Université de Lille

Projet d'Assurance

2018-2019

Sinistralité en assurance

Realisé par :
Nasradine IBRAHIM
Zeinab DIA

Encadré par :
Pr. Yves Arrighi

17 janvier 2020

TABLE DES MATIÈRES

Remerciement

Tout d'abord, ce travail est le fruit d'une collaboration avec ***moi*** et Zeinab tous les deux étudiants parcours MIASHS de l'UFR MIME à l'Université de Lille 3 .Je remercie énormément Monsieur ***Arrighi*** son soutien qu'il nous a donné et également pour sa gentillesse .

CHAPITRE 1

INTRODUCTION

La notion de sinistralité en assurance automobile se mesure en termes de nombre des accidents (fréquences) et de coût de ces accidents. Dans ce marché marqué par les concurrences , l'assureur cherche à sélectionner des éléments qui contribuent à expliquer la sinistralité. On va s'intéresser aux facteurs explicatifs du nombre d'accidents responsables déclarés par l'assuré à son assureur.

Tout d'abord ,on va utiliser le modèle binomial pour illustré la survenue d'un sinistre ou pas en cas d'accidents sur une police d'assurance .
Cette première partie est développée en python .

Deuxièmement , pour répondre à l'importance du nombre d'assurés sans sinistre sur une période d'exercice et à l'hétérogénéité de cette population (absence de sinistre ou sinistre non déclaré), des modèles à « inflation de zéros » sont proposés : le modèle de Poisson et le modèle binomial négatif.
Cette partie est codée en SAS.

CHAPITRE 2

ANALYSE EXPLORATOIRE DE DONNÉES

Notre Table de données contient 100021 individus et 20 variables :

- PolNum : Numéro de la police est une variable numerique
- CalYear : Année calendaire de souscription est une variable numerique constante
- Gender : Genre du conducteur (Homme/Femme) est une variable qualitative
- Type : Type de Vehicule une variable qualitative
- Category : Categorie du vehiculeune variable qualitative
- Occupation : Profession une variable qualitative
- Age : Age du conducteur est une variable quantitative
- Group1 : Groupe du vehicule est une variable quantitative
- Bonus : Bonus-Malus est une variable quantitative
- Poldur : Ancienneté du contrat est une variable quantitative
- Value : Valeur du vehicule est une variable quantitative
- Adind : Indicateur d'une garantie dommages est une variable categorielle binaire
- Subgroup2 : Sous-region d'habitation est une variable qualitative
- Group2 : Region d'habitation est une variable qualitative

- Density : Densité de la population est une variable quantitative
- Expdays : Exposition en jours est une variable quantitative
- Nb1 : Nombre de sinistres RC Materiels est une variable de comptage
- Nb2 : Nombre de sinistre RC corporels est une variable de comptage
- Surv1 : Survenance de sinistres RC Materiels est une variable binaire
- Surv2 : Survenance de sinistres RC Corporels est une variable binaire

Nous disposons :

- **Les caracteristiques du conducteur**{ PolNum,CalYear,genre,Occupation,Age,Bonus,Expdays,group2,Subg et Density }

-**Les caracteristiques du véhicule**{ Type,Category,group1,Value}

-**Le type de contrat**{Poldur et Adind}

-**La sinistralité** {Nb1,Nb2,Surv1,Surv2}

Comme précisé précédemment, nous allons procéder à réaliser deux analyses statistiques :

-Une analyse univariée

-Une analyse bivariée

2.1 Analyse Univariée

Mais avant de procéder à l'analyse univariée, nous vous présentons un tableau de statistiques descrptives de l'ensemble des variables (à l'exception des variables catégorielles).

