Revue des Sciences Humaines & Sociales
Volume 08, Numéro 03, Décembre 2022, P.P. 857-871.

Individualisation du risque en assurance automobile sur la période (2010-2018) - Cas d'étude CASH assurances-

تفريد المخاطر في التأمين على السيارات خلال الفترة 2010–2018– دراسة حالة CASH للتأمينات–

Individualization of risk in automobile insurance over the period (2010-2018) - CASH insurance case study –

AZZOUZI, Ziad[1]; DJOUADI, Ali*[2]

Date :01/ 06/ 2021 - Date d'acceptation : 31/ 10/ 2021 - Date d'édition : 02/ 12/ 2022

ملخص:

تهدف هذه الدراسة إلى نمذجة عدد ومبلغ الخسائر لحوادث السيارات لاقتراح تفريد المخاطر على السيارات من خلال التقسيم الذي سمح لنا بالحصول على فئات مختلفة من المخاطر وفقًا لخصائص المؤمن له، السيارة والضمانات التي حصل عليها. ولهذا الغرض تم استعمال التحليل الوصفي الذي يعد خطوة رئيسية في نمذجة المخاطر، ويوفر لنا الكثير من الحدس حول المتغيرات التي ليست على استعداد لتقسيم محفظتنا. كما تم استخدام النماذج الخطية المعممة (GLM) لإعداد نموذج تسعير في التأمين على السيارات خلال الفترة 2010 – 2018. بينت النتائج أن مساهمة النماذج الخطية المعممة (GLM) تكمن في الفصل بين دراسة عدد ومبلغ الخسائر. النموذج الأول الذي تم بنائه هو انحدار تكرار الخسائر. النموذج الثاني المطبق هو انحدار متوسط مبلغ الخسائر. وبالتالي، فقد استخدمنا بشكل خاص قوانين جاما وبواسون، المستخدمة تقليديًا لنمذجة متوسط عدد ومبلغ الخسائر لحوادث السيارات.

الكلمات المفتاحية: نمذجة؛ التأمين على السيارات؛ عدد الخسائر؛ مبلغ الخسائر؛ نماذج خطية معممة

Abstract:

The objective of this study is to model the average claim severities and frequency claims independently and propose an individualization of the automobile risk by segmentation which has enabled us to obtain different categories of risk according to the characteristics of the insured, the insured vehicle and the coverage. To do this, a descriptive analysis, which is a crucial step in risk modeling, provided us with a large number of intuitions concerning, among others, the variables that are not ready to segment our portfolio. Then a methodology for the implementation of a pricing model in automobile insurance using generalized linear models was carried out on claims history from 2010 to 2018. The first model is a regression of the observed frequency claims. The second model implemented is a regression of the average claim severity. We have used Gamma and Poisson distribution.

Keyword: Risk modelling; Automobile insurance; claim frequency; claim severity; GLM

Résumé :

Ce travail a pour objectif de modéliser la charge de sinistre et proposer une individualisation du risque automobile par une segmentation qui nous a permis d'obtenir différentes catégories de risques selon les caractéristiques de l'assuré, du véhicule assuré et des garanties souscrites.

* Auteur correspondant.
[1] AZZOUZI Zaid, Akli Mohand Oulhad University of Bouira, Laboratory of Development Policies and Prospective Studies: Algeria, z.azzouzi@univ-bouira.dz.
[2] DJOUADI Ali, Akli Mohand Oulhad University of Bouira, Algeria, a.djouadi@univ-bouira.dz.

Pour ce faire, nous avons réalisé une analyse descriptive étant une étape cruciale dans la modélisation du risque et qui nous a fourni un grand nombre d'intuitions concernant, entre autres, les variables qui ne sont pas prêtes à segmenter notre portefeuille. Ensuite une méthodologie de la mise en place d'un modèle de tarification en assurance automobile par les modèles linéaires généralisés a été menée sur un historique de sinistralité allant de 2010 à 2018. Le premier modèle est une régression de la fréquence de sinistres. Le deuxième modèle mis en place est une régression du coût moyen de sinistres. Nous avons ainsi utilisé plus particulièrement les lois Gamma et Poisson.
Mots clés : modélisation; Assurance automobile; sévérité de sinistre; fréquence de sinistre; MLG

Introduction

Environ 50 % du chiffre d'affaires du secteur des assurances, atteignant les 125.5 milliards de dinars en 2020, provenait de l'assurance automobile qui est qualifiée de marché concurrentiel. Dans un contexte de marché concurrentiel, chaque assureur adopte une tarification qui lui est propre et qui se base sur des considérations techniques et commerciales lui permettant de se différentier et d'adapter le prix de risque au profil de son portefeuille d'assurés. Pour ce faire, la segmentation des risques est nécessaire afin de différencier les classes de risques (assurés) qu'il porte à sa charge et raffiner ses modèles de tarification.

