Modélisation des montants des sinistres à l'aide de la méthode GLM (SAS / R)

Article · April 2017

CITATIONS
0

READS
4,245

1 author:

Rym Chekayri
Université Mohammed VI Polytechnique
1 PUBLICATION **0** CITATIONS

SEE PROFILE

EMINES
School of Industrial Management

MOHAMMED VI
POLYTECHNIC
UNIVERSITY

Avril 2017

Modélisation du montant des sinistres - GLM

Auteurs :
BAKSSOU Nedal
CHEKAYRI Rym
NAIM Hamza
TAGHIA Oumayma

Professeur :
BENJELLOUN Saad
Encadrant :
MARRI Fouad

Table des matières

1 Introduction

La question de l'évaluation des fonds propres nécessaires à la viabilité d'une entreprise a toujours suscité des réflexions au sein des communautés scientifique, industrielle et financière. Dans le domaine de l'assurance automobile, cette viabilité est mesurée grâce à ce qu'on appelle la « marge de solvabilité » qui représente le montant des fonds propres que doit impérativement détenir la société d'assurance pour garantir ses engagements vis-à-vis des assurés et pour leur permettre de faire face aux aléas inhérents à son activité.

Depuis la mise en vigueur d'une réglementation relative à la marge de solvabilité dans plusieurs pays (Maroc, France...), l'évaluation des fonds propres nécessaires aux différentes activités d'assurance demeure un sujet de recherche continu. Cette évaluation, qui ne reposait autrefois que sur les données comptables de la société, se fait aujourd'hui grâce à des modèles statistiques qui prédisent le risque de sinistralité en fonction des données des assurés. Cette information permet aux sociétés d'assurance de prendre des dispositions majeures qui se traduisent par la détermination préalable :

— Des montants des primes, tarifés et provisionnés prudemment par mutualisation des risques.

— Des fonds propres, couvrant les risques non mutualisables causés par des erreurs de modèles ou de paramètres.

L'objet du présent article est de proposer une démarche et une modélisation pour l'évaluation des montants des sinistres relatifs à l'activité d'une assurance automobile.Notre démarche confronte ainsi des modélisations, qui comportent nécessairement des aspects théoriques avec des résultats expérimentaux. L'étude que nous avons réalisée à ce sujet a adopté une approche sur la base :

— Des informations sur les assurés d'un échantillon donné et sur leurs biens assurés.

— De résultats portant, exercice par exercice, sur une période pluriannuelle (de 2007 à 2011).

Dans ce qui suit, nous présentons brièvement le traitement opéré au niveau des bases de données. Ensuite, nous partageons les différentes méthodes adoptées pour modéliser notre variable à expliquer que nous comparons sur la base de tests statistiques. Finalement, nous exposons et analysons de manière critique nos résultats.

2 Traitement et analyse des données

Comme nous l'avons précédemment mentionné, notre objectif est de proposer une démarche et une modélisation pour l'évaluation des montants des sinistres relatifs à l'activité d'une assurance automobile. Pour ce faire, nous utilisons le logiciel SAS – *Statistical Analysis System* University Edition et le logiciel GNU R, sur lesquels nous suivons les étapes suivantes :

2.1 Préparation de la base des données

Notre projet s'appuie sur deux bases de données distinctes : la base "Production" qui comporte 145 325 observations et la base "Sinistres" qui en comporte 53 589. Evidemment, ces dernières ne représentent qu'un échantillon d'une population plus grande d'assurés.

- La base "Production" :
Contient des informations sur tous les assurés qui ont contracté une assurance auprès de notre société d'assurance fictive et sur leurs véhicules. Celle-ci contient 10 variables, dont certaines sont relatives à l'assuré : NUMERO_POLICE (Identifiant de l'assuré), Exercice (Année observée), exposition (Ratio de la durée du contrat en mois sur 12), SEXE, Date de naissance, Date obtention du permis. et Date du premier effet. (Début du premier contrat), tandis que d'autres sont liées au véhicule assuré : combustion, puissance_fiscale et Date de Mise en Circulation.

- La base "Sinistres" :
Regroupe des informations sur les sinistres qui ont eu lieu au cours d'un exercice donné. Cette base ne contient que 4 variables : Police (Identifiant de l'assuré), numero_sinistr (Identifiant du sinistre), Charge (Montant versé à l'assuré) et Exercice (Année observée).

Les variables présentes dans les deux bases de données précédentes ne peuvent pas être utilisées par un modèle statistique sans un traitement préalable. En effet la présence de valeurs aberrantes et de valeurs manquantes posent certains problèmes. Nous présentons dans cette section les traitements réalisés.

2.1.1 Traitement de la base "Productions"

Après avoir importé la base, nous affichons les statistiques descriptives des variables [Figure 1] :

Variable	N	Mean	Std Dev	Minimum	Maximum
exposition	145325	0.5304059	0.3751834	0	2.0027397
puissance_fiscale	130298	7.7186219	2.4912825	1.0000000	340.0000000
Exercice	145325	2009.14	1.4207400	2007.00	2011.00
Date_obtention_du_permis_	135263	19910095.19	451517.51	0	20111220.00
Date_de_naissance	135263	19619445.70	602239.36	0	20090502.00
Date_du_premier_effet_	145325	20071853.64	97539.24	0	20111231.00
Date_de_Mise_en_Circulation	130298	19964229.23	212401.33	0	20950106.00
NUMERO_POLICE	145325	810105.37	491991.29	3516.00	1632159.00

FIGURE 1 – Statistiques descriptives de la base "Productions"

En analysant ce tableau, nous comprenons que notre base contient plusieurs valeurs manquantes car le nombre d'observations (N) diffère d'une variable à l'autre. Elle contient également des valeurs aberrantes au niveau de l'exposition *(La plupart des contrats sont annuels donc il est judicieux de se ramener à une exposition inférieure ou égale à 1)* et au niveau des dates *(Toutes ont des 0 comme valeur minimale)*. Nous remarquons aussi que le format des dates n'est pas révélateur *(Numérique)* et qu'il serait préférable de les convertir au format DATE.
Pour y remédier, nous faisons ce qui suit :

1. Suppression des valeurs manquantes, de la forme « . »
2. Supression des doublons
3. Conversion des variables de dates au format DATE
4. Suppression des valeurs aberrantes des dates, de la forme « 0 » ou « . »

A ce niveau, nous avons jugé qu'il était préférable de raisonner en termes d'âge plutôt qu'en terme de date.

5. Ajout des variables d'âge

Date obtention du permis.	→	dodpdate
Date de naissance	→	ddndate
Date du premier effet.	→	ddpedate
Date de Mise en Circulation	→	ddmecdate

6. Suppression des variables aberrantes d'âge, *ex : ageconducteur<18, ageconducteur>70 ou encore agevehicule>50*
7. Suppression des variables de dates non significatives

Etant donné que notre objectif est de modéliser le montant moyen annuel des sinistres, nous avons réalisé un :

8. Regroupement des observations par numero de police et par exercice, tout en ajoutant une nouvelle variable : SommeExpo, qui représente la somme des expositions d'un assuré ayant contracté une assurance pour une courte durée, l'ayant rompue, et la renouvelant au cours du même exercice.

2.1.2 Traitement de la base "Sinistres"

Après avoir importé la base, nous affichons les statistiques descriptives des variables [Figure 2] :

Variable	N	Mean	Std Dev	Minimum	Maximum
numero_sinistr	65358	1.6230527E14	7.4078375E13	2.01E13	2.011E14
Charge	65358	21684.37	47557.70	-50000.00	2400000.00
exercice	65358	2009.17	1.3959522	2007.00	2011.00
Police	65358	816001.92	541326.09	3407.00	1632150.00

FIGURE 2 – Statistiques descriptives de la base "Sinistres"

En analysant ce tableau, nous comprenons que notre base ne contient pas de valeurs manquantes étant donné que le nombre d'observations est le même pour toutes las variables. Néanmoins, nous remarquons que la variable charge a certaines valeurs négatives - aberrantes en d'autres termes. Le traitement opéré au niveau de cette base ne se traduit que par :

1. Suppression des doublons
2. Suppression des valeurs négatives de la variable charge

Etant donné que nous nous intéressons à la charge moyenne annuelle de chaque assuré, nous faisons en sortes de l'expliciter au niveau de la base de donnée :

3. Ajout de la variable : nbSinistres, qui représente le total des sinistres qu'un assuré a eu au cours d'un exercice donné
4. Ajout de la variable : sommeCharge, qui représente la somme des charges versées à chaque assuré au cours d'un exercice donné

2.1.3 Fusion des deux bases

Afin de construire notre modèle, nous avons besoin de combiner la variable à expliquer et les variables explicatives dans une seule base de données. Pour ce faire, nous réalisons ce qui suit :

1. Fusion *(Union mathématique)* des deux bases de données "Productions" et "Sinistres" selon le numéro de police et l'exercice

Puisque la majorité des assurés incrits à la base "Production" n'ont pas eu de sinistres, il est naturel de se retrouver avec un nombre important de valeurs manquantes, donc l'étape suivante consiste au :

2. Remplacement des valeurs manquantes de la variable charge par des « 0 »
3. Ajout d'une nouvelle variable : coutMoyen, qui représente la somme de la charge divisée par le nombre de sinistres pour chaque assuré durant tout exercice.

Notons que c'est cette variable que nous cherchons précisément à expliquer en fonction des autres. Du fait que 88% des observations ont une charge nulle, l'entrainement du modèle serait erroné et aurait tendance à prédire des charges nulles. D'où l'importance de :

4. Suppression de toutes les observations dont la charge est nulle

Pour nous assurer des résultats de cette union, nous affichons encore une fois les statistiques descriptives des variables [Figure 3] :

Variable	N	Mean	Std Dev	Minimum	Maximum
NUMERO_POLICE	4576	815357.48	535365.23	3516.00	1630324.00
Exercice	4576	2009.28	1.3297381	2007.00	2011.00
puissance_fiscale	4576	7.8133741	2.2666323	4.0000000	32.0000000
ageconducteur	4576	44.9219463	11.6325397	18.9712603	69.9986575
agepermis	4576	18.2912249	11.2159613	0.1381096	51.4987671
agevehicule	4576	9.3627396	8.1727680	0.0301370	38.7240548
ancienneteAssurance	4576	2.7923141	3.1441958	0.0164384	40.7404932
SommeExpo	4576	0.7075669	0.3349415	0	1.0000000
nbSinistres	4576	1.0493881	0.2442064	1.0000000	6.0000000
sommeCharge	4576	27863.90	54029.00	0.7800000	987637.53
CoutMoyen	4576	26867.74	53061.22	0.7800000	987637.53

FIGURE 3 – Statistiques descriptives de la base finale

2.2 Analyse exploratoire

Dans ce qui suit, nous essayons de mettre en emphase les facteurs qui favorisent la sinistralité. A travers un regard critique sur la répartition et l'évolution de la variable à expliquer « coutMoyen » en fonction des variables explicatives, nous déterminons les tendances prépondérantes qui nous aiderons par la suite à tester la validité de notre modèle.

