Modélisation et Analyse des Données pour une Tarification Efficace en Assurance Automobile : Une Approche Basée sur la Sélection des Risques.

Auteur : Ibrahima LY

CY Cergy Paris Université

Diplôme universitaire - Data Analyst

Rapport : Data Analyst pour l'Actuariat

Août 2023

Tuteur pédagogique :
Matthieu Cisel

Table des matières

Résumé

Dans ce rapport, nous explorons la tarification et le risque dans l'assurance automobile, en examinant les relations entre des variables telles que l'âge, le bonus-malus, et le type de véhicule. Bien que le modèle initial ait révélé certains liens significatifs, il a également montré des problèmes tels que des résidus studentisés importants et l'hétéroscédasticité. L'étude recommande l'exploration d'un modèle alternatif pour le nombre de sinistres RC matériels et souligne l'absence d'informations clés dans les données. Les résultats fournissent des *insights* pour une segmentation plus précise des clients dans l'assurance automobile et appellent à une analyse et une modélisation plus robustes.

Mots-clés : Assurance automobile, Tarification, Risque, Modélisation, Résidus studentisés, Hétéroscédasticité, Sinistres RC matériels.

Abstract

This study explores pricing and risk in car insurance, examining the relationships between variables such as age, no-claims bonus, and vehicle type. Although the initial model revealed some significant links, it also showed problems such as significant studentized residuals and heteroscedasticity. The study recommends exploring an alternative model for the number of material RC claims and highlights the absence of key information in the data. The results provide insights for more precise client segmentation in car insurance and call for more robust analysis and modeling.

Keywords : Car Insurance, Pricing, Risk, Modeling, Studentized Residuals, Heteroscedasticity, Material RC Claims.

I Introduction Générale

La modélisation et l'analyse des données sont devenues des éléments essentiels dans le domaine de l'assurance automobile en France. Avec une concurrence croissante et la nécessité de comprendre les risques liés aux conducteurs, les assureurs font face au défi constant de proposer des tarifs adaptés à la situation individuelle de chaque assuré. La collecte exhaustive de données pertinentes et une analyse rigoureuse pour en extraire les informations clés sont indispensables. Par ailleurs, l'intégration de méthodes analytiques avancées, telles que l'apprentissage automatique et les algorithmes prédictifs, peut améliorer la compétitivité des compagnies d'assurances tout en offrant un service personnalisé à leurs clients.

Dans ce contexte dynamique et complexe, Cette analyse vise à démontrer comment la modélisation et l'analyse des données peuvent optimiser les pratiques tarifaires en assurance automobile, en mettant l'accent sur l'équilibrage de la rentabilité économique pour les assureurs avec la protection adéquate pour les assurés. En analysant le coût total des sinistres RC matériels pour 100 000 polices d'assurance automobile pour les années 2009 et 2010, cette étude cherche à développer un modèle prédictif pour proposer des primes annuelles équilibrées pour l'année 2011.

Le rapport est organisé en segments essentiels qui offrent une progression logique à travers le sujet de recherche. Il commence par une introduction pour établir le cadre de l'étude, suivi d'une description précise de la méthodologie. Avec une exactitude méthodologique, les résultats obtenus sont ensuite exposés, préparant le terrain pour une discussion détaillée et une réflexion sur les implications pratiques et théoriques de la recherche. La conclusion résume les découvertes principales et suggère des avenues possibles pour des études ultérieures.

En somme, cette étude contribue au domaine de l'assurance automobile en France en explorant les possibilités offertes par les techniques de modélisation et d'analyse de données modernes. La problématique centrale, de comprendre et de quantifier les risques pour une tarification efficace, guide l'ensemble de l'effort de recherche, avec des implications significatives pour les compagnies d'assurance, les assurés, et le secteur dans son ensemble.

Description des Variables

PolNum : numéro de police

CalYear : année calendaire de souscription

Gender : genre du conducteur

Type : type de véhicule

Category : catégorie du véhicule

Occupation : profession

Age : âge du conducteur

Group1 : groupe du véhicule

Bonus : Bonus Malus

Poldur : Ancienneté du contrat

Value : Valeur du véhicule

Adind : Indicateur d'une garantie dommages

SubGroup2 : Sous-région d'habitation

Group2 : Région d'habitation

Density : Densité de population

Expdays : Exposition (en jours)

nb_sin : Nombre de sinistres RC matériels

chg_sin : Coût total des sinistres RC matériels

Description des Bases de Données

[1] Deux jeux de données sont disponibles en ligne :

— http://freakonometrics.free.fr/training.csv

— http://freakonometrics.free.fr/pricing.csv

1. La base training est constituée de 100,000 polices différentes, pour les années 2009 et 2010, et est composée des variables mentionnées ici.

II Méthodologie

La section méthodologie constitue l'épine dorsale de notre étude, détaillant minutieusement les processus et les étapes indispensables à l'analyse et à la modélisation des données relatives au domaine de l'assurance automobile. Cette composante se subdivise en trois sections principales, chacune abordant un aspect essentiel de notre recherche.
Tout d'abord, le volet Manipulation et Prétraitement des Données revêt une importance capitale puisqu'il englobe le traitement initial des données afin d'en garantir leur intégrité ainsi que leur qualité avant toute analyse subséquente. Ensuite, nous explorons dans la section Modélisation diverses techniques et méthodes qui ont été employées pour parvenir à tirer des conclusions pertinentes à partir des données préalablement traitées

1 Présentation et Traitement des Données

Présentation des Jeux de Données :

a) Base de Portefeuille (base_ptf) : Le premier jeu de données, appelé "Base de Portefeuille", contient 100 027 entrées avec un total de 15 colonnes. Ces colonnes couvrent diverses informations telles que le numéro de police, l'année civile, le genre, le type, la catégorie, l'occupation, l'âge, et d'autres caractéristiques pertinentes pour le portefeuille d'assurance.

b) Base des Sinistres (base_sin) : Le deuxième jeu de données, nommé "Base des Sinistres", se compose de 13 300 entrées et trois colonnes. Ces colonnes représentent le Nombre de sinistres RC matériels (nb_sin), Coût total des sinistres RC matériels (chg_sin), et le numéro de police (PolNum).

c) Base d'Exposition (base_expo) : Le troisième jeu de données, que nous appellerons "Base d'Exposition", contient 100 021 entrées et deux colonnes. Les colonnes incluent le numéro de police (PolNum) et le nombre de jours d'exposition (Expdays). L'exposition peut souvent faire référence à la mesure du risque ou à la quantité de temps durant laquelle le risque est évalué.

1.1 Élimination des Doublons

A- Base de Portefeuille (base_ptf) :

La base de Portefeuille initiale contenait 100 027 enregistrements répartis sur 15 colonnes. Après avoir supprimé les doublons basés sur la clé primaire, 27 enregistrements en double ont été supprimés, laissant un total de 100 000 enregistrements uniques.

1.2 Gestion des Valeurs Manquantes

Colonne	Nombre de valeurs manquantes	Pourcentage de valeurs manquantes
SubGroup2	88406	88.41
Value	785	0.78
Gender	5	0.00
PolNum	0	0.00
CalYear	0	0.00
Type	0	0.00
Category	0	0.00
Occupation	0	0.00
Age	0	0.00
Group1	0	0.00
Bonus	0	0.00
Poldur	0	0.00
Adind	0	0.00
Group2	0	0.00
Density	0	0.00

TABLE II.1 – **Nombre et pourcentage de valeurs manquantes par colonne**

Dans notre base de Portefeuille (base_ptf) , la plupart des colonnes sont complètes, à l'exception de **'Gender', 'Value' et 'SubGroup2'**. 'Gender' a très peu de valeurs manquantes, alors que 'Value' en compte environ **0.78%**. La colonne 'SubGroup2' est la plus affectée avec **88.41%** de valeurs manquantes, Cette variable n'apporte pas d'informations pertiantes. Donc on peut la supprimer dans le dataset.

Suppression de la Variable SubGroup2 : La variable *SubGroup2*, qui représente les sous-régions d'habitation, a été supprimée de l'ensemble de données en raison d'un nombre élevé de valeurs manquantes (88.41% soit 88406 NaN sur un total de 100026). Étant donné que la variable *Group2* fournit déjà des informations sur la région d'habitation, le maintien de *SubGroup2* aurait été redondant. De plus, l'imputation d'un si grand nombre de valeurs manquantes aurait pu introduire un biais, tandis que la suppression aurait réduit significativement l'information. Par conséquent, la suppression de cette variable était la démarche la plus prudente pour préserver la qualité et l'intégrité des données.

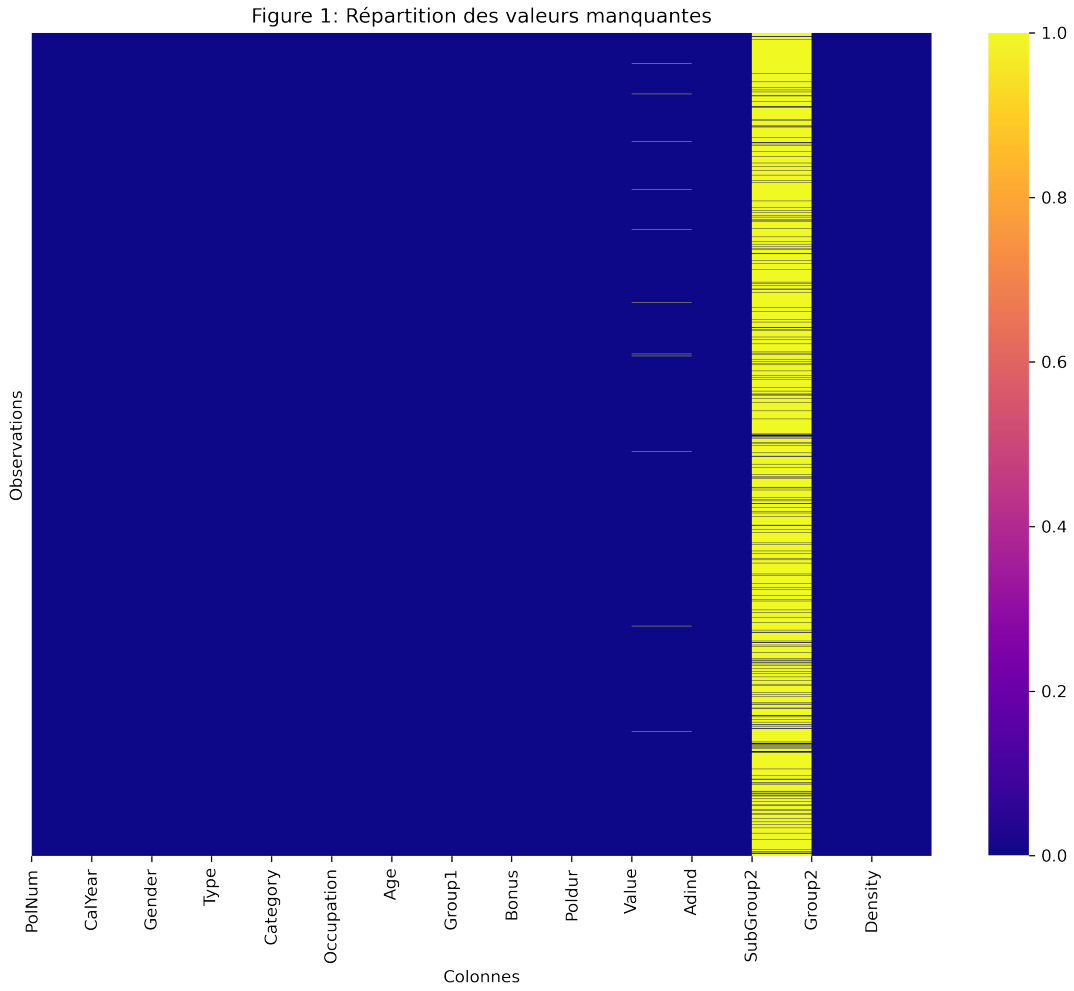


FIGURE II.1 – **Répartition des valeurs manquantes**

Imputation dans la Colonne *Value* : La colonne *Value* a nécessité plusieurs étapes de traitement. Les valeurs non numériques ont été identifiées et remplacées par NaN. Ensuite, ces valeurs manquantes ont été imputées par la médiane de la colonne. Utiliser la médiane minimise l'impact des valeurs aberrantes, préservant la distribution originale de la variable. Cette méthode d'imputation est en accord avec les bonnes pratiques en data science, et offre une solution robuste pour gérer les valeurs manquantes sans introduire de biais.

Suppression des NaN dans la Colonne *Gender* : La colonne *Gender* contenait 5 valeurs manquantes, soit 0.005% de l'ensemble de données. En raison de cette faible proportion, la suppression de ces lignes a été préférée à l'imputation. Cette décision minimise l'introduction de biais potentiel, préservant la validité et l'intégrité de l'analyse. La perte d'informations est négligeable, et cette suppression est en accord avec les principes de data science, qui mettent l'accent sur l'exactitude et la rigueur dans le

traitement des données.

