Tarification en assurance non vie (IARD)

Maximilien Dialufuma V.

Objectifs

Cette présentation a pour objectifs de :

- introduire les notions basiques de la tarification en assurance non vie;
- analyser la prime pure;
- apprendre les méthodes statistiques de la tarification;
- discuter de la tarification à la frontière de la recherche.

C'est quoi la tarification?

- Tout part du concept technique de l'assurance qui est une opération synallagmatique: une partie - l'**assureur** - s'engage..
- ..à payer une somme (ou indemnité, ou prestation) à une autre partie - l'**assuré** pendant toute la période de couverture..
- .. (généralement un an) - en cas de survenance d'un sinistre prédéfini à l'avance.
- En échange, l'**assureur** reçoit une prime (ou *cotisation*) payée par l'assuré à la date définie par le contrat.

C'est quoi la tarification?

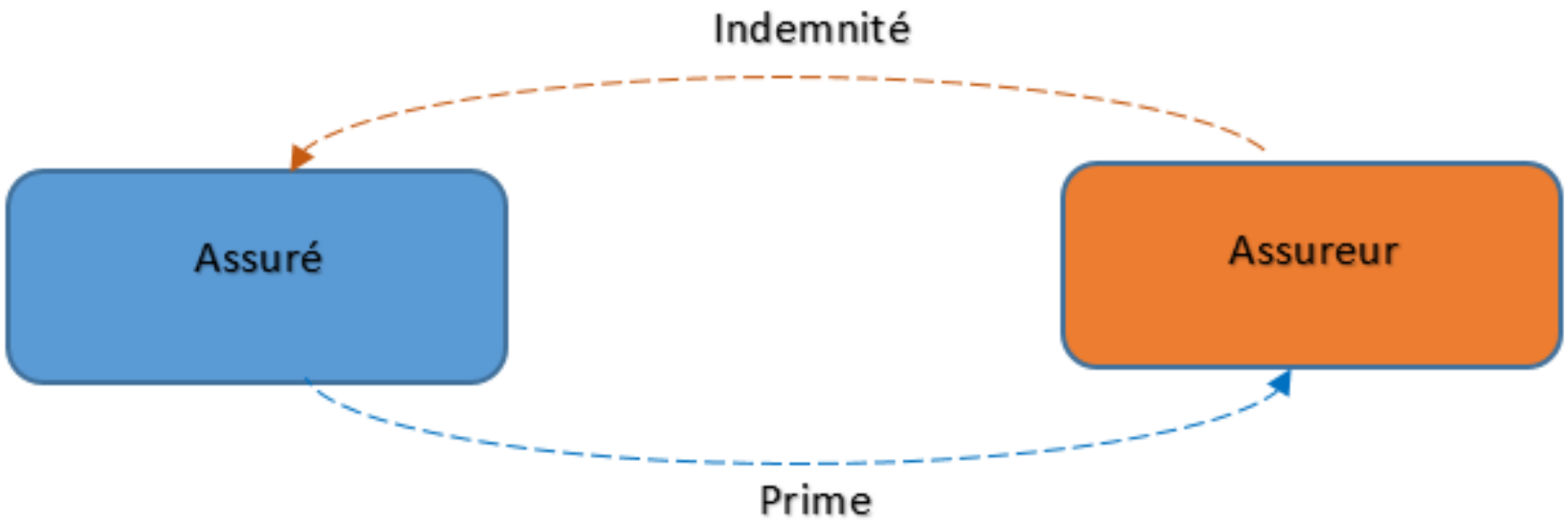


Figure: Schéma d'un contrat d'assurance

C'est quoi la tarification?

- La **tarification**, designe le processus qui consiste à évaluer le niveau de risque d'une personne (physique ou morale) demandant une assurance.
- Le risque constitue la matière première en assurance.
- Le risque doit être assurable i.e. un événement futur, aléatoire, légalement assurable et mesurable.
- Sa réalisation ne dépend pas de la volonté exclusive d'une partie.

Pourquoi la tarification?

- Parce que la compagnie doit déterminer le prix du produit qu'elle souhaite vendre.

$$Prix = Couts + Marge\ de\ profit$$

- Mais le cout réel du produit est donc inconnu au moment où le prix doit être établi, et que..
- .. ce coût réel sera connu après la vente du produit i.e. lorsqu'un sinistre survient.
- C'est ce qu'on appelle l'inversion du cycle de production dans le marché des assurances..

Qui peut faire la tarification et comment ?

- Un actuaire ou un statisticien spécialisé en assurance;
- Estimer ce coût à l'aide de modèles (ou méthodes) statistiques rigoureuses ..
- ..et plus ou moins complexes.

Type de tarification

- On distingue deux types de tarifications.
  - **Tarification à priori** : Elle consiste à évaluer le niveau de risque d'une personne en utilisant les variables dites variables tarifaires (ou facteurs de risque).
  - **Tarification à postériori**: Elle consiste à évaluer le niveau de risque d'une personne selon son historique (ou expérience) de sinistralité pendant une certaine période. C'est qu'on appelle en science actuarielle la théorie de crédibilité.[1]

[1] Bühlmann(1969), Goulet(2008)

Tarification à priori

La tarification à priori utilise certaines informations de l'assuré pour évaluer la prime. On peut citer:

- Les informations sur l'âge, sexe, profession de l'assuré.
- Les informations sur les biens de l'assuré (Ex: l'âge du véhicule, la puissance du véhicule, surface du logement en assurance multirisque).
- Les informations géographiques de l'assuré (Ex: zone d'habitation, densité de la population).

Pourquoi la segmentation?

- En assurance, tous les individus ne sont pas égaux devant le risque : certains sont plus dangereux que d'autres. On parle du portefeuille hétérogène.
- Il faudrait adapter la prime en fonction du niveau de risque représenté par chacun des assurés.

Pourquoi la segmentation?

- La segmentation est une technique que l'assureur utilise pour différencier la prime (ou la couverture), en fonction d'un certain nombre de caractéristiques spécifiques du risque à assurer,..
- .. et ce afin de parvenir à une meilleure concordance entre les coûts qu'une personne déterminée met à charge de la collectivité des preneurs d'assurance et la prime que cette personne doit payer pour la couverture offerte.
- Pour approfondir la lecture sur la techinque de segmentation, on peut consulter Denuit(2005).

Définitions

Pour parler de la prime pure, nous définissons

Exposition au risque

C'est la période de temps (jours), sur une durée donnée, pour lequel l'assuré était couvert par l'assurance.

Fréquence

C'est nombre de sinistres observés par unité d'exposition.

Sévérité moyenne

C'est le coût moyen d'un sinistre.

Notations

- N: variable aléatoire associé à la fréquence;
- Y_i : v.a associé aux coûts des sinistres i.e. les prestations(ou indemnités) versées par l'assureur à l'assuré avec $i = 1, \cdots, N$;
- S : La charge totale par police

$$S = Y_1 + Y_2 + \cdots + Y_N = \sum_{i=1}^{N} Y_i, \;\; Y_i > 0 \qquad (1)$$

Prime pure

Prime pure

C'est le montant nécessaire pour couvrir les sinistres, sans perte ou gain.

- Si nous sommes dans un portefeuille de risques homogènes, alors $\mathbb{E}[S]$ est la candidate par excellence pour la prime pure.
- Si les coûts de sinistres Y_i sont i.i.d et indépendants du nombre de sinistres N, alors

$$\mathbb{E}[S] = \mathbb{E}[N].\mathbb{E}[Y_i] \tag{2}$$

Prime pure

- Si nous sommes dans un portefeuille de risques hétérogènes i.e la fréquence et les couts sont hétérogènes, alors la prime pure devrait être

$$\mathbb{E}[S \mid \Omega] = \mathbb{E}[N \mid \Omega].\mathbb{E}[Y_i \mid \Omega] \tag{3}$$

où Ω représente une information inconnue de l'assuré.

- Comme Ω est inconnue, on va utiliser les facteurs de risque (ou variables tarifaires) qui nous donnent certaines informations de l'assuré.

Prime pure

- Soit $\mathbf{X} = (X_1, \cdots, X_n) \subset \Omega$, l'ensembre des informations de l'assuré que possède l'assureur.
- La prime pure est donnée par :

$$\mathbb{E}[S \mid \mathbf{X}] = \mathbb{E}[N \mid \mathbf{X}] . \mathbb{E}[Y_i \mid \mathbf{X}] \tag{4}$$

- Cette prime permet à l'assureur d'être en moyenne à l'équilibre, car

$$\mathbb{E}[\mathbb{E}[S \mid \mathbf{X}]] = \mathbb{E}[S] \tag{5}$$

- On regarde comment est ce que le risque est partagé entre l'assureur et les assurés par la décompositiion de la variance.

Décomposition de la variance du risque

Table: **Décomposition de la variance entre assureur et assurés**

	Assurés	Assureur
Dépense	$\mathbb{E}[S \mid \mathbf{X}]$	$S - \mathbb{E}[S \mid \mathbf{X}]$
Dépense moyenne	$\mathbb{E}[S]$	0
Variance (risque)	$\text{Var}[\mathbb{E}[S \mid \mathbf{X}]]$	$\mathbb{E}[Var[S \mid \mathbf{X}]]$

La variance de l'assureur est:

$$\begin{aligned}\mathbb{E}[Var[S \mid \mathbf{X}]] &= \mathbb{E}[\mathbb{E}[Var[S \mid \Omega] \mid \mathbf{X}]] + \mathbb{E}[Var[\mathbb{E}[S \mid \Omega] \mid \mathbf{X}]] \\ &= \mathbb{E}[Var[S \mid \Omega]] + \mathbb{E}[Var[\mathbb{E}[S \mid \Omega] \mid \mathbf{X}]]\end{aligned}$$

Décomposition de la variance du risque

- Si tous les facteurs de risque ne sont pas pris en compte, l'assureur intervient pour réparer les conséquences...
- ...fâcheuses du hasard (mutualisation du risque), mais prend aussi en charge les variations de la prime pure exacte...
- ...qui ne sont pas expliquées par les facteurs de risques intégrés au tarif (solidarité).
- Ainsi la variance totale du risque est :

$$\begin{aligned} Var[S] &= \mathbb{E}[Var[S \mid \Omega]] + \mathbb{E}[Var[\mathbb{E}[S \mid \Omega] \mid \mathbf{X}]] + Var[\mathbb{E}[S \mid \mathbf{X}]] \\ &= \textbf{mutualisation} + \textbf{solidarité} + \textbf{assurés} \end{aligned}$$

Prime majorée

- Il est très risqué de faire payer la prime pure, car le montant des sinistres n'est jamais égal à la moyenne et est aléatoire.
- On va ajuster la prime pure en ajoutant un coefficient η qui correspond à une marge de securité,..
- ..une marge pour les profits, une marge pour les dépenses fixes,etc.
- La prime majorée (ou commerciale, ou avec chargement) est donnée par

$$\pi(S) = (1+\eta).\mathbb{E}[S], \ 0 < \eta < 1 \tag{6}$$

Prime majorée

- Considérons une fonction ϕ_n, appelée probabilité de ruine définie par :

$$\phi_n = \mathbb{P}\left(S_1 + S_2 + \cdots + S_N > \pi_1 + \pi_2 + \cdots + \pi_N\right) \qquad (7)$$

où ϕ_n représente probabilité pour l'assureur de ne pas être en mesure d'indemniser ses assurés et
$S_1 + S_2 + \cdots + S_N$ la somme des sinistres des assurrés du portefeuille.

