UCLouvain

Institut de Statistique, Biostatistique et Sciences Actuarielles

Corrélation entre nombres et coûts des sinistres.

Membres du jury :

Prof. DENUIT Michel *Promoteur*
Prof. WALHIN Jean-François *Lecteur*

Mémoire présenté en vue de
l'obtention du master
en sciences actuarielles
par :

NGUYEN Alisson
SPOORENBERG Brieuc

Louvain-la-Neuve
Août 2020

Remerciements

Nous tenons tout d'abord à remercier notre promoteur Monsieur Michel Denuit pour son suivi et ses conseils tout au long de notre recherche. Nous remercions également nos proches pour leur soutien au cours de notre master, ainsi que les personnes ayant pris le temps de lire et commenter notre travail.

Table des matières

Table des figures

Liste des tableaux

Introduction

Habituellement, en assurance non-vie, la fréquence et la sévérité des sinistres sont supposées indépendantes. Un des premiers modèles se basant sur cette hypothèse a été développé par Lundberg 1903 en appliquant 2 distributions fréquemment utilisées par la suite : la loi Gamma pour le montant des sinistres et celle de Poisson pour le nombre de sinistres.
Cette hypothèse, bien que très pratique pour le calcul de la perte attendue qui peut s'exprimer comme le produit de la fréquence et de la sévérité, n'est pas forcément réaliste. En effet, des publications récentes de la littérature actuarielle relâchent cette hypothèse dans le cadre des assurances automobiles. Celles-ci supposent que le nombre et les montants des sinistres sont dépendants. Afin de modéliser cette dépendance, nous avons relevé trois principaux axes d'approche différents : le modèle dépendant en deux parties, le modèle de copules et le modèle à effets aléatoires dépendants.

Le premier type de modèle, qui consiste en une extension intuitive du modèle indépendant, ajoute en facteur explicatif le nombre de sinistres dans le modèle marginal de la sévérité. Ce type de modèle a, par exemple, été utilisé dans les travaux de Garrido et al. 2016, W. Lee et al. 2019 ou Gschlössl et Czado 2007.
Cette méthode d'incorporation de la dépendance ne modifie pas la composante de fréquence par rapport au modèle indépendant. Néanmoins, pour la sévérité, la dépendance requiert l'utilisation du nombre de sinistres comme variable explicative. Les résultats apportés par ces auteurs montrent une amélioration de la prédiction par rapport aux modèles ignorant l'existence d'une dépendance. En prenant le problème dans l'autre sens, Krämer et al. 2013 font l'hypothèse que le nombre de sinistres peut être expliqué par la sévérité. Dans leur recherche, c'est donc le modèle marginal du nombre de sinistres qui est modifié.

Un deuxième type de modèle utilise des copules afin d'étudier la dépendance. Ces dernières permettent de lier plusieurs densités marginales afin de déterminer leur distribution jointe. Elles sont ainsi, par exemple, fréquemment utilisées dans le domaine de la finance pour modéliser la dépendance entre des variables aléatoires. Dans le domaine actuariel, nous retrouvons leur application seulement depuis récemment. Dans le cadre de notre recherche, nous nous intéresserons aux copules bivariées qui nous permettent de lier le nombre de sinistres à leur montant moyen afin d'extraire la structure de dépendance sous-jacente. Tout en permettant des dépendances très flexibles, l'implémentation de ce type de modèle reste relativement simple et peut facilement s'adapter à des structures plus complexes.

Krämer et al. 2013 et Czado et al. 2012, par exemple, utilisent des copules bivariées. Ils emploient notamment la copule Gaussienne, afin de joindre les distributions marginales du nombre de sinistres et de la sévérité moyenne. Cossette et al. 2019 et Frees et al. 2016 quant à eux, proposent d'utiliser les copules afin de modéliser la dépendance entre les variables aléatoires du nombre de sinistres et chaque montant de sinistre individuel. Encore récemment, G. Y. Lee et Shi 2019 utilisent les copules afin de modéliser la dépendance entre le temps séparant deux sinistres et leur valeur. Ceci est une autre façon d'étudier la dépendance entre la fréquence et la sévérité des sinistres. La principale difficulté de ce type de modèle provient de la nécessité de choisir une famille de copules appropriée pour les données. De plus, l'estimation des paramètres ainsi que la simulation de résultats à l'aide de cette méthode peut également nécessiter des calculs importants.

Enfin, la dépendance peut être introduite avec l'utilisation de modèles à effets aléatoires dépendants. Les effets aléatoires permettent de modéliser des variables non observables pour chaque individu et qui ont un effet sur la variable de réponse. De ce fait, comme présenté dans les travaux de Baumgartner et al. 2015, l'utilisation des mêmes effets aléatoires entre la sévérité et le nombre de sinistres permet d'introduire une dépendance entre ces deux variables. Une autre solution, présentée par Jeong et al. 2017 dans leur recherche, est d'utiliser des effets aléatoires corrélés dans les deux modèles marginaux. En comparaison avec les approches présentées précédemment, ici, l'hypothèse est que la dépendance ne provient pas directement d'une relation entre la sévérité et la fréquence mais plutôt de facteurs inconnus qui influencent simultanément ces deux variables. La dépendance serait donc seulement apparente et pas nécessairement réelle.

Même si la majorité des auteurs s'accorde à dire qu'une dépendance existe, les résultats de ces différentes recherches n'ont pas fait émerger de consensus quant au signe de cette dernière. Nous observons en effet que les résultats diffèrent d'une étude à l'autre. Ceci dépend notamment du type d'assurance considéré, des différentes caractéristiques de la base de données utilisée dans le cadre des recherches ou encore du modèle appliqué.
Tout d'abord selon le type de couverture considéré, par exemple, Park et al. 2018 trouvent une dépendance positive dans le cadre de l'assurance responsabilité civile alors qu'en assurance automobile dégâts matériels, ils trouvent une dépendance négative. Les auteurs indiquent que cette différence de signe pourrait être expliquée par le phénomène de "Bonus Hunger", c'est-à-dire que le comportement des individus change en fonction de leur niveau de Bonus-Malus. La différence de comportement entre les deux types de couvertures réside dans le choix que l'assuré a de déclarer son sinistre à son assureur ou non.
Le signe de la dépendance peut également varier selon le modèle considéré. Par exemple, Krämer et al. 2013 et Baumgartner et al. 2015 utilisent les mêmes données mais trouvent des résultats différents. Baumgartner et al. 2015 obtiennent une dépendance négative en utilisant un modèle à effets aléatoires partagés. À l'inverse, Krämer et al. 2013 trouvent une dépendance positive entre les deux variables en travaillant avec les copules. Dans une autre étude, Jeong et al. 2017 observent quant à eux que le signe de la

dépendance s'inverse en introduisant des effets aléatoires à leur modèle de régression.

Les résultats inconsistants semblent se confirmer dans l'ensemble de la littérature sans qu'une tendance concrète entre une dépendance positive, décelée notamment dans les travaux de Czado et al. 2012, Jeong et al. 2017 ou encore Shi et al. 2015 et une dépendance négative, trouvée par exemple dans Oh, Ahn et al. 2019, Schulz 2013 ou encore Baumgartner et al. 2015, ne se dégage.

Dans le cadre de notre recherche, nous présenterons trois différents modèles afin d'essayer de comprendre la dépendance entre la fréquence et la sévérité des sinistres en assurance automobile ainsi que son impact sur la tarification des contrats.

Tout d'abord, nous nous intéresserons au modèle en deux parties dépendant, en prenant comme point de départ la méthodologie développée par Garrido et al. 2016 prenant le nombre de sinistre comme variable explicative de la sévérité. Ce modèle, relativement simple, devrait nous donner une première idée sur l'existence d'une dépendance et du signe de celle-ci. Le modèle présenté dans ce papier a l'avantage de fournir une forme simple et intuitive pour le coût total moyen. Celui-ci peut en effet s'écrire comme le produit de trois termes : la fréquence marginale moyenne, la sévérité marginale moyenne corrigée et un terme de correction pour la dépendance. D'un point de vue pratique, ce modèle a l'avantage d'être très intuitif puisque ce dernier est une simple extension du modèle standard utilisé actuellement.
Néanmoins, celui-ci intègre directement le nombre de sinistres comme facteur explicatif de la sévérité et ne tient pas compte de l'exposition de chaque police lors du calcul de la dépendance. Une police ayant eu un sinistre dans l'année sera interprétée de la même façon qu'une police observée sur une durée d'un mois avec un sinistre.
De plus, cette méthode, par sa construction, n'utilise qu'un seul coefficient pour représenter la dépendance. Ceci n'est pas forcément adapté pour des cas où la structure de cette dernière est plus complexe. Dans un deuxième temps, nous étendrons le modèle afin d'essayer d'apporter une réponse à ces deux limitations grâce à l'utilisation de la fréquence annuelle.

Ensuite, nous appliquerons sur nos données la méthode développée par Krämer et al. 2013 qui intègre les copules aux modèles marginaux. Pour cela, nous utiliserons le package `R CopulaRegression` développé dans le cadre de leur recherche.
Nous avons choisi d'utiliser ce dernier car il permet d'observer l'impact de la dépendance sur la perte du portefeuille et donc in-fine sur la tarification.
Les copules ont comme avantage de proposer une classe flexible de modèles de dépendance y compris entre une variable discrète et continue. Par conséquent, elles permettent de modéliser un grand nombre de structures de dépendance plus ou moins complexes.
De plus, la dépendance ne sera influencée que par le choix de la famille de copules utilisée et non par celui des distributions marginales. Dans le cas que nous considérons, une des

distributions marginales est discrète, de ce fait, le copule n'est pas unique. Il est donc important de choisir une famille de copules adéquate, celle-ci pouvant avoir un effet important sur le résultat obtenu. Nous analyserons donc l'impact du choix de la famille de copules sur la dépendance trouvée.

Finalement, nous intégrerons des effets aléatoires dépendants à notre modèle de régression en nous appuyant sur la méthode présentée par Baumgartner et al. 2015. Par rapport aux démarches présentées précédemment, celle-ci a plusieurs propriétés intéressantes.
Premièrement, elle permet d'introduire une dépendance dont le niveau peut être variable en fonction de la valeur des facteurs explicatifs. Ceci n'est pas possible avec les deux modèles présentés antérieurement.
Deuxièmement, la dépendance n'est pas étudiée sous le même angle que précédemment. Alors que les deux autres modèles présument une dépendance directe, celui-ci suppose que la dépendance observée provient de facteurs inconnus qui influencent à la fois le niveau de la fréquence et de la sévérité. De plus, ce type de modèle appliqué à des données assurantielles, peut être utilisé dans le cadre de la théorie de la crédibilité, par l'ajustement de la distribution de ces effets aléatoires. Cette dernière permet de quantifier l'impact de la dépendance sur la tarification à posteriori.
Pour l'estimation des paramètres, les auteurs utilisent une méthode Bayésienne basée sur un algorithme de Monte-Carlo par chaînes de Markov. De notre côté, nous appliquerons une estimation par maximum de vraisemblance, nécessitant un algorithme d'approximation de l'intégrale. Après avoir vérifié que cette méthode donnait des résultats satisfaisants, nous avons fait ce choix dans le but d'optimiser le temps de calcul nécessaire à l'application de ce modèle. En effet, dans le cadre considéré, les algorithmes Bayésiens ne fournissent que des résultats marginalement meilleurs que d'autres méthodes plus couramment utilisées. Par exemple, Baumgartner 2014 ne relève qu'une amélioration marginale apportée par la méthode Bayesienne par rapport à celle du maximum de vraisemblance. Nous avons donc estimé qu'il était plus efficient d'appliquer un algorithme d'approximation numérique.
Ces modèles généralisés mixtes, en pratique souvent utilisés pour des données longitudinales, seront ici étendus à un contexte bivarié. Ainsi, chaque régression marginale inclut un même effet aléatoire gaussien lié par un paramètre de mise à l'échelle. Le signe de ce dernier nous permettra d'analyser la dépendance. Par la suite, afin d'apporter davantage de flexibilité, les effets aléatoires des deux modèles marginaux seront liés par une loi normale bivariée. La valeur du coefficient de corrélation permettra ici d'interpréter la dépendance.

En parcourant la littérature, nous avons relevé d'autres méthodes intéressantes qui ne seront néanmoins pas appliquées dans le cadre de ce travail. Oflaz et al. 2019, par exemple, introduisent une dépendance par l'existence d'états cachés, sur lesquels aucune information n'est disponible. Selon leur raisonnement, des conditions macro (par exemple économiques, climatiques ou financières) affectent à la fois la fréquence et la sévérité des sinistres. Ceci engendre donc des interactions apparentes entre ces deux composantes. Les auteurs choi-

sissent de modéliser ces états cachés en utilisant un modèle de Markov permettant d'introduire une dépendance à la fois mutuelle et sérielle. Ce type de démarche permet de conserver une grande flexibilité dans le type de données qu'il est possible de traiter, ainsi que dans les explications de la dépendance. Il est intéressant de noter que les études basées sur une approche Bayésienne restent peu courantes pour le problème que nous considérons. Une des principales faiblesses provient de l'interprétation des états cachés. En effet, afin de pouvoir obtenir des résultats intéressants, un grand travail de recherche doit être réalisé pour les relier à des évènements réels. Ce type de considération sortant du cadre de notre recherche, nous avons choisi de ne pas l'appliquer ici.
La dépendance entre la fréquence et la sévérité peut également être étudiée par le biais de la relation entre le temps séparant deux sinistres et leur montant. Par exemple, dans leur travaux G. Y. Lee et Shi 2019 ou Boudreault et al. 2006, les auteurs supposent que la taille du sinistre dépend du temps écoulé depuis le dernier sinistre. Par manque de données disponibles, nous ne pourrons appliquer ce type d'approche.