- L'évolution de coutMoyen par exercice

Pour étudier cette évolution, nous affichons ce qui suit [Figure 4] et [Figure 5] :

Analysis Variable : CoutMoyen						
Exercice	N Obs	N	Mean	Std Dev	Minimum	Maximum
2007	520	520	26025.81	43020.06	18.0000000	490447.08
2008	905	905	30347.01	53736.06	14.0000000	590150.00
2009	1057	1057	28271.39	60150.64	18.0000000	987637.53
2010	973	973	26400.51	47012.65	0.7800000	470900.00
2011	1121	1121	23531.46	54403.28	4.0000000	839094.15

FIGURE 4 – Statistiques descriptives de coutMoyen par exercice

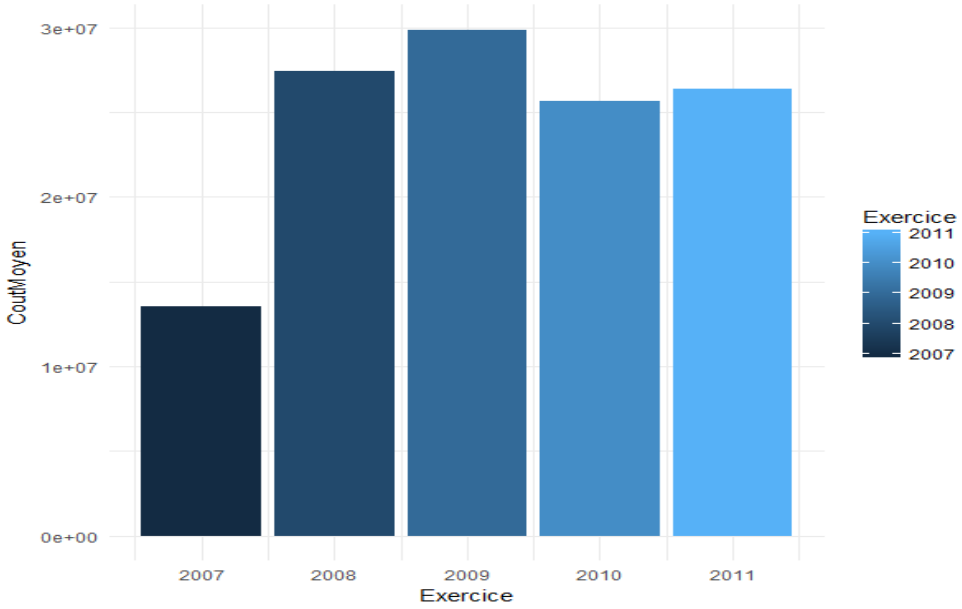

FIGURE 5 – Evolution de la somme de coutMoyen par exercice

Bien que le total des charges moyennes annuelles atteind son maximum au cours de l'année 2009, c'est en 2011 que nous trouvons le plus de sinistres. Remarquons toutefois la moyenne de la variable à expliquer varie peu d'année en année et qu'elle est maximale en 2008.

- L'évolution de coutMoyen par tranche d'âge

Pour déterminer les sous populations les plus critiques, nous divisons notre base de données selon les tranches d'âge que nous jugeons parlantes :

$$
\begin{aligned}
18 \leq \text{AgeConducteur} < 25 &\rightarrow \text{Tranche A} \\
25 \leq \text{AgeConducteur} < 35 &\rightarrow \text{Tranche B} \\
35 \leq \text{AgeConducteur} < 45 &\rightarrow \text{Tranche C} \\
45 \leq \text{AgeConducteur} < 55 &\rightarrow \text{Tranche D} \\
55 \leq \text{AgeConducteur} < 65 &\rightarrow \text{Tranche E} \\
65 \leq \text{AgeConducteur} < 70 &\rightarrow \text{Tranche F}
\end{aligned}
$$

Pour étudier cette évolution, nous affichons ce qui suit [Figure 6] [Figure 7] :

Analysis Variable : CoutMoyen						
trancheAge	N Obs	N	Mean	Std Dev	Minimum	Maximum
A	125	125	29532.96	44512.70	604.0000000	320397.48
B	961	961	27041.28	49277.19	63.9500000	724304.00
C	1274	1274	29189.57	56504.54	0.7800000	590150.00
D	1221	1221	24790.06	41666.52	18.0000000	460008.00
E	813	813	25741.97	63067.44	3.7200000	987637.53
F	182	182	26835.83	69641.25	88.4800000	839094.15

FIGURE 6 – Statistiques descriptives du coutMoyen par tranche d'âge

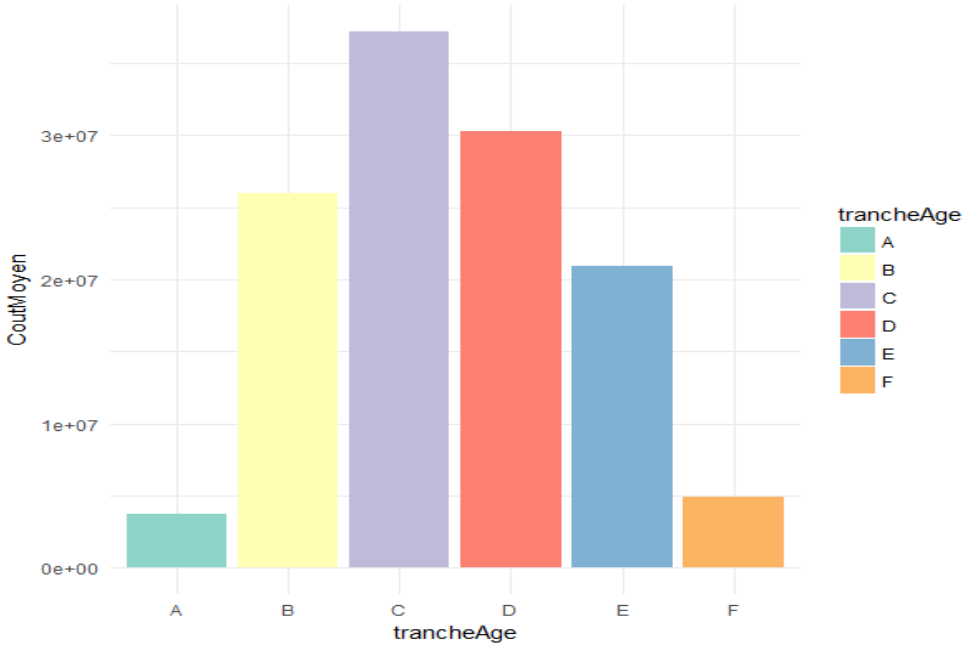

FIGURE 7 – Evolution de la somme du coutMoyen par tranche d'âge

A en juger par les données du tableau, nous déduisons que la sous population des assurés dont l'âge varie entre 35 et 55 ans est la plus large. Toutefois, nous remarquons que les sous populations des plus jeunes *(Entre 18 et 25 ans)* et des plus agés *(Entre 65 et 70 ans)* sont celles qui présentent le plus de risque de sinistralité car leurs moyennes sont relativement importantes même si elles sont minoritaires.

- L'évolution de coutMoyen par sexe

Pour étudier cette évolution, nous affichons ce qui suit [Figure 8] et [Figure 9] :

Analysis Variable : CoutMoyen						
SEXE	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	869	869	19745.72	37094.22	0.7800000	335542.11
M	3706	3706	28543.38	56030.67	1.0000000	987637.53

FIGURE 8 – Statistiques descriptives de coutMoyen par sexe

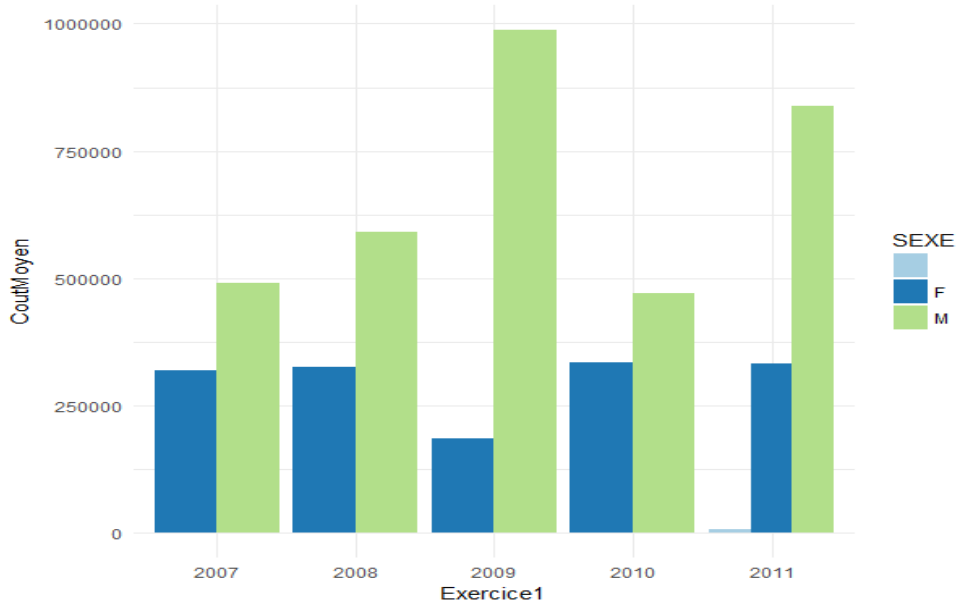

FIGURE 9 – Evolution de la somme de coutMoyen par sexe

Les deux figures précédentes nous permettent de déduire qu'il y a plus d'assurés hommes que de femmes. A en juger par la répartition du total des charges moyennes annuelles, nous remarquons que les hommes bénéficient de plus de remboursement, chose qui est normale étant donné qu'ils sont majoriatires. Cela n'empêche, la moyenne des hommes est largement supérieure à celle des femmes.

- L'évolution de coutMoyen par combustion

Pour étudier cette évolution, nous affichons ce qui suit [Figure 10] et [Figure 11] :

Analysis Variable : CoutMoyen						
Combustion	N Obs	N	Mean	Std Dev	Minimum	Maximum
E	1292	1292	25007.74	48977.61	4.0000000	987637.53
G	3284	3284	27599.51	54573.90	0.7800000	901460.26

FIGURE 10 – Statistiques descriptives de coutMoyen par combustion

La sous population des véhicules Gasoil ayant commis un sinistre représente plus que le double de celles dont la combustion se fait par essence, chose qui est normale étant donné qu'elles sont plus nombreuses en circulation. Malgré cela, leurs variables à expliquer ont à peu près la même moyenne. Néanmoins, nous remarquons qu'elles ont des tendances oposées et que la charge des véhicules Essence est très importante compte tenu de leur proportion. De ce fait, nous déduisons que les assurés *"Essence"* présentent plus de risque de sinistralité.

