Université Montpellier

Faculté d'Economie

PROJET D'ECONOMETRIE APPLIQUEE

THEME

TARIFICATION PAR LA MÉTHODE FRÉQUENCE-COUT EN ASSURANCE AUTOMOBILE

Rédigé par :

MUJIDILA MUAMBA Sylvain
FOPOUSSI KAMGA Béatrice
Étudiants M1 ACTUARIAT

Chargés du cours :

Benoit MULKAY
Jules SADEFO KAMDEM
Professeurs à l'UM

Mars-Avril 2022

Table des matières

DÉCHARGE

Les propos émis dans ce document sont propres aux auteurs. La faculté d'économie de l'Université de Montpellier n'entend donner aucune approbation ni improbation.

Introduction

La tarification est une problématique au cœur du métier de l'assurance. Elle consiste pour une société d'assurance de fixer un prix pour un risque assurable donné. Une mauvaise tarification conduirait probablement au problème d'anti-sélection autrement dit sélection de mauvais risques ou conduire à la ruine de la société. Ce phénomène correspond au fait que si un assureur applique une prime uniforme à tous ses assurés, il ne va probablement attirer que les mauvais risques puisque ces derniers bénéficieront d'un tarif plus avantageux au regard de leurs profils que chez des assureurs ayant adapté leur tarif en fonction du profil, et les bons risques ne souscriront pas puisque leur tarif sera respectivement moins avantageux. Partitionner son portefeuille va permettre de créer de nouveaux sous-portefeuilles dits classes tarifaires au sein desquels les risques peuvent être considérés de même loi et indépendants, et l'assureur pourra adapter ainsi sa tarification.

La réalisation de cette tarification en assurance IARD en général et en assurance automobile en particulier s'appuie classiquement sur l'analyse de la prime pure dans le cadre d'un modèle multiplicatif fréquence X coût, dans lequel l'effet des variables tarifaires sur le niveau du risque est modélisé par des modèles linéaires généralisés.

Plusieurs étapes sont nécessaires pour la mise en place de cette méthode de tarification : la constitution d'une base de données, la distinction de différents types de sinistres ; le choix de variables tarifaires puis, la mise en place du modèle devant expliquer l'effet de caractéristiques de chaque individus de la base données sur la variable expliqué (la fréquence et le cout).

Le modèle de cout-fréquence que nous allons appliquer dans le cadre de ce travail de recherche, nécessite de faire deux modélisations. Une première modélisation du cout de sinistre afin d'estimer le cout moyen et une deuxième modélisation pour la fréquence de survenance de sinistres afin d'estimer la fréquence moyenne de ces derniers pour chaque assuré. Comme tous les individus n'ont pas les mêmes caractéristiques, nous les grouperons dans des classes les plus homogènes possible à l'intérieur et les plus hétérogènes possibles entre elles. Cette classification nous la réaliserons grâce à l'algorithme de classification automatique utilisé en machine learning, l'algorithme de K-means. Ainsi les individus regroupés dans une même classe auront de contrats de même valeurs, donc payeront une même prime.

L'objectif de notre étude est de mettre en place un modèle de tarification pour des contrats d'assurance automobile. Pour y arriver cette étude sera subdivisée 5 chapitres. Une présentation et analyse exploratoire de données utilisées, une présentation de différents modèles et méthodes statistiques qu'on aura à utiliser, présentation des résultats et validation des modèles et en dernier une conclusion.

Chapitre 1

Présentation des données et Analyse exploratoire

Ce chapitre consiste en la présentation de la base de données et la description de cette dernière par les statistiques descriptives.

1.1 Présentation de la base

La base de données utilisée pour les analyse faites dans ce document, est celle fournie par l'institut des actuaires aux participants de la compétition PRICING GAME. Cet institut organise cette compétition depuis 2017 en partenariat avec l'université de Rennes sous la coordination du professeur Arthur Charpentier.

Elle est constituée de de deux sous bases de données. Une première de 100000 observations d'assurés en assurance automobile. Une deuxième base de donnée de réclamations pour les assurés ayant connu de sinistre durant la période d'observation. Dans ce deuxième jeu donnée on retrouve le montant de charge engendrée par chaque sinistre subi par l'assuré et nombre de fois que l'assuré a connu de sinistre.

Pour une description complète de la base de données, vous trouverez le fichier correspondant :

- ☞ cliquer ici pour voir la description complète de données.

.

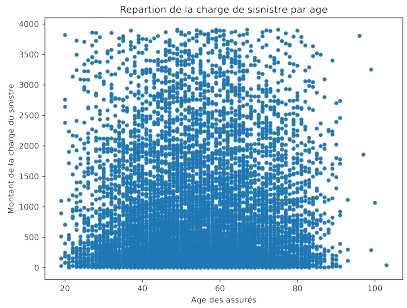
1.2 Analyse descriptive

La base données à notre disposition, comporte 38 variables parmi lesquelles, il y a de variables qualitatives et quantitatives. La synthèse de caractéristique de toutes ses différentes variables est à retrouver en annexe. Dans cette partie nous présenterons en premier, une analyse bivariée de la charge de sinistre avec les différentes variables puis en deuxième celle de la fréquence de nombre de sinistres.

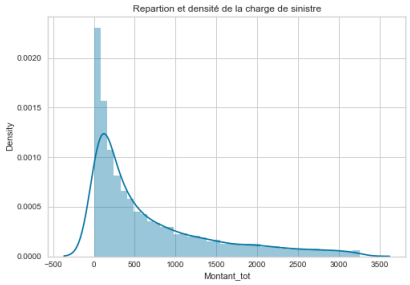
1.2.1 Analyse de la charge de sinistre : Cout

La charge du sinistre représente la somme décaissée par l'assureur pour dédommager l'assuré. Elle est négative lorsque la responsabilité de l'assuré n'est engagée. Dans cette

étude nous ne considérons que les sinistres ayant de cout positifs, c'est à dire les sinistres où la responsabilité de l'assuré est établie.La charge moyenne est de 77,7 euros sur l'ensemble de assurés ayant connu un accident.



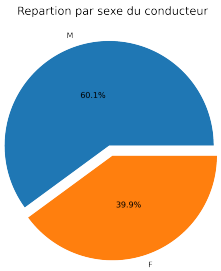
Source :Calculs des auteurs



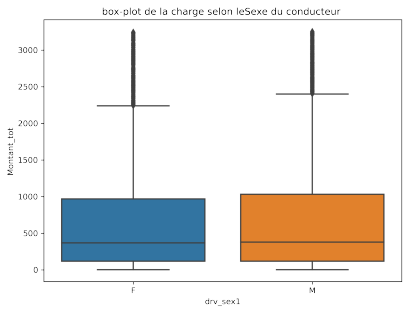
Source :Calculs des auteurs

- Sexe de l'assuré et charge de sinistre : La population de nos données est repartie de 60% des hommes et 39,9 de femmes. La charge de sinistre moyenne est de 747 euros chez les femmes et 774 euros chez les hommes. Les deux valeurs semblent mathématiquement différentes, mais cela n'est pas le cas , statistiquement. Le test de comparaison de moyennes entre deux échantillons nous relève que ces deux valeurs sont statistiquement les même.

  Ainsi charge des indemnisations ne dépend pas du sexe de l'assuré. On peut le remarquer sur le box-plot ci-dessous. Cette indépendance est également confirmée par le test chi2 d'indépendance (p-value=0.48). Au regard de ce qui précède, on peut conclure que risque assuré chez les femmes semble le même que chez les hommes. C'est à dire entre ces deux groupes d'individus les sinistres subi sont de même ampleur.