Dans le cadre de notre recherche, le premier chapitre reprend les concepts de bases nécessaires à celle-ci. Dans le chapitre suivant, la base de données utilisée, provenant du portefeuille en RC automobile d'une compagnie d'assurance française, est d'abord présentée. Dans ce même chapitre, nous validons également les distributions choisies, et enfin, nous présentons les modèles de régression que nous développons. Nous commençons par un modèle en deux parties indépendant puis incluant une dépendance conditionnelle. Ce modèle est ensuite étendu pour inclure la dépendance à l'aide de copules. Pour finir, nous modélisons la dépendance à l'aide d'effets aléatoire liés. Le troisième chapitre est destiné à l'application de ces modèles ainsi qu'à la présentation des différents résultats. Ce chapitre est clôturé par le choix d'un modèle final. Enfin, ce dernier est utilisé dans le quatrième chapitre pour étudier l'impact de la dépendance sur la tarification a priori et a posteriori.

Chapitre 1

Concepts de base

1.1 Symboles

Dans cette section, nous reprenons quelques notations ainsi que leur description utilisées dans le cadre de notre recherche.

Symbole	Description
$\mathbf{m}$	Nombre de polices
$\mathbf{Y_i}$	Nombre de sinistres de la police i, avec $i = 1, ..., m$
$\mathbf{X_{ik}}$	Montant des sinistres de la police i, avec $k = 1, .., Y_i$
$\mathbf{e}$	Exposition
$\mathbf{z_i}$	Variables explicatives pour la sévérité
$\mathbf{w_i}$	Variables explicatives pour la fréquence
$\boldsymbol{\beta}$	Coefficients pour le modèle de sévérité
$\boldsymbol{\alpha}$	Coefficients pour le modèle de fréquence
$\mathbf{S}$	Matrice de design du modèle de fréquence
$\mathbf{R}$	Matrice de design du modèle de sévérité
$\boldsymbol{\mu_{X_{ik}}}$	Prédiction de la sévérité à l'aide du GLM
$\boldsymbol{\mu_{Y_i}}$	Prédiction du nombre de sinistres à l'aide du GLM
$\boldsymbol{\psi}$	Effets aléatoires du modèle de sévérité
$\boldsymbol{\phi}$	Effets aléatoires du modèle de fréquence
$\boldsymbol{\tau}$	Tau de Kendall
$\boldsymbol{\rho}$	Coefficient de corrélation
$\mathbf{\Phi}$	Fonction de distribution de la loi normale

TABLE 1.1.1 – Présentation des différentes notations

1.2 La famille exponentielle

La famille exponentielle est une famille de distribution utilisée pour étendre les distributions disponibles dans les modèles de régression. En effet, dans une régression linéaire classique, nous faisons l'hypothèse que les résidus suivent une distribution normale.
Les modèles de régression linéaire généralisée permettent d'utiliser une des distributions de la famille exponentielle pour les erreurs. Cette extension est particulièrement intéressante en assurance. En effet, pour la sévérité tout comme pour la fréquence, l'hypothèse de normalité n'est pas raisonnable.

Définissons plus formellement cette famille de distribution. La loi de probabilité des distributions appartenant à cette dernière peut s'écrire d'une manière commune. Nous donnons ici la formulation générale, mais elle peut évidemment être reparamétrée pour la rendre plus adaptée aux usages en sciences actuarielles, par exemple.

Ainsi, la densité, $P[Y_i = y]$ pour les distributions discrètes ou $f_{Y_i}(y)$ pour les distributions continues, a la forme suivante :

$$\exp\left(\frac{y\theta_i - a(\theta_i)}{\phi}\right) c(y, \phi) \tag{1.2.1}$$

Avec :

- θ_i les paramètres canoniques. Ce sont ces paramètres qui contiennent les effets des variables explicatives dans le modèle de régression.
- $a(\theta_i)$ la fonction cumulatrice.
- $\phi > 0$ le paramètre de dispersion, lié à la variance.
- $c(y, \phi)$ une fonction indépendante de θ_i.

Pour ces distributions, il est également possible de définir des expressions générales pour la moyenne et la variance. Il peut être effectivement montré qu'elles sont respectivement données par :

$$\mathrm{E}[Y] = a'(\theta_i) \qquad \mathrm{Var}[Y] = a''(\theta_i)\phi$$

Il est aussi utile de définir la fonction de variance qui, elle, est donnée par :

$$\mathrm{V}(\mu_i) = a''(a'^{-1}(\mu_i))$$

avec μ_i, la moyenne pour une police i.
Pour obtenir la variance, il suffit de multiplier la fonction de variance par le paramètre de dispersion (ϕ). La variance est donc une fonction de la moyenne. Lorsque nous modélisons la moyenne en utilisant un modèle de régression, cela permet ainsi d'obtenir indirectement des informations sur la variance. L'estimation de la variance est néanmoins limitée et a peu de liberté.

Toute une série de distributions couramment utilisées appartiennent à cette famille. Le tableau 1.2.1 reprend certaines de ces familles à titre d'exemple. Notons ici que les distributions sont paramétrées en fonction de la moyenne.

Distribution	θ	$a(\theta)$	ϕ	E[Y]	V(μ_i)
Bin(1,μ_i)	$\ln\left(\frac{\mu_i}{1-\mu_i}\right)$	ln(1+exp(θ))	1	μ_i	$\mu_i(1-\mu_i)$
Poi(μ_i)	ln(μ_i)	exp(θ)	1	μ_i	μ_i
Nor(μ_i,σ^2)	μ_i	$\frac{\theta^2}{2}$	σ^2	μ_i	1
Gam(μ_i,α)	$-\frac{1}{\mu_i}$	$-\ln(-\theta)$	$\frac{1}{\alpha}$	μ_i	μ_i^2
Twe(μ_i,ϕ)	$\frac{\mu_i^{1-q}}{1-q}$	$\frac{((1-q)\theta)^{\frac{2-q}{1-q}}}{2-q}$	ϕ	μ_i	μ_i^q, 1<q,2

TABLE 1.2.1 – Distributions appartenant à la famille exponentielle

Dans la suite de ce mémoire, nous nous intéressons aux polices ayant au moins un sinistre. Il est donc utile de définir la distribution de Poisson tronquée en 0, qui sera appelée par la suite Poisson tronquée (ZTP). Sa densité est définie comme la densité conditionnelle d'une Poisson, c'est-à-dire, de façon générale :

$$\begin{aligned} p_Y(y) = P[Y^* = y | Y^* \geq 1] &= \frac{e^{-\lambda}\lambda^y}{y\,!\,(1-e^{-\lambda})} \\ &= \exp\big(\mathrm{y}\ln(\lambda) - \lambda - \ln(1-\mathrm{e}^{-\lambda})\big)\,\frac{1}{y!} \end{aligned} \tag{1.2.2}$$

avec $Y^* \sim Poi(\lambda)$.

Cette dernière formulation montre ainsi que la distribution ZTP fait également partie de la famille des distributions exponentielles et peut être utilisée dans le cadre des GLM, avec :

- $\theta_i = \ln(\lambda_i)$
- $\mathrm{a}(\theta_i) = \lambda_i + \ln(1-e^{-\lambda_i})$
- $\phi = 1$
- $c(y,\phi) = \frac{1}{y!}$

1.3 Les modèles linéaires généralisés (GLM)

Nous présentons désormais les modèles linéaires généralisés (GLM). Ces modèles permettent d'étendre les modèles de régression linéaire classiques en élargissant certaines hypothèses.

Tout d'abord, alors que ces derniers considèrent que les erreurs sont distribuées normalement, les GLM permettent d'utiliser n'importe quelle distribution appartenant à la famille exponentielle. Dans notre cas, cela nous intéresse particulièrement. En effet, l'hypothèse de normalité n'est pas raisonnable lorsque nous modélisons des montants et des nombres de sinistres.

De plus, dans le modèle de régression linéaire la moyenne correspond au score, c'est-à-dire à la combinaison linéaire des facteurs explicatifs. Les GLM permettent en revanche de rendre cette relation plus flexible. Une fonction de lien monotone g est en effet introduite, avec pour rôle de lier le score et la moyenne :

$$\mu_i = g^{-1}(\text{score}_i) = g^{-1}(X^\top \beta) \tag{1.3.1}$$

En actuariat, la fonction logarithmique est souvent prise comme fonction de lien, ce que nous ferons également ici. L'intérêt de ceci réside dans l'interprétation de chaque coefficient estimé. En effet, la moyenne est exprimée à l'aide d'une structure multiplicative. L'effet relatif de chaque coefficient peut donc facilement être déterminé :

$$\mu_i = \exp(\text{score}_i) = e^{\beta_0} \prod_{j=1}^{p} e^{\beta_j x_{ij}} \tag{1.3.2}$$

avec p+1 paramètres de régression.

De plus, la fonction de lien logarithmique contraint la moyenne à être positive, ce qui est adapté aux estimations que nous souhaitons mener.

Une fois la distribution et la fonction de lien choisies, l'estimation des paramètres de régression se fait, en pratique, en utilisant la maximisation de la fonction de vraisemblance du modèle. La log-vraisemblance générale d'une fonction appartenant à la famille exponentielle est de la forme suivante :

$$l(y|\theta, \phi) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - a(\theta_i)}{\phi} \right) + \sum_{i=1}^{n} \ln(c(y_i, \phi)) \tag{1.3.3}$$

Afin de maximiser cette dernière, les p+1 équations à résoudre sont donc données par :

$$\frac{\partial l}{\partial \beta_j} = 0$$
$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i) g'(\mu_i)} = 0 \tag{1.3.4}$$

En général, il n'est pas possible de trouver des solutions fermées pour ces équations. Il est donc nécessaire d'appliquer des algorithmes d'approximation numérique. Dans ce mémoire, ceux-ci sont implémentés à l'aide des logiciels `R` et `SAS`.

1.4 Les copules

Formellement, une copule est une fonction de distribution multivariée dont les distributions marginales sont distribuées uniformément et qui permet de modéliser la structure de dépendance entre les variables. Grâce aux copules, il est ainsi possible de séparer la modélisation de la structure de la dépendance de celle des distributions marginales.
Étant donné que nous nous intéressons à la modélisation de deux variables aléatoires (X et Y), nous nous limitons ici à la présentation du cas bivarié.

Une copule bivariée est une fonction C : $[0,1]^2 \to [0,1]$ qui répond aux trois propriétés suivantes :

1. $\forall u_1, u_2 \in [0,1]$:
   $lim_{u_i \to 0} C(u_1, u_2) = 0$ avec $i = 1, 2$;
2. $\forall u_1, u_2 \in [0,1]$:
   $lim_{u_1 \to 1} C(u_1, u_2) = u_2$ et $lim_{u_2 \to 1} C(u_1, u_2) = u_1$
3. $\forall u_1, u_2, v_1, v_2 \in [0,1]$ et $u_1 \leq u_2, v_1 \leq v_2$:
   $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$

En utilisant le théorème de Sklar, nous savons qu'il existe une copule telle que :

$$F(x, y) = C\big(F_X(x), F_Y(y)\big) \tag{1.4.1}$$

La copule est donc bien une fonction qui donne une distribution jointe $F(x, y)$ ayant comme marginales F_X et F_Y. Sachant les marginales, il est possible de considérer plusieurs structures de dépendance selon le choix de la copule, pour décrire au mieux les données mises à disposition. Dans les cas où les distributions marginales de X et Y sont continues et strictement croissantes, alors la copule C est unique.

En pratique, lorsque nous observons plusieurs variables aléatoires, il est utile de pouvoir déterminer leur distribution jointe. Cette dernière est souvent inconnue et difficile à estimer puisqu'elle dépend à la fois des distributions marginales des variables aléatoires et de leur dépendance. En revanche, il est en général relativement facile d'estimer les distributions marginales de chacune d'entre elles.
Sachant cela, nous voyons tout l'intérêt d'utiliser les copules afin de calculer, à partir des distributions marginales, la distribution jointe prenant en compte la dépendance.

Il existe plusieurs familles de copules. Tout d'abord, nous retrouvons la classe des copules archimédiennes. Les copules appartenant à cette dernière sont définies selon la fonction de distribution exprimée à l'équation (1.4.2). Dans cette famille de copules, nous retrouvons notamment celles de Gumbel, Clayton et Frank.

$$C(u,v) = \phi^{-1}\big(\phi(u) + \phi(v)\big) \tag{1.4.2}$$

où $\phi(.)$ *est appelé le générateur de la copule archimédienne et est une fonction décroissante allant de* $[0,1]$ *à* $[0,\infty]$ *avec* $\phi(0) = \infty$.

Une autre famille de copules est celle des copules elliptiques. Nous retrouvons notamment dans cette famille la copule de Gauss.

Le tableau 1.4.1 reprend les fonctions de distribution des différentes familles de copules.

Famille	Copule $C_\theta(u,v)$	Bornes de θ
Gauss	$\Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v)\middle\vert\theta\right)$	]-1,1[
Gumbel	$\exp\Big(-\big((-\ln u)^\theta + (-\ln v)^\theta\big)^{\frac{1}{\theta}}\Big)$	]1,∞[
Clayton	$\big(u^{-\theta} + v^{-\theta} - 1\big)^{\frac{1}{\theta}}$	]0,∞[
Frank	$-\frac{1}{\theta}\ln\big(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1}{e^{-\theta}-1}\big)$	$\mathbb{R}\setminus\{0\}$

TABLE 1.4.1 – Définition des copules de Gauss, Gumbel, Clayton et Frank

Le coefficient de corrélation traditionnellement utilisé pour mesurer la dépendance a notamment comme désavantage que sa valeur dépend du choix des distributions marginales. Si une transformation non linéaire est appliquée sur les marginales, la valeur de ce dernier variera. Cette variation est problématique car il n'y a pas eu de changement de dépendance. De plus, le coefficient de corrélation se limite à représenter les dépendances linéaires entre les deux variables aléatoires.

