 ans et certaines catégories d'invalides ; Medicaid, ~11% de la population, pour une partie de la population à faible revenus) ; l'existence d'une Sécurité Sociale Nationale étant

souvent couplée à un contrôle important des coûts par l'Etat, ce qui tend à limiter les dépenses de santé
- La part importante des fournisseurs privés dans le parcours de soins (hôpitaux, praticiens…), sans contrôle des coûts par l'Etat

C'est ainsi une sorte de « cercle inflationniste » qui est instauré : les coûts élevés de Santé conduisent mécaniquement à un niveau de primes d'assurance Santé élevées. Et ce niveau de primes élevés prive de fait une partie de la population de la possibilité de se payer une couverture Santé.

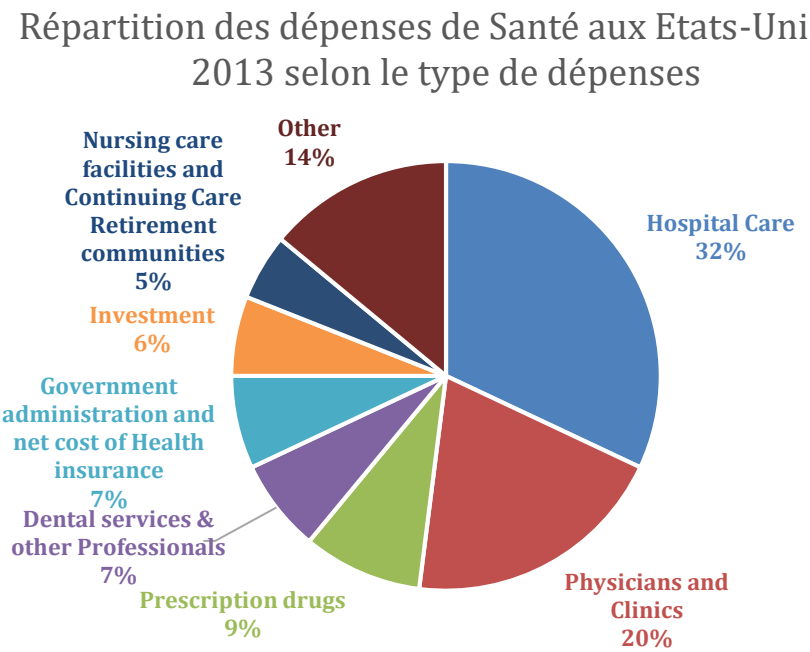
C'est dans cette perspective de réduction des coûts pour les assurés et in fine de permettre au plus grand nombre de s'offrir une couverture Santé individuelle que la réforme dite Obamacare a été pensée et introduite.

Mais avant d'en détailler plus précisément la genèse et les modalités de mise en place, il convient tout d'abord de mieux comprendre l'organisation du système de soins aux Etats-Unis.

1.2 Les différents acteurs du système de soins Américain

1.2.1 Les prestataires de soin

Le graphique ci-dessous illustre la répartition des dépenses de Santé aux Etats-Unis en 2013.



Ce graphique illustre que 68% des dépenses sont concentrées sur 4 postes :

1. Les hôpitaux (‘Hospital Care’) : 32% des dépenses de santé
2. Les médecins (‘Physician and Clinical Services’) : 20%
3. Les médicaments sur ordonnance (‘Prescription Drugs’) : 9%
4. Et les frais dentaires et ceux liés à d’autres spécialités (‘Dental services & other Professionals’) : 7%

Les hôpitaux

En 2017, les Etats-Unis comptaient 5 500 hôpitaux (pour ~900 000 lits), se répartissant comme suit :

- 4 700 “community hospitals” : ces structures qui regroupent les hôpitaux généralistes et les hôpitaux spécialisés délivrent principalement des soins de courte durée pour des pathologies graves. Un des éléments particulièrement notables réside dans la diversité des structures :
  - 59% de ces hôpitaux sont détenus par des organisations caritatives (“non profit”)
  - 21% sont détenus par des compagnies privées
  - 20% sont détenus par les Etats et gouvernements locaux.
- 400 hôpitaux psychiatriques
- 200 hôpitaux fédéraux
- 80 hôpitaux de soins de longue durée
- 10 hôpitaux au sein d’institutions : cette catégorie comprend notamment les hôpitaux au sein de prisons

Les médecins (généralistes et spécialistes) :

Les Etats-Unis comptent environ 740 000 médecins actifs, dont l’organisation du travail a été profondément modifiée au cours des dernières années :

- Quelques décennies auparavant, la grande majorité des médecins exerçaient en cabinet privé et se faisaient rémunérer à l'acte (Fee For Service). Ils pouvaient prodiguer des soins à leurs patients dans leurs cabinets et les faire admettre dans des hôpitaux où ils pouvaient également travailler personnellement. Mais la plupart des médecins ont désormais des contrats négociés avec des tiers, assureurs ou hôpitaux...
- En 2010 et pour la première fois dans l’histoire des États-Unis, le nombre de nouveaux médecins commençant à travailler dans les hôpitaux a dépassé le nombre de ceux ayant choisi de travailler en cabinet.
- En 2015, près de 57% des médecins travaillaient encore en cabinet alors que déjà 33% des médecins travaillaient directement pour un hôpital, dans une tendance qui continue à se renforcer

Dans le système de Santé Américain, un médecin en hôpital n’est ni un employé, ni le propriétaire de l’hôpital ; les médecins fonctionnant en tant qu’entités économiques indépendantes, mais cela leur permet de bénéficier des avantages liées à la mise à disposition de personnel hospitalier et de pouvoir intervenir sur des procédures particulières pratiquées presque exclusivement dans les hôpitaux. Les médecins ne paient pas les hôpitaux pour

bénéficier de leurs installations, car l'hôpital fonctionne plutôt comme « atelier » gratuit où les médecins ont accès à des ressources auxquelles ils n'auraient pas accès en cabinet.

Ce mode de fonctionnement spécifique diffère ainsi des pratiques en France, au Royaume-Uni ou en Allemagne ; en cela qu'aux États-Unis, l'hôpital n'engage pas de médecin, mais doit plutôt les « attirer ». Evidemment, sans le service d'un médecin, aucun hôpital ne peut fournir de traitement médical. Et comme les deux parties n'échangent pas directement d'argent, les hôpitaux doivent offrir aux médecins d'autres avantages pour les attirer, par exemple un environnement de haute technologie, un excellent personnel infirmier et des salles d'opération ou équipements spécifiques. Les hôpitaux ont ainsi pour objectif de se rendre plus attractifs pour les médecins, et de leur alléger la charge de la pratique médicale, tout en gérant l'optimisation et le développement de leurs bénéfices.

Les laboratoires pharmaceutiques

Les Etats-Unis sont le 1er pays au monde en termes de dépenses de médicament sur ordonnance par habitant. Le pays à lui seul représente près de 40% du marché mondial des médicaments. 7 des 15 plus importantes compagnies pharmaceutiques ou biotechnologiques ont leur siège social aux Etats-Unis. Par ailleurs l'ensemble du top 20 des sociétés pharmaceutiques y possèdent des laboratoires de recherche. Ainsi en 2007, sur les 6 500 médicaments en cours de tests cliniques dans le monde, 40% avaient été découverts aux Etats-Unis.

Les compagnies pharmaceutiques sont donc un acteur important du système de Santé aux Etats-Unis et l'absence de système d'encadrement des prix en font un marché clés pour ces acteurs internationaux.

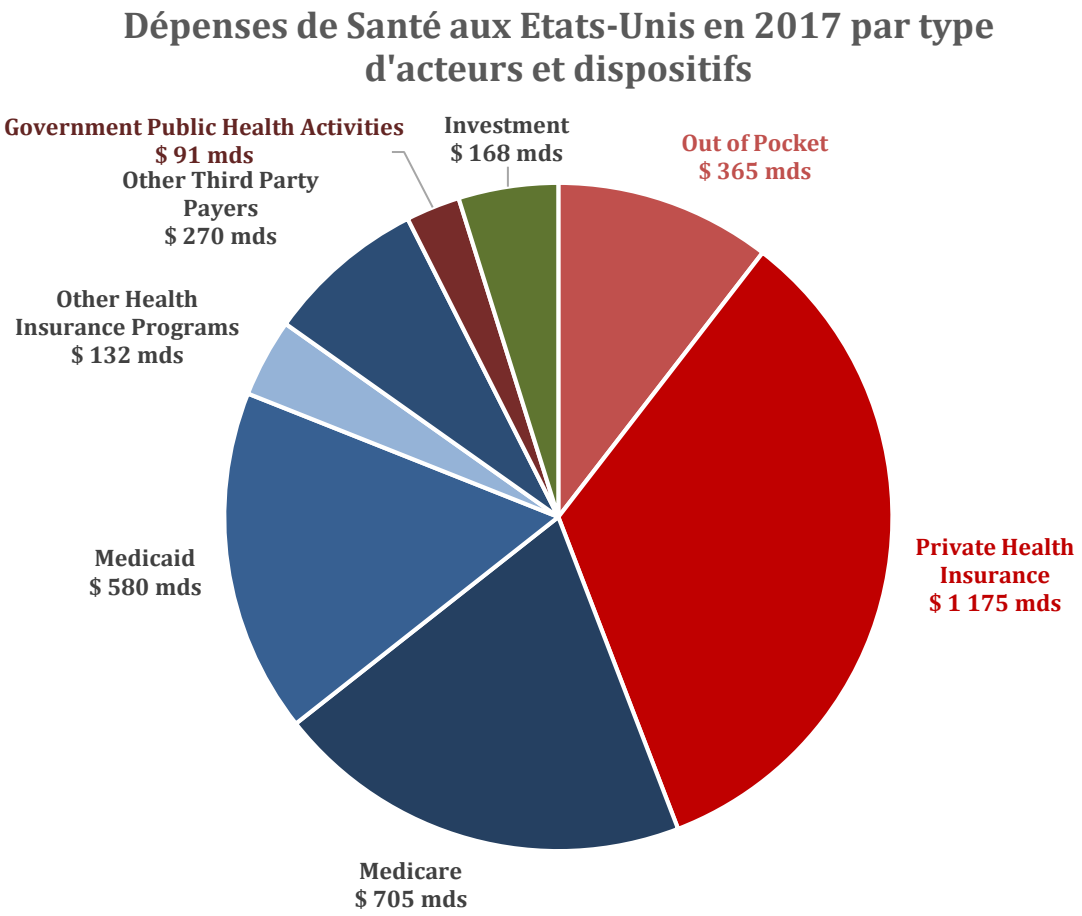
Les autres « fournisseurs »

Dans le cadre de la réglementation fédérale, un fournisseur d'acte de santé, « Health care provider » est défini comme un fournisseur de services médicaux ou de santé ou toute autre personne ou organisation qui fournit, facture ou est rémunéré pour un acte de santé. Ainsi, cette définition inclut par exemple les médecins, les hôpitaux mais également les acteurs suivants :

- 2 200 000 infirmières
- 226 000 pharmacies
- 168 000 dentistes
- 16 100 maisons de retraite

1.2.2 Le financement des dépenses de Santé

Les dépenses de Santé aux Etats-Unis sont réparties sur un nombre conséquent d'acteurs et de dispositifs :



Parmi les « payeurs » du système de Santé aux Etats-Unis, on peut ainsi distinguer 3 grandes catégories :

- Les assurances privées ('Prive Health Insurance')
- Les systèmes publiques ('Medicare' et 'Medicaid' notamment)
- Les bénéficiaires eux-mêmes ('Out Of Pocket')

Le financement des dépenses de Santé par les assureurs privés

a. Les compagnies d'assurance

Les Etats-Unis comptent près de 1 300 compagnies d'assurance Santé. Dans ce marché « hors normes », extrêmement fragmenté, certains acteurs majeurs se distinguent, avec des chiffres particulièrement impressionnants :

	Revenus - 2016	Nombre d'assurés
United Health Group	$ 185 mds	70 M d'assurés
Anthem	$ 89 mds	40 M d'assurés
Aetna	$ 63 mds	23 M d'assurés
Humana	$ 54 mds	14 M d'assurés
Cigna	$ 40 mds	15 M d'assurés

A noter qu'au sein des assureurs privés, des modèles opérationnels extrêmement distincts coexistent, mais il sortirait du champ de l'étude de les exposer ici compte tenu de l'objectif de ce mémoire.

b. Les mutuelles

Parmi les assureurs privés mutualistes, un acteur se distingue tout particulièrement : l'association « Blue Cross Blue Shield » qui est une fédération de 36 assureurs indépendants couvrant au total plus de 100 M d'assurés américains.

c. Les types d'assurance

Au-delà de l'approche traditionnelle de remboursement des frais par acte ou service, différentes approches se sont développées historiquement dans le but de réduire le coût des polices d'assurance et limiter l'inflation médicale.

Health Maintenance Organisation (HMO) : Ces organisations proposent à leurs assurés des formules prépayées au sein d'un réseau de prestataires défini, intégrant :

- Médecine de ville
- Hôpitaux
- Soins de convalescence
- etc.

Les tarifs sont négociés entre l'assureur et les établissements de soin sur la base de formule par capitation plutôt que de paiement à l'acte, à des conditions avantageuses pour l'assureur en échange de la garantie d'un nombre déterminé de patients.

Preferred Provider Organisation (PPO) : Ces dispositifs d'assurances définissent des partenaires privilégiés regroupés au sein d'un réseau. Ces partenaires sont rémunérés à l'acte comme dans les dispositifs d'assurance traditionnels (et non par capitation comme c'est le cas pour les HMO), mais en contrepartie d'un engagement de l'assureur d'orienter préférentiellement les patients vers ces prestataires, des tarifs négociés sont utilisés entre l'assureur et les prestataires.

D'un point de vue de l'assuré, le bénéficiaire a en principe le choix de ses prestataires de soins, toutefois les conditions de remboursement, franchise, etc… sont plus avantageuses s'il choisit d'être soigné au sein d'un établissement du réseau de soin (Preferred Provider)

Exclusive Provider Organisation (EPO) : Ces dispositifs d'assurance fonctionnent selon le principe du PPO mais sont beaucoup plus restrictifs. Les soins ne sont couverts par l'assureur que s'ils sont effectués dans un établissement du réseau de soin (sauf en cas d'urgence ou d'incapacité de l'assureur à proposer un praticien ou établissement dans le voisinage de l'assuré).
Ce type d'approche permet à l'assureur d'augmenter le volume de patients vers ses partenaires et ainsi d'obtenir des tarifs encore plus avantageux que dans le modèle PPO. Le fait d'avoir un réseau de soin fermé impose plus de contraintes aux assurés et leurs parcours de soin, mais ces inconvénients peuvent être compensés par les tarifs d'assurance plus compétitifs.

Point Of Service (POS) : Ce type de dispositif est un hybride entre les approches HMO et PPO. Dans une approche POS, le patient doit choisir un médecin généraliste au sein d'un réseau fermé de praticien. Ce généraliste ou médecin de famille est un passage obligé pour l'initiation de tout parcours de soin. Toutefois le généralise peut proposer soit des soins complémentaires au sein de prestataires du réseau auquel il appartient (rémunéré par capitation tout comme le ferait un HMO) soit auprès d'autres prestataires n'appartenant pas au réseau (rémunéré à l'acte).

Le financement des dépenses de Santé par les organisations gouvernementales

Le Président des Etats-Unis nomme les directeurs des agences de santé majeures à un niveau national, avec l'approbation du Sénat. Les gouverneurs jouent le même rôle pour les agences de santé des Etats.
Le « U.S. Department of Health and Human Services » (HHS) est la principale agence de santé aux Etats-Unis. Cette agence reçoit ses instructions directement du Congrès et de la Maison Blanche en ce qui concerne la gestion des finances, la coordination, la réglementation ou encore la fourniture de services de santé. Le HHS regroupe de nombreuses organisations, la plus importante d'entre elles étant le « Center for Medicare & Medicaid Services » (CMS). Le CMS est l'agence qui administre les programmes Medicare et Medicaid que nous détaillerons ultérieurement.

Le Center for Disease Control and Prevention (CDC) est une agence du HHS et la principale agence de santé publique aux Etats-Unis. Le CDC travaille en collaboration avec des organisations partenaires (institutions académiques, agences gouvernementales, organisations privées...) pour assurer le suivi, la prévention et le contrôle de la santé humaine et environnementale. Le CDC détecte les menaces sanitaires nouvelles ou émergentes, et organise la réponse à ces menaces.

Au niveau des Etats, les organisations gouvernementales jouent un rôle important dans le système de santé. Les Etats accomplissent leurs rôles par le biais de structures organisationnelles variées. La plupart des Etats disposent de départements administratifs similaires à ceux retrouvés au niveau fédéral (Division of Insurance, Department of Health…).

De façon notable, les Etats participent au financement du programme Medicaid et à sa régulation, et établissent également les règles en matière d'assurance Santé privée au sein de l'Etat.

Les Etats-Unis disposent de plusieurs programmes d'assurance public :

1. **Medicare** : ce programme instauré en 1965 par Lyndon B. Johnson est à destination des personnes de plus de 65 ans et couvrait en 2017 près de 52 M de personnes pour $ 705 Mds de dépenses. Le programme est financé via trois sources principales :
    - Les taxes fédérales payées par les citoyens américains
    - Les taxes payées par les employeurs
    - Les mensualités payées par les bénéficiaires du programme
2. **Medicaid** : ce programme à destination des personnes à faibles revenus, couvrait en 2016 près de 71 M de personnes pour près $ 580 Mds de dépenses. Tout comme Medicare, le programme Medicaid a été instauré en 1965. Il s'agit d'un programme co-financé par le gouvernement fédéral mais géré par chacun des Etats, destiné aux individus et familles aux faibles revenus. La plupart des Etats administre le programme Medicaid de façon personnalisée. Les conditions d'éligibilité au programme ne sont pas identiques selon les Etats, plus ou moins restrictives.
3. **Children's Health Insurance Program (CHIP) :** ce programme a pour but de couvrir les dépenses de santé des enfants dont les parents ont des revenus trop importants pour bénéficier du programme Medicaid mais n'ayant toutefois pas les moyens suffisants pour souscrire une assurance privée. En 2014, ce programme couvrait près de 6 M d'enfants pour un budget total d'environ $ 9 Mds

Au-delà de ces trois principaux programmes publics mentionnés ci-dessus, il existe également d'autres programmes publics mais de moindre envergure, destinés à des catégories de populations (ex. vétérans, militaires en activité, indiens d'Amérique,...).

1.2.3 Les instances législatives et exécutives des Etats-Unis

La stricte séparation des pouvoirs législatifs et exécutifs aux Etats-Unis, la diffusion des responsabilités entre le gouvernement fédéral et les états, la structure bi-camérale autonome du congrès (une chambre basse de députés et une chambre haute de sénateurs) constituent autant d'éléments qui contribuent à une grande fragmentation du pouvoir politique, complexifiant toute réforme d'envergure de politique sociale.

Cette organisation des pouvoirs aboutit à davantage de pouvoir de blocage des différentes parties prenantes et confère ainsi à chaque réforme un caractère plus largement négocié afin de prendre en compte les contraintes imposées par les différents « groupes d'intérêt »ou lobbies.

1.2.4 Les lobbies

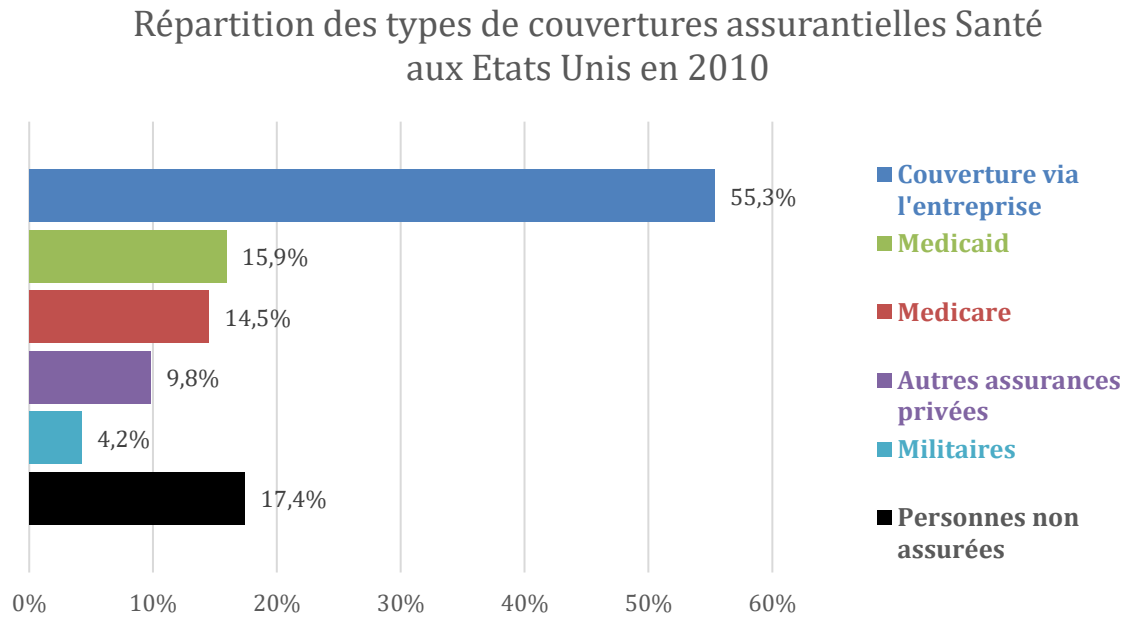
Les différents acteurs du système de Santé sont regroupés au sein de groupes d'intérêts (« lobbies ») disposant de ressources financières conséquentes et d'un important pouvoir d'influence. Parmi les principaux groupes d'influence du secteur de la santé, on peut notamment citer :

- **American Medical Association** (**AMA**) : représentation des professions médicales
- **American Hospital Association** (**AHA**) : représentation des hôpitaux et personnels hospitaliers
- **Health Insurance Association of America** (**HIAA**) et **American Health Insurance Plan** (**AHIP**) : représentation des assureurs Santé
- **Pharmaceutical Research and Manufacturers of America** (**PhRAMA**) : représentation de l'industrie pharmaceutique.

De nombreux travaux s'intéressant aux politiques de santés ont mis en avant l'influence de ces groupes de pression et expliqué les échecs de certaines réformes par les contraintes importantes que ces lobbies ont fait peser sur les décisions publiques.

1.2.5 La couverture assurantielle des américains

Une couverture assurantielle hétérogène aux Etats-Unis



En 2010 :

- 65,1% de la population disposait d'une assurance privée (dans une large majorité il s'agit d'une assurance proposée par l'employeur)
- 34,6% de la population disposait d'une assurance publique
- 17,4% de la population n'était pas couvert par une assurance Santé

1.3 Brève perspective historique de la protection sociale aux Etats-Unis

La culture libérale américaine n'a que rarement encouragé la mise en place d'un système national de Protection Sociale, privilégiant le retour au plein emploi, qui est considéré comme la meilleure des formes de protection. A titre illustratif, le président Hoover, au plus fort de la crise des années 30, indiquait « la prospérité est au coin de la rue », justifiant ainsi la non-intervention de l'état sur les problématiques sociale et économique ainsi que la confiance absolue dans le monde des affaires pour apporter prospérité et protection aux citoyens américains. Toutefois, certaines périodes de bouleversements économiques mirent à mal l'approche libérale et permirent l'introduction par "à coup" de systèmes nationaux.

Successeur du président Hoover, Theodore Roosevelt élu en 1933 lors de la Grande Dépression des années 30 impulse le "Social Security Act" (1935) offrant une pension mensuelle aux retraités, des allocations aux chômeurs, et des garanties en cas d'accident du travail aux salariés.

En 1964, le président Lyndon Johnson proclame la "Guerre contre la pauvreté" qui atteint alors 19% de la population. Le « Social Security Amendment Act » voté en 1965 permet la création de deux programmes nationaux : Medicare et Medicaid.

- Medicaid est un programme financé conjointement par l'état fédéral et les états fédérés et administrés par ces derniers. Comme présenté précédemment, il est destiné aux catégories de population les plus démunies. En termes de fonctionnement, ce sont les « Centers for Medicare and Medicaid Services » (CMS) qui supervisent les programmes dirigés par les États et établissent les besoins nécessaires à la mise en place des services, leur qualité, leur subventionnement ainsi que leurs standards d'éligibilité. Le Congrès et les « Centers for Medicare and Medicaid Services » définissent les grandes lignes et règles du programme, chaque État bénéficie ensuite de marges de manœuvre importantes dans le choix des critères d'admissibilité des bénéficiaires ainsi que dans les garanties de Santé proposées. Principalement financée par des ressources fiscales, la part financée par le gouvernement fédéral variait ainsi entre 50% et 83% du dispositif retenu.
- Medicare est un programme fédéral, placé sous l'autorité du « Department of Health and Human Services » (DHHS). Il couvre les personnes de plus de 65 ans ayant cotisé au moins 10 ans au système de Retraite publique. Depuis 1972, ce programme couvre aussi les personnes souffrant d'invalidité permanente. Originellement, le programme comportait 2 volets : Le « part A » obligatoire couvrant les hospitalisations (avec des limites de durées) et le « Part B » complémentaire couvrant la médecine de ville (docteurs, infirmiers avec un remboursement maximum de 80%). Le part A est financé par des cotisations sociales payées conjointement par l'employeur et l'employé, alimentant le "Hospital Insurance Trust Fund" placé sous l'autorité du trésor. Le part B est financé à 25% par les cotisations des bénéficiaires et 75% par des revenus fiscaux fédéraux. En 2003, Medicare est complété d'une couverture volontaire pour les médicaments Part D. Financé par des cotisations d'assurés versées au DHHS, ce troisième volet voit sa gestion déléguée aux acteurs d'Assurance privés.

Parallèlement à la mise ne place de ces 2 systèmes nationaux, de nombreuses lois fédérales et étatiques soutiennent et encadrent le développement d'Assurances Maladie privées.

Les années 1965-1980 constituent l'âge d'or du Système Public-Privé américain. A la fin des années 70, 70% des américains bénéficient d'une assurance Santé privée (très majoritairement professionnelle), Medicare et Medicaid couvrant la majeure partie du reste de la population. Seuls 12% des moins de 65 ans sont alors sans assurance.