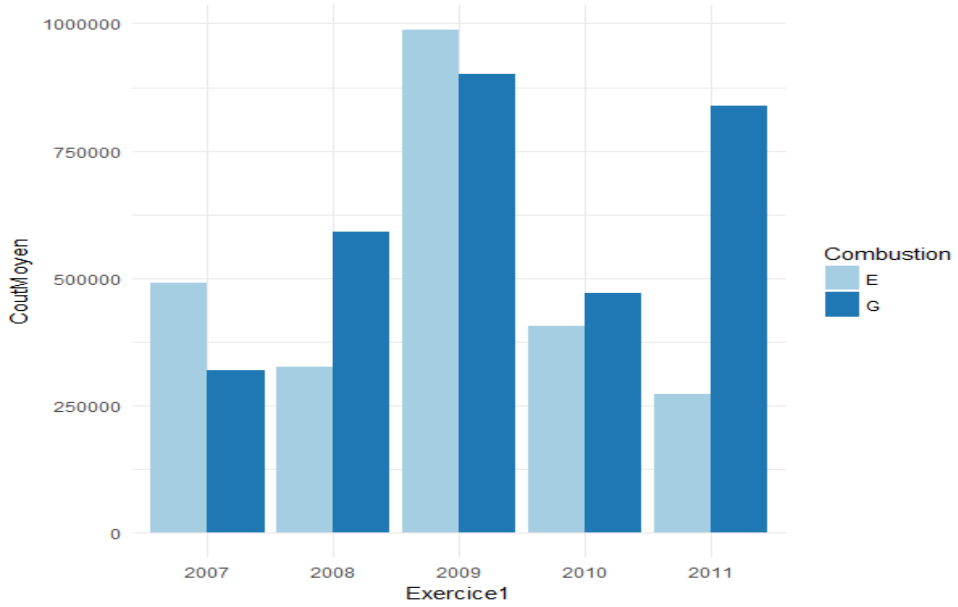

FIGURE 11 – Evolution de la somme de coutMoyen par combustion

- L'évolution de coutMoyen par puissance fiscale

Etant donné que les puissances fiscales de notre population varient de 4 à 32, il est difficile d'afficher le tableau qui regroupe leurs statistiques descriptives. En l'analysant, nous déduisons ce qui suit :

— Les véhicules qui sont les plus nombreuses en circulation et qui, par conséquent, commettent le plus de sinistralité ont une puissance fiscale qui varie entre 6 et 8

— Les véhicules qui ont une très grande puissance fiscale *(Entre 20 et 32)* sont très minoritaires *(4 véhicules au maximum)*. Toutefois, elles ont des moyennes supérieures aux autres.

Nous en concluons que les véhicules qui ont une puissance fiscale importante ne présentent pas un plus grand risque de sinistralité mais leurs dédommagements sont extrêmement

coûteux pour les sociétés d'assurance.

- L'évolution de coutMoyen par tranche d'âge du véhicule
Pour déterminer les sous populations les plus critiques, nous divisons notre base de données selon les tranches d'âge du véhicule que nous jugeons parlantes :

$$\begin{aligned} 0 \leq \text{AgeVehicule} < 9 &\rightarrow \text{Tranche A} \\ 9 \leq \text{AgeVehicule} < 18 &\rightarrow \text{Tranche B} \\ 18 \leq \text{AgeVehicule} < 27 &\rightarrow \text{Tranche C} \\ 27 \leq \text{AgeVehicule} < 39 &\rightarrow \text{Tranche D} \end{aligned}$$

Pour étudier cette évolution, nous affichons ce qui suit [Figure 12][Figure 13] :

Analysis Variable : CoutMoyen						
VAge	N Obs	N	Mean	Std Dev	Minimum	Maximum
A	2766	2766	21431.97	44634.29	0.7800000	901460.26
B	984	984	30653.59	52270.53	24.0000000	590150.00
C	679	679	39699.53	67831.31	14.0000000	724304.00
D	147	147	44536.35	94852.88	138.6000000	987637.53

FIGURE 12 – Statistiques descriptives du coutMoyen par tranche d'âge du véhicule

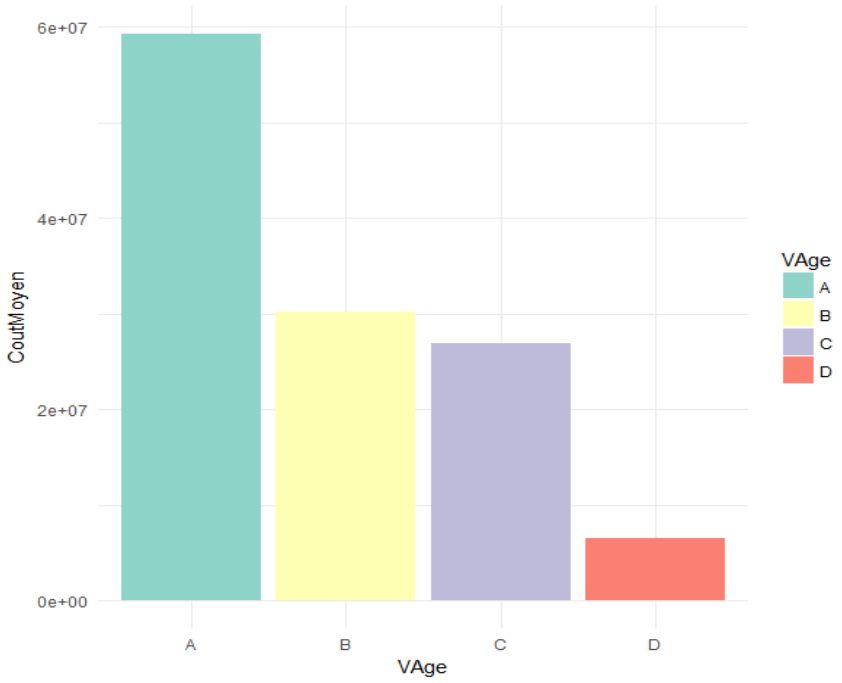

FIGURE 13 – Evolution de la somme du coutMoyen par tranche d'âge du véhicule

En analysant les données du tableau, nous déduisons que les nouveaux véhicules *(L'âge du véhicule varie entre 0 et 9 ans)* sont ceux qui commettent le plus de sinistres car ils

contractent le plus d'assurances. Nous remarquons aussi que plus l'âge des véhicules augmente plus leur nombre diminue. A l'opposé, plus le véhicule vieillit plus sa moyenne de charge moyenne annuelle augmente. De cela, nous déduisons que le risque de sinistralité augmente avec l'augmentation de l'âge du véhicule.

- L'évolution de coutMoyen par ancienneté d'assurance

Pour déterminer les sous populations les plus critiques, nous divisons notre base de données selon les tranches d'ancienneté d'assurance que nous jugeons parlantes :

$$\begin{aligned} 0 \leq \text{ancienneteAssurance} < 3 &\quad \rightarrow \quad \text{Tranche A} \\ 3 \leq \text{ancienneteAssurance} < 6 &\quad \rightarrow \quad \text{Tranche B} \\ 6 \leq \text{ancienneteAssurance} < 9 &\quad \rightarrow \quad \text{Tranche C} \\ 9 \leq \text{ancienneteAssurance} < 41 &\quad \rightarrow \quad \text{Tranche D} \end{aligned}$$

Pour étudier cette évolution, nous affichons ce qui suit [Figure 14][Figure 15] :

Analysis Variable : CoutMoyen						
AncAss	N Obs	N	Mean	Std Dev	Minimum	Maximum
A	3234	3234	27983.07	54608.39	0.7800000	987637.53
B	858	858	24040.19	43609.15	4.0000000	435789.75
C	284	284	25363.97	63651.35	130.7200000	901460.26
D	200	200	23098.43	47504.70	4.0000000	460008.00

FIGURE 14 – Statistiques descriptives du coutMoyen par tranche d'ancienneté assurance

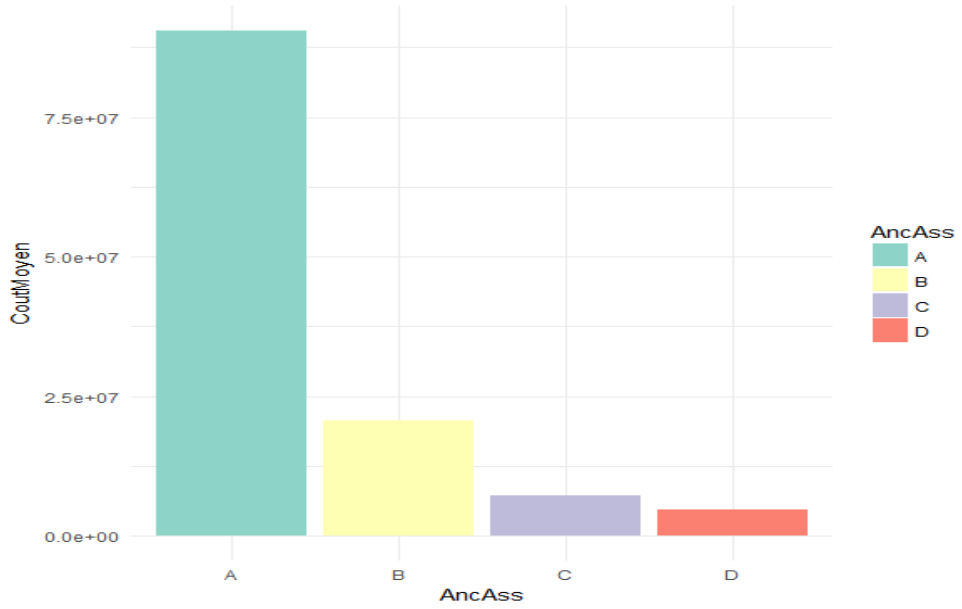

FIGURE 15 – Evolution de la somme du coutMoyen par tranche d'ancienneté assurance

Nous remarquons que notre base de données contient plus de nouveaux contractants que d'anciens. Par conséquent, ces derniers ont d'une part, plus tendance à commetre des sinistres et d'autre part des moyennes élevées par rapport aux anciens clients de la société en question. Nous déduisons que l'expérience du conducteur joue un rôle important quant à la détermination du risque de la sinistralité.

- Analyse de la corrélation

L'analyse de la corrélation nous permet de déterminer la relation entre les variables de notre base de données, pour caractériser par la suite la forme de cette relation (linéaire ou non linéaire, positive ou négative). Pour ce faire, nous commençons par une analyse graphique des nuages de points entre toutes les variables. Cette analyse est présentée par la [Figure 16] .

