Table des matières

1 Description statistique des données

Nous disposons d'une base de données qui contient les caractéristiques de 100 000 polices d'assurance et d'une autre contenant les informations des éventuels sinistres associés.

Pour l'étude de la fréquence nous réalisons une jointure de ces deux bases à partir de l'identifiant de la police et sommons le nombre de sinistres associés si un contrat présente plusieurs sinistres dans l'année. Les sinistres aux montants négatifs ne seront pas traités et sont supprimés en amont de la jointure. Nous disposons alors d'une base contenant les caractéristiques des 100 000 polices auxquelles sont associés le nombre de sinistres subit par l'assuré. Le tableau 1 présente la répartition des contrats en fonction du nombre de sinistres observés. La grande majorité des polices (plus de 88%) n'ont pas subit de sinistre.

L'étude et la modélisation de la sévérité a été réalisée sur les montants de sinistres individuels et non la somme des sinistres ni la moyenne empirique par police. La base qui résulte de la jointure contient alors 13 464 observations, les contrats présentant plusieurs sinistres étant présents plusieurs fois dans la base. De la même manière que pour la sévérité, les sinistres aux montants négatifs ne seront pas traités et sont supprimés en amont de la jointure, nous laissant ainsi avec 13 024 observations.

Table 1 – Répartition des polices en fonction du nombre de sinistres

Nombre de sinistres	Nombre de polices	Proportion (en %)
0	88273	88.27
1	10541	10.54
2	1080	1.08
3	101	0.10
4	5	0.01

Les statistiques de certaines variables jugées importantes pour la suite de la modélisation sont résumées tableau 2.

Table 2 – Statistiques descriptives des principales variables

Statistique	claim_nb	drv_age1	vh_age	vh_din	vh_value	pol_bonus
Minimum	0.00	19.00	1.00	15.00	0	0.50
1er Quantile	0.00	43.00	4.00	68.00	11 950	0.50
Médiane	0.00	54.00	8.00	87.00	16 200	0.50
Moyenne	0.13	54.61	9.48	91.32	18 037	0.54
3ème Quantile	0.00	65.00	13.00	109.00	22 106	0.50
Maximum	4.00	103.00	63.00	507.00	155 498	1.65

pol_coverage	pol_usage	pol_region	vh_make	vh_fuel
Maxi : 64 623	WorkPrivate : 66 055	Ile-de-France : 15 621	Renault : 27 011	Diesel : 54 760
Median2 :17 401	Retired : 26 655	Auvergne : 13 270	Peugeot : 19278	Gasoline : 45 160
Median1 : 94 04	Professional : 7 195	Nouvelle-Aquitaine : 11 323	Citroen : 15 800	Hybrid : 80
Mini : 8 572	AllTrips : 95	Autres : 59 786	Autres : 37 911	

Enfin, les figures 10 et 11 en annexe 3.1, réalisées à partir des montants de sinistres individuels, permettent d'observer une forte asymétrie à gauche. Si certains montants sont très élevés (plusieurs dizaines de milliers d'euros), près de 92% des sinistres sont inférieurs à 3 000€. Ce phénomène d'asymétrie est très fortement souligné par le boxplot des montants dont il résulte que de nombreuses valeurs sont aberrantes (10% sont au dessus du 3ème quartile). Lorsque nous nous penchons sur la part des montants que représentent ces grands sinistres, la figure 12 indique que :

— 95% des plus petits sinistres pèsent pour 62% des montants totaux ;

— les 10% des plus grands sinistres représentent plus de la moitiés des montants en jeu.

Nous relevons ainsi qu'il existe un fort écart entre les montants que nous qualifierons par la suite de "standards" et les montants élevés, ce qui motive notre approche de modélisation par écrêtement.

2 Modélisation de la fréquence

2.1 Première sélection des variables

L'analyse statistique des données permet une première sélection des variables explicatives pour le nombre de sinistres. Pour les variables numériques, une analyse des corrélations permet d'éliminer les variables corrélées et redondantes. Au travers des matrices de corrélations dont la représentation graphique est disponible en annexe figure 2 et des histogrammes du nombre de sinistre par catégories, nous choisissons pour la modélisation de la fréquence un ensemble de variables explicatives. Les caractéristiques liées à la voiture étant très corrélées entre elles, nous choisissons de conserver l'âge du véhicule, sa puissance dynamique, son poids, son alimentation et son constructeur. En ce qui concerne la police, nous étudions l'âge du conducteur, la durée de la police, le type de couverture, le bonus, la région et l'usage.

L'étude statistique des variables explicatives réalisée en amont [1] a permis la construction de catégories pertinentes dans l'explication de la fréquence. Les histogrammes représentant le nombre de polices et le nombre de sinistres selon certaines des variables retenues sont disponibles figures 4 à 8. Quand la fréquence de sinistre semble relativement constante avec l'âge du conducteur, ce n'est pas le cas pour la majorité des autres variables considérées. Nous observons que l'âge du véhicule est inversement proportionnel à la fréquence : plus un véhicule est récent et plus il est probable que l'assuré subisse un sinistre. La valeur du véhicule et sa puissance dynamique sont également des variables clés dans l'explication de la fréquence, plus ces dernières sont élevée et plus le risque d'occurrence d'un sinistre est élevé. Naturellement, le risque de sinistre est d'autant plus accru que l'assuré a choisit une couverture élevée. Le tableau 7 récapitule les variables sélectionnées pour le modèle de fréquence et donne leur catégorisation.

2.2 Modélisation

Une première régression Poisson est réalisée avec les onze variables décrites tableau 7. L'implémentation d'une méthode *stepwise* AIC [2] [4] permet ensuite de réaliser une sélection plus parcimonieuse de ces variables. Le *stepwise* AIC garde neuf des onze variables : il ne conserve pas le constructeur du véhicule et la région. Les modèles étudiés par la suite sont réalisés avec les neuf variables restantes. Le tableau 3 présente les métriques d'adéquation des différentes modélisation.

TABLE 3 – Métriques d'adéquation (pour les mêmes 9 variables explicatives)

	Poisson	Quasi Poisson	BN θ estimé	BN avec θ fixe
log-vraisemblance	-39 659.59		-39 563.76	-39 564.1
AIC	79 435.17		79 245.52	79 244.21
BIC	79 986.92		79 806.78	79 795.96
déviance	55 095.87	55 095.87	49 453.56	49 125.26

	ZI Poisson	ZI Bin. nég	ZM Poisson	ZM Bin. nég.
log-vraisemblance	-39 479.92	-39 481.34	-39 520.19	-39 519.8
AIC	79 191.83	79 196.68	79 272.39	79 273.61
BIC	80 295.33	80 309.7	80 375.89	80 386.62

Bien que le modèle qui minimise le critère AIC soit le modèle zéro-inflaté, nous choisissons le modèle binomial négatif avec θ fixe qui n'a pas un AIC très éloigné du ZI Poisson mais qui a un critère BIC bien plus faible. Le modèle binomial négatif est celui qui a la déviance la plus faible ce qui indique qu'il est le meilleur modèle en termes d'ajustement (entre les quatre modélisations pour qui la déviance est disponible). Les coefficients de cette régression sont détaillés tableau 8. Le détail du calcul du θ est disponible en annexe A.3.

3 Modélisation de la sévérité

3.1 Études des variables

De la même manière que pour la fréquence en partie 2.1, nous étudions les liens entre les montants de sinistres observés et les variables explicatives à disposition. A nouveau, l'étude des corrélations croisées réalisée précédemment permet d'évincer les variables redondantes. Une étude descriptive approfondie mais non-exhaustive a été réalisée [2] afin

1. voir fichier "1_frequence"
2. voir fichier "1_stat_claim.html"

d'en arriver aux catégorisations finales. Les figures 13 à 18 de la partie 3.1 représentent les différents boxplots des montants de sinistres selon les catégorisations retenues. Nous relevons des niveaux médians relativement proches selon les différentes profils de risque. La puissance du véhicule et le bonus de la police d'assurance font toutefois exception : la médiane des montants de sinistres double entre les premières et dernières catégories considérées pour ces variables. De plus, les écarts-types et troisièmes quartiles sont différents quelque soit la variable. Enfin, l'analyse de la part des sinistres en figure 19, en montant et par type de carburant, semble indiquer que les véhicules hybrides sont les moins dangereux puisque 10% des plus grands sinistres représentent 30% de leur charge sinistre, contre près de 60% pour les autres carburants. Bien que le meilleur modèle soit celui de la loi logamma, nous présentons dans la suite du rapport des résultats obtenus à partir d'une loi inverse gaussienne qui présente des résultats plus intéressants dans la tarification commerciale [3]. Les variables finalement retenues et leur catégorisation pour la modélisation sont indiquées dans le tableau 9. Les coefficients obtenus à partir d'une régression loggamma sont fournis dans le tableau 10.