L'enjeu de cette étude est de modéliser la charge de sinistre pour proposer une différentiation du tarif, soit un tarif propre à chaque classe de risque, et prévoir une tarification sur les profils trop risqués, et parallèlement, augmenter la souscription d'affaires nouvelles sur des profils bénéficiaires.

Notre stratégie du travail a été divisée en deux parties. La première partie nous permet de mettre en avant une analyse descriptive qui est une étape cruciale avant toute modélisation des risques en assurance. La deuxième partie viendra aborder la mise en application de notre modèle de tarification.

1- Statistiques exploratoires

Notre étude est basée sur des données provenant du portefeuille d'assurance automobile de la compagnie CASH assurances et notre base de données compte après traitement 76 015 contrats ayant souscrit au moins une garantie dommages.

1-2 Analyse univariée

La répartition des assurés selon le type de garantie dont nous disposons est représentée comme suit :

Tableau -1-**: Répartition des assurés selon le type de garantie**

Garantie	Tous risques	Bris-Glace	Vol-Incendie
Pourcentage (%)	20,91%	41,36%	37,73%

Source: Etabli par l'auteur, 2021

La garantie Bris de Glace occupe la première place dans notre portefeuille avec une part de 41.36% suivie de la garantie Vol-Incendie.

Tableau -2 -**: Répartition de la fréquence des sinistres relatifs à la garantie tous risques**

Nombre sinistres	Fréquence	Pourcentage (%)	Fréquence cumulée	Pourcentage cumulé (%)
0	10 760	68%	10 760	68%
1	4 631	29%	15 391	97%
2	505	3%	15 896	100%

Source: Etabli par l'auteur, 2021

La table 2 présente la répartition de l'échantillon relatif à la garantie tous risques et met en évidence la fréquence de sinistres selon le nombre de sinistres déclarés durant la période 2010-2018.

L'espérance de la variable aléatoire Y est égale à 0.35 dans l'échantillon relatif à la garantie tous risques.

Tableau -3 -: **Répartition et fréquence moyenne des assurés selon le genre**

Genre	F	M
Pourcentage (%)	26,03%	73,97%
Fréquence moyenne	45%	39%
Coût moyen	89,27%	104,69%

Source: Etabli par l'auteur, 2021

Le portefeuille est à dominance masculine, Il y a une majorité écrasante du genre masculin. Ce tableau laisse apparaitre que la fréquence moyenne des femmes est plus élevée à celle des hommes. Cependant, ce constat n'est pas valable pour le cout moyen de sinistres car en termes de sévérité, les sinistres causés par les hommes sont plus importants relativement à l'intensité des femmes.

Tableau -4 -: **Répartition des assurés selon la zone**

Zone	Zone1	Zone2
Pourcentage (%)	96,33%	3,67%

Source: Etabli par l'auteur, 2021

Nous remarquons qu'il y a une majorité écrasante de la classe Zone1. Ce constat pourrait s'expliquer par deux facteurs. Le premier est la disparité en termes de population au sein des zones en Algérie. Le deuxième pourrait s'expliquer par le fait que le taux de pénétration de la zone2 par le réseau commercial de l'entreprise demeure faible. Dans tous les cas de figures, cette variable n'apporte ni de l'information ni de la segmentation. Ainsi donc, elle ne sera pas prise comme une variable de segmentation.

Tableau -5- **: Statistiques des variables qualitatives**

	Moyenne	Ecart-type	Médiane
Age véhicule	9,29	2,30	9
Nb-chevaux	4,84	0,73	5
Age client	43,09	12,48	41

Source: Etabli par l'auteur, 2021

L'âge moyen des véhicules dans notre portefeuille est de 9 ans et la médiane témoigne également que l'âge de la moitié de ces véhicules est supérieur à 9 ans.

2- Analyse bivariée

L'analyse bivariée fût d'une grande importance avant de procéder à la construction des modèles ajustés à nos données. En effet, il est indispensable d'étudier les liens existant entre nos variables. Pour ce faire, nous avons procédé en deux étapes

2-1 Etude de liens entre variables explicatives

Les tests permettant de déceler la corrélation entre variables défèrent selon la nature des variables à tester. Dans notre cas de figure, nous avons fait recours au test d'indépendance du χ^2 **(Denuit & Charpentier, 2005)** et V de cramer car nous ne disposons que des variables qualitatives. Les variables quantitatives ont été transformées en classe (qualitative).

A l'application du Test d'indépendance du χ^2, nous constatons que la p-value est inférieure à 0.05, ce constat nous amène à rejeter l'hypothèse H0 qui suppose l'indépendance entre l'âge du véhicule et celui du client.