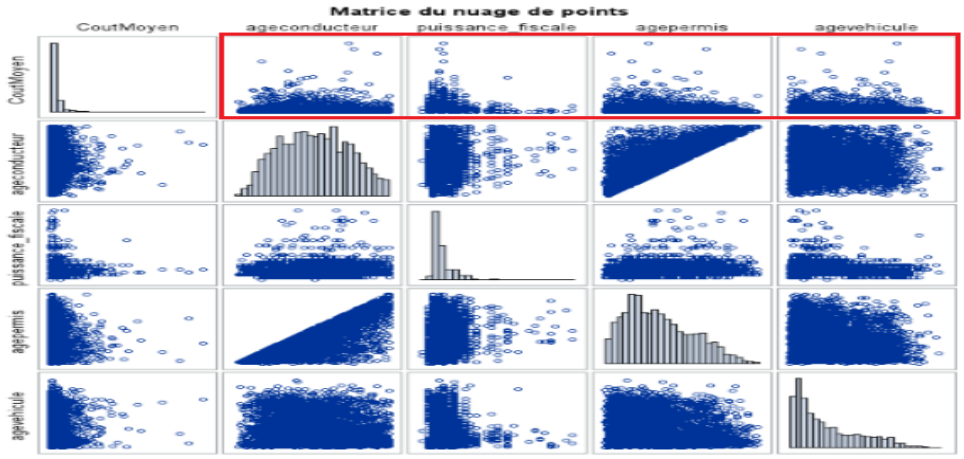

FIGURE 16 – Matrice du nuage de points

De cette matrice, nous pouvons remarquer une faible corrélation entre les variables explicatives et la variable à expliquer, chose que nous essayons de valider à l'aide de la matrice de corrélation de Pearson illustrée par le [tableau 17] .

Coefficients de corrélation de Pearson, N = 4576
Proba > |r| sous H0: Rho=0

	CoutMoyen	ageconducteur	puissance_fiscale	agepermis	agevehicule	ancienneteAssurance
CoutMoyen	1.00000	-0.02956 0.0456	0.00875 0.5540	-0.06802 <.0001	0.14363 <.0001	-0.04445 0.0026
ageconducteur	-0.02956 0.0456	1.00000	0.10797 <.0001	0.77237 <.0001	-0.05976 <.0001	0.26288 <.0001
puissance_fiscale	0.00875 0.5540	0.10797 <.0001	1.00000	0.13918 <.0001	0.00393 0.7903	0.05440 0.0002
agepermis	-0.06802 <.0001	0.77237 <.0001	0.13918 <.0001	1.00000	-0.15266 <.0001	0.28726 <.0001
agevehicule	0.14363 <.0001	-0.05976 <.0001	0.00393 0.7903	-0.15266 <.0001	1.00000	0.00677 0.6471
ancienneteAssurance	-0.04445 0.0026	0.26288 <.0001	0.05440 0.0002	0.28726 <.0001	0.00677 0.6471	1.00000

FIGURE 17 – Coefficients de corrélation de Pearson

Ces résultats nous permettent de rejeter l'hypothèse : *"La dépendance entre la variable à expliquer et les variables explicatives est linéaire"* . De plus, nous déduisons que ces variables explicatives ont un faible pouvoir prédictif même pour le cas non linéaire (de la matrice du nuage de points).

2.3 Les lois

L'analyse exploratoire nous permet de dégager non seulement les différentes relations qui lient les variables explicatives à la variable à expliquer, mais aussi leurs impacts sur l'évolution de celle ci . En revanche, cet analyse nous ne fournit pas d'information supplémentaire sur le comportement de la variable aléatoire "CoutMoyen". Pour y arriver nous utilisons les lois de probabilités.
La variable aléatoire "CoutMoyen" peut prendre toutes les valeurs réelles d'un intervalle I inclus dans R^+. Ainsi elle a une distribution continue sur un intervalle semi-fini, comme l'illustre la figure [18] . Ce qui nous permet de conclure, que cette variable ne peut pas suivre la loi Normale (qui est souvent utilisée). La distribution du "CoutMoyen" nous donne une idée sur des lois, autre que la loi Normale, qui peuvent l'approcher. La loi LogNormale, Gamma, et Exponentielle, ont plus de chance de pouvoir modéliser la distribution de la variable à expliquer.

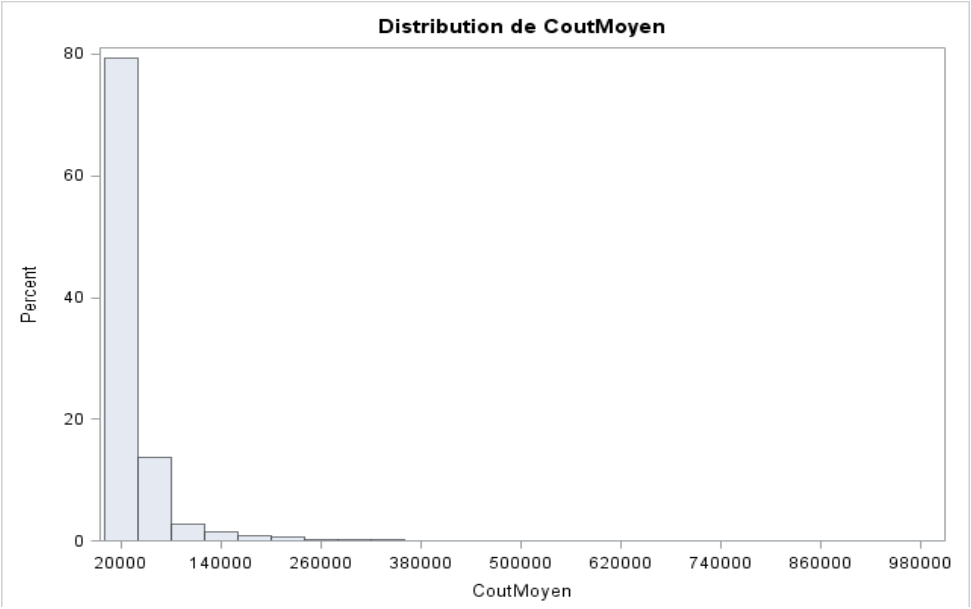

FIGURE 18 – Distribution de la variable "CoutMoyen"

2.3.1 La loi log normale

La loi log normale ou loi de Galton est une loi qui permet la modélisation de données à peu près symétriques ou asymétriques vers la droite. On dit qu'une variable aléatoire X suit une loi log normale quand son logarithme suit une loi normale. La densité de probabilité de cette loi s'écrit de la manière suivante :

$$f(x) = \frac{1}{x}.\frac{1}{\beta\sqrt{2\pi}}.e^{-\frac{1}{2}(\frac{ln(x)-\alpha}{\beta})^2}, x > 0 \quad (1)$$

avec

$$\alpha = E(ln(X)) \text{ (espérance type de } ln(X))$$

et

$$\beta = \sigma_{ln(X)} \text{ (écart type de } ln(X))$$

d'où les moments de la variable aléatoire X sont :

$$la\ moyenne\ :\ \mu_x = E(X) = e^{(\mu_{ln(x)} + \frac{\sigma^2_{ln(x)}}{2})}$$

$$la\ variance\ :\ \sigma^2_x = Var(X) = e^{2.(\mu_{ln(x)} + \sigma^2_{ln(x)})}.(\frac{e^{\sigma^2_{ln(x)}} - 1}{e^{\sigma^2_{ln(x)}}})$$

2.3.2 La loi gamma

On dit qu'une variable aléatoire réelle suit une loi gamma de paramètres λ et a, si et seulement si sa densité de probabilité est donnée par la formule suivante :

$$f(x) = \frac{\lambda^a}{\Gamma(a)}.x^{a-1}.e^{-\lambda x}, x \geq 0 \tag{2}$$

d'où les moments de la variable aléatoire réelle X sont :

$$la\ moyenne\ :\ \mu_x = E(X) = \frac{a}{\lambda}$$

$$la\ variance\ :\ \sigma^2_x = Var(X) = \frac{a}{\lambda^2}$$

2.3.3 La loi exponentielle

On dit qu'une variable aléatoire réelle suit une loi exponentielle de paramètre $\lambda > 0$ quand sa densité de probabilité s'écrit sous la forme suivante :

$$f(x) = \lambda e^{-\lambda x}, x \geq 0 \tag{3}$$

d'où les moments de la variable aléatoire X sont :

$$la\ moyenne\ :\ \mu_x = E(X) = \frac{1}{\lambda}$$

$$la\ variance\ :\ \sigma^2_x = Var(X) = \frac{1}{\lambda^2}$$

2.4 Tests d'adéquations des lois

Supposons avoir un ensemble d'observations qui sont des réalisations indépendantes d'une même variable aléatoire. Les tests statistiques nous permettent d'adopter ou rejeter des hypothèses de modélisation probabiliste de ces réalisations. Nous utiliserons plusieurs tests afin de valider ou réfuter, au seuil de risque choisi, que ces observations sont tirées dans une distribution LogNormale, Gamma ou Exponentielle. Nous présenterons les tests non paramétriques classiques à savoir Kolmogorov-Smirnov, Cramer-Von-Mises et Anderson Darling. Ces tests reposent sur la comparaison de la fonction de répartition théorique $F_0(x)$ à la fonction de répartition empirique $\hat{F}$, sous l'hypothèse H_0.

Soit $(X_1, \ldots, X_n)$ un échantillon de loi P (P est soit LogNormale, Gamma ou Exponentielle). L'hypothèse H_0 : La loi P a pour fonction de répartition F_0. Si la fonction de répartition empirique $\hat{F}$ de l'échantillon est proche de la fonction de répartition théorique F_0, dans ce cas l'hypothèse H_0 est correcte.

2.4.1 Test de "Kolmogorov-Smirnov"

L'adéquation de la fonction de répartition empirique à celle théorique se mesure par la distance de Kolmogorov Smirnov. Afin de la calculer, nous évaluons la différence entre les deux aux points X_i comme suit :

$$D_{KS}(F_0, \hat{F}) = \sup_x |\hat{F}(x) - F_0(x)| \tag{4}$$

L'écart absolu entre F_0 et $\hat{F}$ ne dépend pas de F_0, puisque les images de X_i par F_0 sont des variables aléatoires de loi U(0,1). Il s'avère très compliqué d'établir une expression explicite et simple de D_{KS} d'où l'intérêt d'utiliser le résultat asymptotique suivant (résulte du théorème de Kolmogorov) :

- Proposition :

Sous l'hypothèse H_0 , on a, pour t ≥ 0 :

$$\lim_{n\to\infty} P_{H_0}[\sqrt{n} D_{KS}(F_0, \hat{F}) \leq t] = 1 - 2\sum_{k=1}^{\infty} (-1)^{K+1} exp(-2K^2t^2) \tag{5}$$

La série converge très rapidement.

Si $\sqrt{n} D_{KS}(F_0, \hat{F})$ tend vers ∞ quand n tend vers ∞, alors on rejette l'hypothèse H_0.