	Age	Group1	Bonus	Poldur	Value	Density
count	100021.000000	100021.000000	100021.000000	100021.000000	100021.000000	100021.000000
mean	41.122514	10.692625	-6.921646	5.470781	16454.675268	117.159270
std	14.299349	4.687286	48.633165	4.591194	10506.742732	79.500907
min	18.000000	1.000000	-50.000000	0.000000	1000.000000	14.377142
25%	30.000000	7.000000	-40.000000	1.000000	8380.000000	50.625783
50%	40.000000	11.000000	-30.000000	4.000000	14610.000000	94.364623
75%	51.000000	14.000000	10.000000	9.000000	22575.000000	174.644525
max	75.000000	20.000000	150.000000	15.000000	49995.000000	297.385170

FIGURE 2.1 – Tableau de statistique descriptive

D'après les resultats de cet tableau on peut en tirer :
- L'age des clients varie entre 18 et 75 ans .L'age moyen vaut 41,12 ans et au moins 50% ont un age inferieur à 40 ans .
-La durée moyenne du contrat est de 5 ans .La plus grande ancienneté du contrat vaut 15 ans .

On a décider de recoder les variables qualitatives et de définir des catégories de référence.
-La variable **Gender** a eté codé en binaire et les femmes sont mis dans la categorie des reference.
-La variable **Occupation** prend 1 si l'individu est employé et 0 sinon.
- La variable**Category** prend 1 si la catégorie de voiture est Large
-La variable **Type** prend 1 si la voiture est de type A et 0 sinon
-La variable **Exppdays** est considérée comme la variable offset pour la suite de la modélisation .

2.2 Analyse Bivariée

En faisant une analyse exploratoire sur les variables quantitatives ,on croise les variables et on obtient cette matrice de corrélation suivante :

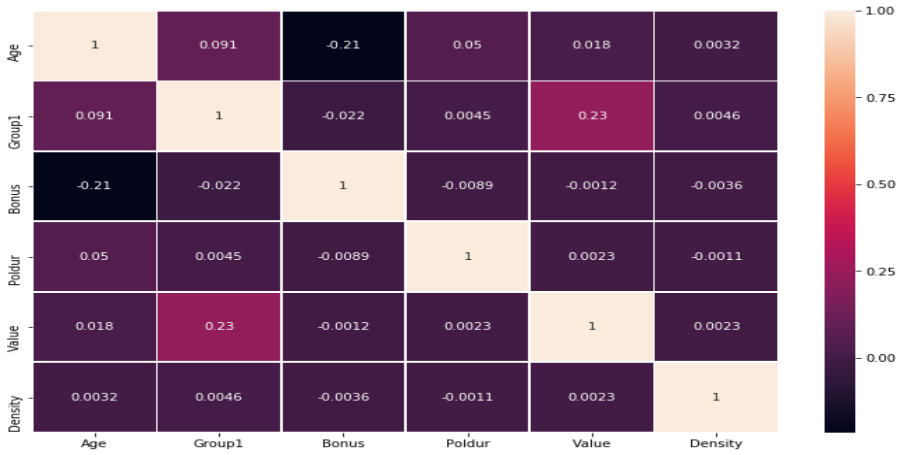

FIGURE 2.2 – Matrice de corrélation des variables quantitatives

Les coefficients de corrélations sont inférieurs à 0.25,ce qui montre que les variables ne sont pas corrélées entre elles .Il n'y aura donc pas un problème de colinéarité dans la partie de la modélisation .

En établissant une table de fréquence entre les variables qualitatives ,on obtient cet résultat suivant :

Genre	Fréquence	Pourcentage	Adind	Fréquence	Pourcentage
Femme	36578	36.57	0	48796	48.79
Homme	63443	63.43	1	51225	51.21
	Occupation			Surv1	
Employé	31150	31.14	0	87744	87.73
Femme au foyer	20010	20.01	1	12277	12.27
Retraité	13167	13.16		Category	
Travailleur indépendant	20372	20.37	Large	35130	35.12
Sans emploi	15322	15.32	Medium	36644	36.64
			Small	28247	28.24

FIGURE 2.3 – Table de fréquences des variables qualitatives

La table des fréquences de chaque modalité des variables nous apprend que la population est composé majoritairement d'hommes (63,43%) contre seulement 36,57% de femmes.

Les professions parmi les hommes et femmes les plus représentées sont les employés (31,14%), suivi des travailleurs indépendants (20.38%) et des femmes au foyer (20.02%), les sans-emplois (15,34%) et La profession la moins fréquente est celle des retraités avec 13,17%. Les voitures qui occupent la plus grande part de la population est celle des voitures de tailles moyennes avec une proportion de 37%, suivie de celle de grandes tailles (35,12%), enfin la catégorie des voitures de petites tailles (28,24%). Il y a un peu plus de client souscrivant à une garantie dommages (51,21%). Enfin il y a près de 12% de sinistres matériels dans la population d'assurées.