Pour les deux raisons citées précédemment, nous avons choisi d'étudier la dépendance à l'aide du tau de Kendall (τ). Ce dernier est une mesure de la corrélation de rang entre deux variables. Cela signifie que le tau de Kendall mesure la différence entre la probabilité de concordance et celle de discordance entre deux variables aléatoires, soit :

$$\tau = P\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - P\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right] \tag{1.4.3}$$

Ainsi, les observations (X_1, Y_1) et (X_2, Y_2) sont dites concordantes si $(X_1 - X_2)(Y_1 - Y_2) > 0$ et discordantes si $(X_1 - X_2)(Y_1 - Y_2) < 0$. Les deux variables sont concordantes si la taille de leurs valeurs respectives sont simultanément grandes ou petites. Dans le cas inverse, nous parlons de discordance.

Pour les variables ayant une distribution marginale continue, le tau de Kendall peut être obtenu de la façon suivante :

$$\tau = 4\int_0^1\int_0^1 C(u,v)\, dC(u,v) - 1 \tag{1.4.4}$$

$C(u, v)$ étant la copule de la fonction de distribution bivariée des variables aléatoires X et Y, ceci confirme que le τ ne dépend pas des distributions marginales mais seulement de la structure de dépendance utilisée.
Nous résumons dans la table 1.4.2 la valeur du tau de Kendall pour chacune des familles de copule présentées dans cette partie.

Famille	τ_θ de Kendall	Bornes du τ
Gauss	$\frac{2}{\pi}\arcsin(\theta)$	$\tau \in \mathbb{R}$
Gumbel	$1 - \frac{1}{\theta}$	$\tau \in [0, \infty[$
Clayton	$\frac{\theta}{\theta+2}$	$\tau \in]0; \infty[$
Frank	$1 - \frac{4}{\theta}\big(1 - D_1(\theta)\big)$	$\tau \in \mathbb{R} \setminus \{0\}$

TABLE 1.4.2 – Valeurs du tau de Kendall

Avec la fonction $D_1(\theta)$, appelée fonction de Debye et définie comme : $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{exp(t)-1}dt$.

Chapitre 2

Modèles de régression

2.1 Présentation de la base de données

Pour la suite de cette recherche, sont brièvement présentées dans cette section les différentes variables contenues dans la base de données et avec lesquelles nous allons travailler.

La base de données utilisée provient du portefeuille responsabilité civile automobile d'un assureur français, fourni dans le package `R : CASDataset`. Celui-ci est composé de 677'991 polices. La table 2.1.1 donne une description des différentes variables contenues dans cette base de données.

Variable	Description
`ClaimNb`	Nombre de sinistres
`Exposure`	Durée du contrat en années
`VehPower`	Puissance véhicule : par ordre croissant
`VehAge`	Age du véhicule
`DrivAge`	Age du conducteur
`BonusMalus`	Système de Bonus et Malus entre 50 et 350 : <100 = bonus, >100 = malus
`VehBrand`	Marque de la voiture
`VehGas`	Diesel ou Essence
`Area`	Zone d'habitat du conducteur : de A pour rural à F pour centre urbain
`Density`	Nombre d'habitants par km^2 de la ville dans laquelle vit le conducteur
`Region`	Région française (découpage avant 2015)
`ClaimAmount`	Montant de chaque sinistre

TABLE 2.1.1 – Description des variables contenues dans la base de données

2.2 Sélection des distributions marginales

Avant de pouvoir appliquer les modèles, il est nécessaire de sélectionner une famille de distribution adaptée à la variable que nous souhaitons modéliser. Dans la littérature actuarielle, il est courant de retrouver la distribution de Gamma ou Gaussienne inverse pour les modèles de sévérité. En ce qui concerne les modèles du nombre de sinistres, la distribution de Poisson ou la distribution Binomiale sont souvent considérées.

Dans ce mémoire, nous faisons le choix d'utiliser la distribution de Gamma pour la sévérité et la distribution de Poisson pour la fréquence. Ces distributions font partie de la famille exponentielle dont la densité est définie à l'équation (1.2.1) ayant comme paramètres ceux retrouvés dans le tableau 1.2.1.

Nous devons cependant nous assurer que ces choix sont raisonnables pour les données que nous utilisons dans le cadre de cette recherche. Pour cela, nous avons d'abord jugé nécessaire d'opérer certains ajustements sur la base de données. Ces derniers, ainsi que la validation des distributions marginales choisies, sont présentés dans cette section.

2.2.1 Ajustements sur la base de données

A la table 2.1.1 présentée précédemment, nous remarquons que certaines variables habituellement utilisées pour ce type d'analyse ne sont pas disponibles. C'est le cas, par exemple, de la variable de genre qui est souvent une variable fortement discriminante. Notons que même s'il n'est désormais plus possible de segmenter selon le sexe pour la prime commerciale, il reste toujours intéressant de l'intégrer dans le tarif technique afin que celui-ci soit le plus complet possible et permette à l'assureur une bonne compréhension du risque. Par conséquent, avec un nombre aussi limité de variables, il ne sera sans doute pas possible de décrire de façon fiable les données.

Cependant, étant donné que l'objectif de cette recherche est de modéliser la dépendance entre le nombre de sinistres et leur sévérité et non pas de construire un tarif le plus précis possible, ceci ne pose pas réellement de problème. Le nombre de variables à disposition devrait être suffisant pour permettre d'extraire une dépendance.

Nous remarquons également que certains ajustements sur notre base de données sont nécessaires. La table 2.2.1 reprend les différentes modifications réalisées sur les données.
Tout d'abord, nous décidons de retirer les sinistres supérieurs à 20'000 €. En effet, les grands sinistres avec des valeurs extrêmes étant rares et pouvant être expliqués en partie par la "malchance", les prendre en compte pourrait fausser notre analyse. De plus, nous utilisons une distribution Gamma pour la sévérité, or celle-ci ne permet pas de représenter correctement les valeurs extrêmes. C'est pourquoi nous nous concentrons donc sur les petits sinistres plafonnés. Nous ne traiterons pas de manière séparée ces sinistres extrêmes. Ceci

aura un impact sur l'estimation de la perte mais nous permettra toujours d'analyser la dépendance qui nous intéresse.
Ensuite, certains montants exacts reviennent fréquemment. Ce type d'observation pourrait s'expliquer par le fait qu'au moment d'un sinistre, la compagnie d'assurance provisionne un montant forfaitaire pour celui-ci tant qu'elle ne connaît pas sa valeur finale. S'agissant d'un portefeuille en responsabilité civile, les sinistres peuvent se développer sur plusieurs années avant d'atteindre leur valeur finale. Prendre en compte ces provisions forfaitaires fausserait également notre analyse, ces montants ne représentant pas la réelle valeur du sinistre. Qui plus est, avoir une masse d'observations sur certaines valeurs pourrait mener à des problèmes de calcul par la suite. Par conséquent, nous retirons de notre étude les sinistres avec les montants concernés.

Valeur retirée	**Raisons**	**Nb. d'observations**
$>$ 20 000	Trop grande valeur : max. 4 Mio	218
1 204	Montant forfaitaire	4 792
1 128.12	Montant forfaitaire	3 056
1 172	Montant forfaitaire	2 072
1 128	Montant forfaitaire	832

TABLE 2.2.1 – Ajustements effectués sur les données

En outre, certaines polices semblent posséder des valeurs aberrantes. En effet, nous pouvons observer, dans la base de données, certains assurés ayant réalisé un grand nombre de sinistres sur une très courte période d'exposition. Suite à cela, nous retirons 4 polices concernées par ce type de problème.
Nous remarquons que certaines polices à sinistres multiples n'ont l'information sur la sévérité que pour certains sinistres, notamment suite aux ajustements effectués. Par conséquent, nous supposons dans le cadre de notre recherche que les sévérités non observées suivent le même comportement que celles qui sont disponibles.
Enfin, pour les polices ayant plus de 4 sinistres dans le portefeuille, nous fixons le nombre de sinistres à 4. En effet, il y a trop peu d'observations pour ces nombres de sinistres, ce qui entraînerait des estimations peu fiables lors de la calibration de nos modèles.

Finalement, nous avons à notre disposition 659'385 observations dont la majorité n'a réalisé aucun sinistre et 15'432 avec des sinistres. La figure 2.2.1 représente la distribution des montants des sinistres suite aux ajustements effectués sur les données.

Nous retirons également la variable `Density` de notre analyse. En effet, la variable `Area` mesure également la densité de population mais sous forme de catégories. La table 2.2.2 représente les valeurs de la variable `Density` en fonction de la variable `Area`.

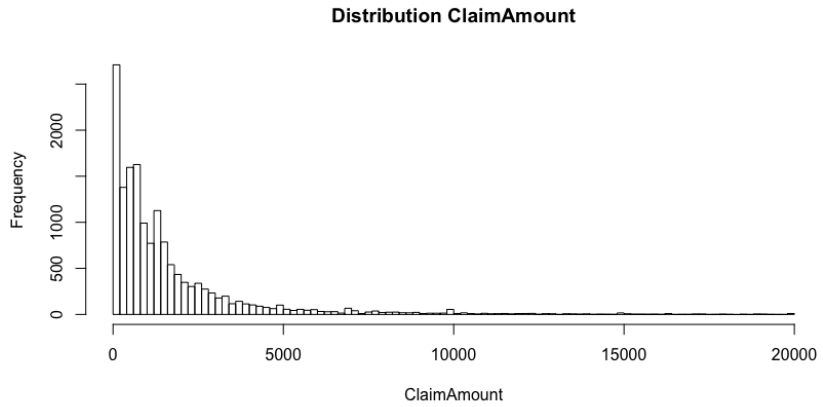


FIGURE 2.2.1 – Distribution des montants des sinistres

Ceci montre bien que les deux variables sont parfaitement liées. Par conséquent, nous avons décidé de ne travailler qu'avec la variable `Area`, afin d'éviter la redondance de l'information.

`Area`	A	B	C	D	E	F
`Density`	[1-50]	[51-100]	[101-500]	[501-1993]	[2001-9850]	[10'008-27'000]

TABLE 2.2.2 – Valeurs de `Density` en fonction de `Area`

2.2.2 Validation des distributions marginales

Premièrement, pour le modèle de sévérité, nous réalisons un modèle simple, sans prendre en compte une quelconque dépendance. Pour cela, nous effectuons la commande suivante sur le logiciel `R` :

```
glm_sev <- glm(ClaimAmount ~ Area_modif_amount+[..],data=
   table_mergee,family=Gamma(link="log"))
```

Sur la figure 2.2.2, nous représentons d'abord les résidus de déviance de ce modèle en fonction des valeurs prédites. Nous pouvons donc constater que l'hypothèse d'une distribution de Gamma pour modéliser la sévérité des sinistres semble adaptée à nos données. En effet, nous n'observons pas de tendance générale dans la distribution des résidus.
Nous pouvons néanmoins observer une concentration de points de valeur tournant autour de -2 et qui semblent former un segment. Cela provient probablement du fait qu'il y a une sur-représentation de très petites valeurs de sinistres dans les données mises à notre disposition.

Ensuite, sur la même figure, est représenté le QQ-plot des résidus. Ces derniers semblent suivre une distribution Gaussienne mise à part dans les queues, où les points sont plutôt éloignés de la droite. De plus, la distribution représentée à la figure 2.2.1 ne semble pas incompatible avec la loi Gamma.

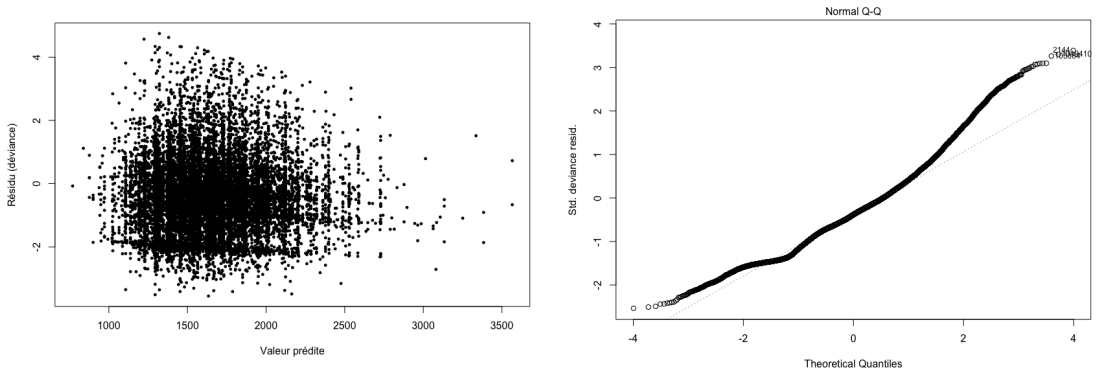


FIGURE 2.2.2 – Résidus de déviance et QQ-plot du modèle de sévérité

Deuxièmement, nous vérifions que l'hypothèse d'une distribution de Poisson est également raisonnable pour le modèle de fréquence. Comme nous considérons directement les nombres de sinistres strictement positifs, nous utilisons la distribution ZTP. Sur `R`, nous avons :

```
glm_freq <- ztp.glm(table_mergee$ClaimNb,S,exposure = exposure, sd.
    error = TRUE)
```

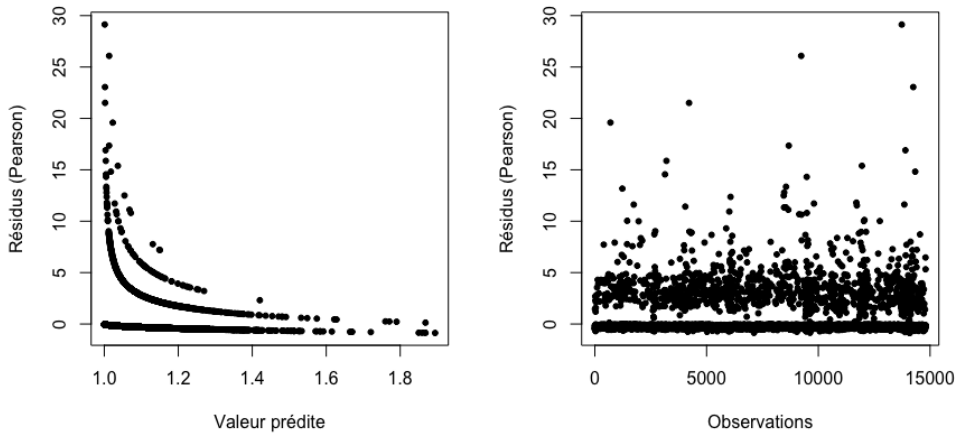


FIGURE 2.2.3 – Résidus de Pearson du modèle de fréquence

Sur la figure 2.2.3, sont représentés les résidus de Pearson de ce modèle. Sur le premier graphique, les courbes observées sont caractéristiques des variables de comptage avec des niveaux bien définis. Nous n'observons pas de tendance générale dans la distribution des résidus. Sur le deuxième graphique, nous pouvons observer que globalement, les résidus prennent peu de valeurs aberrantes.

