

Proposition d'un modèle de tarification en assurance maladie obligatoire à travers le Modèle Linéaire Généralisé, EL KASSIMI, F.¹ et ZAHI, J.²

1. Doctorante, laboratoire de modélisation mathématique et calcul économique, université Hassan premier, f.elkassimi@uhp.ac.ma.
2. Professeur d'enseignement supérieur, université Hassan premier, zahi_ja@yahoo.fr.

Date de soumission : 18/08/2022

Date d'acceptation : 26/10/2022

Résumé :

L'assurance maladie obligatoire est destinée à couvrir, en termes de dépenses et soins médicaux, un ensemble hétérogène d'assurés concernant leurs états de santé ; on pourrait assister à différents niveaux de risque et à des états de santé très variés, allant des bien-portants jusqu'aux personnes sujettes à des niveaux de risque plus élevés, notamment les patients atteints de pathologies chroniques. Cependant les prélèvements obligatoires sont recouverts indépendamment de l'état de santé des assurés, faisant supporter le coût des soins prodigués à des personnes à faible risque au lieu de celles à fort risque. Or, ces prélèvements doivent être assis sur le risque que présente l'assuré, de façon à ce que le taux de cotisation soit proportionnel au risque qu'il fait supporter à l'assurance.

L'objectif de cet article est de proposer un modèle de tarification en assurance maladie, moyennant les modèles traditionnels dits des modèles linéaires généralisés, et ce on se basant sur les caractéristiques des assurés.

Les résultats auxquelles nous nous sommes parvenus démontrent l'effet des facteurs de risques tels que le sexe du bénéficiaire, son âge ou son atteinte d'une affection de longue durée sur les tarifs.

Mots- clés : Assurance non-vie, tarification en assurance maladie, variables tarifaires, modèles linéaires généralisés.

Proposal of a pricing model in compulsory health insurance through the Generalized Linear Model

Abstract:

Compulsory health insurance is intended to cover, in terms of medical care and expenditures, a heterogeneous set of insureds with regard to their health status; there could be different levels of risk and a wide range of health conditions, from the healthy ones to those subject to higher levels of risk, especially patients with chronic conditions. However, compulsory levies are collected independently of the health status of the insured, making low-risk people bear the cost of care instead of high-risked ones.

These levies should be based on the risk of the insured, so that the premium rate is proportional to the risk that the insured is bearing.

The aim of this article is to propose a pricing model for health insurance, using the traditional generalized linear models, based on the characteristics of the insured.

The results show the effect of risk factors such as the beneficiary's gender, age or long-term illness on the tariffs.

Key words: Health insurance, health insurance pricing, pricing variables, Generalized Linear Models.

Introduction :

Généralement l'assurance maladie obligatoire est destinée à couvrir, en termes de dépenses et soins médicaux, un ensemble d'assurés hétérogènes concernant leurs états de santé ; on pourrait assister à différents niveaux de risques et à des états de santé très variés, allant des bien-portants jusqu'aux personnes sujettes à des niveaux de risque très élevés, notamment les patients atteints de pathologies chroniques, ou encore d'une affection de longue durée « ALD ». Cependant les prélèvements obligatoires sont recouverts indépendamment de l'état de santé des assurés, faisant supporter le coût des soins prodigués à l'assureur qui porte à sa charge l'intégralité du risque. Or ces prélèvements doivent être assis sur le risque que présente l'assuré, de façon à ce que le taux de cotisation soit proportionnel au risque qu'il fait supporter à l'assurance.

C'est dans cet esprit actuariel, et en se basant sur le constat que le système de tarification en vigueur des mutuelles de santé n'est pas optimal (OCDE ,2020) que le présent travail répond à la problématique suivante : quel modèle de tarification devrait être mis en place en vue d'instaurer un système d'assurance maladie obligatoire équitable au niveau des mutuelles de santé au Maroc ?

Pour répondre à cette question principale de recherche, notre article combine harmonieusement la théorie et la pratique en s'articulant autour de six sections. Nous nous proposons, dans le présent article, de construire un modèle de tarification optimal basé sur les caractéristiques des adhérents y compris leurs états de santé, ainsi que sur le nombre de prestations de soins et de prises en charge (sinistres). Le but de cette tarification étant de subdiviser les adhérents en plusieurs catégories de telle sorte qu'à l'intérieur d'une catégorie, les adhérents sont plus au moins homogènes en termes de caractéristiques, et de risque bien entendu. L'hétérogénéité au sein d'un portefeuille pose l'assurance face au problème d'anti sélection qui est généralement le résultat de l'application d'une prime unique à l'ensemble du portefeuille, chose qui fait que les mauvais risques vont s'assurer, tandis que les bons vont être découragés par les primes onéreuses, ce qui emmène à une dégradation du résultat Denuit et Charpentier (2004).

Cette tarification dite *a priori* va permettre d'ajuster les cotisations individuelles selon le degré de risque intrinsèque, de sorte que chaque assuré paye une cotisation proportionnelle à sa fréquence de sinistre (prestation) et son coût moyen (la sévérité du sinistre) (Lotsi, Mettle, & Adjorlolo, 2019), tout en prenant en compte la solvabilité de la mutuelle, qui doit préserver son équilibre financier dans le temps. Le modèle proposé s'inscrit dans une approche actuarielle traditionnelle bien fondée et documentée. En s'inspirant de nombreux travaux de recherches, (Fox (2016) ; Ohlsson et Johansson (2010) ; Lemaire (1995) ; Dionne et Vanasse, (1992) ; (Frees, Lee, & Yang, 2016) (Gao, Meng, & Wuthrich, 2018) Pitrebois et al. (2003). Les données sur lesquelles nous allons nous baser pour construire notre modèle sont celles d'une mutuelle de santé qui détient une part importante du marché des mutuelles de santé au Maroc.

Elle sera élaborée dans le présent article une tarification *a priori* à l'aide des modèles GLM, en effet cet article sera dédié à la modélisation de la cotisation de la branche d'assurance maladie de base.