CHAPITRE 3

MODÈLE LOGIT BINOMIALE

Le modèle logit ou régression logistique est un modèle de régression ou la variable dépendante est une variable binaire .Dans notre cas,notre variable dépendante **Surv1** indiquant la survenue ou non de sinistres.

Pour la validation de notre modèle ,on essaye de diviser notre échantillons en deux : Un ensemble d'entraînement 70% de nos données pour évaluer notre modèle et un ensemble de test 30% pour la validation .

3.1 Matrice de Confusion

La *matrice de confusion* est un bon outil utilisé dans l'apprentissage supervisée servant 'a donner des mesures sur la qualité du classication .

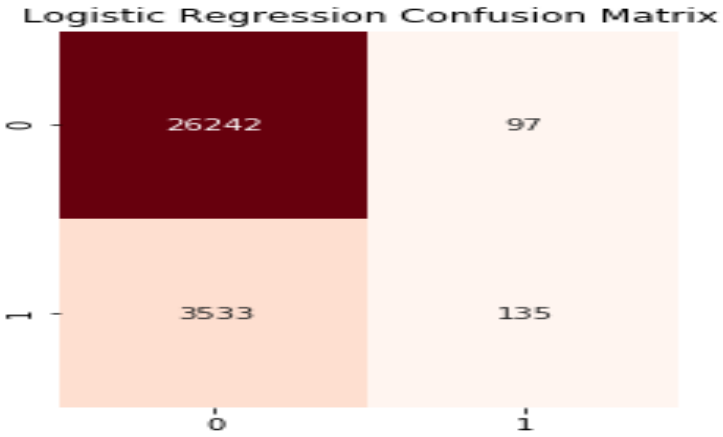

FIGURE 3.1 – Matrice de confusion

Les coefficients sur la diagonale indique les éléments bien classés ici au total notre modèle a bien classé 26377 sur 30085 individus , les coefficients en dehors de ceux que le classifieur a mis dans la mauvaise classe.Il n'a pas parvenu à bien classer 3630 individus s'ils sont sinistrés ou non .

3.2 Interprétation des coefficients et Odds Ratios

Il est important de savoir les caractéristiques des individus qui présentent des sinistres .Pour cela on va se baser sur l'interprétation des coefficients de la régression logistique.

	Coefficient	Rapport de cote
Type	-0.230509	0.794130
Category	0.051710	1.053070
Occupation	-0.079059	0.923985
Age	-0.038287	0.962437
Group1	0.072049	1.074708
Bonus	0.011127	1.011189
Adind	-0.136572	0.872343
Density	0.004453	1.004463

FIGURE 3.2 – Coefficients et Odds ratio

On remarque que les coefficient des variables Category,Bonus,Density sont positifs . Également ,les coefficients des variables Type ,Age sont négatifs .

-Les assurés qui ont des voitures de catégorie large ont tendance à avoir de plus de sinistres que les autres individus de catégories de voiture différentes.

-La survenance du sinistre augmente avec le coefficient du bonus-malus, ce qui est normal car les assurés avec un bonus élevé son ceux qui ont tendance à faire plus d'accidents.
-On voit bien que la survenance de sinistre diminue avec l'âge des assurés.
-On peut dire que plus la densité est élevée ,plus la survenance du sinistre augmente.

3.3 Le Scoring

Par définition , le score est une note attribuée a individu qui mesure le risque ou la probabilité de sinistre de cet individu. Il ne s'agit pas seulement de classer les clients potentiels en deux catégories, les bons et les mauvais assurés, mais de produire un indicateur quantitatif du risque individuel que présente chaque client.

Dans la plupart de temps ,on préfère plus la probabilité de risque que le score attribuée à un assuré.

	risqué avec probabilité	Non risqué avec probabilité
22851	0.250643	0.749357
7782	0.256084	0.743916
6861	0.269461	0.730539
7891	0.273332	0.726668
9492	0.280133	0.719867

FIGURE 3.3 – Probabilité du risque

Par exemple,le 22051eme client présente la probabilité de sinistre le plus faible(0.250643) c'est à dire qu'il est le moins risqué de commettre un sinistre .