De la même manière nous procédons pour l'étude des autres variables dont les résultats sont récapitulés dans le tableau ci-après :

Tableau -6- **: P-value des tests du Khi-deux d'indépendance**

	Age.V	**VV**	**Nb.CH**	**Type- Client**	**Genre**	**Garantie**
Age.C	0.0248	0.3941	0.008248	**0.000156**	0.000084	5.798e-12
Age.V		< 2.2e-16	< 2.2e-16	**< 2.2e-16**	< 2.2e-16	< 2.2e-16
VV			< 2.2e-16	**2.2e-16**	< 2.2e-16	< 2.2e-16
Nb.CH				< 2.2e-16	**< 2.2e-16**	< 2.2e-16
Type - Client					**< 2.2e-16**	**< 2.2e-16**
Genre						< 2.2e-16

Source : Etabli par l'auteur à l'aide de logiciel R, 2021

Les résultats fournis annoncent l'existence de lien entre toutes les variables prises deux à deux à l'exception de la variable âge du véhicule (**Age.C**) et sa valeur (**VV**) avec une p-value de 0.3941 qui est supérieure à 0.05. De ce fait, nous devrions mesurer le degré d'association des variables à l'aide de **l'indice de Cramer** (Rakotomalala, 2011) pour repérer les variables fortement corrélées.

L'indice de Cramer prend ces valeurs dans l'intervalle [0,1], plus qu'il proche de 1 plus que les variables sont corrélées

Tableau -7- **: Indices de Cramer**

	Age.C	**Age.V**	**VV**	**nb.CH**	**Type- Client**	**Genre**	**Garantie**
Age.C	1	0.0101	0.0091	0.0133	0.0150	0.0145	0.0214
Age.V		1	0.3511	0.0842	0.0693	0.0652	0.1399
VV			1	**1**	0.1327	0.1105	0.1119
nb.CH				1	0.1751	0.0985	0.0303
Type - Client					1	0.2102	0.0603
Genre						1	0.0348
garantie							1

Source : Etabli par l'auteur à l'aide de logiciel R, 2021

À première lecture, nous remarquons que les variables prises deux à deux ne sont pas fortement corrélées exception faite pour la variable valeur du véhicule et ses nombres de chevaux (V=1). En effet, ces deux variables sont positivement fortement corrélées. Ceci paraît intuitif puisqu'une voiture puissante aura tendance à se situer dans une classe de valeur élevée.

De plus, il existe un certain lien (V=0.35) entre l'âge du véhicule et sa valeur. Ceci est aussi légitime, car plus l'âge du véhicule augmente plus sa valeur diminue.

D'après l'étude d'association, la variable nb.CH sera éliminée pour la suite de l'étude en raison de l'importance du lien existant avec la variable VV, et ce pour ne pas paramétrer notre modèle par des variables ayant un double effet.

2-2 Etude de liens entre variables explicatives et variable à expliquer

L'étude de lien se fera à l'aide du teste statistique Kruskal-Wallis (Millot, 2011). Le choix du teste est fonction de nombres de modalités de la variable explicative. Nous réalisons le test de Kruskal-Wallis car toutes les variables, après l'élimination de la variable zone, admettent entre trois à sept modalités* .

Les résultats du test pour l'ensemble des variables sont synthétisés dans le tableau ci-dessous :

Tableau -8-**: Récapitulatif des résultats du test Kruskal-Wallis**

	Fréquence	**Coût**
Age.C	p-value = 0.0008544	p-value = 0.0008893
Age.V	p-value < 2.2e-16	p-value < 2.2e-16
VV	p-value < 2.2e-16	p-value < 2.2e-16
Nb.CH	p-value = 2.652e-14	p-value = 1.515e-13
Type. Client	p-value = 8.112e-10	p-value = 0.000000002552
Genre	p-value < 2.2e-16	p-value < 2.2e-16
Garantie	p-value < 2.2e-16	p-value < 2.2e-16

Source : Etabli par l'auteur à l'aide de logiciel R, 2021

La première observation que nous pouvons faire est que toutes nos variables ont une influence significative sur la fréquence et le coût de sinistres (p-value < 0.05), bien que la variable Âge.c affiche une probabilité largement supérieure par rapport aux autres variables. De ce fait, nous attentons un degré d'influence faible de la part de la variable Âge.c sur la fréquence et le coût de sinistres

3- Modélisation de la prime

la modélisation de la prime (Boulanger, 1993) dans notre étude est fondée sur l'hypothèse d'un modèle collectif, dont l'une des propriétés est de permettre l'estimation de la charge moyenne des sinistres en séparant l'étude de cette dernière en l'étude de la fréquence moyenne et du coût moyen de sinistre sous l'hypothèse de l'indépendance entre ces dernières selon la formule :

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

Avec :

S : la charge de sinistre

N : la fréquence (le nombre) de sinistres

* La variable genre compte des valeurs manquantes (non renseignée) et compte tenu de l'information que porte cette variable nous avons fixé leur valeur à « ND » (non disponible) au lieu de les exclure, le non remplissage pouvant être considéré comme une information en lui-même

C : le coût de sinistre.