3.2 Modélisation

3.2.1 Approche par écrêtement

Nous avons relevé en partie 1 un écart important entre montants de sinistres standards et élevés. Nous proposons donc d'écrêter les montants dans un premier temps. Nous choisissons un niveau de séparation $u = 6000$ €. Les sinistres supérieurs à ce montant représentent 2.46% des sinistres de la base et 25.5% de la charge sinistre totale. La charge surcrête devient ainsi $S_u = 120.65$ €.

Dans un second temps, les lois lognormale, gamma, loggamma et inverse gaussienne sont calibrées à partir des variables explicatives retenues en 3.1. Tout comme pour la sévérité, un modèle *stepwise* AIC permet une sélection de variables parmi celles que nous avons retenues. Une étude détaillée de la sélection de variables selon la méthode *stepwise* est disponible [4]. La comparaison des variables sélectionnées par les modèles *stepwise* pour chaque loi nous permet finalement de choisir les variables les plus robustes, à savoir l'âge du conducteur, celui du véhicule catégorisé, la valeur du véhicule en continue ainsi que son type puis finalement des variables portant sur la police : le bonus tel que présent dans la base (continu) et la durée et type de couverture catégorisées. Les métriques d'adéquation obtenues sur ces variables explicatives pour les 4 lois calibrées sont présentées dans la table 4 ci-dessous.

Au vu des métriques d'adéquation, la loi loggamma est la plus intéressante, toutes mesures confondues. Finalement, les résidus pour chaque lois sont représentés dans la figure 20, en annexe B.4. Nous remarquons que les lois gamma et inverse gaussienne semblent moins adaptées à la prédiction de montants plus élevés. Par ailleurs, les modèles lognormal et loggamma proposent des estimations de la moyenne différentes des deux lois précédentes.

TABLE 4 – Métriques d'adéquation pour les mêmes variables explicatives

Loi Lien	Gamma Logarithme	Inv. Gaussienne Logarithme	Lognormale Identité	Loggamma Identité
Log-vraisemblance	-103422.3	-101167.1	-18854.35	-18512.23
AIC	206876.6	202366.2	37740.71	37056.46
BIC	206996.2	202485.8	37860.3	37176.06
Deviance	15068.94	22.29224		
Null Dev.	15664.3	22.87138		

3.2.2 Approche par séparation

Nous proposons également d'implémenter l'approche par séparation mise en avant dans *computational actuarial science with R*[1]. Cette approche [5] consiste tout d'abord à décomposer l'espérance du montant sinistre Y de la façon suivante :

$$E(Y|X) = E(Y|X, Y \leq u) * P(Y \leq u|X) + E(Y|Y > u, X) * P(Y > u|X) \quad (1)$$

où u = 6 000€ est le montant de séparation et X le vecteur des variables explicatives. Le modèle décompose ainsi le coût moyen des sinistres en deux : une première partie "standard" pondérée de la probabilité que ce sinistre soit standard

3. les primes obtenus par des loi logarithmiques sont en effet deux fois moins élevées (voir 3.2.1), menant ainsi à des taux de chargement beaucoup plus importants
4. voir le fichier "2_reg_claim.html"
5. présenté en fin de *markdown* "2_reg_claim.html"

ainsi qu'une seconde partie "importante" pondérée de la probabilité que le sinistre soit important. Nous nous proposons d'implémenter cette approche dans le cas simple où l'âge est utilisé en tant que variable explicative. La probabilité d'observer un sinistre standard (ou important) est calculée par régression logistique sur l'âge lissé par *spline*[3]. Nous pouvons visualiser la probabilité d'observer un sinistre standard en fonction de l'âge de l'assuré sur la figure 21. Nous considérons alors deux modèles gamma, respectivement calibrés sur les sinistres standards ou sinistres importants. Les modèles correspondent alors respectivement à $E(Y|X, Y \leq u)$ et $E(Y|X, Y > u)$. Les mesures d'adéquations des modèles sont représentés ci-dessous.

TABLE 5 – Mesure d'adéquation pour les régressions gamma avec l'âge lissé

Loi	Gamma (sinistres standards)	Gamma (sinistres importants)
Lien	Logarithme	Logarithme
Log-vraisemblance	-103422.3	3207.297
AIC	193703.4	6424.594
BIC	193740.6	6443.451
Deviance	23957.95	6443.451
Null Dev.	24011.24	105.4028

Enfin, nous pouvons visualiser grâce à pour chaque classe d'âge, la part de la prime associée à chaque type de sinistre par notre modèle de séparation. La ligne horizontale figure 22 représente le coût moyen d'un sinistre. La ligne sombre à l'arrière, est une prédiction réalisée sur l'ensemble des données. A chaque âge, la partie sombre représente la part du sinistre moyen liée aux sinistres standard (inférieurs à 6 000€) et la zone plus claire est la part du sinistre moyen due à d'éventuels sinistres importants (supérieurs à 6 000€). Nous remarquons que les montants relatifs aux sinistres importants sont croissants avec l'âge (au moins jusqu'à 60 ans). Par ailleurs, les assurés les plus jeunes ainsi que les plus âgés présentent des montants liés aux deux types de sinistres, alors que les montants prédits pour la classe d'assurés de 30 à 60 ans sont majoritairement liés aux sinistres importants.

Bien que cette méthode propose des interprétations intéressantes, nous retenons, par souci de simplicité, la méthode par écrêtement présentée en 3.2.1 pour la suite du projet.

4 Calculs de primes

4.1 Prime pure

Nous avons modélisé la fréquence par un modèle binomial négatif et la sévérité par un modèle inverse gaussien. Suite à cette modélisation, nous pouvons calculer la prime pure $\Pi(X_i) = E[B_i]E[N_i]$ affectée à chaque police. Figures 23 et 24 sont représentées les primes pures en fonction de l'âge du conducteur et de l'âge du véhicule pour l'année zéro.

4.2 Prime commerciale

Par méthode de bootstrap non-paramétrique tel que proposée dans le cours nous simulons la charge portefeuille sur 10^4 scénarios. L'histogramme et la répartition empirique de cette charge simulée sont disponibles en annexe figure 25. Le tableau 6 récapitule la valeur, les quantiles et les écarts avec la prime pure pour différents quantiles. Nous remarquons que la prime pure protège contre le quantile d'ordre 7.56%, bien en dessous de la moyenne. Nous cherchons à construire une prime sur le principe de la valeur espérée $\Pi(X) = (1+\kappa)E[X]$, où κ doit permettre de couvrir la charge portefeuille dans 95% des cas. La figure 26 représente l'évolution entre le niveau de confiance et le taux de chargement κ appliqué. L'écart entre la prime pure et le quantile d'ordre 95% est de 8.896%. Ainsi, nous pouvons calculer pour chaque police i sa tarification en prime commerciale : $\Pi(X_i) = (1+\kappa)E[B_i]E[N_i]$ où $\kappa = 8.896\%$.

TABLE 6 – Moyenne et quantile en fonction de la somme des primes pures

	valeur (millions d'€)	quant. de S p	Écart (en %)
somme prime pure	13.627	0.076	0.000
moyenne charge ptf.	14.147	0.534	3.811
quant95	14.840	0.950	8.896
quant99	15.195	0.990	11.506
quant995	15.330	0.990	12.491

Références

[1] Arthur Charpentier. *Computational actuarial science with R / edited by Arthur Charpentier.* eng. Chapman Hall/CRC the R series (CRC Press). Boca Raton, FL : CRC Press/Taylor et Francis Group, 2015. ISBN : 9781466592605.

[2] Arthur Charpentier. *GLM et sélection de variables (stepwise).* 2020. URL : https://freakonometrics.hypotheses.org/60727.

[3] Arthur Charpentier. *Logistic Regression with splines.* 2020. URL : https://freakonometrics.hypotheses.org/52771@.

[4] Zhongheng Zhang. « Variable selection with stepwise and best subset approaches ». In : *Ann Transl of Med* (April, 2016). URL : https://atm.amegroups.com/article/view/9706.