2.4.2 Test de "Cramer-Von-Mises" et "d'Anderson-Darling"

Le test de Cramer-Von-Mises et celui d'Anderson Darling, reposent aussi sur l'examination de la distance entre la fonction de répartition empirique et théorique sous l'hypothèse H_0. La différence entre ces deux tests et celui de Kolmogorov-Smirnov est au niveau de la distance utilisée et la manière de la calculer. Pour ces deux tests nous utilisons la distance quadratique $(F_0 - \hat{F})^2$, ainsi que nous ne nous intéressons pas à la distance maximale entre

les deux fonctions de répartition mais nous tenons en compte l'ensemble des observations, d'où l'expression suivante :

$$Q = n \int_{-\infty}^{+\infty} (F_0 - \hat{F})^2 \psi(x) dF_0(x) \tag{6}$$

où $\psi(x)$ est la fonction de pondération qui diffère en fonction du test utilisé (soit de Cramer-Von-Mises ou d'Anderson Darling). Contrairement au test de Kolmogorov Smirnov, la distance calculée dépend de F_0.

- Test de Cramer-Von-Mises

Dans ce cas on a $\psi(x) = 1$ et la statistique du test est :

$$W^2 = \sum_{i=1}^{n} (\frac{2i-1}{2n} - F(x_i))^2 + \frac{1}{12n} \tag{7}$$

avec x_i la i-ème réalisation de la variable aléatoire X, si ces réalisations sont ordonnées en ordre croissant.
Si W^2 est supérieure à sa valeur critique on rejette l'hypothèse H_0.

- Test d'Anderson-Darling

Dans ce cas la fonction de pondération est $\psi(x) = [F_0(x)(1-F_0(x))]^{-1}$ et la statistique est :

$$A^2 = n - \frac{1}{n} \sum_{i=1}^{n} [(2i-1)\log(F_0(x_i)) + (2n+1-2i)\log(1-F_0(x_i))] \tag{8}$$

L'hypothèse H_0 est rejetée si la valeur de W^2 dans le cas d'un test de Cramer von Mises (respectivement A^2 dans le cas d'un test d'Anderson Darling) est supérieure à sa valeur critique. Afin de déterminer cette valeur, il faut se reporter à des tables qui dépendent de la distribution théorique considérée. Vue la difficulté de déterminer cette valeur à partir des tables, nous utilisons les résultats de SAS. Ce dernier fournit la probabilité que W^2 (respectivement A^2) dépasse sa valeur critique. Si cette probabilité (P-value) est supérieure au facteur de risque α, nous rejetons H_0.

2.4.3 Tests empiriques

Nous utilisons les tests d'adéquation cités précédemment afin de déterminer la loi que suit la variable "coutMoyen". La mise en oeuvre de ces tests s'effectue en utilisant l'instruction PROC UNIVARIATE. Celle ci nous permet de générer les tableaux illustrés dans [Figure 19] et [Figure 20] réspectivement. En effet, pour une raison que nous n'avons pas pu déterminer, le programme n'a pas généré le tableau des tests correspondants à la loi Gamma

Goodness-of-Fit Tests for Lognormal Distribution				
Test	Statistique		P-value	
Kolmogorov-Smirnov	D	0.147708	Pr > D	<0.010
Cramer-von Mises	W-Sq	24.290092	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	114.000958	Pr > A-Sq	<0.005

FIGURE 19 – Résultat des tests d'adéquation pour la loi Log Normale

Goodness-of-Fit Tests for Exponential Distribution				
Test	Statistique		P-value	
Kolmogorov-Smirnov	D	0.302612	Pr > D	<0.001
Cramer-von Mises	W-Sq	106.834995	Pr > W-Sq	<0.001
Anderson-Darling	A-Sq	514.745430	Pr > A-Sq	<0.001

FIGURE 20 – Résultat des tests d'adéquation pour la loi Exponentielle

Pour déterminer si les données suivent une loi LogNormal, Gamma ou Exponentielle, on compare la P-value au seuil de signifiance α que nous avons fixé à 5% . Nous constatons que pour les deux tests d'adéquation des lois Exponentiel et LogNormal, toutes les P-values affichées sont faibles si nous les comparons à la valeur $\alpha = 5\%$. Par conséquent on rejette l'hypothèse H_0 pour les trois tests. Autrement dit, la variable "CoutMoyen" ne suit ni la loi Log Normal ni la loi Exponentielle ni la loi Gamma.

Il existe d'autres manières qui permettent de savoir si un ensemble d'observations est compatible avec l'hypothèse H_0. Parmis ces méthodes figurent les histogrammes et le diagramme QQ-Plot.

- Les histogrammes

Les histogrammes permettent d'examiner si une variable suit une loi, si la distribution de la variable est compatible avec la courbe de chacune des lois déterminées au préalable. On peut générer cette distribution comme le montre la [Figure 21]. Ce type de visualisation nous a permis de conclure qu'aucune des lois citées précédemment modélise le mieux la distribution de la variable coutMoyen.

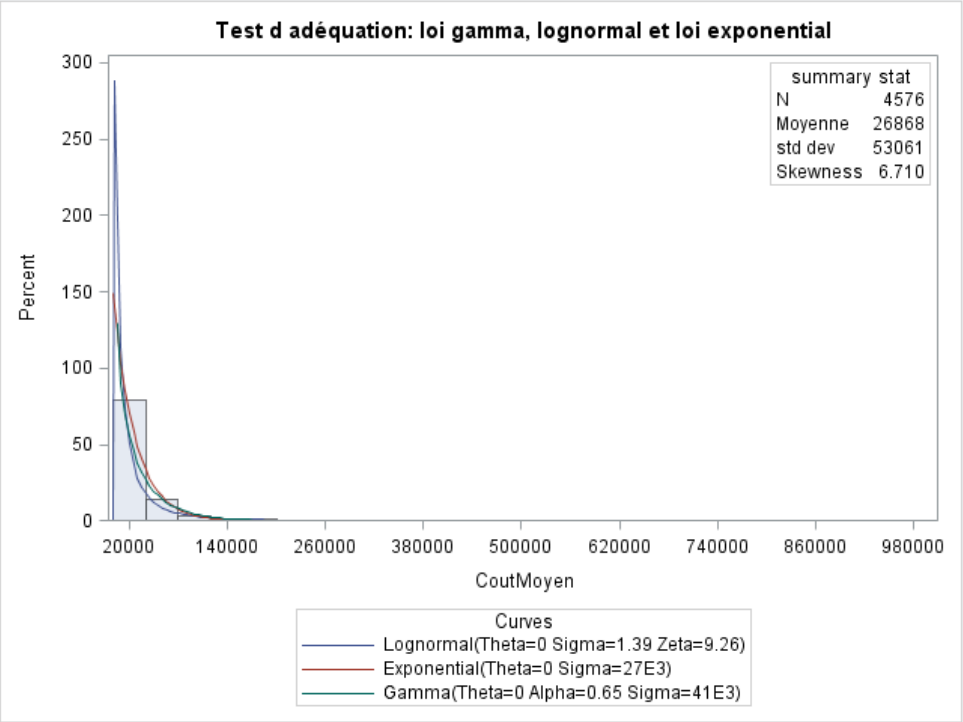

FIGURE 21 – Tests d'adéquation : Histogramme

- QQ-Plot

Le diagramme QQ-Plot est une aide visuelle permettant de savoir si H_0 peut être raisonnablement maintenue tout en se basant sur la comparaison observés aux quantiles théoriques. Les Figures [22] représentent respectivement le diagramme QQ-Plot de la loi Gamma, logNormal et Exponentielle. On remarque que le nuage de points pour toutes les lois ne s'aligne pas sur la première bissectrice. Néanmoins, on constate que la loi lognormal a moins dévié de la première bissectrice que les deux autres lois.

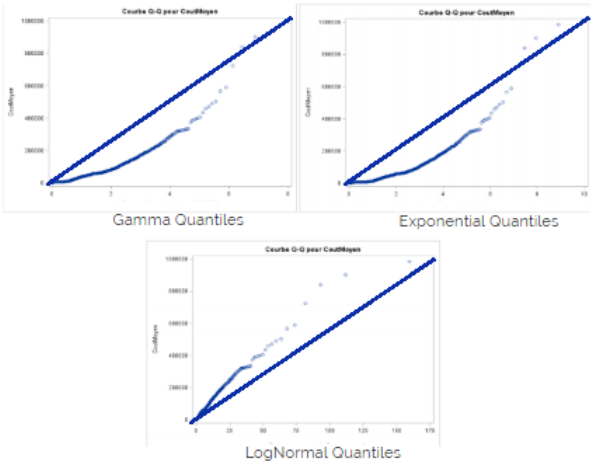

FIGURE 22 – Diagramme QQ-Plot

Les tests d'adéquation effectués soit par comparaison des fonctions de répartition soit par des visualisations nous conduisent à conclure qu'il est difficile de trouver des paramètres qui permettent d'approcher le coût moyen des sinistres par les différentes

lois citées précédemment. Ce qui est susceptible puisque la distribution de cette variable dépend de plusieurs autres variables.

3 Classification des données

3.1 Méthodes de classification et choix

La classification permet la segmentation des observations. On distingue la classification supervisée et la classification non supervisée.

3.1.1 Classification non supervisée

Cette méthode vise le regroupement ou la classification des individus qui ont des caractéristiques semblables. En effet, dans le cas de la classification non supervisée, nous ne disposons pas de mesure de la variable dépendante.

AgeConducteur	CoûtMoyen
20	51 000
21	30 000
22	1 000
23	1 200
60	630
62	800
63	40 000
64	20 680
63,5	30 000

TABLE 1 – Echantillon "BaseFus"

Supposons avoir un ensemble d'observation tableau [1] issue de notre base de donnée, dont nous avons garder les deux variables “CoutMoyen” et “Ageconducteur”.L'application de la classification non supervisée, s'appuiera sur la variable ageconducteur pour créer des groupes homogènes. Ainsi cette classification aboutira au graphe [23]. De ce fait, quelque soit la valeur du coût Moyen, toute observation ayant moins de 40 ans sera classée dans le premier groupe, tandis que toute observation supérieure à 40 ans sera classée dans le deuxième groupe. Ce qui n'a pas d'interêt puisqu'on cherche à créer des groupes homogènes par rapport au coût moyen. Pour y remédier nous utilisons la classification supervisée.

3.1.2 Classification supervisée

Elle permet de décrire la relation entre les variables indépendants et la variable dépendante. Pour y arriver nous disposons de plusieurs outils dont figure les arbres de décisions.