Suite à cette analyse, nous pouvons conclure que les hypothèses d'une distribution de Gamma pour la sévérité et une distribution de Poisson pour la fréquence des sinistres sont adaptées aux données mises à notre disposition.

2.3 Modèle classique en deux parties

Dans cette partie, nous présentons le modèle en deux parties. Il s'agit d'une des méthodes la plus couramment utilisée en pratique sous l'hypothèse d'indépendance entre la sévérité et la fréquence.

Le principe de base de cette approche est de modéliser séparément la sévérité et la fréquence des sinistres afin d'obtenir leur moyenne respective. Grâce à l'hypothèse d'indépendance, il est ensuite possible d'écrire la prime pure comme le produit de ces 2 moyennes. La perte totale d'une police peut donc s'écrire comme :

$$S = \sum_{k=1}^{N} X_k$$

où N est une variable aléatoire représentant le nombre de sinistres et X_k une variable représentant leur montant respectif.

Nous pourrions envisager de modéliser directement la perte totale de la police. Néanmoins, cette approche a plusieurs limites.
Tout d'abord, il ne sera pas possible d'étudier l'impact d'options du contrat sur la perte finale, comme par exemple les franchises ou le système de Bonus-Malus.
De plus, une telle modélisation suppose que les mêmes facteurs influencent à la fois la sévérité et la fréquence et ceci, dans la même direction. Néanmoins, dans la réalité, il est courant que certains facteurs n'aient une influence que sur une seule des composantes de la perte ou alors qu'ils aient une influence opposée sur les deux variables d'intérêt. Comme exemple, nous pouvons prendre le facteur rural-urbain. En assurance automobile dégâts matériels, les assurés vivant en ville ont en moyenne plus d'accidents que ceux vivant à la campagne, cependant, ceux-ci restent aussi en moyenne moins chers. Cet exemple illustre bien le fait que le signe de l'impact d'un certain facteur peut être différent pour les deux grandeurs.

Ces deux limitations motivent donc l'utilisation d'un modèle en deux parties. Il est nécessaire de choisir une distribution marginale adaptée pour chacune des parties. Pour la fréquence, les distributions de Poisson et Binomiale sont régulièrement utilisées. Pour la sévérité, ce sont les distributions de Gamma et Inverse Gaussienne.

Nous appliquons par la suite ce modèle à deux parties sur nos données et l'appelons le "Modèle indépendant en deux parties" dans le cadre de notre recherche. Afin de modéliser la fréquence et la sévérité, nous utilisons un GLM pour chacun des deux modèles marginaux. Ces derniers utilisent la fonction de lien logarithmique.

Nous supposons que le nombre de sinistres pour une police i suit une loi de Poisson : $Y_i \sim \text{Poi}(\lambda_i)$. La densité de celle-ci a la forme suivante :

$$f_{Y_i}(y) = \frac{e^{-\lambda_i}\lambda_i^y}{y!},\ y \in \mathbb{N} \tag{2.3.1}$$

Comme nous ne prenons en compte ici que les polices avec un nombre de sinistres strictement positif, le nombre considéré suit une loi de Poisson tronquée en 0 ($Y_i \sim ZTP(\lambda_i)$). Cela revient à dire que Y_i suit une loi de Poisson conditionnellement au fait que la valeur du nombre de sinistres soit strictement positive. De plus, nous supposons que la distribution du nombre de sinistres ne dépend pas de leur valeur. Ainsi, la loi conditionnelle de Y_i peut s'écrire de la manière suivante :

$$f_{Y_i|Y_i\geq 1}(y) = \frac{f_{Y_i}(y)}{1 - f_{Y_i}(0)} = \frac{\lambda_i^y e^{-\lambda_i}}{y!(1-e^{-\lambda_i})} = \frac{\lambda_i^y}{(e^{\lambda_i}-1)y!} \tag{2.3.2}$$

Comme indiqué à l'équation (1.2.2), cette loi conditionnelle peut s'écrire sous la forme d'une distribution appartenant à la famille exponentielle. La moyenne de cette distribution peut être exprimée en fonction du paramètre de la Poisson (λ_i) :

$$\mu_{\{Y_i|Y_i\geq 1\}} = \frac{\lambda_i}{1-e^{-\lambda_i}} \tag{2.3.3}$$

Et sa variance est déterminée par la formule suivante :

$$\text{Var}[\text{Y}_\text{i}|\text{Y}_\text{i} \geq 1] = \mu_{\{Y_i|Y_i\geq 1\}}\left[1 + \lambda_i - \mu_{\{Y_i|Y_i\geq 1\}}\right] \tag{2.3.4}$$

Concernant la sévérité, nous supposons que celle-ci suit une loi Gamma : $X_{ik} \sim \text{Gamma}(\mu_{ik}, a)$. La densité de la sévérité individuelle X peut donc s'écrire :

$$f_{X_{ik}}(x) = \exp\left(a\left(-\frac{x}{\mu_{X_{ik}}} - ln\,\mu_{X_{ik}}\right)\right)\frac{a^a}{\Gamma(a)}x^{a-1},\ x \in \mathbb{R} \tag{2.3.5}$$

La moyenne de la sévérité est donnée par :

$$\text{E}[\text{X}_\text{ik}] = \mu_{X_{ik}} \tag{2.3.6}$$

Et sa variance par :

$$\text{Var}[\text{X}_\text{ik}] = \frac{\mu_{X_{ik}}^2}{a} \tag{2.3.7}$$

Il est courant de travailler avec la sévérité moyenne, les données individuelles n'étant pas toujours disponibles. Les propriétés de la Gamma assurent que ce choix n'ait pas d'impact sur l'estimation des paramètres. Néanmoins, l'estimation à l'aide des données agrégées n'est pas toujours optimale. Tout d'abord, l'utilisation de données agrégées diminue la précision de l'estimation du paramètre de dispersion de la Gamma. De plus, comme relevé

dans l'étude de W. Lee et al. 2019, dans le cadre de la modélisation de la dépendance, l'utilisation de données moyennes peut révéler l'existence d'une dépendance y compris dans les cas où la fréquence et la sévérité individuelle sont indépendantes. Le modèle suppose en effet une dépendance structurelle entre celles-ci, en faisant dépendre de la fréquence la variance de la sévérité moyenne. Par conséquent, il est préférable d'utiliser les données individuelles lorsque cela est possible.

Nous avons à notre disposition les données sur les montants de chaque sinistre. Pour cette raison, et étant donné que le paramètre de dispersion est utilisé plus loin lors des simulations, nous choisissons de travailler avec les données individuelles dans le cadre de cette recherche.

2.4 Modèles dépendants

2.4.1 Modèle dépendant en deux parties

Afin de développer ce modèle, nous nous basons sur la méthode présentée par Garrido et al. 2016. Dans leur recherche les auteurs modélisent un montant moyen de sinistres. Nous utilisons, quant à nous, le montant individuel de chaque sinistre étant donné que ceux-ci sont mis à notre disposition dans la base de données. Comme expliqué précédemment, cela nous permet d'éviter une perte d'information due à l'agrégation des données, notamment sur la dispersion des valeurs de sinistre.

Dans notre recherche, nous commençons d'abord par suivre la même méthode puis, dans un deuxième temps, nous utilisons la fréquence annuelle comme variable explicative. Ce choix a été fait afin d'améliorer la fiabilité de l'estimation du paramètre de dépendance. En effet, nous pouvons nous interroger sur la fiabilité d'une estimation effectuée à l'aide d'un petit nombre d'observations pour chaque valeur de nombre de sinistres. Cette approche a également l'avantage de prendre en compte la durée d'exposition du contrat lors de la modélisation de la dépendance et donc de modéliser la dépendance entre la fréquence et la sévérité.

Chaque modèle marginal, pour la fréquence et la sévérité, est modélisé à l'aide d'un GLM. Comme pour le modèle indépendant, la fréquence, Y, suit une loi ZTP dont la densité est donnée par l'équation 2.3.2.
Pour le modèle de la sévérité, nous supposons que X suit une loi Gamma dont la densité est donnée à l'équation 2.3.5. De plus, nous continuons de supposer que, conditionnellement aux caractéristiques des assurés et au nombre de sinistres, les montants des sinistres sont indépendants et identiquement distribués. Afin de prendre en compte la dépendance, nous permettons à la distribution de la sévérité individuelle de dépendre du nombre de sinistres.

Nous pouvons désormais exprimer dans le modèle GLM les moyennes de la sévérité et de la fréquence conditionnellement à la valeur des variables explicatives. Pour la sévérité nous avons pour chaque sinistre :

$$\mu^{\nu}_{X_{ik}} = e^{z_i\beta+\nu Y_i} = \mu_{X_{ik}} e^{\nu Y_i} \tag{2.4.1}$$

Pour la fréquence, le modèle marginal permettant de modéliser le paramètre de la distribution de Poisson sous-jacente est :

$$\lambda_i = e_i e^{w_i\alpha} \tag{2.4.2}$$

où e_i représente l'exposition en nombre d'années du contrat..

Afin de calibrer ce modèle nous procédons par maximum de vraisemblance. Les deux modèles marginaux étant indépendants, la fonction à maximiser est donnée par l'équation (2.4.3).

$$l(\alpha, \beta, \nu) = \sum_{i=1}^{m} \ln\left(f_{Y|Y\geq 1}(y_i|\alpha)\right) + \sum_{i=1}^{m}\sum_{k=1}^{Y_i} \ln\left(f_X(x_{ik}|\beta, \nu)\right) \tag{2.4.3}$$

L'estimation des paramètres pour le modèle de fréquence est identique à celle du modèle indépendant en deux parties. En effet, la fonction de vraisemblance pour ce modèle n'est pas influencée par les nouvelles transformations que nous avons effectuées sur le modèle de la sévérité.

Dans une deuxième partie, nous transformons légèrement ce modèle en permettant à la sévérité conditionnelle de dépendre de la fréquence annuelle. Comme indiqué précédemment, ceci nous permet entre autres de prendre en compte la durée d'exposition de chaque police dans la modélisation de la dépendance, ce qui n'est pas le cas dans le premier modèle présenté. Si nous prenons deux polices avec un sinistre, l'une observée sur une année complète et l'autre sur un mois, celles-ci seront considérées de la même manière lors de la modélisation de la dépendance.

Nous remplaçons d'abord dans l'équation (2.4.1) le nombre de sinistres observé par la fréquence annuelle. Parallèlement, nous réalisons un modèle où la fréquence annuelle est catégorisée en trois classes, au lieu de considérer uniquement la fréquence annuelle brute. Dans ce cas, le modèle marginal pour la sévérité est donné par :

$$\mu^{\nu}_{X_{ik}} = e^{z_i\beta+\nu_1 f_{i1}+\nu_2 f_{i2}} = \mu_{X_{ik}} e^{\nu_1 f_{i1}+\nu_2 f_{i2}} \tag{2.4.4}$$

La variable f_{ij} représente la fréquence annuelle observée classée selon trois niveaux. Respectivement, nous avons une fréquence annuelle faible ($j = 0$), moyenne ($j = 1$) et élevée ($j = 2$).

2.4.2 Modèle de copules

Afin de développer ce modèle, nous nous basons sur la méthode présentée par KRÄMER et al. 2013. Celle-ci utilise les copules afin de lier les modèles marginaux. L'intérêt de leur recherche est que celle-ci donne un exemple d'application des copules dans un contexte de régression. Les modèles marginaux sont en effet construits à l'aide de modèles GLM.
De plus, dans le cadre de leur recherche, les auteurs ont développé un package `R` adapté à l'utilisation que nous souhaitons en faire, c'est-à-dire lier sévérité et fréquence. Nous utilisons donc ce package `CopulaRegression` afin d'appliquer ce modèle sur nos données. Nous testons les copules de Gauss, Gumbel, Clayton et Frank. Le package permet de lier une loi marginale continue et une loi marginale discrète, données par des modèles de régression.