- Si les S_i sont i.i.d, alors par la loi des grands nombres on a :

Prime majorée

- Si $\eta > 0$, alors

$$\pi(S_i) > \mathbb{E}[S_i] \ \ et \ \phi_n \underset{n\to\infty}{\longrightarrow} 0. \tag{8}$$

- Si $\eta < 0$, alors

$$\pi(S_i) < \mathbb{E}[S_i] \ \ et \ \phi_n \underset{n\to\infty}{\longrightarrow} 1. \tag{9}$$

- On voit d'après l'équation (7) que la prime majorée permet de tendre la probabilité de ruine vers 0 losque n tend vers l'infini.

Modèles statistiques pour la tarification

- Il exites plusieurs modèles statistiques pour la tarification à priori.
- On peut citer entre autres:
  - Le modèle linéaire généralisé (GLM), voir Denuit and Charpentier(2005), Nelder and Wedderburn(1972),McCullagh and Nelder(1991).
  - Le modèle additif généralisé (GAM), voir Hastie and Tibshirani(1990).
  - Modèle d'apprentissage statistique supervisé (Arbre de décision, Forêt aléatoires, Boosting, Gradien boosting, Baggin,etc..), voir James et al.(2013).
- Le plus utilisé est le modèle linéaire généralisé. Sa mise en oeuvre en assurance était introduite vers les années 90, voir Kaas et al.(2013), Frees(2009), Denuit and Charpentier(2005).

C'est quoi les modèles GLM

- Les modèles linéaires généralisés (GLM) sont une généralisation du modèle linéaire Gaussien et admettant d'autres lois pour la variable endogène que la loi normale.
- Ces lois appartiennent à la famille dite exponentielle, dont la densité s'écrit:

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \qquad (10)$$

où a, b, c sont des fonctions et θ, ϕ les paramètres canonique et de dispersion respectivement.

Famille exponentielle

Proposition

Pour toute distribution de la famille exponentielle, on a :

(i) $\mu = \mathbb{E}[Y] = b'(\theta)$

(ii) $Var[Y] = b''(\theta)a(\phi)$ et $V(\mu) = b''(\theta)$ (fonction de variance)

(iii) Si $Y_1, \cdots, Y_n$ sont des v.a i.i.d dont la distribution est donnée par l'équation (9), alors

$$f(y_1, \cdots, y_n) = \exp\left(\sum_{i=1}^{n} \frac{(y_i\theta_i - b(\theta_i))}{a_i(\phi)} + \sum_{i-1}^{n} c(y_i, \phi)\right)$$

Les lois membres de la famille exponentielle

Il existe plusieurs lois de probabilites qui y sont membres, nous citons entre autres:

- Pour les lois discrètes
  - La loi de Bernoulli $\mathcal{B}(p)$
  - La loi binomiale $\mathcal{B}(n,p)$
  - La loi binomiale négative $\mathcal{B}(r,p)$
  - La loi de Poisson $\mathcal{P}(\lambda)$
- Pour les lois continues
  - La loi normale $\mathcal{N}(\mu,\sigma^2)$
  - La loi Gamma $\mathcal{G}(\alpha,\lambda)$
  - La loi normale inverse $\mathcal{IN}(\mu,\lambda)$
  - La loi de Pareto $\mathcal{P}(\alpha,\theta)$
  - La loi Tweedie $\mathcal{T}_w(\mu,\sigma^2)$

Éléments du GLM

Considérons que nous avons un échantillon ayant

- une variable endogène Y
- p' variables exogènes, $x_1, \cdots, x_p$ pour n individus indépendants

Alors le modèle GLM exige les éléments suivants :

- La distribution de la variable endogène (membre de la famille exponentielle)
- Le prédicteur linéaire $\eta_i = \mathbf{x}_i'\beta$ où $\mathbf{x}_i' = (1, x_{i1}, \cdots, x_{ip})$ et $\beta = (\beta_0, \cdots, \beta_p)$
- La fonction de lien $g(\mu_i) = \eta_i = \mathbf{x}_i'\beta$ où $g(\mu_i) = g(\mathbb{E}[Y_i \mid \mathbf{x}_i'])$

Éléments du GLM

- La fonction de lien $g(.)$ permet de relier l'espérance μ au préditeur linéaire.
- Si $g(\mu_i) = \theta_i$, alors on dit que $g(.)$ est un lien canonique.
- Choisir lien canonique comporte plusieurs avantages lors de l'estimation des paramètres.
- Il existe d'autres liens aussi (lien identité, lien inverse, lien logarithmique, ect.)

Procédure de la modélisation GLM

- Estimation des paramètres (Par la méthode de maximum de vraisemblance)
- Test statistiques (Test de nullité des paramètres, Test de Wald)
- Selection des variables par la méthode (forward, backward et stepwise) pour reternir le meilleur modèle
- Deviance (comparaison entre la vraisemblance du modèle saturé et la vraisemblance du modèle ajusté par un test de khi-deux)
- Choix du modèle par les critères AIC et BIC
- Pour une lecture approfondie, on peut consulter McCullagh and Nelder(1991), Frees(2009), Ohlsson and Johansson(2010).

Présentation de la base de données

- Il s'agit d'une base de données réelle d'une compagnie d'assurance automobile.
- La base est tirée du packages *CASdatasets*("freMPL2", "freMPL4") dans R
- Nous avons joint les deux bases de données pour augmenter le nombre d'observations.
- La base de données comprends 84590 observations et 22 variables.

Visualisation de la base d'assurance automobile

```
'data.frame':   84590 obs. of  22 variables:
 $ Exposure   : num  0.583 0.416 0.583 0.2 0.083 0.375 0.5 0.499 0.218 0.75 ...
 $ LicAge     : int  579 361 366 187 169 170 224 230 169 232 ...
 $ RecordBeg  : Date, format: "2004-06-01" "2004-01-01" "2004-06-01" ...
 $ RecordEnd  : Date, format: NA "2004-06-01" NA ...
 $ VehAge     : Factor w/ 9 levels "0","1","10+",..: 3 2 4 1 2 2 5 5 8 6 ...
 $ Gender     : Factor w/ 2 levels "Female","Male": 2 1 1 2 1 1 2 2 2 1 ...
 $ MariStat   : Factor w/ 2 levels "Alone","Other": 2 2 2 1 2 2 2 2 2 2 ...
 $ SocioCateg : Factor w/ 52 levels "CSP1","CSP16",..: 28 1 1 24 1 1 18 18 22 24 ...
 $ VehUsage   : Factor w/ 4 levels "Private","Private+trip to office",..: 1 3 3 2 3 3 3 3 2 2 ...
 $ DrivAge    : int  83 55 55 34 33 34 53 53 32 38 ...
 $ HasKmLimit : int  0 0 0 0 0 0 0 0 0 0 ...
 $ BonusMalus : int  50 58 72 80 63 63 72 68 50 57 ...
 $ VehBody    : Factor w/ 9 levels "bus","cabriolet",..: 6 6 6 4 5 5 9 9 6 6 ...
 $ VehPrice   : Factor w/ 26 levels "A","B","C","D",..: 14 4 4 11 12 12 12 12 7 2 ...
 $ VehEngine  : Factor w/ 6 levels "carburation",..: 5 5 5 2 2 2 2 2 5 5 ...
 $ VehEnergy  : Factor w/ 4 levels "diesel","eletric",..: 4 4 4 1 1 1 1 1 1 4 ...
 $ VehMaxSpeed: Factor w/ 10 levels "1-130 km/h","130-140 km/h",..: 8 5 5 6 6 6 3 3 5 4 ...
 $ VehClass   : Factor w/ 6 levels "0","A","B","H",..: 4 3 3 5 5 5 1 1 3 2 ...
 $ RiskVar    : int  14 15 15 20 17 17 19 19 19 19 ...
 $ ClaimAmount: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Garage     : Factor w/ 3 levels "Collective garage",..: 2 2 2 2 2 3 2 2 2 2 ...
 $ ClaimInd   : int  0 0 0 0 0 0 0 0 0 0 ...
```

Figure: Base de données

Types des variables dans la base

La base contient différente type des variables à savoir:

- 6 variables quantitatives (Exposure, LicAge, DriveAge, ClaimAmount,etc.)
- 4 variables qualitatives dichotomiques (Gender, Claimind, Maristat, etc.)
- 10 variables qualitatives polytomiques(VehUsage, VehAge, vehMaxSpee, VehPrice, etc.)
- 2 variables temporelles (RecordBeg, RecordEnd)

Corrélation

- Pour les variables quantitatives, on peut regarder s'il existe une relation linéaire entre les variables en calculant le coefficient de corrélation.