3.1.3 Arbre de décision

Les arbres de décisions sont le résultat des algorithmes qui partitionnent les observations sous la forme d'un arbre de décision afin de trouver des classes d'observations ayant

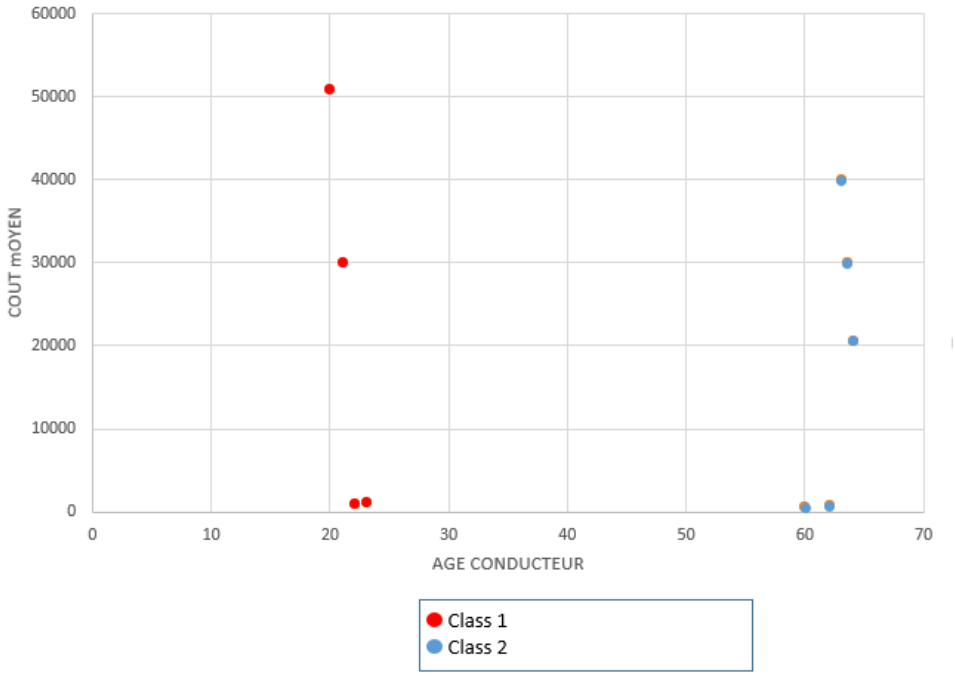

FIGURE 23 – Classification non supervisée appliquée au Coût Moyen

des valeurs de la variable à expliquer les plus homogènes possible.

- Principe de construction

Sur un échantillon de n observations, les données se composent d'une variable à expliquer qui peut être quantitative ou qualitative, et d'un ensemble de variables explicatives. La construction d'un arbre binaire consiste à choisir une valeur seuil d'une variable explicative qui permet de partitionner les observations en deux classes tout en minimisant la variance de la variable à expliquer sur chacun des sous-ensembles. Donc, à chaque nœud correspond un sous-ensemble à caractéristiques similaires. A la racine du nœud initial correspond la totalité des observations, ensuite, la procédure est itéré sur chacun des sous-ensembles. De plus, il faut déterminer un critère d'arrêt permettant de décider qu'un nœud est terminal. En effet, les arbres construits doivent être le plus profond possible, cela a pour effet de mieux prédire les valeurs de la variable à expliquer, néanmoins, il peut engendrer des problèmes de sur apprentissage et de sensibilité au bruit, donc une défaillance pour la prédiction des nouvelles observations. Pour y remédier, il faut supprimer les classes peu représentatives pour garder de bonnes performances prédictives sur les nouvelles observations.

La classification des nouvelles observations consiste à suivre l'arbre en commençant par la racine, ensuite, on effectue les différents tests à chaque nœud, jusqu'à atteindre une classe terminale. L'arbre final affiché résulte d'algorithme de recherche d'arbre optimal. Cet algorithme commence par construire un arbre maximal. Ensuite, il pénalise les sous arbres construits en se basant sur le critère de déviance ou de taux des observations mal-classées. Ce qui permet d'ordonner les sous arbre. Finalement, l'algorithme choisit le sous arbre optimal.

Supposons que l'application de la classification supervisée,à l'échantillon représenté par le tableau [1], aboutira à la classification décrite par le graphe [24] . Nous constatons

que cet arbre a créé des groupes homogènes par rapport à la variable coût moyen. Ce qui permet d'avoir des conclusions concernant la variable à expliquer, par segmentation d'age. A titre d'exemple, les observations ayant moins de 21,5 ans seront classées dans la classe 2, dont la moyenne du coût est 40500. Tandis que les observations ayant plus que 21,5 ans et moins de 40 ans seront affecter à la classe 1.

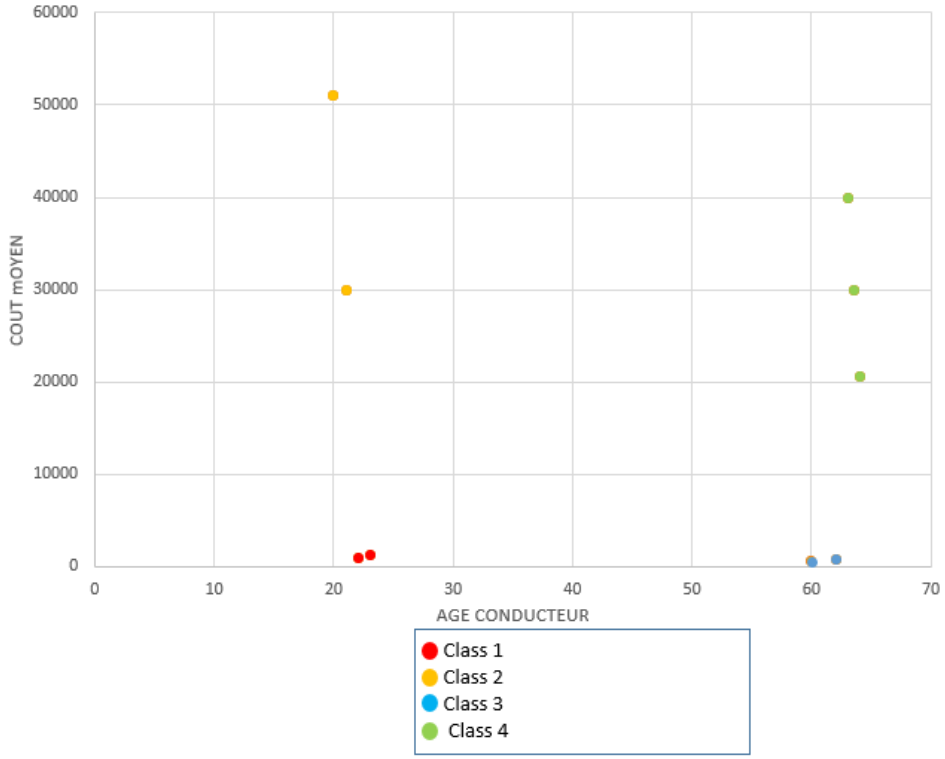

FIGURE 24 – Classification supervisée appliquée au Coût Moyen

Vue l'homogénéité des classes crées par la classification supervisée, nous optons pour cette dernière dans le but de créer les classes sur lesquelles nous allons se baser pour créer le modèle.

3.2 Application sous R

Nous cherchons à créer un modèle qui généralise les profils dans les données existantes et de classer correctement les nouvelles observations. Pour y arriver, nous importons la table "BaseFus", puis nous partitionnons les observations en trois sous ensembles, soit 60% pour l'apprentissage, 30% pour la validation, et 10% pour tester la robustesse du modèle. Par défaut, R affiche l'arbre qui maximise la performance sur l'ensemble de données de validation. Pour appliquer la classification supervisée, nous définissons la variable "Cout-Moyen" comme variable à expliquer et les autres variables comme variables explicatives. Au premier lieu nous avons généré l'arbre de décision du "CoutMoyen" en fonction de chacune des variables explicatives (une seule variable pour chaque arbre de décision). Posons n_i le nombre de feuilles de l'arbre générée par la variable explicative X_i. A titre d'exemple, l'arbre de décision [25] génère quatres feuilles finales pour la variable "Agepermis".

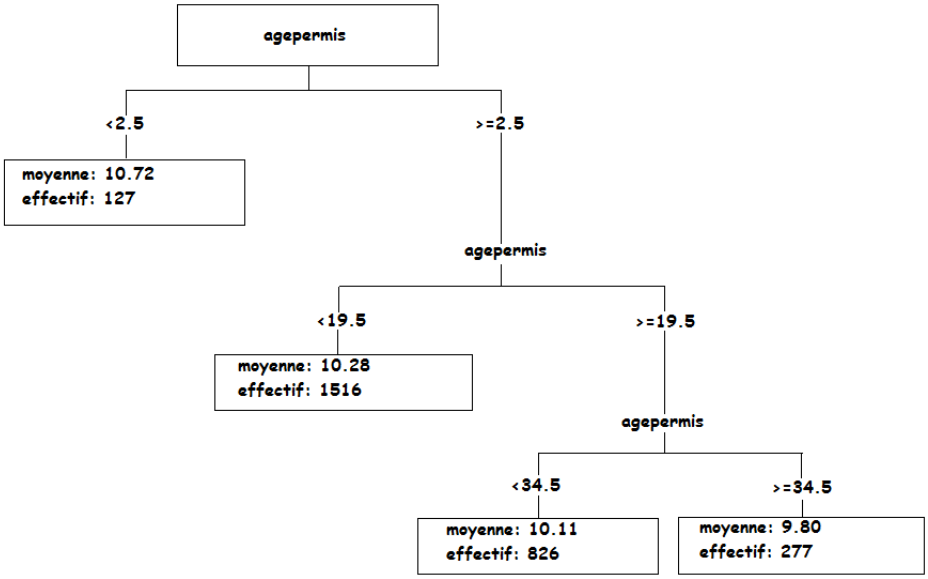

FIGURE 25 – Arbre de décision en fonction "Agepermis"

Après avoir généré les arbres qui correspondent à chacune des variables explicatives, nous créons sur SAS, pour chaque variable X_i, une variable Clus_i qui prend les valeurs de 1 jusqu'à n_i . Par exemple, pour la variable "Agepermis", "Clus_Agepermis" appartient à $[|1, 4|]$. En deuxième lieu, nous avons utilisé, sur R, toutes les variables explicatives à la fois afin de générer l'arbre de décision de la variable à expliquer. L'arbre produit 10 feuilles finales. De là nous pouvons identifier les classe les plus risquées, comme le montre la figure [26]. D'ailleurs si un assuré a un permis de moins de 9,5 ans et un véhicule de plus de 21,5 ans de puissance fiscale supérieure à 7,5 , sera classé parmis les observations ayant une forte probabilité de commettre un sinistre grave. Sur SAS, nous transformons l'arbre obtenue en un vecteur "clus" qui prend les valeurs de 1 jusqu'à 10.