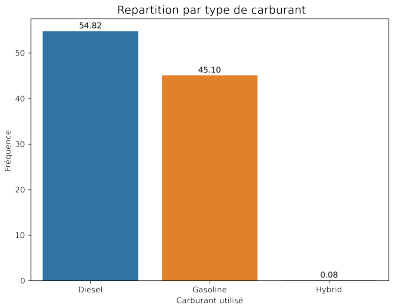
Source :Calculs des auteurs



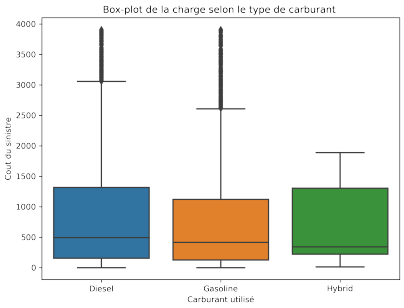
Source :Calculs des auteurs

- Type de carburant et charge de sinistre : La majorité de voitures assurées utilisent comme carburant le Diesel(54,82%) le Gazoline (45,1%) et moins de 1% sont de véhicules hybrides. On a remarqué une dépendance entre ce caractère et la charge de sinistre (P-value =0.01674). Les véhicules utilisant le diesel comme carburant sont ceux ayant en moyenne des cout de sinistre élavés (871 euros) alors que les

voitures hybrides entrainent en moyenne de cout moins faibles (706 euros) contre 787 euros pour gazoline.

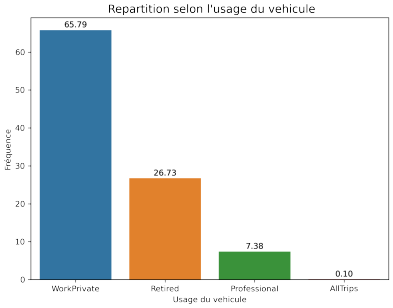


Source :Calculs des auteurs

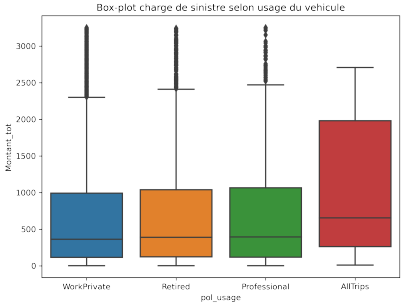


Source :Calculs des auteurs

- Usage du véhicule et charge de sinistre : La police assurance étudiée ici, couvre en majorité les individus qui travaillent (65,79%) mais n'utilisant pas leur véhicules pour l'usage professionnel. Il y a 26,73% des assurés qui sont retraités. Ces derniers utilisent moins leurs véhicules et donc ont moins de chance de connaitre de sinistre. Seulement moins de 8% utilisent leurs véhicules pour un usage professionnel, ces derniers sont les plus exposés car ils utilisent régulièrement leurs véhicules dans le cadre de leur travail. On enregistre en moyenne de cout de sinistre de 922 euros chez les professionnels, 844 euros chez les retraités, 826 euros chez les privés. La catégorie AllTrip enregistre quant à elle une moyenne de charge 1200 euros, ce qui semble beaucoup plus élevé au regard des autres catégories. cette différence est due au manque d'une grande mutualisation dans cette catégorie il y'a moins de 1% d'individus. L'analyse de la variance nous a confirmé le lien de dépendance de la charge de sinistre au regard de ce caractère.



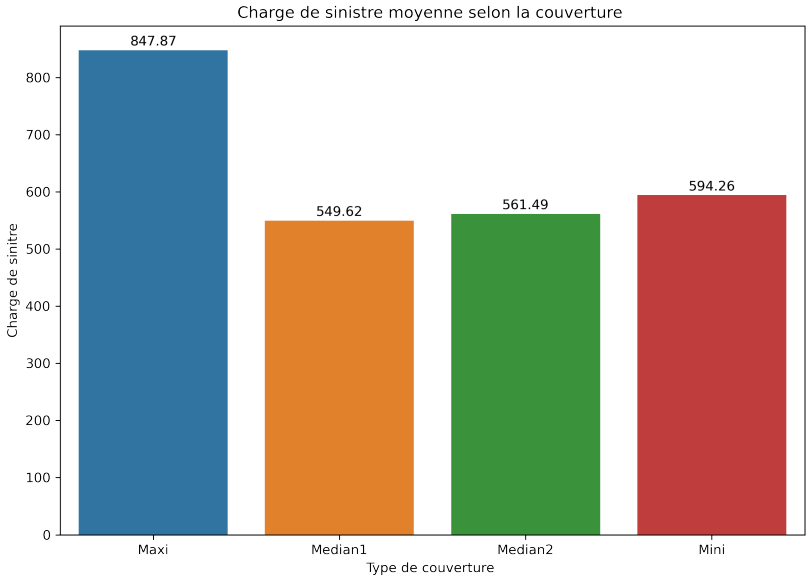
Source :Calculs des auteurs



Source :Calculs des auteurs

- charge de sinistre et type de couverture de la police :

Selon, le type de couverture de la police assurance, les assurés ayant la couverture max ont en moyenne de charge beaucoup plus élevée par rapport à d'autres types de couvertures. Ceci est normale, car le type de couverture représente le pourcentage du montant de sinistre pris en charge par l'assureur. Ce pourcentage est beaucoup plus grand pour la couverture max.



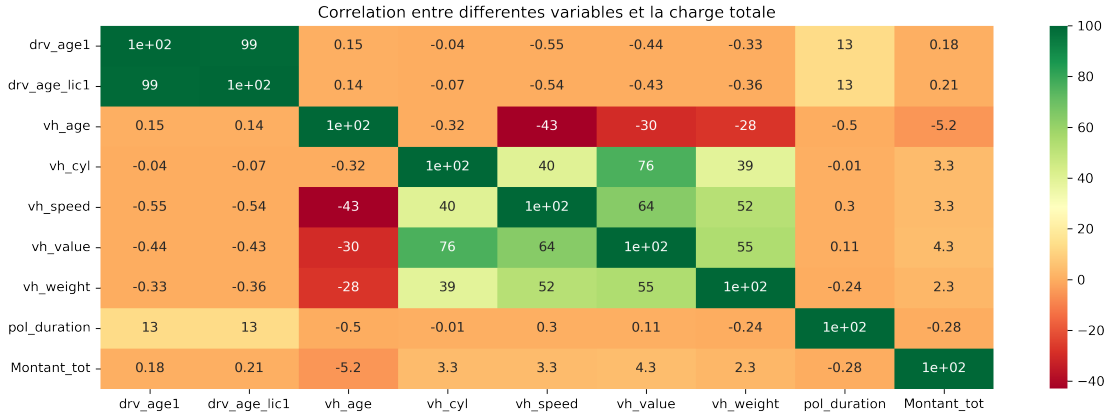
Source :Calculs des auteurs

TABLE 1.1 – ANOVA (charge de sinistre et Type de couverture de la police)

	sum_sq	df	F	PR(>F)
C(pol_coverage)	1.73E+08	3	75.46254	2.88E-48
Residual	7.71E+09	10106	NaN	NaN

- Corrélation entre Variables quantitatives : La liaison entre les différentes variables quantitatives et les montants de charge de sinistres semble très faibles. Ceci ne nous permet pas de sélectionner les variables pertinentes directement pour la modélisation du cout moyen . Nous adapterons les méthodes de sélection progressive pour déterminer le modèle le plus pertinent.

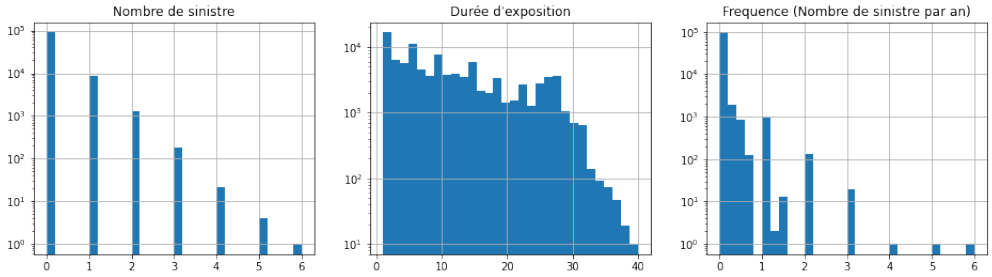
  Aussi, les modèles linéaires généralisés sont de modèles prédictifs, la corrélation linéaire n'est pas très importante, néanmoins elle nous permet de détecter les variables explicative qui sont les plus liées afin de réduire si possible, les nombre de variables dans le modèle selon le principe de parcimonie.



