**RAPPORT FINAL PROJET ECONOMETRIE DE L’ASSURANCE**

**Auteur : HABIB MBOW**

**INTRODUCTION**

Le marché des assurances est un contexte privilégié d’information incomplète entre l’agent (l’assuré) et le principal (l’assureur). L’assureur ne connaît pas tout de l’assuré, et en particulier les risques ﬁnanciers qu’il lui fait courir à travers sa sinistralité. Il s’en suit des opportunités de comportements stratégiques résultant d’une évaluation asymétrique des risques. La littérature économétrique distingue classiquement deux types d’asymétrie d’information : l’aléa moral et la sélection adverse.

L’aléa moral concerne l’inﬂuence de la couverture d’assurance sur le comportement de l’assuré. En situation d’aléa moral, les compagnies d’assurances ne connaissent pas le niveau d’autoprotection des assurés après la souscription du contrat. Les agents peuvent être incités à diminuer leur prévention du risque fondamental quand ils bénéﬁcient d’une bonne couverture d’assurance.

Concernant le phénomène de sélection adverse ou d’antisélection, l’agent qui choisit le contrat d’assurance qu’il considère le plus attrayant, dispose d’un avantage certain en matière d’information par rapport à l’assureur. Confronté à des individus hétérogènes, l’assureur se trouve dans l’incapacité de distinguer les agents selon leur degré de risque. L’assureur ne dispose donc pas d’une information suffisante lui permettant de réaliser une adéquation entre les termes du contrat et les risques individuels. Cette situation a pour conséquence qu’un même contrat d’assurance peut être oﬀert à des agents aux risques diﬀérents. Une tariﬁcation uniforme désavantage les agents de moindre risque, qui paieraient plus cher que ne l’exigent les impératifs liés au coût de leur risque

Au sein d’un portefeuille d’assurance automobile, les assurés ne sont pas tous égaux devant le risque. En eﬀet, certains présentent ont un proﬁl plus dangereux que d’autres. Du fait de cette hétérogénéité, l’assureur est contraint de segmenter le portefeuille aﬁn de constituer des classes de risques, pour ainsi proposer au souscripteur potentiel le meilleur contrat d’assurance en adéquation avec son proﬁl. Il se fonde pour cela sur toutes les informations disponibles et observables liées au conducteur (sexe, âge, ancienneté du permis de conduire...), au véhicule assuré (valeur du véhicule, groupe, type...). Malgré cela, il peut subsister une certaine hétérogénéité au sein de chaque classe, les facteurs observables qui contribuent à expliquer le risque de sinistralité des assurés.

L'objectif de ce document est de proposer une démarche et une modélisation sur la fréquence de sinistres RC matériels relative à l'activité d'une assurance automobile. Notre démarche confronte ainsi des modélisations, qui comportent nécessairement des aspects pratiques avec des résultats expérimentaux. L'étude que nous avons réalisés à ce sujet a adopté une approche sur la base.

Dans ce qui suit, nous présentons brièvement le traitement opéré au niveau de la base de données. Ensuite, nous présenterons les différentes parties adoptées pour modéliser notre variable à expliquer à l’aide de la régression logistique. Finalement, nous exposerons et analyseront de manière critique nos résultats.

**I- TRAITEMENT ET ANALYSE DES DONNEES**

Comme nous l'avons précédemment mentionné, notre objectif est de proposer une démarche et une modélisation pour l'évolution de la survenance de sinistres RC matériels relatifs à l'activité d'une assurance automobile. Pour ce faire, nous utiliserons les logiciels SAS et Excel, sur lesquels nous suivrons les étapes suivantes **:**

**I-1- LES DONNEES**

Toute compagnie d’assurance dispose d’un ﬁchier clients qui enregistre des quantités importantes d’informations. Avant d’opter pour une modélisation économétrique explicative, il est souvent utile d’analyser les données sans formuler d’hypothèses à leur égard. Toute étude sophistiquée d’une base de données doit être précédée d’une étude exploratoire à l’aide de plusieurs outils, certes rudimentaires mais robustes. C’est la meilleure façon de se familiariser avec les données.

La première étape de toute investigation dans les données est l’examen des statistiques univariées des variables aﬁn de détecter d’éventuelles anomalies dans leur distribution (valeurs manquantes, erronées ou atypiques). Dans un second temps, les statistiques bivariées permettent de repérer les incohérences entre les variables. Ces deux étapes permettent l’apurement et le redressement des données (imputation ou non des données manquantes, transformation logarithmique, discrétisation des variables continues, suppression de quelques observations le cas échéant).

**I-2- PRESENTATION DES DONNEES ET STATISTIQUES DESCRIPTIVES**

Notre projet s’appuie sur la base de données, base.xlsx qui contient des informations sur **100021** observations. Un enregistrement correspond à un couple (assuré, véhicule). Si un assuré a deux véhicules, il a deux enregistrements dans le ﬁchier. Chaque assuré dispose dans le ﬁchier de plusieurs informations qui concernent ses propres caractéristiques, les caractéristiques de son véhicule, du contrat d’assurance qu’il a choisi et des sinistres qu’il a déclarés. Ainsi, les variables se répartissent en quatre catégories :

- les caractéristiques propres au conducteur assuré,

- les caractéristiques du véhicule assuré,

- les caractéristiques du contrat

-les variables décrivant la sinistralité déclarée, en termes de fréquences, durant les années 2009 et 2010.