Annexe

A Fréquence

A.1 Premières analyses

FIGURE 1 – Histogramme du nombre de sinistres

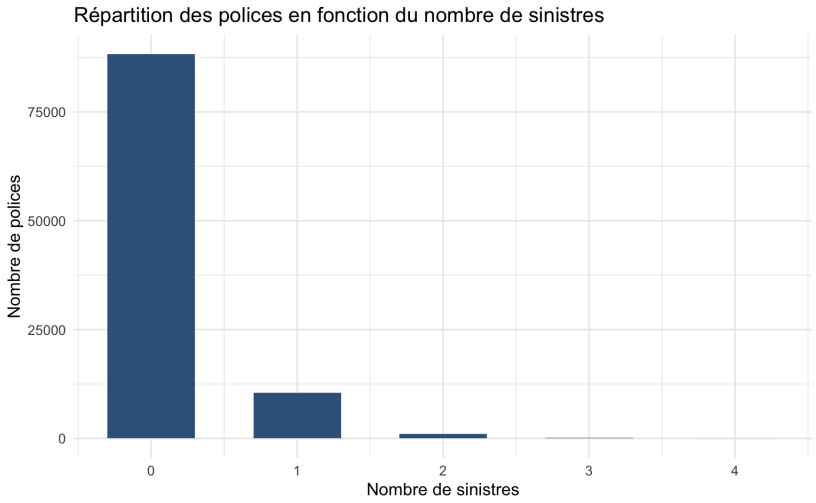


FIGURE 2 – Corrélations des variables de la police

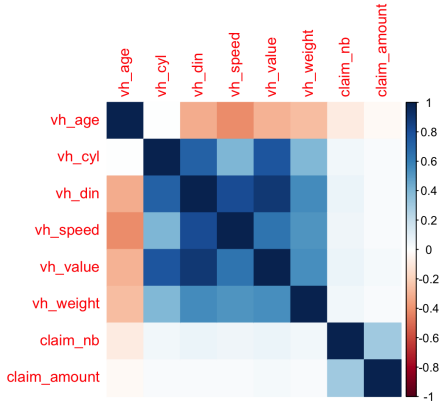


FIGURE 3 – Corrélations des variables du véhicule

A.2 Descriptions des variables retenues catégorisées

FIGURE 4 – Histogramme du nombre de sinistres - Âge des conducteurs

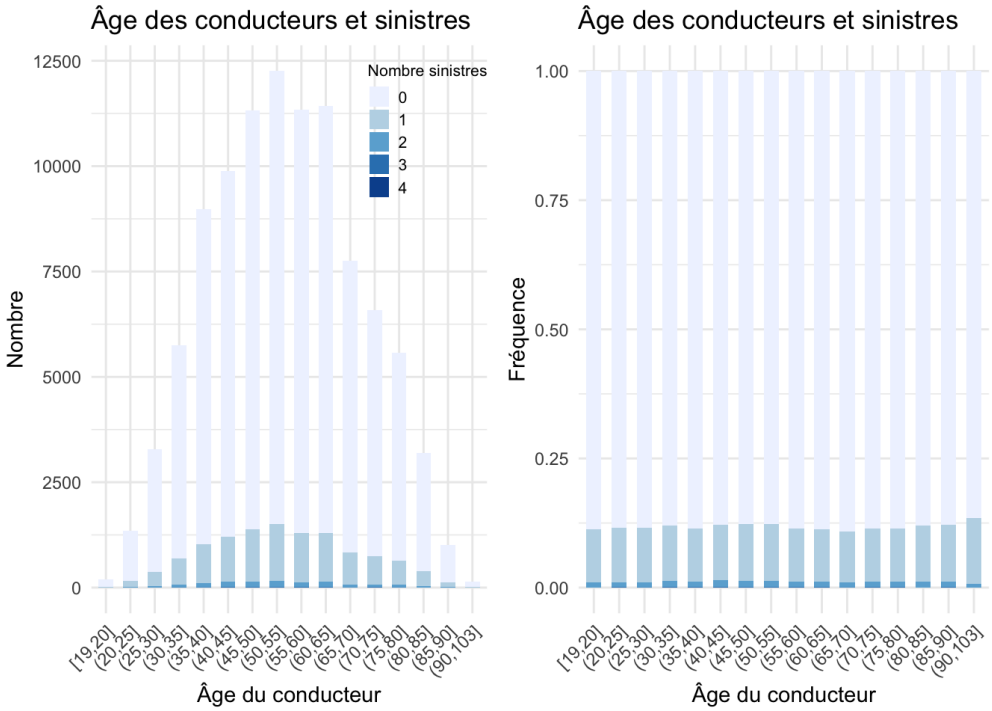


FIGURE 5 – Histogramme du nombre de sinistres - Durée de la police

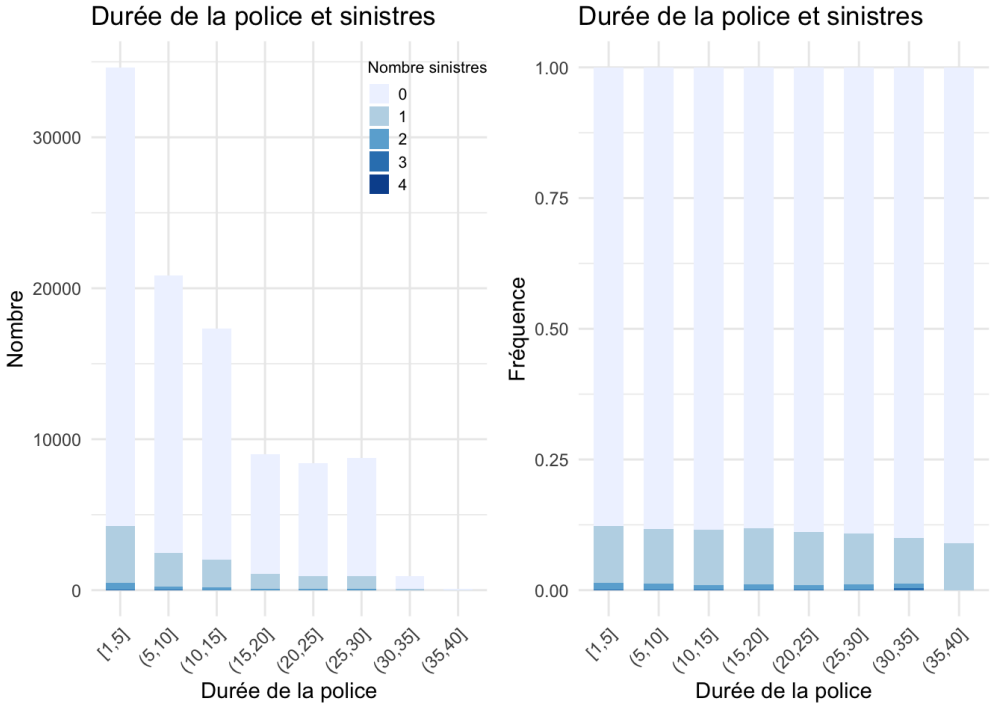


FIGURE 6 – Histogramme du nombre de sinistres - Âge du véhicule

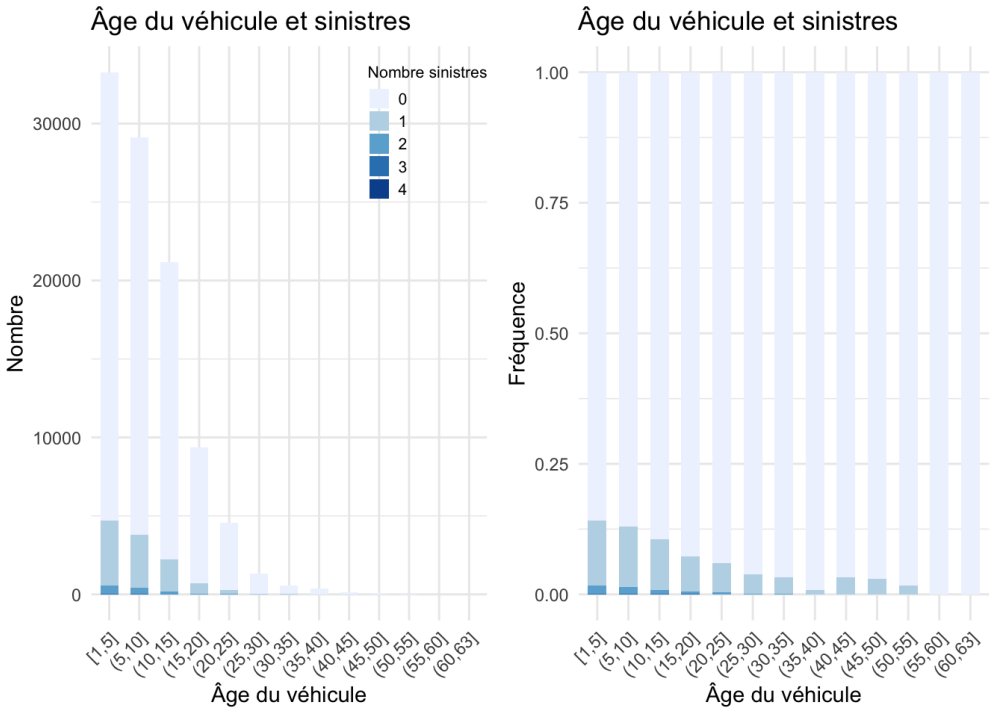


FIGURE 7 – Histogramme du nombre de sinistres - Valeur du véhicule

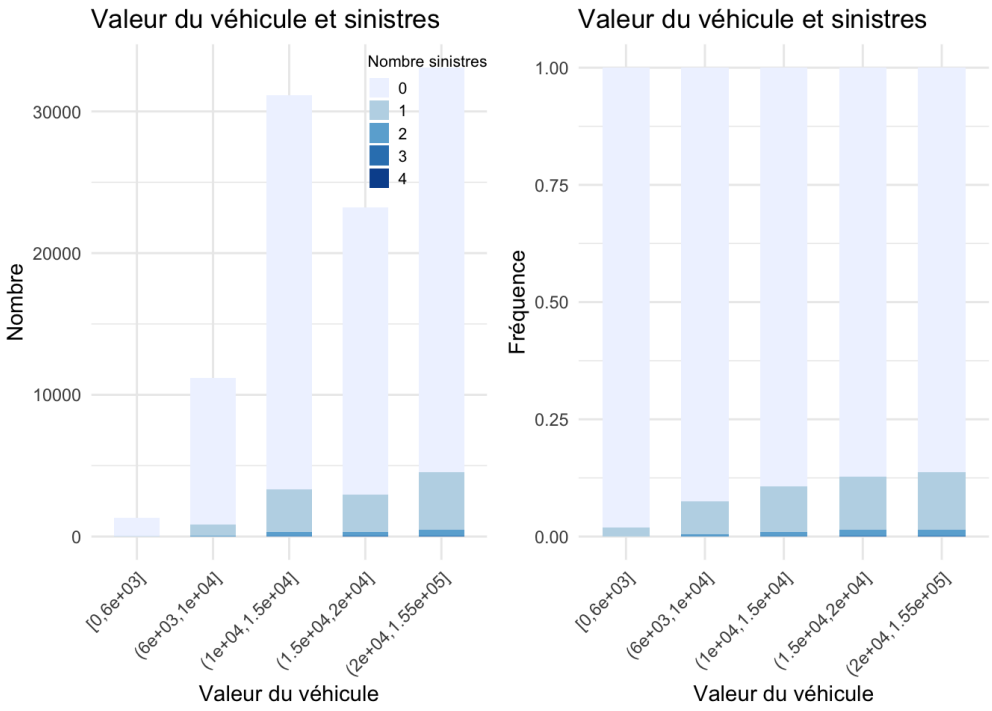


FIGURE 8 – Histogramme du nombre de sinistres - Puissance du véhicule

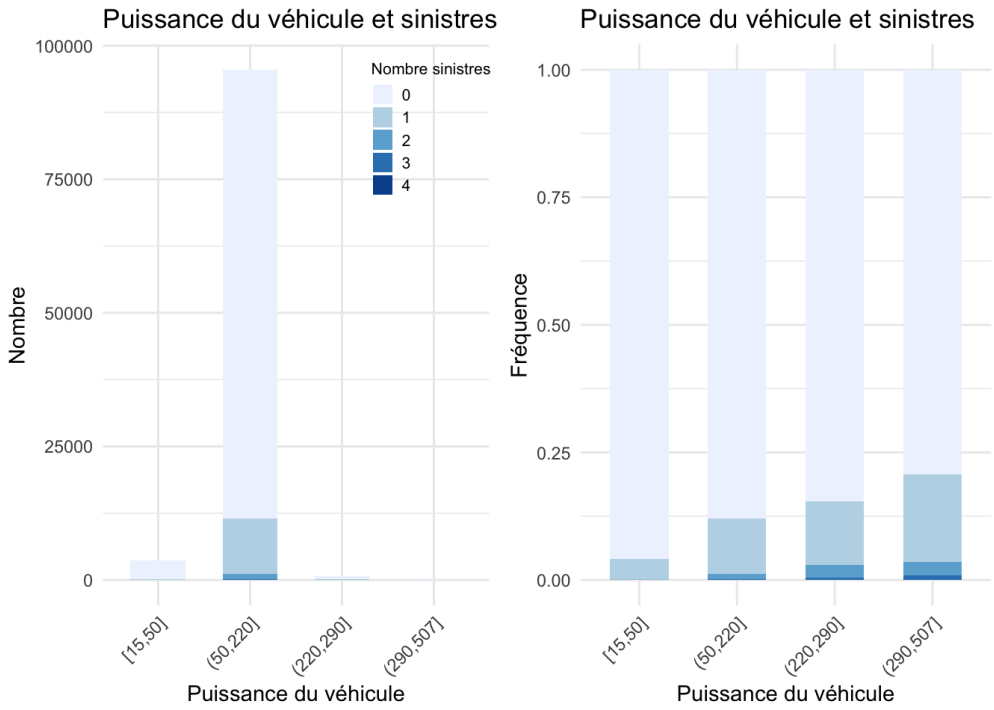


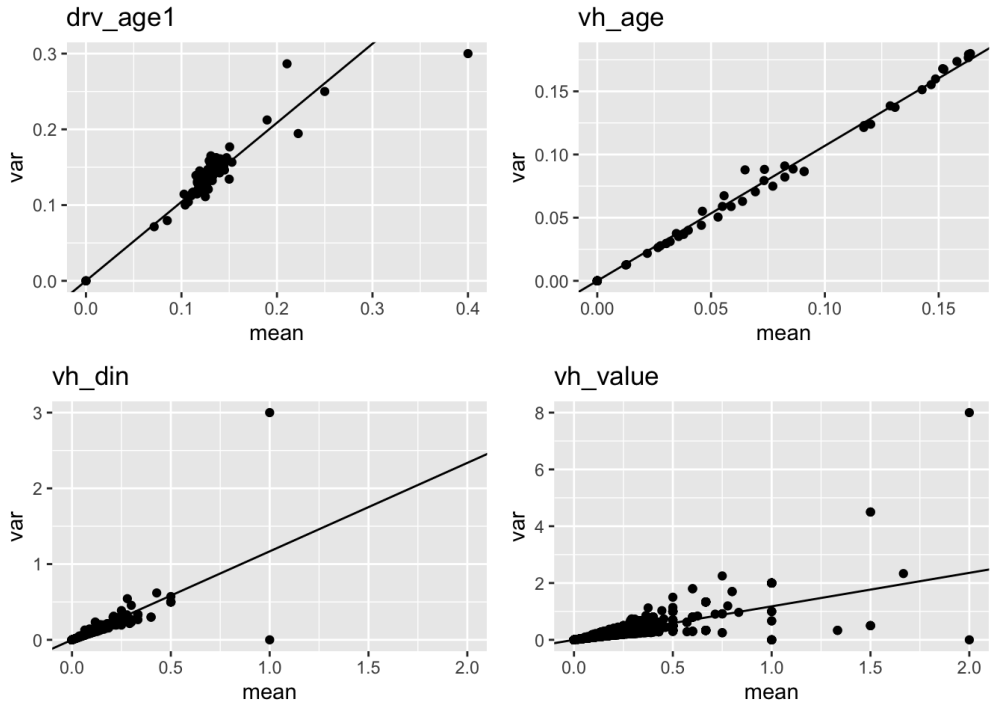
TABLE 7 – Variables retenues et catégorisation pour fréquence

Variable	Catégorisation
drv_age1	Par tranches de cinq ans : [19 ; 20],]20 ; 25], ...,]85 ; 90],]90 ; 103]
vh_age	Par tranches de cinq ans : [1 ; 5],]5 ; 10],...,]55 ; 60],]60 ; 63]
pol_bonus	Par tranches de 0,1 : [0,5 ; 0,6], ...,]1,5 ; 1,6],]1,6 ; 1,65]