Colonne	Nombre de valeurs manquantes
PolNum	0
CalYear	0
Gender	0
Type	0
Category	0
Occupation	0
Age	0
Group1	0
Bonus	0
Poldur	0
Value	0
Adind	0
Group2	0
Density	0

TABLE II.2 – Nombre de valeurs manquantes après traitement

1.3 Détection des Données Incohérentes

Age Statistiques descriptives pour la variable *Age* :

Statistique	Valeur
Count	100022.000000
Mean	41.123463
Std	14.315898
Min	4.000000
25%	30.000000
50%	40.000000
75%	51.000000
Max	250.000000

TABLE II.3 – Statistiques descriptives pour la variable *Age*.

Value Statistiques descriptives pour la variable *Value* :

Statistique	Valeur
Count	100022.000000
Mean	16440.235398
Std	10466.567424
Min	1000.000000
25%	8410.000000
50%	14610.000000
75%	22515.000000
Max	49995.000000

TABLE II.4 – Statistiques descriptives pour la variable *Value*.

Density Statistiques descriptives pour la variable *Density* :

Statistique	Valeur
Count	100022.000000
Mean	117.160264
Std	79.500672
Min	14.377142
25%	50.625783
50%	94.364623
75%	174.644525
Max	297.385170

TABLE II.5 – Statistiques descriptives pour la variable *Density*.

a) Visualisation des valeurs Abérentes :

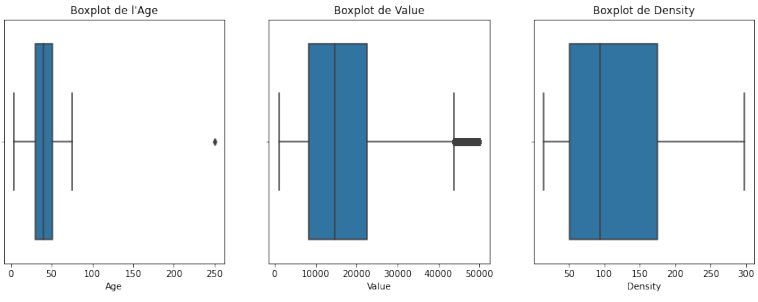


FIGURE II.2 – Boxplot de l'Age, de Value, et de Density

b) Visualisation des valeurs Abérentes :

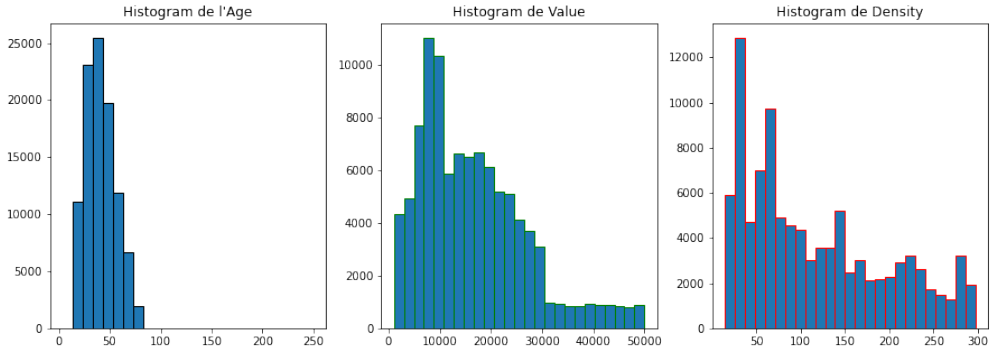


FIGURE II.3 – Histogrammes de l'Age, de Value, et de Density

Genre du Conducteur : La variable *Genre* comprend des valeurs incohérentes comme "H", "F", et "h", ce qui suggère une erreur de saisie. Ces anomalies doivent être corrigées pour assurer la cohérence de l'analyse.

Genre	Nombre
Male	63437
Female	36574
H	5
F	5
h	1

TABLE II.6 – Répartition des valeurs pour la variable *Genre*

Type de Véhicule : La variable *Type* semble cohérente sans valeurs aberrantes évidentes. Les catégories A, B, C, D, E, et F représentent différents types de véhicules.

Type	Nombre
A	27760
B	22090
D	19597
C	13864
E	11170
F	554

TABLE II.7 – Répartition des valeurs pour la variable *Type*

Catégorie du Véhicule : La variable *Catégorie* contient une valeur aberrante "???", qui représente probablement des données manquantes ou une erreur de saisie. Une investigation supplémentaire sera nécessaire pour déterminer comment traiter cette valeur.

Catégorie	Nombre
Medium	36253
Large	32022
Small	31718
???	29

TABLE II.8 – Répartition des valeurs pour la variable *Catégorie*

***Occupation (Profession)* :** La variable *Occupation* représente la profession du conducteur et semble cohérente avec différentes catégories telles qu'Employed, Self-employed, Housewife, Unemployed, et Retired.

Occupation	Nombre
Employed	31149
Self-employed	20372
Housewife	20012
Unemployed	15322
Retired	13167

TABLE II.9 – Répartition des valeurs pour la variable *Occupation*

***Group2 (Région d'Habitation)* :** La variable *Group2* indique la région d'habitation avec plusieurs codes. Aucune anomalie n'est apparente dans cette variable.

Group2	Nombre
L	23730
Q	22389
R	15081
M	7596
U	5365
P	5259
O	5216
T	5197
N	5195
S	4994

TABLE II.10 – Répartition des valeurs pour la variable *Group2*

***Adind (Indicateur d'une Garantie Dommages)* :** La variable *Adind* est une variable dichotomique indiquant la présence (1) ou l'absence (0) d'une garantie dommages. Les valeurs sont cohérentes et ne présentent pas d'anomalies.

Adind	Nombre
1	51225
0	48797

TABLE II.11 – Répartition des valeurs pour la variable *Adind*

1.4 Correction des Données Incohérentes

Variable 'Age' :

Les valeurs aberrantes pour la variable 'Age' ont été identifiées comme étant en dehors de l'intervalle [18, 80]. Nous avons choisi de supprimer ces valeurs pour obtenir une distribution plus représentative de l'âge des conducteurs.

TABLE II.12 – Statistiques descriptives pour l'âge du conducteur

Statistique	Valeur
count	100019
mean	41.122057
std	14.300049
min	18
25%	30
50%	40
75%	51
max	75

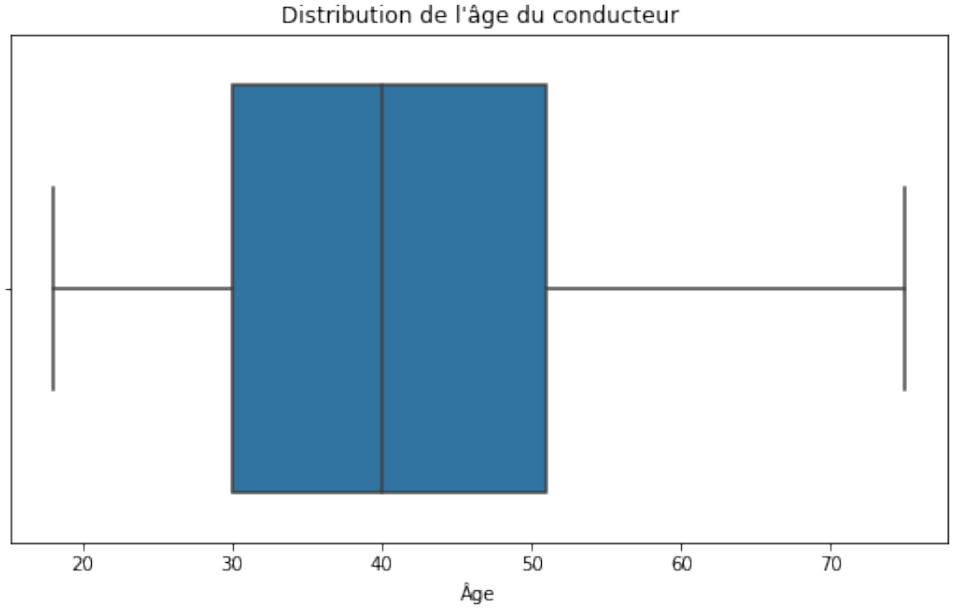


FIGURE II.4 – Distribution de l'âge du conducteur

Variable 'Gender' :

Pour la variable 'Gender', nous avons identifié des valeurs aberrantes telles que 'H', 'F', et 'h'. Ces valeurs ont été remplacées par les valeurs correspondantes 'Male' et 'Female'.

TABLE II.13 – Répartition finale du genre

Genre	Compte
Male	63441
Female	36578

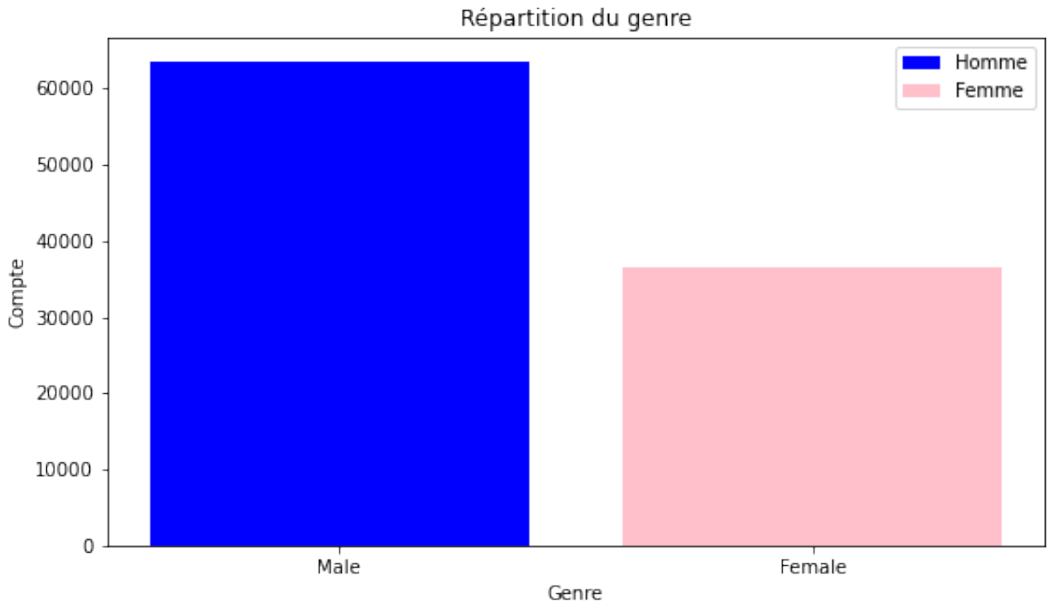


FIGURE II.5 – Répartition du genre

Variable 'Category' :

La correction des valeurs incohérentes pour la variable 'Category' a impliqué l'identification et le remplacement de 29 occurrences de "???" en utilisant l'imputation aléatoire basée sur les proportions existantes des catégories.

TABLE II.14 – Distribution des catégories de véhicules

Catégorie	Compte
Medium	36265
Large	32028
Small	31726

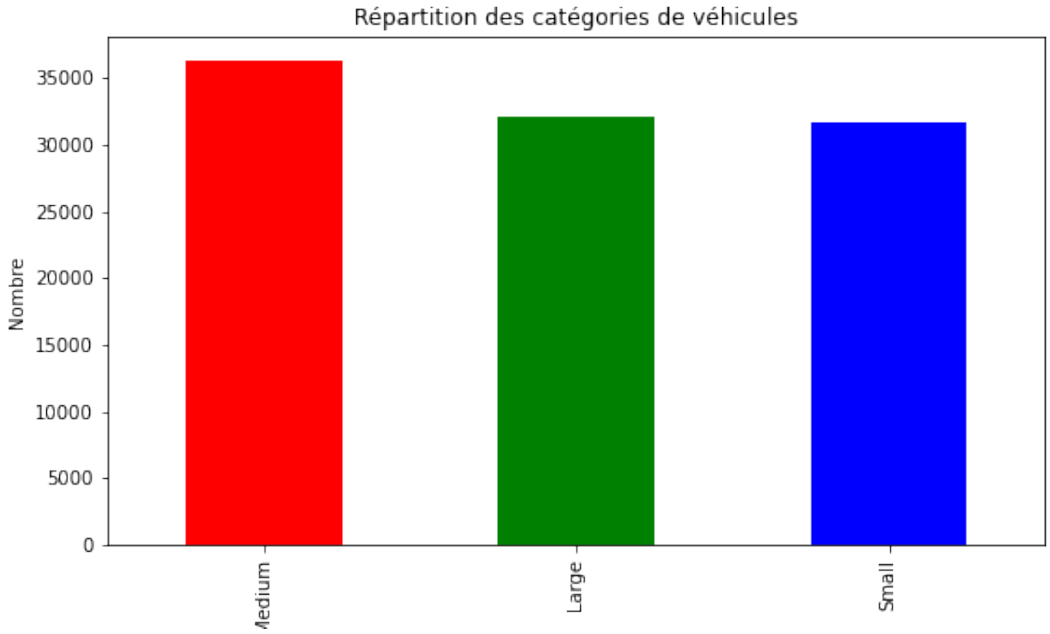


FIGURE II.6 – Répartition des catégories de véhicules

Traitement de la Base Sinistre (base_sin) :

La base de données des sinistres est cruciale pour l'analyse et nécessite un prétraitement pour garantir l'exactitude et la pertinence des informations. Le processus suivant a été appliqué :

Vérification des Doublons : La première étape a consisté à identifier les doublons dans la colonne `PolNum`. Un total de 99 doublons a été trouvé, représentant 197 lignes. Ces doublons ont été consolidés en sommant les valeurs correspondantes de `nb_sin` et `chg_sin` pour chaque numéro de police unique.