Source :Calculs des auteurs

1.2.2 Analyse de nombre de sinistre : Fréquence

Le nombre de sinistre représente le nombre de fois que l'assuré a réclamé à l'assureur une indemnisation. La fréquence c'est nombre de sinistre sur la durée d'exposition qui est la durée depuis écoulée que l'individu est assuré chez le même assureur.



Calculs des auteurs

Les statistiques descriptives n'étant pas très importantes [1]. Nous avons ainsi décidé de nous en passer de cette partie pour l'étude de nombre de sinistre.

Vous trouverez en annexes un tableaux synthétique, permettant de comprendre les différentes relations de cette variable avec les autres variables.

1. Selon ce que le prof Benoit

Chapitre 2

Modèle et méthodes statistiques utilisés

Ce chapitre consiste en la présentation théorique de modèle économétrique et les méthodes statistiques que nous utiliserons tout au long de cet travail. En premier nous présenterons les modèles linéaires généralisés puis la technique de classification en occurrence le clustering de K mean, qui est une méthode utilisée en apprentissage automatique.

2.1 Les modèles linéaires généralisés

2.1.1 Présentation de GLM

Nous avons vu durant le cours d'économétrie théorique que le modèle linéaire général repose sur une hypothèse forte : le terme d'erreur suit une loi normale et de même variance. Nous avons parfois le besoin d'expliquer des variables qui ne suivent pas ce pré-requis.

Le modèle linéaire généralisé est la technique qui permet d'avoir un large choix pour la distribution de la variable à expliquer lorsque cette dernière ne suit pas forcément une loi normale. Les modèles linéaires généralisés dits en anglais GLM ont pris une grande importance dans le domaine de la tarification IARD. En effet, la plupart des modèles de tarification aujourd'hui sont basés sur une estimation par GLM. Comme on le verra par la suite, ces modèles offrent des estimations faciles à mettre en production notamment par un calcul multiplicatif.

L'objet d'un modèle économétrique GLM est d'expliquer la variable Y par les variables $X_1...X_p$. Dans notre cas la variable Y représente le coût ou la fréquence du sinistre. Les $X_1...X_p$ représentent les variables tarifaires internes explicatives telles que (l'age du conducteur, le sexe du conducteur, l'age du véhicule assuré, la valeur du véhicule....) car nous sommes en assurance automobile.

Techniquement un GLM est défini par trois composantes :La Distribution ; Les prédicteurs ; La fonction de lien.

- **La Distribution** : Dans le cadre des GLM, la distribution à choisir pour la variable à prédire fait partie de la famille des lois exponentielles. On suppose que l'échantillon de $Y_i, i = 1...n$ indépendantes et tirées de loi dont la distribution suit la forme suivante :

$$f(y_i, \theta_i, \phi) = exp\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + C(y_i, \phi)$$

  où la valeur θ est le paramètre naturel de la famille exponentielle.

  Les fonctions a, b et c sont spécifiées en fonction du type de loi exponentielle.

On peut citer des lois usuelles qui sont majoritairement utilisées et qui font partie de la famille des lois exponentielles (ex : Normale, Poisson , Gamma , Inverse gaussienne, Binomiale, Binomiale Négative).

- **Les prédicateurs** : Les observations liées aux $Y_i, i = 1...n$ sont les composantes déterministes du modèle et représentées par les $(X_{i,j})i = 1, ..., n; j = 1, ..., p$

  Cela se traduit dans notre cas par les variables tarifaires comme celles citées ci-haut.

  On définit le prédicteur linéaire par : $\eta = X\beta$ avec β les paramètres à estimer du GLM.

- **La fonction de lien** : La fonction de lien est celle qui lie l'espérance μ de Y au prédicteur linéaire. Si on note g, cette fonction de lien , alors :

$$g(E(Y_i)) = g(\mu_i) = \eta$$

  Selon la distribution de la loi de Y, la fonction de lien diffère. Pour une distribution normale la fonction de lien est l'identité, une distribution de Poisson elle est le logarithme de l'espérance et pour la loi binomiale elle est le logit de la probabilité de succès.

  La fonction lien qui associe la moyenne de Y au paramètre naturel θ est appelée fonction lien canonique. Nous avons dans ce cas :

$$g(E(\mu_i)) = \theta_i = \eta_i$$

2.1.2 Estimation du paramètre β par GLM

Pour la plupart des modèles linéaires généralisés, les équations qui déterminent les paramètres au sens du maximum de vraisemblance sont non linéaires et les estimateurs n'ont pas d'autres expressions formulables comme solutions de ces équations. Les logiciels statistiques et économétriques calculent les estimations en utilisant un algorithme itératif pour la résolution d'équations non linéaires.

Un algorithme populaire pour atteindre cet objectif est appelé Fisher scoring et a été proposé initialement pour ajuster des modèles probit. Pour la régression logistique binomiale et pour les modèles log-linéaires de Poisson cet algorithme se simplifie et n'est alors qu'une version du très connu algorithme de Newton-Raphson.

Soit $L_i(\theta_i, \Phi, y_i) = Ln(f(y_i, \theta_i, \phi))$ la log-vraisemblance de la i^{me} observation. on a alors :

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \qquad (1)$$
$$\frac{\partial^2 l}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)} \qquad (2)$$

Pour une loi appartenant à la famille des exponentielles on a le résultat suivant :

$$E(\frac{\partial l}{\partial \theta_i}) = 0 \quad et \quad E(\frac{\partial^2 l}{\partial \theta_i^2}) = 0$$

Alors en utilisant les deux équations ci-haut, nous obtenons :

$$E(y_i) = b'(\theta_i) \quad et \quad Var(y_i) = -b''(\theta_i)a(\phi)$$

La log-vraisemblance de l'échantillon est donnée par :

$$L(\beta) = \sum_{i=1}^{n} Ln(f(y_i, \theta_i, \phi)$$

La maximisation de la log-vraisemblance passe par la résolution de l'équation :

$$\frac{\partial L(\beta)}{\partial \beta_j} = 0 \qquad \forall j = 1...p$$

On peut observer alors pour chacun des L_i la dérivée par rapport à β_j . Pour faciliter les calculs de la dérivé de la log vraisemblance de l'échantillon, on peut écrire cette dernière comme produit de dérivé qu'on peut calculer plus facilement :

$$\frac{\partial L(\beta)}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta} \frac{\partial_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\eta_i}{\beta_i}$$

$\frac{\partial \ell_i}{\partial \theta_i} = [y_i - v'(\theta_i)] / u(\phi) = (y_i - \mu_i) / u(\phi)$
$\frac{\partial \mu_i}{\partial \theta_i} = v''(\theta_i) = \text{Var}(Y_i) / u(\phi)$
$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \quad \text{car} \quad \eta_i = \mathbf{x}_i' \beta$
$\frac{\partial \mu_i}{\partial \eta_i}$ dépend de la fonction lien $\eta_i = g(\mu_i)$

Les équations de la vraisemblance sont :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad j = 1, \ldots, p$$

Ce sont des équations non-linéaires en $\boldsymbol{\beta}$ comme nous l'avons signalé ci-haut et donc leur résolution requiert la méthode de Newton-Raphson.

2.1.3 Ajustement du modèle et qualité du modèle

La déviance est un critère utilisé pour mesurer la qualité d'ajustement des modèles aux données. Cette mesure représente la différence entre la log-vraisemblance du modèle et celle du modèle saturé. Le modèle saturé est défini comme étant le modèle qui estime exactement les données.

$$D = -2(L - L_s tat)$$

qui est le logarithme du carré du rapport des vraisemblances. Il généralise l'usage des sommes de carrés propres au cas Normal et donc à l'estimation par moindres carrés. Asymptotiquement , D suit une loi de $\chi^2 \quad n-p$ degré de liberté. ce qui va nous permettre de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

on peut aussi se baser sur le résidus , du modèle pour valider ce dernier. Dans le cas de modèle GLM, on a deux types de residus :

- les résidus de Pearson
- les résidus de déviance

On peut noter que la somme des carrés des résidus est dans les deux cas, asymptotiquement, un $\chi^2 n - p - 1$ degrés de liberté.