Notre article sera organisé comme suit : dans une première section nous allons présenter le modèle GLM, ainsi que la tarification en assurance maladie tout en s'attardant sur la définition de la prime pure ainsi que les facteurs de risque. Dans une deuxième section nous présenterons notre modèle d'analyse y compris le développement des hypothèses et les modèles conceptuels, il sera également lieu de présenter notre portefeuille et les traitements apportés aux données afin de sélectionner les variables tarifaires. la troisième section sera consacrée à l'ajustement du modèle GLM des coûts moyens aux différentes loi de distributions afin de choisir la mieux adaptée. La section quatre sera dédiée à ajuster le modèle des fréquences tout en adoptant le même procédé et pour l'analyse la performance des deux modèles. Tandis que la dernière section sera dédiée à présenter les résultats finals d'estimation des coûts moyens et des fréquences ainsi que la prime pure.

1. Modèle linéaire généralisé « GLM »

1.1. Présentation du Modèle linéaire généralisé GLM

Depuis le début des années 90, l'outil principal utilisé dans la tarification est celui du Modèle Linéaire Généralisé (GLM), développé par Nelder et Wedderburn (1972), il est considéré comme la norme standard de tarification McCullagh et Nelder (1989) ; Brockman et Wright (1992). Dans leurs travaux initiaux, Nelder et Wedderburn (1972) étendent les techniques de régression linéaire en permettant à la variable objective d'avoir une distribution de la famille exponentielle, notamment les distributions Gamma et Poisson (voir El kassimi et Zahi (2021) pour plus de détail). En fait, la sévérité et la fréquence des sinistres prennent toutes les deux des valeurs positives et leurs distributions sont asymétriques. Tandis que la sévérité des sinistres est souvent représentée par une distribution Gamma, ou de Log normale. Les distributions Binomiale et de Poisson sont particulièrement intéressantes pour ajuster le nombre de sinistres Ohlsson et Johansson (2010) ; Brisard (2014) ; Bellina (2014).

Les GLM sont une extension du modèle linéaire classique, qui est défini par l'équation :

$$y_i = \sum \beta_k x_{ik} + \epsilon_i \quad (\forall i \in [1, n])$$

Où n est le nombre d'observations.

Généralement le Modèle linéaire généralisé est constitué de trois composantes (Fox, 2016).

- Une composante aléatoire, qui spécifie la distribution conditionnelle de la variable endogène y_i (ième des n observations), étant donné les valeurs des variables explicatives k. Ainsi, la composante aléatoire spécifie la distribution de $E[y_i | \{x_1, \dots, x_k\}]$. Cette distribution appartient généralement à la famille exponentielle.

- Un prédicteur linéaire, ou encore (composante systématique), qui est donné par :

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_k X_{ik}$$

- Une fonction de lien inversible, dite canonique qui relie la composante aléatoire au prédicteur linéaire. En d'autres termes, la fonction de lien transforme la valeur attendue de la variable dépendante¹, $\mu_i = E[y_i | \{x_1, \dots, x_k\}]$ en la composante linéaire. Ainsi, on a

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_k X_{ik}$$

Comme la fonction de liaison est inversible, on obtient $\mu_i = g^{-1}(\eta_i)$.

Une propriété importante et utile des distributions de la famille exponentielle est que la variance conditionnelle de y_i est fonction de sa moyenne μ_i et dans certains cas, d'un paramètre de dispersion constant. La fonction principale de ce paramètre de dispersion est d'indiquer la distribution spécifique qui est utilisée Fox (2016) et Navarun (2018).

1.2. Tarification d'un produit d'assurance santé

Après que nous avons introduit les modèles linéaires généralisés et que nous avons expliqué la structure de cette approche actuarielle, le reste de l'article sera consacré à la modélisation de la prime pure pour une mutuelle de santé. En effet, nous allons tout d'abord commencer par définir la prime pure et préciser notre démarche de tarification, puis l'étape suivante sera dédiée à la sélection des variables dites de tarification, ensuite nous allons segmenter le portefeuille étudié en nous basant sur les variables tarifaires sélectionnées. Cette segmentation de la population va nous permettre d'offrir des tarifs diversifiés pour les adhérents de la branche maladie, et ceci selon la structure de la population en matière de la sinistralité.

1.2.1. Prime PURE

La prime pure représente le risque économique qui est transféré de l'assuré à l'assureur. Selon la loi des grands nombres, la perte totale de la compagnie d'assurance (correspond à la somme d'un grand nombre de pertes indépendantes relativement petites), devrait être plus facile à prévoir que la valeur d'une perte individuelle (Ohlsson et Johansson (2010)). Cela signifie que la perte réelle ne doit pas être excessivement éloignée de sa valeur prévue. Raison pour laquelle les actuaires définissent la prime pure comme le coût espéré de tous les sinistres que les assurés vont probablement subir pendant la période de couverture (Denuit et al. (2007)).

Le modèle de tarification que nous allons adopter pour modéliser la prime pure, et celui utilisant l'approche *fréquence × coût moyen* (Sakthivel & Rajitha, 2017) (Lotsi, Mettle, & Adjorlolo, 2019) (Gao, Meng, & Wuthrich, 2018), en effet les actuaires ont coutume d'évaluer la prime pure en utilisant des techniques de régression (Denuit et Lang (2001)). Ces modèles classent les assurés selon leurs caractéristiques en termes de risque (Antonio et Valdez, 2012) ; (Denuit et al. ,

¹ Par conséquent, un GLM peut être interprété comme un modèle de régression non linéaire pour la variable objective.

2007);(Paefgen et al. , 2013) ; (McClenahan, 2001) et fournissent des informations sur les charges spécifiques par classe de risque. Ainsi, la fréquence (F_i) et le coût moyen encore dit « sévérité » (S_i) sont généralement supposées être indépendants, et la prime pure qui en résulte est le produit de la valeur attendue de la fréquence multipliée par celle de sévérité (Denuit et al., 2007) ; (Ohlsson et Johansson (2010) ; (Klugman et al., 2012); (Frees et al., 2016).