Le ﬁchier comprend des variables catégorielles comme le sexe du conducteur, la profession de l’assuré, sa région d’habitation, sa sous-région d’habitation, la formule d’assurance, la catégorie, de véhicule, le type de véhicule et la survenance de sinistres matériels et corporels etc. La base contient aussi des variables quantitatives : l’âge du conducteur, la densité d’habitation, le bonus-malus, la valeur du véhicule, le groupe de véhicule et le nombre de sinistres matériels et corporels.

 .

* **Caractéristiques du conducteur :**

**PolNum :** Numéro de police (unique pour chaque assuré)

**CalYear :** Année calendaire de souscription (2009, 2010)

Proportion : En 2009 : 50.01% et en 2010 : 49.99%

**Genre :** Sexe genre du conducteur (Male, Female)

Proportion : Female : 36.57 %, Male : 63.43%

**Occupation :** Profession (Employed, Housewife, Retired, Self-employed, Unemployed)

Proportion : Employed : 31.14%, Housewife : 20.01%, Retired : 13.16%,

Self-employed : 20.37%, Unemployed : 15.32%

**Age :** Age du conducteur (compris entre 18 ans à 75 ans).

Distribution : Quantile inferieur : 30, Médiane : 40, Moyenne : 41.12, Quantile supérieur : 51

**Bonus :** Bonus-Malus (compris entre -50 et 150)

Distribution : Quantile inferieur : -40, Médiane : -30, Quantile supérieur : 10

**Exppdays :** Exposition en jours (entre 91= 3 mois et 365 = un an)

Distribution : Quantile inferieur : 340, Médiane : 365, Quantile supérieur : 365

**Group2 :** Région d’habitation (modalité prise entre L et U)

Proportion : L : 23.73%, M : 7,60%, N : 5.19, O :5.21%, P : 5.26%, Q :22.38%, R :15.08%, T :4.99%, S :5.19%, U :5.36%

**SubGroup2 :** Sous-région d’habitation.

**Density :** densité de la population (entre 14.38 et 297.39)

Répartition : Quantile inferieur : 50.63, Médiane : 94.36, Quantile supérieur : 174.64

* **Caractéristique du contrat :**

**Poldur :** Ancienneté du contrat

Distribution : Quantile inferieur : 1, Médiane : 4, Quantile supérieur : 9

**Adind :** Indicateur d’une garantie dommage (1 ou 0)

Répartition : 0 : 48.79% et 1 : 51.21%

* **Caractéristique du véhicule :**

**Type :** Le type de véhicule (A, B, C, D, E et F)

Proportion : *A :* 27.75 %, B : 22.09%, C : 13.86%, D : 19.59%, E : 11.17%,

F : 5.54%

**Value :** Valeur du véhicule (compris entre 1000 et 49995)

Distribution : Quantile inferieur : 8380, Médiane : 14610.00, Quantile supérieur : 22575.00

**Group1** : Groupe du véhicule (compris entre 1 et 20).

Distribution : Quantile inferieur : 7, Médiane : 11, Quantile supérieur : 14

**Catégory :** Catégorie du véhicule (Large, Medium, Small)

Proportion : Large : 35.12 %, Medium : 36.64%, Large : 28.24%

* **Caractéristiques des sinistres :**

**Nb1 :** Nombre de sinistre RC matériels (valeur comprise entre 0 et 7)

Distribution : Quantile inferieur : 0, Médiane : 0, Quantile supérieur : 0

**Surv1 :** Survenance des sinistres RC matériels (soit 1 ou 0)

Proportion : 0 : 87.73% et 1 : 12.27%

**Nb2 :** Nombre de sinistre RC corporels (valeur comprise entre 0 et 3)

**Surv2 :** Survenance des sinistres RC corporels (Soit 1ou 0)

Proportion : 0 :95.49% et 1 :4.51%

L’échantillon fait apparaître environ autant d’hommes que de femmes assurées. Les assurés de profession Employed sont plus représentés 31.14% de la population.

Nous constatons que nos classes sont déséquilibrées, et le ratio d'absence de survenance de sinistres RC matériels et celles de RC corporels sont respectivement de 87,73% et 95.49% contre 12.27% et 4.51%. La moyenne d’âge des conducteurs est de 41.12.

Nous remarquons que notre base de données contient plus de nouveaux contractants que d’anciens.

La distribution de Nb1 indique que la plupart des valeurs renseignées sont égale à 0, on a alors une surabondance de 0. La durée d’exposition n’étant pas la même pour tous les assurés, il faudra ajouter un offset à nos modélisations pour prendre en compte que les assurés n’ont pas tous eu la même durée d’exposition. La modalité F du type de véhicule est sous représentée.