A partir des années 1980, plusieurs facteurs mettent à mal le modèle Public-Privé. En premier lieu, les difficultés économiques conduisent à une plus forte part de chômeurs ou de populations défavorisées ne bénéficiant pas nécessairement du programme Medicaid. Deuxièmement, le coût grandissant des dépenses de santé (4-5% de croissance par an en moyenne) se répercutant directement sur les primes polices d'assurance et pesant sur les budgets des entreprises et des salariés à une période de moins grande prospérité. Ainsi les dépenses nationales de santé passent de 4% du PIB en 1950 à 17% en 2009.

Ces nombreux éléments conduisent à une réduction de la part des assurances privées et une augmentation structurelle de la part des individus non assurés (près de 3% par an) pour atteindre 17% de la population en 2010 soit environ 50 M de citoyens américains.

En 1997, le Balance Budget Act (BBA) crée le State Children Health Insurance Program à destination des enfants issus de familles à revenu modeste non éligible à Medicaid.

En 2003, le Medicare Modernization Act introduit un remboursement des médicaments devenu depuis le Medicare Part D.

Il faut attendre 2010 pour voir apparaître la première réforme structurelle de la santé aux Etats-Unis, avec " l'Affordable Care Act" ou Obamacare. Cette réforme visait à encadrer, réguler et corriger les déficiences du système de Santé. La stratégie centrale visait à instaurer une obligation individuelle d'assurance ("Individual Mandate").

1.4 Présentation de l'Affordable Care Act (ACA)

1.4.1 Points clés de la réforme

L'Affordable Care Act (souvent appelé "Obamacare") constitue une réforme sans précédent dans l'histoire de la couverture Santé des populations aux Etats-Unis. Cette loi complexe, de près d'un millier de pages, signée le 23 mars 2010 par le président Obama, a permis la mise en place de changements structurels, mais aussi l'aménagement de dispositifs existants.

Parmi les points majeurs de l'ACA, on peut mentionner tout particulièrement les points suivants :

- ***"Individual Mandate"*** (Introduction d'une obligation individuelle d'assurance) : Prenant acte du déclin depuis les années 80 de l'engagement des employeurs dans la couverture des salariés, l'ACA instaure une obligation individuelle d'être assuré

- ***"Health exchange"*** (place de marché / centrale d'achat au niveau de chaque état) : Ces dispositifs qui concernent les individus, travailleurs indépendants et entreprises de moins de 50 salariés ont pour but de renforcer la concurrence entre les différents fournisseurs d'assurance, de faciliter l'accès à l'information des assurés ainsi que de centraliser les aides possibles pour les assurés.

- **Régulation des assurances** : de nombreuses mesures mises en place ont contribué à mettre fin à des pratiques « d'évitement de risque » de la part des assureurs, ce qui les avait détournés de leur ambition sociale initiale de couverture du risque maladie. Pour cela, plusieurs mesures notables :
  - Mise en place d'un processus encadrant les augmentations de primes et imposant une justification par l'assureur des taux proposés
  - Interdiction des refus selon l'état de santé
  - Encadrement de la part des primes servant aux dépenses de santé (le rapport sinistre à primes doit être supérieur à 85% pour les polices d'assurance collectives et 80% pour les polices d'assurance individuelle)
  - Abolition de la possibilité pour les sociétés d'assurance de ne pas couvrir les maladies pré-existantes ainsi que de définir des montants maximums de remboursement par personne
  - Interdiction pour les assureurs de résilier des polices individuelles, sauf en cas de fraude

- **Subvention pour les individus :** mise en place de subvention et de réductions fiscales pour les individus se situant entre 133% et 400% du seuil de pauvreté fédérale (à titre illustratif le seuil de pauvreté fédéral est fixé à $12 140 pour un individu seul en 2018 et $20 780 pour une famille de 3 personnes)

- ***"Employer requirement"*** (contraintes pour les employeurs) : impose des taxes aux entreprises de plus de 50 salariés qui ne proposeraient pas de couverture Santé à leurs salariés bénéficiant d'aides fédérales

- **Medicaid** : extension de la couverture Medicaid à l'ensemble des individus de moins de 65 situés en deçà de 133% du seuil de pauvreté fédéral.

Il est intéressant de noter que certains des éléments clés de cette réforme n'étaient pas des idées nouvelles, mais celles-ci n'ont pu être effectivement introduites qu'au travers de la mise en place d'une réforme d'envergure.

Le principe des « Health Exchanges » s'appuie sur le principe de managed competition. Ce programme visait à réduire les dépenses de Santé en améliorant les mécanismes de concurrence entre compagnies d'assurance. Les conditions de choix des consommateurs sont renforcées par l'amélioration de leur information et de leur pouvoir de négociation, et la compétition est encouragée par la création des centrales d'achat au sein desquelles différents groupes d'assurés peuvent choisir une police d'assurance parmi un ensemble de plans. L'idée sous-jacente postule que les compagnies d'assurance, placées dans un environnement plus compétitif, vont naturellement réduire leurs coûts pour rester attractive, donc effectuer des réorganisations significatives de la manière dont les soins sont dispensés. Cette approche innovante dans les années 90 était la pierre angulaire d'un projet de réforme extrêmement ambitieux mené par l'administration Clinton en 1994-1995. Cependant, l'absence de recul ou d'expérience sur ce type de dispositif ainsi que de nombreuses autres considérations politiques firent échouer la réforme.

Apres l'échec de cette réforme au niveau fédéral, un certain nombre d'états décidèrent de lancer en 1995 leur propre "Health Exchange". Leur souhait était de démontrer qu'une voie privée était possible et pouvait corriger les défaillances du marché, mieux que toute intervention publique. Mais après plusieurs années, nombre de ces dispositifs périclitèrent. Plusieurs raisons furent mises en avant pour expliquer ces échecs :

- Les plans proposés dans ces centrales d'achat étaient d'avantage contrôlés que ceux n'y appartenant pas
- Ils impliquaient la limitation de la sélection des risques, des prestations minimales et un encadrement des tarifs des polices
- Ces Health Exchanges subissaient ainsi une concurrence 'déloyale' de la part des assureurs qui n'y participaient pas et qui offraient des tarifs plus avantageux aux personnes en bonne santé grâce à la sélection des risques et des profils de leur assuré
- Les centrales se retrouvèrent rapidement dans une spirale négative : moins d'assurés, en moins bonne santé, ce qui diminua leur pouvoir de négociation

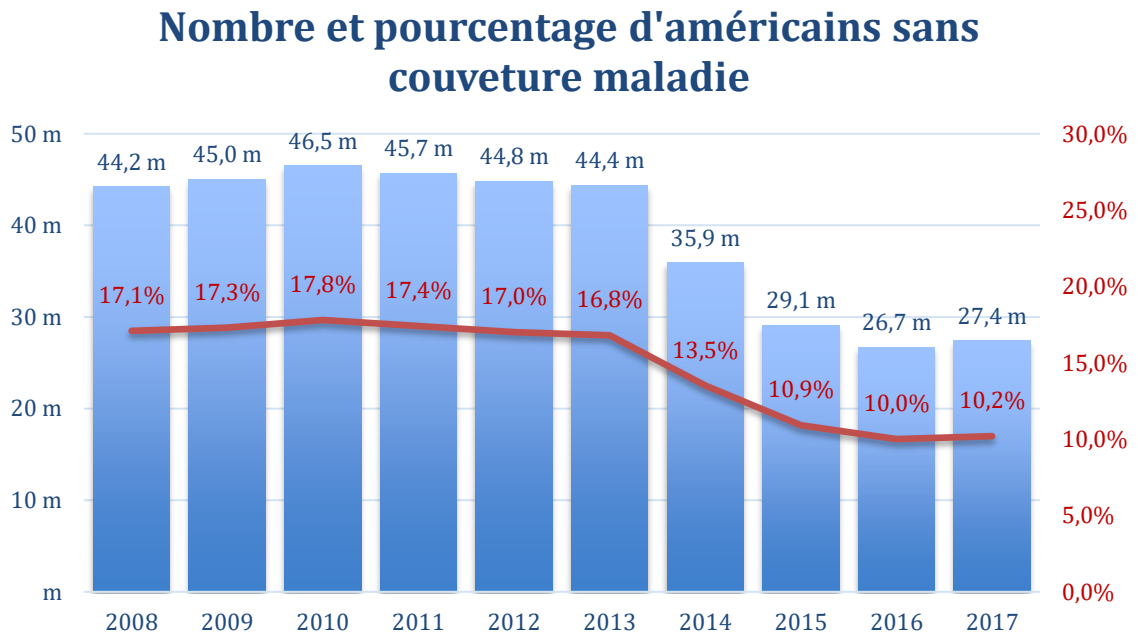
Toutefois, le succès de l'expérimentation menée dans le Massachussetts eut un impact très important sur l'Obamacare. Cet état créa en effet en 2006 un dispositif beaucoup plus ambitieux que la simple mise en place d'un Health Exchange. Il comprenait une obligation individuelle d'assurance ("Individual Mandate") et une obligation d'assurance pour les entreprises de plus de 10 salariés ("Employer Mandate"). Il mit en place des Health Exchanges appelés Health Connectors, qui offraient des polices d'assurance standardisées et limitant les possibilités de sélection des risques. Un système de subventions, pour les PME et les individus en dessous de 300 % du seuil de pauvreté fédéral, pour qui souscrivait au sein des Health Connectors devait rendre cette structure plus attractive que les assurances proposées en dehors de la centrale d'achat et ainsi éviter une concurrence déloyale.

L'idée derrière cela était la suivante : pour que les Health Exchanges fonctionnent, il fallait s'assurer que ces structures restent suffisamment attractives pour les personnes en bonne santé. Le système mis en place au Massachussetts obtint en ce sens d'excellent résultats..

De même, l'Individual Mandate est une proposition apparue à la fin des années 1980 au sein d'un think tank conservateur, l'Heritage Foundation. Cette proposition visait à ce qu'en transférant le coût et le devoir d'assurance directement aux individus, ces derniers soient plus sensibles au prix et à la qualité des soins, contraignant ainsi les fournisseurs de services de santé à être plus compétitifs et améliorer leur service.

1.4.2 Mise en œuvre de la réforme et chiffres clés

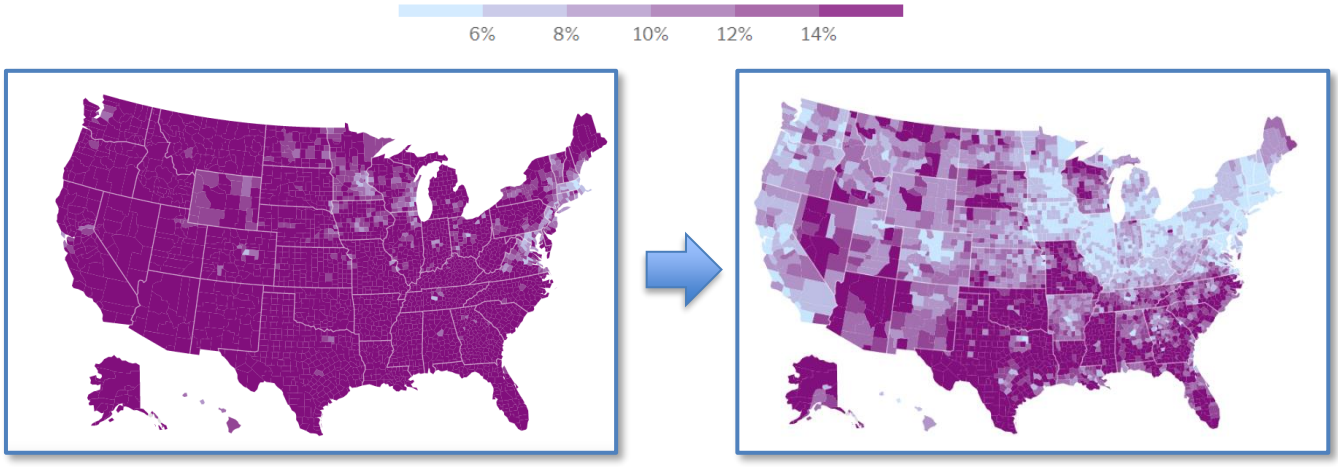
La réforme est lancée en octobre 2013 mais connaît alors de sérieuses difficultés, dont d'ordre tarifaire : 5 % des Américains bénéficiant déjà d'une assurance privée ont vu son coût augmenter, devant souscrire alors une nouvelle assurance. Après des difficultés initiales, la réforme s'est finalement révélée fructueuse, en cela qu'elle aura conduit à une baisse de 7 points de la population non-assurée :



Source : American Community Survey (ACS)

Ainsi, en mars 2015, l'administration américaine souligne une « réduction historique » du nombre d'Américains sans assurance maladie : « la proportion de non-assurés est tombée de 20,3 % à 13,2 % de la population entre le troisième trimestre 2013 et le premier trimestre 2015 ».

Evolution de la population non assurée par Etat entre 2013 et 2016



2013 **2016**

Tous les États ont vu la proportion de non-assurés baisser, même si cette baisse n'a pas été homogène. Elle a été beaucoup plus forte dans les États qui ont décidé d'étendre, dans le cadre de la réforme, l'accès au Medicaid et qui ont cherché à pousser leurs habitants non couverts à souscrire une assurance, comme par exemple la Virginie-Occidentale dont la proportion de non assurés est passée de 20 % en 2013 à 6,8 % en 2016. Au contraire, la proportion d'individus non couverts a peu baissé dans les Etats du Sud, très républicains et qui ont mis en œuvre la réforme avec beaucoup de résistance.

2. Caractéristiques des contrats Santé Obamacare

2.1 Les garanties proposées

Les polices d'assurance proposées dans le cadre de l'ACA doivent respecter un certain nombre de contraintes, tant en termes de garanties qu'en termes de couverture financière ou de partage des dépenses entre assureur et assuré.

En ce qui concerne les garanties minimums (*"Essential Health Benefit"*), l'ACA impose à toute police de couvrir sans limite annuelle ou à vie, les dix garanties suivantes :

Il est bien sûr possible aux assureurs de proposer des produits offrant plus de garanties que le minimum imposé par l'ACA, mais il n'est pas autorisé de se soustraire à une des garanties.

Notons que la couverture des garanties dentaires ou optiques pour les adultes ne fait pas partie des garanties minimum imposées par Obamacare.

Au-delà des garanties en elles-mêmes, il est intéressant de considérer la structure même des garanties proposés et les dispositifs de franchise, coassurance et co-paiement typiquement inclus dans les polices d'assurance américaines.

Le tableau page suivante reprend une synthèse de différentes offres d'assurance Santé proposés par l'assureur Molina en 2019 pour l'état du Texas :

	Choice Bronze	Choice Silver 100	Choice Silver 150	Choice Silver 200	Choice Silver 250	Choice Gold
FEATURES (INDIVIDUAL/FAMILY)						
Annual Medical Deductible	$6,400/$12,800	N/A	$750/$1,500	$3,300/$6,600	$5,350/$10,700	$2,925/$5,850
Annual Prescription Drug Deductible	Included in Medical deductible	N/A	N/A	$400/$800	$400/$800	N/A
Annual Out-of-Pocket Max	$7,900/$15,800	$1,400/$2,800	$2,600/$5,200	$6,300/$12,600	$7,900/$15,800	$5,000/$10,000
BENEFITS[1]						
Emergency Room[2]	40% ▲	10%	20% ▲	30% ▲	30% ▲	20% ▲
Urgent Care	$75	$10	$20	$50	$50	$35
PCP Office Visit	$35	No Charge	$10	$20	$30	$10
Mental Health Services, Outpatient	$35	No Charge	$10	$20	$30	$10
Substance Abuse Services, Outpatient	$35	No Charge	$10	$20	$30	$10
Specialist Office Visit	$80 ▲	$15	$30	$60	$75	$50
Habilitative Services	40% ▲	$15	$30	$60	$75	$50
Rehabilitative Services	40% ▲	$15	$30	$60	$75	$50
Outpatient Surgery	40% ▲	10%	20% ▲	30% ▲	30% ▲	20% ▲
X-rays	$80 ▲	$10	$30	$65	$75	$35
Lab Tests	$40 ▲	$10	$10	$40	$40	$15
Inpatient Hospital Services	40% ▲	10%	20% ▲	30% ▲	30% ▲	20% ▲
Maternity Care	40% ▲	10%	20% ▲	30% ▲	30% ▲	20% ▲
Tier-1 Lower-Cost Generic and Brand Name Drugs[3]	$20	$2	$5	$10	$20	$10
Tier-2 Preferred Generic and Brand Name Drugs[3]	40% ▲	$15	$30	$60	$60	$50
Tier-3 Non-Preferred Brand Name Drugs[3]	50% ▲	20%	30%	40% ▲	40% ▲	30%
Tier-4 Generic and Brand Name Specialty Drugs[3]	50% ▲	20%	30%	40% ▲	40% ▲	30%

KEY: Co-pay | Coinsurance | ▲ Deductible applies | See back cover for details and descriptions.

Source : Molina Health Care

2.2 Les différentes modalités de financement / remboursement

Le tableau présenté précédemment illustre la structure des polices d'assurance Santé aux Etats-Unis. On y note en particulier différentes modalités de financement / remboursement :

- **Les Franchises *"Deductible"*** : ces franchises indiquent le seuil de dépenses à partir duquel l'assuré peut faire appel à l'assureur pour rembourser ses soins.
    - Ainsi pour le premier produit proposé dans le tableau (première colonne "Bronze"), on observe une franchise globale annuelle de $6,400 pour un individu ($12,800 pour une famille)
    - Cette franchise s'applique à la majeure partie des couvertures proposées (comme l'illustre le petit triangle indiqué sur certaines garanties) mais pas à toutes les couvertures proposées
    - Une visite auprès d'un généralise (Primary Care Phisician (PCP) office visit) est ainsi possible pour le produit Bronze sans avoir recours à la franchise annuelle

- **Les Co-assurances (*"Co-insurance"*)** : Dans ce dispositif, l'assuré paye un pourcentage de la dépense totale, quelle qu'elle soit.
    - Dans le produit Bronze (première colonne), l'assuré doit payer lui-même 40% des dépenses de maternité, l'assureur prenant en charge les 60% restants.

- **Les Co paiement *'"Co payment"*** : Dans ce dispositif, l'assuré doit s'acquitter d'un montant fixe par acte, l'assureur prenant en charge le reste des dépenses, sans plafond
    - A titre d'illustration pour le produit Bronze (première colonne), l'assuré doit régler un forfait de $40 pour tout test de laboratoire
    - Il faut préciser que ce montant fixe doit être réglé à chaque acte. Ainsi pour tout nouvel épisode où le patient se verrait prescrire de nouveaux tests laboratoires, le patient devra à nouveau régler la somme de $40.

- **Les plafonds annuels de dépenses "annuals Out of Pocket Maximum"** : Ce dispositif la correspond à la dépense annuelle maximum que l'assuré devra supporter. Au-delà de ce plafond, l'intégralité des dépenses passe à la charge de l'assureur et les dispositifs de Co assurance et Co payment ne s'appliquent plus.

On comprend ainsi la diversité offerte aux assureurs quant aux possibilités de définir des offres spécifiques. Aussi bien en termes de garanties qu'en termes de structures de garanties, la comparaison des offres au sein d'un même assureur et entre assureurs est extrêmement complexe. L'avantage d'une telle diversité est de pouvoir permettre aux populations cherchant à s'assurer de choisir l'offre qui leur sera adaptée, certes au détriment d'une lisibilité globale des différentes polices proposées.

Afin de remédier à cette problématique de comparaison des offres entre elles et d'information du consommateur sur la qualité de l'offre souscrite, une mesure objective obligatoire permettant de comparer les offres a été introduite : la valeur actuarielle (Actuarial Value), que nous détaillons ci-dessous.

2.3. Contraintes réglementaires et impact sur la tarification

2.2.1 L'Actuarial Value

La valeur actuarielle introduite par l'Obamacare est une mesure du pourcentage des dépenses de Santé couvertes par la police d'assurance souscrite.

A partir de 2014, toute police d'assurance Santé individuelle, ou à destination des petites entreprises, commercialisée dans un *market exchange* doit avoir un des 4 niveaux possibles de valeur actuarielle (aussi qualifié de « niveau métallique ») :


60% - police "Bronze"


70% - police "Silver" (Argent)


80% - police "Gold" (Or)


90% - police "Platinum" (Platine)

Cette classification ne s'applique toutefois pas dans quelques cas, par exemple pour les personnes disposant de couvertures d'assurance avant Obamacare.

- En effet, dans certains cas ces personnes ont pu conserver les assurances en place, bien que celles-ci ne satisfassent pas l'intégralité des critères imposé par l'ACA. Cette catégorie est définie comme les "grandfathered plans"
- Ou bien pour les individus de moins de 30 ou les populations pour lesquelles la souscription d'une couverture santé les placeraient dans une difficulté financière. Ces 2 populations peuvent souscrire un "Catastrophic plan" (couverture catastrophique) à un prix moins élevé mais ne couvrant que les sinistres exceptionnels.

Comme il peut être difficile pour un assureur de concevoir une police d'assurance correspondant exactement à un des niveaux métalliques imposés, l'Etat Fédéral publie chaque année les tolérances possibles par niveau. A titre d'exemple en 2014, les polices de niveau "Bronze" devaient avoir une valeur actuarielle entre 58% et 62%, en 2018 cette plage a été étendue à une fourchette comprise entre 56% et 65%.

Le calculateur de valeur actuarielle est défini au niveau fédéral et la méthodologie ainsi que l'outil est mis à jour chaque année.

Eléments pris en compte dans le calcul de l'Actuarial Value

Le calcul de l'Actuarial Value prend en compte uniquement les garanties identifiées comme obligatoire selon l'ACA. Il est possible pour un assureur d'offrir des garanties complémentaires à celle imposées par l'ACA, mais ces garanties ne seront pas prises en compte dans le calcul de la valeur actuarielle.

De plus, dans le cas où la police offre des niveaux de garantie distincts selon que les soins se font au sein du réseau partenaire ou hors réseau, seules les garanties proposées au sein du réseau partenaire sont prises en comptes dans le calcul de l'Actuarial Value

Enfin ce calcul de l'Actuarial Value est conditionné par les différents niveaux de garantie proposés et les éléments externes pouvant réduire les dépenses de l'assuré :

1. Le niveau métallique visé
2. Le montant de dépense moyen pour l'intégralité des assurés pour un niveau métallique donné. Ce montant est indiqué dans des tables fournies avec le calculateur de valeur actuarielle et correspond au dénominateur du calcul de valeur actuarielle.
3. Le montant de dépenses de santé couvertes par l'employeur au travers d'un Health Saving Account ou Health Reimbursement Account. Ces mécanismes proposés par l'employeur sont des comptes sur lesquels les sommes déposées doivent servir à couvrir les dépenses de santé. Le montant déposé sur ces comptes intervient au numérateur du calcul de l'Actuarial Value, comme des dépenses de santé prises en charge au premier dollar par l'assureur.
4. Le calcul des dépenses de santé par acte en-deçà de la franchise (le calcul tient compte de niveau de co-assurance, co-paiement, etc…). L'ensemble des dépenses couvertes par l'assureur en-deçà de la franchise sont intégrées au numérateur de l'Actuarial Value.
5. La détermination du montant de dépenses correspondant au plafond annuel de dépense.
6. Le calcul des dépenses de santé couvertes par la police entre les montants de franchise et le plafond maximum par assuré par an.
7. Le calcul des dépenses de santé couvertes au-delà du plafond annuel de dépenses par an
8. La prise en compte de l'impact des différents réseaux de soins dans le cadre d'offre spécifiant des dépenses prises en compte par type de réseau
9. Le calcul de la valeur actuarielle et du niveau métallique correspondant

2.2.2 Autres éléments : risk adjustment et « filling »

« Filing » validation des prix auprès du régulateur

En contrepartie de l'obligation pour les individus de s'assurer, l'ACA impose un contrôle des prix des différentes polices d'assurance au niveau de chaque état. Ainsi chaque année, tout assureur désirant distribuer des polices d'assurance dans un état doit soumettre au régulateur de cet état la grille tarifaire qu'il souhaite appliquer. En pratique, dès mars de l'année N-1 (N étant l'année de distribution d'une police donnée), les assureurs préparent les grilles de garantie et de tarifs des contrats qu'ils souhaitent distribuer. Ces grilles tarifaires ainsi qu'un mémorandum contenant des justifications quantitatives et qualitatives sont soumis en Juin de l'année N-1, puis revues par le régulateur de chaque état. Selon le niveau d'interventionnisme du régulateur (qui peut varier d'un état à un autre), les régulateurs peuvent décider de valider, commenter ou rejeter les propositions des assureurs. L'objectif du régulateur est double, d'une part s'assurer que les prix et augmentations tarifaires proposés sont « justes » pour le consommateur, d'autre part protéger les assureurs en contrôlant que les prix proposés ne mettent pas en danger leur solvabilité.

Ci-dessous à titre d'illustration la table des matières d'un mémorandum soumis au régulateur pour justifier les grilles tarifaires appliquées :

TABLE OF CONTENTS

ATTACHMENTS	4
GENERAL INFORMATION	5
PART III ACTUARIAL MEMORANDUM	6
PROPOSED RATE INCREASE(S)	6
EXPERIENCE PERIOD PREMIUM AND CLAIMS	9
BENEFIT CATEGORIES	10
PROJECTION FACTORS	10
CREDIBILITY MANUAL RATE DEVELOPMENT	12
CREDIBILITY OF EXPERIENCE	12
PAID TO ALLOWED RATIO	12
RISK ADJUSTMENT AND REINSURANCE	13
NON-BENEFIT EXPENSES AND PROFIT & RISK	15
PROJECTED LOSS RATIO	16
SINGLE RISK POOL	17
INDEX RATE	17
MARKET ADJUSTED INDEX RATES	18
PLAN ADJUSTED INDEX RATES	18
CALIBRATION	19
CONSUMER ADJUSTED PREMIUM RATE DEVELOPMENT	21
AV METAL VALUES	23
AV PRICING VALUES	23
MEMBERSHIP PROJECTIONS	24
TERMINATED PRODUCTS	24
PLAN TYPE	25
WARNING ALERTS	25
EFFECTIVE RATE REVIEW INFORMATION	26

Figure : Table des matières d'un « Actuarial memorandum »

Ci-dessous également un exemple des hypothèses de dérive annuelle par grand poste soumis par un assureur.