Avant de spécifier la prime pure, il convient tout d'abord, de définir la fréquence (F_i) et la sévérité des sinistres (S_i) pour chaque assuré i . La fréquence des sinistres représente le rapport du nombre total de sinistres N_i déclarés pendant une période de couverture par l'exposition totale t_i :

$$F_i = \frac{N_i}{t_i}$$

Où t_i désigne la fraction de la période de la police pour laquelle le souscripteur est assuré (e.g., une exposition $t_i=1$ représente une année complète lorsque N_i est donné pour un contrat d'un an entier). La sévérité des sinistres (S_i) représente le coût moyen des sinistres, exprimé comme le rapport entre la perte totale (L_i) de chaque assuré et le nombre correspondant de sinistres à l'origine de cette perte totale. Elle correspond à l'espérance du remboursement.

$$S_i = \frac{L_i}{N_i}$$

Maintenant, la prime pure (p_i) de l'assuré i peut être définie comme le rapport de la fréquence des sinistres (F_i) et le coût moyen des sinistres (S_i) :

$$p_i = \frac{L_i}{t_i} = \frac{N_i}{t_i} \times \frac{L_i}{N_i} = F_i \times S_i$$

En supposant l'indépendance entre la fréquence et le coût moyen des sinistres, on aura :

$$\pi_i = E(p_i) = E(F_i)E(S_i)$$

Ainsi, en multipliant les estimations des deux modèles, on obtient la prime pure à facturer pour chaque catégorie des assurés.

Le modèle multiplicatif de cette équation est souvent préféré aux modèles qui évaluent la prime pure globale $\left(\frac{L_i}{t_i}\right)$, du fait qu'il permet de mieux comprendre le processus de sinistre.

Premièrement, cette approche fournit une meilleure compréhension des facteurs de risque sous-jacents affectant la sévérité et la fréquence des sinistres. Lorsque l'on modélise directement les primes pures, certains effets intéressants peuvent disparaître en raison d'effets opposés sur ses composantes : par exemple, une variable qui a un impact négatif important sur la fréquence mais un impact positif équivalent sur le coût moyen passerait totalement inaperçue.

Deuxièmement, la sévérité et la fréquence des sinistres suivent généralement des distributions différentes.

Une fois la prime pure est calculée, une marge de risque prenant en compte le risque de modèle et l'aléa pur, ainsi que d'autres éléments de prime (à titre d'exemple, le bénéfice, les commissions, les taxes), sont ajoutés à la prime pure pour aboutir à un tarif commercial (Wüthrich, 2016).

1.2.2. Facteurs de risque

Les modèles de coût moyen et de fréquence pour chaque assuré, sont estimés en incluant les facteurs de risque comme variables explicatives dans les modèles.

Les facteurs de risque les plus couramment utilisés pour déterminer les tarifs de l'assurance maladie sont : L'âge ; Le sexe ; Les informations supplémentaires sur l'assuré : état civil, l'indice de masse corporelle « IMC », tabagisme, maladie préexistante (affection de longue durée) ; Le Poste de soins ; Les informations sur la région géographique dans laquelle réside l'assuré (risques spatiaux). Ces facteurs de risque nous ont conduits à formuler un ensemble d'hypothèses que nous attarderons à présenter dans la section qui suit.

2. Modèle d'analyse

2.1. Développement des hypothèses et modèles conceptuels

Afin de modéliser nos deux modèles de fréquence et de coût moyen, il sera objet de vérifier l'ensemble des hypothèses suivantes.

2.1.1. Développement des hypothèses

H1. L'âge est une fonction croissante de la fréquence des prises en charges, et du coût moyen et donc de la prime pure.

H2. Le fait d'être de sexe féminin augmente la fréquence de consommation des prises en charge ainsi que le coût moyen par acte, est donc présente un facteur d'accroissement de la prime pure.

H3. Un état de santé dégradé, plus particulièrement la présence d'une ALD, conduit à une prolifération des prises en charges, ainsi qu'une augmentation des charges par acte, ce qui engendre un accroissement des tarifs.

H4. Les prises en charges consommées par l'adhérent lui-même sont les plus fréquentes et dont le coût moyen par acte est très élevé, cela revient à supposer que l'attribut : « adhérent » du type de bénéficiaire engendre des tarifs élevés.

H5. La catégorie « retraité » est un facteur d'accroissement à la fois de la fréquence et du coût moyen des prises en charge, la catégorie « retraité » est un facteur d'accroissement des tarifs.

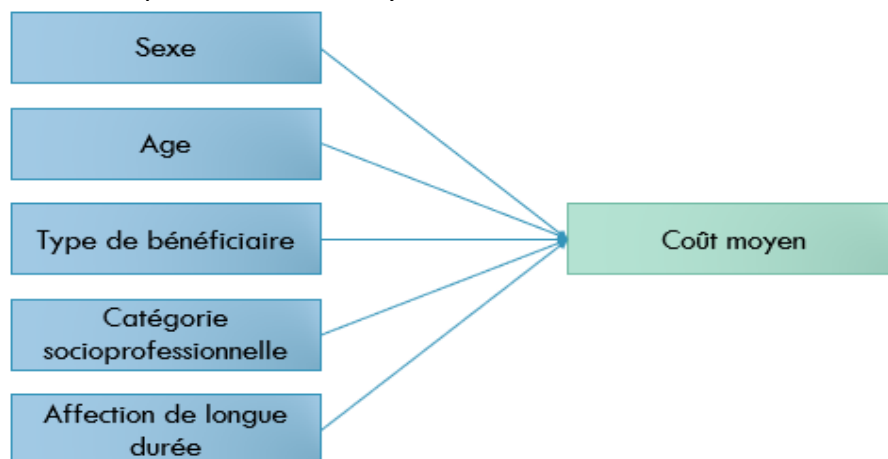
L'ensemble de ces hypothèses nous a permis de concevoir les modèles conceptuels abordés ci-après.

2.1.2. Modèles conceptuels

Comme précité le modèle de tarification que nous allons adopter pour modéliser la prime pure, et celui utilisant l'approche *fréquence × coût moyen*. Suivant cette approche nous aurons deux modèles conceptuels à tester, celui des Fréquences et celui des Coûts moyens.