Notre démarche est la suivante :

Nous commencerons par la construction du modèle de fréquence, puis le modèle de coût. Dans chaque partie, nous allons respecter les étapes suivantes qui sont nécessaires (Atia, 2016) à la validation des modèles.

1. Choix de la distribution théorique ajustée
2. Estimation des paramètres du modèle
3. Sélection des variables
4. Validation du modèle

3-1 Modèle pour la fréquence

Dans notre premier modèle, la variable d'intérêt est la fréquence de sinistres. Notons que la fréquence est égale au nombre de sinistres déclarés divisés par l'exposition, car nous cherchons à une fréquence exprimée à une échelle d'année.

3-1-1 Choix de la distribution théorique

Il consiste à analyser la distribution empirique et la rapprocher à la loi théorique la plus adaptée en procédant par classement sur la base du test d'Anderson - Darling et Kolmogorov - Smirnov. Les résultats de l'ajustement sont illustrés ci-après :

Tableau -9-**: Résultats d'ajustage**

#	**Distribution**	**Kolmogorov Smirnov**	**Anderson Darling**		
		Statistique	Rang	Statistique	Rang
1	**Geometric**	0,90264	2	7129,3	2
2	**Neg.Binomial**	0,92159	3	8140,0	3
3	**Poisson**	0,89775	1	6934,9	1
4	**Bernoulli**	Pas d'ajustage (max > 1)			
5	**Binomial**	Pas d'ajustage			
6	**D. Uniform**	Pas d'ajustage			
7	**Hypergeometric**	Pas d'ajustage			
8	**Logarithmic**	Pas d'ajustage (min < 1)			

Source : Etabli par l'auteur à l'aide de logiciel EasyFit, 2021

Tableau -10- : **Qualité d'ajustement – Synthèse**

#	Distribution	Paramètres
1	**Geometric**	p=0,90264
2	**Neg. Binomial**	n=2 p=0,96
3	**Poisson**	λ=0,10786
4	**Bernoulli**	Pas d'ajustage
5	**Binomial**	Pas d'ajustage
6	**D. Uniform**	Pas d'ajustage
7	**HyperGeometric**	Pas d'ajustage
8	**Logarithmic**	Pas d'ajustage

Source : Etabli par l'auteur à l'aide de logiciel EasyFit, 2021

D'après ces deux tests, notre distribution se rapproche le mieux à une distribution poissoniénne avec un paramètre λ=0.1. Une fois la distribution théorique y est nous passons à l'estimation des coefficients du modèle.

3-1-2 Estimation des coefficients de la régression

Les variables intégrées (avant sélection) au modèle initial ainsi que leur signification sont les suivantes :

Tableau -11-**: variables intégrées au modèle initial de la fréquence de sinistres**

Codification	Signification	Nombre - Modalités
Age.C	Âge du client	5
Age.V	Âge du véhicule	4
VV	Valeur du véhicule	7
Garantie	La garantie souscrite	3
Genre	Le genre du client	3*
Type.client	Type client	3

Source: Etabli par l'auteur, 2021

La fonction de lien appliquée sur la variable à expliquer est la fonction log **(Denuit & Charpentier, 2005)**, ce qui donne la relation à estimer suivante

$$\log([F]) = \beta 0 + \beta 1 x 1 + \cdots + \beta p x p$$

L'estimation des paramètres du modèle de fréquence affiche les résultats suivants :

log ([F]) = -2,87+ 0.2VV2 -0.4VV3 +0.009VV4 -0.02VV5

0.3-VV6 -0.55VV7 +2.23Garantie1 -1.79Ganatie2 -0.05M

0.15-ND -0.06AV2 -0.16AV3 -0.23AV4+0.43Société +0.02AC2

-0.13AC3 -0.02AC4 +0.003AC5 (Voir Annexe 1)

* La variable genre compte des valeurs manquantes (non renseignée) et compte tenu de l'information que porte cette variable nous avons fixé leur valeur à « ND » (non disponible) au lieu de les exclure, le non remplissage pouvant être considéré comme une information en lui-même.