Dans le cadre de ce modèle, et ce afin de limiter le temps de calcul nécessaire pour la calibration des modèles et l'interprétation des résultats, nous nous limitons au cas des copules bivariées. Cela signifie que nous utilisons le montant moyen des sinistres pour chaque police et non plus le montant individuel de chaque sinistre.
Pour la partie fréquence du modèle, cela n'a pas d'impact. Le nombre de sinistre suit toujours une distribution ZTP dont la densité $f_{Y_i}(y)$ est donnée par l'équation (2.3.2). En revanche, pour la sévérité, comme nous utilisons des données moyennes, nous devons pondérer la densité afin de tenir compte du nombre de sinistres. De ce fait, nous avons $X_i \sim \text{Gamma}(\mu_i, Y_i a)$ et la densité $f_{X_i}(x)$ de la Gamma est donnée par l'équation (2.3.5). Cette perte d'information, due à l'utilisation de données agrégées, ainsi que le temps de calcul restant relativement important sont déjà deux inconvénients que nous pouvons relever pour ce modèle.

En utilisant la copule, la distribution jointe entre X et Y est donnée par :

$$F_{X_i,Y_i|\theta}(x,y) = C\big(F_{X_i}(x), F_{Y_i}(y)|\theta\big) \tag{2.4.5}$$

avec θ, le paramètre de la copule en fonction de la famille choisie.
Nous définissons désormais la densité jointe f_{X_i,Y_i} d'une variable continue et d'une variable discrète en fonction de la copule. Pour cela, nous posons tout d'abord $D_1(F_{X_i}(x), F_{Y_i}(y)|\theta)$ comme étant la dérivée partielle de la fonction de distribution jointe par rapport à la valeur de la fonction de distribution de la sévérité. Il est possible de trouver la forme de cette dérivée partielle pour chacune des 4 familles de copules considérées.

$$D_1(F_{X_i}(x), F_{Y_i}(y)|\theta) = \frac{\partial}{\partial F_{X_i}(x)} C(F_{X_i}(x), F_{Y_i}(y)|\theta) \tag{2.4.6}$$

Afin de pouvoir trouver une expression pour la densité jointe, nous avons également besoin de définir la fonction de masse de probabilité du nombre de sinistres conditionnellement à la sévérité.

$$P[Y_i = y|X_i = x, a, \mu_{X_i}, \lambda_i, \theta] = D_1\Big(F_{X_i}(x|a, \mu_{X_i}), F_{Y_i}(y|\lambda_i)|\theta\Big) - D_1\Big(F_{X_i}(x|a, \mu_{X_i}), F_{Y_i}(y-1|\lambda_i)|\theta\Big) \tag{2.4.7}$$

Par définition, la densité jointe entre une variable continue (Y) et une variable discrète (X) est donnée par :

$$\begin{aligned} f_{X_i,Y_i}(x, y|a, \mu_{X_i}, \lambda_i, \theta) &= \frac{\partial}{\partial x} P[X_i \leq x, Y_i = y] \\ &= \frac{\partial}{\partial x} P[X_i \leq x, Y_i \leq y] - \frac{\partial}{\partial x} P[X_i \leq x, Y_i \leq y-1] \\ &= \frac{\partial}{\partial x}\Big(C(F_{X_i}(x|a, \mu_{X_i}), F_{Y_i}(y|\lambda_i)) - C(F_{X_i}(x|a, \mu_{X_i}), F_{Y_i}(y-1|\lambda_i))\Big) \end{aligned}$$

En utilisant les expressions données aux équations (2.4.5), (2.4.6) et (2.4.7), la densité jointe s'obtient comme ceci :

$$f_{X_i}(x|a, \mu_{X_i})\Big(D_1(F_{X_i}(x|a, \mu_{X_i}), F_{Y_i}(y|\lambda_i)|\theta) - D_1(F_{X_i}(x|a, \mu_{X_i}), F_{Y_i}(y-1|\lambda_i)|\theta)\Big) \tag{2.4.8}$$

Nous pouvons remarquer que cette dernière dépend à la fois des paramètres des distributions marginales $(a, \mu_{X_i}, \lambda_i)$ déterminés par les modèles GLM et du paramètre de la copule (θ). Nous utiliserons cette densité afin de maximiser la log-vraisemblance de ce modèle et d'estimer ses paramètres.

Une fois les différentes copules calibrées sur nos données, nous devons choisir laquelle correspond le mieux. A cette fin, nous présentons pour finir le test de Vuong. Ce dernier nous permet de tester deux à deux quel modèle de copule est le plus proche des données. Pour se faire, ce test se base sur l'écart de vraisemblance entre les deux familles de copules considérées. Nous définissons tout d'abord l'écart de log-vraisemblance entre les deux modèles de copules comme :

$$m_i = l_i^{(1)} - l_i^{(2)}, \qquad i = 1, .., n \tag{2.4.9}$$

avec $l^{(1)}, l^{(2)}$ les log-vraisemblances respectives pour les deux familles de copule considérées lors de la comparaison.

Nous définissons également l'écart moyen de log-vraisemblance pour l'ensemble du portefeuille, $\bar{m}$:

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i \tag{2.4.10}$$

En fonction de ces grandeurs, la statistique de test T_V est définie comme :

$$T_V = \frac{\sqrt{n}\bar{m}}{\sqrt{\sum_{i=1}^{n}(m_i - \bar{m})^2}} \to N(0,1) \tag{2.4.11}$$

Il est possible de conclure qu'à un niveau α donné, la famille de copule 1 est plus adapté que la famille 2 lorsque :

$$T_V > \phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{2.4.12}$$

où ϕ représente la fonction de distribution normale standardisée.

2.4.3 Modèle à effets aléatoires

Afin de développer ce modèle, nous nous basons sur les travaux de Baumgartner et al. 2015. Nous allons donc appliquer un modèle mixte sur nos données, incluant à la fois des effets fixes et des effets aléatoires, variant selon chaque individu.

Les effets aléatoires sont habituellement utilisés afin de modéliser des caractéristiques individuelles non observées qui influencent la variable de réponse et que le modèle n'est pas capable de capturer correctement. Ils permettent donc de prendre en compte l'hétérogénéité résiduelle de chaque classe de risque. De nombreux modèles utilisant des données répétées dans le temps ont recours à ces effets aléatoires. En effet, ceux-ci permettent d'introduire une dépendance entre les différentes observations d'un même individu.

Le modèle présenté par Baumgartner et al. 2015 utilise ces effets afin d'introduire une dépendance entre la fréquence et la sévérité des sinistres. La dépendance introduite refléterait donc l'existence de facteurs qui influencent les deux variables de réponse simultanément. L'hypothèse est que la dépendance n'est pas directe, mais n'est qu'apparente à cause de ces effets cachés. Conditionnellement à ces derniers, la fréquence et la sévérité des sinistres sont donc supposées indépendantes.

Pour ce modèle, un même effet aléatoire (ψ_i) est introduit dans le modèle marginal de la fréquence et celui de la sévérité. De plus, afin de faire le lien entre ces deux modèles marginaux, un paramètre d'échelle (s) est utilisé. C'est en analysant le signe de ce paramètre ainsi que sa taille que nous pourrons tirer des conclusions quant à la dépendance entre les deux variables d'intérêt.
Nous pouvons donc exprimer les modèles marginaux de nos deux variables de réponse. Pour la sévérité nous avons :

$$\mu_{ik} = e^{\beta z_i + \psi_i}$$

Le modèle pour le nombre de sinistres est :

$$\lambda_i = e_i e^{\alpha w_i + s\psi_i}$$

où e_i représente l'exposition en nombre d'années du contrat.
Nous supposons que les effets aléatoires suivent une loi normale de moyenne nulle et dont le paramètre de variance est à estimer :

$$\psi_i \sim N(0, \sigma_\psi)$$

Afin d'estimer ce modèle, il est utile de définir sa fonction de vraisemblance. Celle-ci fait l'hypothèse que, conditionnellement aux effets aléatoires (ψ_i), la sévérité et la fréquence sont indépendantes.

$$L(y, x; \beta, \alpha, s, \sigma_\psi) = \prod_{i=1}^{m} \left[\int_{-\infty}^{\infty} f_{Y_i}(y; \alpha, s, \psi_i, w_i) f_\psi(\psi_i; \sigma_\psi) \prod_{k=1}^{y_i} f_{X_{ik}}(x_{ik}; \beta, \psi_i, z_i)\, d\psi_i \right] \tag{2.4.13}$$

Bien que le paramètre de mise à l'échelle (s) permette aux moyennes conditionnelles de la sévérité et de la fréquence de dépendre différemment de l'effet aléatoire (ψ_i), ce modèle ne permet néanmoins que de décrire une dépendance linéaire. Celle-ci n'est en effet représentée que par un terme de mise à l'échelle constant. Dans un deuxième temps, nous proposons une extension de ce modèle afin de répondre à cette limitation.

Dans leur recherche, Jeong et al. 2017 proposent d'utiliser des effets aléatoires suivant une loi normale multivariée. Dans notre contexte, cela revient à utiliser des effets aléatoires distincts pour la fréquence et la sévérité mais qui suivent tous les deux la même loi Normale bivariée de paramètre de corrélation (ρ).

Le modèle marginal de la sévérité reste donc :

$$\mu_{ik} = e^{\beta z_i + \psi_i}$$

Celui pour le nombre de sinistres devient quant à lui :

$$\lambda_i = e_i e^{\alpha w_i + \xi_i}$$

avec $[\psi_i, \xi_i] \sim N\left([0,0], \begin{bmatrix} \sigma_\psi & \rho \\ \rho & \sigma_\xi \end{bmatrix}\right) = N(0, \Sigma)$.

Dans ce modèle, il conviendra donc d'analyser la valeur du ρ afin de pouvoir donner des conclusions à propos de la dépendance.

A nouveau, nous pouvons écrire la fonction de vraisemblance de ce modèle en conservant l'hypothèse d'indépendance conditionnelle :

$$L(y, x; \beta, \alpha, \Sigma) = \prod_{i=1}^{m} \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_i}(y; \alpha, \psi_i, w_i) f_{\psi,\xi}(\psi_i, \xi_i; \Sigma) \prod_{k=1}^{y_i} f_{X_{ik}}(x_{ik}; \beta, \psi_i, z_i)\, d\psi_i\, d\xi_i \right] \tag{2.4.14}$$

Nous verrons au moment de l'application numérique, qu'à cause de la présence des intégrales, il n'est pas trivial de maximiser ces fonctions de vraisemblance. Celles-ci nécessitent donc l'application d'algorithmes d'approximation des intégrales.

Chapitre 3

Application numérique

3.1 Analyse exploratoire

3.1.1 Statistiques descriptives

Avant de commencer à appliquer les modèles, une première analyse de statistiques descriptives est réalisée. Celle-ci a pour but de se donner une meilleure idée de la composition de la base de données et des premières tendances. La description de la base de données ainsi que les différents ajustements effectués sur celle-ci peuvent être retrouvés au chapitre précédent, aux sections 3.1 et 3.2.1. Pour chacune des polices, nous avons une information sur le nombre de sinistres, le montant de chaque sinistre réalisé, la durée d'exposition ainsi qu'un ensemble de variables explicatives.

Sur la figure 3.1.2, les deux premiers histogrammes représentent, respectivement, la distribution de l'exposition au risque mesurée en mois et celle du nombre de sinistres lorsqu'il y a au moins un sinistre. Nous remarquons qu'un grand nombre de polices du portefeuille ont été observées sur une période d'un an (30.79%). Néanmoins, environ deux tiers (68.85%) des assurés sont restés dans le portefeuille sur des périodes inférieures à un an. Ce grand nombre d'observations en-dessous d'un an peut s'expliquer, par exemple, par la méthode de gestion de la compagnie d'assurance. Il est possible que celle-ci encode de nouveaux dossiers à chaque fois qu'un changement dans la situation de l'assuré est effectué. Par exemple, lorsque ce dernier achète un nouveau véhicule ou suite à un changement d'adresse.

Concernant le nombre de sinistres, la table 3.1.1 reprend les différentes informations sur la fréquence relative du nombre de sinistres. Nous observons que 92.34% des assurés ayant réalisé au moins un sinistre n'ont causé qu'un seul sinistre. Ceci n'est pas vraiment surprenant, l'assurance automobile étant une branche à fréquence relativement faible. Plus précisément, dans le portefeuille considéré, la fréquence annuelle moyenne des sinistres ($\frac{\sum NbClaims}{\sum Exposure}$) est de 0.05 pour une police d'assurance et vaut 1.08 lorsque l'assuré a réalisé au moins un sinistre.

	1	2	3	4
Part d'observations	92.338 %	7.196 %	0.405 %	0.061 %
Nbre d'observations	13'667	1'065	60	9

TABLE 3.1.1 – Distribution du nombre de sinistres

Le tableau 3.1.2 indique la moyenne des montants pour chaque nombre de sinistres. Ce tableau permet de dégager une première tendance positive entre les deux variables. Sur l'ensemble du portefeuille, le montant moyen par sinistre augmente avec le nombre de sinistres réalisés. Cette tendance se confirme si nous nous intéressons au graphique 3.1.1. Celui-ci représente la sévérité moyenne ainsi que son intervalle de confiance à 95% en fonction de la fréquence. Nous avons également ajouté une droite de régression linéaire entre les deux grandeurs, dont l'équation est donnée par : $y = 1322.35 + 285.76x$
La tendance positive se confirme ici, à la fois graphiquement, et par la pente positive de la droite de régression. Remarquons tout de même que l'intervalle de confiance s'agrandit avec le nombre de sinistres. Cela signifie que l'estimation des moyennes est de moins en moins fiable, étant donné le peu d'observations disponibles pour les polices multi-sinistrées.

De plus, nous pouvons calculer le coefficient de corrélation entre nos deux variables d'intérêt. Celui-ci a une valeur de 0.04580, ce qui correspond à une légère dépendance positive.
Cependant, le coefficient de corrélation ne nous apporte qu'un premier aperçu sur la relation entre ces deux grandeurs. En effet, ce dernier ne donne uniquement de l'information que sur l'existence d'une relation linéaire entre les deux variables. Il peut exister une dépendance non-linéaire entre celles-ci, qui ne peut être détectée par le coefficient de corrélation.