Notre premier modèle conceptuel est celui des coûts moyens que l'on présente comme suit :

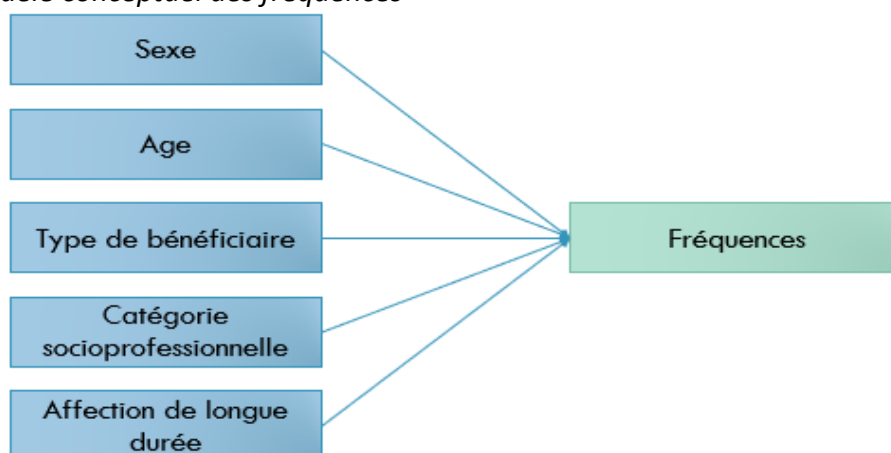
FIGURE 1 *Modèle conceptuel des coûts moyens*



Source : conception de l'auteur

Le deuxième modèle conceptuel que nous envisageons de tester et celui des fréquences, il est donné par la figure suivante :

Figure 2 *Modèle conceptuel des fréquences*



Source : conception de l'auteur

Ces deux modèles conceptuels seront testés dans les sections qui suivent.

2.2. Portefeuille et traitement des données

Les données exploitées sont issues d'une mutuelle d'assurance santé. La base de données contient en outre des informations relatives à 98 000 prises en charges et prestations de soins d'assurance santé observées sur l'exercice (2019). Les variables y présentes ne peuvent être utilisées directement par un modèle statistique, le passage par un traitement préalable s'impose. Pour aboutir à une bonne analyse du portefeuille exploité, il a fallu passer par des vérifications d'éventuelles incohérences qui peuvent nuire à la qualité des résultats. Les valeurs non-cohérentes ont été détectées à l'aide d'un filtre. Dans notre exemple, l'âge a été calculé grâce à la date de naissance des adhérents cela a permis de détecter les valeurs non-cohérentes. Une fois détectées, ces valeurs ont été soit remplacées par la valeur la plus susceptible (on se basant sur les résultats d'une régression linéaire sur la variable à changer), soit elles ont été définies en tant que valeurs manquantes.

La variable quantitative « age_adh » a fait l'objet d'une segmentation en 8 tranches d'âge, ce choix de segmentation est essentiellement basé sur l'avis d'un expert, est justifié en réalité par le fait que l'âge est souvent fonction décroissante de l'état de santé des individus.

2.3. Variables tarifaires

Dans le but de déterminer les descripteurs potentiels de la sinistralité des adhérents nous avons appliqué le processus de sélection *stepwise* aux données dont nous disposons, le portefeuille des adhérents contient 96 935 observations, la sélection *stepwise* nous a conduits à retenir pour la tarification *a priori* les variables tarifaires suivantes : Sexe du bénéficiaire, Tranche d'âge et Risque ALD.

TABLEAU 1 Modalités des variables tarifaires

Variable	Modalité
Sexe	1 : femme
	2 : homme
Tranche d'âge	1 : [00,10 [2 : [10,20[
	3 : [20,30[4 : [30,40[
	5 : [40,50[6 : [50,60[
	7 : [60,70[8 : [plus de 70 ans.
ALD	1 : ALD non
	2 : ALD oui

Source : Conçu par l'auteur

3. Ajustement du modèle du coût moyen

Pour pouvoir élaborer un modèle statistique se basant sur le GLM, nous devons tout d'abord ajuster la variable endogène compte tenu des variables explicatives à une loi de probabilité faisant partie de la famille exponentielle.

Les données relatives aux coûts des sinistres dans l'assurance maladie suivent généralement une distribution non négative et à asymétrie gauche. Cela découle du fait que les coûts ont une borne inférieure (coût > 0), mais qu'ils varient largement dans la partie inférieure. Par conséquent, la distribution gamma, la distribution de poisson et la distribution log-normale donnent souvent de bons résultats lors de la modélisation des coûts moyens des sinistres¹. Dans ce qui suit, ces trois lois de probabilité seront testées sur nos données.

3.1. Modélisation par la famille de loi gamma

Nous commençons par la loi Gamma, bien que le lien canonique de celle-ci soit la fonction inverse, il est plus fréquent d'utiliser un lien logarithmique. En effet la forme multiplicative donne des interprétations simples². La procédure « glm » sous R renvoie les estimations des paramètres du modèle ainsi que la représentativité de chacune.

TABLEAU 2 Estimation des paramètres du modèle gamma

Variables	Significativité	Seuil de significativité
Constante	Oui	0.001
Tage_1	Oui	0.001
Tage_2	Oui	0.001
Tage_3	Oui	0.01
Tage_4	Non	1
Tage_5	Non	1
Tage_6	Non	1
Tage_7	Oui	0.05
Benef sexe_ Féminin	Non	1
Risque Ald_N	Oui	0.001

Source : Conçu par l'auteur

Les résultats de la colonne (test de student) doivent être les plus proches de 0 pour que la variable soit suffisamment explicative car plus le résultat est proche de 0 plus les variables sont significatives.

