

جامعة أكلي محمد اولحاج - البويرة / الجزائر  
مجلة معارف

مجلة علمية دولية محكمة (صنف ج) (Catégorie C) Revue académique

تصدر عن جامعة أكلي محمد اولحاج / البويرة

ر.ت.م.د (ISSN) 1112-7007  
ر.ت.م.د.إ (EISSN) 6804-2600  
رقم الإيداع القانوني: 2006-1369

التاريخ: 2022/02/01  
الرقم: 37/ج.ب.م.م/2022

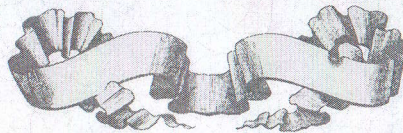
إفادة بالنشر

د. زياد عزوزي المحترم. جامعة البويرة. د. علي جوادني المحترم. جامعة البويرة.

تحية طيبة وبعد...

فإن رئاسة تحرير مجلة معارف، يسعدنا إفادتكم بأن البحث الذي تقدمتم به إلى المجلة: «Etude de la qualité des assurés en assurance automobile sur la période (2018-2010) - Cas d'étude CASH assurances- 16 ماي 2021، قد تم نشره في المجلد 16، العدد 2، الصادر في 31 ديسمبر 2021، وهو موجود في منصة ASJP على الرابط: <https://www.asjp.cerist.dz/en/article/178253> وهي تدعو لكم بمزيد من التوفيق في البذل والعطاء العلمي... كما يسعدنا استمرار التواصل مع المجلة.

سلمت هذه الوثيقة بناءً على طلب المعني لاستعمالها في حدود ما يسمح به القانون.



رئيس التحرير  
معارف  
مجلة علمية  
محكمة  
الجامعة الجزائرية  
أ. شيهاني سمير

**Etude de la qualité des assurés en assurance automobile sur  
la période (2010-2018) -Cas d'étude CASH assurances –**

**Study of the quality of insureds in automobile insurance over  
the period (2010-2018) - CASH assurances case study –**

Ziad AZZOUZI <sup>(1)</sup>; Ali DJOUADI <sup>(2)\*</sup>.

<sup>(1)</sup> Université de Bouira, Laboratoire des Politiques du développement  
et des Études Prospectives ; Algérie.

z.azzouzi@univ-bouira.dz.

<sup>(2)</sup> Université de Bouira, Algérie, a.djouadi@univ-bouira.dz

Date de réception: 16/05/2021; Date d'admission: 17/10/2021; Date de  
publication: 31/12/2021.

**Le résumé :**

Note étude vient proposer un modèle statistique permettant de déterminer le nombre de sinistres de chaque classe de risque en assurance automobile et par conséquent prévoir la qualité des assurés.

Pour ce faire, une 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. Puis, à l'aide des modèles linéaires généralisés (GLM), il a été mis en place une régression de la fréquence de sinistres observés sur un historique de sinistralité allant de 2010 à 2018. À l'aide de ce modèle de fréquence, nous avons décelé la qualité des assurés par la modélisation de la fréquence de sinistres de chaque classe de risque.

**Mots- Clés:** Assurance automobile; Nombre de sinistres; Modélisation; GLM.

---

\* Auteur correspondant.

**Abstract:**

This study suggests a statistical model to determine the number of claims for each risk class in automobile insurance and consequently to predict the quality of the insured.

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, using generalized linear models (GLM), a regression of the observed claims frequency on claims history from 2010 to 2018 was implemented. Using this frequency model, we detected the quality of the insureds by modeling the claims frequency of each risk class.

**Keywords:** Automobile insurance; Number of claims; Risk modeling; GLM.

**ملخص:**

تهدف هذه الدراسة إلى اقتراح نموذج إحصائي يسمح بتحديد عدد الخسائر للحوادث في التأمين على السيارات وبالتالي التنبؤ بجودة المؤمن عليه، ولهذا الغرض تم استعمال التحليل الوصفي الذي يعد خطوة رئيسية في نمذجة المخاطر، ويوفر لنا الكثير من الجدس حول المتغيرات التي ليست على استعداد لتقسيم محفظتنا. كما تم استخدام النماذج الخطية المعممة (GLM) لإجراء انحدار تكرار الخسائر خلال الفترة 2010-2018، حيث تم دراسة نوعية المؤمن عليهم من خلال تقدير عدد الخسائر لكل فئة. الكلمات المفتاحية: التأمين على السيارات؛ عدد الخسائر؛ النمذجة: GLM.