Nous ne pouvons tirer de conclusions de cette dépendance observée ici. Cette dernière peut provenir du fait que l'on ignore les variables explicatives qui ont un effet sur la sévérité et la fréquence des sinistres. C'est pour cela qu'il est utile d'introduire des modèles de régression afin de pouvoir isoler la dépendance réelle entre les deux variables d'intérêt des effets des autres variables explicatives.

	1	2	3	4
Moyenne des montants de sinistres	1605.938	1929.680	1977.644	2411.712

TABLE 3.1.2 – Moyenne des montants de sinistres en fonction du nombre de sinistres

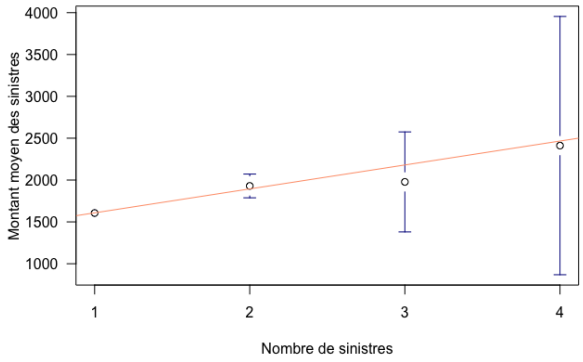
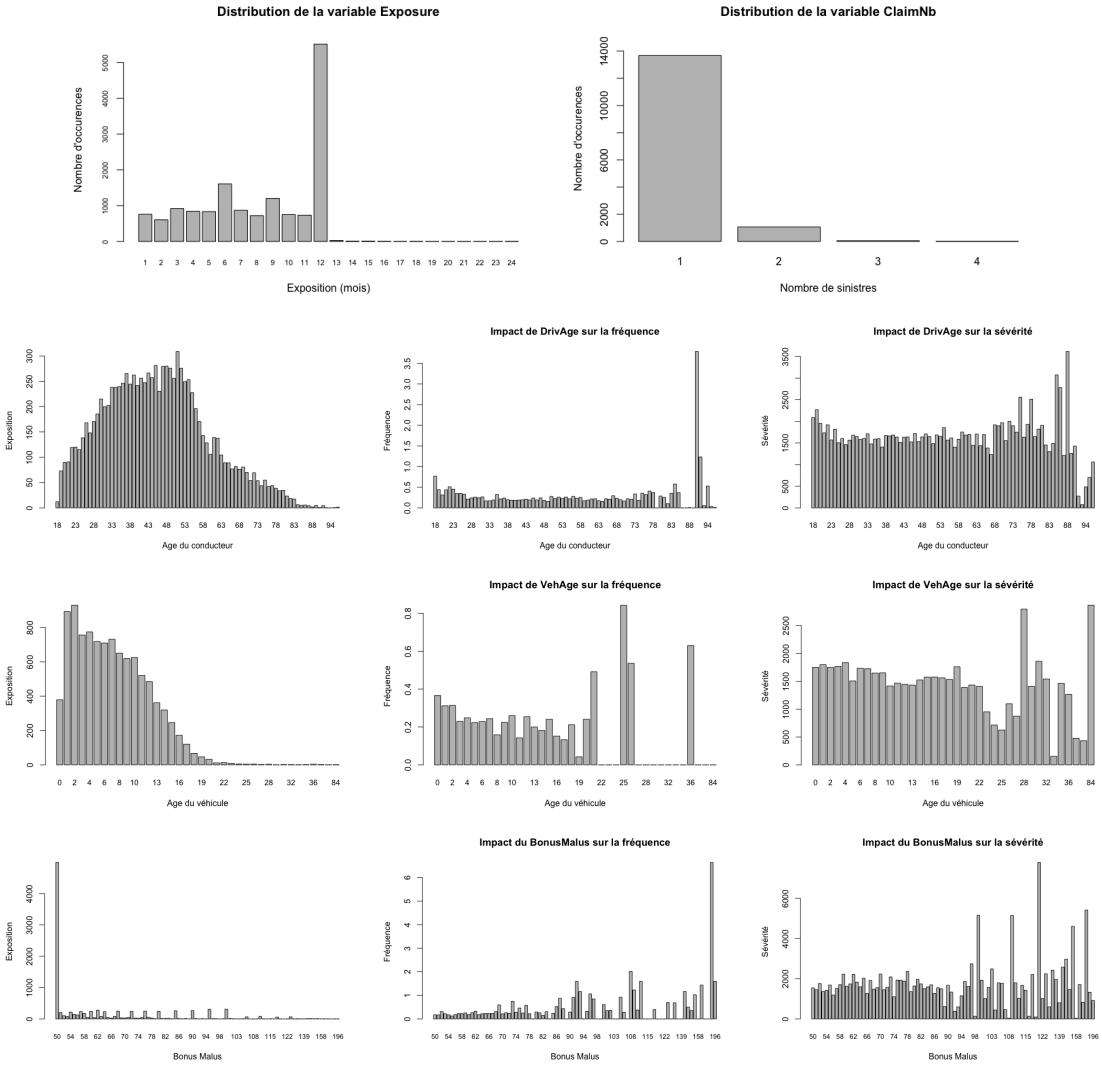


FIGURE 3.1.1 – Évolution du montant moyen des sinistres en fonction de leur nombre

Les autres graphiques de la figure 3.1.2, nous permettent d'effectuer une analyse univariée des différentes variables explicatives contenues dans la base de données via le GLM, de distribution Poisson pour le nombre de sinistres et Gamma pour la sévérité. Nous observons d'abord que pour la marque de voiture (`VehBrand`), les voitures de marque B04 semblent plus risquées que les autres. Ensuite, pour la variable `Area`, nous remarquons que plus nous nous rapprochons de la ville (F), plus la fréquence de sinistres est grande, c'est-à-dire que les habitants de la ville ont tendance à réaliser plus de sinistres. Cependant, ceux-ci ont l'air de coûter moins cher.

Il est important de mentionner que ces premiers résultats permettent uniquement de regarder l'effet de chaque variable explicative séparément sur les deux variables d'intérêt et ne peuvent permettre de tirer des conclusions définitives. En effet, nous ne prenons pas en compte les différentes interactions qu'il pourrait y avoir entre certaines variables. De plus, comme nous pouvons le remarquer sur les graphiques de gauche, certaines observations ne sont que faiblement exposées au risque. Par exemple, pour la puissance des véhicules (`VehPower`), les plus puissantes semblent être plus risquées que les moins puissantes. Cependant, ces profils ne représentent qu'une minorité du portefeuille. Il en va de même pour la variable `Area`, pour laquelle les assurés vivant dans la zone F (ville) sont restés sur une plus courte période dans le portefeuille.


Distribution de la variable Exposure
Nombre d'occurences
Exposition (mois)
Distribution de la variable ClaimNb
Nombre d'occurences
Nombre de sinistres
Impact de DrivAge sur la fréquence
Exposition
Fréquence
Age du conducteur
Impact de DrivAge sur la sévérité
Sévérité
Age du conducteur
Impact de VehAge sur la fréquence
Exposition
Fréquence
Age du véhicule
Impact de VehAge sur la sévérité
Sévérité
Age du véhicule
Impact du BonusMalus sur la fréquence
Exposition
Fréquence
Bonus Malus
Impact du BonusMalus sur la sévérité
Sévérité
Bonus Malus

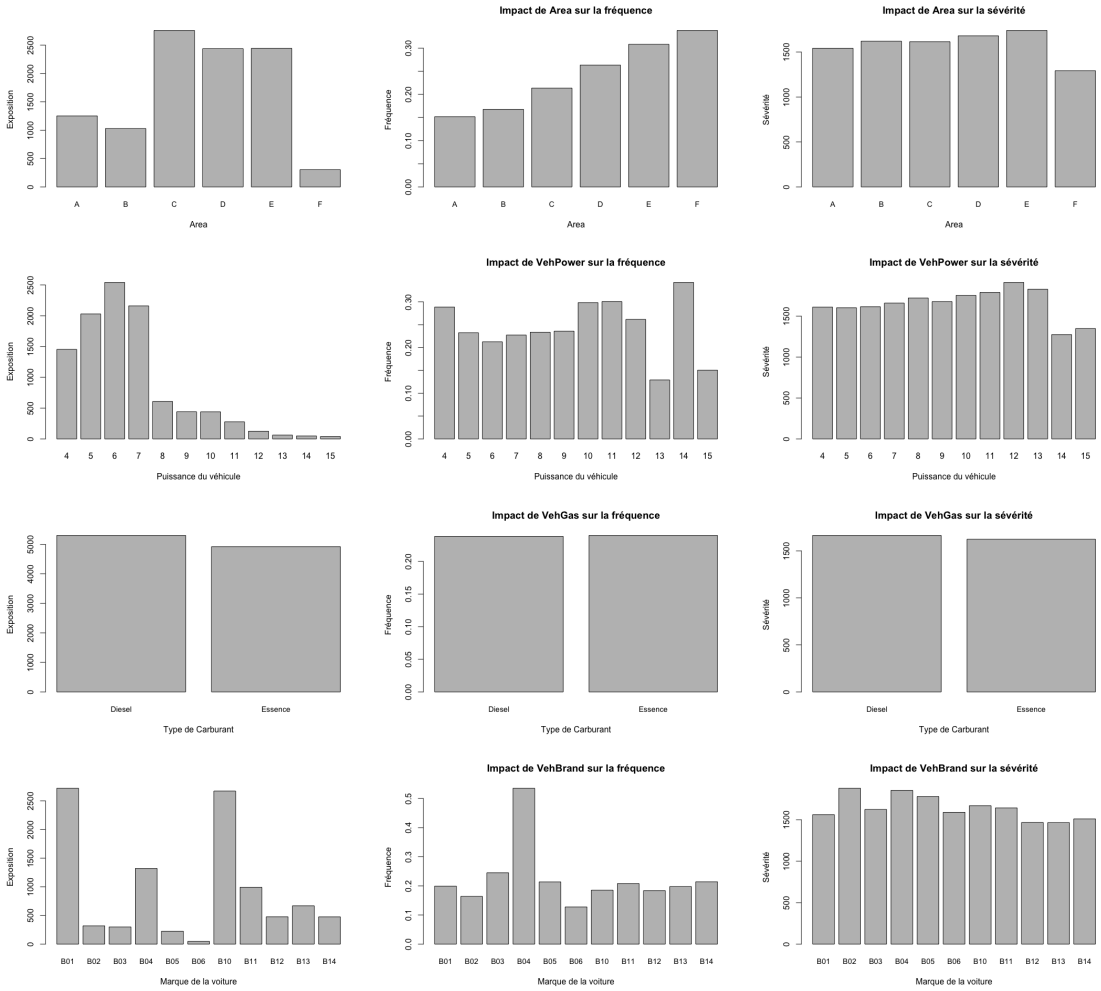


FIGURE 3.1.2 – Statistiques descriptives des variables données : distribution et impact marginal sur la fréquence et la sévérité

3.2 Calibration des modèles de régression

Dans un premier temps, nous commençons par identifier les variables pertinentes pour les modèles marginaux de fréquence et de sévérité.

Nous effectuons une première sélection à l'aide de la fonction `step` de `R`. Cette fonction nous permet de simplifier le modèle en retirant les variables explicatives qui ont peu d'intérêt. Pour cela, elle fonctionne par itération et retire tour à tour chacune des variables explicatives. À chaque tour, la fonction se base sur l'AIC (Akaike Information Criterion) afin de mesurer la qualité du modèle où un AIC plus faible est préféré. Pour éviter l'overfitting, ce critère pénalise la log-vraisemblance ($\ln(L)$) par le nombre de paramètres du modèle $k = p + 1$, comme :

$$\text{AIC} = 2k - 2\ln(L) \tag{3.2.1}$$

Le modèle final obtenu à l'aide de cette fonction a moins de variables et un AIC plus faible. En revanche, logiquement, cette amélioration se fait au prix d'une augmentation de la déviance étant donné que nous retirons certaines variables.

À titre d'exemple nous utilisons le code suivant afin de sélectionner les variables du modèle de sévérité :

```
glm_sev <- step(glm(ClaimAmount ~ Area_modif_amount +[..] , data=
   table_mergee ,family=Gamma(link="log"))
```

Suite à cela, nous retirons ainsi les variables `VehPower`, `VehAge` et `VehGas` pour la fréquence et seulement la variable `VehGas` pour la sévérité.

Une fois les variables pertinentes sélectionnées, il convient de regrouper les niveaux de facteurs afin d'améliorer le modèle en augmentant leur significativité. Tout d'abord, nous étudions les intervalles de confiance des prédictions des coefficients pour chaque niveau de facteurs et regroupons les niveaux dont les intervalles de confiance se superposent. De plus, pour les regroupements moins évidents nous utilisons la fonction `fit.contrast` qui permet de tester la significativité des modalités deux à deux. Nous procédons de manière itérative en regroupant à chaque fois celles avec la plus haute p-valeur. De cette façon, pour deux modalités i et j, un test est effectué sous l'hypothèse nulle :

$$H_0 : \beta_i = \beta_j$$

En parallèle, nous nous appuyons également sur le "Likelihood ratio test" (LRT). Ce dernier permet de tester si la diminution de déviance est significative en incluant ou non un facteur. Nous prenons le modèle H_r sans regroupement, et le modèle H_s avec le regroupement que nous testons. Les estimateurs de maximum de vraisemblance sont respectivement donnés par $\hat{\mu}_r$ et $\hat{\mu}_s$. La statistique de test est donnée par :

$$\text{LRT} = D^*(y, \hat{\mu}_s) - D^*(y, \hat{\mu}_r) \tag{3.2.2}$$

Lorsque LRT $> \chi^2_{r-s,1-\alpha}$, le regroupement est accepté. L'évolution de l'AIC est également surveillée lors des regroupements, ceux le faisant diminuer sont privilégiés.