Nous constatons que les modalités 5, 6 et 7 de la variable tranche d'âge, ainsi que la modalité de la variable Sexe du bénéficiaire ne sont pas significatives. En pratique, pour obtenir le modèle le plus parcimonieux, nous devons regrouper, pour chaque variable, les deux modalités les moins significatives, puis reconduire l'estimation et répéter cette procédure, jusqu'à ne plus avoir de modalité non significative. Mais dans notre cas on ne peut pas regrouper la modalité Sexe du bénéficiaire Féminin, c'est pourquoi on va tester le modèle GLM de la famille log normal pour voir s'il va expliquer le coût moyen par acte par la variable sexe du bénéficiaire.

¹ Dans la pratique, la distribution la plus utilisée pour modéliser les coûts moyens des sinistres est la distribution gamma.

² Charpentier et Dutang R pour l'actuariat 2012.

3.2. Modélisation par la famille de log normal

Pour l'implémentation du modèle de Log-normale, nous avons réalisé une régression linéaire (Gaussienne) sur le logarithme du Coût moyen par acte. L'équation symbolique de ce modèle est la suivante :

$$\text{Log (CM/acte)} = \beta_0 + \beta_1 \times \text{Tag} + \beta_2 \text{ risque ALD} + \beta_3 \times \text{benef sexe} + \varepsilon$$

La sortie relative à ce modèle est fournie par la figure suivante :

TABLEAU 3 Estimation des paramètres du modèle log normal

Variables	Significativité	Seuil de significativité
Constante	Oui	0.001
Tage_2	Oui	0.01
Tage_3	Oui	0.001
Tage_4	Oui	0.001
Tage_5	Oui	0.001
Tage_6	Oui	0.001
Tage_7	Oui	0.001
Tage_8	Oui	0.001
Benef sexe_ Féminin	Oui	0.001
Risque Ald_N	Oui	0.001

Source : Conçu par l'auteur

Les *p-values* des modalités sont toutes inférieures à 0,05 donc elles sont significatives. Ce qui conduit à une bonne explication de la variable dépendante par les variables explicatives.

3.3. Modélisation par la famille de loi poisson

Le modèle de poisson a bien réussi à expliquer la variable « Coût moyen par acte » par les variables catégorielles à savoir le sexe du bénéficiaire, la tranche d'âge et le risque ALD. Nous remarquons que toutes les *p-values* sont inférieurs à 5% ce qui signifie que les coefficients des variables explicatives sont bien significatifs.

TABLEAU 4 Estimation des paramètres du modèle de poisson

Variables	Significativité	Seuil de significativité
Constante	Oui	0.001
Tage_2	Oui	0.001
Tage_3	Oui	0.001
Tage_4	Oui	0.001
Tage_5	Oui	0.001
Tage_6	Oui	0.001
Tage_7	Oui	0.001
Tage_8	Oui	0.001
Benef sexe_ Féminin	Oui	0.001
Risque Ald_N	Oui	0.001

Source : Conçu par l'auteur

Pour vérifier l'hypothèse selon laquelle les trois variables explicatives sont réellement discriminantes, nous allons faire appel aux techniques de sélection des variables. Ces dernières vont être détaillées dans le paragraphe qui suit.

3.4. Sélection des variables explicatives

Plus un modèle contient de variables explicatives, plus il est précis mais moins il est robuste. À l'inverse, moins un modèle a de variables explicatives, plus il est robuste mais moins il est précis. Ainsi on voudrait vérifier si toutes les variables retenues dans le modèle lui sont indispensables ou l'on peut alléger ce dernier en supprimant éventuellement une variable ou plus.

3.4.1. Méthode de type stepwise

La méthode de régression de stepwise dite pas à pas permet soit de rajouter ou de supprimer une variable à chaque stade. Le processus est arrêté quand l'incorporation ou l'élimination d'une variable n'améliore pas le modèle. L'atout de cette démarche est qu'une variable puisse être supprimée puisqu'elle est devenue moins significative suite à l'introduction de variables supplémentaires.

Pour les deux modèles Poisson et Log normal, nous avons lancé la procédure de sélection de variables en utilisant la méthode de stepwise. Le résultat obtenu est que toutes les variables ont été sélectionnées comme étant pertinentes dans le modèle.

3.4.2. Analyse de la variance (ANOVA)

Pour confirmer la pertinence du modèle complet c'est-à-dire celui conservant toutes nos variables explicatives, nous avons réalisé le test d'analyse de la variance. Ce dernier permet de tester l'effet facteur et donc de déterminer si les variables explicatives choisies, sont significatives pour le modèle.

Les résultats de ce test viennent confirmer la conclusion issue du test de Stepwise, autrement dit, nos trois variables à savoir : le risque ALD, le type de bénéficiaire, et la tranche d'âge, sont toutes significatives (toutes les p-values sont petites) et donc elles pourront être gardées dans le modèle, et par conséquent nous allons retenir le modèle de départ Poisson.

4. Ajustement du modèle des fréquences et performance finale

4.1. Ajustement du modèle des fréquences

On va retenir les mêmes variables utilisées précédemment au niveau du modèle des coûts moyens pour modéliser la variable fréquence par bénéficiaire. Nous reproduisons une démarche similaire pour la modélisation de la variable de fréquence des sinistres. Pour pouvoir élaborer un modèle statistique se basant sur le modèle GLM, nous devons tout d'abord ajuster la variable

endogène à une loi de probabilité faisant partie de la famille exponentielle. Les trois modèles classiques permettant de modéliser la charge des sinistres en assurance maladie sont : le modèle de Gamma, de poisson, et de Log-normal. Dans ce qui suit, ces trois lois de probabilité seront testées sur nos données.

Rappelons que la fréquence est le rapport entre le nombre total de sinistres et l'exposition totale t_i .

$$F_i = \frac{N_i}{t_i}$$

Où : N_i représente le nombre de sinistres déclarés par un assuré i . Il est possible de modéliser le nombre de sinistres en supposant que l'exposition de chaque assuré est la même ($t_i = 1$), ce qui correspond à la période d'assurance complète d'un an. Cependant, cela n'est guère réaliste dans la pratique ; dans le cas de notre portefeuille, seulement un total de 136 adhérents ayant une durée de couverture différente de 12 mois, comptes tenus de leurs poids insignifiants (0.001%), nous avons éliminé ces assurés de notre analyse. Comme le nombre espéré de sinistres serait proportionnel à la durée de la couverture, nous ne pouvons pas traiter les sinistres des assurés dont la durée de couverture est inférieure à 12 mois, de la même manière que ceux avec une exposition totale ($t_i = 1$) dans le processus de modélisation. Il est donc crucial de modéliser la fréquence des sinistres, qui tient compte de l'exposition totale (t_i).

