LÉTOCART
Pierre-Nicolas
CHAPERON Romain
NONNENMACHER
Alexandre

ENSAE 3ème année
Année scolaire 2025–2026

Projet Actuariat IARD : Tarification en assurance dommage

Étude de tarification a priori d'un portefeuille automobile : GLM, Validation et Modèle Lasso

Abstract

Ce rapport présente une démarche complète de tarification a priori pour un portefeuille d'assurance automobile à l'aide des données `freMTPLfreq` et `freMTPLsev` du package `CASdatasets`. Les modèles retenus concernent la fréquence et la sévérité des sinistres, estimés via des GLM (une densité de poisson et Gamma sont respectivement retenues) ainsi qu'un modèle pénalisé Lasso. Un tarif final est construit et est comparé selon les différentes approches.

Contents

1 Introduction

1.1 Objectifs du projet

L'objectif de ce projet est de construire un tarif a priori fondé sur les caractéristiques des contrats du portefeuille étudié. Nous estimons dans un premier temps la fréquence et la sévérité des sinistres en utilisant des modèles linéaires généralisés.

Ces modèles sont ensuite comparés à une approche pénalisée de type Lasso, permettant une sélection automatique des variables et une amélioration potentielle de la parcimonie du modèle.

L'ensemble des modèles proposés est évalué et validé à l'aide d'un échantillon test, afin de mesurer leur performance prédictive et de retenir la solution la plus pertinente pour la tarification.

1.2 Description des données

Les données utilisées dans ce projet proviennent des jeux `freMTPLfreq` et `freMTPLsev`, issus du package `CASdatasets`.

Le premier correspond au volet « fréquence » et contient dix variables décrivant les caractéristiques des contrats d'assurance automobile : l'identifiant de police (*PolicyID*), le nombre de sinistres déclarés (*ClaimNb*), la durée d'exposition (*Exposure*), la puissance du véhicule (*Power*), son âge (*CarAge*), l'âge du conducteur (*DriverAge*), la marque du véhicule (*Brand*), le type de carburant (*Gas*), la région d'appartenance (*Region*) ainsi que la densité de population de la zone de résidence (*Density*).

Le second jeu de données, `freMTPLsev`, correspond à la sévérité et contient pour chaque sinistre l'identifiant de police (*PolicyID*) permettant le lien avec les contrats, ainsi que le coût associé (*ClaimAmount*), évalué à une date récente. Au total, les deux bases regroupent les informations relatives à 413,169 contrats d'assurance responsabilité civile automobile, observés pour la plupart sur une année.

2 Exploration des données

2.1 Statistiques descriptives

L'analyse descriptive des données met en évidence des distributions hétérogènes selon les variables explicatives du portefeuille.

Les variables continues telles que l'âge du conducteur (*DriverAge*) et l'âge du véhicule (*CarAge*) présentent des distributions asymétriques : la majorité des conducteurs ont entre 25 et 60 ans, tandis que les véhicules sont principalement récents, avec une concentration entre 0 et 10 ans.

La densité de population (*Density*) est fortement dispersée et montre une distribution très asymétrique, reflétant la grande diversité des zones géographiques couvertes.

Les variables catégorielles comme la marque du véhicule (*Brand*), le carburant (*Gas*) ou la région (*Region*) sont relativement équilibrées, bien que certains groupes (par exemple les marques françaises) soient surreprésentés.

Concernant les variables cibles, le nombre de sinistres (*ClaimNb*) est très faiblement dispersé et présente une distribution fortement concentrée autour de zéro, caractéristique des données de fréquence en assurance, où la grande majorité des assurés n'ont aucun sinistre sur la période. Enfin, le montant des sinistres (*ClaimAmount*) est très asymétrique, avec une majorité de petits sinistres et quelques valeurs extrêmes, typique d'une distribution de coûts lourde et à queue épaisse.