Afin que les résultats soient comparables d'un modèle à l'autre, nous utilisons les mêmes regroupements pour chacun d'entre eux. Ces regroupements sont détaillés en Annexe A.

3.2.1 Modèle indépendant en deux parties

Tout d'abord, nous modélisons la sévérité et la fréquence des sinistres séparément par un GLM. Ce dernier intègre les variables listées à la section précédente et, à l'aide du logiciel R, nous appelons les deux fonctions GLM suivantes pour calibrer nos modèles.

```
glm_freq <- ztp.glm(table_claimNb$ClaimNb,S,exposure = exposure, sd
   .error = TRUE)

glm_sev <- glm(ClaimAmount ~ Area_modif_amount+[...],data=
   table_claimAmount,family=Gamma(link="log"))
```

Les tables 3.2.2 et 3.2.1 reprennent, respectivement, les coefficients estimés pour les modèles de sévérité et de fréquence des sinistres.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	-1.62032	0.12974	-12.489	< 2e-16	***
Area 2	0.28819	0.07652	3.766	0.000166	***
Area 3	0.16704	0.08719	1.916	0.055407	.
VehBrand 2	0.844	0.06474	13.037	< 2e-16	***
Region 2	-0.5314	0.09598	-5.536	03.09e-08	***
Region 3	-0.20434	0.08745	-2.337	0.019463	*
Region 4	0.15328	0.082	1.869	0.061598	.
Region 5	0.4087	0.11666	3.503	0.000459	***
DrivAge(25,50]	-0.35895	0.0941	-3.814	0.000137	***
DrivAge(50,100]	-0.20053	0.10416	-1.925	0.054195	.
BonusMalus(65,100]	0.5242	0.06618	7.921	2.44e-15	***
BonusMalus(100,230]	0.88179	0.11796	7.475	7.70e-14	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring iterations :	7
Null deviance :	6 249.58 on 14 801 degrees of freedom
Residual deviance :	5 853.18 on 14 789 degrees of freedom
AIC :	8 480.85

TABLE 3.2.1 – Modèle indépendant - Fréquence

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	7.20448	0.03580	201.236	< 2e-16	***
Area 2	-0.29195	0.06851	-4.261	2.04e-05	***
VehPower(6,13]	0.05249	0.02366	2.218	0.026559	*
VehPower(13,15]	-0.20337	0.12140	-1.675	0.093899	.
VehAge(7,10]	-0.09478	0.03067	-3.091	0.002000	**
VehAge(10,100]	-0.16810	0.02806	-5.991	2.14e-09	***
DrivAge(35,70]	0.07685	0.02874	2.674	0.007501	**
DrivAge(70,100]	0.25192	0.05808	4.337	1.45e-05	***
VehBrand 2	0.08510	0.03078	2.765	0.005708	**
VehBrand 3	-0.10386	0.03189	-3.257	0.001128	**
Region 2	0.11730	0.03224	3.639	0.000275	***
Region 3	0.24202	0.03252	7.443	1.04e-13	***
Region 4	0.05812	0.03009	1.932	0.053395	.
BonusMalus(60,75]	0.15846	0.03407	4.652	3.32e-06	***
BonusMalus(75,100]	0.25065	0.03237	7.743	1.03e-14	***
BonusMalus(100,120]	0.27271	0.07846	3.476	0.000511	***
BonusMalus(120,230]	0.51822	0.11323	4.577	4.76e-06	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of Fisher Scoring iterations :	6
Null deviance :	23 685 on 15 431 degrees of freedom
Residual deviance :	23 224 on 15 415 degrees of freedom
AIC :	258 672

TABLE 3.2.2 – Modèle indépendant - Sévérité

Une fois ces modèles estimés, nous pouvons regarder les impacts principaux des facteurs. Notre attention se porte d'abord sur l'effet de la variable de puissance du véhicule (`VehPower`) dans le modèle de sévérité. Celui-ci peut être interpellant. En effet, selon notre modèle, les véhicules de puissance intermédiaire provoqueraient en moyenne une augmentation de la sévérité. Les plus puissants, quant à eux, impactent celle-ci à la baisse. Face à ce type d'observation, il pourrait être intéressant pour un assureur de chercher une explication grâce aux informations supplémentaires dont il dispose. Dans cette base de données, cette estimation peut provenir du faible nombre de véhicules puissants contenus dans le portefeuille, comme nous l'avons observé sur la figure 3.1.2.

Nous pouvons ensuite comparer les coefficients estimés pour les deux modèles. Tout d'abord, en ce qui concerne l'âge du conducteur (`DrivAge`), les plus âgés auraient une diminution du nombre de sinistres moindre par rapport aux autres assurés et en moyenne des sinistres plus chers. Les individus de grand âge semblent donc plus risqués. À nouveau,

cette observation est à nuancer étant donné la faible exposition de ces profils dans le portefeuille considéré (cfr. figure 3.1.2.).

Comme il a été expliqué précédemment, nous pouvons également observer que le signe de l'impact de certains facteurs peut être différent pour les deux grandeurs. Ainsi, pour la variable `Area`, nous observons que celle-ci a une influence opposée sur les deux variables d'intérêt. En effet, les assurés vivant en ville ont en moyenne plus d'accidents que ceux vivant ailleurs. En revanche, ceux-ci restent en moyenne moins chers.

De plus, certains facteurs ont une influence uniquement sur une des composantes de la perte totale. Par exemple, l'âge du véhicule (`VehAge`) est un facteur qui a une influence uniquement sur la sévérité. Il ressort de ce modèle que plus un véhicule est ancien, plus le montant des sinistres va diminuer. Un véhicule datant d'il y a plus de 10 ans aura en moyenne une sévérité inférieure de l'ordre de 15% par rapport à un nouveau véhicule.

Intuitivement, il n'y a pas de raison que la région (`Region`) influence la fréquence et la sévérité. Pourtant, les coefficients associés à cette variable sont significatifs dans les deux modèles. Celle-ci permet probablement de capturer des effets d'autres variables non observables qui influencent directement le risque, variant par région. Nous pouvons citer comme exemple le niveau de richesse moyen, la proportion de jeunes, le niveau de consommation d'alcool ou encore l'état des infrastructures routières.

Il est également intéressant de regarder l'effet du Bonus-Malus (`BonusMalus`), sa valeur étant utilisée afin d'ajuster la prime a posteriori et étant donc censée refléter le risque de la police. Nous observons un effet positif significatif, autant pour la sévérité que pour la fréquence.
Pour rappel, s'agissant de données françaises, un niveau de Bonus-Malus se situant entre 50-100 correspond à une réduction de la prime et donc à un Bonus. A l'inverse, pour un niveau supérieur à 100, un Malus est octroyé. La valeur positive des coefficients est donc intuitivement attendue, un niveau plus élevé indiquant une police plus risquée. Ainsi, par rapport à une personne se trouvant dans la classe de Bonus-Malus la plus basse, un assuré dans la plus élevée aura une augmentation de sévérité de 70% et de fréquence de 140%.

3.2.2 Modèles dépendants

3.2.2.1 Modèle dépendant en deux parties

Dans un premier temps, nous commençons par appliquer la méthode proposée par Garrido et al. 2016. Nous ajoutons donc le nombre de sinistres comme variable explicative de la sévérité individuelle. Le modèle GLM pour le nombre de sinistres reste le même que pour le modèle indépendant. Nous appelons la fonction GLM suivante afin de calibrer le modèle :

```
glm_sev_fonctionFreq <- glm(ClaimAmount ~ Area_modif_amount+ [...]+
   ClaimNb,data=table_claimAmount,family=Gamma(link="log"))
```

Dans la table 3.2.3, sont repris les coefficients estimés pour la sévérité.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	7.10774	0.04795	148.227	< 2e-16	***
Area 2	-0.29342	0.06842	-4.288	1.81e-05	***
VehPower(6,13]	0.05411	0.02364	2.289	0.022109	*
VehPower(13,15]	-0.19858	0.12124	-1.638	0.101470	
VehAge(7,10]	-0.09217	0.03063	-3.009	0.002624	**
VehAge(10,100]	-0.16622	0.02803	-5.929	3.11e-09	***
DrivAge(35,70]	0.07514	0.02871	2.618	0.008864	**
DrivAge(70,100]	0.24581	0.05803	4.236	2.29e-05	***
VehBrand 2	0.06996	0.03086	2.268	0.023372	*
VehBrand 3	-0.10257	0.03185	-3.221	0.001282	**
Region 2	0.11192	0.03225	3.471	0.000520	***
Region 4	0.23122	0.03260	7.092	1.38e-12	***
Region 5	0.05235	0.03009	1.740	0.081903	.
BonusMalus(60,75]	0.15660	0.03405	4.599	4.28e-06	***
BonusMalus(75,100]	0.24715	0.03239	7.631	2.47e-14	***
BonusMalus(100,120]	0.27086	0.07843	3.454	0.000554	***
BonusMalus(120,230]	0.49440	0.11338	4.361	1.31e-05	***
ClaimNb	0.09288	0.03052	3.043	0.002343	**

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of Fisher Scoring iterations :	7
Null deviance :	23 685 on 15 431 degrees of freedom
Residual deviance :	23 206 on 15 414 degrees of freedom
AIC :	258 659

Table 3.2.3 – Modèle dépendant en deux parties (nombre de sinistres) - Sévérité

Premièrement, nous pouvons remarquer que l'ajout du nombre de sinistres en variable explicative de la sévérité améliore le modèle de manière significative. En effet, l'AIC et la déviance diminuent, passant respectivement de 258'672 à 258'659 et de 23'224 à 23'206. Nous observons que le nouveau coefficient représentant le nombre de sinistres est significatif et a une valeur positive. Ainsi, chaque sinistre supplémentaire augmenterait, en moyenne, la sévérité de 9.7%. Selon ce premier chiffre, il semble donc qu'il existe une dépendance positive entre la sévérité et le nombre de sinistres.

Nous pouvons néanmoins nous demander quelle serait l'interprétation d'une telle mesure. En effet, en considérant le nombre de sinistres comme variable explicative, nous ne prenons jamais en compte, lors de l'estimation de la dépendance, le nombre d'années pendant lesquelles le contrat a été en cours. De plus, nous pouvons nous interroger sur la fiabilité d'un coefficient estimé sur le nombre de sinistres qui a très peu de valeurs supérieures à 1.

Pour ces raisons, nous préférons estimer le modèle en ajoutant la fréquence annuelle de sinistres de la police comme variable explicative. De cette façon, nous pouvons prendre en compte la durée d'exposition de chaque police dans la modélisation de la dépendance. Nous avons d'abord essayé d'ajouter les fréquences brutes ($\texttt{Frequence} = \frac{Y_i}{e_i}$) dans le modèle conditionnel des coûts, à l'aide du code `R` suivant :

```
glm_sev_fonctionFreq <- glm(ClaimAmount ~ Area_modif_amount+ [...]+
   Frequence,data=table_claimAmount,family=Gamma(link="log"))
```

Ce type de modèle nous permet donc d'étudier le lien entre sévérité et fréquence annuelle.

Parallèlement à ce dernier modèle, nous avons testé la division des contrats selon différentes catégories de fréquence annuelle. Ainsi, trois catégories de fréquence ont été créées selon les quantiles de la distribution de la fréquence :

1. Fréquences faibles
2. Fréquences moyennes
3. Fréquences élevées

Nous obtenons alors des classes avec un nombre suffisant d'individus.

De ce fait, les fréquences inférieures à 1.0753 sinistres par an sont catégorisées comme "faibles", celles comprises entre 1.0753 et 2.0833 sont "moyennes" et le reste est labélisé comme "élevées". Afin de calibrer le modèle, la variable de fréquence annuelle sous forme de catégories, `FrequenceModif`, est créée et nous faisons donc appel au code `R` suivant :

```
table_mergee$FrequenceModif[table_mergee$frequence<=1.075269]='F'
table_mergee$FrequenceModif[[...]>1.075269 & [...]<=2.083333]='M'
table_mergee$FrequenceModif[[...]>2.083333]='E'

glm_sev_fonctionFreq <- glm(ClaimAmount ~ Area_modif_amount+[...]+
   FrequenceModif,data=table_mergee,family=Gamma(link="log"))
```

La table 3.2.4 reprend l'AIC et la déviance des trois différents modèles décrits dans cette section. Comme nous pouvons l'observer, par rapport au modèle considérant le nombre de sinistres comme variable explicative ou encore celui tenant compte plutôt des fréquences annuelles brutes, l'AIC et la déviance s'améliorent fortement dans le modèle classant la fréquence annuelle en trois catégories. En effet, en utilisant les fréquences brutes comme variables, nous obtenons un coefficient en fonction d'une variable continue, ce qui suppose une relation strictement linéaire entre la fréquence annuelle et la sévérité. Ceci donne un modèle moins bon que celui où cette variable est transformée en plusieurs catégories. Cela semble donc confirmer qu'un modèle utilisant la fréquence annuelle explique mieux nos données et davantage lorsque celle-ci est catégorisée.