**I-3-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 « Surv1 » en fonction des variables explicatives, nous déterminons les tendances prépondérantes qui nous aiderons par la suite à la validité de notre modèle

* **La répartition de Surv1 par Age\_Conduct**

Pour déterminer les sous populations les plus critiques, nous divisons notre base de données selon les classes d’âge du conducteur que nous jugeons parlantes :

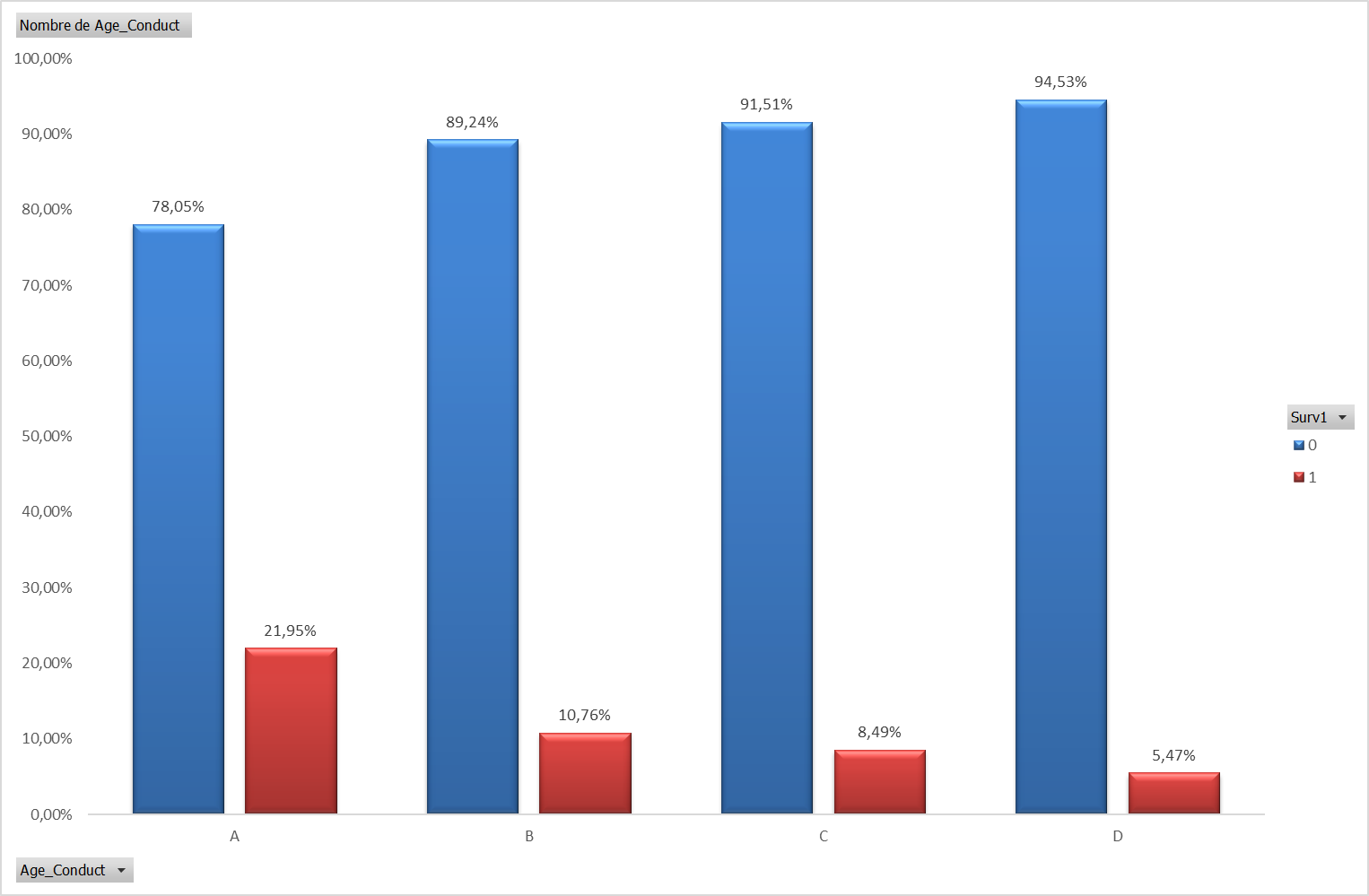
0 ≤ Age <30 → Age\_Conduct A

30 ≤ Age <45 → Age\_Conduct B

45 ≤ Age <60 → Age\_Conduct C

60 ≤ Age ≤75 → Age\_Conduct D

Pour étudier cette évolution, nous aﬃchons ci-dessous



En Analysant ce graphe, les conducteurs les plus jeunes ont une fréquence sinistre plus marquée. La fréquence sinistre décroit avec la maturité du conducteur.

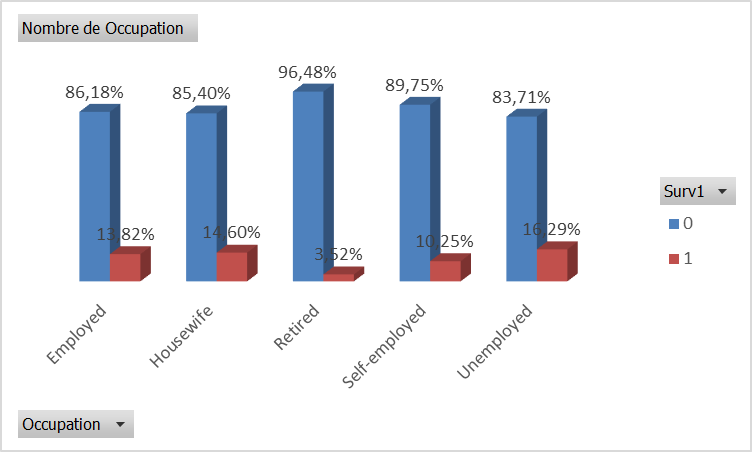
* **Répartition en pourcentage de Surv1 par rapport à Gender**

Pour étudier cette répartition, nous aﬃchons le graphe ci-dessous.