pol_coverage	Les catégories Median1 et Median2 sont regroupées dans une seule catégorie Median
pol_duration	Par tranches de cinq ans : [1 ;5 20], ...,]35 ; 40]
pol_usage	Seule la catégorie Professional est distinguée des autres catégories regroupées dans Other
pol_region	Pas de regroupement
vh_din	Les catégories suivantes sont considérées : [15 ; 50],]50 ; 220],]220 ; 290],]290 ; 507]
vh_value	Les catégories suivantes sont considérées : [0 ; 6 000],]6 000 ; 10 000],]10 000 ; 15 000],]15 000 ; 20 000],]15 000 ; 20 000],]20 000 ; 550 000]
vh_fuel	Pas de regroupement
vh_make	Regroupement identique à celui du cours à l'exception que les véhicules japonais et coréens sont classés dans la catégorie Other : A, B, ..., G

A.3 Calcul du θ loi binomiale négative

Premièrement nous réalisons la régression $s_n^2(x) = 0 + x.m_n(x)$. Les droites de régressions obtenues sont disponibles figure 9. Au vu de ces régressions, le lien linéaire le plus net entre la moyenne et la variance est observé pour l'âge du véhicule. C'est sur cette variable que nous réalisons le calcul du θ avec la régression $s_n^2(x) = 0 + 1.m_n(x) + \frac{1}{\theta}.m_n^2(x)$. Nous obtenons $\theta = \frac{1}{0.53477} = 1.869963$.