2.2 Premières visualisations

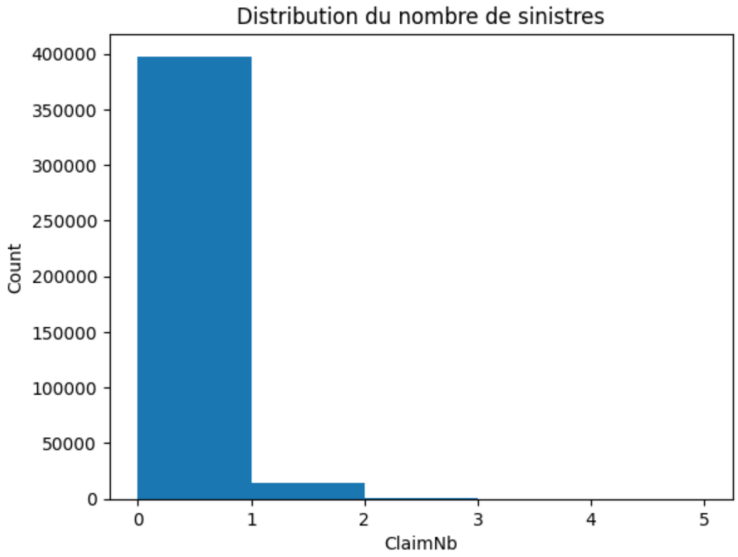

Figure 1: Histogramme du nombre de sinistres.

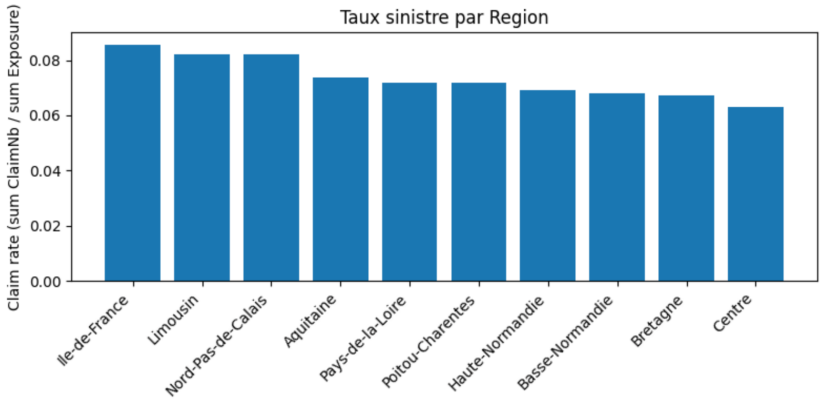

Figure 2: Histogramme du taux de sinistre par région

3 Méthodologie

3.1 Découpage apprentissage / test

Pour construire et évaluer nos modèles de tarification, nous avons séparé l'ensemble des observations en deux sous-échantillons distincts. 75% des données ont été utilisées pour l'apprentissage des modèles, tandis que les 25% restantes ont été réservées à leur validation. Cette répartition a été effectuée au moyen d'un tirage aléatoire, de manière à garantir que les deux échantillons soient représentatifs du portefeuille initial.

3.2 Choix des modèles

Pour modéliser la fréquence des sinistres, nous avons retenu un GLM de type Poisson avec une fonction de lien logarithmique, ce qui permet de modéliser une variable de comptage tout en intégrant un offset pour tenir compte de la durée d'exposition. La sévérité des sinistres a été estimée à l'aide d'un GLM Gamma également associé à une fonction de lien log, adaptée aux montants strictement positifs et à leur forte asymétrie.

Les différents modèles ont été comparés à l'aide du critère AIC ainsi que des tests de significativité, permettant d'identifier les variables explicatives les plus pertinentes. Enfin, un modèle pénalisé Lasso, implémenté via le package `glmnet` et optimisé par validation croisée, a été utilisé afin d'évaluer l'apport d'une sélection automatique des variables et d'étudier la parcimonie des modèles obtenus.

4 Modélisation de la fréquence

4.1 Modèle Poisson et analyse de la surdispersion

Après étude de la densité de la fréquence de sinistres par contrat, nous avons choisi de modéliser la fréquence des sinistres par une loi de Poisson. On note Y_i le nombre de sinistres observés pour le contrat i, et X_i le vecteur des covariables (ex. âge du conducteur, type de véhicule, densité de la population, région, etc.). Le modèle de Poisson s'écrit :

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{avec} \quad \lambda_i = \exp(X_i^\top \beta), \tag{1}$$

où λ_i est le nombre moyen de sinistres attendu pour le contrat i, et β le vecteur des coefficients à estimer.

4.1.1 Prise en compte de la durée d'exposition

Dans notre portefeuille, chaque contrat n'est pas observé sur une durée identique : la variable `Exposure` indique le nombre d'années (ou fractions d'années) pendant lesquelles le risque a été couvert. Pour modéliser correctement la fréquence annuelle, nous incluons donc un *offset* $\log(\text{Exposure}_i)$ dans le modèle de Poisson.