Annual Trend Assumptions		
	Utilization	Unit Cost
Inpatient Facility	0.0%	4.2%
Outpatient Facility	1.5%	3.6%
Professional	1.5%	0.0%
Other Medical	1.2%	0.0%
Prescription Drug	1.5%	7.5%

Figure : Illustration des hypothèses de dérive soumis par un assureur au régulateur

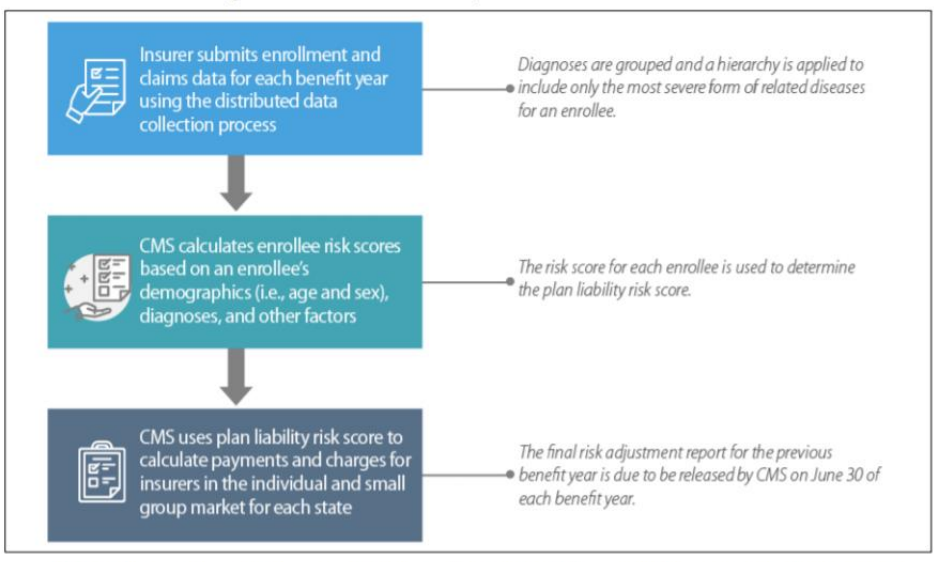
« Risk adjustment » ou mécanisme de péréquation

L'ACA interdit aux assureurs de mettre en place des dispositifs de sélection du risque ou de limiter les garanties en fonction de maladies préexistantes. Il est également impossible pour un assureur de résilier une police individuelle ou de procéder à des majorations ciblées sur une police en particulier. Ces contraintes peuvent introduire des déséquilibres importants entre assureurs en fonction du profil des populations ayant souscrit chez eux. Afin de lisser ces inégalités potentielles, un mécanisme de péréquation appelé « Risk Adjustment » a été mis en place.

Grâce à ce mécanisme un assureur ayant une population plus à risque (ex : maladies chroniques, cancers,…) recevra une partie des primes d'un assureur ayant une population en meilleurs santé (ex : population jeune sans maladie chronique). Ce dispositif est donc neutre au global.

	Risk Adjustment
What the program does	Redistributes funds from plans with lower-risk enrollees to plans with higher-risk enrollees
Why it was enacted	Protects against adverse selection and risk selection in the individual and small group markets, inside and outside the exchanges by spreading financial risk across the markets
Who participates	Non-grandfathered individual and small group market plans, both inside and outside of the exchanges
How it works	Plans' average actuarial risk will be determined based on enrollees' individual risk scores. Plans with lower actuarial risk will make payments to higher risk plans. Payments net to zero.
When it goes into effect	2014, onward (Permanent)

Figure : description du mécanisme de risk adjustment



Source: CRS-developed flow chart of the permanent risk adjustment program based on information from CMS.
Note: CMS denotes Centers for Medicare & Medicaid Services.

Figure : processus de soumission des informations permettant le calcul du « Risk adjustment »

2.2.3 Critères utilisables pour la tarification

In fine, seuls 5 facteurs peuvent être pris en compte pour déterminer le niveau des primes au sein d'Obamacare

- **Le lieu de résidence :** les états définissent des zones délimitant des marchés locaux (« rating areas »). Les garanties peuvent différer d'un état à l'autre, et la morbidité de la population ainsi que le coût des services d'une « rating area » à l'autre.
- **L'âge** : la prime des assurés les plus âgés est au maximum le triple de celle des plus jeunes (21 ans)
- **La population assurée sur le contrat :** Personne seule vs famille
- **Fumeur vs. non-fumeur :** les assureurs peuvent augmenter les primes de 50% pour les fumeurs
- **Le « niveau métallique » du plan :** niveau de couverture

3. Portefeuille considéré pour l'étude

3.1 Présentation des données

Nous disposons de plusieurs bases de données d'expériences détaillées sur la période de 2016 à 2018, relatives à l'exposition ainsi qu'à la sinistralité observée :

- 1 base de données sur les personnes assurées (l'exposition)
    - Informations sur les populations assurées entre 2016 à 2018, comprenant 64 variables pour chaque observation (informations sur les personnes couvertes et les produits souscrits mois par mois), avec 1,4m d'observations en 2016, 1,7m en 2017 et 2,7m en 2018.
- 3 bases de données de sinistres
    - 3 bases sur les sinistres médicaux (hors pharmacie), pour les mêmes années 2016 à 2018, comprenant 160 variables pour chaque observation avec respectivement 3,2, 2,3 et 2,7m d'observations
    - 1 base pour les sinistres pharmacie de 2016 à 2018, comprenant 98 variables pour chaque observation, et avec 2,7m d'observations au total
- 1 base de données sur les primes payées par les clients
    - Reprenant les données de cotisation de 2016 à 2018, avec 16 variables retraçant l'historique de primes mensuelles de chaque assuré avec 3,7m d'observations

3.1.1 Hypothèses et traitements effectués

Les bases de données sont issues des systèmes de gestion du partenaire, et peuvent ainsi comporter certaines anomalies / limites, comme par exemple de possibles erreurs humaines lors de la saisie d'information par le gestionnaire, les données manquantes ou la possible perte d'informations liée à des mises à jour ou non d'informations propres aux assurés.

Néanmoins, les analyses menées n'ont pas mis en évidence de manques manifestes de données et celles-ci ont ainsi été réputées fiables. Mais les données ont été expurgées de certains changements de situation qui auront eu lieu en cours d'année. Par exemple, 256 assurés (0,08%) ont changé d'Etat au cours des 3 années observées ; et 4 820 assurés ont changé de niveau métallique. Ces cas-là ont ainsi été censurés des données exploitées afin de ne pas biaiser les résultats.

3.1.2 Les bases exploitables

Les bases de données ont été organisées selon le schéma suivant :

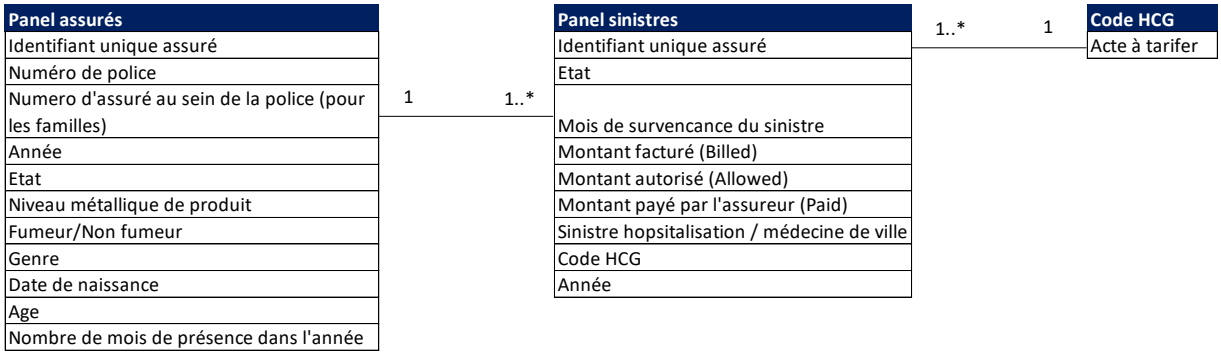


Figure : modèle de données utilisé

Pour le panel des assurés, nous avons réduit la base initiale des expositions pour ne conserver que les champs que nous retenons ensuite pour l'étude. Une ligne dans cette base correspond à un bénéficiaire. A un même bénéficiaire peuvent correspondre 3 lignes correspondant aux années d'exposition. La présence de l'assuré au cours d'une année est capturée par la variable 'nombres de mois de présence dans l'année'.

Pour le panel des sinistres, nous avons réduit la base initiale des sinistres pour ne conserver que les champs que nous retenons pour l'étude. Une ligne dans cette base correspond à un sinistre. Chaque sinistre est associé de manière unique à un assuré à travers la variable 'identifiant unique assuré'. Chaque assuré peut ainsi avoir plusieurs sinistres.

Enfin comme nous l'expliciterons plus loin dans la partie application, nous avons construit une table de correspondance afin d'associer chaque code HCG (Health Cost Guidelines : table de classification des actes de soins comprenant plus d'une centaine de catégorie) à 14 actes à tarifer. Ainsi à plusieurs code HCG peuvent correspondre un acte ou garantie à tarifer.

3.2. Statistiques descriptives

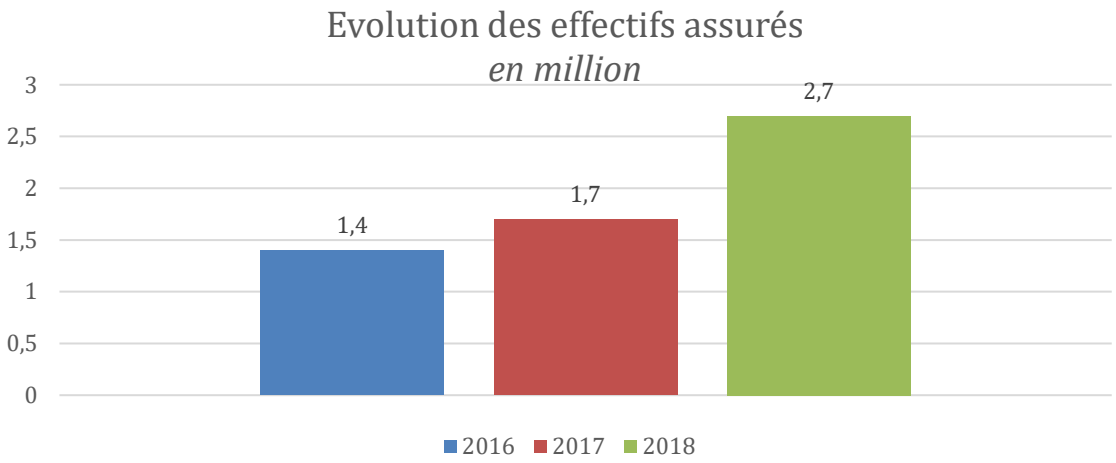
3.2.1 Populations assurées

Nous disposons de bases de données concernant des assurés sur les années 2016, 2017, 2018 sur 6 Etats :

- La Californie (CA)
- Le New Jersey (NJ)
- L'état de New York (NY)
- L'Ohio (OH)
- Le Tennessee (TN)
- Le Texas (TX)

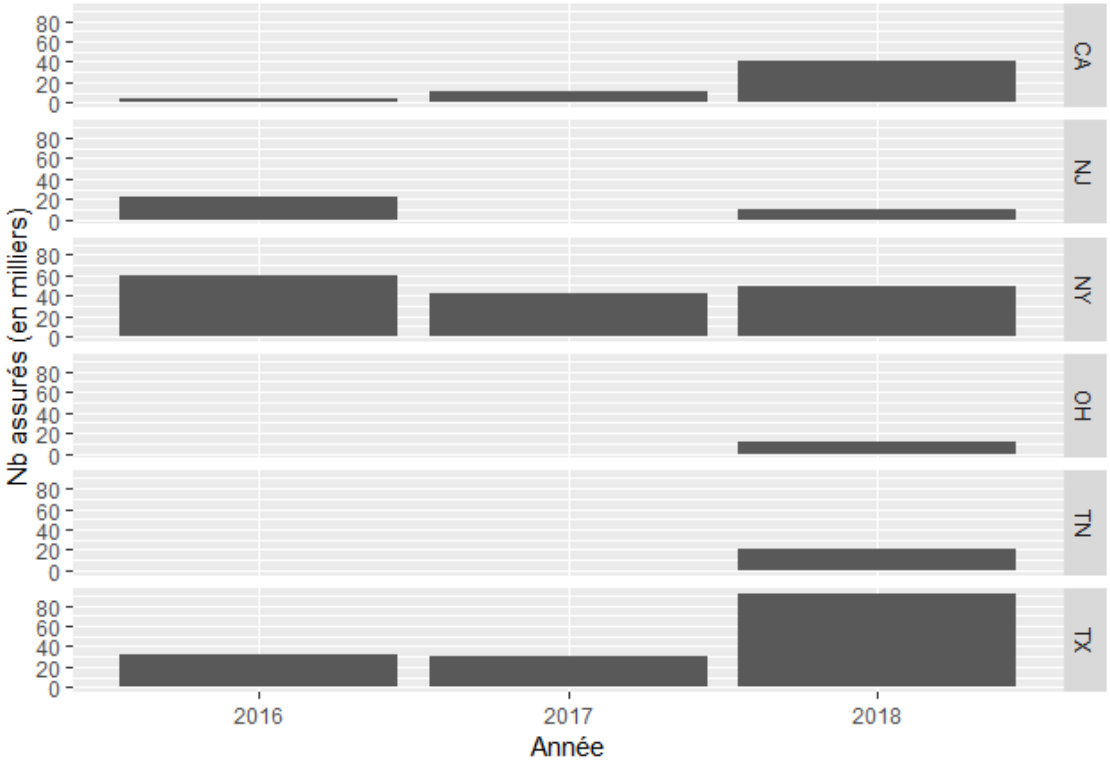
Effectifs assurés

Le graphique ci-dessous illustre l'évolution du nombre des effectifs x mois assurés sur la période d'observation :



A noter qu'il faut entendre par « effectifs assurés », la somme des présences associées à chaque bénéficiaire au cours des années d'observations : par exemple, un assuré présent 6 mois dans une année comptera pour 6 dans le décompte relatif à l'année en question dans le graphique ci-dessus.

Cette évolution s'explique lorsque nous observons l'évolution des effectifs assurés par état, illustrée ci-après



Graphique : Base des assurés par état et par année

En particulier nous constatons que le portefeuille mis à disposition a connu des évolutions diverses :

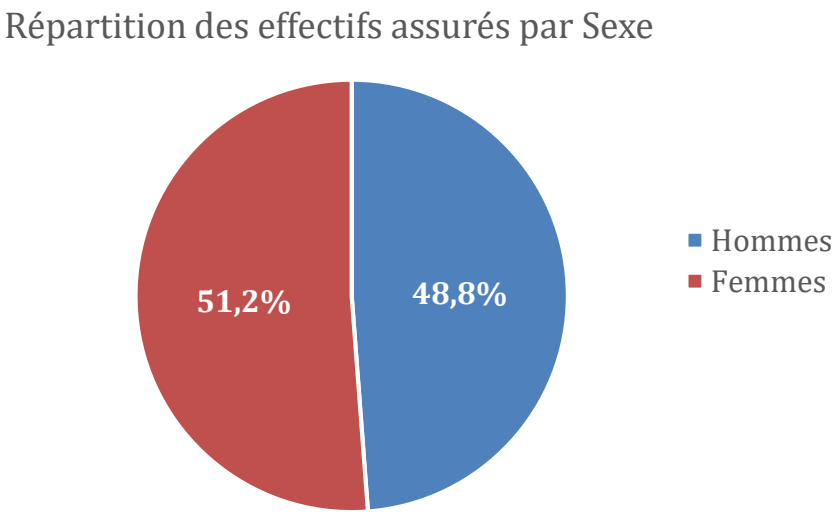
- Des états dont la population assurée a cru :
  - En Californie (CA), passage de 4K assurés en 2016 à 40K en 2018
  - Au Texas (TX), passage de 32 K assurés en 2016 à 92K assurés en 2018
- Des états dont l'activité a été lancée en 2018
  - Le Tennessee (TN), avec 21K assurés en 2018
  - L'Ohio (OH), avec 11K assurés en 2018
- Des états dont la population assurée a décru
  - A New-York (NY), passage de 60K assurés en 2016 à 49K assurés en 2018
  - Dans le New Jersey (NJ), 23K assurés en 2016, puis une sortie de l'état en 2017 et un retour en 2018 avec 10K assurés

Il est à noter que les statistiques relatives à cette année ne sont pas complètes car ayant été reçues en Décembre 2018. L'exposition (et la sinistralité) associées étant ainsi censurées sur la fin de l'année 2018, c'est toute l'année 2018 qui sera exclue par la suite pour notre étude de tarification.

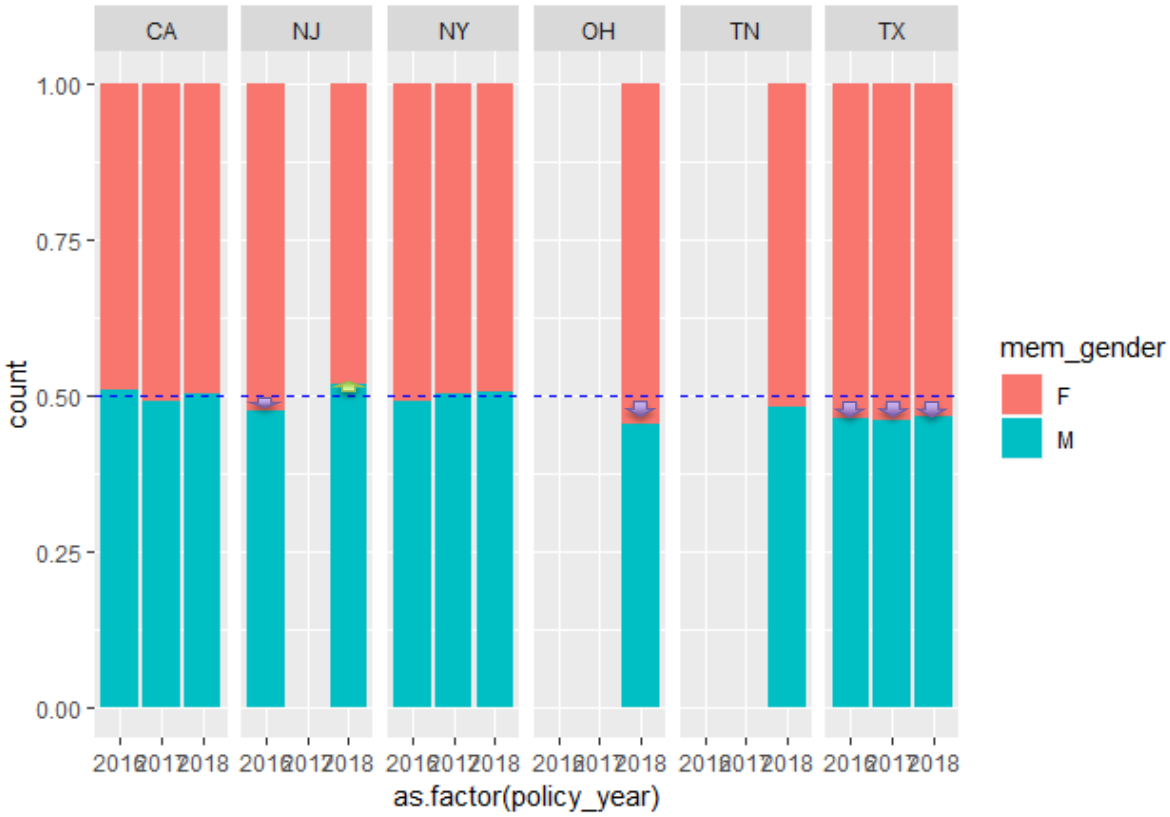
Compte tenu de notre besoin d'historique pour cette étude, nous retiendrons les années 2016 et 2017 ; et nous restreindrons l'analyse sur les 4 états suivants : New York, Texas, Californie et le New Jersey.

Répartition Hommes/Femmes

En termes de répartition des effectifs assurés par sexe, le diagramme en secteur présenté ci-dessous illustre une très légère sur-représentation des femmes par rapport aux hommes, avec 51,2% des effectifs.



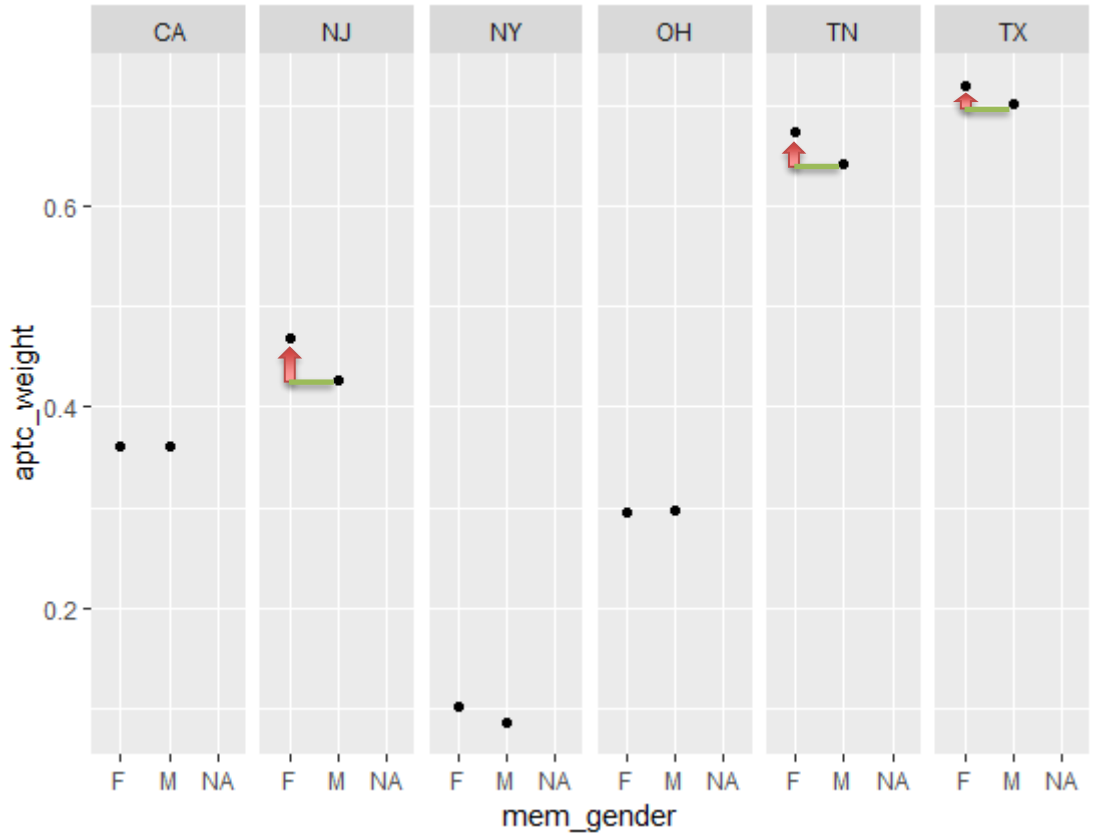
De manière plus précise, l'histogramme ci-dessous représente la répartition homme/femme par état au cours du temps et pour chaque état :



On observe que pour les états de Californie, New Jersey, New York et Tennessee, le mix homme /femme est relativement équilibré et stable au cours du temps.

Toutefois pour les états Ohio et du Texas, on constate une proportion plus importante de femmes dans les portefeuilles. Cette singularité pourrait être expliquée par les mécanismes d'aide mis en place au niveau de ces états.

Mais cette explication semble contredite lorsque l'on observe le graphique ci-dessous qui représente le pourcentage de subvention ("Advanced Premium Tax Credit" ou APTC) par état et par sexe.



Graphique : pourcentage de subvention dans la prime totale par sexe pour chaque état

On observe bien sur le Texas un part de subvention ("Advanced Premium Tax Credit" ou APTC) plus importante chez les femmes que chez les hommes. L'hypothèse sous-jacente étant que les femmes, plus favorisées financièrement, souscrivent plus fortement ce produit. De plus, le niveau de subvention élevé (> 60% de la prime) tend à indiquer des populations plutôt défavorisées et donc a priori plus sensibles au prix. Un constat similaire pourrait être dressé dans le Tennessee, où la proportion de subvention est plus importante pour les femmes que pour les hommes, et on observe dans le portefeuille assuré une proportion également plus importante de femmes.

Mais au contraire, dans l'Ohio où les pourcentages de subvention sont identiques entre hommes et femmes, on observe dans le portefeuille une proportion plus importante de

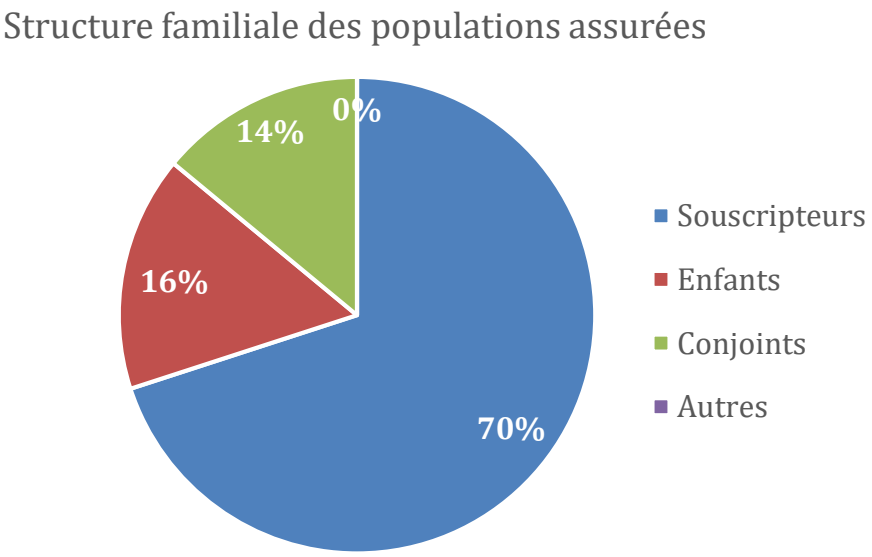
femmes. Mais le plus faible nombre de personnes assurées dans cet état (11K assurés en 2018) pourrait représenter un biais statistique. Des conclusions similaires pourraient être tirées pour le New Jersey.

On observe bien un équilibre Homme Femme dans le portefeuille significatif en Californie, ce qui est bien en ligne avec un niveau de subvention observé dans cet état dans le graphique ci-dessus.

Afin de confirmer cette hypothèse, nous pourrions vérifier si dans certains états le niveau de subvention dépend effectivement du sexe.