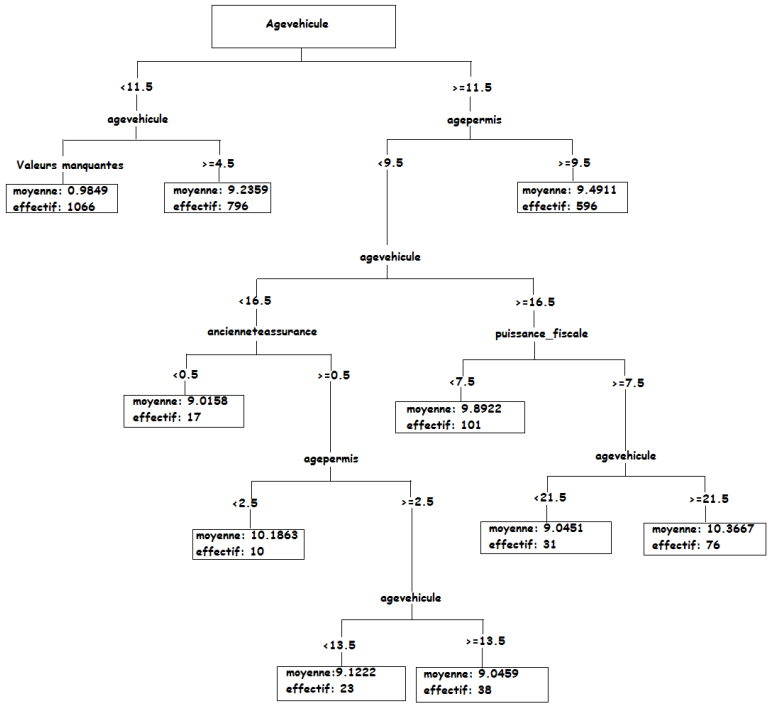

FIGURE 26 – Arbre de décision en fonction de toutes les variables explicatives

Nous utilisons les variables "Clus" et "Clus_i", au lieu des variables explicatives, pour créer le modèle GLM. En effet, nous créons deux modèles GLM, le premier utilise les variables "Clus_i", tandis que le deuxième utilise le vecteur “Clus”. Lors de la création des deux modèles, nous avons constaté que les deux méthodes de classification donnent des résultats similaires. Or le deuxième s’avère plus simple à manipuler et à interpréter. De ce fait nous utilisons le vecteur "Clus" pour modéliser le "CoutMoyen".

4 Modèles linéaires généralisés

4.1 Cadre général de GLM

Les modèles linéaires généralisés permettent d’étudier la liaison entre une variable dépendante ou réponse Y et un ensemble de variables explicatives ou prédicteurs $X_1, X_2, ..., X_n$. Ils englobent les modèles *linéaire général et log-linéaire* ainsi que les régressions *logistique et de Poisson.*
Ces modèles, communément appelés GLM, sont formés de trois composantes :

— *La composante aléatoire,* qui est la variable indépendante à laquelle est associée une loi de probabilité.

— *La composante déterministe,* qui est une combinaison linéaire des variables explicatives $X_1, X_2, ..., X_n$.

— *La fonction « Lien »,* qui décrit la relation fonctionnelle entre la combinaison linéaire des variables $X_1, X_2, ..., X_n$ et l’espérance mathématique de la variable de réponse Y.

- La composante aléatoire :

Identifie la distribution de probabilités de la variable à expliquer. On suppose que l'échantillon statistique est constitué de p variables aléatoires $Y_1, Y_2, ..., Y_p$ indépendantes admettant des distributions issues d'une *structure exponentielle*. Cela signifie que leur famille de densités s'écrit sous la forme :

$$f_Y(y|\theta) = a(\theta)b(y)\ \exp[\ \phi(\theta) \cdot T(y)\ ] \tag{9}$$

où :

— ϕ est appelé *paramètre naturel* de la famille exponentielle

— $T(y)$, $b(y)$, $\eta(\theta)$, et $a(\theta)$ sont précisées

A noter que cette formulation inclut la plupart des lois usuelles comportant un ou deux paramètres : *gaussienne, gaussienne inverse, gamma, Poisson, binomiale...*

- La composante déterministe :

Celle-ci est exprimée sous forme d'une combinaison linéaire :

$$\eta = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{10}$$

où β est un vecteur de n paramètres, appelé *prédicteur linéaire*. L'estimation des paramètres β_j est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé ou en minimisant sa déviance *(que nous allons introduire dans la section "critère de déviance").*

- Le lien :

La troisième composante exprime une relation fonctionnelle entre la composante aléatoire et le prédicteur linéaire. Elle spécifie comment l'espérance mathématique de Y notée E[Y] est liée au prédicteur linéaire construit à partir des variables explicatives. De ce fait, on obtient :

$$g(\mu = E[Y]) = \eta \tag{11}$$

où g représente la fonction lien, supposée monotone et différentiable. Si $g = log$ on dit que celle-ci est *log-linéaire*, sinon si $g(\mu) = log(\frac{\mu}{1-\mu})$ on dit qu'elle représente un *logit* et qu'elle modélise le rapport des chances, sinon g peut être simplement la fonction identité.

En général, un modèle GLM s'écrit sous la forme d'une combinaison linéaire à laquelle s'ajoute un terme d'erreur (ϵ), qui appartient lui-même à une famille exponentielle. Celui-ci est non centré sur 0 et sa variance est dépendante des X_i.

$$g(\mu = E[Y]) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon \tag{12}$$

Une fois que ce modèle est identifié, celui-ci doit être évalué et ajusté sur la base des différences entre observations et estimations. Plusieurs critères peuvent être appliqués à cet effet, dont nous ne citerons que *le critère de déviance* et *le test de Pearson.*

- Critère de déviance :

Avant d'expliquer le concept de ce test, introduisons la notion de *vraisemblance* qui représente la probabilité qu'un échantillon observé se réalise. Soit l'échantillon $Y_1, Y_2, ..., Y_p$ un échantillon à p observations de la variable Y. Alors la vraisemblance d'observer cet échantillon en fonction des β_i, supposés connus, se définit par :

$$V(Y; \beta_1, \beta_2, ..., \beta_n) = \prod_{i=1}^{p} f(y_i; \beta_1, \beta_2, ..., \beta_n) \tag{13}$$

tel que :

$$f(y_i; \beta_1, \beta_2, ..., \beta_n) = \begin{cases} f_{\beta_1,\beta_2,...,\beta_n}(y) & \text{si } Y \text{ est continue} \\ P_{\beta_1,\beta_2,...,\beta_n}(Y = y) & \text{si } Y \text{ est discrète} \end{cases} \tag{14}$$

La log-vraisemblance se définit à son tour, par :

$$L(Y; \beta_1, \beta_2, ..., \beta_n) = \ln V(Y; \beta_1, \beta_2, ..., \beta_n) \tag{15}$$

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 se fait à l'aide de la déviance D avec des vraisemblances ou à l'aide de la déviance normalisée D' des log-vraisemblances L et L_{sat} :

$$D = \sqrt[2]{\frac{V}{V_{sat}}} \tag{16}$$

$$D' = -2(L - L_{sat}) \tag{17}$$

Lors de l'ajustement d'un GLM, l'objectif serait d'approcher la valeur de D à 1 ou de minimiser D'.

- Test de Pearson :

Un test du χ^2 est également utilisé pour comparer les valeurs observées y_i à leur prévision par le modèle. La statistique du test est définie par :

$$\chi^2 = \sum_{i=1}^{p} \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} \tag{18}$$

où $\hat{\mu}_i$ est l'espérance des $\hat{y}_i$ prédits par le modèle GLM.
Cette statistique est distribuée approximativement selon une loi χ^2 à $n - p$ degrès de liberté si le modèle étudié est exact.

4.2 Modélisation des coûts moyens annuels des sinistres

Comme nous l'avons précédemment mentionné, notre objectif est de prédir les charges moyennes annuelles des sinistres au profit d'une société d'assurance. Pour ce faire, nous appliquons la méthode de régression linéaire généralisée à la base de données « baseTrain » qui est la concaténation des deux bases : Part_Train et Part_Validate.

Dans ce cas, notre variable indépendante n'est autre que coutMoyen tandis que notre variable explicative est *clus*, que nous avons obtenue grâce à la classification supervisée introduite à la section 3.3.

Sur le logiciel SAS, nous avons utilisé la procédure GENMOD tout en fixant un ensemble de paramètres, dont :

- La loi de la variable à expliquer :

Etant donné que *coutMoyen* suit une loi continue, nous avons décidé de construire deux modèles : l'un avec une loi *Gamma* et l'autre avec une loi *log-normale*. En ce qui concerne la deuxième loi choisie, notons que celle-ci ne fait partie de la famille exponentielle. Par conséquent, nous avons tout d'abord défini une nouvelle variable « logCout », égale au logarithme de coutMoyen, à laquelle nous avons appliquée une loi *Normale*.

- La fonction lien :

Pour la loi Gamma, nous avons opté pour la fonction communément utilisée *Log* alors que pour la loi normale, nous avons appliqué la fonction *identité*.

Cette procédure nous permet d'obtenir deux tableaux, le premier [Figure 27]affiche les critères d'évaluation de la qualité d'ajustement du modèle et le deuxième donne le prédicteur linéaire β et l'intervalle de confiance des différents paramètres β_j

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4109	7538.7298	1.8347
Scaled Deviance	4109	5014.5329	1.2204
Pearson Chi-Square	4109	16595.7324	4.0389
Scaled Pearson X2	4109	11038.9745	2.6865
Log Likelihood		-45764.5849	
Full Log Likelihood		-45764.5849	
AIC (smaller is better)		91551.1698	
AICC (smaller is better)		91551.2341	
BIC (smaller is better)		91620.7269	

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4109	7752.1572	1.8866
Scaled Deviance	4109	4119.0000	1.0024
Pearson Chi-Square	4109	7752.1572	1.8866
Scaled Pearson X2	4109	4119.0000	1.0024
Log Likelihood		-7146.9548	
Full Log Likelihood		-7146.9548	
AIC (smaller is better)		14315.9095	
AICC (smaller is better)		14315.9738	
BIC (smaller is better)		14385.4666	

FIGURE 27 – Les critères de la loi gamma (à gauche) et de la loi normale (à droite)

En comparant les données des deux tableaux , on remarque que la déviance normalisée et la valeur de χ^2 (test de Pearson) de la loi normale sont plus proches de 1 que celles de la loi gamma. Par conséquent nous avons opté pour *le modèle de la loi normale* présenté par la [Figure 28] .