Le modèle complet devient alors :

$$\log(\lambda_i) = \log(\text{Exposure}_i) + X_i^\top \beta,$$

ce qui revient à modéliser non pas le nombre de sinistres bruts N_i, mais le taux annuel de sinistres par unité d'exposition,

$$\mu_i = \frac{\lambda_i}{\text{Exposure}_i}.$$

Ainsi, les coefficients β s'interprètent comme des effets multiplicatifs sur le taux de sinistres, et la prédiction finale tient automatiquement compte de la durée d'exposition via :

$$\hat{\lambda}_i = \text{Exposure}_i \times \exp(X_i^\top \hat{\beta}).$$

4.1.2 Diagnostic de surdispersion

Le modèle de Poisson suppose que $\text{Var}(Y_i) = \lambda_i$. En pratique, la variance observée peut être supérieure à la moyenne, ce qui indique une surdispersion. Un diagnostic simple est le rapport du chi-deux de Pearson sur les degrés de liberté :

$$\hat{\phi} = \frac{\sum_i \frac{(Y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}}{n - p - 1}. \tag{2}$$

Si $\hat{\phi} > 1$, le modèle Poisson sous-estime la variance, et il est conseillé d'envisager :

- un modèle binomial négatif,
- ou un ajustement des erreurs standards (*quasi-Poisson*).

Ainsi, comme notre calcule de la dispersion indique que $\hat{\phi} = 1{,}73$, alors, nous avons mit en évidence une sur-dispersion significative. Un modèle de Poisson standard sur ces données conduit à sous-estimer la variance réelle des estimateurs.

Pour pallier ce problème et sélectionner rigoureusement les variables, nous optons pour une approche Quasi-Poisson. Ce modèle permet de relâcher la contrainte de variance.

4.1.3 Estimation des coefficients

Le modèle Quasi-Poisson nous permet de valider la significativité de nos variables explicatives.

les résultats obtenus sont résumé dans le graphique ci-dessous:

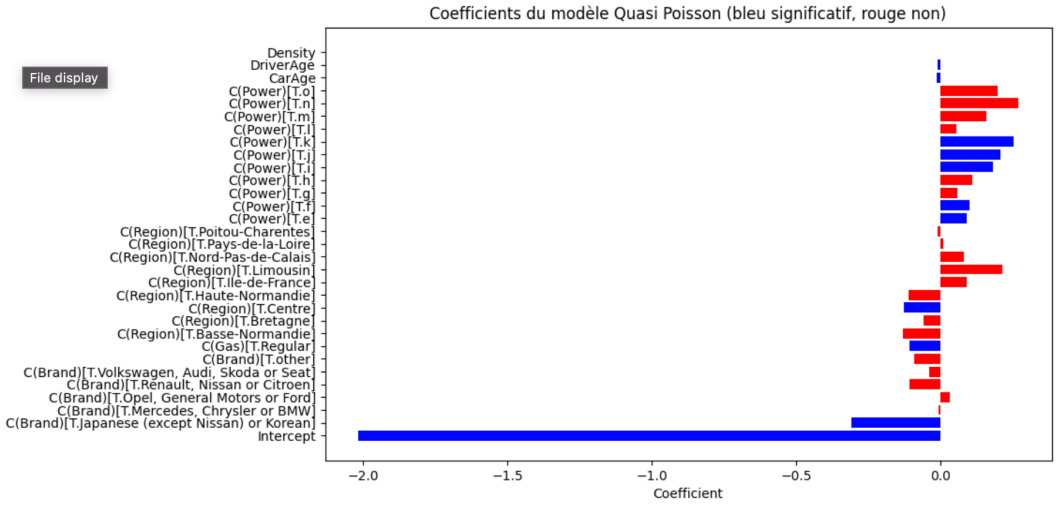

Figure 3: Grapgique de selection de variable via le modèle Quasi-poisson

4.2 Sélection du modèle via AIC

Nous venons d'utiliser une loi de quasi poissson pour faire de la sélection de variable. Ce modèle s'y prête bien, néanmoins pour comparer plusieurs modèles avec une métrique telle que l'AIC, ceci n'est pas possible avec une loi quasi poisson, qui n'a pas de fonction de vraisemblance. Nous allons donc comparer un modèle de Poisson, et un modèle de loi Binomiale Négative, basés sur la sélection de variables effectuée précedemment, pour déterminer quel modèle est le plus performant.

```
1. CRITÈRE D'INFORMATION D'AKAIKE (AIC) :
   AIC Poisson : 101491.2028
   AIC Binomial Négatif : 101429.2191

2. ERREUR ABSOLUE MOYENNE PONDÉRÉE :
   MAE Poisson sur Test : 0.133483
   MAE Binomial Négatif sur Test : 0.133521
```