Lorsque l'on examine les distributions de probabilité pour le modèle des fréquences, plusieurs choix sont possibles. La distribution la plus couramment utilisée pour les variables de réponse qui sont des comptages est la distribution de Poisson. Il est également possible d'utiliser les modèles Gamma ou de Log-normal.

4.2. Sélection des variables explicatives

Pour sélectionner ou comparer la qualité d'ajustement de plusieurs modèles économétriques¹, diverses mesures sont disponibles. Le critère le plus communément utilisé est celui d'Akaike également appelé critère d'information d'Akaike « AIC » (Akaike, 1974,); ce critère donne une estimation de la qualité de l'ajustement d'un modèle.

D'après les résultats de cet indice, la loi Log normale permet une meilleure modélisation de la fréquence par bénéficiaire que la loi de Poisson. Donc, c'est cette loi que nous avons choisie au final pour la modélisation des fréquences.

¹ Il convient de noter que le critère AIC n'est pas un critère permettant de juger de la qualité de l'ajustement d'un modèle, mais plutôt il sert de comparaison de modèles : le modèle qui minimise l'AIC est considéré comme étant le meilleur.

4.3. Performance finale du modèle

Examinons tout d'abord les erreurs résiduelles à l'aide de la mesure de la racine de la moyenne des carrés des résidus, ou RMSE (Root Mean Square Error) :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

Un outil statistique crucial est le RMSE. Cette statistique donne des informations par rapport à la dispersion et la qualité de la prédiction. L'erreur résiduelle peut être reliée à la variance d'un modèle. Généralement, le RMSE est difficile à interpréter parce que l'on ne sait pas si la variance est faible ou forte. Pour remédier à cet effet et donner du sens à l'indicateur, il est avantageux de normaliser les valeurs du RMSE pour que cet indice soit exprimé comme un pourcentage de la moyenne des observations.

TABLEAU 5 Performance des modèles de tarification

MODÈLES DE SÉVÉRITÉS		MODÈLES DE FRÉQUENCES	
RMSE GLM	MOYENNE	RMSE GLM	MOYENNE
3 890,53	121 476,345	1,1158	22,1957

Le RMSE issu du modèle GLM du modèle des sévérités est de l'ordre de 3 890,53 ce qui est relativement faible car la moyenne des observations est de 121 476,345. Le même constat est valide pour le modèle des fréquences. En effet, dans le premier cas, la variance du modèle correspond à seulement 3.2% de la moyenne des observations alors que dans le second cas, la variance atteint moins de 5% de la moyenne des observations.

Nous constatons ainsi que les deux modèles de GLM performant clairement bien. Rappelons néanmoins que ces outputs sont amplement conditionnés par les données utilisées, ainsi que la configuration apportée des paramètres. Pour les deux modèles de sévérités et de fréquences, l'algorithme de GLM semble très performant dans la mesure où il permet de réduire la variance pour les deux modèles. Ces résultats tendent donc à favoriser les modèles linéaires généralisés. Cependant, insistons sur le fait que l'ordre de grandeur du RMSE n'est pas réellement interprétable opérationnellement dans le cadre concret de la tarification maladie, ce qui limite davantage les conclusions formulées.

5. Résultats de l'étude

5.1. Utilisation des résultats

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

Les coefficients du modèle étant estimés, ils traduisent soit une majoration par rapport à la consommation de l'agent de référence s'ils sont positifs, soit une réduction s'ils sont négatifs. Donc, si une modalité a un paramètre supérieur à zéro, cela signifie que cette dernière contribue à l'aggravation du risque par rapport au cas de référence où elle prend pour chaque variable une modalité à paramètre nul.

Les deux modèles GLM, avec comme fonction de liaison un lien logarithmique ont la structure et les coefficients ajustés suivants :

TABLEAU 6 Coefficients des modèles de fréquence et sévérité

Modèle du coût moyen			Modèle des fréquences	
Variable	Modalité	Estimation des paramètres	Modalité	Estimation des paramètres
Constante	-	5.6279600	-	1,149919
	0	-0.3079047	0	-0,020949
Sexe	1	0.0000	1	0.000000
	0	0.0000	1	-0,031319
Tranche d'âge	2	0.1621994	2	-0,023256
	3	0.1442904	3	0.036026
	4	-0.0778397	4 et 5	0.08876
	5	0.2156819		
	6	0.1192249	6, 7 et 8	0.000000
	7	-0.0157275		
	8	0.2404354		
RISQUE ALD	1	0.0306972	1	-1,080260
	0	0.0000	0	0.00000

Source : Conçu par l'auteur

5.2. Prédiction du coût moyen

En prenant la fonction de lien inverse, à savoir l'exponentielle, des deux côtés, on obtient une fonction de prédiction GLM multiplicative avec les interprétations globales suivantes :

- la constante modélise le coût moyen des sinistres d'un adhérent "moyen", en effet les lignes comportent uniquement des 0, correspondent aux modalités de l'individu de référence pour lequel est calculé la prime de base :
 1. Sexe de l'assuré : Sexe_masculin ;
 2. Classe d'âge : tranche d'âge =1 ;
 3. Risque_ d'affection de longue durée : risque_ALDoui ;

Nous trouvons ainsi la prime de base suivante : $\exp(5.62796) = 278.09$ dhs.