Vérification des Valeurs NaN : L'étape suivante a été la vérification de l'existence de valeurs NaN (Not a Number) dans la base de données. Aucune valeur NaN n'a été trouvée, garantissant ainsi que toutes les entrées sont complètes et prêtes pour l'analyse.

Résultat du Prétraitement : Après ces étapes de prétraitement, la base de données des sinistres est prête pour l'analyse ultérieure. La structure et la qualité des données ont été affinées pour refléter avec précision l'information sur les sinistres, en éliminant les ambiguïtés et en assurant une représentation fidèle de la réalité.

Traitement de la Base Exposition (base_Expo) :

La base de données Exposition contient des informations cruciales sur les numéros de police et les jours d'exposition. Un nettoyage minutieux a été effectué pour assurer l'exactitude des données :

— **Vérification des Doublons :** La colonne `PolNum` a été inspectée pour détecter des doublons. 21 doublons ont été identifiés et supprimés, laissant un total de 100 000 lignes uniques dans la base de données.

— **Vérification des Valeurs NaN :** Un contrôle a été effectué pour s'assurer qu'il n'y a pas de valeurs NaN dans la base de données. Aucune valeur NaN n'a été trouvée, ce qui garantit que toutes les entrées sont complètes.

— **Analyse des Données Incohérentes :** Une vérification supplémentaire des données incohérentes et aberrantes peut être effectuée si nécessaire. Par exemple, s'assurer que toutes les valeurs dans la colonne `Expdays` sont positives et cohérentes avec les attentes du domaine.

— **Résultat du Prétraitement :** Après ces étapes, la base de données Exposition est prête pour l'analyse ultérieure, toutes les ambiguïtés étant éliminées et les données représentant fidèlement la réalité.

1.5 Fusion des Bases de Données

Fusion des Bases de Données :

La fusion des bases de données a été une étape cruciale dans la préparation des données pour l'analyse. Après le nettoyage et la vérification des trois bases de données distinctes :

La base du portefeuille, la base d'exposition, et la base des sinistres, elles ont été fusionnées en une base de données unique en utilisant la colonne identifiant les numéros de police.

La démarche a été réalisée en deux étapes majeures :

— **Première Jointure :** La première étape a consisté en une fusion de la base de données du portefeuille avec celle de l'exposition. Cette jointure interne a été effectuée sur la colonne contenant les numéros de police, assurant que seules les lignes correspondantes de chaque base de données ont été combinées.

— **Deuxième Jointure :** La base de données résultante de la première jointure a ensuite été fusionnée avec la base de données des sinistres, également par une

jointure interne sur la même colonne. Ceci garantit que les informations sur les sinistres sont correctement alignées avec les autres données.

La base de données finale ainsi obtenue contient toutes les informations nécessaires pour l'analyse ultérieure. Cette démarche assure que toutes les données sont alignées correctement selon les numéros de police, et que toutes les informations pertinentes sont disponibles dans un format unifié et cohérent.

2 Modélisation Statistique

2.1 Analyse Descriptive des Variables :

2.1.1 Variables Numériques : Âge, Bonus-Malus, Valeur du Véhicule

Statistique	Age	Bonus	Value
Compte	12277.0	12277.0	12277.0
Moyenne	34.972	22.283	17088.836
Écart-type	13.298	57.691	10721.574
Min	18.0	-50.0	1005.0
25%	24.0	-20.0	8625.0
Médiane	32.0	0.0	15410.0
75%	44.0	60.0	23225.0
Max	75.0	150.0	49995.0
Skewness	0.847	0.655	0.919

TABLE II.15 – Statistiques descriptives des variables quantitatives

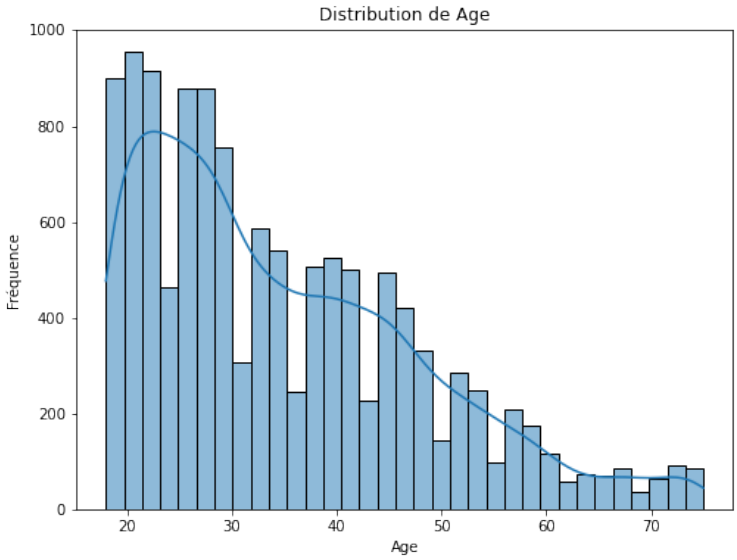


FIGURE II.7 – Distribution de l'âge

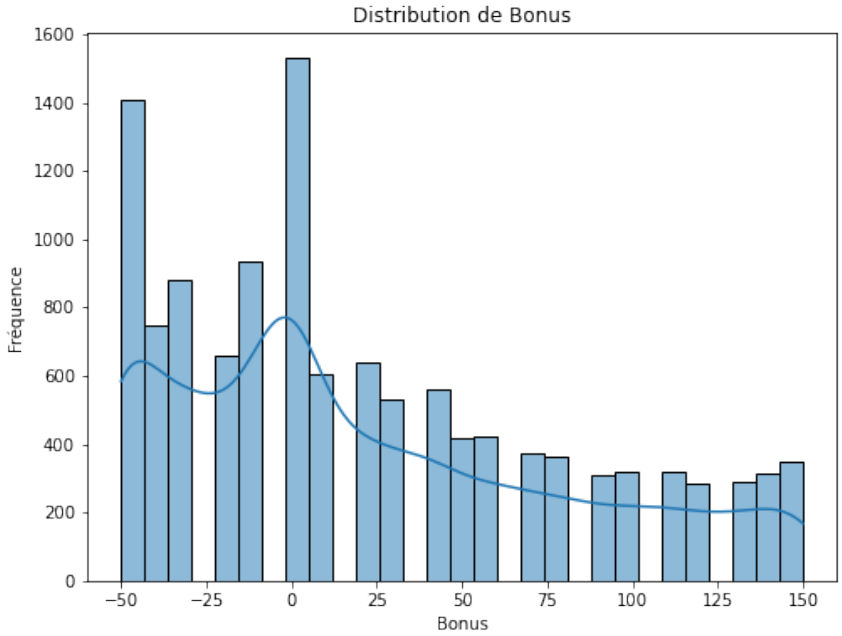


FIGURE II.8 – Distribution du bonus

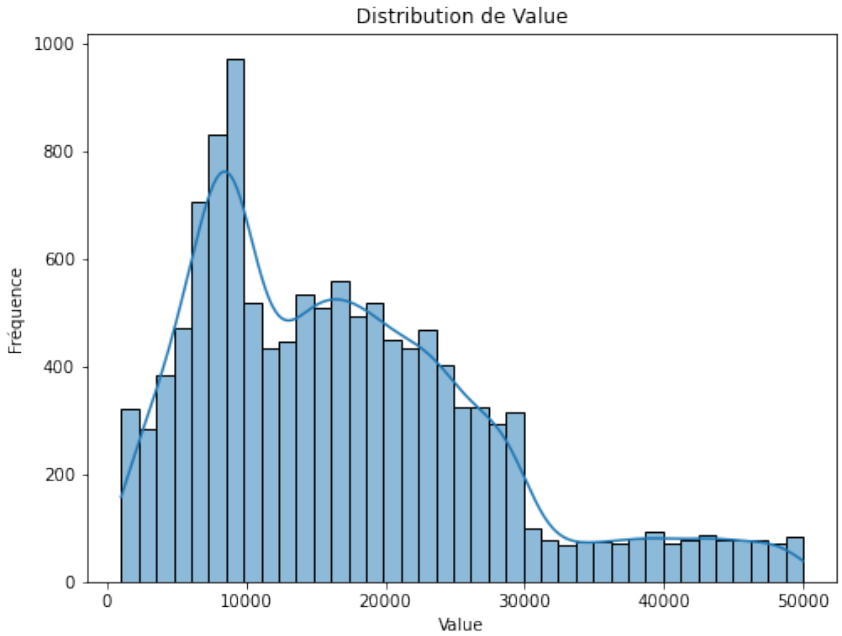


FIGURE II.9 – Distribution de la valeur

2.1.2 Catégorielles : Genre, Type de Véhicule, Occupation

Type	Fréquence	Pourcentage
A	0.244604	24.460373
D	0.212104	21.210393
B	0.207624	20.762401
C	0.139936	13.993647
E	0.128370	12.837012
F	0.067362	6.736173

TABLE II.16 – Table de fréquence pour Type

Category	Fréquence	Pourcentage
Medium	0.358475	35.847520
Large	0.343406	34.340637
Small	0.298118	29.811843

TABLE II.17 – Table de fréquence pour Category

Occupation	Fréquence	Pourcentage
Employed	0.350737	35.073715
Housewife	0.238087	23.808748
Unemployed	0.203307	20.330700
Self-employed	0.170074	17.007412
Retired	0.037794	3.779425

TABLE II.18 – Table de fréquence pour Occupation

Gender	Fréquence	Pourcentage
Male	0.680541	68.054085
Female	0.319459	31.945915

TABLE II.19 – Table de fréquence pour Gender

Variable	Statistique	p-valeur
Age	0.9258	0.0
Bonus	0.9194	0.0
Value	0.9327	0.0

TABLE II.20 – Résultats du test de Shapiro-Wilk pour les variables continues

Évaluation de la Distribution des Variables Continue : Les distributions des variables continues ont été évaluées à l'aide du test de Shapiro-Wilk pour tester la normalité. Les résultats du test sont présentés dans ce tableau II.20.

2.2 Tableau de Corrélation entre Variables Numériques

Le tableau ci-dessous montre les coefficients de corrélation entre les variables numériques Age, Bonus, Value, Group1, Density, et Expdays. Ces coefficients mesurent la force et la direction de la relation linéaire entre les paires de variables.

	Age	Bonus	Value	Group1	Density	Expdays
Age	1.0000	-0.0489	0.0353	0.1207	0.0045	0.0165
Bonus	-0.0489	1.0000	-0.0151	-0.0220	-0.0317	-0.0134
Value	0.0353	-0.0151	1.0000	0.2232	0.0101	-0.0116
Group1	0.1207	-0.0220	0.2232	1.0000	-0.0124	-0.0096
Density	0.0045	-0.0317	0.0101	-0.0124	1.0000	-0.0176
Expdays	0.0165	-0.0134	-0.0116	-0.0096	-0.0176	1.0000

TABLE II.21 – Coefficients de corrélation entre les variables numériques

Ensuite, nous allons utiliser des tests statistiques pour identifier les relations entre les caractéristiques des conducteurs et des véhicules et le bonus/Malus en assurance. Plus précisément, nous avons réalisé un test ANOVA (Analyse of Variance) pour comparer la moyenne du bonus entre les différentes catégories des variables Type, Category, et Gender.

ANOVA pour Type

Nous avons comparé le bonus entre les différents types de véhicules. Le test ANOVA a produit les résultats suivants :

— Valeur de F : $F = 3.1214$

— Valeur de p : $p = 0.0442$

ANOVA pour Category

Nous avons comparé le bonus entre les différentes catégories de véhicules. Les résultats sont les suivants :

— Valeur de F : $F = 1.0743$

— Valeur de p : $p = 0.3416$

ANOVA pour Gender

Nous avons comparé le bonus entre les genres de conducteurs. Les résultats sont les suivants :

— Valeur de F : $F = 6.9856$

— Valeur de p : $p = 0.0082$

Test du Chi-carré

Pour examiner la relation entre les variables catégorielles, nous avons utilisé le test du Chi-carré. Ce test statistique non paramétrique a été appliqué pour les combinaisons suivantes de variables :

1. Genre et Type
2. Genre et Catégorie
3. Type et Catégorie

Les résultats des tests de chi-carré sont présentés dans ce Tableau 2.2.