3-1-3 Sélection des variables

Nous distinguons parmi les méthodes de sélection conjointe des variables non significatives, les méthodes de sélection ascendante, descendante et pas à pas mixte. Dans notre cas, la sélection sera faite à l'aide de la méthode de sélection descendante. (Voir Annexe 2)

La variable Age.C a été sélectionnée comme une variable non significative puisque son retrait a baissé le critère d'AIC d'une manière significative. Cela, en fait, coïncide avec l'étude de l'influence des variables explicatives sur la sinistralité où la p-value associée à cette variable était relativement grande par rapport à celles des autres variables.

3-1-4 Validation du modèle

La validation du modèle doit être guidée par la théorie économique. Cela revient à vérifier le signe des coefficients estimés ainsi que leurs ordres de grandeur en examinant les écarts entre les valeurs théoriques (obtenues avec le modèle) et les valeurs observées. Puis, il convient de contrôler la légitimité de ce modèle par sa déviance.

3-1-4-1 Signe des coefficients

Pour contrôler les signes des coefficients du modèle, nous allons vérifier la position des coefficients des modalités de la variable j par rapport au coefficient de sa modalité de référence. En effet, à l'étape de l'estimation, le coefficient de la modalité de référence est fixé à 0 par construction. Avec un lien logarithmique, nous pouvons établir une relation entre le coefficient de la modalité de la variable j et le coefficient de sa modalité de référence.

$$\begin{cases} \textbf{si} \quad \beta > 0 \quad \Rightarrow \quad \exp(\beta) > \exp(0) \\ \textbf{Sinon} \quad \Rightarrow \quad \exp(\beta) < \end{cases}$$

β : coefficient de modalité

En référence à ce qui précède et d'après la positivité de la variable à expliquer, nous obtenons la règle suivante pour contrôler le signe des coefficients de notre modèle de fréquence:

Si le coefficient βj associé à la modalité j est positif, la fréquence moyenne de cette modalité doit être supérieure à la fréquence moyenne de la modalité de référence correspondante. De même, si ce coefficient est négatif, cela témoigne que la fréquence moyenne de la modalité y afférente est inférieure à la fréquence moyenne de la modalité de référence correspondante.

Suivant cette règle, les hommes ont une fréquence moyenne inférieure à celle des femmes (modalité de référence) du fait que le coefficient associé est négatif.

En effet, toutes les modalités ayant un coefficient positif dans notre modèle est caractérisée par une fréquence moyenne supérieure à la fréquence moyenne de la modalité de référence à l'exception de la variable valeur du véhicule « VV » qui déroge à cette règle. Pour ne pas fausser la modélisation, nous avons jugé utile de l'éliminer de notre modèle de fréquence. Nous affichons dans le tableau ci-après la fréquence moyenne de chaque modalité en comparaison avec la modalité de référence.

Tableau -11-**: la fréquence moyenne par modalité en comparaison avec la fréquence moyenne de référence**

Modalité	Coefficients	Fréquence moyenne observée	Fréquence moyenne de référence
M	-0.08329	11%	14%
ND	-0.19613	8%	
Tous risques	2.22831	35%	4%
Vol- Incendie	-1.79362	1%	
AV2	-0.01324	10%	11%
AV3	-0.09059	9%	
AV4	-0.13283	3%	
Société	-0.53585	7%	9%

Etabli par l'auteur à l'aide de logiciel R, 2021

En ce qui concerne la progression des coefficients, Une comparaison de la fréquence moyenne observée avec la progression des coefficients estimés laisse apparaitre une adéquation quasi-parfaite entre ces derniers. En effet, plus l'âge du véhicule augmente plus la fréquence diminue.

3-1-4-2 Ordre de grandeur des coefficients

Il s'agit de voir si la moyenne des valeurs théoriques est proche de la moyenne des valeurs observées dans un premier temps, ensuite de relever l'écart entre les valeurs du modèle et les valeurs observées en rapportant la somme des valeurs observées à la somme des valeurs du modèle

La moyenne des valeurs prédites (0,09) est très proche de la moyenne des valeurs observées du portefeuille (0,10).

3-1-4-3 Déviance du modèle

Si le modèle est en bonne adéquation avec les données, la déviance standardisée doit être proche de la valeur n-p **(CHikhi & CHavance, 2012)** (n étant le nombre d'observations et p variables explicatives).