Pour discriminer deux ou plusieurs modèles candidats (loi de la variable explicative), nous effectuerons des tests d'adéquation tel que le test chi2, le test Kolmogorov-smirnov, le test Ks, le test de Man Wittney... Nous prendrons comme loi Y, celle qui s'adapte au mieux au données puis nous estimerons les coefficient de ce dernier.

2.2 Présentation de la méthode fréquence-cout

Pour chaque contrat ou police, la prime est fonction de variables dites de tarification. Généralement, on considère des informations sur l'assuré, comme l'âge ou le sexe pour un particulier, ou le secteur d'activité et le nombre de salariés pour une entreprise, des informations sur le bien assuré, comme l'âge du véhicule, la puissance ou la marque en assurance auto, la surface du logement en multirisque habitation, le chiffre d'affaire de l'entreprise en perte d'exploitation,des informations géographiques comme le revenu moyen dans la commune ou le département, la densité de population, etc.

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

$$S = Y_1 + Y_2 + ..Y_N = \sum_{i=1}^{N} Y_i$$

La prime pure est E(S) = E(N)* E(Yi) dès lors que les coûts individuels sont i.i.d., indépendants du nombre de sinistres. Dans le cas où la fréquence et les charges sont hétérogènes, l'hétérogénéité étant caractérisée par une information , la prime pure devrait être :

$$E(S|\Omega) = E(N|\Omega) * E(Yi|\Omega)$$

. L'espérance de cout de chaque sinistre est alors estimée soit par un modèle gamma ou log-normal et l'espérance de nombre de sinistre par un modèle de poisson ou binomiale négative.

2.3 Classification par K-Mean

Le K-means (k-moyennes) est un algorithme non supervisé de clustering ou regroupement. Ainsi, on n'essaie pas d'apprendre une relation de corrélation entre un ensemble de variables explicatives dites features X d'une observation et une valeur à prédire Y d'une variable expliquée, comme c'est le cas pour l'apprentissage supervisé. L'apprentissage non supervisé va plutôt trouver des patterns dans les données. Notamment, en regroupant les choses qui se ressemblent.

Pour faire ce regroupement, l'algorithme K-Means, se base sur un critère dit critère de similarité ou dissimilarité. ainsi donc, deux individus qui se ressemblent auront une distance de similarité réduite et une distance de disimilarité grande.

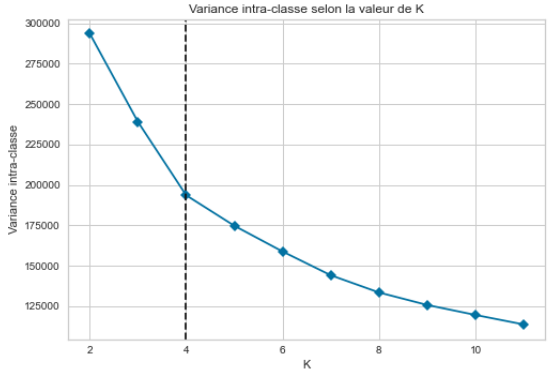
Le critère les plus utilisé en littérature statistique est la distance Euclidienne ; elle est

calculée pour deux observation x_1 et x_2 dans un espace vectoriel V^n par :

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n} (x_{1j} - x_{2j})^2}$$

D'autre type de distance sont aussi telles que distance de Manhattan, distance de mahalanobis...

Quel nombre de classe faut-il retenir ? pour répondre à cette question , il faut effectuer différent modèles avec de valeur de K différentes et choisir la valeur de K qui minimise la variance intra-classe de façon significative.



Calculs des auteurs

En regardant l'évolution du taux d'erreur , c'est à dire la variance inter-classe on remarque une formation du coude à partir de K=4, donc on prendra comme nombre de classe classe 4. On formera ainsi 4 groupes dits des classes tarifaires dans lesquels , le risque sera le plus homogène et donc une même prime pure.

TABLE 2.1 – Moyennes selon les classes

Classe	drv_age1	vh_value	vh_age	vitesse	Sinistre	Montant
0	54.55	11678.58	20.58	142.93	0.06	29.30
1	42.08	16169.57	7.57	169.61	0.12	58.59
2	54.24	31005.62	6.45	200.99	0.13	69.32
3	68.69	16238.02	7.61	169.67	0.11	59.15

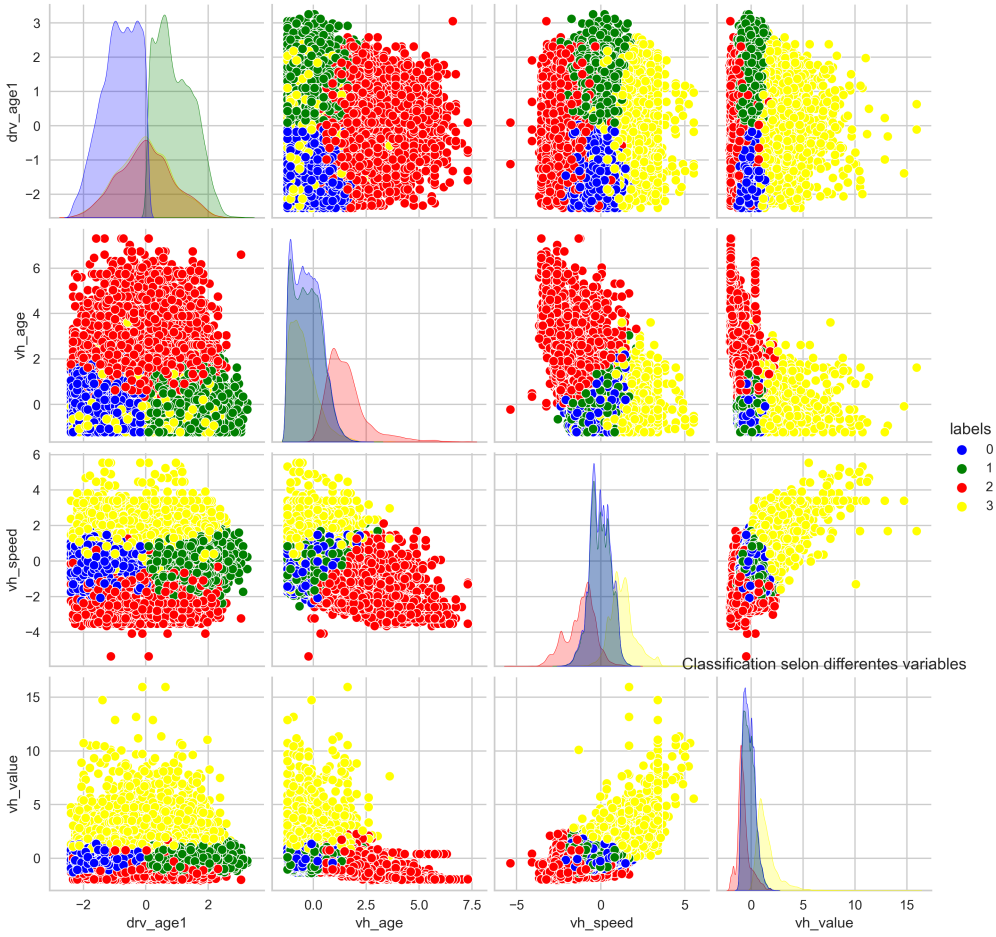
En observant la classification des individus sur le graphique ci-dessous et en se référant au tableau ci-haut, il en ressort les conclusions suivantes :

- Les assurés classés dans les groupe 0, sont ceux ayant les véhicules anciens et qui coutent en moyenne moins chers. Ils ont en moyenne peut de sinistre et ces derniers ont engendré de charges moyennes faibles.
- Les assurés classés dans les groupe 1 sont les moins agés avec de véhicules relativement moins usagés et de vitesse moyenne.
- Les assurés classés dans le groupe 2, sont ceux ayant les véhicules en moyenne moins usagés et qui coutent en moyenne plus chers. Leurs véhicules sont aussi les plus

rapides. Ces assurés, sont ceux ayant en moyenne plus de sinistres et ces derniers engendre de charges en moyennes plus lourdes pour l'assureur.