Combinaison	χ^2 Valeur	Valeur p
Genre et Type	1.3086	0.9340
Genre et Catégorie	2.0736	0.3546
Type et Catégorie	9.7337	0.4642

TABLE II.22 – Résultats du test du Chi-carré

2.3 Modèles de Tarification par Régression Linéaire Multiple

2.3.1 Encodage des Variables Catégorielles

Dans le processus de préparation des données pour la modélisation statistique, il est essentiel de convertir les variables catégorielles en un format qui puisse être fourni aux algorithmes de modélisation. Pour ce faire, nous avons utilisé l'encodage à chaud, également connu sous le nom de *one-hot encoding*, sur les variables catégorielles suivantes : Gender, Type, Category, Occupation, et Group2.

L'encodage à chaud est une représentation binaire des variables catégorielles. Pour chaque catégorie unique dans une variable catégorielle, une nouvelle colonne binaire est créée, où la présence de la catégorie est représentée par un 1, et l'absence par un 0. Par exemple, pour la variable *Gender*, deux nouvelles colonnes seraient créées, une pour "Homme" et l'autre pour "Femme". Si un individu est un homme, la colonne correspondante aurait une valeur de 1, et la colonne pour "Femme" aurait une valeur de 0.

Cette méthode d'encodage permet aux modèles de comprendre et d'utiliser l'information contenue dans les variables catégorielles sans faire d'hypothèse sur l'ordinalité des catégories. L'encodage à chaud est particulièrement utile pour les variables catégorielles où les catégories n'ont pas d'ordre intrinsèque, comme c'est le cas pour les variables catégorielles dans notre ensemble de données.

Cette étape de transformation des données est cruciale pour l'analyse ultérieure, et permet d'appliquer des techniques de régression linéaire aux données, en tenant compte des effets des variables catégorielles sur le coût total des sinistres RC matériels.

2.3.2 Modèle de régression lineaire

Dans cette étude, un modèle de régression linéaire a été utilisé pour étudier les relations entre la variable dépendante Coût total des sinistres RC matériels (chg_sin) et plusieurs variables indépendantes. L'objectif est de déterminer l'effet de chaque variable indépendante sur la variable dépendante tout en contrôlant les autres variables.

Variable	Valeur
Dépendante	chg_sin
R-squared	0.080
Modèle	OLS
R-squared ajusté	0.078
Méthode	Least Squares
F-statistic	35.63
Date	Wed, 23 Aug 2023
Prob (F-statistic)	1.67e-196
Heure	11 :56 :46
Nombre d'observations	12277
AIC	2.031e+05
Df résidus	12246
BIC	2.033e+05
Df modèle	30
Type de covariance	nonrobust

TABLE II.23 – Détails de la régression OLS

Test	Valeur
Omnibus	7582.713
Durbin-Watson	1.980
Prob(Omnibus)	0.000
Jarque-Bera (JB)	119391.728
Skew	2.705
Kurtosis	17.287
Condition Number	1.18e+16

TABLE II.24 – Statistiques de diagnostic de la régression OLS

Variable	Coefficient	Std Err	t-value	P>|t|	[0.025	0.975]
const	-5.091e+04	1.49e+04	-3.412	0.001	-8.02e+04	-2.17e+04
CalYear	58.6476	17.076	3.435	0.001	25.176	92.119
Age	-11.6510	0.747	-15.593	0.000	-13.116	-10.186
Group1	9.9547	1.935	5.145	0.000	6.162	13.747
Bonus	-0.0421	0.149	-0.284	0.777	-0.333	0.249
Poldur	-4.9726	1.898	-2.620	0.009	-8.692	-1.253
Value	0.0014	0.002	0.875	0.382	-0.002	0.004
Adind	-133.1689	17.769	-7.494	0.000	-168.000	-98.338
Density	2.3251	0.285	8.147	0.000	1.766	2.885
Expdays	0.7093	0.152	4.679	0.000	0.412	1.006
Gender_Female	-2.551e+04	7459.494	-3.419	0.001	-4.01e+04	-1.09e+04
Gender_Male	-2.54e+04	7459.540	-3.405	0.001	-4e+04	-1.08e+04
Type_A	-8372.2217	2486.641	-3.367	0.001	-1.32e+04	-3498.013
Type_B	-8396.9214	2486.573	-3.377	0.001	-1.33e+04	-3522.846
Type_C	-8483.9585	2486.487	-3.412	0.001	-1.34e+04	-3610.052
Type_D	-8538.3519	2486.640	-3.434	0.001	-1.34e+04	-3664.145
Type_E	-8541.8135	2486.620	-3.435	0.001	-1.34e+04	-3667.647
Type_F	-8575.5476	2486.591	-3.449	0.001	-1.34e+04	-3701.437
Category_Large	-1.699e+04	4973.076	-3.417	0.001	-2.67e+04	-7244.500
Category_Medium	-1.699e+04	4973.069	-3.416	0.001	-2.67e+04	-7240.877
Category_Small	-1.693e+04	4972.988	-3.404	0.001	-2.67e+04	-7179.585
Occupation_Employed	-1.034e+04	2983.795	-3.467	0.001	-1.62e+04	-4496.150
Occupation_Housewife	-1.042e+04	2983.909	-3.492	0.000	-1.63e+04	-4569.649
Occupation_Retired	-9764.8001	2984.011	-3.272	0.001	-1.56e+04	-3915.667
Occupation_Self-employed	-1.027e+04	2983.813	-3.443	0.001	-1.61e+04	-4424.772
Occupation_Unemployed	-1.011e+04	2983.982	-3.387	0.001	-1.6e+04	-4257.981
Group2_L	-5095.9380	1492.419	-3.415	0.001	-8021.315	-2170.561
Group2_M	-5098.6055	1491.728	-3.418	0.001	-8022.628	-2174.583
Group2_N	-4960.5172	1491.653	-3.326	0.001	-7884.393	-2036.642
Group2_O	-5016.4561	1492.621	-3.361	0.001	-7942.230	-2090.683
Group2_P	-4998.9972	1493.486	-3.347	0.001	-7926.464	-2071.530
Group2_Q	-5211.9526	1491.912	-3.493	0.000	-8136.336	-2287.570
Group2_R	-5248.9530	1491.902	-3.518	0.000	-8173.316	-2324.590
Group2_S	-5040.2841	1493.449	-3.375	0.001	-7967.679	-2112.889
Group2_T	-5107.4983	1492.778	-3.421	0.001	-8033.579	-2181.417
Group2_U	-5129.6127	1492.349	-3.437	0.001	-8054.852	-2204.374

TABLE II.25 – Coefficients et statistiques de la régression OLS

Dans notre analyse, nous avons utilisé une régression linéaire ordinaire (OLS) pour examiner la relation entre le coût total des sinistres RC matériels (variable dépendante `chg_sin`) et diverses variables indépendantes, y compris l'âge, le type de véhicule, la catégorie, l'occupation, et la région d'habitation.

Le choix d'un modèle OLS est motivé par sa facilité d'interprétation et son utilisation courante dans l'analyse des relations linéaires.

Variables explicatives La variable dépendante, `chg_sin`, est expliquée par les va-

riables suivantes :

— **CalYear**, **Age**, **Group1**, **Bonus**, **Poldur**, **Value**, **Adind**, **Density**, **Expdays** : Variables continues.

— **Gender**, **Type**, **Category**, **Occupation**, **Group2** : Variables catégorielles.

Tests statistiques Nous avons effectué divers tests pour évaluer la qualité du modèle, notamment le test F pour l'ensemble du modèle, les tests t pour les coefficients individuels, et des tests tels que Jarque-Bera pour examiner la normalité des résidus.

2.3.3 Retirer les Variables Non Significatives et Refaire la Régression

Dans le modèle initial, deux variables, à savoir 'Bonus' et 'Value', ont été identifiées comme non significatives. Leur p-value était supérieure au seuil choisi pour la significativité, soit 5%, indiquant que ces variables n'avaient pas de lien statistiquement significatif avec le nombre de sinistres RC matériels (chg_sin). Par conséquent, ces variables ont été retirées du modèle pour améliorer l'interprétation et l'efficacité du modèle. (voir Figure III.1)

2.4 Diagnostics du Modèle de Régression MCO

Dans cette étude, nous utilisons le modèle de régression des moindres carrés ordinaires (MCO) pour analyser les données. Pour diagnostiquer ce modèle, nous effectuons plusieurs tests.

2.4.1 Détection de la Normalité des Résidus

Points à fort effet de levier : Nous identifions les points à fort effet de levier en utilisant la diagonale de la matrice chapeau. Les points avec des valeurs supérieures à un seuil spécifique (dans notre cas, 0.05) sont considérés comme ayant un fort effet de levier.

Résidus Studentisés : Nous identifions également les résidus importants en utilisant les résidus studentisés externes, et nous considérons les résidus dont la valeur absolue est supérieure à 2 comme significatifs.

Test de Normalité des Résidus : Enfin, nous utilisons le test de Shapiro-Wilk pour vérifier la normalité des résidus de notre modèle.

2.4.2 Détection de l'Homoscédasticité

L'homoscédasticité est une des hypothèses clés dans la régression linéaire. Elle stipule que la variance des erreurs est constante à travers tous les niveaux de variables

explicatives. En d'autres termes, la dispersion des résidus doit être approximativement la même pour toutes les valeurs prédites.

Pour tester cette hypothèse, nous utilisons à la fois des méthodes graphiques et statistiques. Graphiquement, nous traçons les résidus en fonction des valeurs prédites et recherchons une dispersion constante. Statistiquement, nous utilisons le test de Breusch-Pagan pour vérifier si la variance des résidus est constante ou non.

2.4.3 Lignes d'Ajustement

L'analyse des résidus est une étape essentielle dans l'évaluation de l'ajustement d'un modèle de régression. Elle permet d'identifier les éventuelles non-linéarités et les problèmes d'hétéroscédasticité. Le graphique des résidus contre les valeurs prédites a été tracé pour examiner visuellement ces aspects. Les résidus doivent idéalement être distribués de manière aléatoire autour de la ligne de zéro résidu, sans montrer de tendance claire. Une dispersion des résidus peut indiquer une hétéroscédasticité, tandis qu'une distribution non aléatoire peut signaler des non-linéarités dans les données.(voir Figure III.11)

2.4.4 Détection de la muticollinéarité

La multicolinéarité est un phénomène dans lequel deux ou plusieurs des variables explicatives dans un modèle de régression multiple sont hautement corrélées. Si ces variables sont en corrélation étroite, il devient difficile pour le modèle de déterminer l'effet de chaque variable indépendante sur la variable dépendante.

Détection Pour détecter la multicolinéarité, nous utilisons le facteur d'inflation de la variance (VIF), qui est une mesure de la corrélation et de la force de cette corrélation entre les variables explicatives. Le VIF est calculé comme l'inverse de la proportion de la variance de la variable explicative qui n'est pas expliquée par les autres variables explicatives.

Un VIF de 1 signifie qu'il n'y a pas de corrélation entre la variable explicative en question et les autres variables explicatives. Plus la valeur du VIF est élevée, plus la multicolinéarité est forte. Une valeur de VIF supérieure à 5 ou 10 est souvent considérée comme indiquant une multicolinéarité problématique.

Traitement : Si la multicolinéarité est détectée, diverses méthodes peuvent être utilisées pour l'atténuer, y compris :

— Supprimer une des variables corrélées.

— Utiliser une analyse en composantes principales pour réduire la dimensionnalité.

— Combiner les variables corrélées en une seule variable, si cela a un sens dans le contexte de l'analyse.

2.4.5 Influence des Observations Individuelles

L'analyse de l'influence des observations individuelles est cruciale pour comprendre comment chaque observation affecte les estimations des coefficients dans le modèle. Cette analyse aide à détecter les points atypiques qui peuvent avoir un impact significatif sur le modèle.

Mesure de l'Influence : Nous utilisons la distance de Cook, notée par D, pour mesurer l'influence des observations individuelles. La distance de Cook quantifie l'effet de la suppression d'une observation sur les estimations des coefficients. Une valeur élevée de la distance de Cook pour une observation particulière signifie que cette observation a une grande influence sur les estimations.

Une règle courante est que si la distance de Cook d'une observation dépasse $\frac{4}{n-k-1}$, où n est le nombre d'observations et k est le nombre de variables explicatives, alors l'observation peut être considérée comme influente.

Nous avons également tracé un graphique de la distance de Cook pour chaque observation, ce qui permet de visualiser rapidement les observations qui peuvent avoir une influence importante sur le modèle. (voir Figure III.12)

2.5 Validation Croisée

Afin d'évaluer la robustesse et la fiabilité de notre modèle de régression linéaire, nous avons employé la méthode de validation croisée à 10-folds. La validation croisée est une technique statistique qui vise à comprendre comment le résultat d'un modèle statistique se généralisera à un ensemble de données indépendant. Elle est particulièrement utile dans les situations où les données disponibles sont limitées.

Dans cette méthode, les données sont divisées en 10 parties égales, appelées "folds". Le modèle est alors entraîné sur 9 de ces folds et testé sur le fold restant. Ce processus est répété 10 fois, chaque fold étant utilisé exactement une fois comme ensemble de test. Les 10 résultats sont ensuite moyennés pour obtenir une estimation unique de la performance du modèle. Cette approche permet d'obtenir une mesure plus précise et moins biaisée de la capacité du modèle à se généraliser à de nouvelles données.