3.4 La courbe ROC

La courbe ROC mesure la performance d'un classificateur binaire.

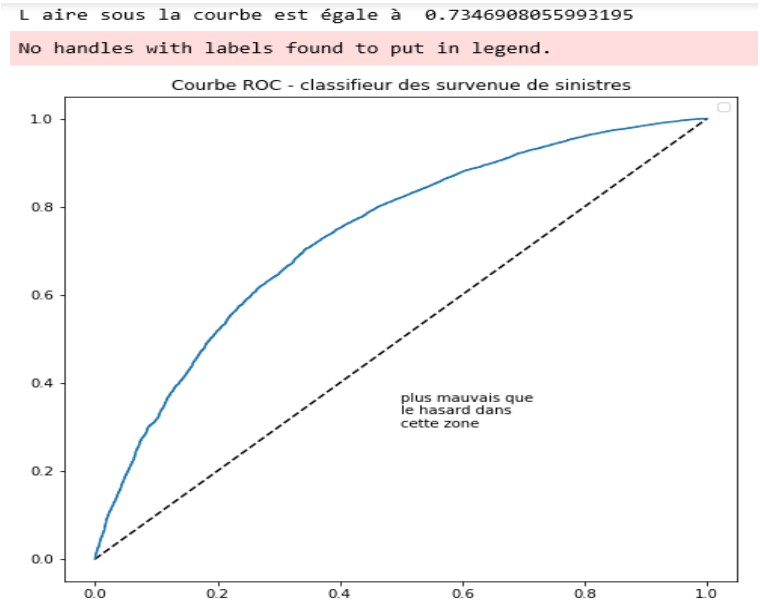

FIGURE 3.4 – La courbe ROC du modèle logit

La précision du modèle qu'une observation soit mieux prédite qu'une prédiction purement aléatoire est présentée par l'aire sous la courbe ROC.

L'aire sous la courbe pour ce modèle est égale a 0.7346 est supérieure à 0.7, cela signifie que la survenance d'un sinistre a une probabilité de 73,46%.

CHAPITRE 4

MODÈLE DE COMPTAGES

4.1 Recodage des variables

Afin d'ameliorer

4.2 Les Régressions de Poisson et Binomial Négatif

Les principes de base de la régression de Poisson, c'est donc une forme de analyse de régression ici pour modéliser les données de comptage et dans le cas particulier si tous les facteurs explicatifs les variables sont catégoriques, puis nous modélisons essentiellement un tableau de contingence. Et le modèle modélise essentiellement les fréquences attendues. Le modèle précise également comment la variable de comptage se rapporte évidemment à l'une de ces variables explicatives ou par exemple au niveau des variables catégorielles. Les modèles de Poisson sont une forme de modélisation linéaire généralisée.

Les résultats issus des modèles de régression de Poisson sont valides si :

-Les réponses sont indépendantes.

-Les réponses sont distribuées selon une loi de Poisson, de paramètre Lambda.

-Il n'existe pas de sur-dispersion.

4.2.1 Interprétation des coefficients

D'apres les sorties SAS pour les modèles de poisson et binomiale negative ,on obtient cet tableau suivant :

		Poisson	Négatif Binomiale
		coefficient	coefficient
constante		-7.1657	-7.2004
Age		-0.0301	-0.0293
Value	[10000,15000]	0.0521	0.0545
	[15000,20000]	0.112	0.1119
	[20000,30000]	0.1674	0.1681
	>30000	0.2097	0.2127
Density		0.0042	0.0042
Poldur		-0.0244	-0.0243
Bonus		0.0099	0.01
Gender	Female	-0.3404	-0.3345
Occupation	Housewife	0.264	0.2653
	Retired	-0.4667	-0.4823
	Self-employed	-0.1455	-0.1439
	Unemployed	0.1961	0.1947
Adind	0	0.1151	0.1159
Scale		1	0.3637

FIGURE 4.1 – Modèle de poisson et binomial négatif

On remarque tous les coefficient des variables augmentent pour le modèle binomiale negative que le modèle de poisson . . On constate alors une augmentation du sinistre avec le bonus, les femmes au foyer, les chômeurs, l'absence de la garantit dommage, et enfin avec la valeur du véhicule. La sinistralité diminue avec l'ancienneté du contrat, les femmes , l'âge et pour les retraités.