Le modèle estimé est comparé avec le modèle dit saturé, c'est-à-dire le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. Cette comparaison est basée sur l'expression de la déviance D des log-vraisemblances $\boldsymbol{l}$ et $\boldsymbol{l}_{\text{gaț}}$:

$$D = -2(\boldsymbol{l}\text{-}\boldsymbol{l}_{\text{gaț}})$$

Qui est le logarithme du carrée du rapport des vraisemblances. Ce rapport remplace ou "généralise" l'usage des sommes de carrées propres au cas gaussien et donc à l'estimation par moindres carrées. On montre qu'asymptotiquement, D suit une loi du $\boldsymbol{x^2}$ à n-p degrés de liberté ce qui permet de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

L'approximation **(Besse, 2003)** de la loi du $\boldsymbol{x^2}$ Peut-être douteuse. En pratique, sachant que l'espérance d'une loi du $\boldsymbol{x^2}$ est son nombre de degrés de liberté et, connaissant les aspects

approximatifs des tests construits, l'usage est souvent de comparer les statistiques avec le nombre de degrés de liberté. Le modèle peut être jugé satisfaisant pour un rapport D/ddl plus petit que 1.

Dans notre cas de figure, la déviance standardisée (20677) est toujours inférieure au nombre de degrés de liberté (71658). Ainsi donc, notre modèle est pertinent.

3-2 Modèle pour le coût

Dans cette partite, nous nous intéressons à la construction du modèle dont la variable d'intérêt est les coûts individuels Yi.

3-2-1 Choix de la distribution théorique

Les deux modèles les plus classiques permettant de modéliser (Charpentier, 2010) les coûts individuels de sinistre sont le modèle Gamma sur les coûts individuels Yi et le modèle log-normal sur les coûts individuels Yi.

Compte tenu que ces deux lois sont fréquemment utilisées dans la modélisation des coûts, nous avons choisi de nous servir de la loi Gamma pour modéliser nos coûts.

3-2-2 Estimation des coefficients de la régression

Une fois le modèle et la fonction de lien choisis, il convient d'estimer les paramètres de ce modèle. (Voir Annexe N°4)

Si nous interprétons les paramètres de la régression, nous pouvons dire, en prenant comme exemple l'âge du véhicule, que le coût moyen pour les véhicules âgés entre 0 à 2 ans est caractérisé par une tendance haussière bien qu'il soit en diminution à partir de la 3ème année de circulation. Ce constat ayant été relevé dans la partie analyse univariée s'expliquerait soit par le fait que les assurés négligeaient la déclaration des petits sinistres ,soit par l'acquisition de l'expérience au fur et à mesure avec le vieillissement du véhicule assuré et par conséquent la diminution de la charge de sinistre supportée par l'assureur.

De plus, les hommes ont un coût moyen supérieur à celui des femmes. En effet, l'apport des modèles linéaires généralisés (GLM) réside dans la séparation entre l'étude de coût et l'étude de la fréquence.

3-2-3 Sélection des variables et validation du modèle

Comme précédemment, nous commençons par vérifier le signe des coefficients estimés ainsi que leurs ordres de grandeur en examinant les écarts entre les valeurs théoriques (obtenues avec le modèle) et les valeurs observées. Puis, il convient de contrôler la légitimité de ce modèle par sa déviance.

Étant donné que la même démarche du modèle de la fréquence s'impose, et pour ne pas répéter ce qui a été déjà présenté dans la section précédente, il convient de présenter uniquement les résultats tirés :

Les variables explicatives retenues pour notre modèle de coût sont :

- **Pour l'assuré** : son genre et son type
- **Pour le véhicule** : sa classe d'âge.

Les variables explicatives retenues pour notre modèle de fréquence sont :

- **Pour l'assuré** : son genre
- **Pour le véhicule** : sa valeur et sa classe d'âge.

4- Résultats et Comparaison tarifaire Homme/Femme

Une fois la fréquence moyenne E(N) et le coût moyen E(C) sont estimés, il convient de déduire une estimation du taux de prime. En effet, la combinaison des résultats des deux modèles linéaires généralisés nous permet de déduire la prime pure pour chaque segment tarifaire. Pour mettre en évidence les résultats du travail, nous verrons à travers le tableau suivant la prime pure de la garantie **Bris de Glace** (BDG) par classe de valeur du véhicule en croisement avec le genre du conducteur. Le reste des résultats ne sera pas présenté pour des raisons de confidentialité.

Tableau -11-**: Estimation du taux de prime par valeur du véhicule et genre du conducteur**

PRIME PURE en (DA)	BDG/H	BDG/F
AV1/VV1	630,4665	535,788771
AV1/VV2	869,2754	738,735544
AV1/VV3	979,2725	832,214317
AV1/VV4	1205,332	1024,32588
AV1/VV5	1232,981	1047,82304
AV1/VV6	1311,277	1114,36127
AV1/VV7	1598,003	1358,02946

Etabli par l'auteur à l'aide de logiciel R, 2021

Nous avons vu dans la modélisation de la fréquence et du coût des sinistres que les femmes sont caractérisées d'une part par une fréquence élevée, et d'autre part par un coût moindre par rapport aux hommes. La combinaison de ces deux résultats témoigne que la prime payée par les femmes est inférieure à celle payée par les hommes.