Figure 4: Résultats critère AIC + Erreur absolues moyenne pondérée sur le groupe Test

Ces résultats permettent de conclure sur l'utilisation du modèle. On choisira un modèle de Poisson pour la modélisation de la fréquence, puisqu'une loi Binomiale Négative n'apporte pas de gain significatif en AIC.

4.3 Validation sur les données test

La loi Binomial négative présente une performance sur le jeu de test légèrement moins bonne qu'un modèle Poisson. On décide donc de valoriser la simplicité et l'interprétabilité, le modèle de Poisson étant plus simple qu'une Binomiale Négative.

5 Modélisation de la sévérité

5.1 Statistique descriptive

5.1.1 Analyse globale

Nous menons notre analyse statistiques pour l'ensemble des polices ayant déclaré au moins un sinistre et ne considérons pas la fréquence des sinistres.

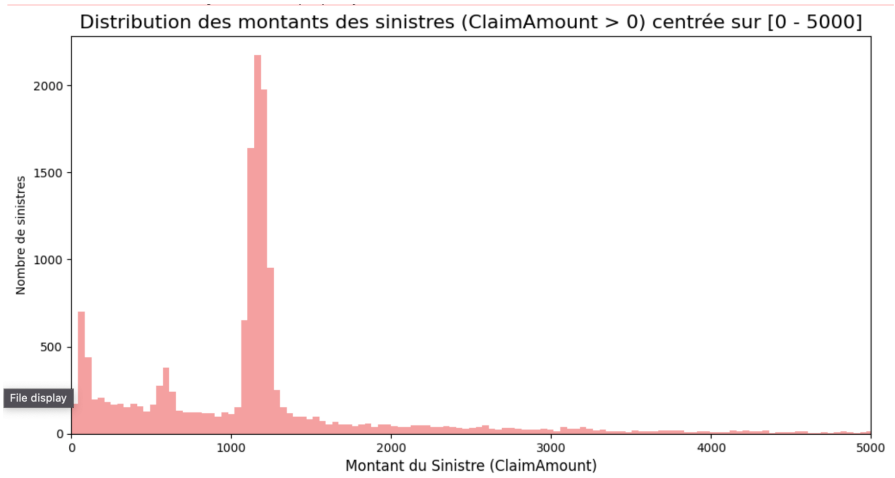

Figure 5: Distribution des montant des sinistres

La visualisation de la distribution révèle une distribution multimodale avec un pic notable entre 1 100 € et 1 300 €, potentiellement lié au montant forfaitaire de recours inter-compagnies (convention IRSA).

Le montant moyen des sinistres est de 2 130 €. La médiane (1 156 €) très inférieure à la moyenne et l'écart-type très élevé (21 064 €) par rapport à la moyenne indiquent une distribution extrêmement asymétrique et fortement étalée à droite.

5.1.2 Analyse des variations de la distribution en fonction des variables explicatives

Nous intuitons des effets croisé entre la marque (Brand) et la puissance (Power). Nous excluons les valeurs supérieures au 1 % supérieur[1].

[1]Principalement pour des raisons de lisibilité. La queue de distribution est analysée en détail par la suite