En analysant ce graphe, nous remarquons que les hommes subissent plus de sinistralité, chose qui est

normale étant donné qu’ils sont majoritaires et que les femmes plus prudentes.

**https://www.cessda.eu/**Pour étudier cette répartition, nous aﬃchons le graphe ci-dessous



Les retraités présentent le plus petit pourcentage de sinistre. Par contre, les personnes aux chômages en possèdent le plus grand.

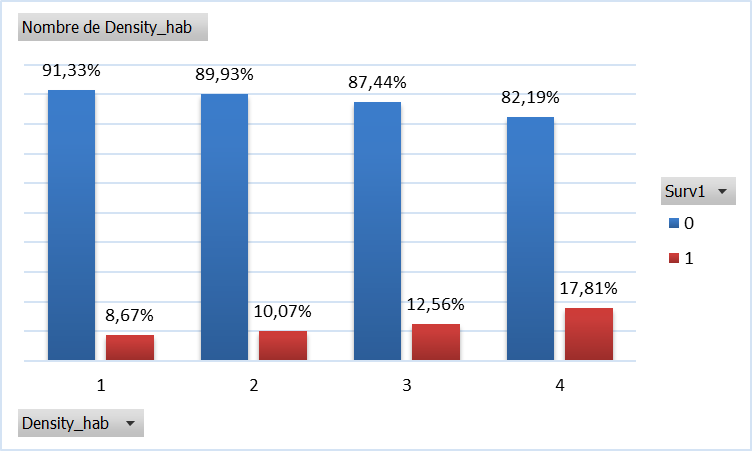
* **Répartition en pourcentage de Surv1 par Density\_hab**

Ce type de segmentation vise à avoir des classes uniformes. Le premier élément correspond au premier quantile et ainsi de suite jusqu’au dernier élément. Par cette méthode, chaque classe est composée de 25% de l’effectif.

Définition de Density\_hab

1 si Density< 51 ; 2 si 51 ≤ Density<94.4; 3 si 94.4 ≤ Density<174.6 ; 4 sinon

Pour étudier cette répartition, nous affichons le graphe ci-dessous



Le pourcentage de sinistre Rc matériels atteint son maximum au cours de Density\_hab 4.

Nous remarquons que la survenance de sinistre matériels augmente avec Density\_hab

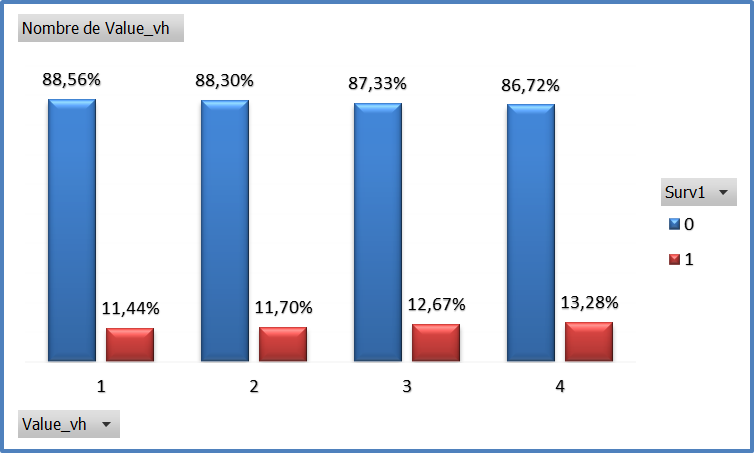
* **Répartition en pourcentage de Surv1 par rapport à Value**

Ce type de segmentation vise à avoir des classes uniformes. Le premier élément correspond au premier quantile et ainsi de suite jusqu’au dernier éléments. Par cette méthode, chaque classe est composée de 25% de l’effectif

Définition de Value\_vh :

0 si value ≤ 8380 ; 1 si 8380<value ≤ 14610.0; 2 si 14610> value ≤ 22575 ; 4 sinon

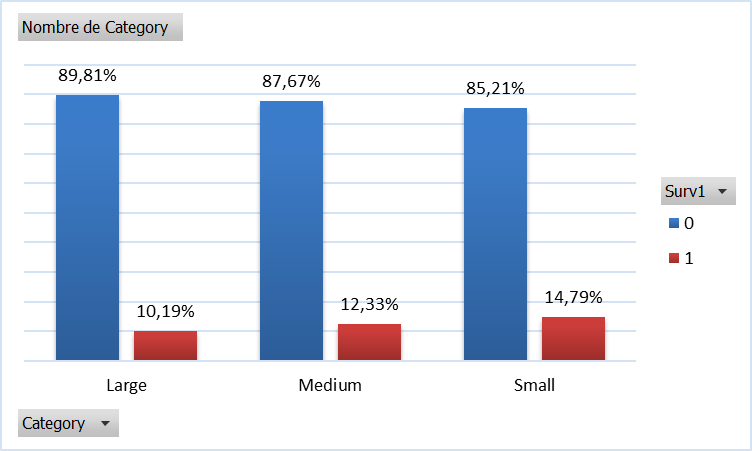
Pour étudier cette répartition, nous affichons le graphe ci-dessous



Plus la valeur du véhicule est importante plus la probabilité d’avoir sinistre augmente.