	Exposure	LicAge	DrivAge	BonusMalus	RiskVar	ClaimAmount
Exposure	1.00000000	0.050384804	0.052797823	-0.04868046	-0.011077793	0.020264993
LicAge	0.05038480	1.000000000	0.928254727	-0.51651709	-0.056045707	-0.004186191
DrivAge	0.05279782	0.928254727	1.000000000	-0.46639230	-0.038935074	-0.001229624
BonusMalus	-0.04868046	-0.516517093	-0.466392303	1.00000000	0.044532385	0.015637727
RiskVar	-0.01107779	-0.056045707	-0.038935074	0.04453239	1.000000000	0.008904155
ClaimAmount	0.02026499	-0.004186191	-0.001229624	0.01563773	0.008904155	1.000000000

Figure: Coefficient de corrélation

- On voit que les variables DrivAge et LicAge sont fortement corrélées. Il faudrait écarter une variable lors de la modélisation pour éviter le sur-apprentissage.

Corrélation

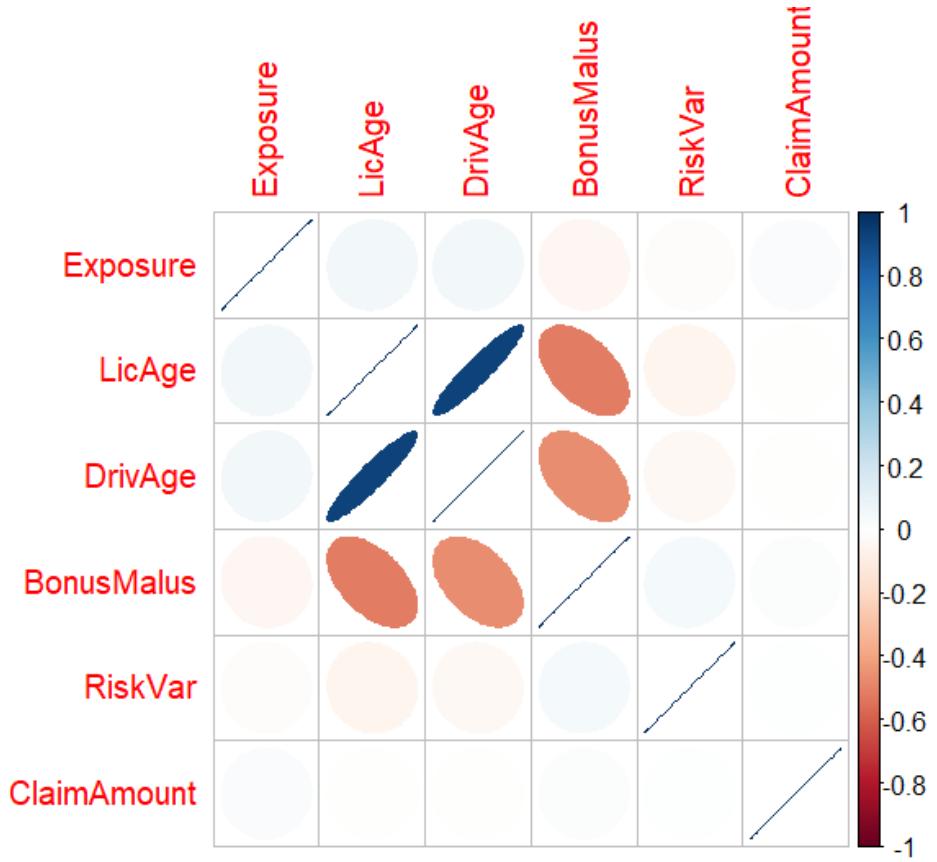


Figure: Corrélogramme

Analyse en composantes principales (ACP)

- ACP permet de réduire la dimension d'un jeu de données tout en conservant le plus d'information possible.
- On va chercher une combinaison linéaire des variables qui maximise la variance, i.e. celle qui retient le maximum de l'information contenue dans le jeu de données.

$$\mathbf{Y} = \mathbf{XA} \tag{11}$$

où $\mathbf{A} = (\alpha_1, \cdots, \alpha_p)$ est un vecteur propre de la matrice variance-covariance Σ de $\mathbf{X}$, $\mathbf{Y}$ est une composante principale.

- Chaque vecteur propre α_k est associé à une valeur propre λ_k, qui constitue la k^{eme} plus grande valeur propre de Σ.

Analyse en composantes principales (ACP)

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Variance	2.311	1.019	0.999	0.972	0.629	0.070
% of var.	38.515	16.976	16.656	16.202	10.489	1.162
Cumulative % of var.	38.515	55.491	72.147	88.349	98.838	100.000

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2
1	3.304	3.156	0.005	0.912	0.213	0.000	0.004
2	1.020	0.917	0.000	0.808	-0.076	0.000	0.006
3	1.069	0.645	0.000	0.365	0.307	0.000	0.083
4	1.987	-1.256	0.001	0.400	-0.358	0.000	0.032
5	1.825	-0.954	0.000	0.273	-0.714	0.001	0.153
6	1.307	-0.840	0.000	0.413	-0.073	0.000	0.003
7	1.389	-0.055	0.000	0.002	0.249	0.000	0.032
8	1.362	0.062	0.000	0.002	0.241	0.000	0.031
9	2.010	-0.683	0.000	0.115	-0.379	0.000	0.036
10	1.791	-0.233	0.000	0.017	0.788	0.001	0.194

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2
Exposure	0.101	0.445	0.010	0.632	39.184	0.399
LicAge	0.952	39.206	0.906	-0.017	0.029	0.000
DrivAge	0.935	37.849	0.875	-0.010	0.009	0.000
BonusMalus	-0.715	22.114	0.511	0.025	0.061	0.001
RiskVar	-0.094	0.380	0.009	0.127	1.588	0.016
ClaimAmount	-0.012	0.006	0.000	0.776	59.128	0.602

Figure: Resumé de l'ACP

Analyse en composantes principales (ACP)

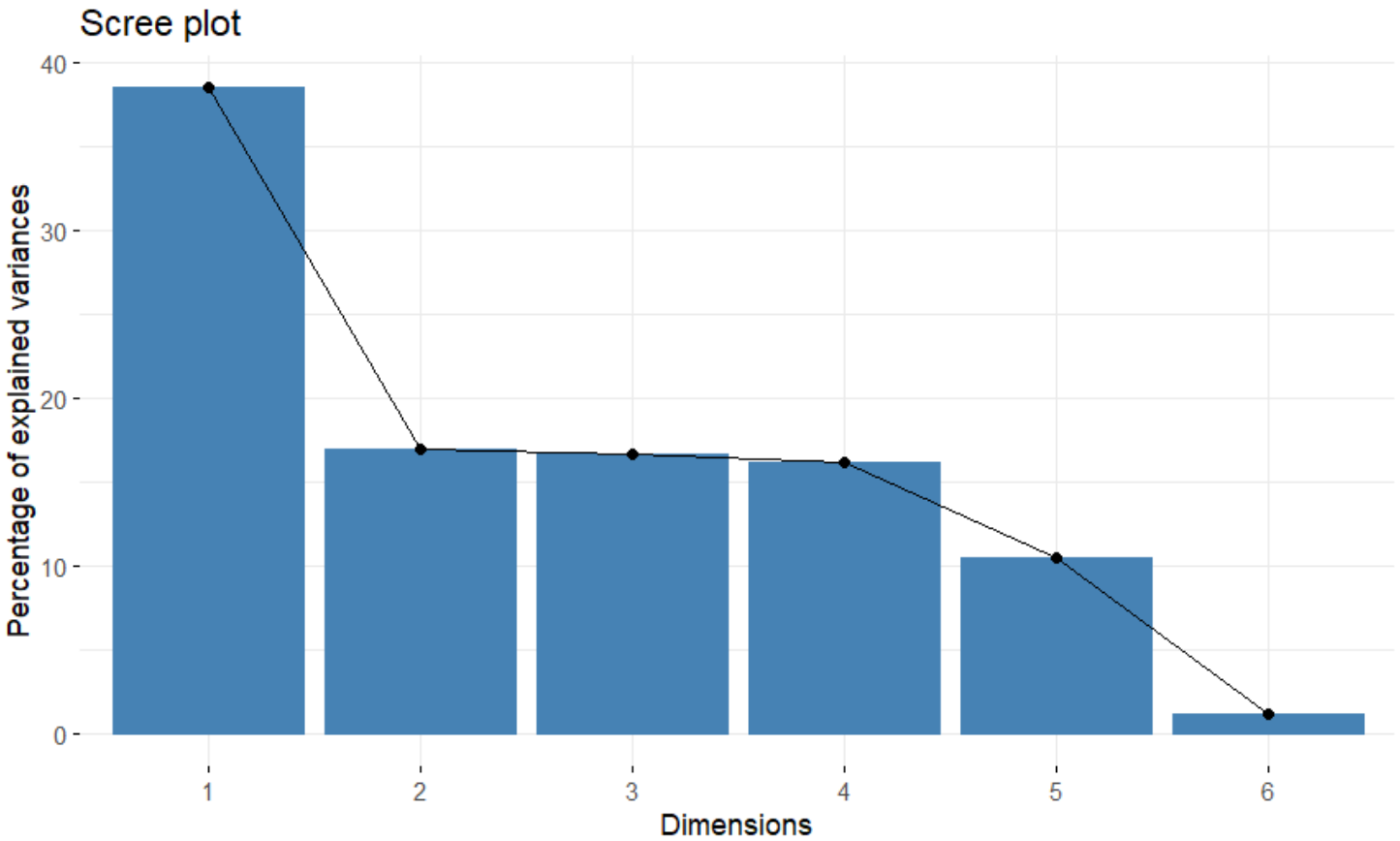


Figure: Proportion de variance expliquée par composantes

Analyse en composantes principales (ACP)