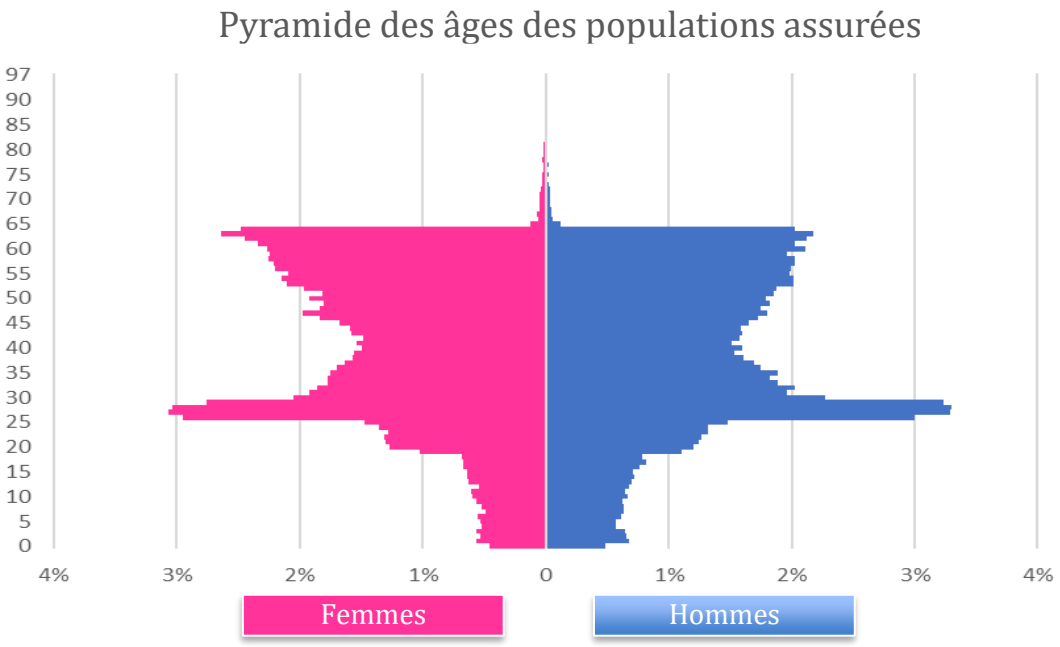
Composition familiale assurée

La composition familiale des populations assurées a également été analysée et est restituée dans le diagramme en secteur ci-dessous :



Très majoritairement, les contrats étudiés couvrent les souscripteurs (70% des effectifs assurés), mais également leurs enfants (16%) et leurs conjoints (un peu moins de 14%).

Profils d'âge



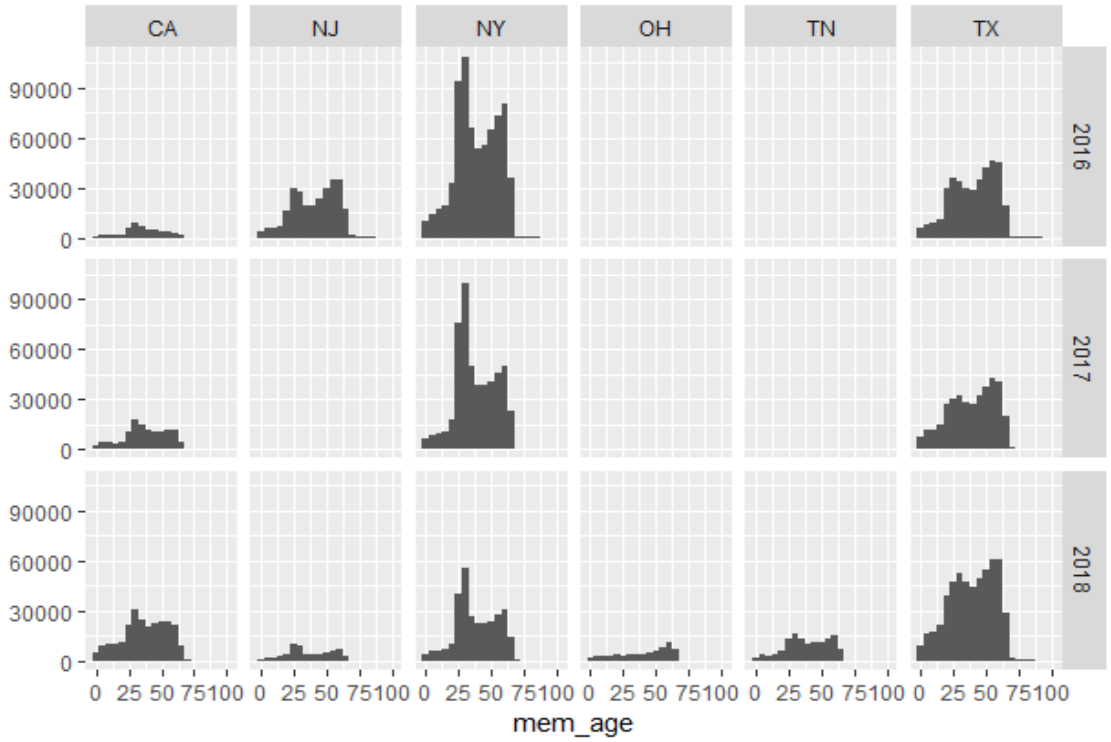
On constate des répartitions d'effectifs par âge avec les caractéristiques suivantes :

- une queue de distribution dans les âges jeunes inférieurs à 20 ans
- un « saut » autour de 25 ans
- un creux autour de 37 ans
- une chute abrupte à partir de 65 ans

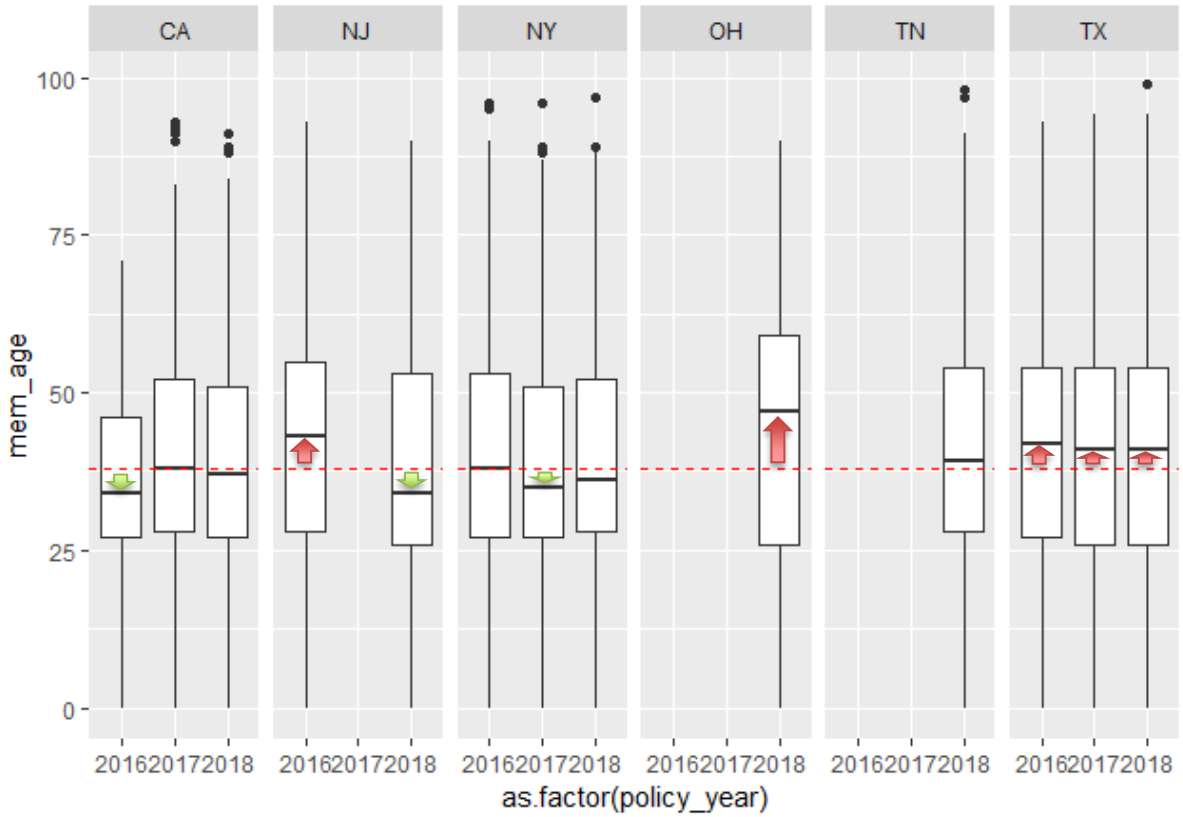
Plusieurs facteurs permettent d'expliquer un tel profil. En particulier :

- Les effectifs en dessous de 20 ans correspondent aux enfants des assurés principaux
- Le pic autour de 25 ans correspond à la fin des études et l'entrée dans la vie active, cette tranche de la population jeune ne pouvant plus être couverte par les polices d'assurance de leurs parents, mais n'étant pas nécessairement déjà couverte par une police d'assurance fournie par un employeur, et avec l'obligation de s'assurer du fait de l'*Individual Mandate* (obligation individuelle d'assurance) imposée par l'Obamacare
    - A titre d'illustration le taux de chômage en janvier 2019 chez les 20-24 ans était selon le *Bureau of Labor Statistic* de 7,6% contre 3,3% pour la tranche plus de 25 ans, avec un minimum atteint sur la tranche 35-44 ans à 2,9%.
- Les populations entre 30 et 65 ans sont le plus souvent actives et couvertes par des polices d'assurances de groupes souscrites par leur employeur.
- Les populations de plus de 65 ans sont en principe couvertes au travers du système de santé publique Medicare, ceci qui explique la chute abrupte constatée à partir de cet âge.

Ce « profil » de pyramide des âges se retrouve globalement dans chacun des états de l'étude :



En termes de profil d'âge des effectifs assurés, il est intéressant de noter des différences importantes entre les populations assurées selon les états, ainsi que l'évolution au sein d'un même état au cours du temps. L'analyse en box plot ('diagramme en boîte') ci-dessous illustre la distribution des âges par état sur les 3 années en base. La ligne rouge pointillée indique l'âge médian aux Etats-Unis (38 ans) en 2017.



Graphique : diagramme en boîte des distributions d'âges des assurés par états au cours du temps.

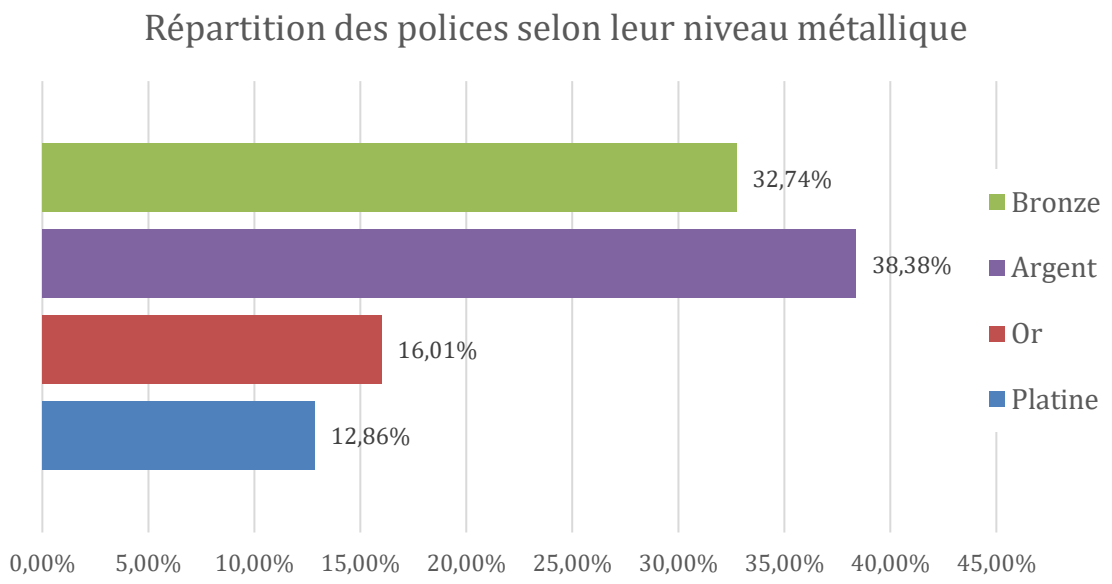
Plusieurs éléments sont à noter :

- Sur les états de Californie et New York, l'âge médian du portefeuille considéré est inférieur à l'âge médian du portefeuille observé
- Les effectifs assurés dans le Texas ont un âge médian supérieur à l'âge médian à la médiane du portefeuille étudié
- Sur l'état du New Jersey, l'âge médian du portefeuille considéré est supérieur en 2016 puis inférieur en 2018 à l'âge médian aux Etats-Unis. Ceci pourrait s'expliquer par une évolution tarifaire (relative à la concurrence) qui aurait rendu les polices plus attractives en 2016 pour les populations plus âgées, et a contrario pour les populations plus jeunes en 2018.

Il n'est pas surprenant d'observer ces différences de positionnement selon les états ou l'année d'observation. En effet, les produits proposés par les différents assureurs étant enregistrés à l'échelle d'un comté (ou "rating area"), l'environnement concurrentiel à la fois en termes de nombres d'assureurs présent dans chaque zone qu'en termes de prix proposés par ces assureurs peut varier à l'échelle du comté.
De plus, comme les prix et les assureurs présents sont décidés à l'échelle d'une année, chaque année pour un même comté peut donner lieu à un bouleversement de l'équilibre concurrentiel.
Ces évolutions potentiellement importantes de l'environnement concurrentiel, permettent ainsi d'expliquer les évolutions constatées sur le profil des assurés couvert par un assureur donné.

Niveau métallique

La répartition des polices par niveau métallique a été analysée et est restituée dans le graphique ci-dessous :

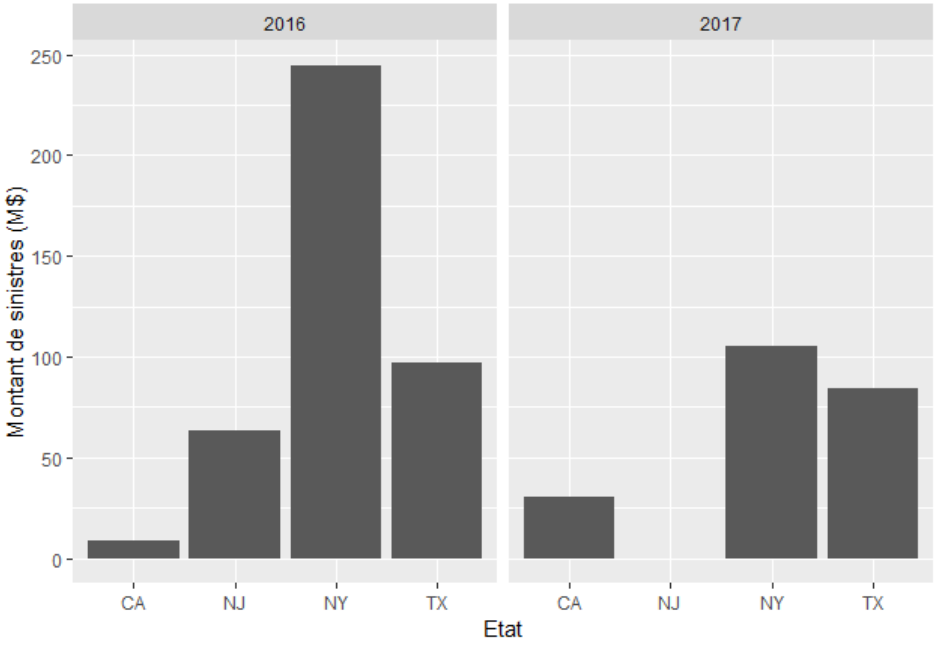


On observe ainsi que c'est le niveau métallique « Argent » qui est le plus souscrit (près de 38% des contrats), devant le niveau bronze à près de 33%. Les contrats « Or » et « Platine », offrant les niveaux les plus élevés de remboursement mais également les plus chers sont près de 2 fois moins souscrits, avec 16% pour les contrats « Or » et 13% pour les contrats « Platine »

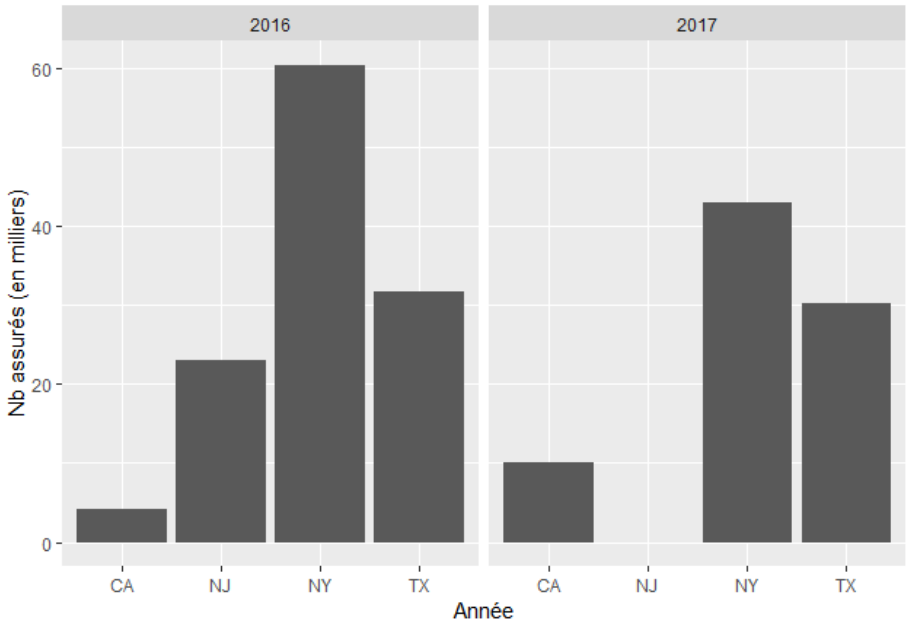
3.2.2 Base de sinistres

Nous nous intéressons dans cette section aux statistiques descriptives sur les données de sinistres à disposition, qui sont des données des sinistres détaillés sur les années 2016, 2017, 2018 sur 5 états différents.

Le graphique ci-dessous illustre les montants des sinistres totaux payés par état pour 2016 et 2017.

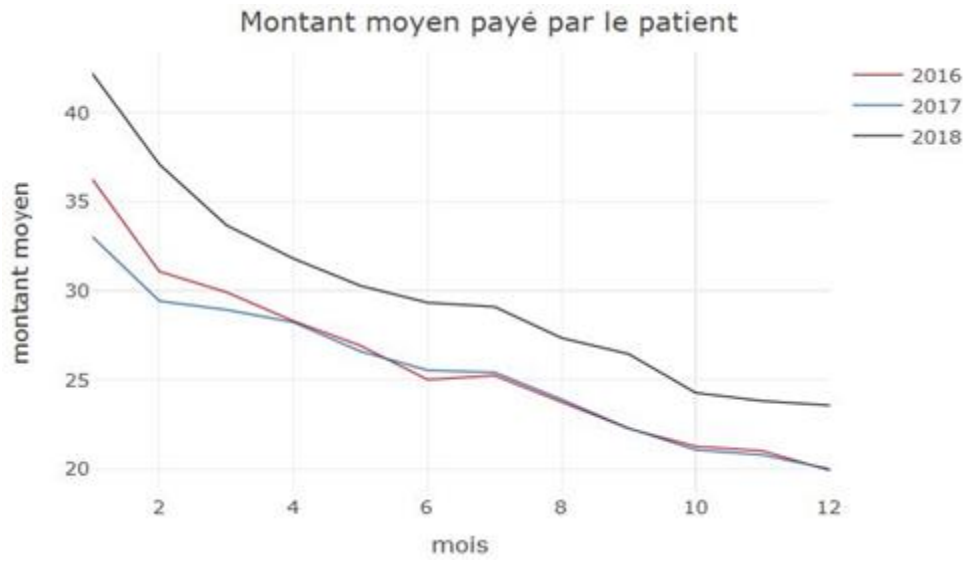


A titre de référence, le nombre d'assurés pour chacun des états sur la même période est représenté ci-dessous. On observe naturellement des montants de sinistres en cohérence avec les expositions constatées sur les périodes.

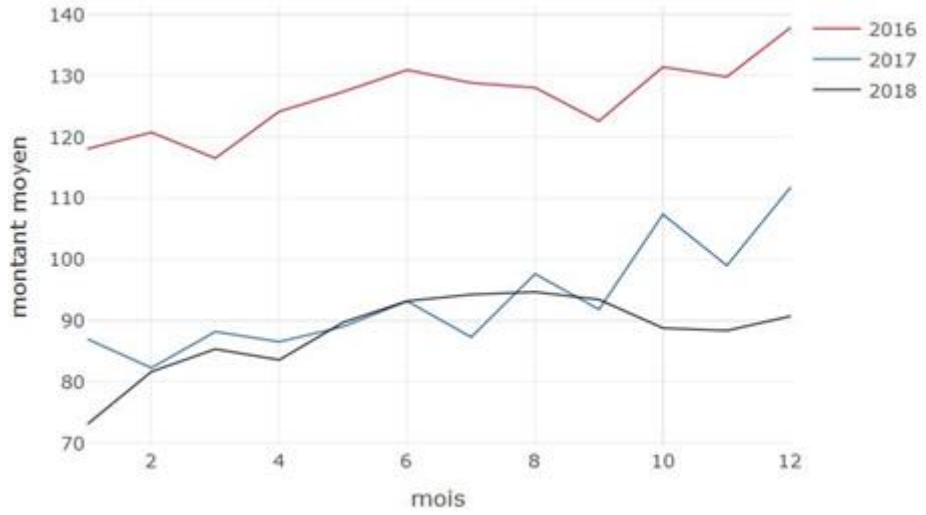


Ensuite si l'on observe l'évolution des sinistres payés mois par mois, des tendances intéressantes se dégagent, directement en lien avec la structure des garanties proposées.

Les graphiques ci-après montrent le montant moyen payé par l'assuré et par l'assureur mois par mois :



Graphique : Montant moyen payé par l'assuré par mois (USD)

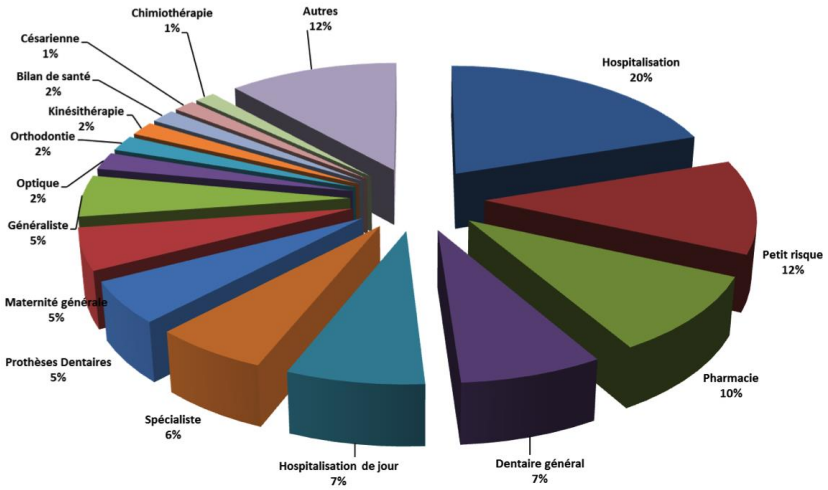


Graphique : Montant moyen payé par l'assureur mois par mois (USD)

On observe ainsi l'effet franchise annuelle ("Deductible") : au fur et à mesure de l'année, les dépenses cumulées des assurés atteignent le seuil annuel entraînant une diminution de la charge relative supportée par l'assuré et une accélération en parallèle de la charge pour l'assureur.

Il est également intéressant de noter une évolution importante du montant moyen payé par l'assuré mois par mois en 2018, cela s'explique par une évolution globale des montants de franchises sur les produits proposés par l'assureur en 2018.

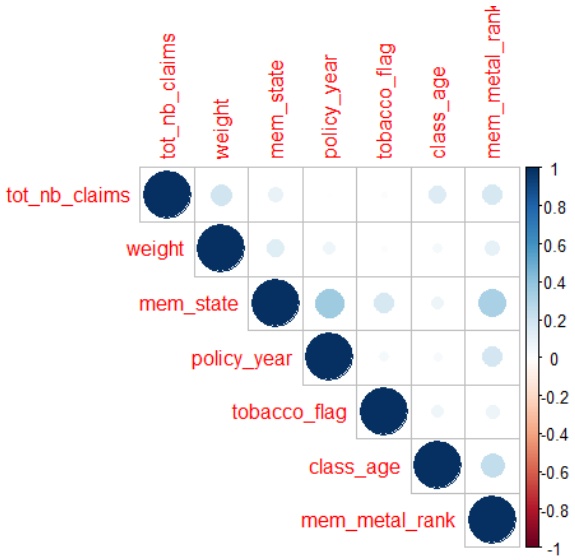
La répartition des frais réels par postes médicaux



On note que l'hospitalisation représente le 1er poste de frais avec près de 20% du total. Les frais de remboursement de pharmacie représentent quant à eux 10% et les frais dentaires 7%. On comprend également la grande diversité de postes de remboursements

Analyse des corrélations entre le nombre de sinistres et les variables tarifaires

Pour comprendre de manière macro les corrélations qui peuvent exister entre le nombre de sinistre et les variables tarifaires, l'analyse de corrélation suivante a été réalisée :



On observe ainsi par exemple que le nombre de sinistres est positivement corrélé, bien naturellement, à la durée d'exposition de l'année ('weight') ainsi qu'au niveau métallique du contrat ; et dans une moindre mesure à l'âge des populations et à l'état.

4. Choix et élaboration du modèle de tarification

4.1 Les variables d'intérêt et les variables explicatives

Comme vu précédemment, l'Obamacare prévoit 5 variables tarifaires : le lieu de résidence (« rating areas »), l'âge, la population assurée sur le contrat (ex. souscripteur seul vs. famille), le statut fumeur/non-fumeur, le « niveau métallique » du plan (~niveau de couverture).

Ces 5 variables doivent expliquer les variables réponses du modèle, variables suivantes dites « d'intérêt » :

- Charge annuelle de consommation médicale
- Nombre de sinistres dans l'année
- Coût moyen d'un sinistre dans l'année
- Occurrence d'un sinistre

Méthodologiquement, compte tenu du fait que les données étudiées concernent plusieurs années, la question de l'indépendance temporelle se pose pour mesurer les coûts et incidences. En effet, nous avons à notre disposition des données relatives à des individus ayant été présents plusieurs années. Les réalisations de consommation annuelle de ces individus sont a priori corrélées, ce qui pose un problème d'un point de vue théorique car les estimations du modèle implémenté (GLM) se font par la méthode du maximum de vraisemblance ; et cette méthode du maximum de vraisemblance est basée sur l'hypothèse d'indépendance entre les observations.