**1. INTRODUCTION**

L'économie s'est toujours intéressée à la réalisation de l'équilibre général sur le marché dans un environnement certain de façon à satisfaire l'offre et le demandeur.

Face à la détention de l'information par l'assuré, l'assureur se trouve avec une sinistralité très élevée qu'il ne pourra pas supporter étant donné l'inadéquation des primes demandées aux assurés risqués, ce qui déséquilibrait son bilan. C'est pourquoi, la segmentation des

risques est nécessaire pour différencier les classes de risques (assurés) qu'il porte à sa charge.

Tant en tarification qu'en souscription, la segmentation permettra une sélection des risques et donnera à l'assureur le recul nécessaire afin de prendre les mesures adéquates envers chaque risque qu'il prend à sa charge et de calibrer ses modèles de tarification en fonction des informations tirées de l'historique de la sinistralité de chaque classe de risque.

L'enjeu de cette étude est de modéliser le nombre de sinistres pour proposer une différenciation du tarif, soit une adaptation des primes au profils de risque ou à chaque classe de risque, et prévoir une surprime 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 fréquence.

## **2. Statistiques exploratoires:**

### **2.1 Analyse univariée:**

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 76015 contrats ayant souscrit au moins une garantie dommages.

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**

<b>Garantie</b>	<b>Tous risques</b>	<b>Bris-Glace</b>	<b>vol-Incendie</b>
<b>Pourcentage (%)</b>	20,91%	41,36%	37,73%

**Source** : Etabli par l'auteur, 2021

**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%

**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.

**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
<b>Age véhicule</b>	9,29	2,30	9
<b>Nb-chevaux</b>	4,84	0,73	5
<b>Age client</b>	43,09	12,48	41

**Source** : Etabli par l'auteur, 2021

## **2.2 Analyse bivariée :**

### **2.2.1 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<sup>(1)</sup> du  $\chi^2$  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  $\chi^2$ , 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 :

(1) Michel Denuit, Arthur Charpentier, "Mathématiques de l'assurance non-vie (Tarification et provisionnement)", ECONOMICA, 2005, tome2, page 117.

**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	<b>0.000156</b>	0.000084	5.798e-12
Age.V		< 2.2e-16	< 2.2e-16	< <b>2.2e-16</b>	< 2.2e-16	< 2.2e-16
VV			< 2.2e-16	<b>2.2e-16</b>	< 2.2e-16	< 2.2e-16
Nb.CH				< 2.2e-16	< <b>2.2e-16</b>	< 2.2e-16
Type Client					< <b>2.2e-16</b>	< <b>2.2e-16</b>
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.V**) 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**<sup>(2)</sup> pour repérer les variables fortement corrélées.

L'indice de Cramer prend ces valeurs dans l'intervalle [0,1], plus qu'il est 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
-------	-------	----	-------	-------------	-------	----------

(2) Ricco Rakotomalala. Cours de statistiques,– Étude des dépendances, Variables qualitatives, Université Lumière Lyon 2, 2011, page 15

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

### 2.2.2 Etude de liens entre variables explicatives et variable à expliquer

L'étude de lien se fera à l'aide du teste statistique Kruskal-Wallis<sup>(3)</sup>. Le choix du teste est fonction de nombres de modalités de la variable explicative. Nous réalisons le test de Kruskal-Wallis car

(3) Gaël Millot, Comprendre et réaliser les tests statistiques à l'aide de R, 3ème édition, de boek, page 511.

toutes les variables, après l'élimination de la variable zone, admettent entre trois à sept modalités.

Les résultats sont synthétisés dans le tableau ci-dessous :

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

	Fréquence
Age.C	p-value = 0.0008544
Age.V	p-value < 2.2e-16
VV	p-value < 2.2e-16
Nb.CH	p-value = 2.652e-14
Type. Client	p-value = 8.112e-10
Genre	p-value < 2.2e-16
Garantie	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 (p-value < 0.05), bien que la variable  $\hat{A}ge.c$  affiche une probabilité largement supérieure par rapport aux autres variables. De ce fait, nous attendons un degré d'influence faible de la part de la variable  $\hat{A}ge.c$  sur la fréquence.