- Les assurés classés dans le groupe 3, sont les plus âgés.

Illustration 2.1 – *Classification selon differentes variables*



Calculs des auteurs

Chapitre 3

Modélisation et Résultats

3.1 Modélisation de la fréquence de sinistre

Comme nous l'avons énoncé précédemment, la variable nombre de sinistre est supposée suivre une loi Binomiale, loi Binomiale négative ou une loi de poisson.

3.1.1 Adéquation à une loi Binomiale

Considérons une expérience aléatoire Y à deux états possibles : Y=1 (succès) avec Probabilité p et Y=0 (échec) avec probabilité 1-p.

Considérons maintenant X une V.A qui consiste a compter le nombre de succès obtenu obtenu au bout de n expériences indépendantes. X suit alors la loi binomiale de paramètres (n,p). On la note $X \sim B(n,p)$.

$$P(X=k) = C_n^k p^k (1-p)^k$$

La vraisemblance s'écrit :

$$L(n,p) = \prod_{i=1}^{N} C_n^{k_i} p^{k_i} (1-p)^{n-k_i}$$

En maximisant la log-vraisemblance on trouve le paramètre p de cette loi.

$$l(n,p) = \ln(L(n,p)) = \sum_{i=1}^{N} \left[\ln\left(C_n^{k_i}\right) + k_i \ln(p) + (n-k_i)\ln(1-p) \right]$$

On derive ainsi la log vraisemblance par rapport au parametres p puis on égalise sa derivé à 0.

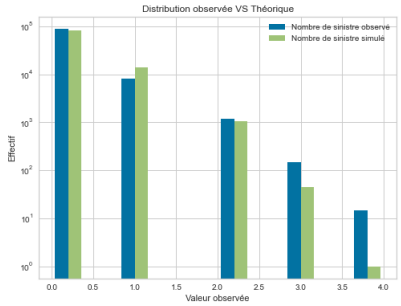
$$\frac{\partial l(n,p)}{\partial p} = \sum_{i=1}^{N} \left[\frac{k_i}{p} - \frac{(n-k_i)}{1-p} \right] = 0$$

On obtient p :

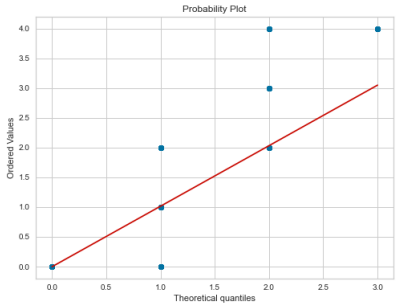
$$p = \frac{\sum_{i=1}^{N} k_i}{Nn}$$

Par application numérique, on retrouve dans notre Cas : $p = 0.028$

Nous avons ainsi simulé un échantillon d'une loi Binomiale(4,0.028) de taille 98620 que nous avons comparer à la distribution observée dans nos données.



Source :Calculs des auteurs



Source :Calculs des auteurs

En regardant le graphique ci-haut , on remarque des écarts entre la distribution empirique et la distribution théorique de la loi binomiale de paramètre p. ce qui ne nous permet pas de nous prononcer sur l'adéquation entre les deux distributions.

3.1.2 Adéquation à une loi de Poisson

Considérons un événement ayant un nombre moyen d'occurrence dans le temps λ. Alors la variable aléatoire qui comptabilise le nombre d'occurrence survenue est dite suivre la loi de poisson de paramètre λ. on la note $X \sim P(\lambda)$. C'est la loi dite des événement rares, tels que les accidents ; le défaut de crédit...

La probabilité d'avoir exactement , K occurrence est :

$$P(X = k) = \frac{\lambda^k}{k!} e^{-k}$$

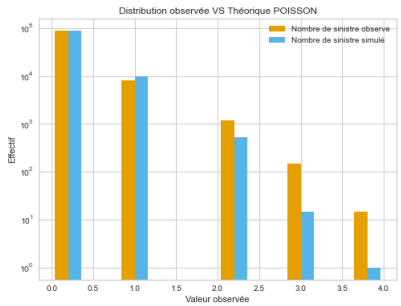
Par maximum de vraisemblance on trouve le paramètre λ.

$$L(x, \lambda) = \prod_{i=1}^{N} \frac{\lambda_i^k}{k!} e^{-\lambda}$$

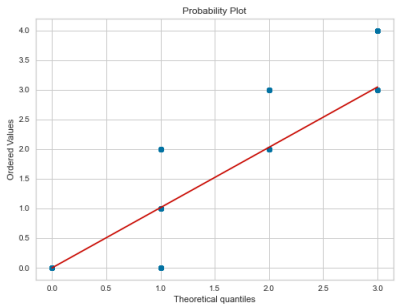
$$ln(L(x, \lambda) = l(x, k) = \sum_{i=1}^{N} k_i ln(\lambda) - \lambda - ln(k!)$$

On dérive et on égalise cette log-vraisemblance à zero pour trouver le paramètre λ qui la maximise. $\frac{\partial l(x,k)}{\partial \lambda} \Leftrightarrow = \frac{\sum_{i=1}^{N} k_i}{N} = E(X)$ Le paramètre est égal à l'espérance (moyenne). Une particularité de cette loi, est que sa variance est égale à sa moyenne. Dans notre cas ces deux paramètre valent respectivement 0.115 et 0.13.

Nous avons ainsi simulé une distribution de poisson de mêmes paramètres pour comparer avec nos observation.



Source :Calculs des auteurs



Source :Calculs des auteurs

La loi de poisson semble bien s'adapter aux données par rapport à la loi Binomiale.

3.1.3 Adéquation à une loi de Binomiale Négative

Soit une expérience qui consiste en une série de tirages indépendants, donnant un succès avec probabilité p. Cette expérience se poursuit jusqu'à l'obtention d'un nombre donné n de succès. La variable aléatoire représentant le nombre d'échecs avant l'obtention de ces n succès, suit alors une loi binomiale négative. Ses paramètres sont n, le nombre de succès attendus, et p, la probabilité d'un succès.

On la note : $X \sim BN(n, p)$.

$$P(X = k) = \binom{k}{k+n-1} P^n (1-p)^k$$

Par maximisation de la vraisemblance on trouve le paramètre p de cette loi.

$$L(x, p) = \prod_{i=1}^{N} C_{k_i+n-1}^{k_i} P^n (1-p)_i^k$$

$$ln(L(x, p) = l(x, p) = \sum_{i=1}^{N} (ln(C_{k_i+n-1}^{k_i}) + nln(p) + kln(1-p))$$

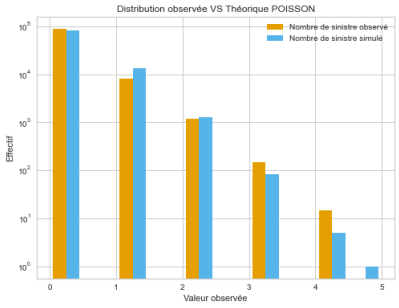
on dérive et on égalise la dérive à zero :

$$\frac{\partial l(x, p)}{\partial p} = \sum_{i=1}^{N} \frac{n}{p} - \sum_{i=1}^{N} \frac{k_i}{1-p} = 0$$

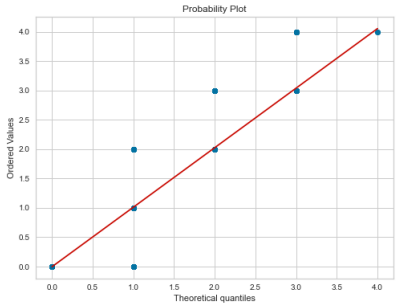
on trouve le paramètre :

$$p = \frac{n}{n + E(x)}$$

Nous avons simulé un échantillon de suivant une loi binomiale négative de même paramètre que notre distribution, comme avec les autres loi ci-hauts.