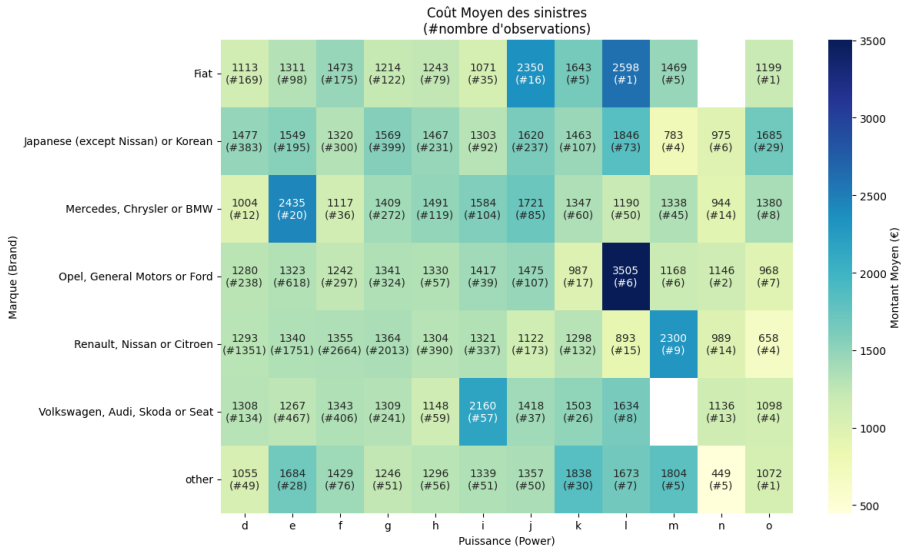

Figure 6: Coût moyen des sinistres en fonction de la marque et de la puissance

La table croisée montre que le montant moyen des sinistres (écrêté au 99e percentile) reste relativement constant pour les premières classes de puissance, avec des variations notables uniquement pour les puissances élevées (à partir de h).

Ceci suggère que l'effet de la marque pourrait être indirectement capturé par la puissance du véhicule.

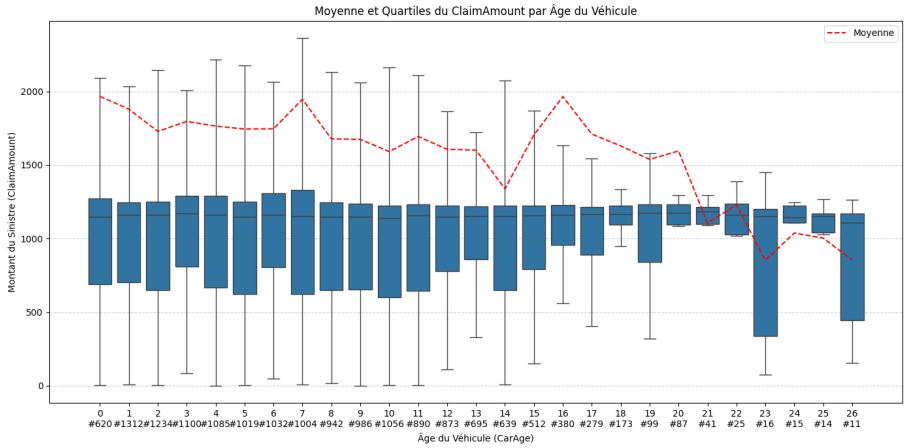

Figure 7: Moyenne et quantile des sinistres par âge de la voiture

L'analyse par box-plot de la sinistralité en fonction de l'âge du véhicule montre une légère décroissance en moyenne de la sévérité avec l'âge du véhicule. Les véhicules plus récents ont tendance à avoir un coût de sinistre légèrement supérieur, bien que les intervalles de confiance se chevauchent largement, signalant une variation limitée.

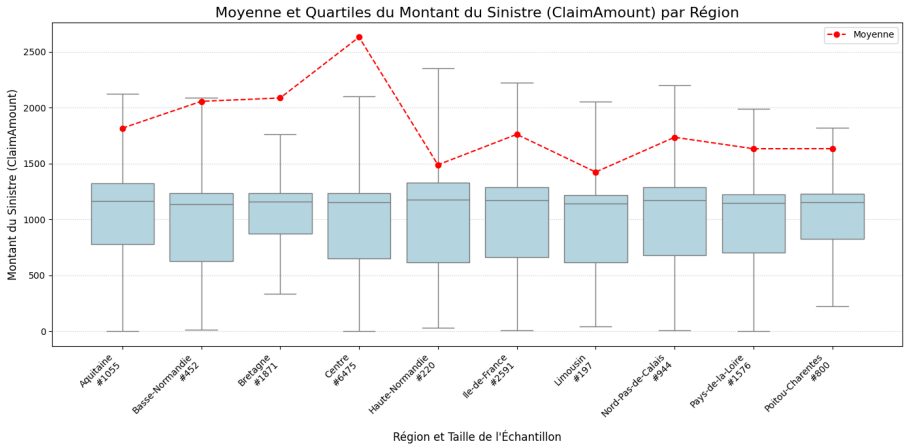

Figure 8: Moyenne et quantile des sinistres par région

L'analyse de la sinistralité en fonction de la région montre une distribution similaire des montants des sinistres entre les regions. Nous constatons cependant en comparant la moyenne avec les écart-type, que les sinistres extrèmes sont porté par les régions Centre et Bretagne.