Asymptotiquement, la dépendance sérielle a peu d'impact sur la moyenne (cf. la convergence de l'estimateur des paramètres du modèle supposant l'indépendance sérielle). Mais la variance est quant à elle impactée par cette dépendance. Des méthodes de type « GEE » permettent d'améliorer la variance asymptotique de l'estimateur, mais ne seront pas étudiées dans ce mémoire.

Autre point méthodologique : les interactions entre variables explicatives

Dans la modélisation, au-delà de la prise en compte de l'impact de chacune des variables explicatives, des effets croisés liés aux interactions entre ces variables devront également être pris en compte. En effet, le modèle GLM estime l'influence de chacune des variables explicatives de façon indépendante. Mais il peut être utile de mesurer les effets liés à différents croisements.

A titre d'illustration, en tarifant sans interaction en fonction du sexe et du statut fumeur vs non-fumeur, la différence relative entre le tarif des femmes et celui des hommes sera la même, quelque soit selon le statut fumeur ou non fumeur. En intégrant un effet d'interaction sexe x statut fumeur vs non fumeur, la spécificité des croisements sera prise en compte et les écarts relatifs ne seront plus les mêmes. Ces éléments sont présents à titre méthodologique, mais ne seront pas étudiés spécifiquement dans notre modélisation.

4.2 Analyse des modèles envisageables

Les variables explicatives étant déterminées, plusieurs approches de tarification peuvent être envisagées. Nous présentons dans la suite l'approche « Fréquence x Coût moyen » et l'approche « Probabilité x Charge », qui peuvent être abordées d'un point de vue déterministe ou par l'intermédiaire des Modèles Linéaires Généralisés (GLM).
La robustesse de l'approche déterministe diminue avec l'ajout de variables explicatives, car la taille des échantillons est réduite dans les cellules tarifaires formées par les nombreux croisements de modalités des variables. Nous lui préférons donc la modélisation des variables d'intérêt via la théorie des Modèles Linéaires Généralisés.

4.3 L'approche « Fréquence x Coût moyen »

L'approche Fréquence x Coût moyen est une approche largement utilisée pour la tarification de contrats Santé.

4.3.1 Calcul de la Fréquence

La fréquence annuelle de sinistre correspond au nombre de sinistres annuels observés. En santé, un sinistre se traduit par un ou plusieurs actes de soins. La fréquence se calcule en rapportant le nombre d'actes médicaux à l'exposition. Cette mesure est ici déterministe.

En cas de « surdispersion », la fréquence peut également être obtenue en utilisant un modèle linéaire généralisé, via une régression de Poisson voire une régression binomiale négative, où les réalisations de la variable d'intérêt correspondent au nombre d'actes par individu et par année, que l'on annualise en divisant cette quantité par l'exposition (1 pour 1 an, 0,5 pour 6 mois...).

4.3.2 Calcul du Coût moyen

Il s'agit ici du coût moyen d'un acte de soin réalisé durant l'année considérée. Il peut être estimé de façon déterministe en ramenant la charge annuelle de consommation médicale au nombre d'actes de soin dans l'année.
Une régression gamma ou lognormale peut aussi être envisagée, les réalisations de la variable réponse représentant alors les coûts moyens par individu et par année.

4.3.3 Calcul de la prime pure

La prime pure est obtenue en multipliant fréquences et coûts moyens par cellule tarifaire. Actuariellement, le risque que l'on considère dans l'approche Fréquence x Coût moyen est la charge annuelle de sinistre. Celle-ci peut s'écrire sous la forme de la variable aléatoire suivante :

$$C = \begin{cases} \sum_{k=1}^{N} C_k \ si\ N > 0 \\ 0\ si\ N = 0 \end{cases}$$

Avec :

- C: Charge sinistre annuelle d'un individu
- N: Variable aléatoire représentant le nombre de sinistres dans l'année
- $Ck:$ Coût du k^{ème} sinistre

La prime pure que l'on veut déterminer est l'espérance mathématique de la charge sinistre *C*. En supposant les $(C_k)_{k\geq 1}$ indépendants et identiquement distribués (*iid*) et indépendants de la variable *N*, on a :

$$E[C] = \sum_{k\geq 1} P_r(N = k) E\left[\sum_{i=1}^{k} C_i\right]$$

$$= \left(\sum_{k\geq 1} P_r(N = k) k\right) \times E[C_k]$$

$$E[C] = E[N] \times E[C_k]$$

$$\Leftrightarrow$$

$$Prime\ Pure = Fréquence \times Coût\ moyen$$

4.3.4 Limites liées à l'utilisation d'une approche Fréquence x Coût moyen

Les hypothèses liées à l'approche Fréquence x Coût moyen sont fortes puisqu'elles reposent d'une part sur l'indépendance entre le nombre de sinistres et les coûts $(C_k)_{k\geq 1}$ et d'autre part sur le caractère iid des variables $(C_k)_{k\geq 1}$ du processus de coûts. Il s'agit de 2 hypothèses qui ne sont pas toujours vérifiées en pratique et qui constituent donc les principales limites de cette approche.

4.4 L'approche Probabilité de consommer x Charge de consommation

Pour modéliser la probabilité de consommer, cette approche s'appuie sur l'implémentation d'une régression logistique et effectue une étude globale de la loi des frais réels par poste médicaux pour la modélisation de la charge annuelle de consommation.

4.4.1 Calcul de la Probabilité de consommer

La probabilité de consommer au moins une fois dans l'année correspond, à l'exposition près, à la proportion de bénéficiaires sinistrés. Pour modéliser cette probabilité, un modèle de régression binomial est utilisé. L'idée est de considérer les réalisations de consommation comme des réalisations d'une variable binomiale et d'estimer la probabilité de consommer par la probabilité associée à cette variable.

En pratique, on cherche à modéliser la probabilité d'occurrence ($\mu_i = q_i$) de cet évènement (ici probabilité de consommer) sachant la catégorie ou classe de facteur explicatif. Il s'agit alors d'une variable d'intérêt à valeur dans [0,1] alors que le prédicteur linéaire ($\eta_i = x_i^t \boldsymbol{\beta}$) a des valeurs dans $\mathbb{R}$. Trois principales fonctions de lien sont alors utilisées pour se ramener de $\mathbb{R}$ à [0,1], dont notamment le lien logit. Ce lien logit est le plus utilisé en pratique pour la modélisation de probabilités, il s'agit du lien canonique associé au modèle de régression binomiale :

$$g(q_i) = ln\left(\frac{q_i}{1-q_i}\right)$$

Ce modèle logit peut être vu comme un modèle de régression linéaire utilisant une variable latente $Y^* \in \mathbb{R}$, variable binomiale dont on cherche à modéliser la probabilité de succès $Y = 1_{[Y^* \geq 0]}$.

Formellement, cette probabilité peut être considérée comme l'espérance mathématique de la variable aléatoire suivante :

$$IND = 1_{\{N>0\}} \begin{cases} 1 \; si \; N > 0, N \; nombre \; de \; sinistres \; dans \; l'anéne \\ 0 \; si \; N = 0 \end{cases}$$

4.4.2 Calcul de la charge de consommation

Cette charge totale par année et par individu sera modélisée par une régression gamma. Les réalisations sur lesquelles on s'appuie pour la modélisation sont des quantités strictement positives. Nous considérons en effet la consommation annuelle des assurés ayant consommé au moins une fois. C'est la stricte positivité des observations qui justifie l'usage d'une loi comme la loi gamma dans la modélisation.

4.4.3 Calcul de la prime pure

La prime pure est obtenue en multipliant classe par classe la probabilité de consommation par la charge totale annuelle que l'on aura calculée. En effet, actuariellement le risque considéré dans l'approche Probabilité de consommer x Charge de consommation est le suivant :

$$C = IND \times C^{+}, C^{+} \; ayant \; même \; loi \; que \; C|N > 0$$

Avec :

- *IND :* variable aléatoire binaire représentant le fait de consommer ou non,
- C^{+} : variable aléatoire représentant la charge annuelle de consommation sachant qu'il y a eu consommation.

La prime pure est alors l'espérance mathématique de ce risque et s'écrit :

$$\begin{aligned} E[C] &= E[IND \times C^{+}] \\ &= Pr(IND = 1) \times E[C^{+}|IND = 1] + Pr(IND = 0) \times E[C^{+}|IND = 0] \\ &= Pr(N > 0) \times E[C^{+}|N > 0] + Pr(N = 0) \times E[C^{+}|N = 0] \\ &= Pr(N > 0) \times E[C^{+}|N > 0] \\ &\Leftrightarrow \end{aligned}$$

$$Prime\ Pure = Probabilité\ de\ consommer \times Charge\ de\ consommation$$

4.4.4 Rationnel du choix de l'approche

Pour notre modélisation tarifaire nous privilégions une approche mixte, sur la majorité des postes nous utilisons une approche « Fréquence x Coût moyen » et pour certains une approche « Probabilité de consommer x Charge de consommation ».

En effet, pour les actes de fréquence (tels que les consultations généralistes ou spécialistes, les analyses de laboratoires…), il est utile d'un point de vue métier de développer une approche « Fréquence x Coût moyen ». La fréquence nous renseigne sur le niveau de morbidité de la population assurée, le coût moyen dépend quant à lui des négociations entreprises avec les différents fournisseurs de soin. Enfin, les franchises étant pour ce type de soin appliquées au niveau de l'acte (ex : « co assurance » de 20% par consultation de généralise, « co pay » de $40 par consultation de spécialiste,…), il est également plus aisé d'un point de vue métier pour établir la prime pure de manipuler la quantité du coût moyen de l'acte.

Pour les actes à faible fréquence et charge importante (ex : hospitalisation, hospitalisation d'urgence…), une approche « Probabilité de consommer x Charge de consommation » est également intéressante d'un point de vue métier. En effet pour ces épisodes plus rares mais coûteux, la quantité d'actes liés à l'hospitalisation importe moins que la charge globale liée à

l'épisode de consommation. En effet l'assureur négocie soit au travers de contrats préétablis, soit de manière ponctuelle des rabais de la charge globale auprès des hôpitaux.

Dans la partie Application, nous illustrerons notre approche théorique par la présentation d'une approche « Fréquence x Coût moyen » sur un acte, puis l'approche « Probabilité de consommer x Charge de consommation » sur un autre acte.

Une fois les primes pures déterminées, il faut évaluer l'impact de la prise en compte de plafonds et de franchises sur les tarifs. C'est l'objet de la partie suivante.

.

4.5 La prise en compte de franchises et plafonds contractuels dans les tarifs

Comme vu dans les sections précédentes avec notamment la notion de franchise "déductible", la prise en compte de franchises est un élément impactant la tarification, celle-ci permettant entre autres de diminuer l'aléa moral ; l'assuré devant, dans ces cas de franchises spécifiques, payer entièrement les frais en cas de sinistre tant qu'un seuil annuel minimal n'aura pas été atteint.

De manière générale, il existe 2 méthodes pour prendre en compte les franchises dans la tarification, le principe étant que Les comportements des assurés peuvent changer en fonction de la présence ou non de franchises et plafonds dans les termes des garanties qu'ils souscrivent : L'une privilégie la diminution du biais associé aux estimations avec une augmentation de la volatilité, et inversement pour l'autre.

Exploiter les plafonds et franchises comme des variables explicatives

La 1ère approche requiert la disponibilité de données exploitables concernant les franchises ou plafonds relatifs à chaque contrat. Ces données seraient alors intégrées aux variables explicatives. Cette approche permet de s'affranchir de l'aléa moral (cf. changement de comportement de consommation des assurés selon la présence de plafonds et de franchises). Elle permet également de faire des prévisions tarifaires intégrant directement des tarifs et franchises de différents niveaux.

Cette approche ne peut ici être retenue en raison du manque de données exhaustives sur les caractéristiques des contrats.

En effet, la « perte » de données exploitables (car n'ayant pas toutes les informations requises) réduirait de plus de la moitié les données à disposition, ce qui en ferait une approche possédant peu de biais mais ayant une plus forte volatilité.

Exploiter a posteriori les plafonds et franchises

La 2nde approche consiste à négliger l'impact de la présence de franchises et de plafonds sur la consommation médicale des assurés. Cette approche est effectivement biaisée par le fait de ne pas prendre en compte la limitation de l'aléa moral, typiquement une consommation supérieure en cas d'absence de franchise. Et c'est même un double effet qui peut exister si on néglige ce changement de comportement : on viendrait en effet modéliser une consommation moyenne franchisée sur des données de consommation qui le sont implicitement du fait de l'aléa moral, et qui se trouve réduit par ces limites.

La prime pure en présence de plafond et franchise contractuels

En notant θ le taux de remboursement (franchise = 1 - θ) et ω le plafond de remboursement, la charge annuelle de consommation plafonnée et franchisée s'écrit comme suit :

$$C^+(\theta,\omega) = \theta C^+ 1_{\{\theta C^+ < \omega\}} + \omega\, 1_{\{\theta C^+ \geq \omega\}}$$

$$= \begin{cases} \theta C^+ \; si \; \theta C^+ < \omega \\ \omega \; si \; \theta C^+ \geq \omega \end{cases}$$

L'espérance mathématique de ce risque s'écrit :

$$\begin{aligned} E[C^+(\theta,\omega)] &= E[C^+(\theta,\omega)|\,\theta C^+ < \omega]P_r(\theta C^+ < \omega) + E[C^+(\theta,\omega)\mid \theta C^+ \geq \omega\,]\,P_r(\theta C^+ \geq \omega) \\ &= \theta\, E\left[C^+|\, C^+ < \tfrac{\omega}{\theta}\right] P_r(\theta C^+ < \omega) + \omega P_r(\theta C^+ \geq \omega) \\ &= \theta\, E\left[C^+|\, C^+ < \tfrac{\omega}{\theta}\right] F_{c^+}\left(\tfrac{\omega}{\theta}\right) + \omega\,(1 - F_{c^+}\left(\tfrac{\omega}{\theta}\right)) \end{aligned}$$

Et la prime pure devient :

$$E[C(\theta,\omega)] = P_r(N > 0) \times \left[\theta E\left[C^+ \mid C^+ < \frac{\omega}{\theta}\right] F_{c^+}\left(\frac{\omega}{\theta}\right) + \omega\left(1 - F_{c^+}\left(\frac{\omega}{\theta}\right)\right)\right]$$

Nous avons ainsi explicité notre démarche de tarification :

- L'approche de tarification Fréquence x Coût Moyen sur les actes de fréquence.
- L'approche Probabilité x Charge annuelle de consommation pour les actes à faible fréquence et charge importante.
- Les franchises et plafonds contractuels pris en compte a posteriori de la modélisation. En d'autres termes, les fréquences et coûts ainsi que les probabilités et charges annuelles sont dans un 1er temps modélisées sans tenir compte de leurs possibles influences sur les comportements. Leur impact est ensuite intégré aux lois modélisées

Les différentes quantités d'intérêts sont modélisées à l'aide de la théorie des Modèles Linéaires Généralisés, dont les fondements théoriques sont présentés dans la section suivante.

5. Les modèles linéaires généralisés pour la tarification

5.1 La formalisation du modèle linéaire gaussien

Il s'agit du modèle de régression suivant :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \chi_{ij} + \epsilon_i, \quad \epsilon_i \sim Nor(0, \sigma^2), i \, \epsilon \{1, \ldots, n\}$$

Ce qui s'écrit sous forme matricielle

$$Y = X\beta + \epsilon$$

Avec

- n : Nombre d'observations,
- p : Nombre de variables explicatives,
- $Y = (Y1, \ldots, Yn)t$: Variables réponses à expliquer supposées indépendantes et non identiquement distribuées,
- $Xj = (x1j, \ldots, xnj)t$, $j \in \{1, \ldots, p\}$: $j^{ème}$ variable explicative
- $X = \begin{bmatrix} 1 & x_1^t \\ \ldots & \ldots \\ n & x_n^t \end{bmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$: Matrice reprenant les variables explicatives,
- $\beta = (\beta 0, \ldots, \beta p)t$: Paramètres du modèle à estimer,
- $\epsilon = (\epsilon 1, \ldots, \epsilon n)t$: Bruit blanc d'écart type σ, $\epsilon_i \sim Nor(0, \sigma^2) \Rightarrow Yi \sim Nor\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right)$

5.1.1 Estimation des paramètres

Dans le modèle défini dans la section précédente, on suppose que les y_i observés sont des réalisations de variables $Yi \sim Nor\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right)$. Les paramètres β_j peuvent alors être estimés par maximum de vraisemblance. La vraisemblance associée au modèle s'écrit comme suit :

$$\begin{aligned} L(\beta, \sigma | y) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} exp\left(\frac{-1}{2\sigma^2}(y_i - x_i^t\beta)^2\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(\frac{-1}{2\sigma^2}(y - X\beta)(y - X\beta)^t\right) \end{aligned}$$

On montre alors que l'estimateur de maximum de vraisemblance du vecteur β est solution des équations dites normales :

$$X^t X\beta - X^t Y = 0 \iff (X^t X)^{-1} X^t Y$$

Un premier estimateur de σ s'écrit :

$$\hat{\sigma}^2 = \frac{\hat{\in}\hat{\in}^t}{n}$$

Mais ce dernier étant biaisé, on lui préfèrera :

$$\hat{\sigma}^2 = \frac{\hat{\in}\hat{\in}^t}{n-p-1}$$

La valeur ajustée de la variable réponse *Y* s'écrit :

$$\hat{Y} = \boldsymbol{X}\,\hat{\beta}$$
$$\hat{Y} = X(X^t X)^{-1} X^t Y$$

La matrice $H = X(X^t X)^{-1} X^t$ est alors appelée matrice de prédiction. Il s'agit de la matrice de projection du vecteur d'observation *Y* sur l'hyperplan des variables explicatives. $\hat{Y}$, valeur ajustée de *Y,* est la projection de *Y* sur l'espace des variables explicatives. Intuitivement, c'est la meilleure approximation que l'on peut faire de *Y* compte tenu des informations disponibles via les variables explicatives. Le vecteur des résidus est estimé par :

$$\hat{\in} = Y - \hat{Y} = (I - H)Y$$

Les équations normales peuvent encore s'écrire :

$$\sum_{j=1}^{p} X_i^t (y_i - \beta^t x_i) = 0, \quad i \in \{1, \ldots, n\}$$

Ecrites sous cette forme, les équations normales ont une interprétation intuitive. Les résidus associés au modèle s'écrivent $\in_i = y_i - \beta^t x_i$. L'écriture des équations normales correspond à l'orthogonalité entre le vecteur des résidus du modèle et le plan des variables explicatives.

La projection du vecteur des résidus sur l'hyperplan des variables explicatives est nulle, ce qui signifie intuitivement qu'il n'y a plus d'information à tirer des variables explicatives pouvant apporter de l'information sur les résidus.

5.1.2 Validation du modèle

La justesse de l'estimation peut être mesurée par le coefficient de détermination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
$$= 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad R^2 \epsilon\, [0; 1]$$

Un R^2 proche de 1 indique que l'ajustement est de bonne qualité. En effet ce coefficient correspond au rapport de la somme des carrés expliquée à la somme des carrés totale. Un R^2 proche de 1 équivaut donc à une somme des carrés expliquée proche de la somme des carrés totale et témoigne ainsi d'une perte minime d'information dans la modélisation.

Cet estimateur a cependant le défaut de tendre systématiquement vers 1 avec l'ajout de variables explicatives, on lui préférera donc le coefficient de détermination ajusté suivant, pénalisé par le nombre p de variables explicatives du modèle :

$$R^2 = 1 - \frac{n-1}{n-p-1}\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Les principales hypothèses du modèle linéaire gaussien sont les suivantes :

$$Y = \boldsymbol{X}\,\beta + \epsilon$$
$$\epsilon \sim \boldsymbol{Nor}(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$$

- X déterministe
- Rang(X) = p + 1 < n

Sous ces hypothèses, on montre que $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(XX^t)^{-1})$. Ce résultat permet d'avoir des intervalles de confiance et d'effectuer des tests d'hypothèse sur les paramètres estimés. Sous ces hypothèses, on a en effet $\hat{\beta}_j \sim N_{p+1}\left(\beta_j, \sigma^2(XX^t)^{-1}_{jj}\right)$ et donc :

$$\frac{\hat{\beta}_j - \beta_j}{C} \sim t_{n-p-1}$$

Où t_{n-p-1}suit une loi de Student à *n-p-1* degrés de liberté et $S^2 = \frac{\hat{\epsilon}\hat{\epsilon}^t}{n-p-1}$ estimateur sans biais de σ^2.

Test de significativité

Pour tester la significativité de la *j*ème variable explicative et décider de l'intégrer ou non au modèle de régression, on peut effectuer le test suivant :

$$H_0 : \beta_j = 0 \text{ contre } H_1 : \beta_j \neq 0$$

L'hypothèse nulle H_0 est alors rejetée au seuil α si $\left|\frac{\hat{\beta}_j}{S\sqrt{(X^tX)^{-1}_{jj}}}\right| > t_{1-\frac{\alpha}{2};\, n-p-1}$ avec $t_{1-\frac{\alpha}{2};\, n-p-1}$ quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi de Student à $n-p-1$ degrés de liberté.

Intervalles de confiance

Dans certains cas, il peut être intéressant d'avoir un intervalle de confiance autour des paramètres estimés pour mesurer le risque associé aux estimations. Comme $\hat{\beta}_j \sim N_{p+1}\left(\beta_j, \sigma^2 (XX^t)^{-1}_{jj}\right)$, un intervalle de confiance au seuil α autour de $\hat{\beta}_j$ est alors :

$$\beta_j \in \left[\hat{\beta}_j - t_{1-\frac{\alpha}{2};\, n-p-1} S\sqrt{(XX^t)^{-1}_{jj}} \;;\; \hat{\beta}_j + t_{1-\frac{\alpha}{2};\, n-p-1} S\sqrt{(XX^t)^{-1}_{jj}}\right]$$

Le modèle linéaire gaussien a longtemps été utilisé pour quantifier l'impact de variables explicatives sur des variables d'intérêt. Mais ce modèle n'est souvent pas adapté à la modélisation en assurance. Nous présentons dans la section qui suit la théorie des modèles linéaires généralisés qui offrent plus de possibilité en termes de modélisation et sont plus adaptés aux problématiques assurantielles.

5.1.3 Limites du modèle linéaire gaussien et généralisation

Le modèle linéaire gaussien n'est en effet souvent pas adapté aux problématiques d'assurance car présentant par exemple les insuffisances suivantes :

- Il s'agit d'une loi continue et à valeur dans $\mathbb{R}$. Or en assurance, on s'intéresse la plupart du temps au nombre de sinistres à valeurs dans $\mathbb{N}$, au coût d'un sinistre à valeurs dans $\mathbb{R}^+$ ou à la probabilité d'avoir un sinistre à valeurs dans [0,1]. Il est parfois possible d'appliquer de bonnes transformations à la variable réponse afin de se ramener à une modélisation par le modèle linéaire gaussien avant d'effectuer les transformations inverses pour avoir les ajustements souhaités, mais cela induit d'autres biais.
- La relation linéaire entre la variable réponse et les variables explicatives n'est pas nécessairement adaptée à toutes les modélisations et impose d'importantes limitations.
- L'homoscédasticité supposée dans le modèle linéaire gaussien impose aussi certaines limites et ne traduit pas nécessairement la réalité des variables dont on souhaite étudier le "comportement".

Les modèles linéaires généralisés sont une double généralisation du modèle linéaire classique et pallient les importantes limitations qu'il impose. Nous noterons dans la suite $\eta = X\beta$ le score du modèle et $\mu = E(Y)$ l'espérance de la variable d'intérêt Y

5.2 La 1ère généralisation du modèle linéaire classique : les lois de la famille exponentielle comme loi pour la variable réponse

Le modèle linéaire classique est souvent inadapté quant à la loi qu'il associe aux variables d'intérêt. Grâce aux GLM, il est possible de leur associer d'autres lois que la loi normale. Ces lois font partie de la famille exponentielle qui offre un cadre commun d'estimation et de modélisation.

La famille exponentielle

Une variable *Y* a une loi faisant partie de la famille exponentielle si sa densité peut se mettre sous la forme :

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

Avec :

- θ *:* Paramètre naturel, aussi appelé paramètre canonique ou encore paramètre de la moyenne
- ϕ *:* Paramètre de dispersion. Il n'existe pas pour certaines lois de la famille exponentielle, notamment lorsque la loi de Y ne dépend que d'un seul paramètre (on pose dans ces cas ϕ = 1). Sinon il s'agit d'un paramètre de nuisance qu'il faut estimer. Comme son nom l'indique, ce paramètre est lié à la variance de la loi. C'est aussi un paramètre très important dans la mesure où il contrôle la variance et donc le risque. Dans certains cas une pondération est nécessaire pour accorder des importances relatives aux différentes observations et le paramètre ϕ est remplacé par $\frac{\phi}{\omega}$, ω étant un poids connu à priori.
- *S :* Support de la loi, sous-ensemble de $\mathbb{R}$ ou $\mathbb{N}$
- La fonction *b*(.) (respectivement c(.)) est une fonction de θ (resp. de ϕ et $y \in S$). La fonction *b*(.) doit être 2 fois dérivable.

Les densités des lois Normale, de Poisson, Binomiale et Gamma peuvent se mettre sous cette forme et font ainsi partie de la famille exponentielle.

Les lois de la famille exponentielle sont très utilisées en pratique du fait de certaines propriétés, notamment celles concernant leur espérance et leur variance. En effet pour une variable *Y* dont la loi fait partie de la famille exponentielle, on a le résultat suivant :

$$E(Y) = b'(\theta) \quad et \quad V(Y) = b''(\theta)\frac{\phi}{\omega}$$

Ces propriétés sur les estimateurs sont des propriétés importantes pour le modèle GLM.

Le paramètre naturel ou de la moyenne θ est directement lié à la moyenne ($\mu = E(Y) = b'(\theta)$) et le paramètre de dispersion ϕ à la variance de la variable d'intérêt. La variance se décompose en un facteur $b''(\theta)$ dépendant uniquement de θ et donc de la moyenne μ et d'un autre facteur dépendant du paramètre de dispersion. Le premier facteur est appelé fonction variance ($V(\mu) = b''(\theta)$). Cette fonction représente la variance à un facteur près et on peut montrer qu'elle caractérise la loi de Y. Après estimation de la moyenne, elle permet d'avoir la variance après estimation du paramètre de dispersion s'il existe. Pour les lois classiques de la famille exponentielle, la dépendance entre la moyenne et la fonction variance (et donc la variance) est très simple.