* **Répartition en pourcentage de Surv1 par rapport à Category**

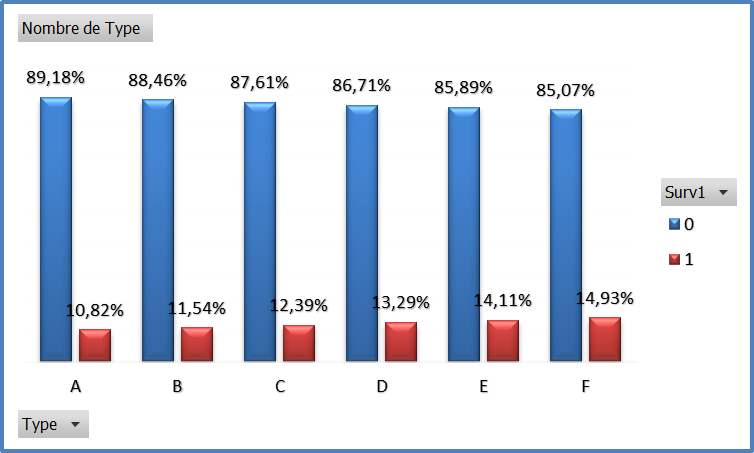
Pour étudier cette répartition, nous affichons le graphe ci-dessous



Le graphe montre que plus le véhicule est volumineux plus sa probabilité d’avoir un sinistre est plus petite. Il y’a donc un aspect apparent de dépendance entre ces modalités et surv1.

* **Répartition en pourcentage de Surv1 par rapport à Type**

Pour étudier cette répartition, nous affichons le graphe ci-dessous

****

En analysant les données du graphe nous déduisons que les véhicules de type F sont ceux qui commettent plus de sinistralité Rc matériels. Nous remarquons la fréquence de sinistres matériels croit avec les types de véhicules (A à F).

* **Répartition en pourcentage de Surv1 par rapport Poldur\_4cat**

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

Définition de Poldur\_4cat :

1 si Poldur < 1 ; 2 si 1 ≤ Poldur ≤ 5 ; 3 2 si 5<Poldur ≤ 10 ; 4 sinon

Pour étudier cette répartition, nous aﬃchons le graphe ci-dessous

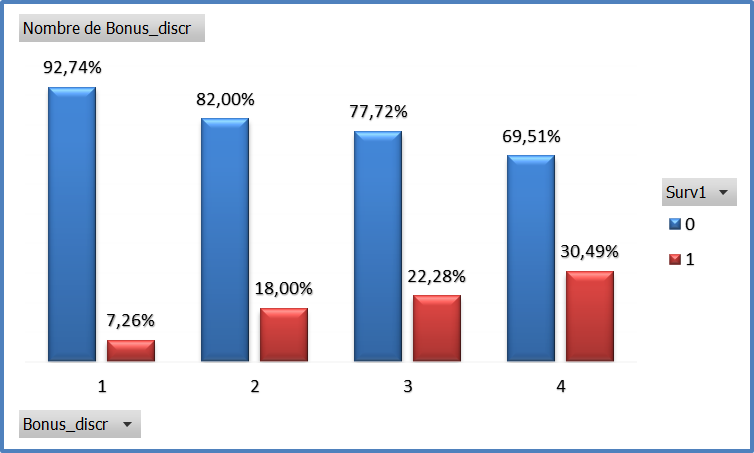
En analysant le graphe, les nouveaux contractants ont d’une part, plus tendance à commettre de sinistre. Nous déduisons que l’expérience du conducteur joue un rôle important quant à la probabilité du risque de la sinistralité.

* **Répartition en pourcentage de Surv1 par rapport à Bonus\_discr**

Définition de Bonus\_discr

1 si Bonus < 0 ; 2 si 0 ≤ Bonus < 50 ; 3 si 50 ≤ Bonus < 100 ; 4 sinon

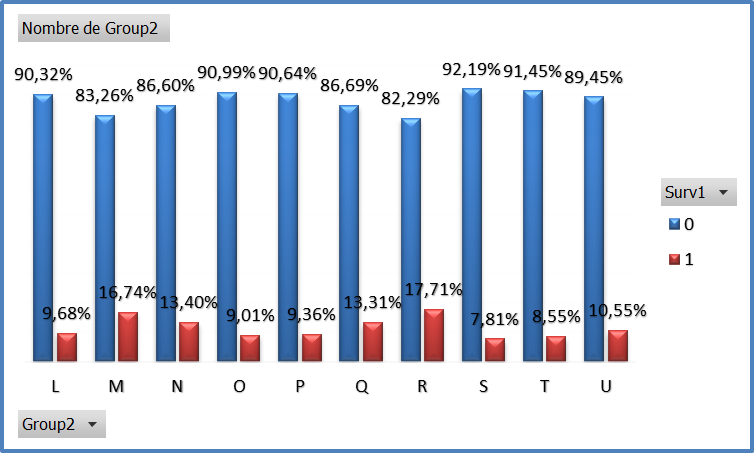
Pour étudier cette répartition, nous aﬃchons le graphe ci-dessous

****

On s’aperçoit que plus le bonus augmente plus le pourcentage de sinistre augmente.