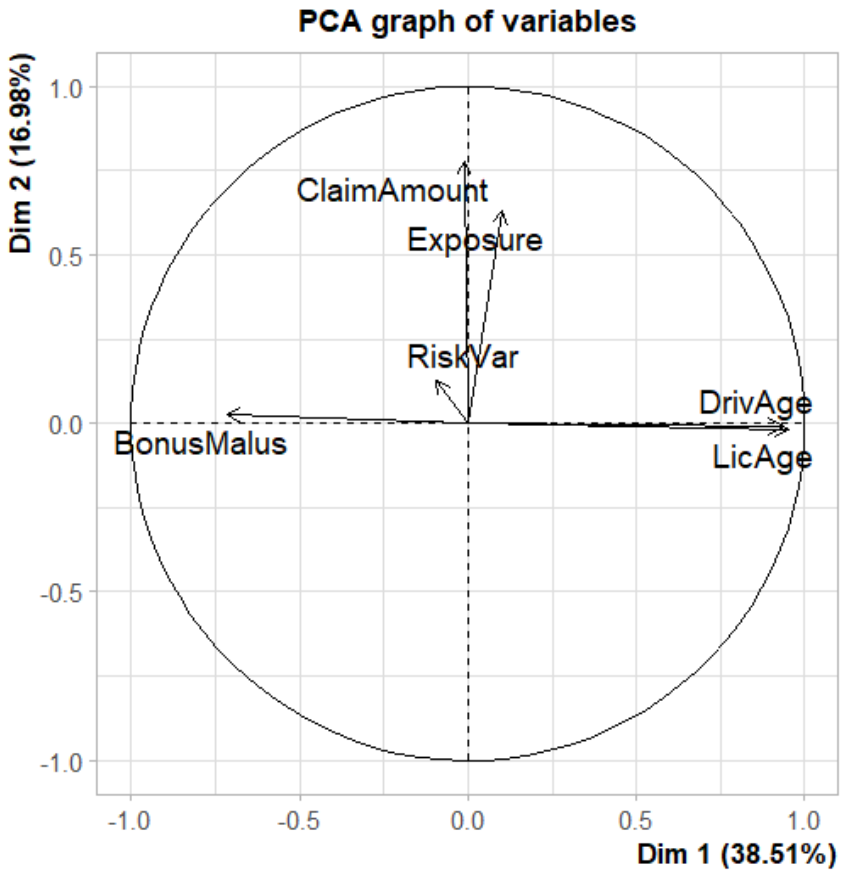


Figure: Représentation des variables sur les premiers axes de l'ACP

Analyse de correspondance(AC ou AFC)

- L'analyse de correspondance est la version ACP pour les variables catégorielles.
- Elle permet de représenter graphiquement le tableau de contingence dont les coordonnées sont les éléments des profils lignes et colonnes du tableau.
- À partir du tableau de contingence, on évalue s'il existe une dépendance significative entre les catégories par le test du χ^2.
- L'analyse de correspondance peut être binaire(resp. multiples) lorsqu'il s'agit de deux variables catégorielles (resp. trois ou plusieurs).

Analyse de correspondance binaire

	ClaimInd=0	ClaimInd=1
Male	52059	1584
Female	30035	912
Alone	22693	726
Other	59401	1770
HasKmLimit=0	71525	2231
HasKmLimit=1	10569	265

Figure: Tableau de contingence

```
        Pearson's Chi-squared test

data:  tableau
X-squared = 13.577, df = 5, p-value = 0.01853
```

Figure: Test d'indépendance de Khi-deux

Analyse de correspondance binaire

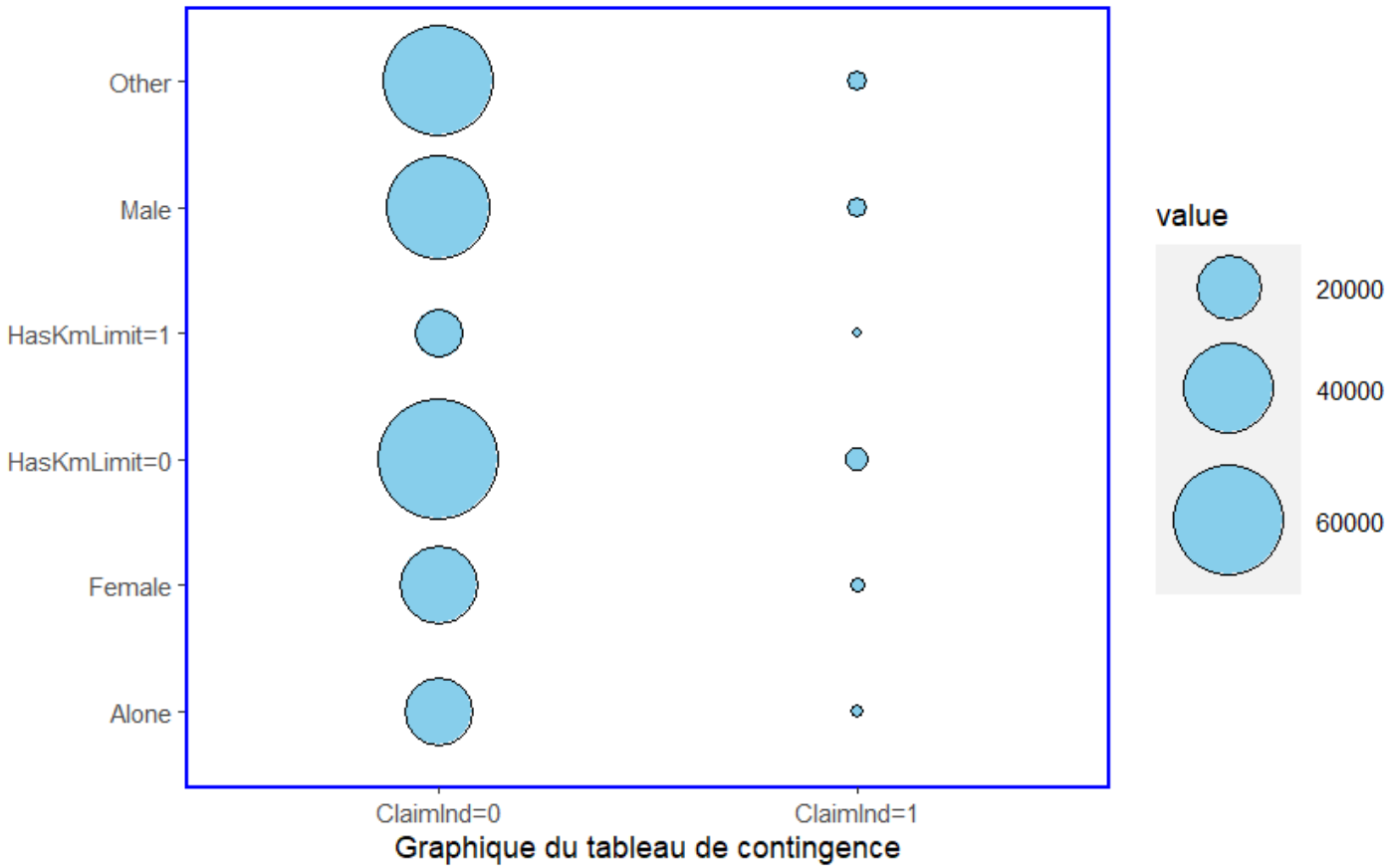


Figure: Représentation graphique du tabeau de contingence

Analyse de correspondance binaire

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	5.350175e-05	100	100

Figure: Valeurs propres

- Les valeurs propres permettent de déterminer le nombre d'axes principaux à retenir.
- Il y a qu'un seul axe à considérer dans notre cas, donc on ne peut extraire le graphe de l'AC pour nos variables catégorielles.

Modélisation de la frequence des sinistres

- Nous allons modéliser la probabilité d'observer (au moins) un sinistre (ClaimInd).
- Cela revient à dire que la réalisation d'un sinistre s'observe comme la réalisation d'une v.a discrète qui suit une loi de Bernoulli $\mathcal{B}(p)$.
- $y_i =$ "*ClaimInd*" , avec $y_i \in [0, 1]$
- Le modèle GLM approprié pour ce cas est le modèle communement appelée régression logistique.

Modélisation avec la régression logistique

On constitue les éléments suivants pour le modèle de régression:

- La variable endogène $y_i \mid x_i \sim \mathcal{B}(p_i)$ avec $p_i = \mathbb{P}[y_i = 1 \mid x_i] = \mathbb{E}[y_i \mid x_i]$
- Le prédicteur linéaire $\eta_i = \mathbf{x}_i'\beta$
- La fonction de lien canonique $g(\mathbb{E}[y_i \mid x_i]) = (p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\beta$
- L'estimateur de β est

$$\hat{p}_i = g^{-1}(\mathbf{x}_i'\hat{\beta}) = \frac{\exp(\mathbf{x}_i'\hat{\beta})}{1+\exp(\mathbf{x}_i'\hat{\beta})} \tag{12}$$

Modélisation avec la régression logistique

Nous utilisons la technique de modélisation par validation croisée.

- Elle consiste à séparer les données de la base en échantillons d'entrainement pour les estimations et test pour les prédictions du modèle.
- On choisit d'utiliser 80% des données pour l'échantillon d'entrainement..
- .. et 20% des données pour l'échantillon test.
- Cette technique est efficace parce qu'elle permet d'éviter le sur-apprentissage du modèle, voir James et al.,(2013).