Conclusion

A travers cette étude nous confirmons que la détention d'un maximum d'informations pertinentes sur le risque à assurer est cruciale pour l'assureur pour pouvoir maitriser la segmentation et raffiner ses modèles de tarification en assurance automobile. En effet, l'introduction de nouveaux critères dans les tarifs permet de trouver de nouveaux segments rentables et de répondre au jeu de la concurrence.

Un traitement particulier a été apporté, dans un premier temps, à l'étude des données. Il est en effet important de s'attarder, dans le cadre de la mise en place de modèles statistiques, sur la qualité des données et les liens existant entre nos variables de tarification pour veiller à la qualité de notre modélisation. En effet, l'analyse descriptive, qui est une étape cruciale dans la modélisation du risque, nous a fourni un grand nombre d'intuitions concernant, entre autres, les variables qui ne sont pas prêtes à segmenter notre portefeuille comme la variable zone (zone géographique) du fait que la modalité zone1 représentant les wilayas de la région Nord contient à elle seule 97% des observations de la variable zone. De même, elle nous a permis également de faire état de la situation actuelle du portefeuille et de détecter d'éventuelles erreurs ayant préalablement échappé lors du traitement de la base de données.

Nous avons vu ensuite une méthodologie de la mise en place d'un modèle de tarification en assurance automobile par les modèles linéaires généralisés. D'après les résultats fournis, nous avons montré que l'apport des modèles linéaires généralisés (GLM) réside dans la séparation entre l'étude de coût et l'étude de la fréquence.

L'application de l'approche fréquence-coût nous a permis de dégager certains points importants concernant la tarification du risque automobile et de répartir la charge de sinistre d'une manière plus équitable entre les différents segments relevés suivant leurs caractéristiques et en fonction du risque qui leur est associé.

Il s'agit d'une approche de tarification dans l'évolution constante des méthodes de tarification et de l'individualisation du risque automobile dans le souci d'adapter le tarif avec l'évolution du portefeuille et du risque assuré. Cette individualisation du risque vise en premier lieu à fidéliser les bons assurés, car chaque assuré paie pour son risque, et à faire face à la pression de la concurrence par la détection de nouveaux segments rentables en deuxième lieu.

Annexes :

Annexe 1

Figure -1-**: Estimation des coefficients du modèle initial de fréquence**

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0539  -0.3020  -0.1972  -0.1130   4.6987

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -2.870965   0.063559 -45.170  < 2e-16 ***
VVVV2                                 0.021277   0.045498   0.468   0.6400
VVVV3                                -0.043982   0.048850  -0.900   0.3679
VVVV4                                 0.000942   0.051540   0.018   0.9854
VVVV5                                -0.025959   0.055306  -0.469   0.6388
VVVV6                                -0.311729   0.067334  -4.630 3.66e-06 ***
VVVV7                                -0.554265   0.066070  -8.389  < 2e-16 ***
garantieDASC                          2.236835   0.034929  64.039  < 2e-16 ***
garantievol-incendie                 -1.792740   0.083706 -21.417  < 2e-16 ***
genreM                               -0.056617   0.040003  -1.415   0.1570
genreND                              -0.156548   0.038689  -4.046 5.20e-05 ***
Age.VAV2                             -0.060808   0.035710  -1.703   0.0886 .
Age.VAV3                             -0.161385   0.031805  -5.074 3.89e-07 ***
Age.VAV4                             -0.230465   0.058217  -3.959 7.54e-05 ***
type.clientSociété                   -0.436404   0.041429 -10.534  < 2e-16 ***
Age.CAC2                              0.024659   0.035442   0.696   0.4866
Age.CAC3                             -0.013872   0.036402  -0.381   0.7032
Age.CAC4                             -0.025873   0.039115  -0.661   0.5083
Age.CAC5                              0.003849   0.041612   0.093   0.9263
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 32181  on 71666  degrees of freedom
Residual deviance: 20527  on 71648  degrees of freedom
AIC: 32978

Number of Fisher Scoring iterations: 7
```

Source : Etabli par logiciel R, 2021

Annexe 2

Figure -2-**: Processus de sélection des variables**

```
Start:  AIC=32977.7
Fréquence ~ type.client + garantie + genre + Age.V + Age.C +
    VV

               Df Deviance   AIC
- Age.C         4    20529 32972
<none>              20527 32978
- genre         2    20548 32995
- Age.V         3    20559 33004
- type.client   1    20646 33095
- VV            6    20675 33113
- garantie      2    31133 43580