* **Répartition en pourcentage de Surv1 par rapport à Group2**

Pour étudier cette répartition, nous affichons le graphe ci-dessous.



Nous remarquons les conducteurs appartenant à la région d’habitation M sont ceux qui commettent plus de sinistralité.

Dans la totalité des cas les p-critiques du test d’indépendance du khi deux entre la variable à expliquer et les variables qualitatives sont strictement inférieurs à 0.0001. Donc on rejette l’hypothèse d’indépendance.

Pour minimiser l’impact des grandes valeurs, les variables numériques : Density et value ont été mises en échelle logarithmique et notées ln\_density et ln\_value.

Afin d’avoir une variable offset comprise entre 0 et 1, on met la variable exppdays sous une forme **offset**= exppdays/max (exppdays²) avec max (exppdays)=365.

**I-4ANALYSE DE CORRELATION**

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 entre toutes les variables quantitatives. Cette analyse est présentée par

Les graphes ci-dessous des variables quantitatives de notre base de données.



Les variables Age et Group1 semblent être distribuées de la même façon, de même que le nombre de sinistres RC matériels et RC corporels. Ces résultats nous permettent de rejeter l’hypothèse de la dépendance entre les variables qualitatives.

Avant de s’attaquer à la modélisation de la fréquence d’accident, on modélisera à l’aide régression logistique la dichotomisation de celle-ci (surv1).

Surv1 :

* + - * 1 si la fréquence des sinistres est supérieure ou égale à 1
      * 0 sinon

**II- MODELISATION**

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, cette analyse nous ne fournit pas d’information supplémentaire sur le comportement de la variable “Surv1”.

**II-1- MODELISATION DE SURV1**

La variable surv1 prend 1 si un agent a subi au moins un accident responsable civile, 0 sinon. Elle est une dichotomisation de la variable Nb1 (nombre de sinistre responsable). Ce qui nous permet de conclure, que cette variable ne peut pas suivre la loi binomiale et la modélisation de cette variable est un problème de classification binaire supervisée. Les modèles adéquats pour ce genre de problématique sont entre autres, la régression logistique et la régression probit. Ces deux modélisations donnent des résultats à peu près similaires, dans la suite on fera le choix du modèle logit.

La régression logistique fut historiquement la première méthode utilisée, notamment en en assurance, pour modéliser la survenance ou la non survenance d’un événement binaire Y (Y=0 ou Y=1) en fonction des caractéristiques de l’assuré X = (X1, X2, X3, …, Xn). Les exemples d’utilisation de cette méthode dans d’autres domaines sont nombreux. En banque, ce modèle peut être utilisé pour déterminer l’acceptation ou le refus de crédit à un client en fonction des caractéristiques de celui-ci. Dans notre cas nous cherchons à modéliser la survenance d’un accident responsable civile à partir des informations relatives à l’assurés.

Soit Y la variable à expliquer et X = (X1, X2, X3, …, Xn) le vecteur de variables explicatives. La variable Y prend deux modalités 0 et 1. La probabilité que Y prenne la valeur 1 sachant X est et lorsqu’elle prend 0 est =1-.

On définit la variable Latente :

Où Y= 1 si et Y=0 si (s est un seuil)

Le vecteur à estimer est le vecteur dont les composants ont le même signe que ceux de .

Les p\_critiques du test de significativité globale de nos modèles sont strictement inférieurs à 0.0001, on en déduit que pour tous les modèles, les variables explicatives ont simultanément une influence sur la probabilité de survenance du sinistre et ceci pour les tests de vraisemblance, de score et de Wald.

L’un des indicateurs associés à la courbe ROC est l’AUC (Area Under the Curve). Il s’agit de l’aire comprise entre la courbe de discrimination et la première bissectrice. L’AUC correspond à la probabilité pour qu’un individu soit bien classé par le test en fonction du seuil choisi. Plus l’AUC est grand et meilleur est le modèle. Alors que le test de Hosmer-Lemeshow est basé sur une statistique qui mesure la qualité d’ajustement des modèles.

Affichons le tableau ci-dessous.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AUC | AIC | SC | -2logL |
| **Modèle 1** | 0.7513 | 73913.717 | 73923.230 | 73911.717 |
| **Modèle 2** | 0.7667 | 73913.717 | 73923.230 | 73911.717 |
| **Modèle 3** | 0.7180 | 73913.717 | 73923.230 | 73911.717 |
| **Modèle 4** | 0.7591 | 74489.087 | 74498.600 | 74487.087 |

Le test d’Hosmer-Lemeshow et l’air sous la courbe de ROC désignent que le modèle 2 comme meilleur modèle.