Source :Calculs des auteurs



Source :Calculs des auteurs

Au regard des graphiques de différentes distributions, nous remarquons que la distribution binomiale négative, est celle qui ressemble beaucoup plus à nos données empiriques. Cette hypothèse est également confirmé par le test de χ^2 d'adéquation. C'est la distribution qui a le plus petit AIC [1]

1. Akaiké's Information Criteriom

TABLE 3.1 – Résultat du test d'adéquation

Chi-squared statistic :		2023.643	10.41067	3340.171
DDL Chi2 distribution		2	1	2
Chi-squaredp-value :		0	0.001252891	0
Chi-squared table :				
	obscounts	theo pois	theo nbinom	theo binom
<=0	89125	88189.81488	89112.595	88088.6522
<=1	8147	9858.08698	8216.9798	10042.09168
<=2	1182	550.98131	1098.5881	476.99866
>2	166	21.11684	191.8371	12.25746
Goodness-of-fit criteria				
		1-mle-pois	2-mle-nbinom	3-mle-binom
Akaike's Information Criterion		72636.62	**71381.54**	73076.9
Bayesian Information Criterion		72646.12	**71400.54**	73086.4

3.1.4 Modélisation : Binomiale Négative

Comme dit dans le paragraphe précédant, Nous allons modéliser la fréquence de nombre de sinistre par un modèle linéaire généralisé famille BINOMIALE NÉGATIVE. Soit Y cette fréquence, notre modèle s'écrit :

$$g(E(Y_i)) = \sum_{j}^{k} \beta_{j,i} X_{j,i}$$

La fonction g [2] étant logarithmique, on retrouve l'espérance de la fréquence pour l'individu i par :

$$E(Y_i) = \exp(\sum_{j}^{k} \beta_{j,i} X_{j,i})$$

.

Nous avons mis en place plusieurs modèles, les différents critères de discrimination, nous ont conduit a choisir comme modèle adéquat :

$$E(Y) = \exp(\beta_1 Vh\_type\_commerciale + \beta_2 vh\_type\_tourisme + \beta_3 pol\_coverage_m axi +$$

$$\beta_4 pol\_coverage\_meadiane + \beta_5 pol\_coverage\_Median2 + \beta_6 pol\_coverage\_Mini + \beta_7 drv\_age1 +_8 vh\_age + densit)$$

Nous avons fait une validation croisée, pour nous assurer de la qualité de prédiction de notre modèle. la base de données à été séparée de façons aléatoire en deux sous bases. Nous avons estimé les paramètres sur les données dites d'entrainement, puis avons prédit avec ces paramètres les valeurs pour les observation de la base de test. Nous avons remarqué que le modèle prédit bien les valeurs pour ces observations, avec une faible erreur.

2. Fonction de lien de loi BN

TABLE 3.2 – Indicateurs de Qualité du modèle par validation croisée

DONNEES	MSE	MAE	MEAN	MEAN PREDICT
Donnees d'entrainnement	0.139556	0.207691	0.116326	0.116326
données de test	0.140538	0.207602	0.116185	0.11623
données globlales	0.139752	0.207673	0.116269	0.116308

Les coefficients du modèle sont tous significatifs sauf celui de la variable densité. on est obligé de garder cette variable dans le modèle car son retrait dégrade la qualité d'ajustement. L'interprétation de coefficient n'a pas grande importance. Car le but étant la prédiction et non l'explication de la fréquence du sinistre. Toutefois ces coefficients sont considérés comme de semis élasticité. C'est à dire la variation en pourcentage de l'espérance de de nombre sinistre pour l'augmentation d'une unité pour les variables quantitatives. Mais elle s'interprète comme variation en pourcentage de l'espérance de nombre de sinistre du fait de posséder une modalité quelconque toute chose égale par ailleurs pour les variables quantitatives pour les variables qualitatives.

TABLE 3.3 – Résultat du modèle

Dep. Variable :	Claim_NbT	**No. Observations :**	98977
Model :	GLM	**Df Residuals :**	98969
Model Family :	NegativeBinomial	**Df Model :**	7
Link Function :	log	**Scale :**	1.0000
Method :	IRLS	**Log-Likelihood :**	-36440.
Date :	Thu, 31 Mar 2022	**Deviance :**	43911.
Time :	23 :38 :21	**Pearson chi2 :**	1.05e+05
No. Iterations :	6		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.3783	0.012	-113.072	0.000	-1.402	-1.354
vh_type_Commercial	-0.6377	0.022	-28.648	0.000	-0.681	-0.594
vh_type_Tourism	-0.7406	0.013	-55.197	0.000	-0.767	-0.714
pol_coverage_Median1	-0.1329	0.035	-3.810	0.000	-0.201	-0.065
pol_coverage_Median2	-0.1692	0.027	-6.210	0.000	-0.223	-0.116
pol_coverage_Mini	-0.6104	0.045	-13.714	0.000	-0.698	-0.523
drv_age1	-0.0316	0.010	-3.184	0.001	-0.051	-0.012
vh_age	-0.2997	0.012	-25.142	0.000	-0.323	-0.276
densité	0.0024	0.010	0.245	0.807	-0.017	0.022

3.1.4.1 Validation du modèle

Deux indicateurs, nous servent à la validation de notre modèle. La déviance et le test de khi deux de Pearson.

la déviance est donnée par : $D^* = 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{\mu}i}\right) - (yi - \hat{\mu}_i)\right)$

La statistique de Pearson est donnée par : $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}i)^2}{V(\hat{\mu}i)}$

Les hypothèses du test : $\begin{cases} H0 : \text{Un bon ajustement entre le modèle et les données} \\ H1 : \text{Pas un bon ajustement.} \end{cases}$

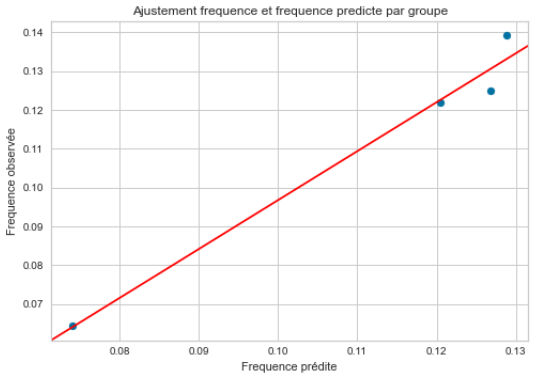
Sous l'hypothèse nulle (H_0) ces deux statistique suivent asymptotiquement la loi $\chi^2_{(n-p-1)}$

En modèle Binomiale et Poisson , on calcule ces statistiques sur les données regroupées pour obtenir des prédictions (μ_i) plus grandes et ainsi augmenter la fiabilité de la loi asymptotique qui devient χ^2_{Cp-1} où C est le nombre de groupes disponibles.

Ainsi, nous allons prendre pour chaque classe d'individus obtenue, lors de notre classification effectuée dans le chapitre précédant, la moyenne de la fréquence et celle de la fréquence prédite. On remplacera dans les formules précédentes Y et μ par leurs moyennes.

Par application numérique, on retrouve les statistiques calculées : $\chi^2 = 0.307$ et une Déviance presque nulle (0.00216), qui sont inférieures à la statistique théorique.

Ainsi , on conclut à un bon ajustement du modèle à nos données. on valide ce modèle pour la prédiction de la fréquence d'apparition de sinistres.



Calculs des auteurs

3.2 Modélisation du cout de sinistre

Le cout ou la charge de sinistre est modélisé selon une loi gamma ou une loi log-normale dans d'autres cas. Ainsi nous allons tester parmi les deux familles de lois, et trouver celle qui s'ajuste mieux à nos données empiriques. Après nous modéliserons selon la loi ayant le meilleur ajustement.

3.2.1 Adéquation à la loi log-normale

Une variable aléatoire X est dite distribuée selon une loi log-normale, si la variable aléatoire Y égale $ln(X)$ suit une loi normale de mêmes paramètres μ et σ^2. On la note $X \sim Log - (\mu, \sigma^2)$.