5.1.3 Analyse de la queue de distribution

L'analyse des montants les plus élevés montre que le sinistre le plus coûteux (2 036 833 €) représente à lui seul 5,9 % du coût total des sinistres du portefeuille. Une analyse de la queue de distribution montre que les sinistres les plus sévères ont essentiellement lieu dans la région Centre et Bretagne.

5.2 Séléction du modèle GLM

Compte tenu de la distribution asymétrique des coûts, ainsi que de la forte autour de 1 200 €, le choix d'un modèle linéaire généralisé (GLM) avec une loi Gamma, Gaussienne ou Inverse Gaussienne est envisagé pour modéliser la sévérité.

Nous privilégions par ailleurs l'utilisation d'une fonction de lien logarithmique, notamment au regard de la distribution. Toutefois, nous testons également les fonctions de lien canoniques, bien que ces dernières ne garantissent pas que les prédictions de sévérité demeurent strictement positives.

Distribution	**Lien**	**AIC**
Gamma	Log	204 331.86
Gamma	Canonical (Inverse Power)	204 349.20
Inverse Gaussian	Log	208 832.29
Gaussian	Log	218 881.89
Gaussian	Identity	218 882.02

Table 1: Comparaison des couples distribution–lien selon l'AIC

Le modèle présentant l'AIC le plus faible que nous retenons est la Loi Gamma avec une fonction de lien Logarithmique.

Le modèle indique que de nombreuses catégories de facteurs (Marque, Région, Carburant, et la variable DriverAge) ne sont pas statistiquement significatives.

Aussi, nous souhaitons adopter les règles de séléction de variables suivantes :

- Nous nous fixons un seuil de la p-valeur de $P > 0,05$.

- Nous imposons que l'ensemble des variables ordinales présente un effet monotone. Cette contrainte s'applique en particulier à la variable «puissance». En effet, nous souhaitons éviter qu'un niveau de puissance y, tel que $x > y > z$ (en rappelant que la puissance est bien une variable catégorielle ordonnée), soit associé à une prime p_y telle que $p_x > p_y < p_z$.

L'ensemble de ces règles a pour objectif de garantir non seulement la robustesse des résultats, mais également la cohérence du modèle. Notons que nous faisons le choix de conserver l'intégralité des coefficients satisfaisant les deux règles précédemment énoncées, quelle que soit leur valeur (y compris lorsqu'elle est infinitésimale).

Nous retenons ainsi le modèle suivant :

$$\log(\mathbb{E}[S_i]) = \beta_0 + \beta_1 \cdot \mathbf{1}_{\{\texttt{Brand\_Agg}=\text{Ref\_Brand}\}} + \beta_2 \cdot \texttt{CarAge} + \beta_3 \cdot \texttt{Density} \tag{3}$$

Où la variable 'Brand_Agg = Ref_Brand' désigne toute les marques de voitures à l'exception des marques Japonaises (sauf Nissan) et Coréennes.

Nous obtenons les coefficients suivants :

	coef	**std err**	**z**	**P> |z|**	**[0.025, 0.975]**
Intercept	7.5290	0.038	197.418	0.000	[7.454, 7.604]
C(Brand_Agg)[T.Ref_Brand]	-0.1538	0.040	-3.821	0.000	[-0.233, -0.075]
CarAge	-0.0079	0.003	-3.144	0.002	[-0.013, -0.003]
Density	-8.911e-06	2.61e-06	-3.414	0.001	[-1.4e-05, -3.79e-06]

Table 2: Résumé des résultats de la régression sévérité séléctionné

6 Construction du tarif

6.1 Principe

Le tarif a priori est construit selon la décomposition classique

$$\text{Prime Pure}_i = \widehat{\text{Fréquence}}_i \times \widehat{\text{Sévérité}}_i.$$

Nous faisons l'hypothèse standard d'indépendance entre le nombre de sinistres N et le coût unitaire S, de sorte que

$$E[NS] = E[N] \times E[S].$$

Les modèles retenus dans les sections précédentes sont :

- un **GLM de Poisson** (avec offset log(Exposure)) pour la fréquence ;
- un **GLM Gamma à lien logarithmique** pour la sévérité.