Nous reprenons quelques résultats dans le tableau ci-dessous :

Loi	$V(\mu)$
Normale	*1*
Poisson	μ
Gamma	μ^2
Binomiale	$\mu(1-\mu)$

Au-delà de la loi de la variable réponse dont on a parlé dans cette section, une caractéristique importante des GLM est la fonction de lien entre l'espérance mathématique de Y au prédicteur linéaire construit à partir des variables explicatives. C'est l'objet de la section suivante

5.3 La deuxième généralisation du modèle linéaire classique : la fonction de lien

Dans le modèle linéaire gaussien nous avions $\mu = \eta$ avec $\mu = E[Y]$ et $\eta = X\beta$. La fonction de lien relie μ au prédicteur linéaire η par le biais de la relation $g(\mu) = \eta$ (lien noté g). Avec l'introduction de cette fonction monotone et dérivable, on s'autorise une dépendance non linéaire entre la variable réponse et les variables explicatives.

Aussi, le prédicteur linéaire peut théoriquement être dans un espace qui peut ne pas coïncider avec l'espace de la variable réponse. Il peut par exemple prendre ses valeurs dans $\mathbb{R}$ alors que la variable d'intérêt a des valeurs dans $\mathbb{R}^+$(modélisation d'un coût) ou dans [0,1] (modélisation d'une probabilité). Grâce à une fonction de lien « bien choisie », il est possible de palier à cette limite.

Lien canonique

La fonction de lien canonique associée à une loi de la famille exponentielle est définie comme la fonction de lien vérifiant $g(\mu) = \theta = \eta \Leftrightarrow g(.) = b'^{-1}(.)$. Le principal intérêt du lien canonique réside dans la simplification qu'il induit au niveau de l'estimation des paramètres.

Ci-dessous un tableau reprenant les fonctions de lien canonique associées à quelques lois classiques de la famille exponentielle :

Loi	*Lien canonique g(μ)*
Normale	*μ*
Poisson	*ln(μ)*
Gamma	*1/μ*
Binomiale	$ln(\frac{\mu}{1-\mu})$

5.3.1 La formalisation du modèle GLM

On considère le modèle de régression suivant :

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = x_i^t \beta = \eta_i, \quad i \in \{1, \ldots, n\}$$

Avec :

- n : nombre d'observations,
- p : Nombre de variables explicatives,
- g : fonction de lien liant le prédicteur linéaire $x_i^t \beta$ à la moyenne de μ_i de Y_i,
- $\mu i = \mathrm{E}[\mathrm{Y_i}]$ où les Y_i sont les variables réponses à expliquer que l'on suppose indépendantes et non identiquement distribuées
- $\beta = (\beta_0, \ldots, \beta_\mathrm{p})^\mathrm{t}$: Paramètres du modèle à estimer
- $(\mathrm{x_{1j}}, \ldots, \mathrm{x_{nj}})^\mathrm{t}, \mathrm{j} \in \{1, \ldots, \mathrm{p}\}$: *j*ème variable explicative.

La loi que l'on associe à la variable réponse *Y* fait partie de la famille exponentielle. La densité des Y_i s'écrit alors :

$$f(y_i|\theta_i, \phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \phi)\right), \quad y_i \in S$$

Avec :

- θ_i *:* Paramètre naturel associé à $\mathrm{Y_i}$. Ce paramètre dépend de $\beta = \left(\beta_{0,\ldots,}\beta_p\right)^t$
- ω_i : Poids affecté à l'observation
- ϕ *:* Paramètre de dispersion. Il s'agit 'un paramètre de « nuisance » à estimer. Il est pris commun à tous les $\mathrm{Y_i}$, mais cela ne signifie pas qu'ils ont la même variance $(Var(Y_i) = b''(\theta_i)\frac{\phi}{\omega_i}$

Equations de vraisemblance et estimation des paramètres

Dans notre modèle, nous choisissons à priori la fonction de lien *g* ; les paramètres à estimer sont alors les β_i , $i \in \{1,\dots,p\}$ et le paramètre de dispersion ϕ.

Les coefficients β

L'estimation des β_i se fait en maximisant la vraisemblance du modèle. Avec l'hypothèse d'indépendance des Y_i, la log-vraisemblance du modèle s'écrit :

$$L(\theta(\beta)|y,\phi) = \sum_{i=1}^{n} ln\left(f(y_i|\theta_i,\phi)\right) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + \sum_{i=1}^{n} c(y_i,\phi)$$

Rechercher les $\beta_j, j \in \{1,\dots,p\}$ qui maximisent la vraisemblance revient à rechercher les β_j vérifiant :

$$\frac{\partial}{\partial\beta_j} L(\theta(\beta)|y,\phi) = \sum_{i=1}^{n} \frac{\partial}{\partial\beta_j} ln\left(f(y_i|\theta_i,\phi)\right) = 0$$

Si g est la fonction canonique, les équations de vraisemblance deviennent :

$$\sum_{i=1}^{n} \omega_i\,(y_i - \mu_i)x_{ij} = 0, \;\; j \in \{1,\dots,p\}$$

Cette équation traduit l'orthogonalité entre le vecteur des résidus du modèle et le plan des variables explicatives. Intuitivement, cela signifie qu'un maximum d'information a pu être tiré des observations et donc que les résidus ne contiennent plus d'information pouvant être captée par le plan explicatif.

5.3.2 La validation du modèle

Dans le modèle linéaire gaussien, l'ajustement du modèle à un jeu de données consiste en une projection orthogonale des observations du vecteur *Y* des observations sur l'hyperplan des variables explicatives. En d'autres termes, on démontre l'égalité suivante grâce au théorème de Pythagore :

$$\sum\nolimits_{i=1}^{n} (y_i - \bar{y})^2 = \sum\nolimits_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum\nolimits_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Cette équation traduit le fait que dans le modèle linéaire gaussien, la variation totale est exactement l'addition de la variation résiduelle à la variation expliquée. Le coefficient de détermination est alors le rapport de la variation expliquée à la variation totale et constitue une

mesure de la qualité d'ajustement du modèle : A nombre de variables explicatives fixé, plus il est proche de 1, meilleur est le modèle.

Cette égalité n'est plus vérifiée dans le cadre des modèles linéaires généralisés. Pour mesurer la qualité d'ajustement du modèle, on compare le modèle à un modèle "idéal" où l'on aurait autant de variables explicatives que d'observations. Ce modèle s'appelle le modèle saturé et est caractérisé par le fait que $y_i = \hat{\mu}_i$.
La comparaison des modèles ajustée et saturée se fait par comparaison des vraisemblances. La vraisemblance associée au modèle saturé est maximale car ne dépendant pas des paramètres à estimer.
L'idée sera de comparer l'éloignement du modèle ajusté par rapport au modèle saturé. Plus cet éloignement sera petit, meilleur sera le modèle. Cet éloignement est traduit par la notion de déviance.

La déviance

On note respectivement $L(y|y)$ et $L(\hat{\mu}|y)$ les vraisemblances des modèles saturé et ajusté. On estime qu'un modèle est "bon" si les deux vraisemblances précitées sont relativement proches. Introduisons la statistique du rapport de vraisemblance suivante :

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

Notre modèle sera ainsi jugé "bon" si Λ est proche de 1 ou encore si ln(Λ) est proche de 0. On note D la déviance et on a :

$$D = 2ln(\Lambda)\phi$$

L'idéal serait donc d'avoir $D = 0$ mais cela ne peut être le cas, le modèle saturé étant un idéal inatteignable en pratique. La statistique $D^* = \frac{D}{\phi}$ est asymptotiquement de loi χ^2_{n-p-1}, avec n nombre d'observations et p nombre de variables explicatives. Ce résultat permet d'avoir un seuil critique au-delà duquel on considérera qu'un modèle n'est pas en adéquation avec les données. Un modèle sera alors considéré comme mauvais au seuil α si :

$$D_{obs} > \chi^2_{n-p-1;1-\alpha} \text{ , avec } \chi^2_{n-p-1;1-\alpha} \text{ quantile d'ordre } 1-\alpha \text{ d'une loi } \chi^2_{n-p-1}$$

Pour aller plus loin : Afin d'avoir une déviance proche de 0, on peut étudier les « résidus de variance », mais ceux-ci ne sont pas détaillés ici dans ce mémoire.

Nous avons ainsi présenté des méthodes d'estimation de paramètres d'un GLM et quelques éléments nous permettant de juger de sa qualité. Pour affiner encore, il existe des tests d'hypothèses et intervalles de confiance classiques permettant d'aller plus loin dans l'analyse, notamment les intervalles de confiance de Wald.

5.4 La modélisation de la probabilité de consommer dans l'année

Dans l'approche de tarification retenue, l'idée est de modéliser la probabilité de consommer dans l'année et ensuite la charge annuelle de consommation des assurés ayant « consommé » suivant leurs caractéristiques. Nous nous intéressons dans un premier temps à la probabilité de consommer.

5.4.1 Intuition

En pratique, on dispose de réalisations binaires relatives à l'occurrence d'un évènement ou non, et on chercher à modéliser la probabilité d'occurrence ($\mu_i = q_i$) de cet évènement sachant la catégorie ou classe de facteur explicatif. Il s'agit alors d'une variable d'intérêt à valeur dans [0,1] alors que le prédicteur linéaire ($\eta i = x_i^t \beta$) a des valeurs dans $\mathbb{R}$

Trois principales fonctions de lien sont alors utilisées pour se ramener de la droite des réels à [0,1] :

- **Le lien logit :**

$$g(q_i) = ln\left(\frac{q_i}{1-q_i}\right)$$

Il s'agit du lien canonique associée au modèle de régression binomiale.

- **Le lien probit :**

$$g(q_i) = \Phi^{-1}(q_i)$$

Φ étant la fonction de répartition de la loi normale centrée réduite.

- **Le lien log log complémentaire :**

$$g(q_i) = ln\big(-ln(1-q_i)\big) = \eta_i \iff q_i = 1 - exp\big(-exp(\eta_i)\big)$$

5.4.2 Introduction d'une variable latente pour la modélisation

Les modèles logit et probit peuvent être vus comme des modèles de régression linéaire utilisant une variable latente $Y^* \in \mathbb{R}$, la variable binomiale dont on cherche à modéliser la probabilité de succès étant $Y = 1_{[Y^* \geq 0]}$.
En considérant le modèle de régression linéaire $Y * = X\beta + \varepsilon$ avec ε bruit blanc gaussien, la probabilité de succès peut s'écrire :

$$
\begin{aligned}
q &= P[Y = 1|X = x] \\
&= P[Y *\geq 0|X = x] \\
&= P[X\beta + \varepsilon \geq 0|X = x] \\
&= 1 - \Phi^{-1}(-x^t\beta) \\
q &= \Phi(x^t\beta)
\end{aligned}
$$

Cette modélisation correspond donc au modèle probit qui est le modèle de régression binomiale avec Φ^{-1} comme fonction de lien.

En introduisant la fonction de répartition f suivante :

$$
f \left| \begin{array}{c} \mathbb{R} \to [0;1] \\ \mathrm{x} \to \dfrac{exp(x)}{1 + exp(x)} \end{array} \right.
$$

Un raisonnement analogue au précédent avec cette fois un modèle dont le résidu a une loi de fonction de répartition f mène au modèle de régression logistique, qui est le modèle de régression binomiale avec lien logit.

Nous choisissons comme modèle la régression logistique qui se formalise comme suit :

$$
g(q_i) = ln\left(\frac{q_i}{1 - q_i}\right) = \eta_i \iff q_i = \frac{exp(\eta_i)}{1 + exp(\eta_i)}
$$

Il s'agit d'un modèle de régression binomiale avec le lien logit. La loi binomiale est naturellement choisie pour la modélisation en raison du caractère binaire de la variable réponse.
Nous choisissons ensuite la fonction logit comme fonction de lien car il s'agit du lien canonique associé au modèle de régression binomiale et que cela induit des simplifications en termes d'estimation comme énoncé plus haut.

Estimations

Supposons que les réponses $y_1, \ldots, y_n$ soient des réalisations de variables aléatoire $Y_1, \ldots, Y_n$ de loi $Bin(m_i, q_i)$. La vraisemblance associée à ce modèle s'écrit donc :

$$
\mathcal{L}(\beta|y) = \prod_{i=1}^{n} \binom{m_i}{y_i} q_i^{y_i}(1 - q_i)^{m_i - y_i}
$$

En passant au logarithme, maximiser la vraisemblance revient à maximiser la quantité suivante selon β :

$$
l(\beta|y) = \sum_{i=1}^{n} \left[y_i ln\left(\frac{exp\ \eta_i}{1 + exp\ \eta_i}\right) + \left(m_{i-y_i}\right) ln\left(\frac{1}{1 + exp\ \eta_i}\right) \right]
$$

De sorte à estimer $\hat{q}_i = \frac{exp\,\hat{\eta}_i}{1+exp\,\hat{\eta}_i}$

Les grandes lignes de la modélisation de la probabilité de consommation étant introduites, nous nous intéressons dans la suite à la charge annuelle de consommation qui est la quantité complémentaire nous permettant d'établir des primes pures.

5.5 La modélisation de la charge annuelle de consommation

Nous nous intéressons dans cette section à la modélisation de la charge annuelle de consommation de la population assurée ayant « consommé ». Il s'agit d'une quantité strictement positive à laquelle il faut associer une loi de probabilité plausible pour la modélisation.

5.5.1 Le choix de la loi gamma

Les lois les plus courantes utilisées pour cette modélisation sont la loi gamma et la loi log-normale. Notre choix se porte sur la loi gamma, même si la loi log-normale aurait très bien pu être utilisée pour la modélisation.

La densité de la loi gamma peut s'écrire comme suit :

$$\frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu} exp\left(-\frac{\nu y}{\mu}\right) d(log\ y), \quad y \geq 0, \nu > 0, \mu > 0$$

Avec cette paramétrisation, la variable Y est d'espérance μ et de variance μ^2.

5.5.2 Le choix du lien log

Variables positives et lien logarithmique

Pour des variables d'intérêt positives, un lien qui est très souvent utilisé est le **lien logarithmique** de sorte que $\mu_i = exp(\eta_i)$. On parle alors de régression log-linéaire. Ce lien est très utilisé en pratique du fait que les effets additifs des variables sur le prédicteur linéaire η_i se transcrivent par des effets multiplicatifs sur la variable expliquée. Il est alors possible d'intégrer la positivité de variable d'intérêt tout en tenant compte de la possible négativité du prédicteur linéaire η_i.

Une lecture facilitée des tarifs par des facteurs multiplicatifs

Le principal avantage du lien logarithmique est d'ordre pratique. Prenons l'exemple d'un modèle de régression avec un lien logarithmique avec une variable explicative à *p* modalités. L'une des modalités est alors prise comme modalité de référence et le modèle de régression s'écrit :

$$log(\mu_i) = \beta_{ref} + \sum_{j=1}^{p-1} \beta_j \mathbf{1}_{\{i \in j\}}$$

Avec :

- p : nombre de modalités
- β_{ref} : coefficient associé à la modalité de référence,
- β_j $(j \neq ref)$: coefficients associés aux autres modalités
- $\mathbf{1}_{\{i \in j\}}$: Indicatrice valant 1 si la i^{ème} observation a la modalité j et 0 sinon

Après estimations des différents coefficients on arrive alors à :

$$\hat{\mu}_i = exp\left(\hat{\beta}_{ref} + \sum_{j=1}^{p-1} \hat{\beta}_j \mathbf{1}_{\{i \in j\}}\right)$$

Dans ce modèle simple, pour un individu ayant la modalité k ≠ ref, le tarif sera :

$$\hat{\mu}_i = exp\left(\hat{\beta}_{ref}\right) \times exp\left(\hat{\beta}_k\right)$$

Avec :

- $exp\left(\hat{\beta}_{ref}\right)$: Tarif de référence
- $exp\left(\hat{\beta}_{ref}\right)$: Correctif à appliquer pour les individus ayant la modalité k

La grille de tarification est alors facile à lire et à interpréter, l'idée étant de fournir un ajustement tarifaire en fonction d'une référence. Nous préférons ainsi le lien logarithmique au lien canonique de la loi gamma qui est la fonction inverse privilégiant ainsi l'interprétation à la simplification des estimations.

5.5.3 La régression log-gamma

Nous choisissons donc comme modélisation de la charge annuelle de sinistres, la régression log-gamma qui se formalise comme suit :

$$log(\mu_i) = X_i^t \beta$$

Estimations

Supposons que les réponses $y_1, ..., y_n$ soient des réalisations de variables aléatoire $Y_1, ..., Y_n$ de loi gamma avec de moyenne μ_i et de variance $\frac{\mu_i^2}{\nu}$. La vraisemblance associée à ce modèle pour l'estimation de μ_i s'écrit donc :

$$\mathcal{L}(\beta|y) = \prod_{i=1}^{n} \frac{1}{\Gamma(\nu)} \left(\frac{\nu y_i}{\mu_i}\right)^{\nu} exp\left(-\frac{\nu y_i}{\mu_i}\right) \frac{1}{y_i}$$

En passant au logarithme, maximiser la vraisemblance revient à maximiser la quantité suivante selon β :

$$l(\beta|y) = \sum_{i=1}^{n} \left[\nu \left(-\frac{y_i}{\mu_i} - ln\mu_i\right)\right]$$

Il s'agit donc de résoudre le système d'équation :

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \left[\nu \left(-\frac{y_i}{exp\, \eta_i} - \eta_i\right)\right] = 0 \iff \sum_{j=1}^{n} x_{ij} \left(1 - \frac{y_i}{\mu_i}\right) = 0, \quad j \in [0, p]$$

De sorte à estimer $\mu_i = exp(\hat{\eta}_i)$

Nous avons ainsi présenté les éléments théoriques liés aux modèles linéaires généralisés, dont les modèles linéaires gaussiens sont très utilisés car simples à mettre en œuvre. Cependant, comme précisé, ils sont souvent peu adaptés à la modélisation de variables d'intérêt en assurance, d'où l'intérêt des modèles linéaires généralisés.

Nous avons ensuite présenté la modélisation des variables d'intérêt retenues pour la tarification : La probabilité de consommation étant modélisée par une régression logistique et la charge annuelle de consommation par une régression log-gamma.

Nous allons désormais illustrer des applications des théories présentées par le biais de modélisations et de tarifications de différents postes médicaux.

5.6 La modélisation de la fréquence annuelle de consommation

Nous nous intéressons dans cette section à la modélisation de la fréquence annuelle de consommation de la population assurée. Il s'agit d'une quantité strictement positive à laquelle il faut associer une loi de probabilité plausible pour la modélisation.

Le choix de la loi de Poisson

Les lois les plus courantes utilisées pour cette modélisation sont la loi de Poisson et la loi binomiale négative. A noter que la loi binomiale négative évalue le nombre d'échecs avant l'obtention de n succès dans une expérience où la probabilité de succès est p. Elle peut aussi s'interpréter comme un mélange de lois de Poisson lorsque le paramètre lambda suit une loi gamma, ce qui s'interprète comme la prise en compte d'une hétérogénéité non observable. Notre choix se porte sur la loi Poisson.

La densité de la loi de Poisson peut s'écrire comme suit :

$$f(y \mid \lambda) = \frac{\lambda^y}{y!} e^{-\lambda} = \exp(y \log \lambda - \lambda - \log y!) \ , y \in \mathbb{N}$$

Avec cette paramétrisation, la variable Y est d'espérance λ et de variance λ.

Prise en compte de l'exposition et variable offset

Dans un modèle collectif, on a besoin de connaître le nombre de sinistres survenus sur une police d'assurance. Dans l'optique de tarifer un contrat, il faut pouvoir prédire le nombre de sinistres qui surviendront, en moyenne, l'année suivante. Or si certaines polices n'ont été observées que 6 mois dans la base, il convient de pondérer la fréquence de sinistre par l'exposition. Compte tenu de la propriété multiplicative d'un processus de Poisson, une police observée 1 an aura, en moyenne, 4 fois plus de sinistres qu'une police observée 3 mois. Dans le cas d'un modèle log-Poisson, il est alors naturel de supposer que :

$$Y|\boldsymbol{X} \ \sim \ \mathcal{P}(\exp[\boldsymbol{X}\beta + log(e)])$$

Où e désigne l'exposition, mesurée en année

5.7 La modélisation du coût moyen d'un acte

Nous nous intéressons dans cette section à la modélisation du coût moyen d'un acte. Il s'agit d'une quantité strictement positive à laquelle il faut associer une loi de probabilité plausible pour la modélisation. Pour des raisons identiques à celles développées précédemment pour la modélisation de la charge annuelle de consommation, nous retenons pour la modélisation du coût moyen une loi gamma et un lien log.

Nous allons désormais illustrer des applications des théories présentées par le biais de modélisations et de tarifications de différents postes médicaux.

6. Application

Dans ce chapitre, nous illustrons l'application des modèles linéaires généralisés à la tarification. Nous décrivons dans un premier temps le choix et la structuration des données retenues pour la modélisation, puis nous nous intéressons à la tarification de deux actes afin d'illustrer deux approches que nous avons souhaité développer. Le premier acte est la consultation d'un généraliste (« *Primary care Physician* ») que nous modéliserons en « Coût moyen x Fréquence », le deuxième acte est l'hospitalisation (« *Inpatient hospital services* ») que nous modéliserons en « Probabilité de consommer x Charge de consommation ». Enfin nous présenterons l'introduction de plafonds et franchise à la tarification.

6.1 Choix et structuration des variables tarifaires

L'Affordable Care Act étant régi par un dispositif fédéral, un certain nombre de contraintes s'imposent aux assureurs tant en termes de structuration des garanties proposées d'une part, qu'en termes de variables utilisables pour la tarification de la police d'un assuré d'autre part.

6.1.1 Définition des garanties à tarifer

La base de données des sinistres bien qu'extrêmement riche et détaillée ne nous permet pas de réaliser une agrégation simple par garantie. En effet, plusieurs champs pourraient en principe nous guider sur la nature de l'acte mais nous ne disposons pas d'une table permettant un passage aisé d'une description fine à une maille plus agrégée.

Parmi les champs éligibles, nous retenons notamment le champ de classification HCG.

La classification HCG ("Health Cost Guidelines") est une classification propriétaire développée aux Etats Unis par le cabinet de conseil actuariel Milliman. Cette classification est utilisée par la quasi-totalité des acteurs santé en plus de leur classification interne afin de faciliter les rapports réglementaires, les comparaisons entre données de marché, etc.

Nous disposons d'une table de référence des codes HCG listant les 143 codes et leur description. Ci-dessous à titre d'illustration un extrait de la table de correspondance :

Code HCG	Description
I11a	HIP Medical - General
O51a	HOP Preventive - General
P21a	PHY Maternity - Normal Deliveries

A partir de ces 143 codes, il convient d'effectuer des regroupements pour réduire à un tableau d'environ 10 à 15 garanties à tarifer pour être capable de proposer une offre similaire à l'exemple présenté en section 3.1.

Ci-dessous la distribution des principaux codes HCG dans la base sinistre par montant (« Allowed ») et quantité :

Code HCG	Montant de sinistres (%)
I12	13.1
I11a	8.4
O12a	5.5
P32d	4.3
P32c	4.1
O11a	4.0
O16a	3.9
P66	2.6
P63c	2.5
Autres	51.2

Code HCG	Nb de sinistres (%)
P63c	23.6
P32c	7.0
P32d	6.6
P63a	5.4
O15	5.2
Autres	51.8

Tableau : Distribution des sinistres par montant et quantité selon la classification « Health Cost Guidelindes »

On observe qu'un nombre limité de codes parmi les 144 concentre une partie importante des montants et des quantités (9 codes constituent plus de 50% des sinistres en montant, 5 codes constituent plus de 50% des sinistres en nombre).

A titre indicatif le sinistre les plus représentés en montant et nombre sont respectivement :

- I12 : Acte de chirurgie en hôpital
- P63c : Test en laboratoire indépendant

Une analyse détaillée des sinistres et des exercices de regroupement successifs conduit à établir le maillage des sinistres selon les 14 actes suivants :

Rang	Acte (garantie à tarifer)	Montant de sinistres (%)	Nb de sinsitres (%)
1	Inpatient hospital services	28.9	2.8
2	Outpatient surgery	19.1	8.1
3	Specialist office visit	10.1	18.1
4	Lab tests	9.0	43.2
5	Radiology & Imaging	7.6	5.9
6	Pharmacy	6.5	4.3
7	Emergency room	5.2	1.8
8	Maternity care	4.4	0.2
9	PCP office visit	4.1	7.0
10	Mental health services, Outpatient	2.7	3.8
11	Rehabilitation services	1.1	3.7
12	Urgent care	0.9	0.9
13	Others	0.2	0.0
14	Substances abuse services, Outpatient	0.1	0.2

6.1.2 Choix des variables explicatives pour la tarification

La réglementation de l'Affordable Care Act impose la liste des variables utilisables par les assureurs pour établir le prix de la police. Ces variables sont au nombre de cinq :

- **Age** : variable discrète positive
- **Fumeur vs Non fumeur** : variable catégorique a 2 niveaux
- **Classe de produit (niveau métallique)** : variable catégorique à 5 niveaux (Bronze, Silver, Gold, Platinum et Catastrophic)
- **Police individuelle / familiale** : variable catégorique à 2 niveaux
- **Lieux de résidence (état)** : variable catégorique

La contrainte fédérale présentée ci-dessus peut être ajustée au niveau de chaque Etat. En effet chaque Etat peut ajouter des limitations supplémentaires sur la manière donc les cinq variables peuvent influencer la tarification (ex : définir une limite maximum entre la différence de prix pour des fumeurs ou des non-fumeurs, etc..) .