2.6 Réduction de Dimensionnalité par Analyse des Composantes Principales (ACP)

Dans notre analyse, nous avons utilisé l'ACP pour réduire la dimensionnalité des données. L'ACP transforme les données en un nouvel ensemble de composantes qui sont non corrélées et capturent la variance dans l'ordre décroissant.

Au début, l'ACP a été appliquée sans standardiser les variables, ce qui a résulté en une seule composante expliquant presque toute la variance. Cela était dû aux échelles

différentes des variables, et a été corrigé en standardisant les données.

La standardisation a été effectuée en utilisant la moyenne et l'écart type de chaque variable. Après la standardisation, l'ACP a été appliquée à nouveau, et les composantes résultantes ont expliqué une variance plus équilibrée. La figure III.13 montre une représentation des deux premières composantes principales.

2.7 Segmentation des Polices d'Assurance par Clustering (K-means)

La méthode de classification K-means a été choisie pour segmenter notre ensemble de données en différents groupes ou clusters. Le choix de K-means est motivé par sa simplicité et son efficacité pour identifier les structures cachées dans les données. L'algorithme fonctionne en partitionnant les données en k clusters où chaque observation appartient au cluster ayant la moyenne la plus proche.

Pour déterminer le nombre optimal de clusters, nous avons utilisé la méthode du coude, qui consiste à calculer la somme des carrés intra-clusters (WCSS) pour une gamme de valeurs de k et à identifier le point où l'ajout de clusters supplémentaires n'apporte pas d'amélioration significative. La ligne reliant les points extrêmes de la WCSS est tracée, et la distance de chaque point à cette ligne est calculée. Le nombre optimal de clusters est le point avec la distance maximale à cette ligne. Dans notre cas, le nombre optimal de clusters trouvé est de 4.

La visualisation des clusters, ainsi que les tableaux résumant les caractéristiques de chaque cluster, sont présentés dans la section des résultats (voir la section 1.9).

La Figure III.14 présente la visualisation des clusters, où chaque couleur représente un cluster différent.

2.8 Analyse du Nombre de Sinistres Moyen par Âge

L'analyse du nombre moyen de sinistres en fonction de l'âge du conducteur révèle une tendance intéressante qui peut être cruciale pour la tarification des polices d'assurance automobile. Le Tableau II.26 présente le nombre moyen de sinistres par âge des conducteurs.

TABLE II.26 – Nombre moyen de sinistres par âge

Âge	**nb_sin**
18	1.2875
19	1.2763
...	...
75	1.2051

III Résultats

1 Présentation des Principaux Résultats

Les statistiques descriptives présentées dans le tableau II.15 offrent un aperçu important des principales variables numériques de l'étude. L'âge moyen des conducteurs est d'environ 35 ans avec un écart-type de 13.3, indiquant une distribution relativement étendue autour de la moyenne. Le coefficient d'asymétrie positif pour l'âge, le bonus, et la valeur du véhicule suggère que la distribution est légèrement étirée vers la droite.

Le bonus-malus varie de -50 à 150 avec une moyenne de 22.28 et un écart-type de 57.69. Cette large plage de valeurs met en évidence la diversité des profils de risque dans le portefeuille d'assurance.

La valeur des véhicules présente également une grande variabilité, avec une moyenne de 17088.84 et un écart-type de 10721.57. La distribution de cette variable est également légèrement asymétrique vers la droite, comme le montrent les histogrammes de la figure II.9.

Les tables de fréquence pour les variables catégorielles, telles que présentées dans les tableaux II.19, 2.1.2, II.17, et II.18, fournissent des informations clés sur la composition du portefeuille d'assurance.

Le sexe est presque divisé en deux tiers d'hommes (68.05%) et un tiers de femmes (31.95%). Les types de véhicules sont diversifiés, avec une légère prédominance des types A, B, et D. La catégorie de véhicule est également bien répartie, avec une légère prédominance de la catégorie Medium (35.85%). En ce qui concerne l'occupation, la majorité des assurés sont employés (35.07%), suivis par les femmes au foyer (23.81%) et les chômeurs (20.33%).

Ces observations peuvent avoir des implications importantes pour la tarification et le ciblage des produits d'assurance, et elles sont en ligne avec les objectifs et la problématique de l'étude.

Les résultats du test de Shapiro-Wilk (voir tableau II.20) indiquent que les variables continues 'Age', 'Bonus' et 'Value' ne suivent pas une distribution normale ($p < 0.05$). Ceci peut avoir des implications sur les méthodes statistiques à utiliser ultérieurement dans l'analyse, car de nombreuses méthodes supposent une distribution normale des données.

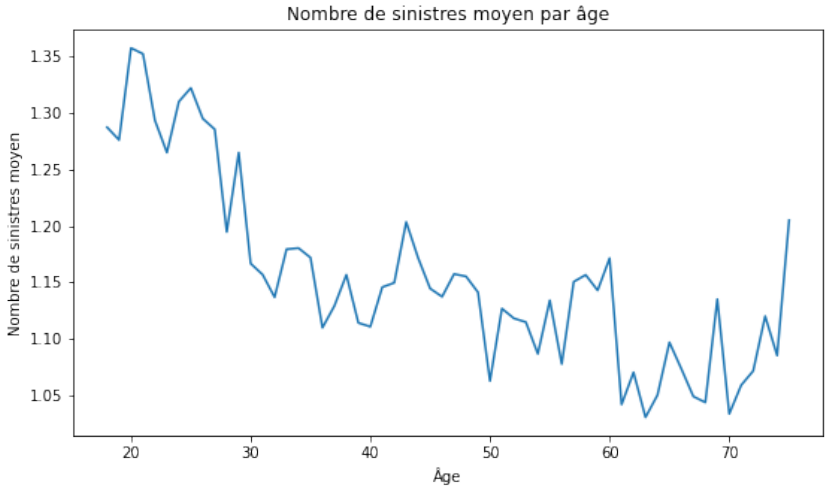


FIGURE III.1 – Nombre de sinistres moyen par âge

Interprétation

- **Jeunes conducteurs :** On observe une tendance relativement stable du nombre de sinistres pour les conducteurs âgés de 18 à 25 ans. Cette stabilité pourrait indiquer que l'âge, dans cette tranche, n'est pas un facteur significatif dans la fréquence des sinistres.
- **Conducteurs d'âge moyen :** Entre 26 et 50 ans, une légère tendance à la baisse du nombre de sinistres est remarquée, liée peut-être à une amélioration des compétences de conduite.
- **Conducteurs plus âgés :** Après 50 ans, la tendance devient moins claire. Cette fluctuation peut être associée à des facteurs divers tels que la diminution des réflexes avec l'âge.
- **Conducteurs âgés :** Notamment, une augmentation notable du nombre de sinistres à 75 ans peut nécessiter une attention particulière dans la tarification.

Ces résultats mettent en évidence l'importance de l'âge comme variable dans la tarification de l'assurance automobile. Ils suggèrent que l'âge peut être utilisé comme un indicateur de risque, avec des adaptations spécifiques selon les tranches d'âge.

Interprétation de la Corrélation entre Variables Numériques

La matrice de corrélation (voir Tableau 2.2 dans la section Méthodologie) et la carte thermique ci-dessous fournissent une compréhension claire des relations entre les variables numériques étudiées.

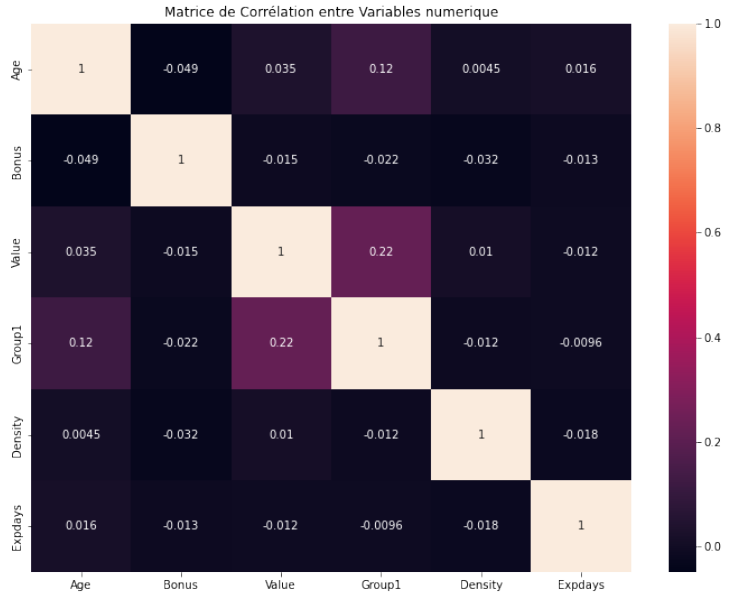


FIGURE III.2 – Matrice de Corrélation entre Variables numérique

L'analyse de la matrice de corrélation dévoile des informations pertinentes concernant le risque d'assurance automobile. La tendance la plus notable est que la prime d'assurance ou le BonusMalus diminue généralement avec l'âge. Cela peut être attribué à l'expérience accumulée et à une diminution des comportements à risque au volant avec l'âge. Cependant, cette tendance est influencée par d'autres facteurs tels que l'historique de conduite, le type de véhicule, et la localisation géographique, et il existe des exceptions.

Les relations clés dans cette analyse sont :

— **Âge et Bonus/Malus** (-0.048884) : Une faible corrélation négative indique une minuscule tendance pour que les conducteurs plus âgés aient un bonus légèrement plus élevé (ou malus légèrement plus bas), bien que la relation soit faible.

— **Âge et Group1** (0.1207) : L'âge du conducteur est légèrement associé à la catégorie de groupement.

— **Valeur et Group1** (0.2232) : La valeur du véhicule a une relation modérée avec la catégorie de groupement.

— **Densité et Bonus** (-0.0317) : Une corrélation négative très faible suggère une légère inversion entre la densité de la population dans la région et le bonus.

Les autres coefficients sont proches de zéro, révélant des relations linéaires faibles ou inexistantes entre ces paires de variables. Ces résultats soulignent l'importance de comprendre l'interaction entre les variables pour la modélisation et la prévision dans

l'étude du risque d'assurance automobile.

Type de Véhicule

L'analyse ANOVA a montré une différence significative dans la distribution de bonus par type de véhicule (voir Figure III.3). La valeur de p de 0.0442 indique que cette différence est statistiquement significative au niveau de 0.05.

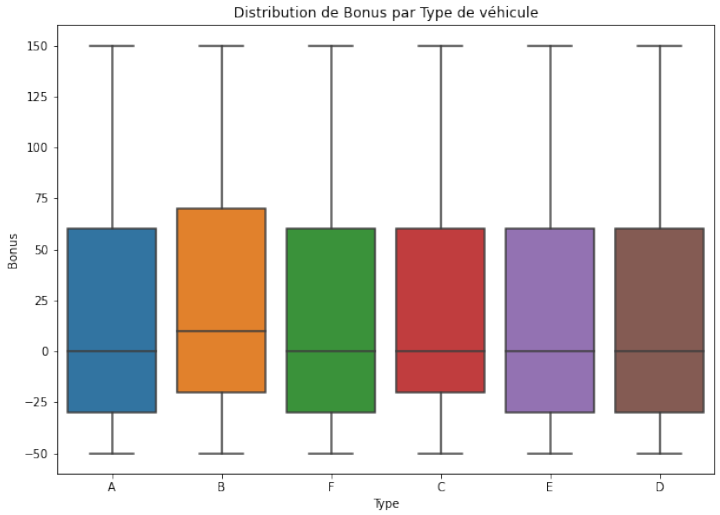


FIGURE III.3 – Distribution de Bonus par Type de véhicule

Catégorie de Véhicule

La comparaison des bonus par catégorie de véhicule n'a pas révélé de différence significative (voir Figure III.4). La valeur de p de 0.3416 est supérieure au seuil de 0.05, indiquant une absence de différence significative.

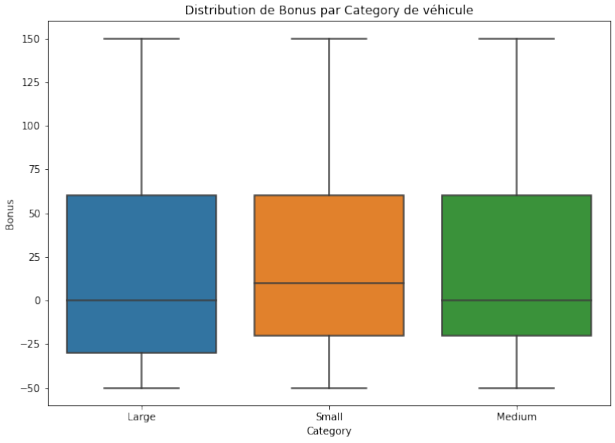


FIGURE III.4 – Distribution de Bonus par Catégorie de véhicule

Genre du Conducteur

L'analyse ANOVA a révélé une différence significative dans la distribution de bonus par genre du conducteur (voir Figure III.5). La valeur de p de 0.0082 indique que cette différence est statistiquement significative au niveau de 0.01.