### 3. Modélisation de la fréquence des sinistres :

La construction du modèle de fréquence fait appel aux étapes suivantes qui sont nécessaires à sa validation, à savoir<sup>(4)</sup> :

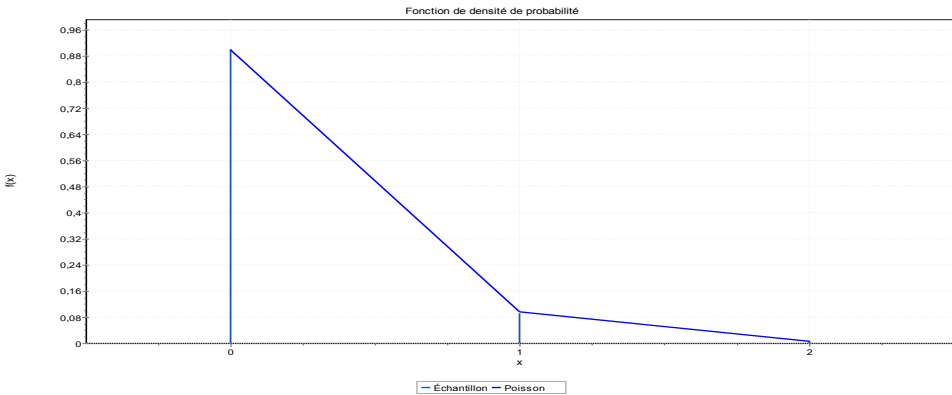
1. Choix de la distribution théorique ajustée
2. Estimation des paramètres du modèle
3. Sélection des variables
4. Validation du modèle

#### 3.1 Choix de la distribution théorique :

(4) ATIA Rachel. Mise en place de modèles de tarification alternatifs face à la suppression réglementaire d'une variable tarifaire en automobile, Master actuariat de dauphine, Paris, 2016, page 36

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 teste de **Anderson-Darling et Kolmogorov-Smirnov**<sup>(5)</sup>. Les résultats de l’ajustement sont illustrés ci-après :

**Figure 1: Ajustement de la distribution théorique à une distribution de la loi de poisson**



**Tableau 9 : Résultats d'ajustage**

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

(5) Gaël Millot, Comprendre et réaliser les tests statistiques à l’aide de R, 3ème édition, de boek, 2015, page 624

**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	$\lambda=0,10786$
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

### 3.2 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*

\* 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 =

<b>Type.client</b>	Type client	3
--------------------	-------------	---

**Source** : Etabli par l'auteur, 2021

La fonction de lien appliquée sur la variable à expliquer est la fonction  $\log^{(7)}$ , ce qui donne la relation à estimer suivante

$$\log ([F]) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

**Coefficients estimés** :  $\log ([F]) = -2,87 + 0.2VV_2 - 0.4VV_3 + 0.009VV_4 - 0.02VV_5 - 0.3VV_6 - 0.55VV_7 + 2.23\text{Garantie}_1 - 1.79\text{Ganatie}_2 - 0.05M - 0.15ND - 0.06AV_2 - 0.16AV_3 - 0.23AV_4 + 0.43\text{Société} + 0.02AC_2 - 0.13AC_3 - 0.02AC_4 + 0.003AC_5$  (Voir Annexe 1)

### 3.3 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.

### 3.4 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.

disponible) au lieu de les exclure, le non remplissage pouvant être considéré comme une information en lui-même.

(7) Michel Denuit, Arthur Charpentier, " Mathématiques de l'assurance non-vie (Tarification et provisionnement) ", ECONOMICA, 2005, tome2, page 98

Puis, il convient de contrôler la légitimité de ce modèle par sa déviance.

### 3.4.1 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.

$$\left\{ \begin{array}{l} \text{si } \beta > 0 \Rightarrow \exp(\beta) > \exp(0) \\ \text{Sinon} \quad \Rightarrow \exp(\beta) < \exp(0) \end{array} \right.$$

$\beta$  : 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  $\beta_j$  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 12 : la fréquence moyenne par modalité en comparaison avec la fréquence moyenne de référence**

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

**Source** : Etabli par l'auteur, 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.

### **3.4.2 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).

### 3.4.3 Déviance du modèle<sup>(8)</sup>

Si le modèle est en bonne adéquation avec les données, la déviance standardisée doit être proche de la valeur  $n-p$  ( $n$  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  $l$  et  $l_{\text{gat}}$  :

$$D = -2(l - l_{\text{gat}})$$

Qui est le logarithme du carré du rapport des vraisemblances. Ce rapport remplace ou "généralise" l'usage des sommes de carrés propres au cas gaussien et donc à l'estimation par moindres carrés. On montre qu'asymptotiquement,  $D$  suit une loi du  $\chi^2$  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<sup>(9)</sup> de la loi du  $\chi^2$  Peut-être douteuse. En pratique, sachant que l'espérance d'une loi du  $\chi^2$  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  $D/\text{ddl}$  plus petit que 1.