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	9.8182	0.1921	9.4417	10.1947	2612.18	<.0001
clus	1	1	-0.8174	0.1951	-1.1998	-0.4351	17.56	<.0001
clus	2	1	-0.6068	0.1963	-0.9916	-0.2221	9.56	0.0020
clus	3	1	-0.3112	0.1975	-0.6984	0.0759	2.48	0.1151
clus	4	1	-0.8562	0.3446	-1.5315	-0.1808	6.17	0.0130
clus	5	1	0.0358	0.2201	-0.3956	0.4672	0.03	0.8708
clus	6	1	0.3241	0.3931	-0.4464	1.0945	0.68	0.4097
clus	7	1	-0.0657	0.2744	-0.6036	0.4722	0.06	0.8108
clus	8	1	0.4322	0.2331	-0.0247	0.8890	3.44	0.0637
clus	9	1	-0.6367	0.3037	-1.2321	-0.0414	4.39	0.0361
clus	10	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	1.3719	0.0151	1.3426	1.4018		

FIGURE 28 – Le prédicteur linéaire du modèle GLM choisi

Donc, notre modèle s'écrit sous la forme :

$$E[logCout] = 9,8182 - 0,8174 \times \mathbf{1}(clus_1) - 0,6068 \times \mathbf{1}(clus_2) - 0,3112 \times \mathbf{1}(clus_3) - 0,8562 \times \mathbf{1}(clus_4)$$

$$+0,0358 \times \mathbf{1}(clus_5) + 0,3241 \times \mathbf{1}(clus_6) - 0,0657 \times \mathbf{1}(clus_7) + 0,4322 \times \mathbf{1}(clus_8) - 0,6367 \times \mathbf{1}(clus_9)$$

Avec $\mathbf{1}(clus_1)$ est l'indicatrice de clus1 et $Scale = \sigma^2$

De l'équation précédente on déduit que les assurés qui appartiennent aux *clus5, clus6 et clus8* ont un risque de sinistralité supérieur aux autres.

D'où si un assuré appartient à clus6, l'espérance de son logcout augmente de 0.3241, donc son cout est multiplié par $e^{(0.3241+1.3719/2)}$. A l'inverse, si un assuré appartient à clus1 l'espérance de son logcout diminue de 0.8174, donc son cout est multiplié par $e^{(-0.8174+1.3719/2)}$

Une fois que nous avons fait les prédictions (comme nous l'expliquerons à la section 5) nous avons jugé que notre modèle n'est pas assez bon. Donc nous avons pris l'initiative de rajouter les variables initiales de notre base de données à celles présentées ci-dessous en vue de l'améliorer. Le modèle obtenu est présenté dans la [Figure 29]

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	9.8079	0.2119	9.3926 10.2232	2142.96	<.0001
clus	1	1	-0.7213	0.1968	-1.1071 -0.3355	13.43	0.0002
clus	2	1	-0.5066	0.1978	-0.8943 -0.1189	6.56	0.0104
clus	3	1	-0.2342	0.1995	-0.6253 0.1569	1.38	0.2405
clus	4	1	-0.8755	0.3438	-1.5493 -0.2018	6.49	0.0109
clus	5	1	0.0030	0.2196	-0.4274 0.4335	0.00	0.9889
clus	6	1	0.2899	0.3921	-0.4786 1.0584	0.55	0.4597
clus	7	1	-0.0962	0.2738	-0.6328 0.4404	0.12	0.7252
clus	8	1	0.4016	0.2325	-0.0541 0.8574	2.98	0.0841
clus	9	1	-0.6251	0.3029	-1.2189 -0.0314	4.26	0.0391
clus	10	0	0.0000	0.0000	0.0000 0.0000	.	.
SEXE	F	1	-0.1885	0.0568	-0.2997 -0.0772	11.02	0.0009
SEXE	M	0	0.0000	0.0000	0.0000 0.0000	.	.
ageconducteur		1	0.0032	0.0029	-0.0025 0.0088	1.20	0.2739
agepermis		1	-0.0068	0.0032	-0.0130 -0.0006	4.59	0.0322
ancienneteAssurance		1	-0.0190	0.0072	-0.0332 -0.0049	6.98	0.0083
Scale		1	1.3679	0.0151	1.3387 1.3978		

FIGURE 29 – Le prédicteur linéaire du modèle GLM amélioré

5 Validation du modèle

Une fois le modèle est établi, nous pouvons tester sa robustesse tout en effectuant des prédictions sur la base " Part_Test". Cette base contient 10 % des observations initiales qui ne sont pas utilisées pour construire le modèle. Afin d'obtenir les prévisions, nous utilisons l'instruction PROC PLM sous SAS. Cette fonction agit sur les résultats du modèle établi. Ainsi il est nécessaire d'ajouter l'argument "store" qui permet de stocker les résultats dans une nouvelle base appelée "p8.mod1", à la fonction PROC GENMOD. La fonction PROC PLM fournit, pour de nouvelles observations, la valeur prédite et son intervalle de confiance. Il faut garder en mémoire qu'il faut effectuer un certain nombre de transformations sur les valeurs prédites et leurs intervalles de confiance. En effet, la fonction PROC PLM nous fournit la valeur que prend $E[log(CoutMoyen)] = \mu$. Ainsi, nous pouvons nous ramener à $E[CoutMoyen]$ via la transformation suivante : $e^{(\mu+\sigma^2/2)}$ ou $\sigma^2 = 1.38$ est donnée par le modèle PROC GENMOD. Le tableau [30] représente un échantillon de la base de données Part_test et les prévisions du CoutMoyen générées par le modèle GLM. En comparant les valeurs estimées aux valeurs réelles, nous remarquons que notre modèle surestime la valeur de la variable à expliquer, ceci résulte de l'importance du paramètre β_0 , ainsi que la domination des coefficients des identités des clusters sur ceux des autres variables explicatives.

	ancienneteAssurance	CoutMoyen	cluster	ClAgePermis	ClAgeVehicule	Clanciennete	clus	logcout	Lower 95% Confidence Limit	Upper 95% Confidence Limit	prediction
1	1.138109589	10150	1	2	1	2	1	9.2252289845	9.0300965817	9.2098838411	18104.941869
2	2.669260274	9093.63	1	2	1	2	1	9.1153294474	8.7363998867	8.9642844571	13825.80665
3	4.2718630137	6000	3	2	2	2	3	8.6995147482	9.4612528399	9.7221588294	29017.542348
4	1.7936712329	6000	1	2	1	2	1	8.6995147482	9.0092944857	9.2055239372	17878.590445
5	4.0109589041	6000	1	2	1	2	1	8.6995147482	8.7293388396	8.9444204442	13640.923603
6	1.7404931507	54650	1	2	1	2	1	10.908704494	9.0212433495	9.1913988666	17859.147271
7	10.557424658	1405.06	1	3	1	2	1	7.2478352855	8.7520780403	9.0193935225	14323.914776
8	4.474109589	34650	3	2	2	2	5	10.453053005	9.5968168511	10.026589728	36157.992278
9	2.8660273973	7716	3	2	2	2	2	8.951051374	9.2087749105	9.4013531272	21785.80781
10	0.6078630137	6150	2	1	2	1	2	8.7242073608	9.210774313	9.4706197957	22576.098888
11	7.3321369863	37453.77	1	3	1	2	1	10.530862651	8.8345312002	9.0323025224	15023.439012
12	9.7744931507	2825	3	3	2	2	3	7.9462636436	9.2653506257	9.5401220418	24021.062623
13	6.7076164384	26847.24	3	2	2	2	3	10.197918301	9.3815457454	9.602434734	26263.626542
14	1.4576712329	34750.6	1	4	1	2	2	10.455952117	9.0427117944	9.3220053706	19270.073633
15	1.2323835616	28718.94	3	2	2	2	5	10.265312114	9.6383083266	10.060762444	37552.130196
16	0.3688767123	6000	1	1	1	1	1	8.6995147482	9.0351429519	9.2582561423	18595.02865
17	1.2241643836	3606.07	1	2	1	2	1	8.1903738157	8.7895812636	9.0255328128	14639.91583
18	1.0109589041	119500	1	4	1	2	1	11.69107165	8.8857314923	9.1109186358	16030.926856
19	1.7103561644	45534.87	1	3	1	2	1	10.726233685	8.7311092192	8.959043039	13753.190728
20	0.4521917808	5804.58	1	3	1	1	1	8.6664025401	8.9748945698	9.1502519304	17094.691405

FIGURE 30 – Prédiction de la base "Part_test"

6 Amélioration du modèle

Après avoir analysé notre modèle au moyen de mesures et tests statistiques diverses, nous parvenons au fait que notre modèle n'est pas assez robuste pour être adopté par une société d'assurance.

Effectivement, nous avons démontré à travers les exemples de cas présentés que les prédictions du modèle linéaire généralisé élaboré ne sont pas fiables. Rappelons que dans une même classe (cluster), nous trouvons d'importantes différences entre les charges moyennes annuelles réelles que le modèle n'arrive pas toujours à détecter. Par conséquent celui-ci a souvent tendance à prédire une seule charge annuelle pour tous les clients appartenant au même groupe.
Ceci est dû à plusieurs causes, dont la principale est le *manque de variables explicatives réellement significatives* : l'analyse exploratoire que nous avons présenté dans la section 2.2 a démontré qu'aucune des variables n'influait sur la charge. De notre côté, nous ne sommes pas parvenus à trouver une combinaison de variables qui soit parlante malgrè les différentes méthodes de classification utilisées. Toutes ces méthodes ne nous ont pas aidé à diviser notre base de données en groupes assez petits pour cibler les bonnes estimations et assez grands pour éviter le phénomène de surapprentissage.

En vue d'améliorer ce modèle, nous proposons d'ajouter à notre base de données de nouvelles variables, comme : *l'usage du véhicule* (de service, familial, personnel,...) étant donné que le comportement du conducteur change en fonction de son rapport avec le véhicule qu'il conduit ; *l'historique des sinistres du client* (conditions du sinistre, fautif ou pas, ...) car tout conducteur est susceptible de commettre une erreur similaire à l'une précédemment commise ; *la région où conduit le client* et *sa densité de population* parce qu'il est évident que la probabilité d'accident d'un client habitant Casablanca est supérieure à celle d'un client habitant Ben Guerir.
Une autre perspective d'amélioration proposée est de combiner plusieurs algorithmes relevant de l'analyse prédictive pour effectuer les prédictions et de ne pas dépendre uniquement de la régression linéaire généralisée.

References

[1] Etude de la tarification et de la segmentation en assurance automobile-Guillaume GONNET .

[2] Modelisation de la fréquence des sinistres en assurance automobile–Olga A. VASECHKO, Michel GRUN-REHOMME, Noureddine BENLAGHA .

[3] Statistique de l'assurance - Arthur CHARPENTIER.

[4] Flottes automobiles :Un nouveau modèle de tarification.Impact de la conservation sur la distribution du ratio sinistres à primes - NGUYEN Thi To-Vong .

[5] Provisionnement pour sinistres à payer : analyses et modelisations sur données détaillées-Magali KELLE .

[6] Processus de surveillance et de majoration des contrats flottes d'entreprise d'AXA France - Romain BOYER CHAMMARD .

[7] Modélisation du coût des sinistres extrêmes en assurance automobile - Marianne VEGNI Modélisation de la fréquence des sinistres en assurance automobile .

[8] Modélisation des distributions de sinistres - Hélène Cossette,Vincent Goulet,Michel Jacques,Mathieu Pigeon .

- Liens Internet :

[9] http://maths.cnam.fr/IMG/pdf/Presentation_MODGEN_02_2007.pdf.

[10] https://jonathanlenoir.files.wordpress.com/2013/12/modeles-lineaires-generalises-glm.pdf.

[11] https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-mlg.pdf.

[12] http://support.sas.com/.

[13] https://communities.sas.com.