FIGURE 9 – Étude de la relation moyenne-variance de la fréquence pour différentes variables



A.4 Régression binomiale négative

TABLE 8 – Coefficients pour la loi binomiale négative avec θ fixe

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	-3.0249	0.3331	-9.08	0.0000
drv_age1G(20,25]	0.1280	0.2358	0.54	0.5873
drv_age1G(25,30]	0.2148	0.2316	0.93	0.3535
drv_age1G(30,35]	0.3630	0.2302	1.58	0.1148
drv_age1G(35,40]	0.3432	0.2296	1.50	0.1349
drv_age1G(40,45]	0.4393	0.2295	1.91	0.0556
drv_age1G(45,50]	0.4380	0.2294	1.91	0.0563
drv_age1G(50,55]	0.4416	0.2294	1.92	0.0543
drv_age1G(55,60]	0.3698	0.2297	1.61	0.1074
drv_age1G(60,65]	0.3629	0.2297	1.58	0.1142
drv_age1G(65,70]	0.3052	0.2306	1.32	0.1857
drv_age1G(70,75]	0.3608	0.2309	1.56	0.1181
drv_age1G(75,80]	0.3618	0.2314	1.56	0.1179
drv_age1G(80,85]	0.4025	0.2336	1.72	0.0849
drv_age1G(85,90]	0.4386	0.2447	1.79	0.0731
drv_age1G(90,104]	0.4206	0.3217	1.31	0.1911
vh_ageG(5,10]	-0.0899	0.0215	-4.17	0.0000
vh_ageG(10,15]	-0.2711	0.0260	-10.44	0.0000
vh_ageG(15,20]	-0.6158	0.0413	-14.92	0.0000
vh_ageG(20,25]	-0.7604	0.0645	-11.78	0.0000
vh_ageG(25,30]	-1.0665	0.1440	-7.41	0.0000
vh_ageG(30,35]	-0.9535	0.2515	-3.79	0.0001
vh_ageG(35,40]	-2.2170	0.5924	-3.74	0.0002
vh_ageG(40,45]	-0.8237	0.4794	-1.72	0.0857
vh_ageG(45,50]	-0.9167	0.6004	-1.53	0.1268
vh_ageG(50,55]	-1.5021	1.0146	-1.48	0.1387
vh_ageG(55,60]	-9.3643	72.3479	-0.13	0.8970
vh_ageG(60,67]	-9.6858	108.4862	-0.09	0.9289
pol_durationG(5,10]	-0.0386	0.0252	-1.53	0.1250
pol_durationG(10,15]	-0.0837	0.0273	-3.07	0.0021
pol_durationG(15,20]	-0.0511	0.0343	-1.49	0.1363
pol_durationG(20,25]	-0.1278	0.0362	-3.53	0.0004
pol_durationG(25,30]	-0.1474	0.0359	-4.11	0.0000
pol_durationG(30,35]	-0.1488	0.1008	-1.48	0.1399
pol_durationG(35,42]	-0.4009	0.3250	-1.23	0.2174
vh_valueG(6e+03,1e+04]	0.5931	0.2447	2.42	0.0154
vh_valueG(1e+04,1.5e+04]	0.7464	0.2475	3.02	0.0026
vh_valueG(1.5e+04,2e+04]	0.8601	0.2480	3.47	0.0005
vh_valueG(2e+04,Inf]	0.8776	0.2480	3.54	0.0004
vh_dinG(50,220]	0.1355	0.0971	1.39	0.1631
vh_dinG(220,290]	0.3383	0.1354	2.50	0.0125
vh_dinG(290,555]	0.7633	0.2256	3.38	0.0007
pol_bonusG(0.6,0.7]	0.1832	0.0409	4.47	0.0000
pol_bonusG(0.7,0.8]	0.1721	0.0494	3.48	0.0005
pol_bonusG(0.8,0.9]	0.2358	0.0694	3.40	0.0007
pol_bonusG(0.9,1]	0.4157	0.0816	5.10	0.0000
pol_bonusG(1,1.1]	0.3400	0.2343	1.45	0.1468
pol_bonusG(1.1,1.2]	0.4298	0.2017	2.13	0.0331
pol_bonusG(1.2,1.3]	-0.3259	0.5151	-0.63	0.5269
pol_bonusG(1.3,1.4]	0.1399	0.5943	0.24	0.8138
pol_bonusG(1.4,1.5]	0.8962	0.6268	1.43	0.1528
pol_bonusG(1.5,1.6]	0.7632	1.0472	0.73	0.4661
pol_bonusG(1.6,1.65]	-9.6483	220.4120	-0.04	0.9651
pol_coverageGMedian	-0.2227	0.0216	-10.32	0.0000
pol_coverageGMini	-0.6450	0.0415	-15.56	0.0000
pol_usageGProfessional	0.0998	0.0336	2.97	0.0029
vh_fuelGasoline	-0.1153	0.0218	-5.30	0.0000
vh_fuelHybrid	0.0121	0.2801	0.04	0.9655

B Statistiques pour la sévérité

B.1 Premières analyses

FIGURE 10 – Fonction de répartition empirique

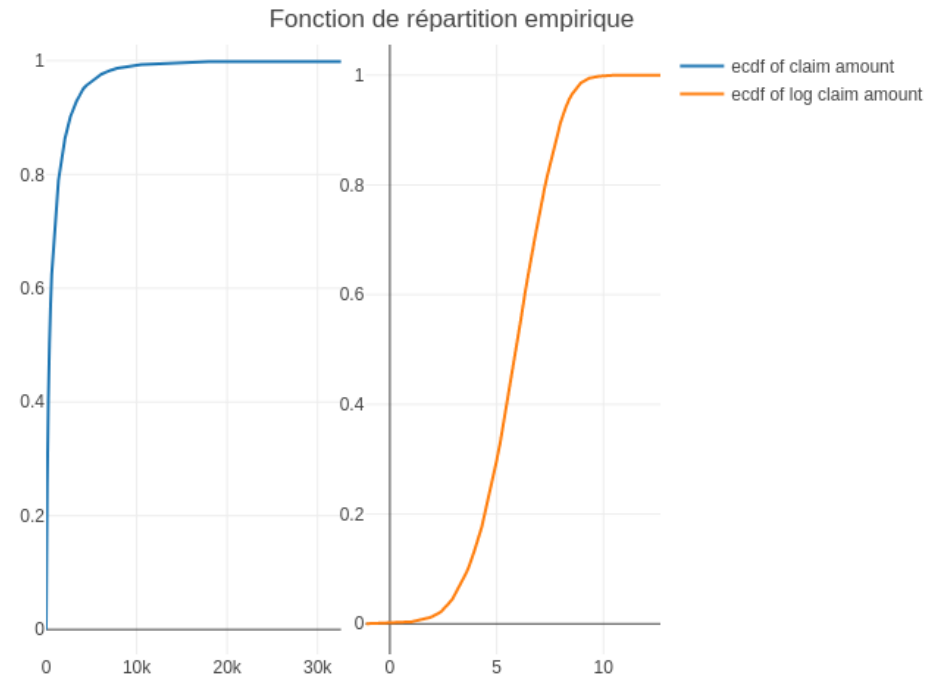


FIGURE 11 – Boxplot des montants et log montants

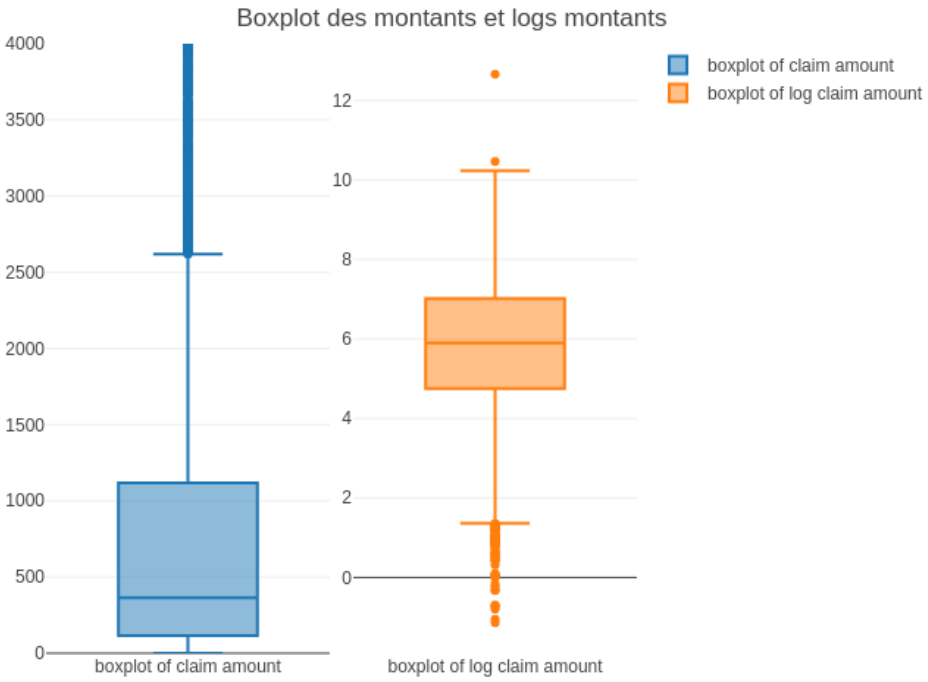
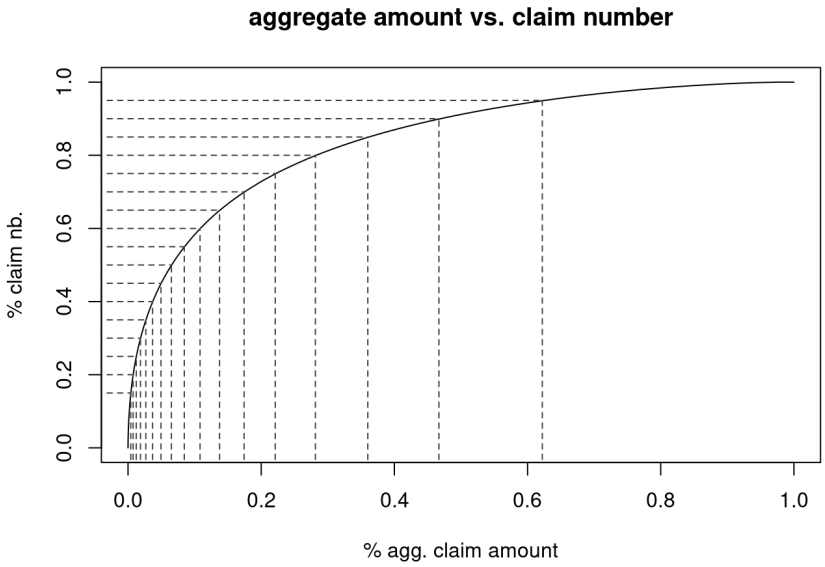