Modélisation avec la régression logistique

- Nous avons retenu le modèle optimal qui contient le plus petit AIC parmis plusieurs modèles ajustés..
- .. en utilisant la technique de sélection des variables (forward, backward et stepwise)

```
Call:
glm(formula = ClaimInd ~ VehAge10 + VehAge6 + VehAge8 + SocioCategCSP2 +
    SocioCategCSP7 + VehUsagePrivate_trip_to_office + VehUsageProfessional +
    VehUsageProfessional_run + DrivAge + HasKmLimit + BonusMalus +
    VehBodystation_wagon + VehPriceL + VehEngineinjection + VehMaxSpeed140_150_km_h +
    VehMaxSpeed150_160_km_h + VehMaxSpeed160_170_km_h + VehMaxSpeed170_180_km_h +
    VehMaxSpeed220_km_h + VehClassB + VehClassM1 + VehClassM2 +
    RiskVar + GarageNone, family = binomial(link = "logit"),
    data = cbind.data.frame(ClaimInd = data1.app$ClaimInd, data1.app.reg))
```

Figure: Meilleur modèle retenu

Modélisation avec la régression logistique

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -4.983657   0.179367 -27.785  < 2e-16 ***
VehAge10                      -0.120059   0.060605  -1.981 0.047591 *
VehAge6                       -0.218813   0.080552  -2.716 0.006599 **
VehAge8                       -0.127082   0.080832  -1.572 0.115913
SocioCategCSP2                 0.200035   0.094399   2.119 0.034088 *
SocioCategCSP7                 1.632878   0.790042   2.067 0.038751 *
VehUsagePrivate_trip_to_office 0.140855   0.057938   2.431 0.015051 *
VehUsageProfessional           0.219294   0.070352   3.117 0.001826 **
VehUsageProfessional_run       0.323093   0.172570   1.872 0.061173 .
DrivAge                        0.005558   0.001858   2.992 0.002770 **
HasKmLimit                    -0.110027   0.077911  -1.412 0.157889
BonusMalus                     0.013026   0.001116  11.672  < 2e-16 ***
VehBodystation_wagon          -0.403913   0.129406  -3.121 0.001801 **
VehPriceL                      0.186987   0.088157   2.121 0.033916 *
VehEngineinjection             0.077942   0.051628   1.510 0.131122
VehMaxSpeed140_150_km_h        0.159669   0.095361   1.674 0.094059 .
VehMaxSpeed150_160_km_h        0.181027   0.074470   2.431 0.015062 *
VehMaxSpeed160_170_km_h        0.141407   0.070618   2.002 0.045239 *
VehMaxSpeed170_180_km_h        0.201085   0.068582   2.932 0.003368 **
VehMaxSpeed220_km_h            0.273272   0.103955   2.629 0.008570 **
VehClassB                     -0.337634   0.062287  -5.421 5.94e-08 ***
VehClassM1                    -0.275187   0.065493  -4.202 2.65e-05 ***
VehClassM2                    -0.116835   0.074497  -1.568 0.116807
RiskVar                        0.011979   0.004960   2.415 0.015731 *
GarageNone                     0.195851   0.058099   3.371 0.000749 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17998  on 67658  degrees of freedom
Residual deviance: 17759  on 67634  degrees of freedom
AIC: 17809
```

Modélisation avec la régression logistique

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 9803  236
         1 6613  263

               Accuracy : 0.5951
                 95% CI : (0.5877, 0.6025)
    No Information Rate : 0.9705
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0173

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.59716
            Specificity : 0.52705
         Pos Pred Value : 0.97649
         Neg Pred Value : 0.03825
             Prevalence : 0.97050
         Detection Rate : 0.57954
   Detection Prevalence : 0.59350
      Balanced Accuracy : 0.56211
```

Figure: Performance du modèle

Modélisation avec la régression logistique

- La qualité du modèle est mesurée par l'erreur quadratique moyenne

$$\mathbb{E}\left[(Y-\hat{Y})^2\right] = 0.02857095 \simeq 2.8\%$$

- On voit que l'erreur de prédiction sur l'échantillon test est très petit et que la précision (accuracy) du modèle est environ 60%.
- La probalibité d'observer (au moins) un sinistre vaut 0.02950079.
- L'air sous la courbe de ROC est environ 60%.
- On conclut que le modèle est acceptable.

Modélisation du coût des sinistres

- Nous allons modéliser la le coût des sinistres (ClaimAmount).
- Cela revient à dire que la réalisation du coût d'un sinistre s'observe comme la réalisation d'une v.a continue.
- La lois Gamma $\mathcal{G}(\alpha, \lambda)$ est une lois classique pour modéliser les données continues positives à queue épaisse.
- $y_i =$ "*ClaimAmount*", avec $y_i \in \mathbb{R}^+$
- Le modèle GLM approprié pour ce cas est le modèle communement appelée régression gamma.

Modélisation avec la régression gamma

On constitue les éléments suivants pour le modèle de régression:

- La variable endogène $y_i \mid x_i \sim \mathcal{G}(\alpha, \lambda)$ avec $\alpha,\ \lambda > 0$
- Le prédicteur linéaire $\eta_i = \mathbf{x}_i'\beta$
- La fonction de lien logarithmique
  $g(\mathbb{E}[y_i \mid x_i]) = \ln(\mathbb{E}[y_i \mid x_i]) = \mathbf{x}_i'\beta$
- L'estimateur de β est

$$\mathbb{E}[y_i \mid x_i] = g^{-1}(\mathbf{x}_i'\hat{\beta}) = \exp(\mathbf{x}_i'\hat{\beta}) \tag{13}$$

Modélisation avec la régression gamma

- Nous utilisons la technique de modélisation par validation croisée.
- Nous avons retenu le modèle optimal qui contient le plus petit AIC parmis plusieurs modèles ajustés..
- .. en utilisant la technique de sélection des variables (forward, backward et stepwise)

```
glm(formula = ClaimAmount ~ DrivAge + HasKmLimit + RiskVar +
    VehAge1 + VehAge10 + VehAge2 + VehAge5 + VehPriceE + VehPriceI +
    VehPriceJ + VehPriceK + VehPriceM + VehPriceN + VehPriceO +
    VehPriceQ + VehEnginedirect_injection_overpowered + VehEngineinjection +
    VehEngineinjection_overpowered + VehEnergyregular + VehMaxSpeed150_160_km_h +
    VehMaxSpeed160_170_km_h + VehMaxSpeed180_190_km_h + VehClassB +
    VehClassH + VehClassM2 + VehUsagePrivate_trip_to_office +
    VehUsageProfessional_run + VehBodymicrovan + SocioCategCSP4 +
    SocioCategCSP5 + SocioCategCSP7, family = Gamma(link = "log"),
    data = cbind.data.frame(ClaimAmount = data1.app.sinistre$ClaimAmount,
        data1.app.sinistre.reg), weights = VehBodymicrovan +
        VehBodycoupe + SocioCategCSP4 + SocioCategCSP5 + SocioCategCSP7 +
        SocioCategCSP9)
```

Figure: Meilleur modèle retenu

Modélisation avec la régression gamma

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               7.383319   0.677019  10.906  < 2e-16 ***
DrivAge                                   0.014442   0.005662   2.551  0.01107 *
HasKmLimit                               -0.159809   0.215620  -0.741  0.45896
RiskVar                                   0.018559   0.016143   1.150  0.25087
VehAge1                                   0.328137   0.263447   1.246  0.21354
VehAge10                                  0.093088   0.258903   0.360  0.71934
VehAge2                                  -0.056506   0.249890  -0.226  0.82120
VehAge5                                   0.488589   0.257525   1.897  0.05840 .
VehPriceE                                -0.365843   0.337016  -1.086  0.27824
VehPriceI                                 0.452912   0.289924   1.562  0.11891
VehPriceJ                                 0.328157   0.267134   1.228  0.21989
VehPriceK                                -0.080717   0.293977  -0.275  0.78377
VehPriceM                                -0.232044   0.351574  -0.660  0.50957
VehPriceN                                 0.026796   0.336201   0.080  0.93651
VehPriceO                                -0.163929   0.359066  -0.457  0.64821
VehPriceQ                                -0.095098   0.385895  -0.246  0.80545
VehEnginedirect_injection_overpowered    -0.628649   0.546460  -1.150  0.25056
VehEngineinjection                        0.265469   0.383540   0.692  0.48918
VehEngineinjection_overpowered           -0.111574   0.461272  -0.242  0.80898
VehEnergyregular                         -0.849864   0.302084  -2.813  0.00511 **
VehMaxSpeed150_160_km_h                  -0.393068   0.252600  -1.556  0.12035
VehMaxSpeed160_170_km_h                   0.250647   0.261243   0.959  0.33783
VehMaxSpeed180_190_km_h                  -0.153553   0.252847  -0.607  0.54395
VehClassB                                -0.034883   0.231841  -0.150  0.88046
VehClassH                                 0.205526   0.294252   0.698  0.48523
VehClassM2                               -0.275346   0.227198  -1.212  0.22615
VehUsagePrivate_trip_to_office            0.021499   0.179985   0.119  0.90497
VehUsageProfessional_run                 -1.939015   0.857907  -2.260  0.02426 *
VehBodymicrovan                           0.227835   0.297853   0.765  0.44470
SocioCategCSP4                           -0.344154   0.251654  -1.368  0.17210
SocioCategCSP5                           -0.736672   0.347925  -2.117  0.03475 *
SocioCategCSP7                           -3.160070   1.240319  -2.548  0.01116 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 2.694847)