L’aire sous la courbe ROC ci-dessous implique une meilleure qualité ajustement et un meilleur pouvoir discriminant de notre modèle

0.00

0.25

0.50

0.75

1.00

Sensibilité

0.00

0.25

0.50

0.75

1.00

1

- Spécificité

**Courbe ROC pour le modèle sélectionné**

Aire sous la courbe = 0.7667

Les catégories de référence pour notre meilleur modèle estimé sont :

Gender :’Male’, Age\_Conduct : ‘B’, Type : ‘A’, Occupation : ’Employed’, Density\_hab : ‘1’, Group1 : ‘10’, Poldur\_4cat :’1’ et Adind :’1’.

Les variable explicatives retenues sont :

Gender, Type, Occupation, Age\_Conduct, Group1, Bonus, Density\_hab, Poldur\_4cat, Adind.

Les modalités de paramètres estimés dont les p-crtique du test de maximum de vraisemblance sont supérieurs au seuil de 5% sont :

Occupation : Self-employed (0.3580)

Type : C (0.0510)

On déduit que ses modalités ne contribuent pas à l’estimation de la probabilité de survenance du sinistre.

Les autres paramètres estimés sont significativement différents de zéro à un seuil de 5%.

**Interprétations :**

Le paramètre estimé des assurés des sexe féminin étant négatif indique que les assurées ont une propension plus petite à avoir un accident que les assurés ayant les mêmes caractéristiques.

Les paramètres estimés de la variable Poldur\_4cat étant décroissant indiquent que plus la durée de contrat augmente plus la probabilité de sinistre diminue.

Pour toutes les autres variables étant fixées, les types de véhicule B et C sont les moins disposés à avoir un sinistre et les types E et F en ont la plus grande propension.

Comme remarquer précédemment, les personnes appartenant à la classe des retraités sont les moins sujettes à subir un sinistre, consolidant ce fait, les paramètres estimés des classes d’âge indiquent que la propension à avoir un sinistre diminue lorsque l'âge augmente.

Plus la densité est élevée dans la zone d’habitation de l’agent, plus la probabilité d’avoir sinistre augmente.

Les assurés dont les véhicules sont du groupe se situe entre 10 à 20 sont plus exposés au sinistre que ceux qui sont inférieur à 10, toute chose égale par ailleurs.

Evidemment, plus le bonus est grand (plus les personnes sont en malus), plus la probabilité d’avoir un sinistre est grande pour des valeurs, pour toutes les autres variables fixées. Notons, que la variable bonus est endogène à la variable surv1.

Les assurés de profession Employed sont les moins sujettes à subir un sinistre par rapport à la profession Unemployed.

**II-2- MODELISATION DE LA FREQUENCE DE SINISTRES NB1**

La variable Surv1 étant une dichotomisation de la fréquence des sinistres, sa modélisation n’est pas suffisante à expliquer la variable Nb1. En effet, elle présente une perte d’information (incapacité de savoir le nombre de sinistre exact si surv1=1) et les résultats sont affectés par le choix de seuil(s) de coupure (dans notre cas *s=5*%)*.* D’où la nécessité de modéliser la fréquence des sinistres par des modèles de comptages (Poisson, Binomial négatif). On gardera ces valeurs explicatives pour les modélisations futures.

* **Modèle de Poisson**

Dans un modèle de Poisson, la probabilité pour qu’une variable aléatoire Y (dans notre cas Nb1) prennent la valeur yi (avec yi un entier naturel) pour un assuré i ayant les caractéristiques Xi est donnée par :

Ici =

Le paramètre a estimé est le paramètre.

Une propriété remarquable de la loi poisson est que : E (/)=V (/)*== e*x*p(B)*. Cette hypothèse dite équi-dispersion est restrictive. Dans la pratique, on a plutôt une sur dispersion de la fréquence de sinistre (la variance est plus grande que l’espérance) ce qui est causé par la surabondance de valeurs nulles.

L’utilisation alors de ce type modélisation implique souvent une sous-estimation de la variance. D’où l’intérêt d’utiliser un modèle moins restrictif, qui prendrais en compte cette sur dispersion.

* **Modèle binomiale négatif :**

Dans un modèle binomial négatif, la sur dispersion est prise en compte par un paramètre alpha.

En utilisant les mêmes notations, on peut définir par :

E (YiXi)-V (YiXi)=\*2pour =0 (équi-dispersion), on peut montrer que le modèle binomial négatif se ramène au modèle de poisson.

Soit la probabilité que la variable dépendante Yi= .

Ici, *v*=1/

* **Comparaisons des deux modèles**

Les premiers éléments permettant de comparer l’ajustement des modèles sont l’AIC le BIC, où les plus faibles sont les mieux. Ceux-ci sont utilisés pour comparer les modèles entre eux en fonction du nombre de paramètres inclus ainsi que de la taille de l’échantillon. Par contre, il est irréalisable de comparer tous les modèles entre eux, car ceux-ci ne sont pas nichés les uns dans les autres. Selon (Atkins & Gallop, 2010), le modèle de Poisson est niché dans le modèle NB alors que le ZIP est niché dans le ZINB. Il est alors possible de comparer le POISSON avec le NB et le ZIP avec le ZINB.