- Le coût moyen de sinistres prédit pour un bénéficiaire enfant de sexe masculin, atteint d'une ALD est égale à 277,82dh ou $\exp(5.627)$.
- Les prévisions sont 73,56 % moins élevées pour les femmes que pour les hommes, puisque $\exp(-0.307) = 0,73$. Une femme consomme moins qu'un homme.
- Comme $\exp(0,030) = 1,030$, les prévisions sont supérieures de 3 % pour les adhérents non atteints d'une ALD.
- Pour les Adhérents des tranches d'âge 2, 3, 5, 6 et 8, les prédictions s'élèvent respectivement à 17%, 15%, 24% ,12% et 72% de celles de la première tranche d'âge. On constate une nette tendance à l'augmentation du coût moyen du sinistre avec l'âge, sauf pour la quatrième et la septième tranche d'âge où les prévisions sont inférieures respectivement de 92% et 98%.

5.3. Prédiction de la fréquence

Tout d'abord les modalités : « Sexe_masculin, risque_ALDoui, tranche d'âge =6,7 et 8 » sont les modalités de référence ;

Nous trouvons la fréquence moyenne : $\exp(1,149919) = 3.16$

- La fréquence de sinistres prédite pour un bénéficiaire de sexe masculin dont l'âge est supérieur à 60ans, atteint d'une ALD est égale à 3,15 ou $\exp(1,149)$.
- Les prévisions sont 98 % inférieurs pour les femmes que pour les hommes, parce que $\exp(-0,020) = 0,98$, et donc ce sont les hommes qui consomment fréquemment les PEC.
- Comme $\exp(-1,080) = 0,34$ les prévisions diminuent de 34 % pour les adhérents non atteints d'une ALD. Est donc ce sont les bénéficiaires atteints d'une pathologie chronique qui consomment fréquemment les PEC.
- Pour les adhérents faisant partie des tranches d'âge 3 et (4et5) les prédictions s'élèvent respectivement à 3% et 8% de celles de la catégorie de référence. Tandis que pour la première et la deuxième tranche d'âge ; les prévisions baissent de 96% et 97% respectivement. On constate donc une nette tendance à l'accroissement de la fréquence de sinistre avec l'âge.

5.4. Application du GLM au portefeuille

Une fois que nous avons obtenu les coefficients à appliquer pour les deux modèles de fréquence et du coût moyen, avec les lois et les variables sélectionnées, nous obtenons la prime annuelle d'un adhérent donné i après inversion par la fonction exponentielle, en utilisant les deux fonctions GLM :

La fonction de lien logarithmique étant choisie pour nos deux modèles nous obtenons alors :

Pour le modèle de fréquence

$$\text{Log}(Freq/benf) = (\beta_0 + \beta_1 \times Tag + \beta_2 \text{ risque ALD} + \beta_3 \times benef \text{ sexe} + \varepsilon)$$

$$Freq/benf = \exp(\beta_0 + \beta_1 \times Tag + \beta_2 \text{ risque ALD} + \beta_3 \times benef \text{ sexe} + \varepsilon)$$

Pour le modèle du coût moyen nous obtenons :

$$\text{Log}(CM/acte) = \beta_0 + \beta_1 \times Tag + \beta_2 \text{ risque ALD} + \beta_3 \times benef \text{ sexe} + \varepsilon$$

$$CM/acte = \exp(\beta_0 + \beta_1 \times Tag + \beta_2 \text{ risque ALD} + \beta_3 \times benef \text{ sexe} + \varepsilon)$$

En effet, pour un individu donné, la prime pure est donnée par le produit des deux grandeurs : la fréquence et le coût moyen.

$$\text{Prime pure} = Freq/benf \times CM/acte$$

Prime pure=

$$\exp(\beta_0 + \beta_1 \times Tag + \beta_2 \text{ risque ALD} + \beta_3 \times benef \text{ sexe} + \varepsilon) \times \exp(\beta_0 + \beta_1 \times Tag + \beta_2 \text{ risque ALD} + \beta_3 \times benef \text{ sexe} + \varepsilon)$$

Conclusion et perspectives :

Le présent article propose une approche de tarification intégrant l'état de santé des adhérents comme facteur déterminant de la cotisation. Une analyse préliminaire de l'état de santé des adhérents du portefeuille justifie l'utilisation d'une telle information. Pour ce faire, nous avons introduit la variable ALD indiquant l'atteinte ou non de l'adhérent d'une affection de longue durée. La méthodologie adoptée pour élaborer le tarif de la mutuelle santé se base sur le modèle de la prime pure, qui est le résultat du produit des deux modèles, de fréquence et de sévérité.

Ces deux modélisations font appel aux techniques des modèles linéaires généralisés GLM. Toutefois, il convient de signaler que la mise en place de l'approche GLM nécessite un processus de modélisation approprié. En effet, nous avons commencé par identifier les lois théoriques qui ajustent correctement les observations des fréquences et des coûts moyens des sinistres. Ensuite, nous avons estimé séparément le coût moyen et la fréquence des sinistres afin d'isoler les effets des facteurs sur ces derniers. Cette méthodologie d'estimation a conduit à obtenir une meilleure compréhension de l'influence des facteurs sur le risque. Une analyse approfondie des résultats de la modélisation a permis de montrer l'impact de l'état de santé des adhérents sur la structure tarifaire du portefeuille étudié.

Le modèle a permis d'obtenir une cotisation qui repose sur la sinistralité des adhérents. Autrement dit, le tarif obtenu a pris en considération les caractéristiques individuelles des adhérents.

Le présent article nous a permis de valider les hypothèses associées au sexe de l'adhérent, à son âge, et à la présence d'une ALD et de rejeter celles liées au type du bénéficiaire, et à sa catégorie socioprofessionnelle. Toutefois, certains points n'ont pas pu être traité en profondeur et certaines pistes d'amélioration demeurent inexplorées : premièrement les sinistres extrêmes peuvent être traités au moyen de techniques appropriées, en procédant à un écrêtement des valeurs extrêmes (cf. voir la théorie des valeurs extrêmes pour plus de détails). Deuxièmement le recours à la théorie de la crédibilité, notamment *via* le système *Bonus-Malus* représenterait un argument actuariel solide afin d'apporter les corrections nécessaires à la prime pure *a priori*. Enfin, le périmètre de ce travail se limite au calcul des cotisations. Les chargements techniques et les coûts de distribution ne font pas objet d'étude. Or, les frais et les commissions de distribution sont des sujets importants afin d'agir sur la stabilité financière.