    Null deviance: 956.26  on 505  degrees of freedom
Residual deviance: 764.12  on 474  degrees of freedom
AIC: 9269.1
```

Modélisation avec la régression gamma

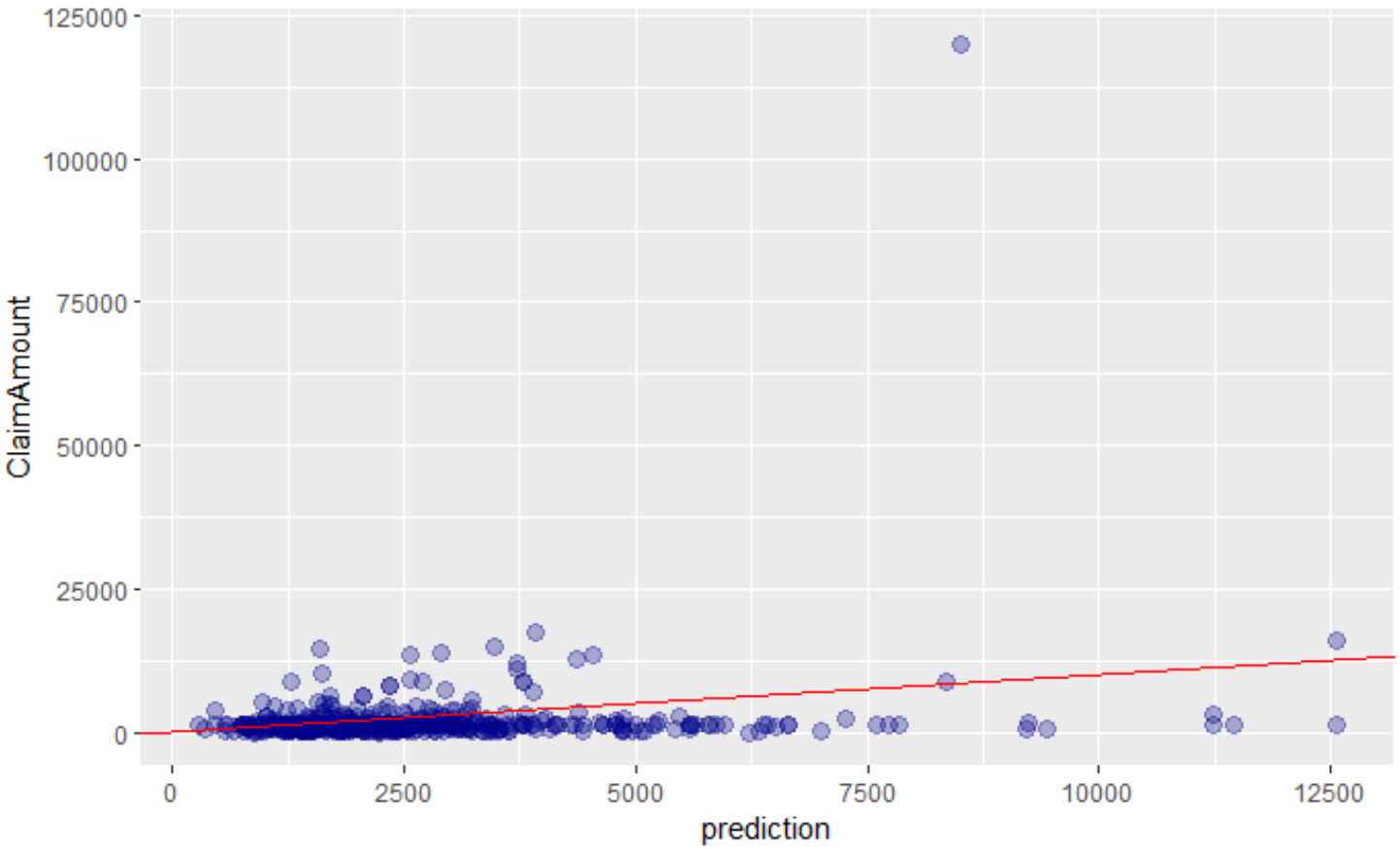


Figure: Performance du modèle pour la prédiction

Modélisation avec la régression gamma

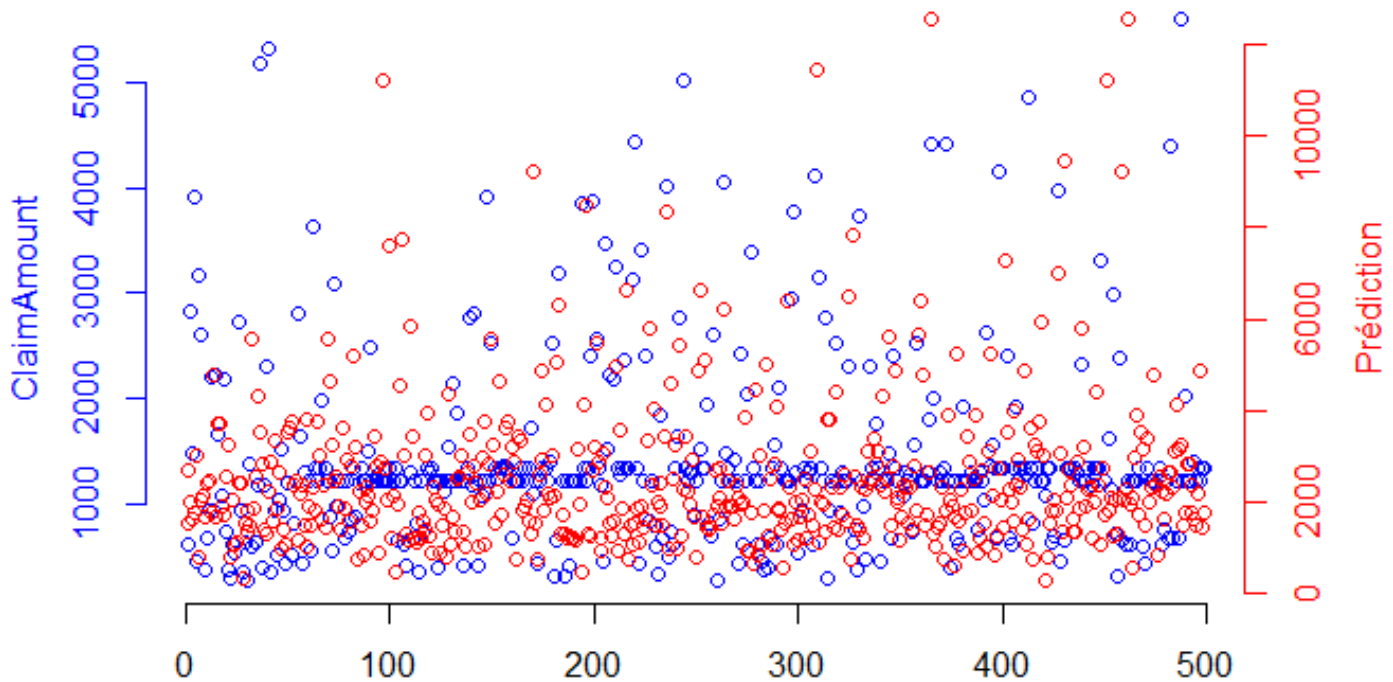


Figure: Comparaison de valeurs prédites et les vraies valeurs

Modélisation avec la régression gamma

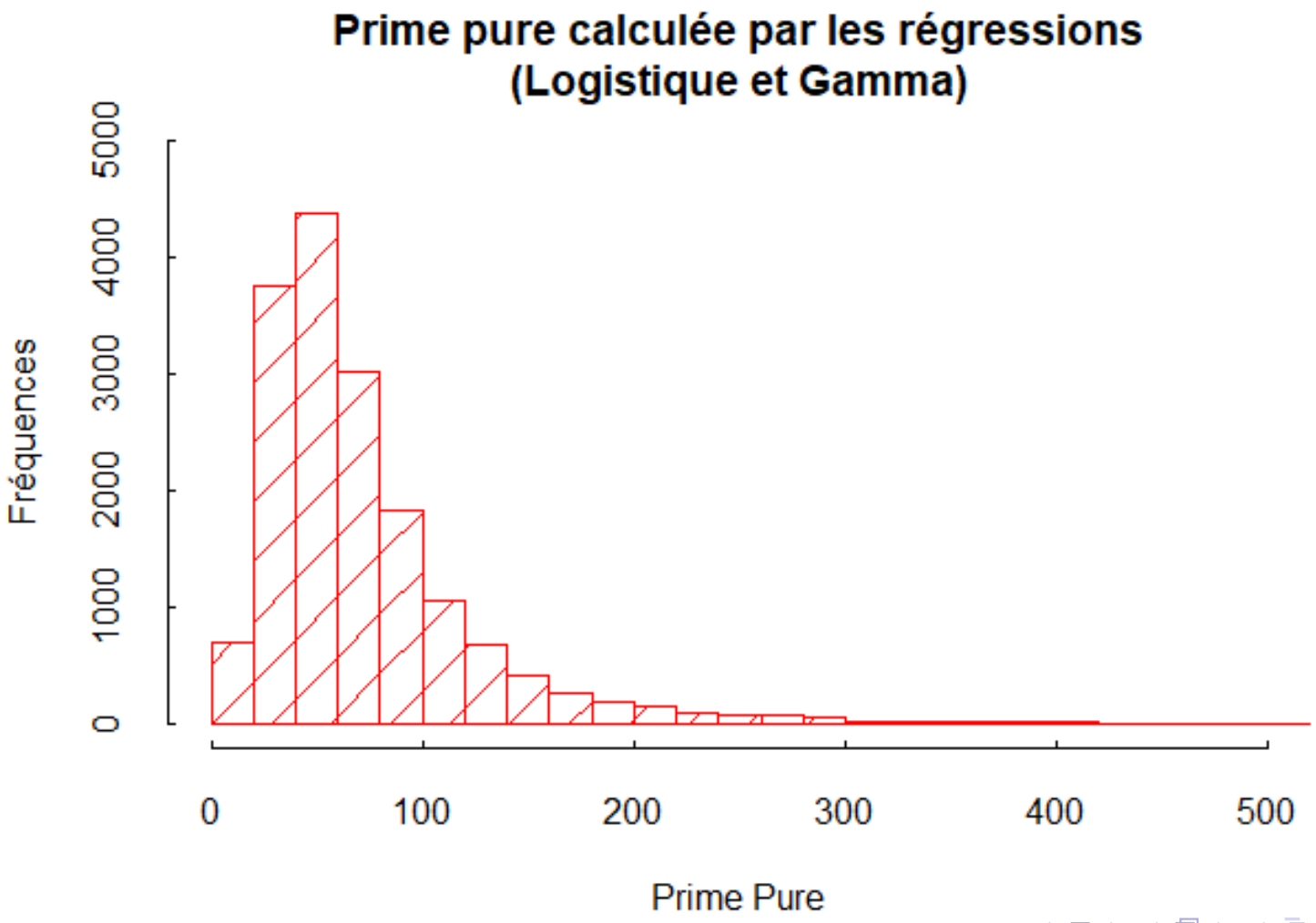
- D'après les deux graphiques, on voit que le modèle s'adapte bien aux données...
- et ses prédictions sont bonnes.
- On conclut que le modèle est acceptable.
- Nous allons maintenant calculer la prime pure en utilisant les deux modèles de régression (logistique et gamma).

Calcul de la prime pure

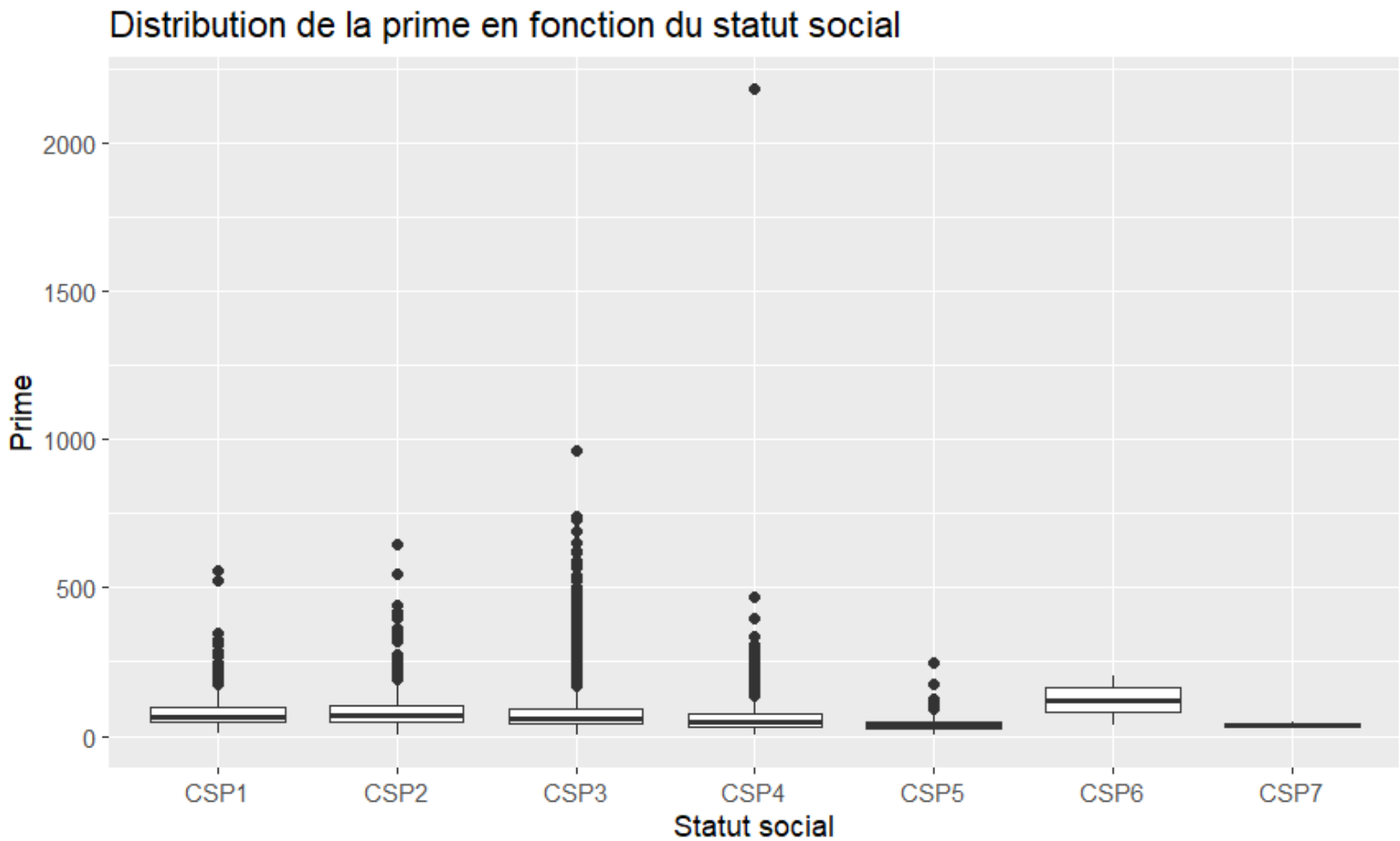
$$\begin{aligned}
\mathbb{E}\left[S_i \mid \mathbf{X}_i\right] &= \mathbb{E}\left[N_i \mid \mathbf{X}_i\right].\mathbb{E}\left[Y_i \mid \mathbf{X}_i\right] \\
&= \frac{\exp\left(\mathbf{x}_i'\hat{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\hat{\beta}\right)}.\exp\left(\mathbf{x}_i'\hat{\beta}\right) \\
&= (Predict.ClaimInd)_i.(Predict.ClaimAmount)_i \\
&= 73.7467
\end{aligned}$$

- L'assureur fera payer aux assurés cette prime pour n'est pas perdre d'argent en moyenne.
- La prime pure est repartie entre 1.581 et 2178.716.
- Avec une moyenne de 73.7467, le premier quartile (resp. troisième) est de 39.048 (resp. 88.130).

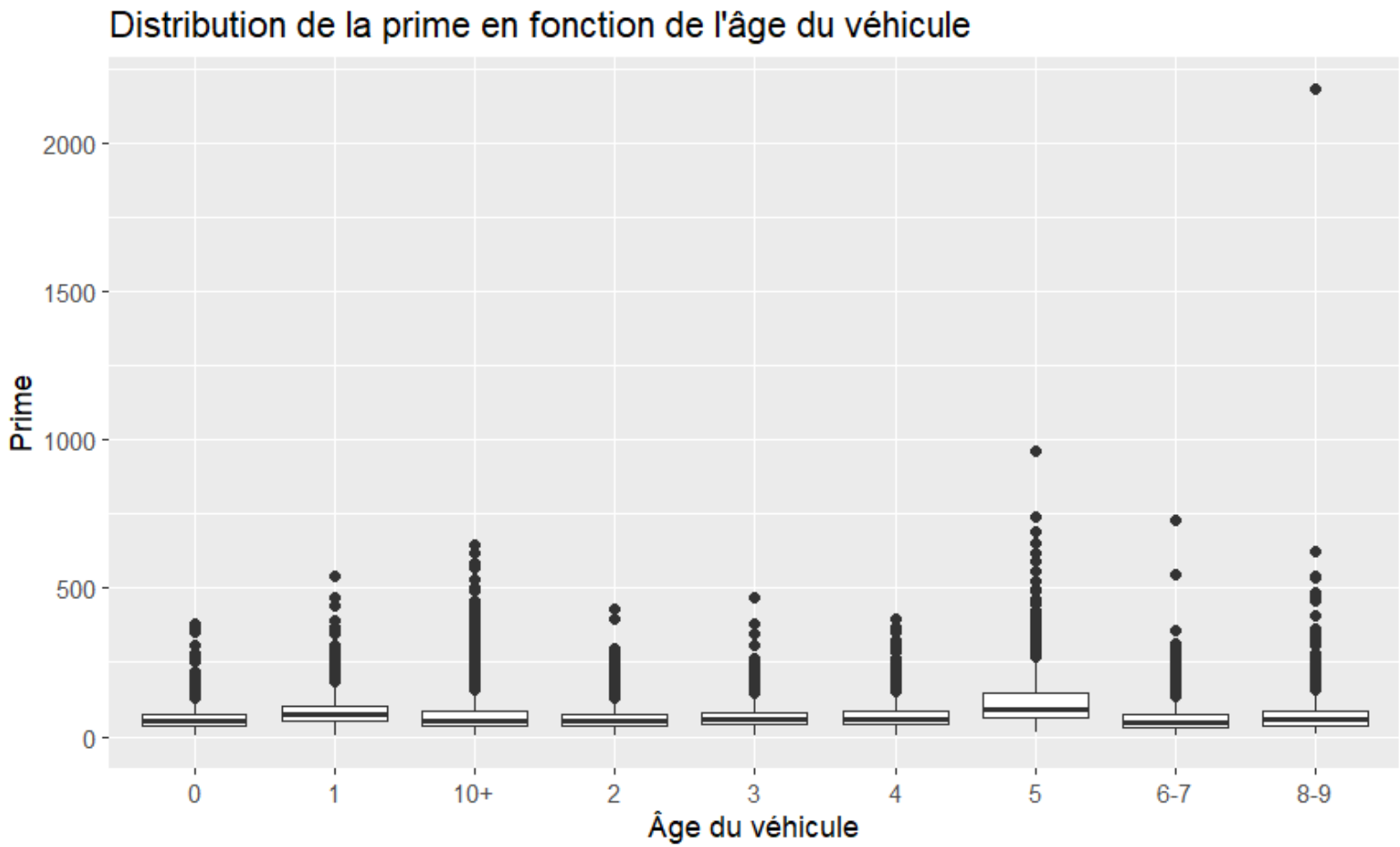
Distribution de la prime pure



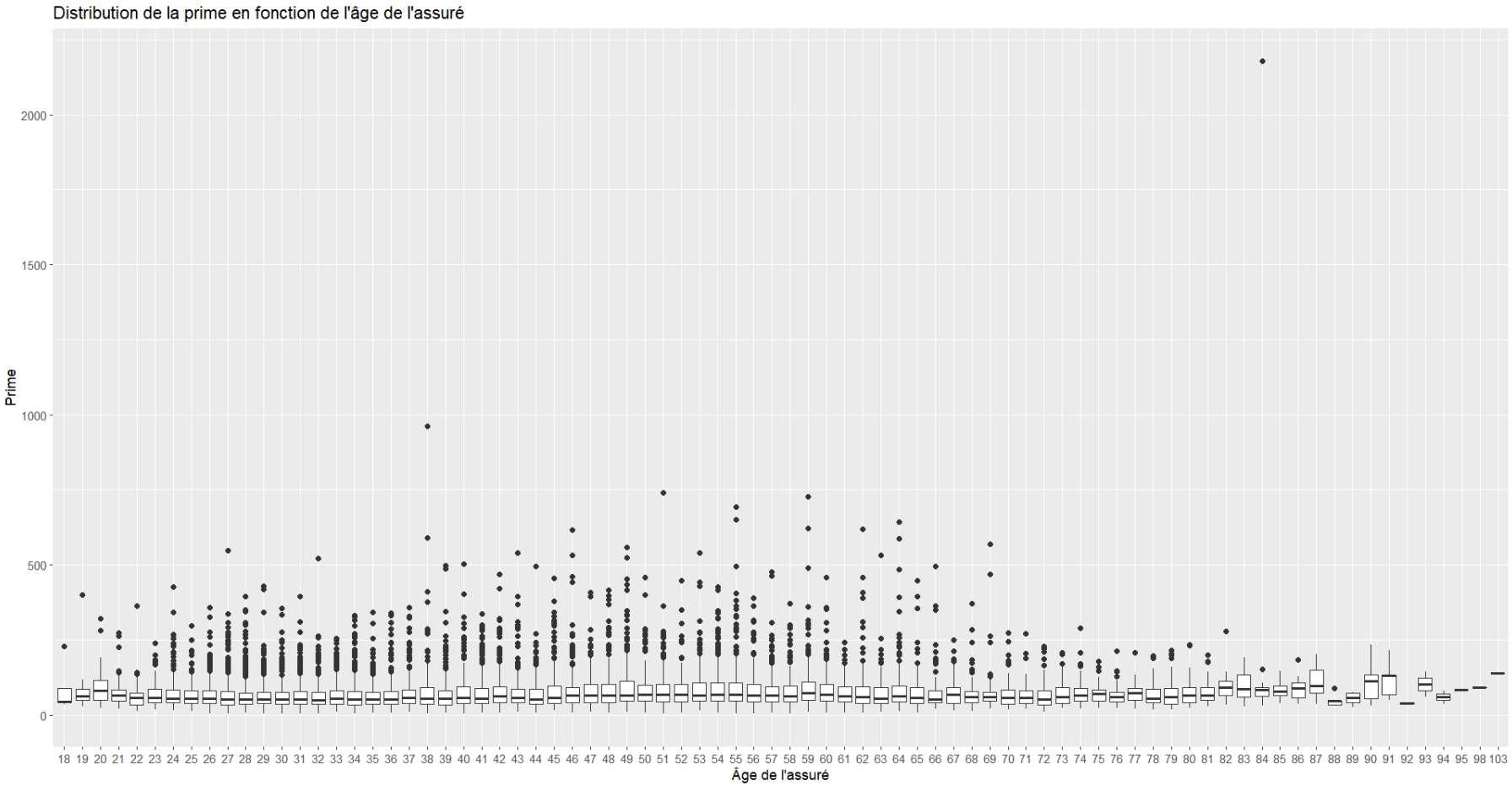
Répartition de la prime pure



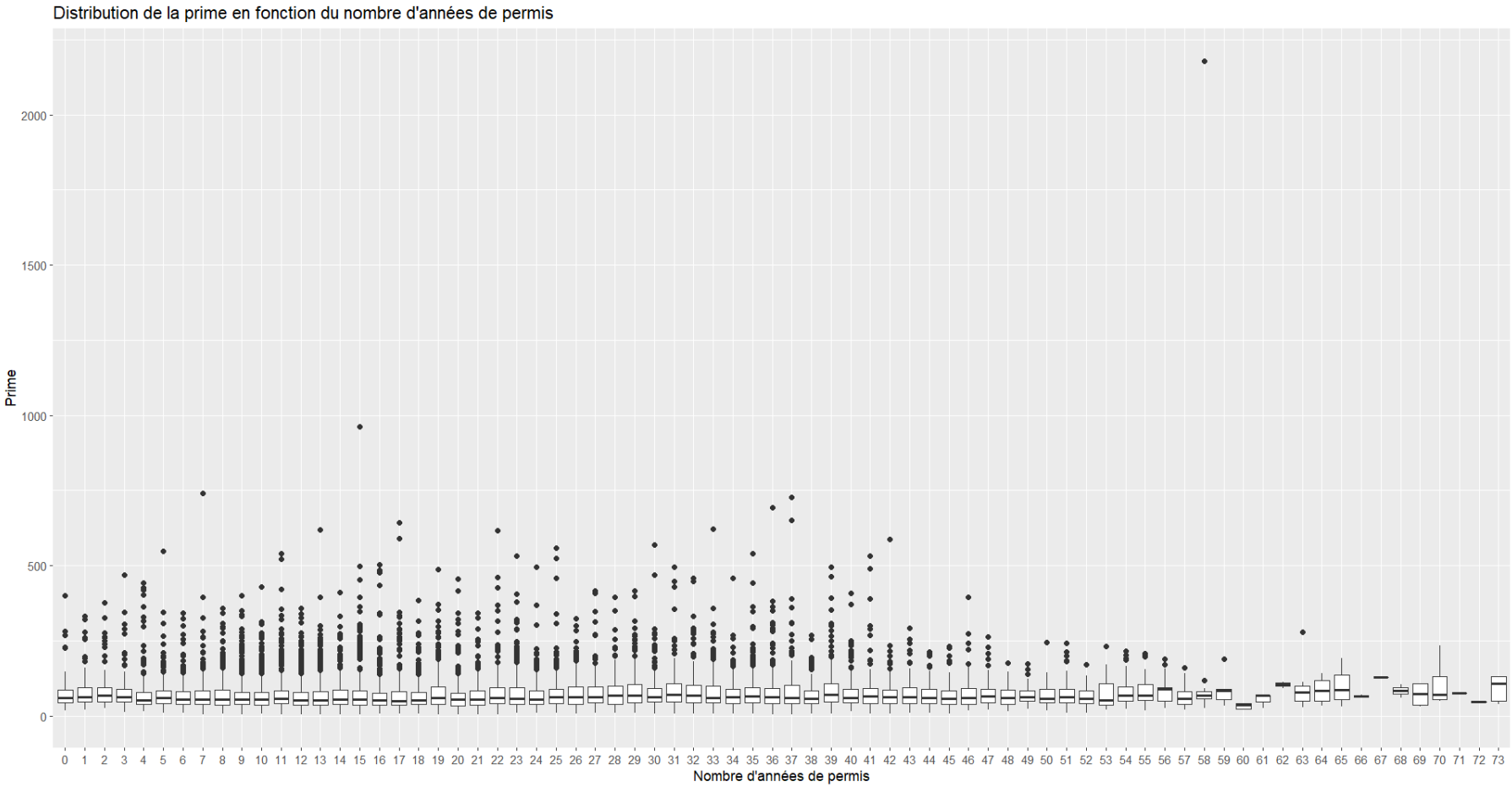
Répartition de la prime pure



Répartition de la prime pure



Répartition de la prime pure



Prime pure et majorée

- La compagnie pourrait ajouter une marge de sécurité η, avec $0 < \eta < 1$..
- .. à la prime pure pour éviter que la compagnie fasse faillite.
- Il est possible de simuler le coefficient η pour que..