La densité de x est donnée par :

$$f_x(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) = \frac{1}{x} f_X(\ln(x);\mu,\sigma)$$

Avec $x \gg 0$.

La vraisemblance est quant à elle donnée par :

$$L(x,\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right)$$

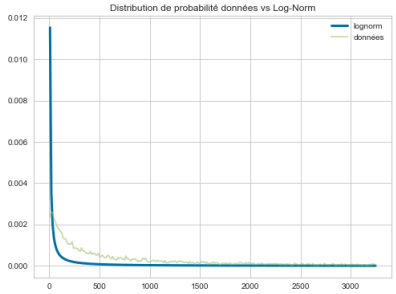
La log-vraisemblance elle, est donnée par :

$$ln(L(x,\mu,\sigma)) = \sum_{i=1}^{N}\left[ln(\frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right))\right]$$

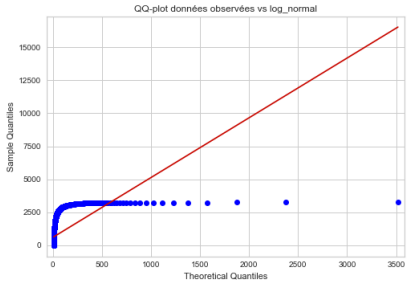
Par maximum de vraisemblance on trouve les deux paramètres μ et σ
$\mu = \frac{\sum_i^n ln(x_i)}{N}$ et $\sigma^2 = \frac{\sum_i^N (ln(x_i)-\mu)^2}{N}$.

Pour nos données empiriques, les deux paramètres sont respectivement 5.74 pour la moyenne μ et 1.21 pour l'écart type σ.

Nous avons ainsi tester l'adéquation de nos données à une loi log normale de même paramètres. Les données empirique semblent ne pas s'ajuster à une loi log-normale Au regard de la distribution de probabilité observée et celle de la loi log-normale. Aussi le QQ-plot montre un décalage important des données observées par rapport à la droite d'Henry.



Source :Calculs des auteurs



Source :Calculs des auteurs

Pour confirmer le diagnostic graphique fait ci haut, nous avons fait des tests statistiques (shapiro wilk, Aderson Darling, Agostino , liliefors , Jarque Bera, Kolmogorov-Smirnov) d'adéquation à une loi normale, pour le logarithme de la charge totale de sinistre. Ces test ont tous à l'exception du test de shapiro-wilk rejeté l'hypothèse de normalité pour le log du Montant total. Au regard de ces résultats , nous concluons que la distribution de cout sinistre ne peut être issue d'une distribution log-normale.

3.2.2 Adéquation à une loi Gamma

Une variable aléatoire X suit une loi gamma de paramètres λ et α notée $X \sim \gamma(\lambda,\alpha)$. Si la fonction de densité de x est donnée :

$$f(x,\lambda,\alpha) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \mathrm{e}^{-\lambda x} x^{\alpha-1} 1_{x\geqslant 0}$$

La vraisemblance pour cette famille de loi est donnée pour les x positifs, par la fonc-

tion :

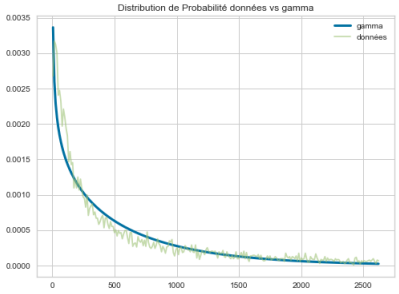
$$L(x,\lambda,\alpha) = \prod_{i=1}^{N} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \mathrm{e}^{-\lambda x_i} x_i^{\alpha-1}$$

et sa log-vraisemblance est donnée par :

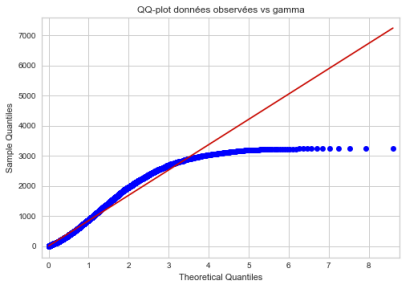
$$ln(L(x,\lambda,)) = \sum_{i=1}^{N} ln\left(\frac{\lambda^{\alpha}}{\Gamma(\alpha)} \mathrm{e}^{-\lambda x_i} x_i^{\alpha-1}\right)$$

Par la méthode de maximum de vraisemblance on trouve les deux paramètres du modèle : $\lambda = \frac{\alpha}{E(x)}$ et $\alpha = \frac{E(X)}{V(x)}$

Nous allons simuler une distribution de loi Gamma de mêmes paramètres que nos données empiriques et faire de comparaisons de deux.

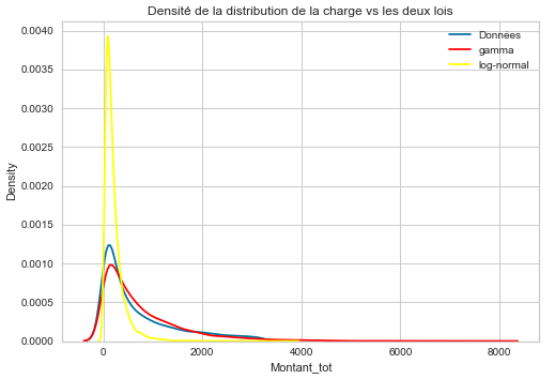


Source :Calculs des auteurs



Source :Calculs des auteurs

En regardant le résultat graphique ci haut, on peut dire que la gamme s'adapte mieux à nos données empiriques par rapport à la loi log-normale. Nous allons par la suite procéder à la modélisation de charge de sinistre, par un modèle de famille gamma.



Calculs des auteurs

3.2.3 Modélisation : Gamma

La loi Gamma est celle qui s'ajuste mieux à données empiriques. Nous allons par suit modéliser le cout pour estimer le cout moyen par un GLM de famille gamma. Ainsi on aura pour Y égal au cout du sinistre :

$$f(E(Y_i)) = \sum_{j}^{k} \beta_{j,i} X_{j,i}$$

La fonction f[3] étant logarithmique, on retrouve l'espérance de la fréquence pour l'individu i par :

$$E(Y_i) = \exp(\sum_{j}^{k} \beta_{j,i} X_{j,i})$$

.

Plusieurs modèles on été estimés. Sur base de critère de discrimination des modèle (AIC, BIC) Nous avons trouvé comme variables pertinentes à garder dans notre modèle final :

TABLE 3.4 – Variables retenues pour le modèle

Assuré	drv_sex1_M	Police	pol_coverage_Maxi pol_coverage_Median1 pol_coverage_Mini pol_usage_WorkPrivate
Véhicule	vh_fuel_Gasoline vh_type_Commercial vh_type_Tourism vh_age vh_value vh_speed vh_weight	Region	densité

La validation croisée du modèle donne de résultats très bons. Ce qui nous rassure qu'on a pas un problème de sur-apprentissage ou sous-apprentissage. Les estimateurs obtenus donnent des bon résultats que ça soit sur l'échantillon de test que sur celui d'apprentissage.