FIGURE 12 – Part de sinistres en nombre et montants



B.2 Descriptions des variables retenues catégorisées

TABLE 9 – Variables retenues et catégorisation pour sévérité

Variable	**Catégorisation**
drv_age1	Pas de regroupement (en continue)
drv_age1	Par tranches de 10 ans : [18 ; 20],]20 ; 230], ...,]80 ; 90],]90 ; 103]
drv_age1	Les catégories suivantes : [18 ; 25],]25 ; 45], ...,]45 ; 103]
vh_age	Pas de regroupement (en continue)
vh_age	Les catégories suivantes sont considérées : [1 ; 10],]10 ; 25]]25 ; 30],]30 ; 100]
pol_bonus	Pas de regroupement (en continue)
pol_bonus	Les catégories suivantes :[0.5 ; 0.9],]0.9 ; 1.2],]1.2 ; 1.4],]1.4 ; 1.95]
pol_coverage	Les catégories Median1 et Median2 sont regroupées dans une seule catégorie Median
pol_duration	Par tranches de cinq ans : c(1,10, 20,25,30, 42) [1 ; 10],]10 ; 20],]20 ; 25],]25 ; 30],]30 ; 42]
vh_din	Pas de regroupement (en continue)
vh_din	Les catégories suivantes sont considérées : [15 ; 50],]50 ; 220],]220 ; 555]
vh_value	Pas de regroupement (en continue)
vh_type	Pas de regroupement