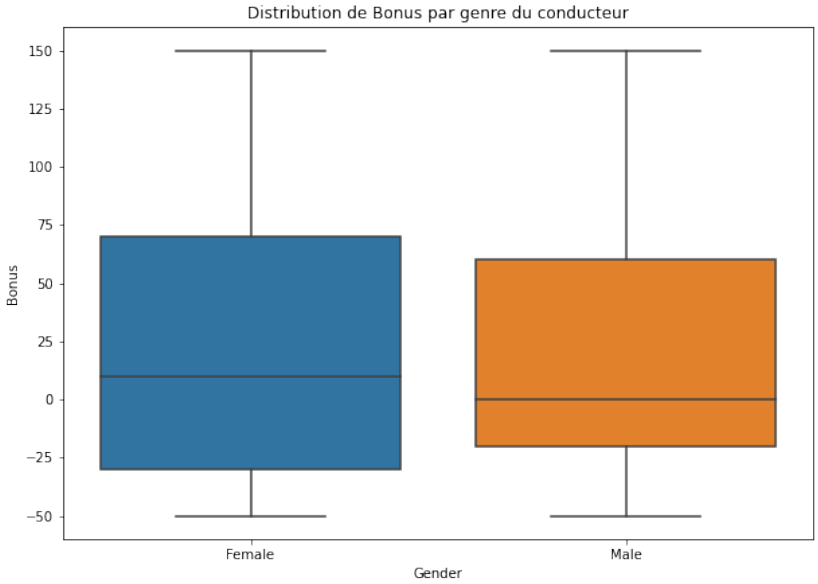


FIGURE III.5 – Distribution de Bonus par genre du conducteur

Les analyses effectuées ont permis de mettre en évidence l'effet du type de véhicule et du genre du conducteur sur le bonus en assurance. Aucun effet significatif n'a été détecté pour la catégorie du véhicule.

Interprétation des résultats du Test du Chi-carré

Genre et Type

Le résultat de $\chi^2 = 1.3086$ avec une valeur p de $p = 0.9340$ pour le test entre Genre et Type indique qu'il n'y a pas de relation significative entre ces deux variables (voir Figure III.6).

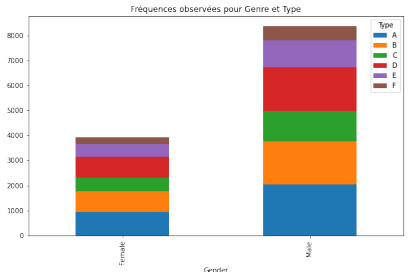


FIGURE III.6 – Fréquences observées pour Genre et Type

Genre et Catégorie

La relation entre Genre et Catégorie a été testée, avec un résultat de $\chi^2 = 2.0736$ et une valeur p de $p = 0.3546$, indiquant une absence de relation significative entre ces variables (voir Figure III.7).

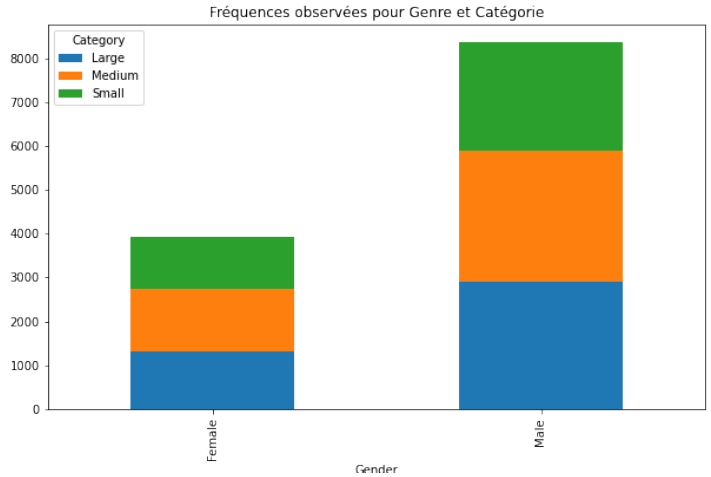


FIGURE III.7 – Fréquences observées pour Genre et Catégorie

Type et Catégorie

Le test entre Type et Catégorie a produit un résultat de $\chi^2 = 9.7337$ avec une valeur p de $p = 0.4642$, indiquant qu'il n'y a pas de relation significative entre ces deux variables (voir Figure III.8).

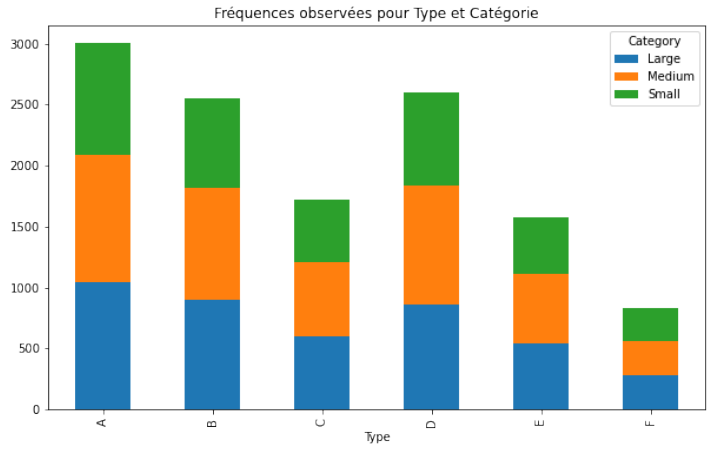


FIGURE III.8 – Fréquences observées pour Type et Catégorie

1.1 Résultats des Tests (voir Figure 2.3.2 et 2.3.2)

— **R-squared** : Les résultats de la régression montrent un R^2 de 0,080 et un R^2 ajusté de 0,078, indiquant que le modèle explique environ 8% de la variation dans le

coût total des sinistres.

— **Test F** : La statistique F de 35.63 et sa probabilité associée proche de zéro indiquent que le modèle est globalement significatif.

— **Jarque-Bera** : Le test de Jarque-Bera avec une p-value de 0.00 indique que les résidus ne suivent pas une distribution normale, ce qui peut nécessiter un examen plus approfond.

1.2 Interprétation des Coefficients Continus

Nous présentons ci-dessous l'interprétation des coefficients qui sont significatifs au seuil de 5% : (voir Figure III.5)

— **const** : La constante du modèle est de -50910, significative à 1%, indiquant la valeur de `chg_sin` lorsque toutes les autres variables sont à zéro.

— **CalYear** : Pour chaque augmentation d'un an, `Coût total des sinistres RC matéeriels` augmente de 58.6476, ceteris paribus.

— **Age** : Pour chaque année supplémentaire de l'âge, `Coût total des sinistres RC matéeriels` diminue de 11.6510, ceteris paribus.

— **Group1** : Pour chaque augmentation d'un dans le groupe du véhicule, `Coût total des sinistres RC matéeriels` augmente de 9.9547, ceteris paribus.

— **Poldur** : Pour chaque année supplémentaire de l'ancienneté du contrat, `Coût total des sinistres RC matéeriels` diminue de 4.9726, ceteris paribus.

— **Adind** : Si l'indicateur de garantie dommages est activé, `Coût total des sinistres RC matéeriels` diminue de 133.1689, ceteris paribus.

— **Density** : Pour chaque unité supplémentaire de densité de population, `Coût total des sinistres RC matéeriels` augmente de 2.3251, ceteris paribus.

— **Expdays** : Pour chaque jour supplémentaire d'exposition, `Coût total des sinistres RC matéeriels` augmente de 0.7093, ceteris paribus.

— **Cependant, deux coefficients, Bonus et Value,** ne sont pas significatifs au seuil de 5%, avec des valeurs de p de 0,777 et 0,382, respectivement.

1.3 Interprétation des Coefficients Catégoriels

— **Type** : Les coefficients pour les différents types de véhicules (A à F) sont négatifs, indiquant une réduction du coût des sinistres RC matériels pour ces types par rapport à une catégorie de référence, ceteris paribus.

— **Category** : Les coefficients pour les catégories de véhicules (Large, Medium, Small) sont négatifs et similaires en magnitude, indiquant une réduction du coût des sinistres RC matériels pour ces catégories par rapport à une catégorie de référence, ceteris paribus.

— **Occupation** : Les coefficients pour les différentes professions (Employed, Housewife, Retired, Self-employed, Unemployed) sont négatifs, indiquant des différences dans le coût des sinistres RC matériels selon la profession de l'assuré par rapport à une profession de référence, ceteris paribus.

— **Group2** : Les coefficients pour les différentes régions d'habitation (L à U) sont négatifs et varient légèrement en magnitude, indiquant des différences dans le coût des sinistres RC matériels selon la région d'habitation de l'assuré par rapport à une région de référence, ceteris paribus.

— **Gender** : Les coefficients pour Homme et Femme sont similaires en magnitude mais négatifs, ce qui peut nécessiter une interprétation contextuelle. Cela pourrait indiquer une différence dans le coût des sinistres RC matériels selon le sexe de l'assuré par rapport à une catégorie de référence, ceteris paribus.

1.3.1 Tableau de Comparaison

Ci-dessous le tableau comparant les mesures clés des deux modèles.

Mesure	Modèle Initial	Modèle Ajusté
R^2	0.080	0.080
AIC	2.031e+05	2.031e+05
BIC	2.033e+05	2.033e+05
Durbin-Watson	1.980	1.980

TABLE III.1 – Comparaison des mesures clés entre le modèle initial et le modèle ajusté

Après avoir retiré les deux variables non significatives, le modèle ajusté explique une proportion similaire de la variance dans le nombre de sinistres RC matériels, comme le montre la valeur R^2 de 0.080. Cela suggère que le modèle ajusté maintient la capacité explicative tout en étant plus simple et plus facile à interpréter.

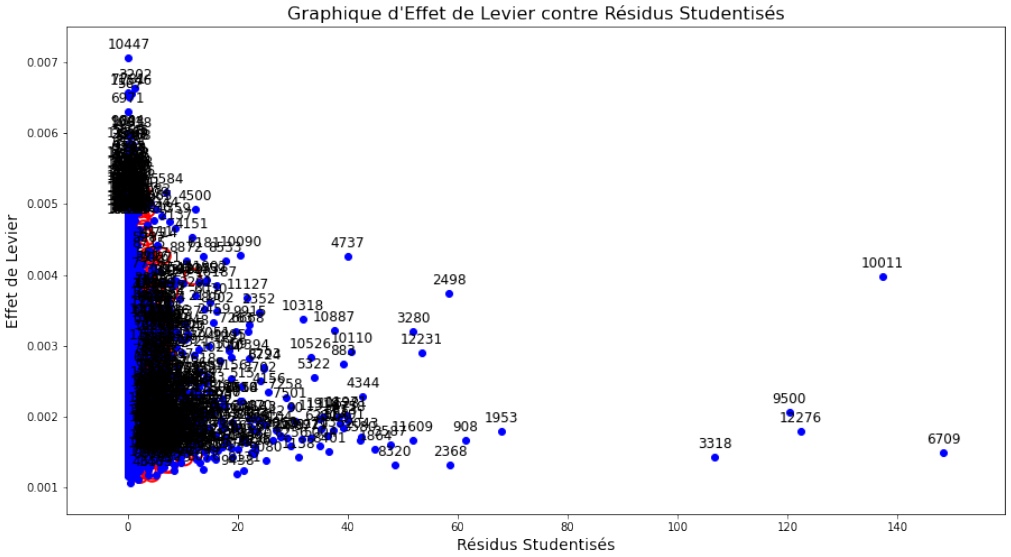
Les autres statistiques, telles que la statistique Durbin-Watson et les tests d'adéquation du modèle (Omnibus, Jarque-Bera), sont également restées cohérentes entre les deux modèles, indiquant que le modèle ajusté continue de répondre aux hypothèses sous-jacentes de la régression linéaire.

En analysant le nombre de sinistres RC matériels, le modèle ajusté offre une vision plus claire des facteurs qui contribuent à la rentabilité économique pour les assureurs et la protection adéquate pour les assurés, soutenant ainsi notre objectif d'optimiser les pratiques tarifaires.

A- Points à fort effet de levier et Résidus Studentisés

Nous avons trouvé que notre modèle ne contient pas de points à fort effet de levier, car aucun point ne dépasse le seuil de 0.05. Cependant, nous avons identifié un grand nombre de résidus studentisés importants.

Graphique d'Effet de Levier contre Résidus Studentisés : La Figure III.9 illustre le graphique de l'effet de levier contre les résidus studentisés.



Cliquez ici pour voir le graphique interactif.