On affiche ici ci-dessous un tableau de comparaison de deux modèles :

	Poisson	Négatif Binomial
Log Likelihood	-36850.52	-36723.77
BIC	77869.32	77627.34
AIC	77726.63	77475.13

FIGURE 4.2 – Comparaison de poisson et binomial négatif

On voit que le AIC du modèle binomial negatif égale à 77475.13 est inférieur à celui du modèle de poisson égale à 77726.63 , on remarque que le modèle basé sur la loi binomiale négatif est plus ajusté à nos données que le modèle de Poisson.

4.2.2 Sur-dispersion

En calculant le ratio residual deviance / ddl . On voit que ce ratio est supérieur à 1 et permet de mettre en évidence la présence d'une sur-dispersion.

Dans l'ajustement du modèle, nous pouvons remarquer un problème de sur-dispersion des données pour les raisons suivantes :

• Une déviance par rapport au nombre de degrés de liberté .

• La variance est supérieur à la moyenne.

Les Conséquences de la sur-dispersion sont :

• Les estimations de l'erreur standard, de la statistique du khi-deux et de la valeur-p ne seront pas justes.

• Précisément, l'erreur standard sera sous-estimée et la statistique du khideux surestimée.

• Les coefficients du modele seront non biaisés mais ils risquent d'être déclarés trop souvent significatifs.

4.3 Modèles à Zéro-Inflation

Du fait d'une abondance de valeurs nulles et de la présence de quelques valeurs extrêmes, la variance est supérieure à la moyenne dans notre cas et comme la partie sur la sur-dispersion , cette situation donne sens une sous estimation des écarts types et on rejette souvent l'hypothèse nulle de non significativité des coefficients du modèle.

Pour les modèles de comptage , de nombreux statisticiens proposent désormais le modèle binomial négatif à zéro inflations et le modèle de poisson à zéro inflation. Ces modèles sont conçus pour faire face à des situations où il y a un nombre «excessif» de personnes avec un décompte de 0. Par exemple, dans une étude où la variable dépendante est «le nombre de fois qu'un élève a eu une absence non excusée», la grande majorité des étudiants peuvent avoir une valeur de 0.

Dans notre cas , la répartition des nombres dans la variable dépendante **Nb1** est donnée dans le graphique suivant :

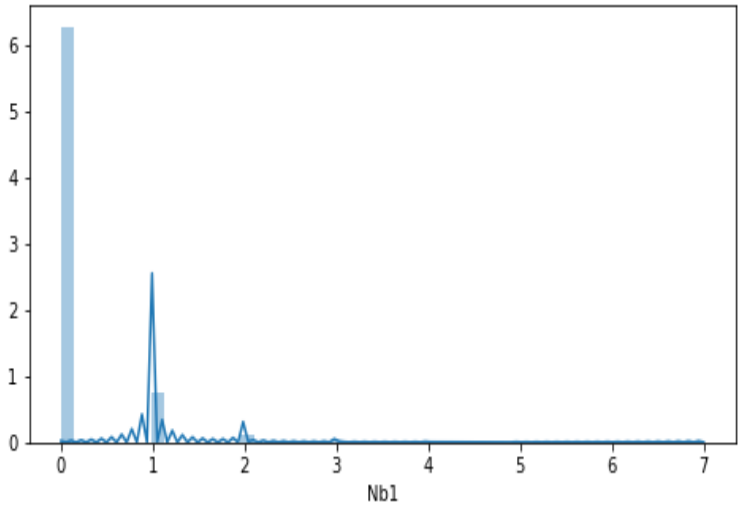

FIGURE 4.3 – Frequence de zero dans Nb1

Le graphe montre que le nombre de zeros est de 87447 soit 87.5% de la variable **Nb1**.