- .. la somme des primes soit supérieure à la somme des montants des sinistres.
- Il faudrait surveiller la probabilité ruine ϕ_n, car le contrôle mener par l'autorité de régularisation (exemple : arca[2]) aux assureurs..
- ..consiste aussi à surveiller cette probabilité afin de protéger les assurés.

[2] www.arca.cd

Discussion

- Les problèmes sur le biais, la discrimination et l'équité de la prime.
- Comment s'assurer qu'un modèle de tarification ne discrimine pas en fonction de certaines informations sensibles (sexe, âge, etc.) ? voir Grari et al.(2022), Frees and Huang(2021).
- Les données télématiques de conduite automobile sont utilisées dans la tarification mais..
- la question sur la transparence de ces données entraîne des problèmes de confidentialité.
- On peut facilement distinguer les bons conducteurs en fonction de leur style de conduite grâce aux cartes thermiques télématiques..
- .. et sur les séries temporelles de trajets individuels, voir Gao et al.(2022).

Conclusion

- Nous avons présenté le processus de la tarification dans le cas d'une assurance non vie, le choix du modèle et le calcul de la prime pure.
- Il existe plusieurs modèles statistiques qui tiennent comptent d'autres paramètres. Dans le cas où la linéarité n'est pas garantie.
- Dans un marché concurentiel, chaque assureur utilise ses méthodes statistiques pour la tarification mais..
- .. il est plus prudent de choisir un modèle qui segmente avec parcimonie.

Conclusion

Car une segmentation trop poussée pourrait conduire à :

- une spirale de segmentation toujours croissante : les assureurs doivent toujours s'aligner sur les autres compagnies pour ne pas attirer que les mauvais risques.
- une inassurabilité : il est plus facile d'identifier des mauvais assurés, qui auront de la difficulté à payer leur prime.
- des frais de fonctionnement plus élevés : on se retrouve dans une situation où les assurés changent de compagnie fréquemment.

[1] H. Bühlmann(1969). Experience rating and credibility. *ASTIN Bulletin: The Journal of the IAA*, 5(2):157–165.