Step:  AIC=32971.64
Fréquence ~ type.client + garantie + genre + Age.V + VV

               Df Deviance   AIC
<none>              20529 32972
- genre         2    20550 32989
- Age.V         3    20561 32998
- type.client   1    20648 33089
- VV            6    20677 33107
- garantie      2    31146 43585
```

Source : Etabli par logiciel R, 2021

Annexe 3

Figure -2-:**Estimation des coefficients du modèle de fréquence après sélection des variables**

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0402  -0.3019  -0.1972  -0.1129   4.6934

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.873314   0.060003 -47.886  < 2e-16 ***
VVVV2                                   0.021631   0.045492   0.475   0.6344
VVVV3                                  -0.043257   0.048842  -0.886   0.3758
VVVV4                                   0.001417   0.051536   0.027   0.9781
VVVV5                                  -0.025566   0.055301  -0.462   0.6439
VVVV6                                  -0.311374   0.067329  -4.625 3.75e-06 ***
VVVV7                                  -0.554191   0.066065  -8.389  < 2e-16 ***
garantieDASC                            2.237303   0.034928  64.054  < 2e-16 ***
garantievol-incendie                   -1.793036   0.083705 -21.421  < 2e-16 ***
genreM                                 -0.055999   0.039993  -1.400   0.1614
genreND                                -0.156493   0.038673  -4.047 5.20e-05 ***
Age.VAV2                               -0.060128   0.035702  -1.684   0.0922 .
Age.VAV3                               -0.161318   0.031807  -5.072 3.94e-07 ***
Age.VAV4                               -0.229834   0.058209  -3.948 7.87e-05 ***
type.clientSociété                     -0.435396   0.041414 -10.513  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 32181  on 71666  degrees of freedom
Residual deviance: 20529  on 71652  degrees of freedom
AIC: 32972

Number of Fisher Scoring iterations: 7
```

Source : Etabli par logiciel R, 2021

Annexe 4

Figure -2-: **Estimation du modèle de coût après sélection.**

```
Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             9.213971   0.122958  74.936  < 2e-16 ***
type.clientSociété                     -0.008411   0.083058  -0.101  0.91934
garantieDASC                            0.822334   0.069379  11.853  < 2e-16 ***
garantievol-incendie                    0.882791   0.173966   5.074 4.00e-07 ***
VVVV2                                   0.321401   0.091231   3.523  0.00043 ***
VVVV3                                   0.440133   0.098179   4.483 7.50e-06 ***
VVVV4                                   0.648156   0.103575   6.258 4.17e-10 ***
VVVV5                                   0.671297   0.110641   6.067 1.38e-09 ***
VVVV6                                   0.598090   0.134458   4.448 8.82e-06 ***
VVVV7                                   0.931332   0.131119   7.103 1.36e-12 ***
genreM                                  0.246230   0.081418   3.024  0.00250 **
genreND                                 0.255395   0.079213   3.224  0.00127 **
Age.VAV2                                0.117016   0.071702   1.632  0.10274
Age.VAV3                               -0.040201   0.063573  -0.632  0.52718
Age.VAV4                               -0.076456   0.116179  -0.658  0.51050
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 3.739255)

    Null deviance: 9188.5  on 6046  degrees of freedom
Residual deviance: 8179.7  on 6032  degrees of freedom
AIC: 140097

Number of Fisher Scoring iterations: 7
```

Source : Etabli par logiciel R, 2021

Bibliographie

1. Atia, R. (2016). Mise en place de modèles de tarification alternatifs face à la suppression règlementaire d'une variable tarifaire en automobile. *Master actuariat de dauphine*. Paris.
2. Besse, P. (2003, Janvier). Data mining II. Modélisation Statistique & Apprentissage. *Publications du Laboratoire de Statistique et Probabilités Université Paul Sabatier*, p. 110.
3. Boulanger, F. (1993). Individualisation du risque en assurance automobile. *JOURNAL DE LA SOCIETE STATISTIQUE DE PARIS*, p. 31.
4. Charpentier, A. (2010, Decembre 28). Statistique de l'assurance. *HAL* , p. 50.
5. CHikhi, M., & CHavance, M. (2012, Juin). ESTIMATION DU MODELE LINEAIRE GENERALISE ET APPLICATION. *Sciences & Technologie A*, p. 19.
6. Denuit, M., & Charpentier, A. (2005). *Mathématiques de l'assurance non-vie (Tarification et provisionnement).* ECONOMICA.
7. Millot, G. (2011). *Comprendre et réaliser les tests statistiques à l'aide de R.* Bruxelles: De boek.
8. Rakotomalala, R. (2011, Mars). Cours de statistiques. *Étude des dépendances, Variables qualitatives*. France: Université Lumière Lyon 2.