Il s'agit de deux sous populations : les assurés qui n'ont eu aucun sinistre dans l'année, et ceux qui ont eu un sinistre et qui ne l'ont pas déclaré.

Pour la comparaison de deux modèles à zéro inflations ,on illustre les résultats de SAS dans cet tableau suivant :

Variable		ZIP	ZINB
		coefficient	coefficient
constante		-6.882	-7.143
Age		-0.028	-0.027
value_5cat	[10000,15000]	0.0523	0.0543
	[15000,20000]	0.1081	0.1105
	[20000,30000]	0.1665	0.1678
	>30000	0.2095	0.2127
Density		0.0042	0.0042
Poldur		-0.0245	-0.0244
Bonus		0.0072	0.0084
Gender	Female	-0.3348	-0.3348
Occupation	Housewife	0.2621	0.2636
	Retired	-0.4939	-0.5064
	Self-employed	-0.1462	-0.1443
	Unemployed	0.1933	0.1949
Adind	0	0.1132	0.1115
Scale		1	0.2932

FIGURE 4.4 – Comparaison de ZIP et ZINB

On peut remarquer que le nombre de sinistre croit avec l'augmentation du coefficient du bonus, la densité de population, la valeur du véhicule. Ainsi que l'augmentation de ce nombre si l'assuré est un chômeur.

Par contre la sinistralité baisse avec l'âge, l'ancienneté du contrat, et enfin

pour les femmes assurées, les retraités et pour les travailleurs indépendants.

4.4 Test de Vuong

Le test de Vuong permet de comparer les modèles à inflation zéro utilisé précédemment deux à deux.

Comparaison des modèles poisson et ZIP

L'efficacité ou la précision des modèles via ce test de comparaison de deux modèles ZIP et Poisson nous donne dans un premier temps la valeur |z|=8.1 avec un seuil significatif de 5% et donc on a retenu le modèle ZIP.

Comparaison NB et ZINB

Ce test de comparaison entre le modèle binomial négatif avec le ZINB, pour le résultat on a trouvé |z|=2.6 avec un seuil significatif 5% donc on a retenu le modèle binomial négatif à zero inflations.

4.5 Les prévisions

Afin d'effectuer les prévisions sur les nouvelles données du fichier pricing.csv, nous avons utilisé la procédure PLM de SAS qui nous permet d'obtenir les prévisions à partir de modèle estimé au préalable. En ce qui nous concerne, nous l'avons utilisé avec les modèles logistique, de poisson, binomial négatif et les 2 modèles à inflation zéro.

Nous avons fait les prévisions sur les mêmes variables utilisées dans notre modélisation. Nous avons rencontré un problème en ce qui concerne la durée d'exposition (Expdays). En effet la variable n'existe pas dans la base pricing.csv, puisque ce sont les données des nouveaux clients donc ils ne possèdent pas encore de durée d'exposition.

Nous avons émis l'hypothèse que tous les assurés ont une durée d'exposition

d'un an (365 jours), ce qui nous aidera à d'estimer la survenue de sinistres sur la période définie pour les nouveaux assurés. Nous avons utilisé la valeur suivante comme offset : log(365) = 2,56.

Variable	Moyenne	Ecart type	Minimum	Maximum
surv1	0.134544	0.108523	0.006918	0.745913
FreqPois	0.164121	0.170568	0.007396	2.262881
FreqNegBin	0.164087	0.17081	0.007511	2.270658
FreqZIP	0.163954	0.167796	0.007232	2.031418
FreqZINB	0.163946	0.167134	0.006935	2.023109

FIGURE 4.5 – Les prévisions

La probabilité de survenance de sinistre estimé est de 0,745 environ pour le modèle logistique. Et nous observons des moyennes à peu près équivalentes, autour de 0,16 pour les modèles de poisson, binomial négatif et à inflation zéro.

CHAPITRE 5

CONCLUSION

Si les modèles actuels sont suffisamment sophistiqués pour que l'on puisse les considérer comme des outils utiles et performants et non plus comme des curiosités théoriques, il ne faut pas oublier qu'un modèle a ses limites et ne donne qu'une image imparfaite de la réalité. Les modèles doivent être utilisés de façon souple, sans y croire complètement à la limite.