Nous affichons le tableau suivant

|  |  |  |
| --- | --- | --- |
|  | **Poisson** | **Binomial-**  **Négatif** |
| **AIC** | 75672.7743 | 75574.5470 |
| **BIC** | 766062.8128 | 75974.0987 |
| Gender Female | -0.3304 | -0.37199 |
| Gender Male | 0.0000 | 0.0000 |
| Age\_Conduct A | 0.8696 | 0.8713 |
| Age\_Conduct C | 0.0079 | 0.0081 |
| Age\_ Conduct D | 0.1047 | 0.1070 |
| Age\_Conduct B | 0.0000 | 0.0000 |
| Occupation Housewife | 0.2601 | 0.2595 |
| Occupation Retired | -0.9678 | -0.9677 |
| Occupation Self-employed | -0.1442 | -0.1438 |
| Occupation Unemployed | 0.11882 | 0.1859 |
| Occupation Employed | 0.0000 | 0.0000 |
| Bonus | 0.0103 | 0.0104 |
| Poldur\_4cat 2 | -0.0787 | -0.0779 |
| Poldur\_4cat 3 | -0.1901 | -0.1904 |
| Poldur\_4cat 4 | -0.3185 | -0.3157 |
| Poldur\_4cat 1 | 0.0000 | 0.0000 |
| Density\_hab 2 | -0.0826 | 0.1624 |
| Density\_hab 3 | -01901 | 0.4181 |
| Density\_hab 4 | -0.3185 | 0.8665 |
| Densty\_hab 1 | 0.0000 | 0.0000 |
| Type B | 0.0809 | 0.0792 |
| Type C | 0.1347 | 0. 1371 |
| Type D | 0.2545 | 0.2552 |
| Type E | 0.3269 | 0.3242 |
| Type F | 0.3755 | 0.3799 |
| Type A | 0.0000 | 0.0000 |

Les signes des paramètres estimés étant la même dans les deux modèles, les interprétations restent les mêmes.

Le nombre de sinistre est moins importante pour les assurés femmes par rapport aux hommes. On constate une diminution de la sinistralité avec l ’ancienneté du contrat et une augmentation avec le coefficient Bonus-malus. Les retraités ont une faible probabilité d’avoir des accidents par rapports aux autres catégories socio-professionnelles : chômeur et ménagère avec les mêmes caractéristiques. Plus la densité est élevée dans la zone d’habitation de l’agent, plus la probabilité qu’il ait des accidents responsables civile est grande, toutes choses égales par ailleurs.

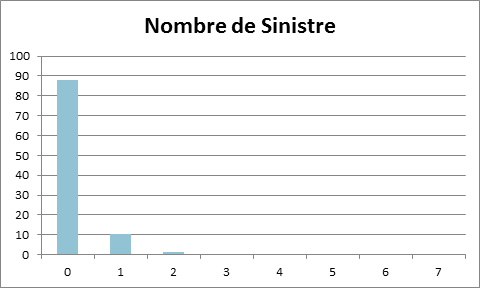
On constate aussi que les signe des paramètres estimés étant la même avec la modélisation de Surv1. Il y’a une sur dispersion dans nos données, on choisit le modèle binomial négatif au profit du modèle poisson. En effet, le modèle binomial-négatif a un AIC et un BIC plus petit que ceux du modèle de poisson. Nous pouvons donc dire que le modèle basé sur la loi binomiale négative est plus adapté à nos données.

Les modèles de comptages simples (Poisson et Binomiale négatif) supposent que les assurés ont systématiquement déclarés la survenue d’un accident. Alors, comme dit précédemment ce n’est pas toujours le cas.

**Modèles ZIP ET Modèle ZINB :**

En bref, les modèles à Inflation zéros (Zéro Inflated en anglais) sont des modèles de comptage qui sont utilisés pour modéliser les variables de comptage avec des zéros excessifs. En outre, la théorie suggère que les zéros en excès sont générés par un processus distinct des valeurs de comptage et que les zéros en excès peuvent être modélisés indépendamment.

Le graphique ci-dessous témoigne de la surabondance de zéro dans notre base de données.



En abscisse le nombre de sinistre, en ordonnées le pourcentage d’assurés.

Plus en détail, considérons la variable bi=0, l’assuré i n’as pas déclaré d’accidents et bi=1, le cas contraire. La variable qi est la probabilité de bi=0.

**Modélisation ZIP :**

La densité de la distribution Y s’écrit :

Pour y nulle, on a :

Pour y non nulle, on a :

**Modélisation ZINB**

Ici, la densité de Y s’écrit :

**Comparaison des deux modèles.**

Pour choisir, le meilleur modèle, on utilisera les critères AIC et BIC.

|  |  |  |
| --- | --- | --- |
|  | **Modèle ZI. Poisson** | **Modèle ZI Bin- Négatif** |
| **AIC** | 75561.8368 | 75576.3863 |
| **BIC** | 75970.9016 | 75994.9642 |

Le modèle de poisson à inflation de zéros est adapté pour la modélisation de nos données car ce modèle a le plus petit BIC.

**CONCLUSION**

Pour les compagnies d’assurances, la modélisation économétrique permet de limiter et contrôler le risque, ainsi plusieurs modèles sont développés afin de fournir une meilleure information pour l’aide à la décision.

Plusieurs modèles étaient testés sur la base de données sur laquelle on a travaillé afin de retenir le meilleur modèle. D’après les statistiques fournies le binomial négatif est considéré comme le meilleur modèle dans nos modélisations.