Après validation hors-échantillon, ces deux modèles sont réajustés sur l'ensemble des données d'apprentissage pour produire la tarification finale.

La figure ci-dessous illustre la distribution des primes pures obtenues à partir du modèle ajusté en fonction de la durée d'exposition au risque. Nous faisons l'hypothèse d'acquisition progressive c'est à dire que la prime est acquise au prorata temporis de l'exposition. Ainsi, pour les contrats dont la durée d'exposition est particulièrement faible l'acquisition d'une prime également très faible (une fraction de la prime annuelle complète). Elle présente une forte asymétrie, avec une concentration importante des valeurs entre 0 et 150 €, tandis que la partie

droite décroît rapidement. Cette forme est cohérente avec une tarification issue d'un modèle multiplicatif : la majorité des assurés présentent une exposition et un risque modérés, tandis que les valeurs plus élevées correspondent à des profils combinant des facteurs défavorables (âge conducteur, type de véhicule, densité, etc.).

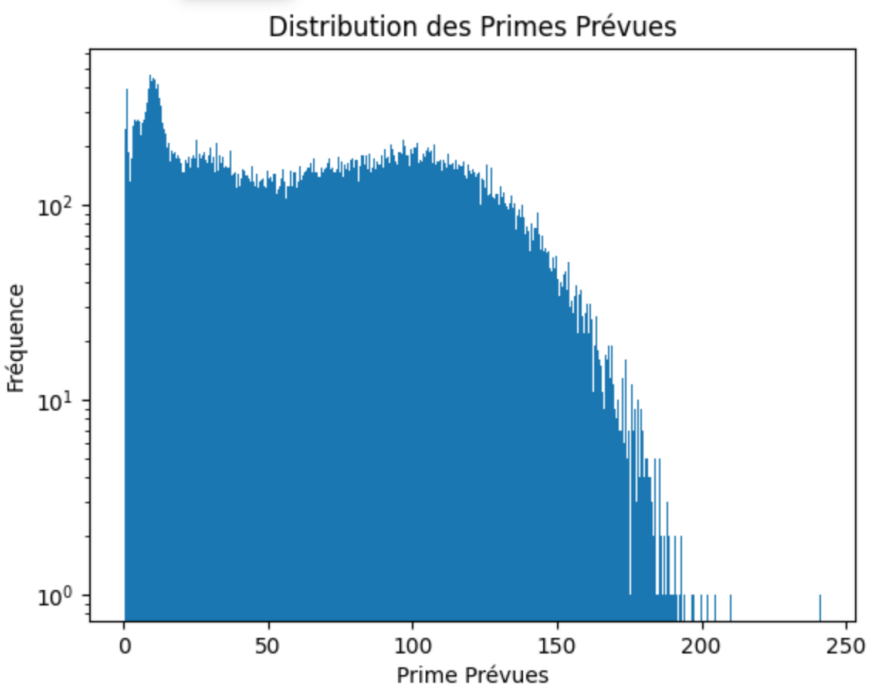

Figure 9: Distribution empirique des primes pures prévues par le modèle fréquence × sévérité.

6.2 Analyse du tarif obtenu et diagnostic

À partir des prédictions issues des deux modèles, nous calculons la prime pure moyenne et la comparons au coût moyen observé du portefeuille.

Le *Loss Ratio* (ou ratio de sinistralité) correspond au rapport entre le coût total des sinistres observés et le montant total des primes pures prédites. Un Loss Ratio égal à 100 % traduit un équilibre technique, tandis qu'un ratio supérieur à 100 % indique un sous-tarif du portefeuille.

L'analyse fait apparaître :

- un **loss ratio global égale à 128%** , indiquant un **sous-tarif**.
- une forte influence des sinistres extrêmes : les **0,1% des sinistres les plus coûteux** représentent une part disproportionnée du coût total, alors qu'ils ne génèrent qu'une infime part des primes.

Lorsque l'on retire le top 0,1% des sinistres les plus élevés, le loss ratio redevient proche de l'équilibre, montrant que :

- le modèle fréquence × sévérité décrit correctement les **petits et moyens sinistres**,
- mais une **modélisation spécifique de la queue lourde** serait pertinente (Pareto, modèles mixtes, troncature, etc.).