Sans cette contrainte fédérale sur les variables explicatives, de nombreuses approches seraient envisageables afin d'optimiser le choix de ces variables. Parmi les approches que nous pourrions envisager, on peut citer :

- **Analyse en composante principale (ACP) :** l'analyse en composante principale est particulièrement adaptée lorsque l'on cherche à synthétiser de larges volumes de données. En effet, l'ACP permet de dégager des axes principaux (ou facteurs) qui sont des combinaisons linéaires des variables initiales, hiérarchisées et indépendantes les unes des autres.
- **K-Mean (ou K-moyenne) :** l'algorithme K-mean est un algorithme de classification non supervisée qui permet de regrouper des données « similaires » en groupes ou cluster dont le nombre est défini à l'initialisation. La base de cet algorithme repose sur le choix d'une distance, d'un nombre de groupe et d'une position initiale.
- **Classification ascendante hiérarchique** : ce type d'algorithme permet également de regrouper des données de dimension importante. La base de cet algorithme repose sur le choix d'une distance et d'un mode de regroupement. Les deux points les plus proches sont identifiés puis agrégés en un nouveau point selon la méthode définie. L'algorithme procède ensuite au rapprochement successif de tous les points. Un intérêt de ce type d'algorithme est qu'il permet par exemple de choisir ensuite aisément un nombre de groupes en fonction d'une distance.

A titre d'illustration nous avons mené une approche de classification hiérarchique afin de regrouper les pathologies selon leur coût moyen (pour chaque sinistre nous disposons de ou des pathologies associées, référencées selon la classification ICD 10 (*« International Classification of Diseases, revision 10 »*)).

Ainsi les 20 premières pathologies en 2017 en termes de dépenses sont indiquées dans le graphique ci-dessous. Les 3 premières correspondent respectivement à :

- Autisme infantile
- Test de détection de tumeur
- Insuffisance rénale

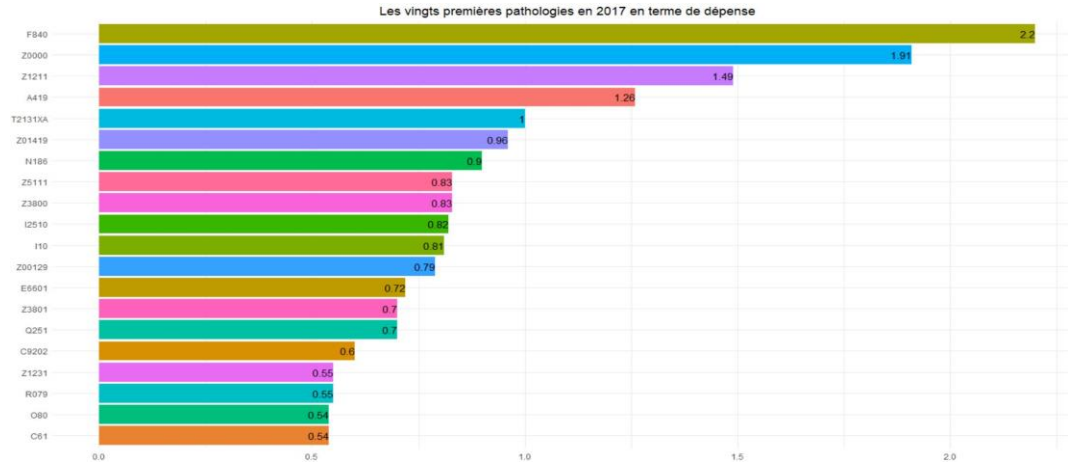


Figure : 20 premières pathologies en termes de dépense en 2017 selon classification ICD10

On peut ensuite établir la classification des pathologies telle qu'illustrée ci-dessous

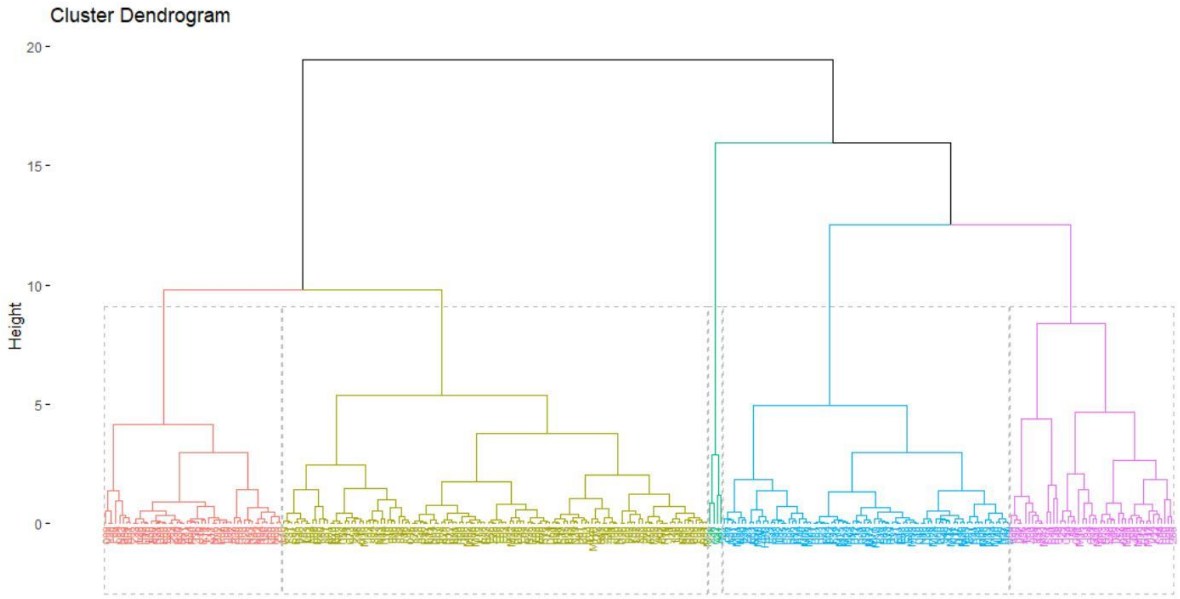


Figure : dendogramme de la classification ascendante hiérarchie des pathologies en fonction de leur coût moyen

6.2 La tarification d'un acte classique : la consultation d'un généraliste

6.2.1 Modélisation du coût moyen

Dans cette section nous modélisons le coût moyen d'une consultation d'un généraliste « PCP (Primary Care Physician) office visit ».

Le modèle

Nous implémentons ici une régression gamma avec une fonction de lien logistique. Les variables explicatives considérées sont l'âge, l'état, le fait d'être fumeur ou non, la classe de produit souscrite, l'année de soin.

Variable réponse	Coût moyen de l'acte, variable continue strictement positive
Variable explicatives	Etat, classe d'âge, fumeur/non fumeur, classe de produit, année de soin
Fonction lien	Fonction logarithme
Loi	Gamma
Poids	Temps de présence dans l'année

Table : Consultation d'un généralise « PCP Office visit » : explication du coût moyen

La validation du modèle

La régression effectuée sous R conduit au résultat ci-dessous :

```
Call :
glm(formula = tot_allowed ~ class_age + mem_state + tobacco_flag +
    mem_metal_rank + policy_year, family = Gamma(link = "log"),
    data = reg_tab_cout, weights = reg_tab_cout$weight)

Deviance Residuals :
     Min        1Q    Median        3Q       Max
-4.0605   -0.2107   -0.0316    0.1427    4.2421


 Coefficients :
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         4.697289  0.004508    1041.925 < 2,00E-16 ***
class_age[21,25)    0.076256  0.004594    16.598   < 2,00E-16 ***
class_age[25,30)    0.108691  0.003649    29.786   < 2,00E-16 ***
class_age[30,35)    0.112768  0.003567    31.613   < 2,00E-16 ***
```

class_age[35,40)	0.114148	0.003638	31.378	<	2,00E-16	***
class_age[40,45)	0.113361	0.003608	31.421	<	2,00E-16	***
class_age[45,50)	0.111221	0.003355	33.153	<	2,00E-16	***
class_age[50,55)	0.111774	0.003152	35.465	<	2,00E-16	***
class_age[55,60)	0.113140	0.003030	37.336	<	2,00E-16	***
class_age[60,65)	0.123284	0.002986	41.287	<	2,00E-16	***
class_age[65,100)	0.161677	0.008934	18.097	<	2,00E-16	***
mem_stateNJ	-0.478662	0.003990	-119.959	<	2,00E-16	***
mem_stateNY	-0.224815	0.003024	-74.348	<	2,00E-16	***
mem_stateOH	-0.304103	0.004885	-62.255	<	2,00E-16	***
mem_stateTN	-0.397940	0.004275	-93.081	<	2,00E-16	***
mem_stateTX	-0.163935	0.002936	-55.839	<	2,00E-16	***
tobacco_flagY	0.039551	0.005977	6.617		3.66e-11	***
mem_metal_rank2	0.005482	0.003723	1.473		0.140847	
mem_metal_rank3	-0.006389	0.003701	-1.726		0.084328	.
mem_metal_rank4	0.007670	0.003463	2.215		0.026793	*
mem_metal_rank5	-0.012653	0.003641	-3.475		0.000512	***
mem_metal_rank6	-0.012859	0.005226	-2.461		0.013869	*
policy_year2017	0.146322	0.002509	58.308	<	2,00E-16	***
policy_year2018	0.136306	0.002112	64.550	<	2,00E-16	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.1121907)

Null deviance : 27530 on 220979 degrees of freedom
Residual deviance : 23683 on 220956 degrees of freedom
AIC : 1793638

Number of Fisher Scoring iterations : 5

Table : Résultat de la régression sous R pour l’estimation du coût moyen de l’acte « Consultation généraliste »

Le résultat de cette régression appelle plusieurs commentaires.

En premier lieu sur la qualité globale du modèle, on constate une déviance de 27 530 pour 220 979 degrés de liberté. Ce niveau de déviance sur degré de liberté est relativement faible (12%), indiquant une bonne qualité d’ajustement du modèle. En effet la déviance est le double de la différence entre la log-vraisemblance du modèle saturé et celle du modèle proposé. Comme le modèle saturé ne dépend pas des paramètres du modèle, minimiser la variance revient à maximiser la vraisemblance du modèle proposé.

Dans un second temps, on s’intéresse au test de significativité des différentes variables, on constate qu’hormis certaines catégories de classe de produit, toutes les variables sont significatives.

Enfin on peut également s'intéresser au test de l'effet de chaque facteur par une analyse de type Anova au test du Chi2. On obtient ainsi le résultat suivant :

Analysis of Deviance Table

Model : Gamma, link : log

Response : tot_allowed

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			220979	27530		
class_age	10	253.87	220969	27276	< 2.2e-16	***
mem_state	5	3020.21	220964	24256	< 2.2e-16	***
tobacco_flag	1	3.16	220963	24253	1.121e-07	***
mem_metal_rank	5	16.41	220958	24236	< 2.2e-16	***
policy_year	2	553.18	220956	23683	< 2.2e-16	***

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

Table : Résultat de l'analyse ANOVA au test du Chi2 pour l'estimation du coût moyen de l'acte « Consultation généraliste »

Les résultats de test de significativité attestent de la pertinence du choix des variables pour l'explication de la variable d'intérêt. Chacune des variables participe pas à pas à une réduction de la déviance du modèle par rapport au modèle saturé.

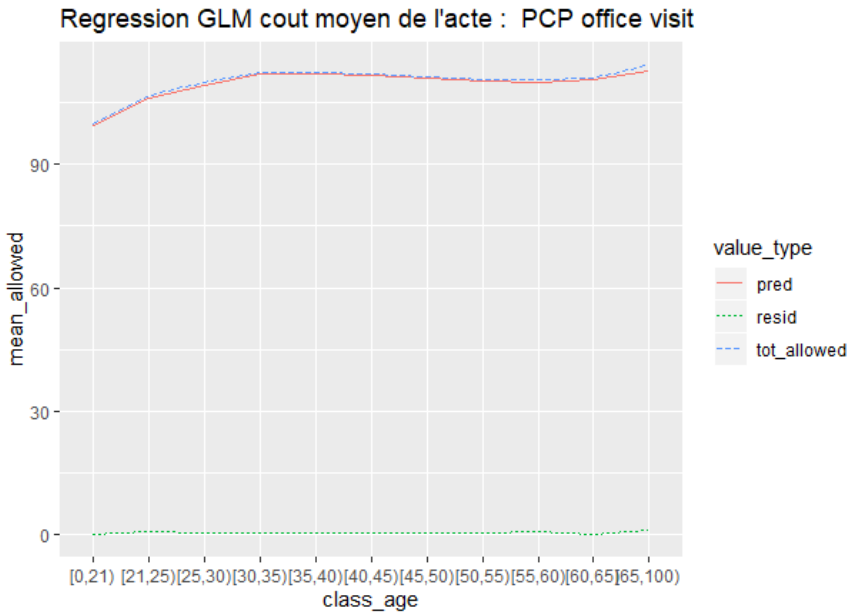


Figure : Coût moyen de l'acte Consultation Généraliste en fonction de la classe d'age (valeur réelle, valeur prédite, résidus)

La figure ci-dessus présente par classe d'âge le coût moyen constaté d'une consultation généraliste (courbe bleu), le coût moyen prédit (courbe rouge) et les résidus moyen (courbe verte). On observe ainsi que le coût moyen d'une consultation de généraliste est légèrement au-dessus de $100 et que ce coût évolue légèrement en fonction de l'âge.

6.2.2 Modélisation du taux d'incidence

Dans cette section nous modélisons le nombre moyen de consultations d'un généraliste « PCP (Primary Care Physician) office visit » par assuré.

Le modèle

Nous implémentons ici une régression de poisson avec une fonction de lien logistique. Les variables explicatives considérées sont l'âge, l'état, le fait d'être fumeur ou non, la classe de produit souscrite, l'année de soin.

Variable réponse	Taux d'incidence de l'acte, variable continue strictement positive
Variable explicatives	Etat, classe d'âge, fumeur/non fumeur, classe de produit, année de soin
Fonction lien	Fonction logarithme
Loi	Poisson
Offset	Temps de présence dans l'année

Table : Consultation d'un généralise « PCP Office visit » : explication du nombre de consultation par assuré

La validation du modèle

La régression effectuée sous R conduit au résultat ci-dessous :

```
Call :
glm(formula = tot_nb_claims ~ class_age + mem_state + tobacco_flag +
    mem_metal_rank + policy_year, family = poisson(link = "log"),
    data = reg_tab_nb, offset = log(weight))
```

Deviance Residuals :

Min	1Q	Median	3Q	Max
-1.1776	-0.5365	-0.1445	0.3257	11.

Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.163870	0.013324	12.299	< 2e-16	***
class_age[21,25)	-0.192628	0.014298	-13.473	< 2e-16	***
class_age[25,30)	-0.110813	0.011075	-10.005	< 2e-16	***
class_age[30,35)	-0.083720	0.010811	-7.744	9.65e-15	***

class_age[35,40)	-0.049837	0.010805	-4.612	3.98e-06	***
class_age[40,45)	-0.016901	0.010527	-1.606	0.1084	
class_age[45,50)	0.009024	0.009691	0.931	0.3518	
class_age[50,55)	0.071054	0.008951	7.938	2.06e-15	***
class_age[55,60)	0.098707	0.008570	11.518	< 2e-16	***
class_age[60,65)	0.128848	0.008393	15.351	< 2e-16	***
class_age[65,100)	0.151588	0.022962	6.602	4.07e-11	***
mem_stateNJ	0.130490	0.011829	11.032	< 2e-16	***
mem_stateNY	-0.041028	0.009466	-4.334	1.46e-05	***
mem_stateOH	0.246920	0.014121	17.486	< 2e-16	***
mem_stateTN	0.207070	0.012736	16.259	< 2e-16	***
mem_stateTX	0.181467	0.008938	20.303	< 2e-16	***
tobacco_flagY	0.039488	0.016075	2.457	0.0140	*
mem_metal_rank2	-0.121078	0.010604	-11.418	< 2e-16	***
mem_metal_rank3	-0.065594	0.010569	-6.206	5.42e-10	***
mem_metal_rank4	-0.194630	0.009997	-19.468	< 2e-16	***
mem_metal_rank5	-0.444520	0.010904	-40.765	< 2e-16	***
mem_metal_rank6	-0.386310	0.016904	-22.853	< 2e-16	***
policy_year2017	-0.016882	0.007201	-2.344	0.0191	*
policy_year2018	-0.054577	0.006060	-9.007	< 2e-16	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance : 112418 on 220979 degrees of freedom
Residual deviance : 103268 on 220956 degrees of freedom
AIC : Inf

Number of Fisher Scoring iterations : 5

Table : Résultat de la régression sous R pour l’estimation du taux d’incidence de l’acte « Consultation généraliste »

Le résultat de cette régression appelle plusieurs commentaires.

En premier lieu sur la qualité globale du modèle, on constate une déviance de 112 418 pour 220 979 degrés de liberté. Ce niveau de déviance sur degré de liberté est relativement faible (51%), indiquant une bonne qualité d’ajustement du modèle.

Dans un second temps, on s’intéresse au test de significativité des différentes variables, on constate qu’hormis certaines catégories de classe d’âge, toutes les variables sont significatives.

Enfin on peut également s’intéresser au test de l’effet de chaque facteur par une analyse de type Anova au test du Chi2. On obtient ainsi le résultat suivant :

Analysis of Deviance Table

Model : poisson, link : log

Response : tot_nb_claims

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			220979	112418		
class_age	10	2688.1	220969	109730	< 2e-16	***
mem_state	5	2535.8	220964	107194	< 2e-16	***
tobacco_flag	1	4.1	220963	107190	0.04233	*
mem_metal_rank	5	3831.7	220958	103358	< 2e-16	***
policy_year	2	90.5	220956	103268	< 2e-16	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table : Résultat de l’analyse ANOVA au test du Chi2 pour l’estimation du taux d’incidence de l’acte « Consultation généraliste »

Les résultats de test de significativité attestent de la pertinence du choix des variables pour l’explication de la variable d’intérêt. Chacune des variables participe pas à pas à une réduction de la déviance du modèle par rapport au modèle saturé.

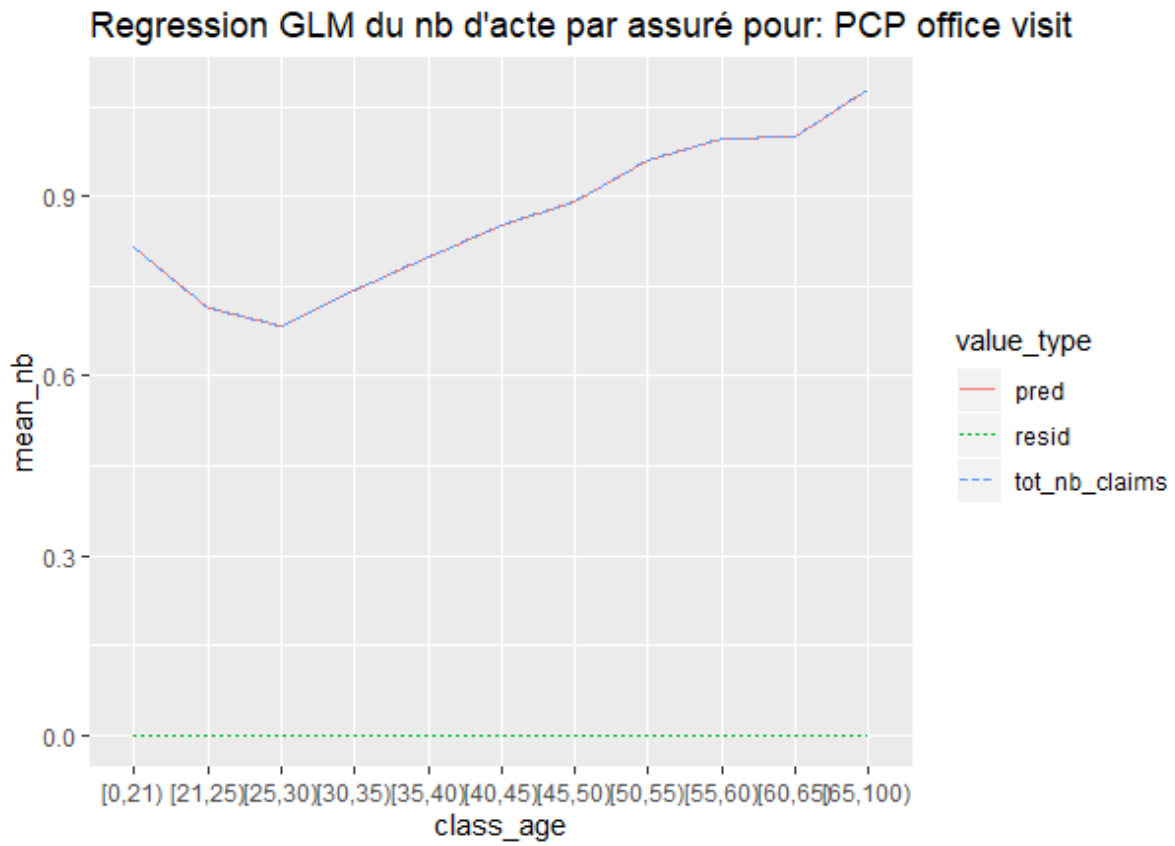


Figure : Nombre moyen d’acte par assuré pour l’acte Consultation Généraliste en fonction de la classe d’âge (valeur réelle, valeur prédite, résidus)

La figure ci-dessus présente par classe d'âge le nombre moyen d'acte par assuré, en valeur moyenne constatée (courbe bleu), valeur moyenne prédite (courbe rouge) et les résidus moyens (courbe verte). On observe ainsi que la fréquence moyenne de consultation d'un généraliste pour un assuré est d'environ 0,87 et augmente avec l'âge.

6.2.3 Modélisation de la prime pure

Une fois les coûts moyens et fréquence annuelle de consommation médicale projetées, il est possible de faire des projections de primes pures par multiplication des 2 quantités précitées.

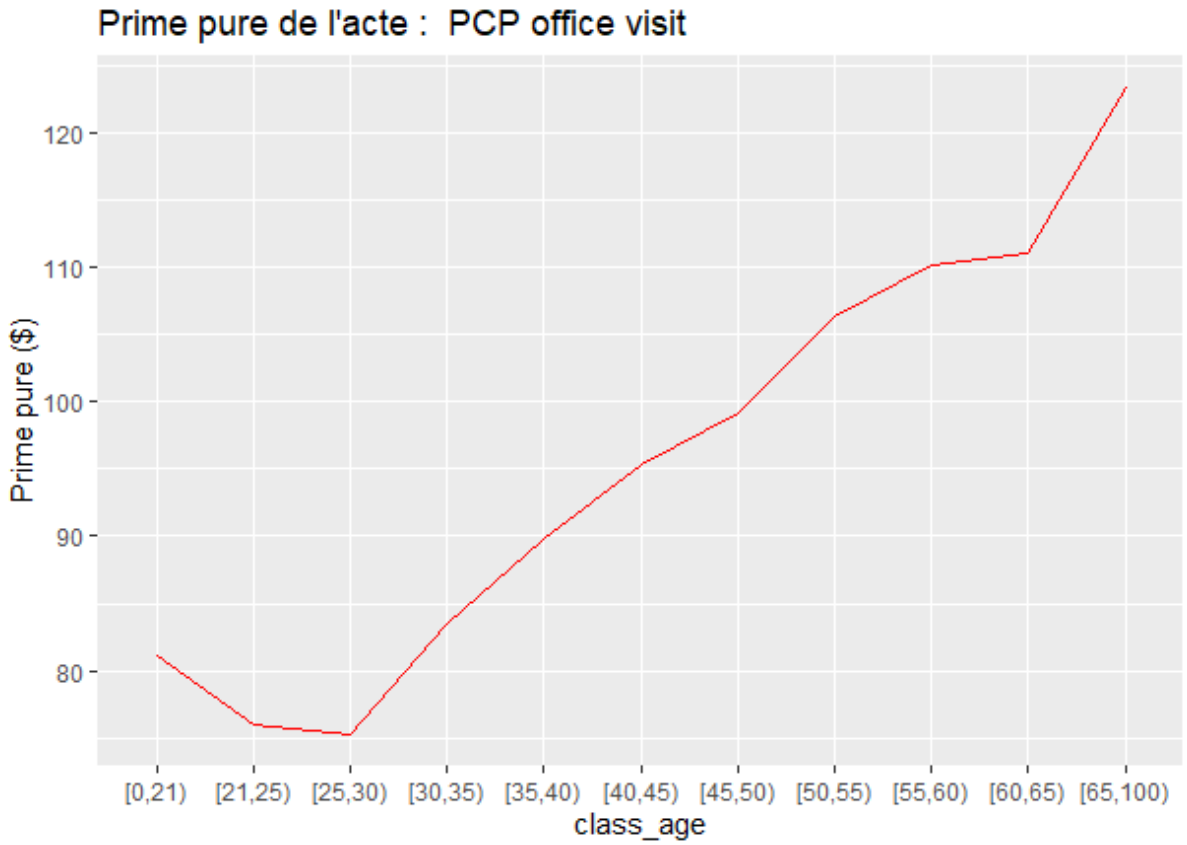


Figure : Prime pure de l'acte consultation généralise (« PCP office visit ») en fonction de l'age

Le graphique ci-dessus montre l'évolution de la prime pure de l'acte consultation généralise (« PCP office visit »). On observe comme on pouvait s'y attendre à une évolution croissante de cette prime pure en fonction de l'âge de l'assuré.

6.3 La tarification de la garantie hospitalisation

Pour la tarification de cette garantie nous utilisons l'approche « Probabilité de consommer x Charge de consommation »

6.3.1 Modélisation de la charge annuelle de consommation

Dans cette section nous modélisons la charge annuelle de consommation de l'acte « Inpatient hospital services »

Le modèle

Nous implémentons ici une régression gamma avec une fonction de lien logistique. Les variables explicatives considérées sont l'âge, l'état, le fait d'être fumeur ou non, la classe de produit souscrite, l'année de soin.