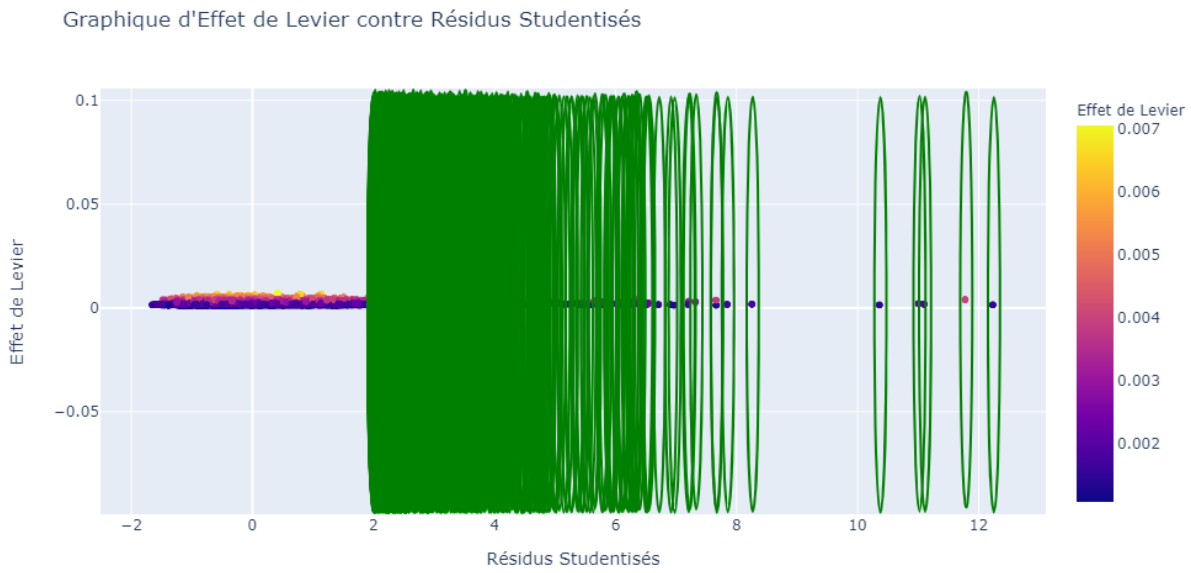


FIGURE III.9 – Graphique d'Effet de Levier contre Résidus Studentisés

Test de Normalité des Résidus : Le test de Shapiro-Wilk sur les résidus donne une

statistique de 0.7989 et une valeur-p de 0.0, ce qui indique que les résidus ne suivent pas une distribution normale.

Interprétation L'absence de points à fort effet de levier et la présence de résidus studentisés importants, ainsi que la non-normalité des résidus, peuvent indiquer des problèmes potentiels avec le modèle, tels que des spécifications de modèle incorrectes ou la présence d'outliers. Une analyse plus approfondie et des modifications de modèle peuvent être nécessaires.

1.4 Homoscédasticité

Les résultats du test de Breusch-Pagan sont les suivants :

$$\begin{aligned} \text{Lagrange multiplier statistic} &: 360.84441141683527 \\ \text{p-value} &: 8.704140244809434 \times 10^{-57} \\ \text{f-value} &: 13.246189529891623 \\ \text{f p-value} &: 1.6552280363006226 \times 10^{-60} \end{aligned}$$

La valeur p est très proche de zéro, ce qui nous amène à rejeter l'hypothèse nulle de l'homoscédasticité. Cela signifie que la variance des résidus n'est pas constante à travers les niveaux de la variable explicative, et nous avons donc un problème d'hétéroscédasticité.

Le graphique des résidus contre les valeurs prédites (voir Figure III.11) confirme ce résultat, montrant une dispersion variable des résidus.

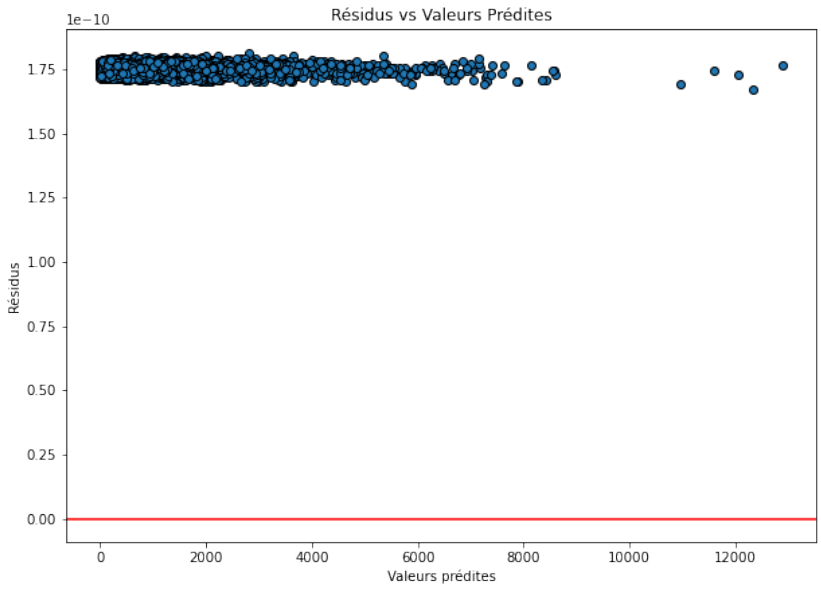


FIGURE III.10 – Résidus vs Valeurs Prédites

1.5 Diagnostic des Résidus

Le graphique des résidus contre les valeurs prédites montre une distribution homogène des résidus autour de la ligne de zéro résidu pour la majeure partie de l'échantillon (jusqu'à 8000 sur l'axe des x). Cela indique une bonne adéquation du modèle à ces données. Cependant, une dispersion visible des résidus au-delà de ce point suggère une possible hétéroscédasticité, indiquant que le modèle peut ne pas capturer toute la complexité des données dans cette plage. Des investigations supplémentaires peuvent être nécessaires pour comprendre la nature de cette dispersion et éventuellement ajuster le modèle en conséquence.

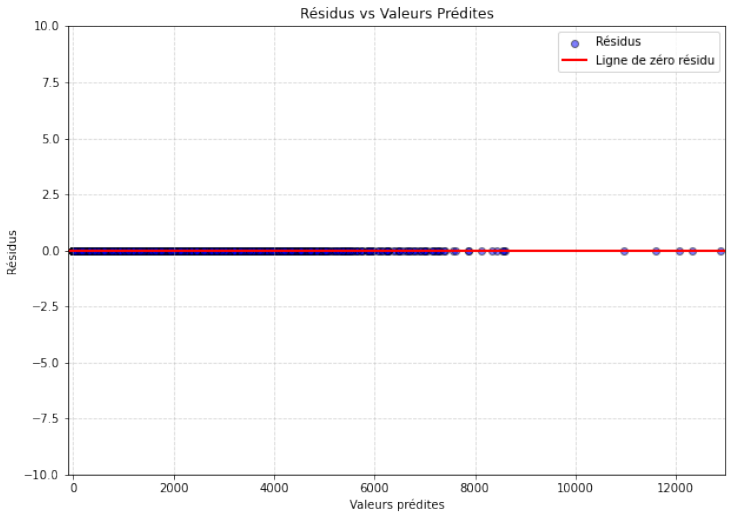


FIGURE III.11 – Résidus vs Valeurs Prédites

1.6 Influence des Observations Individuelles

Nous avons calculé la distance de Cook pour chaque observation afin d'évaluer leur influence sur le modèle. Voici quelques valeurs notables :

```
cooks_d
0       1.669117e-07
1       1.065646e-04
2       7.544387e-07
3       4.867379e-07
4       4.660176e-05
...
12276   6.452027e-03
```

La plupart des valeurs de la distance de Cook sont proches de zéro, indiquant une faible influence. Toutefois, quelques observations ont des valeurs plus élevées, nécessitant une investigation plus approfondie pour déterminer si elles sont des points atypiques qui peuvent affecter les résultats du modèle.

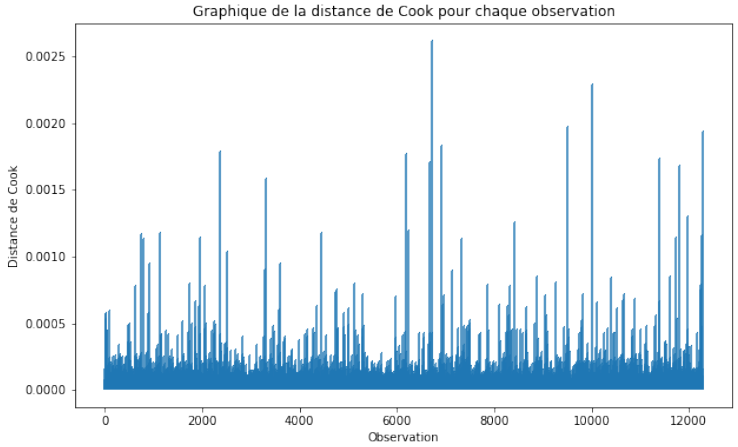


FIGURE III.12 – Graphique de la distance de Cook pour chaque observation

Ce graphique montre la distance de Cook pour chaque observation. Les observations avec des distances de Cook élevées peuvent être considérées comme ayant une influence importante sur les estimations des coefficients du modèle. Il peut être utile d'examiner ces observations de plus près pour comprendre pourquoi elles ont une telle influence.

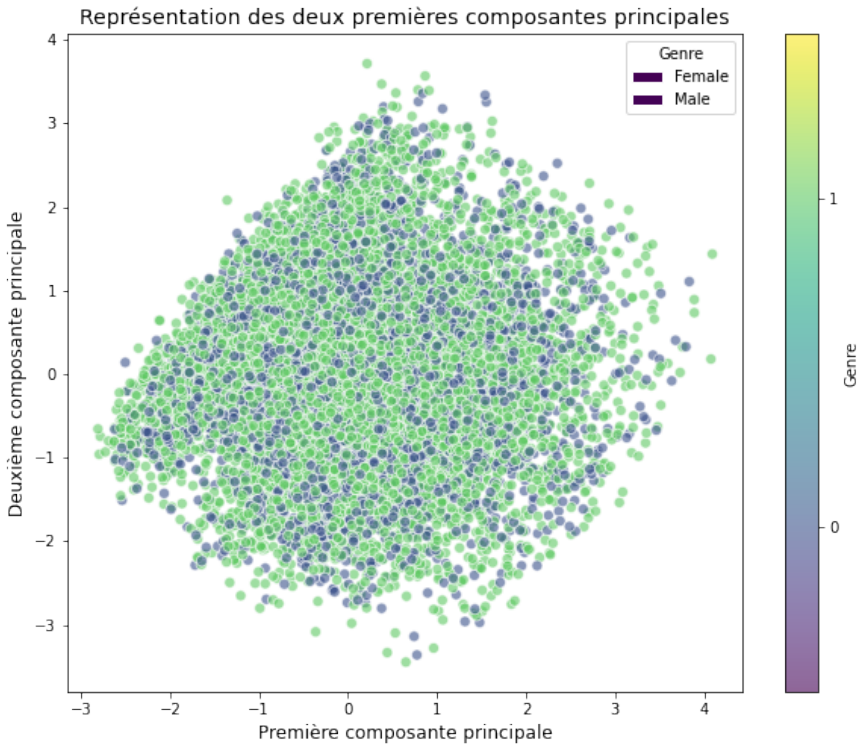
1.7 Résultats de la Validation Croisée

Les résultats de la validation croisée à 10-folds sur notre modèle de régression linéaire sont présentés ci-dessous :

```
Scores pour chaque fold: [0.0787, 0.0333, 0.0598, 0.0588, 0.0940, 0.0910,
0.0930, 0.1010, 0.0626, 0.0743]
Score moyen: 0.0747
```

Ces scores représentent le coefficient de détermination R^2 pour chaque fold, mesurant la proportion de la variance dans la variable dépendante qui est prévisible à partir des variables indépendantes. Le score moyen de 0.0747 indique que le modèle explique en moyenne 7.47% de la variance dans la variable dépendante 'chg_sin' dans chaque fold.

Bien que ces résultats soient positifs, ils sont relativement faibles, ce qui peut indiquer que le modèle n'est pas très puissant pour expliquer la variable dépendante avec



H

FIGURE III.13 – Représentation des deux premières composantes principales.

les variables explicatives choisies. Cela pourrait suggérer la nécessité de réexaminer les variables incluses, d'envisager d'autres techniques de modélisation, ou d'explorer des transformations supplémentaires des variables.

1.8 Résultats de l'ACP

Après avoir standardisé les données, les deux premières composantes principales ont expliqué respectivement 15% et 12.8% de la variance totale. Les coefficients de la première composante principale étaient :

$$-0.0014, 0.5841, 0.5101, 0.0498, 0.4793, 0.0136, 0.0114, -0.1101, 0.3925 \quad \text{(III.1)}$$

Cela suggère que la première composante est fortement influencée par les variables *Age*, *Group1*, *Adind*, et *Value*.