7 Modèle pénalisé Lasso

7.1 Validation croisée

Nous estimons des modèles pénalisés Lasso à l'aide de `glmnet`, séparément pour :

- la fréquence (famille Poisson),
- la sévérité (famille Gamma).

La validation croisée (4-fold) permet d'obtenir deux valeurs importantes pour le paramètre de régularisation λ :

- $\lambda_{\min}$: minimise l'erreur de validation croisée ;
- λ_{1se} : modèle plus parcimonieux avec performance quasi-équivalente.

7.2 Variables retenues

Les résultats obtenus sont cohérents avec ceux des GLM :

- pour la **fréquence**, le Lasso conserve principalement `DriverAge`, `Region`, `Gas`, `Brand` et `Density`;
- pour la **sévérité**, seules `CarAge` et `Density` sont sélectionnées.

7.3 Comparaison avec les GLM

D'après les résultats hors-échantillon du notebook :

- la performance prédictive du Lasso est similaire aux GLM non pénalisés, avec un Loss ratio de 121% ;
- aucun gain clair n'est observé sur le MAE ;
- le Lasso nous semble surtout utile comme outil de validation et de robustesse des choix de variables.

Ainsi, même si le Lasso confirme la stabilité des effets mis en évidence précédemment, les modèles GLM standard restent préférés pour la tarification finale.

8 Conclusion

Ce projet avait pour objectif de construire un tarif a priori pour un portefeuille d'assurance automobile en modélisant séparément la fréquence et la sévérité des sinistres. Nous avons utilisé des modèles linéaires généralisés (GLM), complétés par une approche pénalisée Lasso afin d'évaluer la robustesse et la parcimonie des modèles obtenus.

Pour la fréquence, plusieurs modèles ont été comparés, notamment le Poisson, le quasi-Poisson et la Binomiale Négative. Bien que la sur-dispersion ait été confirmée, le modèle de Poisson s'est révélé être le meilleur compromis entre performance, simplicité et interprétabilité, en particulier sur le jeu de test. Le quasi-Poisson a été utilisé pour la sélection de variables, mais le modèle final retenu est un GLM Poisson.

Pour la sévérité, l'analyse descriptive a mis en évidence une distribution très asymétrique, fortement influencée par un petit nombre de sinistres extrêmes. La comparaison par AIC a clairement montré que le modèle Gamma avec lien logarithmique était le plus adapté. Les résultats ont confirmé que seules les variables `CarAge` et `Density` avaient un impact significatif sur le coût conditionnel des sinistres.

La construction du tarif, obtenue par la combinaison fréquence $\times$ sévérité, a révélé un loss ratio global supérieur à 100 %, indiquant un sous-tarif structurel. Toutefois, l'analyse de la queue lourde a montré que ce déséquilibre provenait principalement d'un très petit nombre de sinistres extrêmes. Une fois ces sinistres extrêmes exclus, les modèles se révèlent cohérents et stables, confirmant la pertinence des GLM pour la majorité du portefeuille. Une modélisation dédiée des sinistres majeurs (Pareto, troncature, modèles mixtes) constituerait une amélioration logique.

Le modèle Lasso a été utilisé comme outil de validation, permettant de confirmer la pertinence des variables sélectionnées dans les GLM. Bien qu'il offre une meilleure parcimonie, aucune amélioration significative n'a été constatée en termes de performance prédictive. Les GLM restent donc mieux adaptés à la tarification, notamment en raison de leur interprétabilité.

En conclusion, ce projet a permis de construire un cadre complet et réplicable de tarification a priori pour un portefeuille automobile. Les modèles obtenus sont cohérents, robustes et conformes aux pratiques actuarielles. Parmi les pistes d'amélioration, on retiendra principalement :

- la modélisation avancée de la queue des coûts (Pareto, EVT, modèles mixtes) ;
- l'exploration de modèles non linéaires pour la fréquence (GBM, Random Forest) ;
- l'intégration d'interactions pertinentes entre variables (puissance $\times$ marque, âge conducteur $\times$ densité) ;
- l'étude du comportement du portefeuille sur plusieurs années pour évaluer la stabilité temporelle du tarif.

Ces extensions permettraient d'obtenir un tarif encore plus précis et mieux calibré, notamment pour les profils à risque extrême.