Notre travail présente ainsi plusieurs apports scientifiques :

- Sur le plan pratique : la proposition d'un modèle de tarification en assurance-maladie en utilisant l'âge, le sexe, et l'atteinte ou non d'une affection de longue durée, comme variables tarifaires, ce qui a permis de démontrer l'effet de ces variables, à la fois sur les fréquences et les coûts moyens, et donc leur impact direct sur la prime pure (le tarif).
- En termes d'implications managériales, les résultats de cette étude suggèrent aux responsables des mutuelles de santé, naviguant désormais dans un environnement concurrentiel marqué par la montée en puissance de la digitalisation et l'émergence de nouvelles sources de données, qu'une structure tarifaire diversifiée, qui soit ajustée au niveau du risque des assurés, est nécessaire pour garantir la solvabilité et la pérennité de leurs structures. En effet, l'adoption d'une pratique actuarielle pour la mutuelle de santé, exige le passage par la prise en compte des facteurs de risque influençant à la fois la fréquence de la consommation des prises en charge et les coûts moyens y afférents. La mise en place de cette pratique conduira à l'instauration d'un système de tarification actuarielle qui incitera les adhérents à prendre soins davantage de leur santé, améliorant par voie de conséquence la solvabilité des mutuelles de santé.

Bibliographie :

- Akaike, H. (1974.). *A new look at the statistical model identification*. *IEEE Transactions on Automatic Control*, 19(6); 716-723.
- Antonio, K., & Valdez, E. A. (2012). *Statistical Concepts of a Priori and a Posteriori Risk Classification in Insurance*. *Asta-Advances in Statistical Analysis*, 96(2), 187-224.
- Bellina, R. (2014). *Méthodes d'apprentissage appliquées à la tarification non-vie*. Université Claude Bernard, Lyon 1, Mémoire.
- Brisard, E. (2014). *Pricing of Car Insurance with Generalized Linear Models*. Université Libre de Bruxelles, Thesis.
- Brockman, M. J., & Wright, T. S. (1992). *Statistical Motor Rating: Making Effective Use of Your Data*. *Journal of the Institute of Actuaries*, 119(03), 457-543.
- Burnham, K. P., & Anderson, D. R. (2004). *Multimodel Inference*. *Sociological Methods & Research*, 33(2), 261-304.
- Charpentier, A., Denuit, M., & Elie, R. (2015). *SEGMENTATION ET MUTUALISATION LES DEUX FACES D'UNE MÊME PIÈCE ? Risques n° 103*, 19-23.
- Denuit, M., & Charpentier, a. (2004). *Mathématiques de l'assurance non-vie*. *Economica*.
- Denuit, M., & Lang, S. (2004). *Non-life Rate-making with Bayesian GAMs*, *Insurance. Mathematics and Economics*, 35(3), 627-647.
- Denuit, M., Marechal, X., Pitrebois, S., & Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons, Ltd.
- Denuit, M., Marechal, X., Pitrebois, S., & Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons, Inc., Hoboken.
- Dionne, G., & Vanasse, C. (1992). *Automobile insurance ratemaking in the presence of asymmetrical information*. *Journal of applied econometrics*.
- Fox, J. (2016). *Applied Regression Analysis & Generalized Linear Models (Third Edition ed.)*. Sage Publications.
- Frees, E., Lee, G., & Yang, L. (2016). *Multivariate Frequency-Severity Regression Models in Insurance*. *Risks*, 4(1).
- Gao, G., Meng, S., & Wuthrich, M. (2018). *Claims Frequency Modeling Using Telematics Car Driving Data*. *Scandinavian Actuarial Journal*.
- Gschlößl, S., & Czado, C. (2007). *Spatial modelling of claim frequency and claim size in non-life insurance*. *Scandinavian Actuarial Journal*, 202-225.
- Jarque, C. M., & Bera, A. K. (1987). *A test for normality of observations and regression residuals*. *International Statistical Review* 55, 163–172.
- Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). *Loss models: from data to decisions*. John Wiley & Sons.
- Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance*. Boston-Dotrecht-London: Kluwer Academic Publishers, ISBN 0-7923-9545-X.

Lorenzoni, L., Marino, A., Morgan, D., & James, C. (2019). *Health Spending Projections to 2030: New results based on a revised OECD methodology*. OECD Health Working Papers, n° 110, OECD Publishing, paris.

Lotsi, A., Mettle, F., & Adjorlolo, P. K. (2019). *Application of Bühlmanns-Straub Credibility Theory in Determining the Effect of Frequency-Severity on Credibility Premium Estimation*. ADRRI Journal of Physical and Natural Sciences, 1-24.

McClenahan, C. (2001). *Ratemaking*. Casualty Actuarial Society, 4th edition.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. CRC Press, 2nd edition.

Navarun, J. (n.d.). *Towards Machine Learning: Alternative Methods for Insurance Pricing – Poisson-Gamma GLM's, Tweedie GLM's and Artificial Neural Networks*.

Nelder, J. A., & Wedderburn, R. W. (1972). *Generalized Linear Models*, Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384.

Ohlsson, E., & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. EAA SERIES. Springer.

Ohlsson, E., & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, Berlin.

Paefgen, J., Staake, T., & Thiesse, F. (2013). *Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach*. Decision Support Systems, 56(1), 192-201.

Pitrebois, S., Denuit, M., & Walhin, J. F. (2003). *Setting a Bonus-Malus Scale in the Presence of Other Rating Factors: Taylor's Work Revisited*. Astin Bulletin 33(02), 419-436.

Sakthivel, K. M., & Rajitha, C. S. (2017). *Artificial intelligence for estimation of future claim frequency in non-life insurance*. Global Journal of Pure and Applied Mathematics, 13, 10.

Wüthrich, M. V. (2016). *Market-Consistent Actuarial Valuation*. Springer.