[2] A. Charpentier, M. Denuit, and R. Elie(2015). Segmentation et mutualisation, les deux faces d'une même pièce? *Risques*, 103:57–64, 2015.

[3] A. Charpentier(2014). *Computational actuarial science with R*. CRC press.

[4] M. Denuit and A Charpentier(2005). *Mathématiques de l'Assurance Non-Vie. Tome I: Principes Fondamentaux de Théorie du Risque*. .

[5] M. Denuit and A Charpentier(2005). *Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement*. .

[6] M. Denuit(2005). Quand la différenciation tarifaire est-elle

techniquement justifiée? *Monde de l'Assurance, Dossier spécial*, 1:39661.

[7] E.W. Frees and F. Huang(2021). The discriminating (pricing) actuary. *North American Actuarial Journal*, pages 1–23.

[8] E.W. Frees(2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press.

[9] G. Gao, S. Meng, and M.V. Wüthrich(2022). What can we learn from telematics car driving data: a survey. *Insurance: Mathematics and Economics*.

[10] V. Goulet(2008). Credibility theory. *Encyclopedia of Quantitative Risk Analysis and Assessment*, 1.

[11] V. Grari, A. Charpentier, S. Lamprier, and M. Detyniecki(2022). A fair pricing model via adversarial learning. *arXiv preprint arXiv:2202.12008*.

[12] T.J. Hastie and R.J. Tibshirani(1990). *Generalized additive models*. Chapman et Hall.

[13] G. James, D. Witten, T. Hastie, and T. Tibshirani(2013). *An introduction to statistical learning*, volume 112. Springer.

[14] R. Kaas, M. Goovaerts, J. Dhaene, and M. Denuit(2008). *Modern actuarial risk theory: using R*, volume 128. Springer Science & Business Media.

[15] P. McCullagh and J.A. Nelder(1991). *Generalized linear models*. Routledge.

[16] J.A. Nelder and W.M.R. Wedderburn(1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

[17] E. Ohlsson and B. Johansson(2010). *Non-life insurance pricing with generalized linear models*, volume 174. Springer.

[18] G. Saporta(2006). *Probabilités, analyse des données et statistique*. Editions technip.