FIGURE 13 – Boxplot Montant - Âge du conducteur

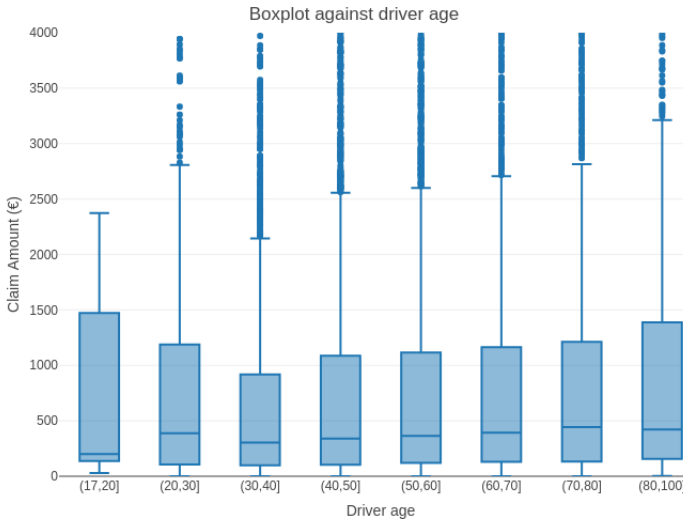


FIGURE 14 – Boxplot Montant - Âge du véhicule

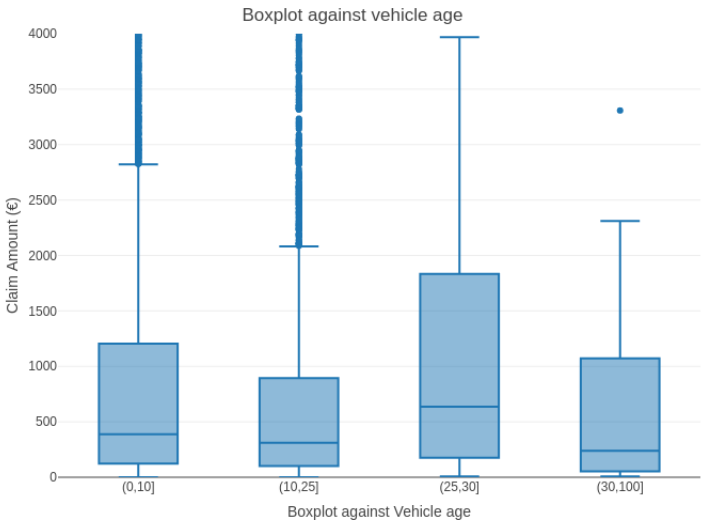


FIGURE 15 – Boxplot Montant - Valeur du véhicule

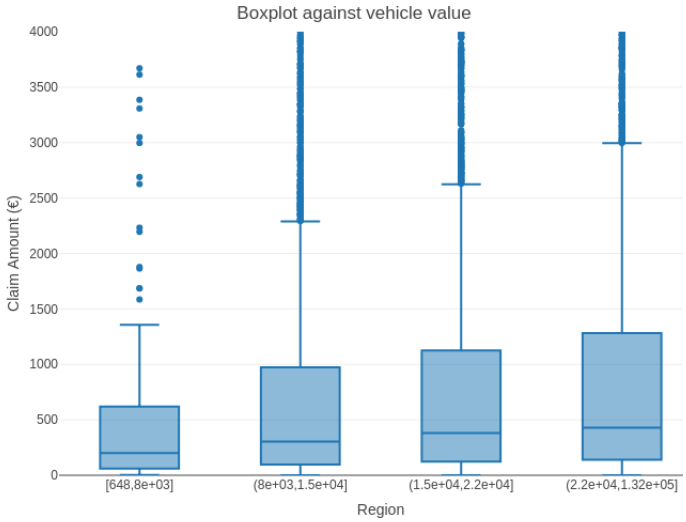


FIGURE 16 – Boxplot Montant - Puissance du véhicule

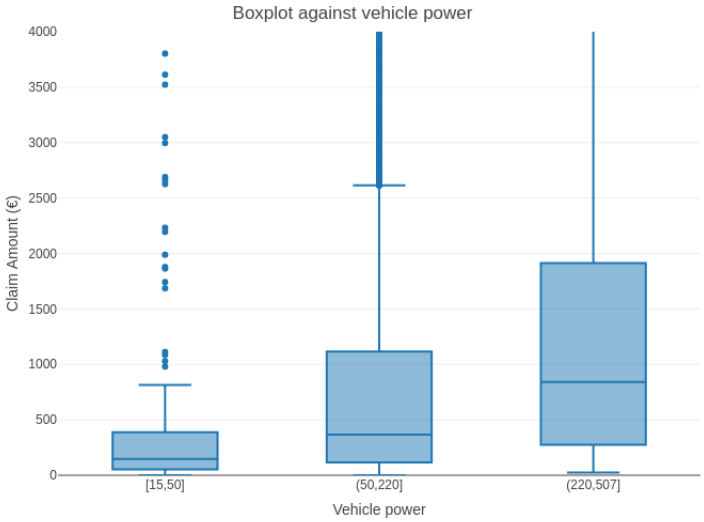


FIGURE 17 – Boxplot Montant - Bonus police

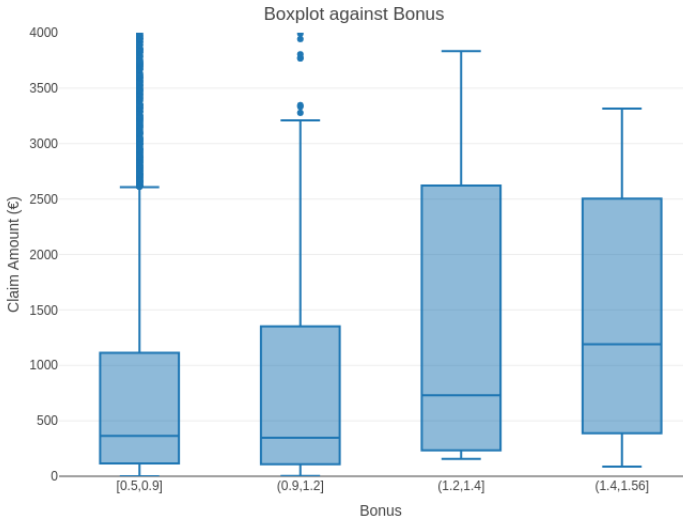


FIGURE 18 – Boxplot Montant - Couverture police

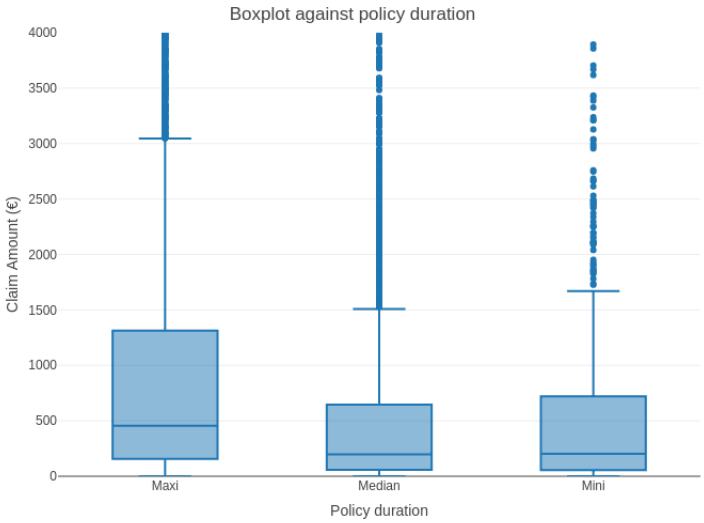
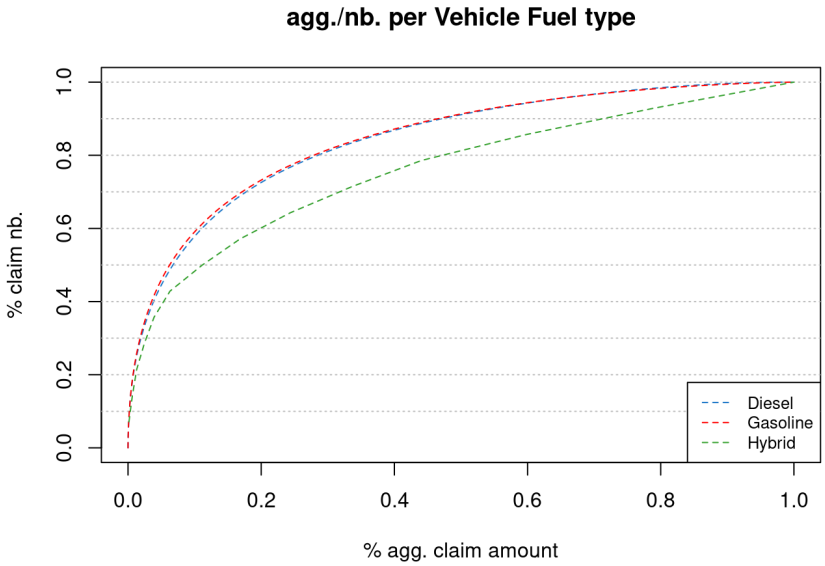


FIGURE 19 – Part de sinistres en nombre et montants selon le carburant



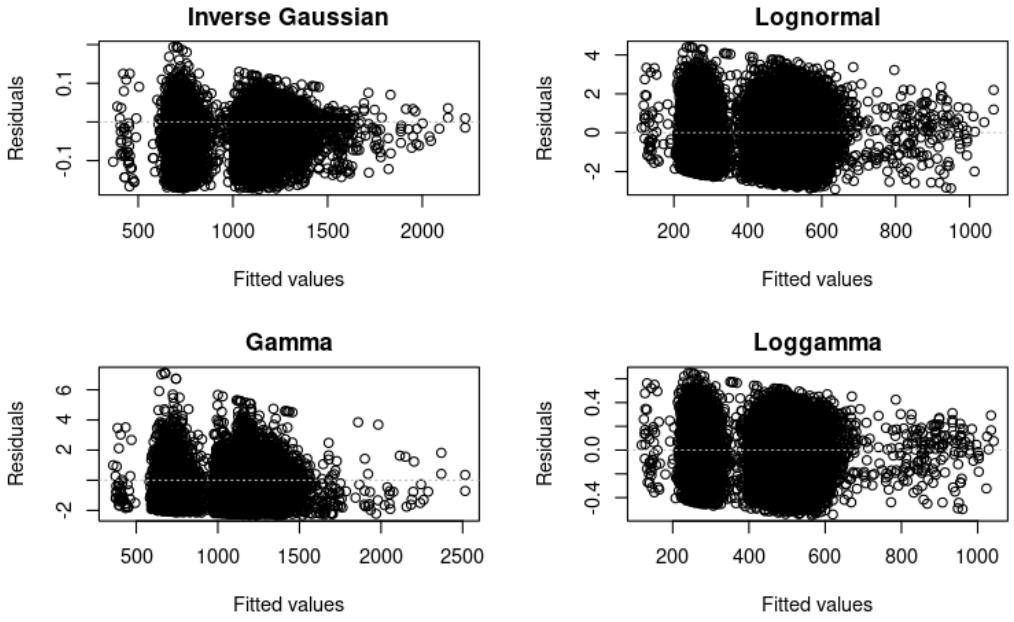
B.3 Régression Inverse Gaussienne

TABLE 10 – Coefficients pour la loi inverse gaussienne

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.4690	0.1419	45.58	0.0000
drv_age1	0.0022	0.0009	2.58	0.0099
vh_ageG(10,25]	-0.0960	0.0265	-3.62	0.0003
vh_ageG(25,30]	0.3249	0.2141	1.52	0.1293
vh_ageG(30,100]	-0.0670	0.2212	-0.30	0.7619
vh_dinG(50,220]	0.2763	0.0958	2.88	0.0039
vh_dinG(220,555]	0.3650	0.1710	2.14	0.0328
vh_value	0.0000	0.0000	3.52	0.0004
vh_typeTourism	-0.0788	0.0426	-1.85	0.0644
pol_bonus	0.2824	0.1217	2.32	0.0204
pol_durationG(10,20]	0.0336	0.0282	1.19	0.2330
pol_durationG(20,25]	-0.0337	0.0438	-0.77	0.4422
pol_durationG(25,30]	0.1267	0.0469	2.70	0.0069
pol_durationG(30,42]	-0.0142	0.1203	-0.12	0.9062
pol_coverageGMedian	-0.3830	0.0256	-14.97	0.0000
pol_coverageGMini	-0.3962	0.0477	-8.31	0.0000

B.4 Résidus et estimations

FIGURE 20 – Graphiques des résidus par loi



B.5 Séparation

FIGURE 21 – Probabilité d'avoir un sinistre standard, étant donné qu'un sinistre est survenu, en fonction de l'âge du conducteur

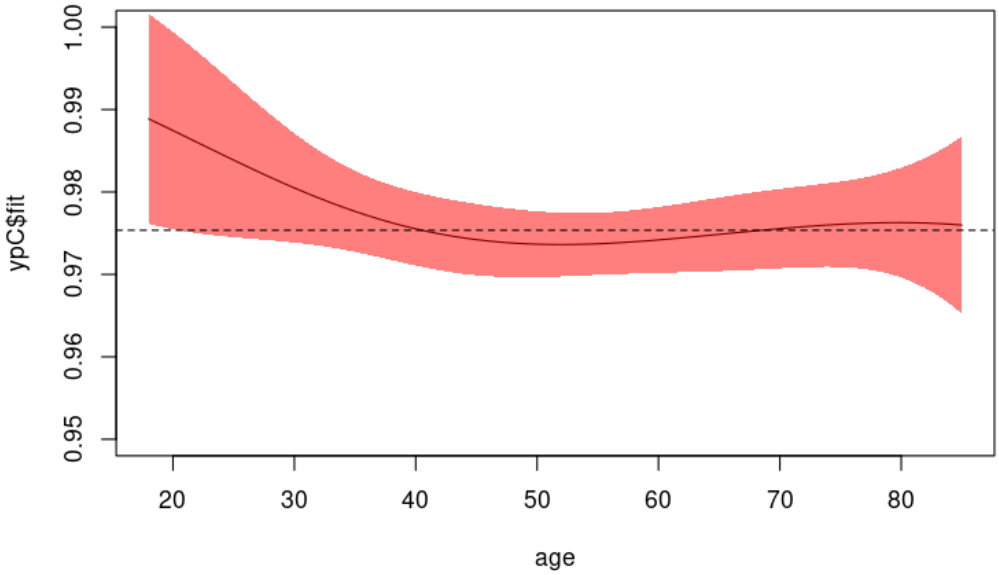
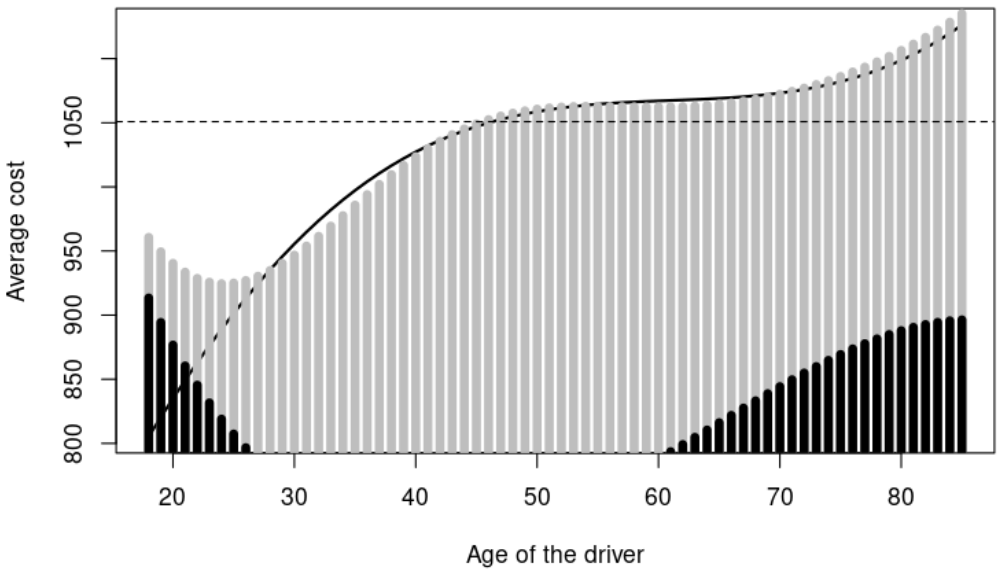