TABLE 3.5 – Indicateurs de Qualité du modèle par validation croisée

DONNEES	MSE	MAE	MEAN	MEAN PREDICT
Donnees d'entrainnement	567575.89	585.31	686.71	686.71
données de test	594610.58	598.59	702.61	694.11
données globlales	572621.18	588.41	689.89	689.89

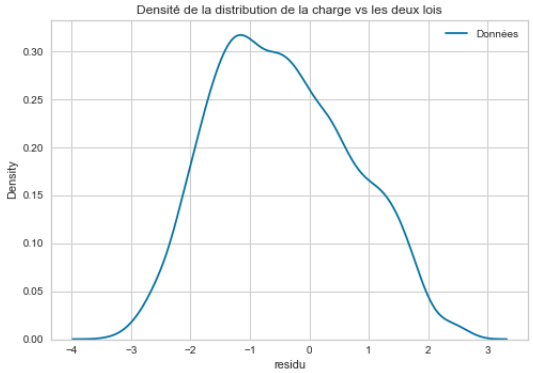
Le modèle gardé est celui ayant le plus petit AIC, au regard des tous les modèles que nous avons testés. Certaines variables ont de coefficients non significatifs statistiquement. on est obligé de garder ces variables dans le modèle car leur retrait dégrade la qualité d'ajustement. L'interprétation de coefficient n'a pas grande importance . Car le but étant la prédiction et non l'explication du cout espéré de sinistre. Toutefois ces coefficients sont considérés comme de semis élasticité. C'est à dire la variation en pourcentage de l'espérance de de nombre sinistre pour l'augmentation d'une unité pour les variables quantitatives. Mais elle s'interprète comme variation en pourcentage de l'espérance de nombre de sinistre du fait de posséder une modalité quelconque toute chose égale par ailleurs pour les variable quantitatives et variation en pourcentage de l'espérance du cout suite à l'augmentation d'une unité supplémentaire toute choses égale par ailleurs pour les variables quantitatives.

3. Fonction de lien de loi gamma

3.2.3.1 Validation du modèle

Comme pour la de modélisation de la fréquence de sinistre. Nous allons nous baser sur la déviance et les résidus pour la validation de notre modèle.

Les résidus studentisés , suivent la loi de student, confirmé par le test de Khi d'ajustement P_value inférieure à 0.05. La test de rapport de vraisemblance confirme également le bon ajustement du modèle aux données empiriques. Ainsi , on conclut à un bon ajustement du modèle à nos données. on valide ce modèle pour la prédiction de l'espérance de cout de sinistres.



Calculs des auteurs

TABLE 3.6 – Résultat du modèle

Dep. Variable :	Montant_tot	**No. Observations :**	9852
Model :	GLM	**Df Residuals :**	9838
Model Family :	Gamma	**Df Model :**	13
Link Function :	inverse_power	**Scale :**	1.2791
Method :	IRLS	**Log-Likelihood :**	-73791.
Date :	Fri, 01 Apr 2022	**Deviance :**	15283.
Time :	00 :09 :37	**Pearson chi2 :**	1.26e+04
No. Iterations :	8		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.0020	6.6e-05	29.813	0.000	0.002	0.002
drv_sex1_M	-4.744e-05	3.34e-05	-1.421	0.155	-0.000	1.8e-05
vh_fuel_Gasoline	6.063e-05	3.85e-05	1.577	0.115	-1.47e-05	0.000
vh_type_Commercial	-0.0002	6.65e-05	-3.110	0.002	-0.000	-7.65e-05
pol_coverage_Maxi	-0.0006	5.73e-05	-10.214	0.000	-0.001	-0.000
pol_coverage_Median1	7.814e-05	9.27e-05	0.843	0.399	-0.000	0.000
pol_coverage_Mini	3.681e-05	0.000	0.335	0.738	-0.000	0.000
pol_usage_WorkPrivate	-8.265e-06	3.41e-05	-0.242	0.809	-7.52e-05	5.87e-05
drv_age1	-2.925e-05	1.67e-05	-1.750	0.080	-6.2e-05	3.51e-06
vh_speed	-2.905e-05	2.65e-05	-1.098	0.272	-8.09e-05	2.28e-05
vh_value	-8.533e-05	1.93e-05	-4.412	0.000	-0.000	-4.74e-05
vh_weight	5.433e-05	1.92e-05	2.833	0.005	1.67e-05	9.19e-05
vh_age	7.716e-05	2.28e-05	3.390	0.001	3.25e-05	0.000
densité	1.426e-05	1.63e-05	0.876	0.381	-1.77e-05	4.62e-05

3.2.4 Calcul de la prime

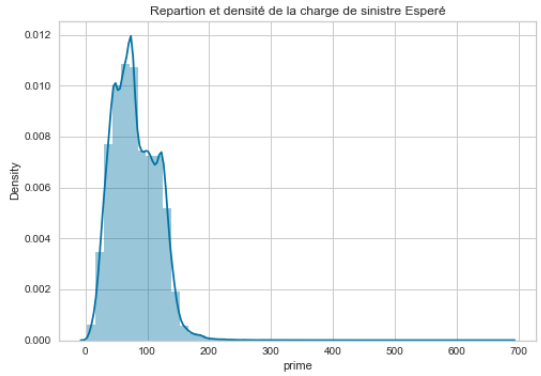
Sous l'hypothèse d'indépendance entre le cout la fréquence de sinistre, la prime pure par le produit de ce deux grandeurs. Comme avons avons classifié nos individus en classes tarifaires, dans le but de la mutualisation, nous aurons ainsi de tarif par classe et non de tarifs individuels.

TABLE 3.7 – Prime pure par classe tarifaire

Classe tarifaire	**Primes**	**Charge chargé**
0	83.702259	72.525346
1	84.106694	72.100866
2	46.520522	34.03876
3	96.944503	88.410036

La prime pur estimée correspond presque parfaitement à la charge moyenne constatée dans chaque classe. Donc le modèle que nous avons mis en place est jugé bon. Il peut être ainsi utilisé pour calculer la prime que devrait payer tout nouveau assuré ou pour le renouvellement de contrat pour les anciens assurés. Il faut alors affecter ce dernier dans

une classe tarifaire à l'aide de l'algorithme de K-means que nous avons également mis en place pour ce fait.



Calculs des auteurs

Conclusion

Au terme de ce travail, nous avons eu à mettre en place un outil de tarification en nous basant sur la régression linéaire généralisée. Nous avons trouvé que que la fréquence de sinistre s'ajustait dans le cadre de nos données empiriques à une loi binomiale négative et le cout de sinistre à une loi gamma.Pour tenir compte d'hétérogénéité des observation, nous avons mis en place un algorithme de classification, pour ainsi tenir compte de la mutualisation. Notre outil est valide au regard de tous les tests faits au cours de différentes modélisations.

Bibliographie

[1] Aasness, J. B. Holtsmark, (1993), Consumer Demand in a General Equilibrium Model for environmental Analysis, Discussion Papers No. 105, Statistics Norway

[2] Denuit Charpentier (2004, 2005), Mathématiques de lAssurance non-vie, Tome 1 et 2

[3] KOUO Kasséa Kévin Axel , Tarification Automobile : GLM vs Réseaux de Neurones, mémoire actuaire , 2017

[4] Quijano, O., and Garrid, J.Generalised linear models for aggregate claims ; to tweedie or not ? Concordia University, Montreal, Canada(2014)

[5] Olga A. VASECHKO, Michel GRUN-RÉHOMME,Noureddine BENLAGHA, (2009), modélisation de la fréquence de sinistre en assurance automobile

[6] Charpentier Arthur (2010), Statistique de l'assurance

[7] Frédéric PLANCHET,Antoine MISERAY (2017), Tarification IARD Introduction aux techniques avancées

[8] Eric Wajnberg (2011),Introduction au Modèle Linéaire Généralisé (Generalized Linear Model ; GLM)

Annexes

Table 3.8 – Tableaux croisé : Nombre de sinistres vs autres caracteristiques

	sinistres	**0**	**1**	**2**	**3**	**4**	**5**	**6**
pol$_c$*ouv*	Maxi	57.438	5.968	0.953	0.13	0.012	0.001	0
	Median1	8.546	0.791	0.103	0.016	0.003	0	0.001
	Median2	15.863	1.373	0.202	0.029	0.004	0.003	0
	Mini	8.043	0.45	0.064	0.004	0.003	0	0
sexe	F	35.923	3.427	0.5	0.077	0.008	0.001	0.001
	M	53.967	5.155	0.822	0.102	0.014	0.003	0
vh_type	Commercial	8.929	0.776	0.126	0.017	0.003	0.001	0
	Tourism	80.961	7.806	1.196	0.162	0.019	0.003	0.001
vh_fuel	Diesel	48.602	5.215	0.849	0.13	0.017	0.004	0
	Gasoline	41.216	3.36	0.472	0.049	0.005	0	0.001
	Hybrid	0.072	0.007	0.001	0	0	0	0