Variable réponse	Charge annuelle de consommation, variable continue strictement positive
Variable explicatives	Etat, classe d'âge, fumeur/non fumeur, classe de produit, année de soin
Fonction lien	Fonction logarithme
Loi	Gamma
Poids	Temps de présence dans l'année

Table : Garantie hospitalisation «Inpatient hospital services » : explication de la charge annuelle de consommation

La validation du modèle

La régression effectuée sous R conduit au résultat ci-dessous :

```
Call:
glm(formula = tot_allowed ~ class_age + mem_state + tobacco_flag +
    mem_metal_rank + policy_year, family = Gamma(link = "log"),
    data = reg_tab_cout)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9559  -2.4886  -1.6145  -0.0601  20.2682

Coefficients :
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)       10.01626  0.11676     85.785   < 2e-16  ***
class_age[21,25)  -0.36057  0.13565     -2.658   0.007864 **
class_age[25,30)  -0.38384  0.10265     -3.739   0.000185 ***
class_age[30,35)  -0.53917  0.09716     -5.549   2.90e-08 ***
class_age[35,40)  -0.22444  0.09894     -2.268   0.023313 *
class_age[40,45)  -0.08578  0.10032     -0.855   0.392510
class_age[45,50)  0.04592   0.09423     0.487    0.626053
class_age[50,55)  0.23008   0.08838     2.603    0.009242 **
class_age[55,60)  0.31662   0.08509     3.721    0.000199 ***
class_age[60,65)  0.41795   0.08282     5.047    4.53e-07 ***
```

class_age[65,100)	0.42217	0.17790	2.373	0.017648	*
mem_stateNJ	-0.57089	0.10353	-5.514	3.53e-08	***
mem_stateNY	-0.20045	0.08259	-2.427	0.015230	*
mem_stateOH	-1.30647	0.09526	-13.714	< 2e-16	***
mem_stateTN	-0.26534	0.11383	-2.331	0.019759	*
mem_stateTX	-0.50385	0.08088	-6.229	4.75e-10	***
tobacco_flagY	0.20822	0.11111	1.874	0.060941	.
mem_metal_rank2	-0.06556	0.07687	-0.853	0.393730	
mem_metal_rank3	-0.02190	0.08046	-0.272	0.785504	
mem_metal_rank4	0.07114	0.07732	0.920	0.357514	
mem_metal_rank5	0.08017	0.08144	0.984	0.324924	
mem_metal_rank6	0.29713	0.15781	1.883	0.059738	.
policy_year2017	-0.20829	0.05831	-3.572	0.000355	***
policy_year2018	-0.22761	0.05007	-4.546	5.49e-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 10.20919)

Null deviance: 122435 on 28343 degrees of freedom
Residual deviance: 116872 on 28320 degrees of freedom
AIC: 558069

Number of Fisher Scoring iterations: 14

Table : Résultat de la régression sous R pour l’estimation de la charge annuelle de consommation pour l’acte hospitalisation

Le résultat de cette régression appelle plusieurs commentaires.

En premier lieu sur la qualité globale du modèle, on constate une déviance de 122 435 pour 28 343 degrés de liberté. Ce niveau de déviance sur degré de liberté est relativement élevé (431%), indiquant une faible qualité d’ajustement du modèle.

Dans un second temps, on s’intéresse au test de significativité des différentes variables, on constate que certaines classes d’âges, l’état et l’année de soin sont significatives mais que la catégorie de produit et le facteur fumer/non fumeur sont peu significatifs.

Enfin on peut également s’intéresser au test de l’effet de chaque facteur par une analyse de type Anova au test du Chi2. On obtient ainsi le résultat suivant :

Analysis of Deviance Table

Model: Gamma, link: log

Response: tot_allowed

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)

NULL			28343	122435		
class_age	10	2162.48	28333	120273	< 2.2e-16	***
mem_state	5	3036.98	28328	117236	< 2.2e-16	***
tobacco_flag	1	42.89	28327	117193	0.04039	*
mem_metal_rank	5	91.61	28322	117101	0.11013	
policy_year	2	229.59	28320	116872	1.308e-05	***

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

Table : Résultat de l'analyse ANOVA au test du Chi2 pour l'estimation de la charge annuelle de consommation pour l'acte « Hospitalisation »

Les résultats de test de significativité indiquent que les variables classe d'âge, année de soin et état sont pertinentes et contribuent pas à pas à une réduction de la déviance du modèle par rapport au modèle saturé. Les variables fumeur / non fumeur et catégorie de produit apparaissent moins pertinentes pour notre modélisation des facteurs explicatifs de la charge annuelle de consommation pour l'acte Hospitalisation.

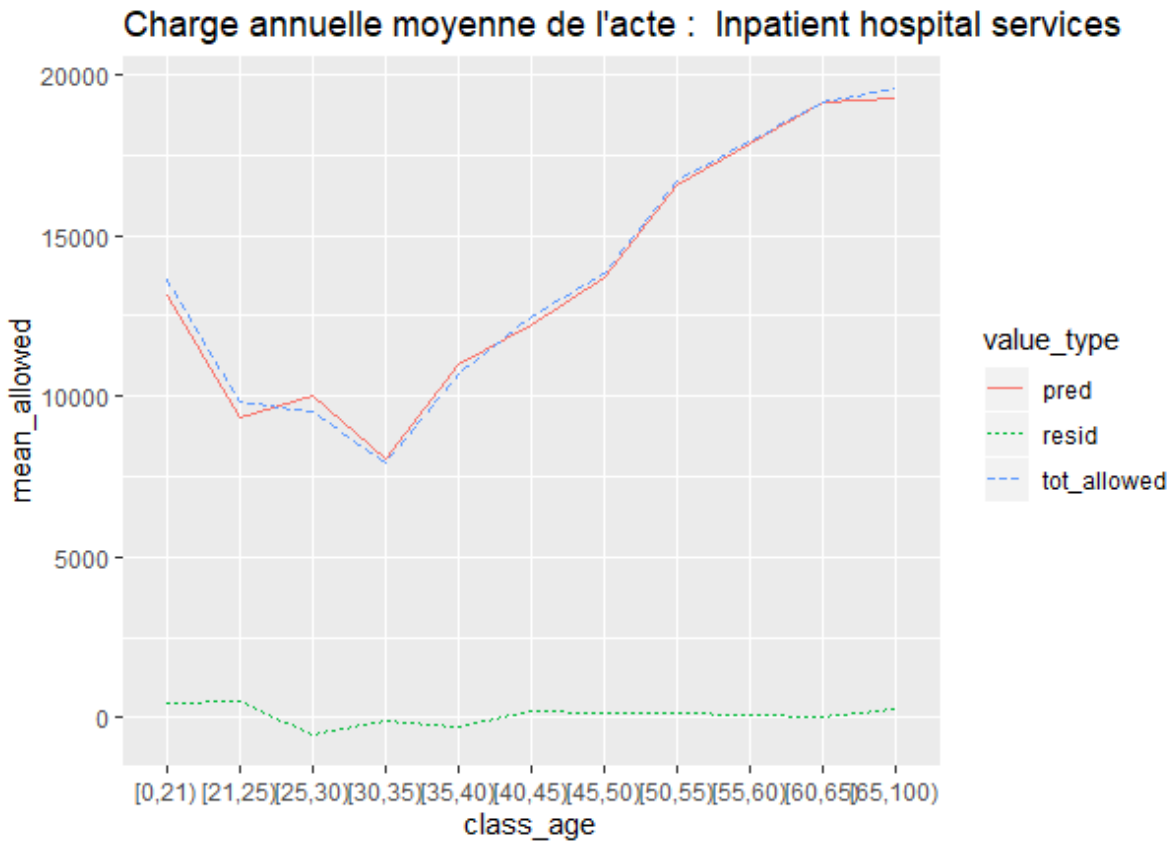


Figure : Charge annuelle moyenne pour la population consommante pour l'acte Hospitalisation en fonction de la classe d'âge (valeur réelle, valeur prédite, résidus)

La figure ci-dessus présente par classe d'âge la charge annuelle moyenne pour la population consommante pour l'acte Hospitalisation (courbe bleu), la charge moyenne prédite (courbe rouge) et les résidus moyen (courbe verte). On observe ainsi que la charge annuelle moyenne de consommation pour l'acte Hospitalisation est d'environ $ 14 700 et que cette charge évolue de manière importante en fonction de l'âge.

6.3.2 Modélisation de la probabilité de consommer

Dans cette section nous modélisons la probabilité de consommer par assuré pour l'acte Hospitalisation « Inpatient hospital services ».

Le modèle

Nous implémentons ici une régression logistique avec une loi binomiale et une fonction de lien logistique (logit). Les variables explicatives considérées sont l'âge, l'état, le fait d'être fumeur ou non, la classe de produit souscrite, l'année de soin.

Variable réponse	Variable binaire indiquant le fait d'avoir consommé ou non
Variable explicatives	Etat, classe d'âge, fumeur/non-fumeur, classe de produit, année de soin
Fonction lien	Fonction logit
Loi	Binomiale
Poids	Temps de présence dans l'année

Table : Hospitalisation « Inpatient hospital services » : modélisation de la probabilité de consommer dans l'année

La validation du modèle

La régression effectuée sous R conduit au résultat ci-dessous :

```
Call:
glm(formula = flag_garantie ~ class_age + mem_state + tobacco_flag +
    mem_metal_rank + policy_year, family = binomial(link = "logit"),
    data = reg_tab_proba)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2623  -0.3332  -0.2491  -0.1864   3.2185

Coefficients :
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)       -2.88954    0.03816  -75.716   < 2e-16  ***
class_age[21,25)   0.05824    0.04303    1.354   0.17587
class_age[25,30)   0.37937    0.03305   11.478   < 2e-16  ***
class_age[30,35)   0.61384    0.03111   19.730   < 2e-16  ***
class_age[35,40)   0.65565    0.03171   20.679   < 2e-16  ***
class_age[40,45)   0.65951    0.03208   20.561   < 2e-16  ***
class_age[45,50)   0.80832    0.03010   26.857   < 2e-16  ***
class_age[50,55)   1.00943    0.02829   35.677   < 2e-16  ***
class_age[55,60)   1.15844    0.02732   42.406   < 2e-16  ***
```

class_age[60,65)	1.26161	0.02664	47.353	< 2e-16	***
class_age[65,100)	1.40833	0.05848	24.081	< 2e-16	***
mem_stateNJ	0.18871	0.03357	5.622	1.89e-08	***
mem_stateNY	-0.15078	0.02688	-5.610	2.02e-08	***
mem_stateOH	2.11921	0.03174	66.777	< 2e-16	***
mem_stateTN	0.55233	0.03628	15.223	< 2e-16	***
mem_stateTX	0.35093	0.02597	13.514	< 2e-16	***
tobacco_flagY	0.36210	0.03709	9.764	< 2e-16	***
mem_metal_rank2	-0.55657	0.02532	-21.985	< 2e-16	***
mem_metal_rank3	-1.03394	0.02661	-38.854	< 2e-16	***
mem_metal_rank4	-1.26579	0.02562	-49.404	< 2e-16	***
mem_metal_rank5	-1.71103	0.02698	-63.423	< 2e-16	***
mem_metal_rank6	-2.03417	0.05078	-40.061	< 2e-16	***
policy_year2017	0.05691	0.01895	3.002	0.00268	**
policy_year2018	-0.09935	0.01639	-6.063	1.33e-09	***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 231584 on 634189 degrees of freedom
Residual deviance: 214100 on 634166 degrees of freedom
AIC: 214148

Number of Fisher Scoring iterations: 7

Table : Résultat de la régression sous R pour l'estimation de la probabilité de consommer dans l'année pour l'acte « Hospitalisation »

Le résultat de cette régression appelle plusieurs commentaires.

En premier lieu sur la qualité globale du modèle, on constate une déviance de 231 584 pour 634 189 degrés de liberté. Ce niveau de déviance sur degré de liberté est relativement faible (36%), indiquant une bonne qualité d'ajustement du modèle.

Dans un second temps, on s'intéresse au test de significativité des différentes variables, on constate qu'hormis certaines catégories de classe d'âge, toutes les variables sont significatives.

Enfin on peut également s'intéresser au test de l'effet de chaque facteur par une analyse de type Anova au test du Chi2. On obtient ainsi le résultat suivant :

Analysis of Deviance Table

Model: binomial, link: logit

Response: flag_garantie

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			634189	231584	

class_age	10	6819.7	634179	224765	< 2.2e-16	***
mem_state	5	4623.1	634174	220142	< 2.2e-16	***
tobacco_flag	1	72.9	634173	220069	< 2.2e-16	***
mem_metal_rank	5	5881.7	634168	214187	< 2.2e-16	***
policy_year	2	86.6	634166	214100	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table : Résultat de l’analyse ANOVA au test du Chi2 pour l’estimation de probabilité de consommer dans l’année pour l’acte « Hospitalisation »

Les résultats de test de significativité attestent de la pertinence du choix des variables pour l’explication de la variable d’intérêt. Chacune des variables participe pas à pas à une réduction de la déviance du modèle par rapport au modèle saturé.

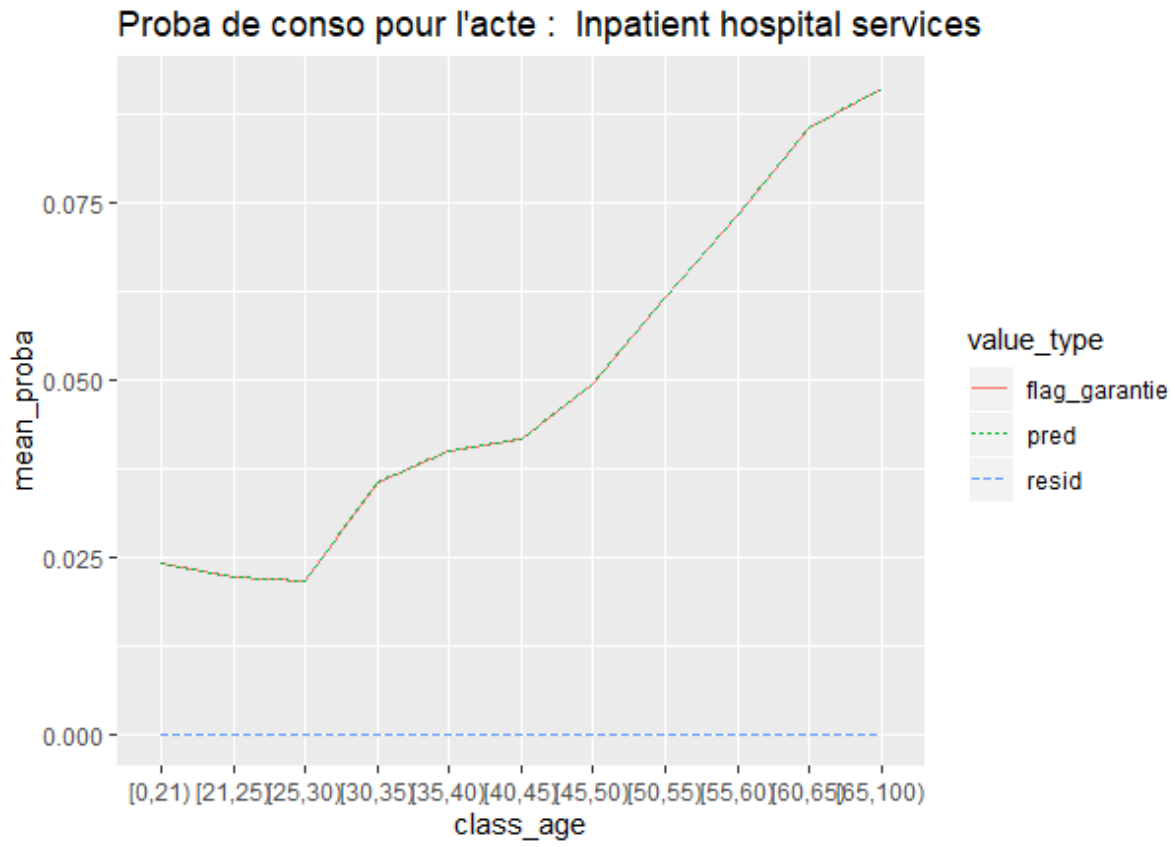


Figure : Probabilité de consommer dans l’année par assuré pour l’acte Hospitalisation en fonction de la classe d’âge (valeur réelle, valeur prédite, résidus)

La figure ci-dessus présente par classe d’âge la probabilité de consommer par assuré, en valeur moyenne constatée (courbe bleu), valeur moyenne prédite (courbe rouge) et les résidus moyens (courbe verte). On observe ainsi que la probabilité de consommer moyenne pour l’acte Hospitalisation pour un assuré est d’environ 4,5% et augmente avec l’âge.

6.3.3 Modélisation de la prime pure

Une fois les probabilités de consommer et la charge de consommation de la population consommante projetées, il est possible de faire des projections de primes pures par multiplication des 2 quantités précitées.

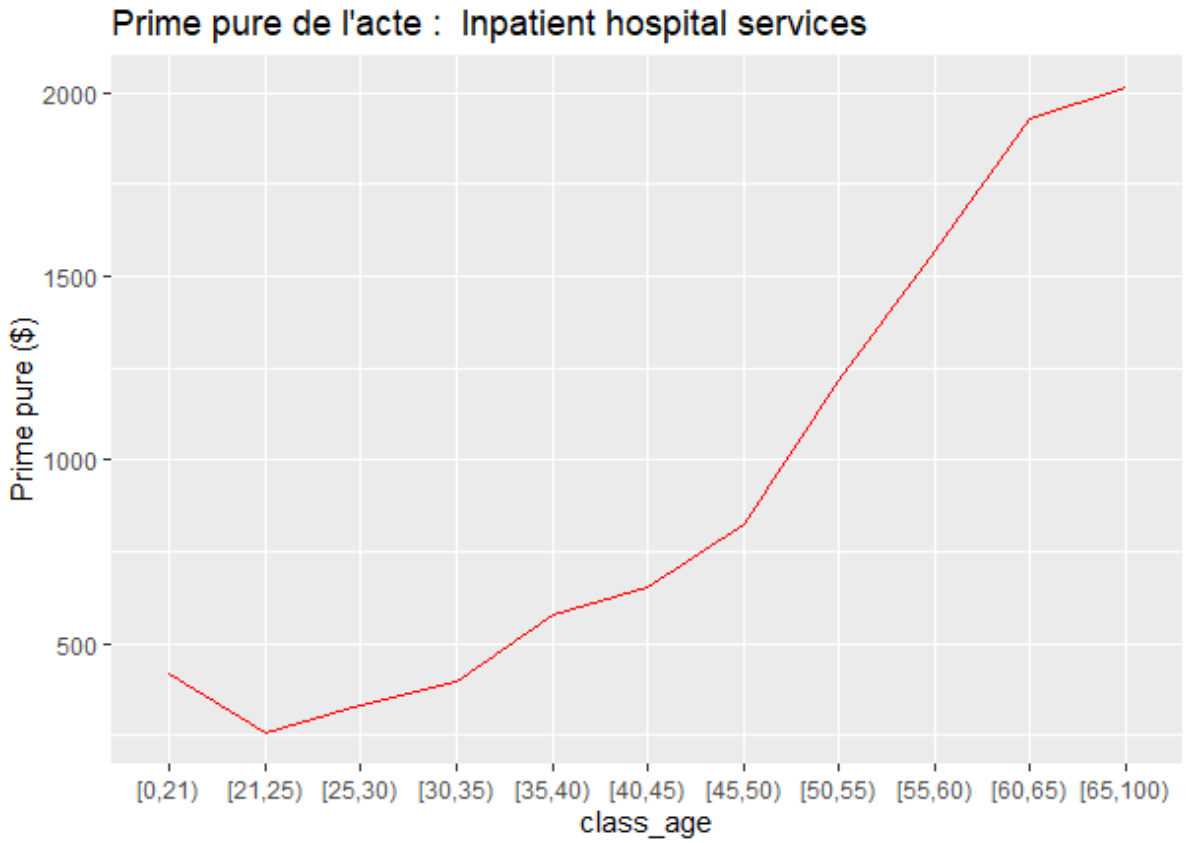


Figure : Prime pure de l'acte Hospitalisation (« Inpatient Hospital Services ») en fonction de l'âge

Le graphique ci-dessus montre l'évolution de la prime pure de l'acte Hospitalisation (« Inpatient hospital services »), on observe comme on pourrait s'y attendre à une évolution croissante de cette prime pure en fonction de l'âge de l'assuré et une prime pure moyenne aux environs de $650.

Pistes de réflexion pour aller plus loin

Les applications opérationnelles des modèles implémentés montrent des résultats satisfaisants via leur capacité de prédiction. Cela étant, d'autres approches et outils pourraient venir compléter, affiner et in fine améliorer ces premières modélisations. Plusieurs pistes sont explicitées par la suite en ce sens.

Le fait de tester plusieurs approches permet en effet d'identifier les points forts et limites de chaque modèle, afin de déterminer le modèle optimal à retenir, tout en tenant compte des contraintes métiers ou opérationnelles. Le modèle pourrait par ailleurs être confronté à celui d'ores et déjà en place. Les hypothèses liées aux lois utilisées pourraient être testées empiriquement, afin de mettre en exergue les éventuels inconvénients des approches proposées. Plusieurs lois ou fonctions liens des GLM pourraient être appliquées et étudiées comparativement. Le regroupement de certains actes, à partir de méthodes statistiques, pourrait également présenter un intérêt, tout comme l'étude de l'attritionnel et des atypismes. L'exploitation des franchises permettrait la mesure de l'antisélection. De même, il pourrait être intéressant de challenger les contraintes imposées par le gouvernement en testant d'autres variables explicatives. Par ailleurs, certains profils de risque pourraient être détectés et analysés plus spécifiquement.
Afin de valider le modèle, une base de test pourrait être mise en place (par ex. sur 1/3 des observations) et évaluée par rapport à une base d'apprentissage (composée des 2 autres tiers des observations). Enfin, un autre aspect qui serait intéressant à étudier : l'inflation (médicale et tarifaire).

Conclusion

Dans le cadre de ce mémoire, nous avons étudié la mise en place d'une tarification des garanties santé dans le cadre de l'Obamacare. Pour cela, nous disposions de bases de données d'expérience liées à l'exposition et la sinistralité de ~2millions de personnes assurées de 2016 à 2018 dans 6 états Américains. Pour la modélisation tarifaire, nous avons privilégié une approche mixte : Pour les actes de fréquence (ex. consultations généralistes ou spécialistes, analyses de laboratoires…), une approche « Fréquence x Coût moyen » ; et pour les actes à faible fréquence et charge importante (ex : hospitalisation, hospitalisation d'urgence…) une approche « Probabilité de consommer x Charge de consommation ».

Afin de mettre en œuvre les deux approches, nous nous sommes intéressés à la tarification de deux actes : Le premier acte étant la consultation d'un généraliste (« Primary care Physician ») que nous modélisons donc en « Coût moyen x Fréquence », le deuxième acte étant l'hospitalisation (« Inpatient hospital services ») modélisé en « Probabilité de consommer x Charge de consommation ». Nous avons en amont mené une approche de classification hiérarchique afin de regrouper les pathologies selon leur coût moyen.

Le coût moyen d'une consultation d'un généraliste « PCP (Primary Care Physician) office visit » a été modélisé via une régression gamma avec une fonction de lien logistique. Les variables explicatives considérées sont l'âge, l'état, le fait d'être fumeur ou non, la classe de produit souscrite, l'année de soin. Cette contrainte des variables utilisables par les assureurs pour établir le prix est liée à la réglementation de l'Affordable Care Act. La modélisation de la fréquence a été réalisée via une régression de poisson avec une fonction de lien logistique. La prime pure a été obtenue en multipliant fréquences et coûts moyens.

Pour la tarification de la garantie hospitalisation, nous utilisons l'approche « Probabilité de consommer x Charge de consommation », où nous modélisons la charge annuelle de consommation de l'acte « Inpatient hospital services » au travers d'une régression gamma avec une fonction de lien logistique. La probabilité de consommer par assuré est modélisée avec une régression logistique avec une loi binomiale et une fonction de lien logistique (logit). La projection de prime pure est ensuite obtenue par multiplication des probabilités de consommer et la charge de consommation de la population consommante projetées.

A noter que les hypothèses liées à l'utilisation de modèles GLM sont basés sur la maximisation de la fonction de vraisemblance en supposant l'indépendance entre les observations. Pour aller plus loin, il conviendrait de vérifier dans quelle mesure ces hypothèses sont vérifiées dans le jeu de données à disposition. Par ailleurs, l'utilisation de méthodes GEE permet de tenir compte de cette corrélation en intégrant une matrice de corrélation aux équations de vraisemblance. Une fois les primes pures obtenues, des plafonds et franchises annuels pourront été intégrés à la tarification par troncature des lois modélisées par segment tarifaire.

Références et bibliographie

BARNETT J., BERCHICK E., HOOD E. (2018) Health Insurance Coverage in the United States: 2017. In “Current Population Reports” P60-264. U.S. Government Printing Office, Washington D.C.

BEAUSSIER A. (2016) La santé aux Etats Unis: une histoire politique. Presse de Sciences Po, Paris.

BUTLER S., HAISLMAIER E. (1989) A National Health System for America. The Heritage Foundation, Washington D.C.

CHARPENTIER A. (2010) Statistique de l’assurance. Université de Rennes 1 et Université de Montréal.

CHARPENTIER M., DENUIT M. (2004) Mathématiques de l'assurance non-vie : Tome 1 : Principes fondamentaux de théorie du risque. Economica, Paris.

CHARPENTIER M., DENUIT M. (2005) Mathématiques de l’assurance non-vie : Tome 2 : Tarification et Provisionnement. Economica, Paris.

DAEMMRICH A. (2011) U.S. Healthcare Reform and the Pharmaceutical Industry. Working paper. Harvard Business School, Boston.

DIOMANDE A. (2013) Tarification de garanties santé liées à un portefeuille d'expatriés. Mémoire d’actuariat, Paris.

DROESBEKE J., LEJEUNE, M., SAPORTA G. (2005) Modèles statistiques pour données qualitatives. Editions TECHNIP, Paris.

MC CULLAGH P., NELDER J. (1989) Generalized Linear Models. Springer, New-York.

OHLSON E., JOHANSSON B. (2010) Non-Life Insurance Pricing with Generalized Linear Models. Springer, New-York.

PENG R. (2019) R Programming for Data Science. Leanpub, Victoria.

PENG R. (2016)] Exploratory Data Analysis with R. Leanpub, Victoria.

RIGOLLET P. (2016) 18.650 Statistics for Applications - Fall 2016. Massachusetts Institute of Technology, Boston.

SEZER M., BAUER F. (2017) Introduction to the U.S. Health Care System. In “Crossing Borders - Innovation in the U.S. Health Care System” coordonné par SCHMID A., SINGH S. Bayreuth.

https://aspe.hhs.gov/poverty-guidelines (Normes de pauvreté publiées par le U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES), site consulté le 10 février 2020.

https://www.cms.gov/CCIIO/Resources/Regulations-and-Guidance/Downloads/Final-2019-AV-Calculator-Methodology.pdf (Méthodologie de calcul de la valeur actuarielle pour les produits d’assurance proposés dans le cadre de l’Affordable Care Act), site consulté le 8 mars 2018.

https://www.kff.org/health-reform/fact-sheet/summary-of-the-affordable-care-act (Présentation synthétique de l'Affordable Care Act), site consulté le 12 juin 2018.

https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-growth-opportunity-for-private-health-insurance-companies (Les opportunités de croissance pour les compagnies d'assurance privées), site consulté le 19 mars 2018.