FIGURE 22 – Coût moyen d'un sinistre, en fonction de l'âge du conducteur, u =6 000€



C Calculs de primes

C.1 Prime pure

FIGURE 23 – Primes pures - Âge du conducteur

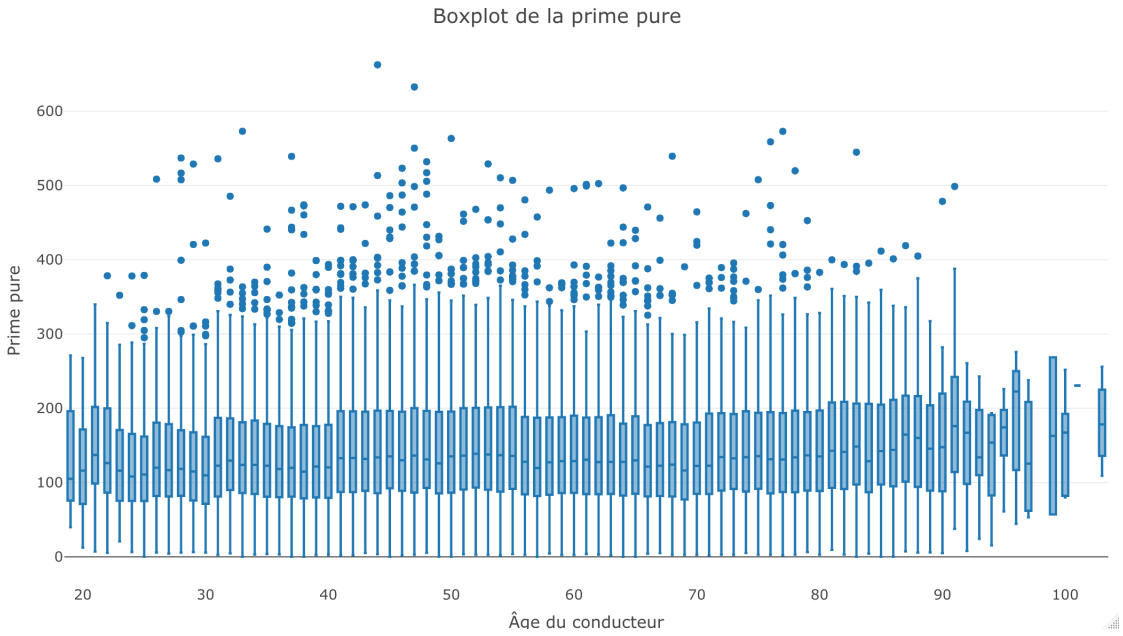
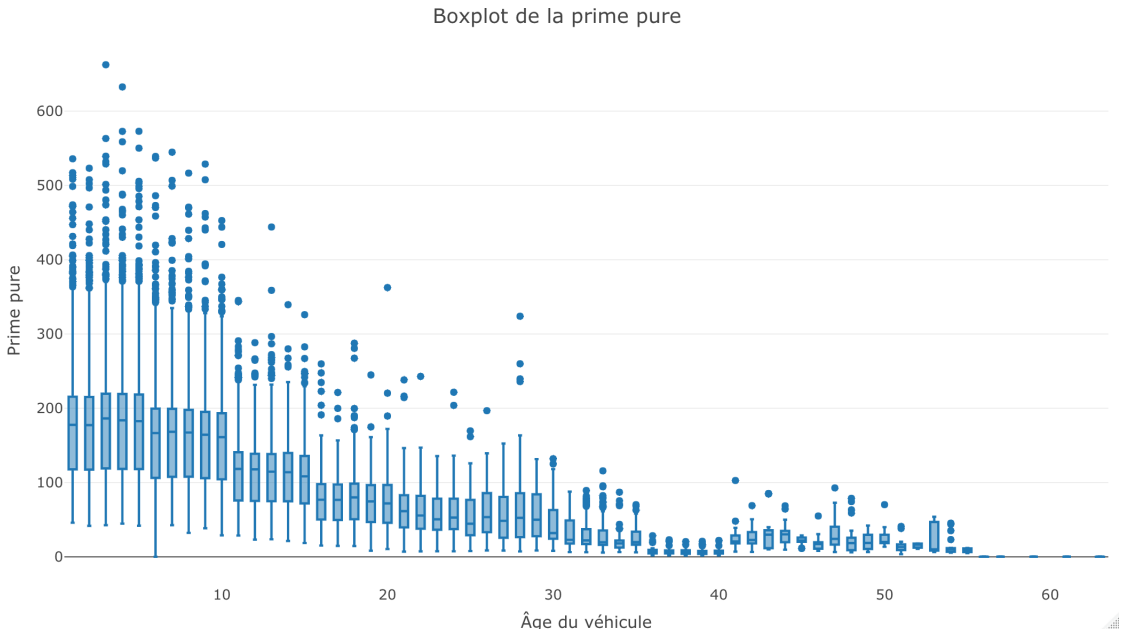


FIGURE 24 – Primes pures - Âge du véhicule



C.2 Prime commerciale

FIGURE 25 – Simulation de la charge sinistre globale du portefeuille de validation sur 10^4 scénarios

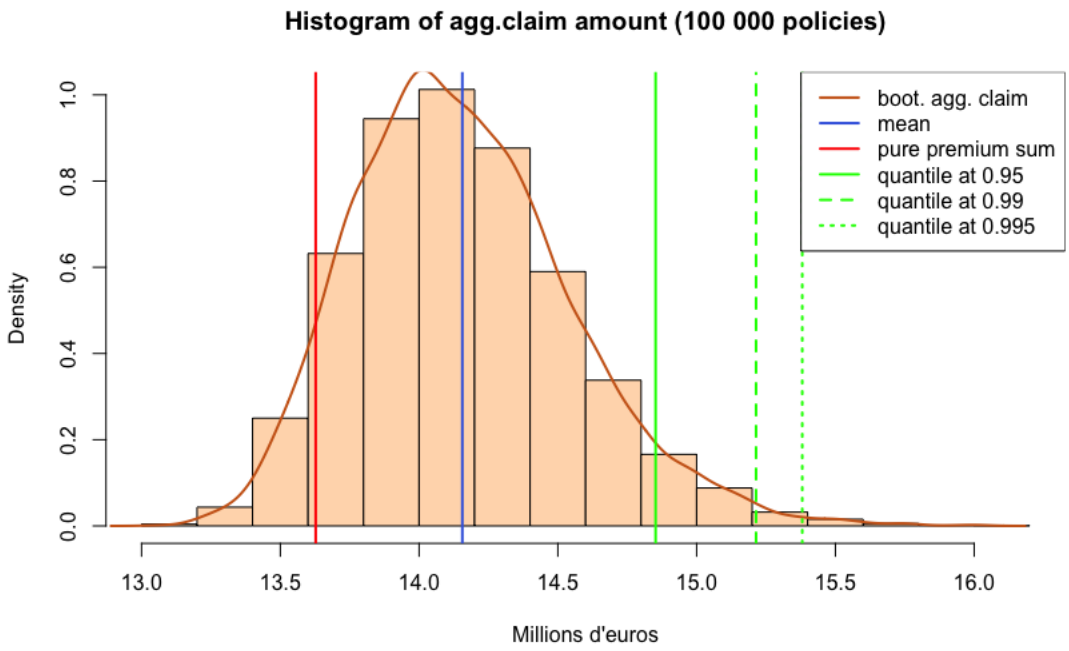


FIGURE 26 – Niveau de confiance en fonction du taux de chargement