---

(8) Malika CHIKHI, Michel CHAVANCE, Estimation du modèle linéaire généralisé et application, Sciences & Technologie, université Mentouri Algérie A – N°35, 2012, page 19

(9) PHILIPPE BESSE, Data mining II. Modélisation Statistique & Apprentissage, Laboratoire de Statistique et Probabilités — UMR CNRS C5583 Université Paul Sabatier — 31062 – Toulouse cedex, page 110.

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.

#### 4. La conclusion

La détention d'un maximum d'informations pertinentes sur le risque à assurer est cruciale pour l'assureur pour pouvoir maîtriser 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.

Dans un premier temps il était important de s'attarder, dans le cadre de la mise en place de modèles statistiques, sur la qualité des données et les liens existants entre nos variables de segmentation pour veiller à la qualité de notre modélisation.

Nous avons vu une méthodologie de la mise en place d'un modèle de fréquence en assurance automobile par les modèles linéaires généralisés.

L'application de ce modèle nous a permis de déceler la qualité de l'assuré par la détermination de sa probabilité de survenance de risque, ce qui confère à l'assureur la possibilité de développer son portefeuille tout en veillant à l'équilibre de son bilan. 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.

#### 5. Les ouvrages

- (1) Philippe Besse, Data mining II. **Modélisation Statistique & Apprentissage, Statistique et Probabilités** — UMR CNRS C5583 Université Paul Sabatier — 31062 – Toulouse, 2003.
- (2) Gaël Millot, **Comprendre et réaliser les tests statistiques à l'aide de R**, 3ème édition, de boek, 2015.

- (3) Michel Denuit, Arthur Charpentier, **Mathématiques de l'assurance non-vie (Principes fondamentaux de théorie du risque), tome 1**, ECONOMICA, 2004.
- (4) Michel Denuit, Arthur Charpentier, **Mathématiques de l'assurance non-vie (Tarification et provisionnement), tome 2**, ECONOMICA, 2005.
- (5) Murielle PAULIN, **économie et organisation de l'assurance**, Séfi édition, Québec, Canada, 2007.
- (6) Jean Luc Besson et Christian Partrat, **Assurance non vie - modélisation et simulation**, Economica, Paris 2004.

#### **Articles et Thèses de recherche**

- (1) Malika CHIKHI, Michel CHAVANCE, **Estimation du modèle linéaire généralisé et application**, Sciences & Technologie, université Mentouri Algérie A – N°35, 2012.
- (2) ATIA Rachel. **Mise en place de modèles de tarification alternatifs face à la suppression réglementaire d'une variable tarifaire en automobile**, Master actuariat de dauphine, 18 janvier 2016.
- (3) Rakotomalala, R. **Cours de statistiques (Étude des dépendances, Variables qualitatives)**, Université Lumière Lyon 2, 2011.
- (4) Olfa N. Ghali, **Un modèle de tarification optimal pour l'assurance automobile dans le cadre d'un marché réglementé : application à la Tunisie**, Cahier de recherche 01-09, Décembre 2001.

#### **Sites**

- (1) <http://www.ressources-actuarielles.net/memoire>

#### **6. Annexes :**

## Annexe 1

**Figure 2: 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 ***
VWV2            0.021277   0.045498   0.468  0.6400
VWV3           -0.043982   0.048850  -0.900  0.3679
VWV4            0.000942   0.051540   0.018  0.9854
VWV5           -0.025959   0.055306  -0.469  0.6388
VWV6           -0.311729   0.067334  -4.630 3.66e-06 ***
VWV7           -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 3 : processus de sélection des variables**

```

Start: AIC=32977.7
fréquence ~ type.client + garantie + genre + Age.V + Age.C +
  VV
- Age.C          Df Deviance   AIC
<none>          4      20529 32972
- genre          2      20527 32978
- Age.V          3      20548 32995
- type.client    1      20559 33004
- VV             6      20646 33095
- garantie       2      20675 33113
-               2      31133 43580

Step: AIC=32971.64
fréquence ~ type.client + garantie + genre + Age.V + VV
<none>          Df Deviance   AIC
- genre          2      20529 32972
- Age.V          3      20550 32989
- type.client    1      20561 32998
- VV             6      20648 33089
- garantie       2      20677 33107
-               2      31146 43585
  
```

Source : Etabli par logiciel R, 2021

Annexe 3

**Figure 3 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