1.9 Résultats des Clusters

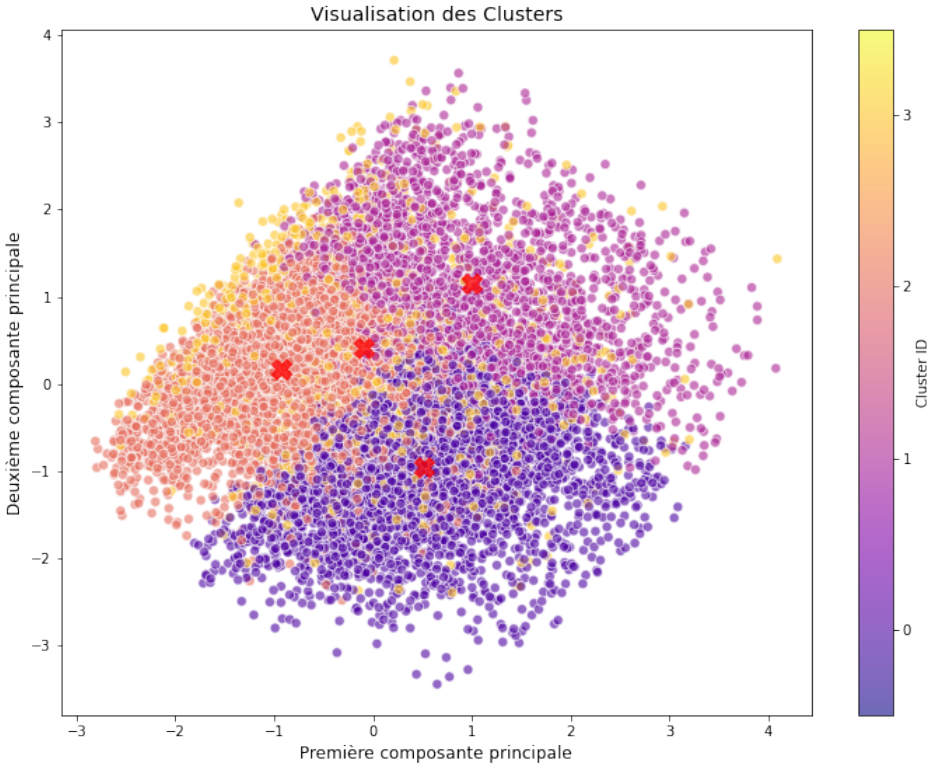


FIGURE III.14 – Visualisation des Clusters

Chaque cluster représente un segment particulier des données et est caractérisé par différentes propriétés statistiques :

— **Cluster 0 :** Ce cluster comprend 4136 observations. Les caractéristiques notables incluent une valeur moyenne de 13655, une durée moyenne de police de 5.05, et une densité moyenne de 139.5. D'autres détails statistiques peuvent être trouvés dans le Tableau III.2.

— **Cluster 1 :** Avec 2298 observations, ce cluster présente une valeur moyenne plus élevée de 32262 et une durée moyenne de police de 5.01. D'autres caractéristiques sont résumées dans le Tableau 1.9.

— **Cluster 2 :** Ce cluster, comprenant 4603 observations, se distingue par une âge moyen plus jeune de 30.4 et une valeur moyenne plus faible de 12697. Plus de détails sont présentés dans le Tableau 1.9.

— **Cluster 3 :** Avec 1240 observations, ce cluster présente des caractéristiques distinctes résumées dans le Tableau 1.9.

Ces clusters représentent différentes catégories ou segments au sein de l'ensemble de données, ce qui peut fournir des insights précieux pour des analyses ultérieures ou des actions ciblées.

TABLE III.2 – Statistiques pour le Cluster 0

	count	mean	std	min	25%	50%	75%	max
PolNum	4.14e+03	2.00e+08	6.23e+04	2.00e+08	2.00e+08	2.00e+08	2.00e+08	2.00e+08
CalYear	4136	2009	0.50	2009	2009	2010	2010	2010
Gender	4136	0.67	0.47	0	0	1	1	1
Age	4136	38.53	13.10	18	28	37	47	75
Group1	4136	11.25	4.42	1	8	11	14	20
Bonus	4136	24.40	61.87	-50	-30	10	70	150
Poldur	4136	5.05	4.51	0	1	4	8	15
Value	4136	13655.31	7217.60	1010	7918.75	12977.50	19333.75	38485
Adind	4136	1	0	1	1	1	1	1
Density	4136	139.50	84.30	14.38	63.83	129.67	212.42	297.39
Expdays	4136	359.23	19.11	247	365	365	365	365
nb_sin	4136	1.18	0.49	1	1	1	1	7
chg_sin	4136	755.95	827.96	0.18	209.79	508.50	1021.46	12878.37
cluster	4136	0	0	0	0	0	0	0

TABLE III.3 – Statistiques pour le Cluster 0

	count	mean	std	min	25%	50%	75%	max
PolNum	2298	2.00e+08	62223	2.00e+08	2.00e+08	2.00e+08	2.00e+08	2.00e+08
CalYear	2298	2009.50	0.50	2009	2009	2009	2010	2010
Gender	2298	0.68	0.47	0	0	1	1	1
Age	2298	38.16	14.32	18	26	36	48	75
Group1	2298	15.19	3.42	1	13	16	18	20
Bonus	2298	16.39	57.31	-50	-30	0	60	150
Poldur	2298	5.01	4.42	0	1	4	8	20
Value	2298	32262.03	9228.20	6215.00	25476.25	29737.50	39985.00	49990.00
Adind	2298	0.35	0.48	0	0	0	1	1
Density	2298	145.85	83.62	14.38	68.85	138.51	218.98	297.39
Expdays	2298	357.42	23.61	210	365	365	365	365
nb_sin	2298	1	0.63	1	1	1	1	7
chg_sin	2298	916.48	1098.91	0.42	231.39	560.48	1219.85	12324.74
cluster	2298	1	0	1	1	1	1	1

TABLE III.4 – Statistiques pour le Cluster 1

	count	mean	std	min	25%	50%	75%	max
PolNum	4603	2.00e+08	62197	2.00e+08	2.00e+08	2.00e+08	2.00e+08	2.00e+08
CalYear	4603	2009.52	0.50	2009	2009	2010	2010	2010
Gender	4603	0.68	0.47	0	0	1	1	1
Age	4603	30.42	11.55	18	22	27	36	75
Group1	4603	10.45	4.34	1	7	10	14	20
Bonus	4603	22.45	53.27	-50	-10	0	50	150
Poldur	4603	4.90	4.52	-6	1	4	8	31
Value	4603	12697.28	6768.97	1005.00	7540.00	11470.00	17582.50	39935.00
Adind	4603	0	0	0	0	0	0	0
Density	4603	137.50	83.91	14.38	61.94	126.14	210.19	297.39
Expdays	4603	359.81	17.73	254	365	365	365	365
nb_sin	4603	1	0.58	1	1	1	1	7
chg_sin	4603	965.66	1063.18	1.14	254.41	623.93	1302.92	12055.25
cluster	4603	2	0	2	2	2	2	2

TABLE III.5 – Statistiques pour le Cluster 2

	count	mean	std	min	25%	50%	75%	max
PolNum	1240	2.00e+08	61557	2.00e+08	2.00e+08	2.00e+08	2.00e+08	2.00e+08
CalYear	1240	2009.50	0.50	2009	2009	2010	2010	2010
Gender	1240	0.70	0.46	0	0	1	1	1
Age	1240	34.10	12.83	18	24	31	42	75
Group1	1240	11.94	4.44	1	9	12	15	20
Bonus	1240	25.55	59.01	-50	-20	10	70	150
Poldur	1240	5.07	4.53	0	1	4	8	15
Value	1240	16723.82	9944.10	1005.00	8952.50	15310.00	22860.00	49995.00
Adind	1240	0.42	0.49	0	0	0	1	1
Density	1240	143.71	84.88	17.88	66.10	138.51	216.49	297.39
Expdays	1240	190.17	47.77	91	153	195	229	278
nb_sin	1240	1	0.37	1	1	1	1	4
chg_sin	1240	783.19	875.95	0.36	209.58	510.97	1029.95	7264.91
cluster	1240	3	0	3	3	3	3	3

TABLE III.6 – Statistiques pour le Cluster 3

2 Discussion des Résultats

2.1 Distribution des Variables

L'âge, le bonus-malus, la valeur du véhicule et autres variables clés montrent une distribution étendue, reflétant une diversité des profils de risque dans le portefeuille d'assurance. La diversité des types et catégories de véhicules, ainsi que l'occupation des assurés, permet une compréhension plus profonde des besoins des clients et de la manière de les servir.

2.2 Corrélation entre Variables

La corrélation entre l'âge et le bonus/malus, bien que faible, indique une relation potentielle entre l'expérience de conduite et le risque d'assurance. Les autres corrélations offrent des insights intéressants mais sont généralement faibles.

2.3 Analyse ANOVA et Tests du Chi-carré

Une différence significative dans la distribution de bonus par genre et type de véhicule, mais pas par catégorie, permet d'identifier les domaines spécifiques où les différenciations peuvent être faites. Le manque de relation significative dans les tests du Chi-carré indique que certaines variables catégorielles peuvent ne pas être aussi pertinentes pour la prédiction que d'autres.

2.4 Résultats des Tests et Coefficients

Le modèle ajusté avec un R^2 de 0,080 montre que les variables sélectionnées expliquent une partie de la variance, mais il y a encore une grande quantité d'information inexpliquée. Les coefficients révèlent l'importance de certaines variables, comme la densité de population, l'âge, le type de véhicule, et plus encore.

2.5 Diagnostics du Modèle

— **Effet de levier :** Aucun problème détecté, aucune valeur n'excède le seuil de 0.05.

— **Résidus studentisés :** De nombreux résidus studentisés importants ont été trouvés, ce qui peut indiquer des problèmes potentiels avec le modèle.

— **Test de Normalité des Résidus :** Le test de Shapiro-Wilk indique que les résidus ne suivent pas une distribution normale. Cette non-normalité peut être symptomatique de problèmes dans la spécification du modèle.

— **Homoscédasticité :** Le test de Breusch-Pagan rejette l'hypothèse d'homoscédasticité, signalant un problème d'hétéroscédasticité.

— **Influence des Observations Individuelles :** La distance de Cook indique que la plupart des observations ont une faible influence, mais certaines valeurs plus élevées nécessitent une investigation plus approfondie.

2.6 Résultats de l'ACP et des Clusters

Les clusters identifiés et les résultats de l'ACP offrent des insights sur la structure des données. Il peut être utile de considérer des transformations des variables, d'exa-

miner de plus près les observations influentes, et de reconsidérer la spécification du modèle.

IV Recommandations

1 Limites et Recommandations

1.1 Ciblage et Tarification

Les assureurs peuvent utiliser ces données pour segmenter leurs clients plus efficacement, offrant des produits personnalisés basés sur l'âge, le type de véhicule, l'occupation, et d'autres facteurs pertinents.

1.2 Amélioration du Modèle

Considérant le faible R^2, l'exploration de variables supplémentaires ou l'utilisation de techniques de modélisation plus sophistiquées peut aider à améliorer la précision de la prédiction.

1.3 Considération des Biais Potentiels

Une analyse plus profonde sur l'impact du genre peut être nécessaire pour s'assurer que la tarification est équitable et ne reflète pas de biais non fondés.

1.4 Utilisation d'un Second Modèle pour le Nombre de Sinistres RC Matériels (nb_sin)

Étant donné que la variable "Nombre de sinistres RC matériels" (nb_sin) est discrète et non négative, une modélisation différente de la régression par les moindres carrés ordinaires (MCO) pourrait être nécessaire. Le modèle Poisson ou le modèle binomial négatif peuvent être explorés, qui sont spécifiquement conçus pour gérer des variables dépendantes discrètes. La comparaison de ce modèle avec le modèle initial peut révéler si cette approche est plus robuste et précise dans la prédiction des sinistres.

1.5 Limitations

Le manque d'information sur le Nombre de sinistres RC corporels, le Coût total des sinistres corporels, et la Prime pure constitue une limitation majeure de cette étude. L'intégration de ces variables dans les futures recherches pourrait fournir des insights plus profonds et conduire à des modèles plus robustes et pertinents.

V Conclusion

L'étude a mis en lumière la relation entre différentes variables telles que l'âge, le bonus-malus, le type de véhicule et le risque d'assurance. Toutefois, un faible R^2 et d'autres problèmes diagnostiques ont été identifiés, suggérant que le modèle actuel nécessite des améliorations.

Cette recherche contribue à une meilleure compréhension des facteurs de risque en assurance automobile. Les assureurs peuvent utiliser ces informations pour segmenter leurs clients de manière plus efficace, offrant ainsi des produits plus personnalisés et une tarification plus précise. La suggestion d'explorer un modèle différent pour le nombre de sinistres RC matériels pourrait également ouvrir de nouvelles voies de modélisation dans ce domaine.

Des méthodes de modélisation alternatives, des transformations de variables et l'examen des observations influentes pourraient améliorer la performance du modèle.

L'étude a révélé la complexité du paysage de risque dans l'assurance automobile et a souligné l'importance d'une modélisation attentive et éthique. Bien que des défis demeurent, notamment en termes de qualité et de complétude des données, cette recherche ouvre la voie à des innovations significatives dans le domaine de l'assurance automobile. La poursuite de l'exploration, en combinant des analyses riches et des techniques de modélisation robustes, est cruciale pour l'évolution future de la pratique de l'assurance automobile.

Table des figures

Liste des tableaux

Bibliographie

[1] G. Saporta. (2011). *Probabilités, analyse des données et Statistique*. TECHNIP. p. 125, 155-168.

[2] S. Tufféry. *Modélisation prédictive et Apprentissage statistique avec R*. p. 13.