Variable explicative de la sévérité	AIC	Déviance
`ClaimNb`	258'659	23'206
`Frequence`	258'547	23'068
`FrequenceModif`	258'349	22'823

TABLE 3.2.4 – AIC et déviance selon la variable explicative testée

Finalement, nous reprenons à la table 3.2.5, les différents résultats pour le dernier modèle tenant compte des trois catégories de fréquence annuelle. Si nous nous intéressons aux valeurs des coefficients estimés, un contrat avec une fréquence moyenne aura une sévérité attendue plus élevée de 20% par rapport à un contrat avec une fréquence faible. Ce surplus de sévérité monte jusqu'à 50% pour un contrat avec une fréquence de sinistres élevée. Les assurés qui réalisent plus fréquemment des sinistres ont donc en moyenne une sévérité plus importante.

En outre, nous pouvons également étudier les changements qu'entraîne l'introduction de la dépendance sur la valeur des coefficients des autres variables explicatives.
Cependant, en introduisant le nombre de sinistres dans le modèle de sévérité, tous les coefficients du modèle changent. Afin de pouvoir isoler les coefficients sensibles à cette introduction, nous procédons en deux étapes.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	7.00306	0.03803	184.152	< 2e-16	***
Area 2	-0.29730	0.06751	-4.404	1.07e-05	***
VehPower(6,13]	0.05202	0.02332	2.231	0.025708	*
VehPower(13,15]	-0.19515	0.11962	-1.631	0.102818	
VehAge(7,10]	-0.10477	0.03022	-3.467	0.000527	***
VehAge(10,100]	-0.18663	0.02767	-6.744	1.60e-11	***
DrivAge(35,70]	0.11404	0.02836	4.021	5.82e-05	***
DrivAge(70,100]	0.30744	0.05736	5.360	8.42e-08	***
VehBrand 2	0.01606	0.03072	0.523	0.601038	
VehBrand 3	-0.09878	0.03142	-3.144	0.001671	**
Region 2	0.10067	0.03178	3.167	0.001541	**
Region 3	0.20476	0.03215	6.369	1.96e-10	***
Region 4	0.04112	0.02967	1.386	0.165848	
BonusMalus(60,75]	0.15171	0.03361	4.513	6.43e-06	***
BonusMalus(75,100]	0.20306	0.03228	6.291	3.23e-10	***
BonusMalus(100,120]	0.27588	0.07734	3.567	0.000362	***
BonusMalus(120,230]	0.51429	0.11165	4.606	4.14e-06	***
Fréquence moyenne	0.19071	0.02841	6.713	1.97e-11	***
Fréquence élevée	0.40454	0.02801	14.443	< 2e-16	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of Fisher Scoring iterations :	6
Null deviance :	23 685 on 15 431 degrees of freedom
Residual deviance :	22 823 on 15 413 degrees of freedom
AIC :	258 349

TABLE 3.2.5 – Modèle dépendant en deux parties (fréquence annuelle) - Sévérité

Tout d'abord, nous introduisons la fréquence annuelle comme variable explicative en figeant les autres bêtas du modèle indépendant. Pour cela, nous figeons ces derniers en offset et faisons donc appel au code `R` suivant :

```
offset.indep <- log(glm_sev$fitted.values)
glm_offset <- glm(ClaimAmount ~ FrequenceModif+offset(offset.indep)
   ,data=table_claimAmount,family=Gamma(link="log"))
```

Ensuite, nous figeons les bêtas obtenus relatifs à la fréquence annuelle, en laissant les autres libres :

```
glm_freq.offset <- glm(ClaimAmount ~ Area_modif_amount+[...]+offset
   (betas_frequence),data=table_claimAmount,family=Gamma(link="log"
   ))
```

Dans la table 3.2.6, nous comparons la variation de ces derniers coefficients par rapport au cas indépendant.

Coefficient	Variation	Coefficient	Variation
(Intercept)	0%	VehBrand 3	-7%
Area 2	0%	Region 2	-7%
VehPower6,13]	4%	Region 3	-13%
VehPower(13,15]	-3%	Region 4	-29%
VehAge(7,10]	10%	BonusMalus(60,75]	-3%
VehAge(10,100]	10%	BonusMalus(75,100]	-14%
DrivAge(35,70]	30%	BonusMalus(100,120]	-0%
DrivAge(70,100]	14%	BonusMalus(120,230]	-5%
VehBrand 2	-64%		

TABLE 3.2.6 – Variation des coefficients

Par rapport au modèle indépendant, tous les coefficients ont vu leur variance légèrement diminuer (-1%). En revanche, nous n'observons pas d'effet consistant sur la valeur même des coefficients. Certaines variables comme la région (`Region`), la marque du véhicule (`VehBrand`) ou le niveau de Bonus Malus (`BonusMalus`) perdent globalement de l'importance. En revanche, d'autres variables comme l'âge du véhicule (`VehAge`) ou du conducteur (`DrivAge`) gagnent en importance.

Bien que donnant une première idée sur la dépendance entre la fréquence et la sévérité, ce modèle se trouve néanmoins rapidement limité.
Premièrement, il ne permet que de modéliser l'influence directe de la fréquence sur le montant des sinistres. C'est pour cette raison qu'un modèle basé sur l'utilisation d'effets aléatoires est présenté par la suite. Ces derniers permettent de modéliser l'éventuelle dépendance apparente entre les deux grandeurs, celle-ci pouvant provenir de variables non disponibles lors de la réalisation de l'étude et ayant un effet simultané sur la fréquence et la sévérité. Deuxièmement, cette approche modélise une structure de dépendance relativement simple. Les modèles basés sur les copules, que nous présentons ensuite, nous permettent d'analyser des structures de dépendance plus complexes sans nécessiter de faire des hypothèses trop restrictives quant à leur forme.

3.2.2.2 Modèle de copules

Afin d'implémenter le modèle de copules, nous utilisons, comme indiqué précédemment, le package R `copulaRegression` développé dans Krämer et al. 2013. Ce package permet de calibrer des copules liant deux modèles de régression. Le premier, suit une distribution Gamma pour la sévérité et le second, une distribution ZTP pour la fréquence.

Afin de calibrer la copule, l'algorithme procède en deux étapes principales. Tout d'abord, afin d'obtenir des valeurs initiales efficientes, les deux modèles de régression marginaux sont ajustés séparément et sont ensuite liés par une copule. En décomposant le processus en deux étapes comme ceci, il n'est pas possible de prendre en compte l'impact de la copule sur la valeur des coefficients des deux modèles marginaux. Dans un deuxième temps, l'algorithme maximise la vraisemblance de la densité jointe, en utilisant les valeurs initiales calculées précédemment. La densité jointe, donnée à l'équation (2.4.8), dépend à la fois du paramètre de la copule, θ, et des paramètres des modèles de régression. L'optimisation est donc faite simultanément sur ces différentes composantes.

Nous réalisons la calibration pour quatre copules différentes : Gauss, Clayton, Gumbel et Frank. À titre d'exemple, l'appel du package pour la copule de Gauss se fait à l'aide de la commande suivante :

```
family = 1
copulaGauss <- copreg(x,y,R,S,family,exposure,sd.error=T,joint=T)
```

La table 3.2.7 reprend le paramètre de la copule estimé ainsi que le tau de Kendall (τ) qui permet de comparer les valeurs de la dépendance modélisée.
Tout d'abord, il semble que la copule de Gumbel ne soit pas capable de s'adapter correctement aux données. Lorsqu'une des marginales est discrète, il n'est pas toujours possible de trouver le paramètre adéquat de la copule. Pour la suite, nous ne présenterons pas les résultats pour cette copule. Parmi les trois autres copules, la valeur du τ estimée est relativement proche. Dans notre cas, le choix de la copule n'a donc pas d'influence excessive sur la valeur du τ.

	Gauss	Clayton	Gumbel	Frank
θ	0.0473	0.0617	1.0082	0.2473
	(0.0157)	*(0.6034)*	*(0.4585)*	*(0.1148)*
τ	0.0301	0.0299	0.0081	0.0272
	(0.0097)	*(0.1804)*	*(0.2385)*	*(0.0126)*

Table 3.2.7 – θ et τ estimés pour chaque copule

À première vue, le coefficient de corrélation (τ) estimé peut sembler faible. Toutefois, une des distributions marginales étant discrète, le support du τ est restreint et n'est plus

donné par [-1,1]. Nous cherchons à trouver la valeur maximum que le τ pourrait prendre ici. Cette borne supérieure est donnée par Denuit et al. 2019 comme :

$$\tau \leq 2E[F(Y-)]$$

$F(Y-)$ représente la valeur de la fonction de distribution juste avant chaque saut. Cette valeur est calculée pour les données utilisées à la table 3.2.8.

Y	Nb. observé	f(y)	F(Y)	F(Y-)
1	13'667	0.9234	0.9234	0
2	1'065	0.0719	0.9953	0.9234
3	60	0.0041	0.9994	0.9953
4	9	0.0006	1	0.9994

Table 3.2.8 – Calcul de la valeur maximale du τ

Dans notre cas, la borne supérieure est de 0.1422. Les valeurs du τ estimé représentent environ 20% de cette valeur maximum, ce qui décrit une dépendance plus importante que ce que les valeurs brutes laissent apparaître. Comme dans le modèle dépendant en deux parties précédent, la dépendance entre la sévérité et la fréquence détectée par ce modèle est positive.

Afin de choisir la copule la plus adaptée, nous les comparons à l'aide du test de Vuong. Comme décrit plus haut à l'équation (2.4.11), ce test compare les copules deux à deux en comparant leur log-vraisemblance respective.
La table 3.2.9 reprend les résultats de la statistique de test ainsi que la copule préférée pour chaque couple. Afin de confirmer les résultats de ce test, nous reprenons dans la table 3.2.10 l'AIC du modèle pour chaque copule.

Pour rappel, ce modèle de copule utilise la sévérité moyenne de chaque police alors que le modèle dépendant en deux parties, présenté précédemment utilise le montant individuel de chaque sinistre. Les valeurs des AIC de ces deux modèles ne sont donc pas directement comparables.

		Copule 2		
		Gauss	Clayton	Frank
	Gauss		*Gauss*	*Gauss*
Copule 1			(+1.76)	(+1.97)
	Clayton	*Gauss*		*Frank*
		(-1.76)		(-0.94)
	Frank	*Gauss*	*Frank*	
		(-1.97)	(+0.94)	

Table 3.2.9 – Copule sélectionnée deux à deux

	Indépendant	Dépendant
Gauss	257'691	255'986
Clayton	257'691	255'993
Frank	257'691	255'991

TABLE 3.2.10 – AIC pour chaque copule

En revanche, pour observer l'effet de l'introduction de la dépendance dans ce modèle, nous pouvons comparer les modèles de copule "Indépendant" et "Dépendant" de la table 3.2.10. L'AIC diminue systématiquement suite à la prise en compte de la dépendance. Pour la copule de Gauss, il passe de 257'691 à 255'986. Cela suggère que le modèle s'améliore lorsque les deux distributions marginales sont dépendantes.

A l'aide des résultats du test de Vuong et des valeurs de l'AIC, nous pouvons en conclure que la copule de Gauss est la plus adaptée pour nos données. Ceci est en ligne avec les valeurs des erreurs standards du θ et du τ données dans la table 3.2.7. La copule Gaussienne a en effet l'erreur standard relative la plus faible.
Nous nous concentrons donc sur cette dernière. Les tables 3.2.11 et 3.2.12 résument les coefficients estimés pour les modèles marginaux de la fréquence et de la sévérité liés par cette copule. Les résultats du modèle de la copule de Gauss indépendant sont détaillés à l'annexe B.

Par rapport aux coefficients obtenus avec le modèle de copule indépendant, nous observons que pour le modèle marginal de la sévérité, l'intégration de la dépendance a un effet important sur la valeur des différents coefficients. Cet effet n'est néanmoins pas consistant pour tous les facteurs, certains voient leur valeur augmenter alors que d'autres la voient diminuer. L'effet sur la variance est en revanche beaucoup plus clair. En effet, l'erreur standard des coefficients diminue pour la plupart. Cette baisse de variance entraîne une hausse globale de la significativité des coefficients. Par exemple, l'âge du conducteur (`DrivAge`) et la variable de Bonus-malus (`BMS`) gagnent fortement en significativité.

Concernant la fréquence, nous observons que les coefficients restent sensiblement les mêmes par rapport au modèle de copule indépendant. En intégrant la dépendance, le modèle de copule ne semble pas avoir beaucoup d'effet sur la composante fréquence du modèle. Au contraire, le modèle marginal de la sévérité, semble, lui, fort influencé par les informations introduites par la dépendance, permettant de réduire de façon significative l'aléa de la plupart des coefficients de ce modèle.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	-1.61408	0.12949	-12.465	< 2e-16	***
Area 2	0.29037	0.0764	3.801	0.000144	***
Area 3	0.16635	0.08708	1.91	0.056094	.
VehBrand 2	0.83151	0.06488	12.816	< 2e-16	***
Region 2	-0.52844	0.09572	-5.52	3.38e-08	***
Region 3	-0.20413	0.08727	-2.339	0.019328	*
Region 4	0.14927	0.08189	1.823	0.068338	.
Region 5	0.40413	0.11657	3.467	0.000526	***
DrivAge(25,50]	-0.35581	0.09386	-3.791	0.000150	***
DrivAge(50,100]	-0.2022	0.10395	-1.945	0.051746	.
BonusMalus(65,100]	0.52841	0.06605	8	1.33e-15	***
BonusMalus(100,230]	0.8843	0.11761	7.519	5.51e-14	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.2.11 – Modèle de copule de Gauss - Fréquence

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	7.20096	0.02881	249.913	< 2e-16	***
Area 2	-0.30924	0.05624	-5.499	3.83e-08	***
VehBrand 2	0.04653	0.02546	1.827	0.067652	.
VehBrand 3	-0.10054	0.02593	-3.877	0.000106	***
Region 2	0.11672	0.02636	4.429	9.49e-06	***
Region 3	0.23697	0.02674	8.863	< 2e-16	***
Region 4	0.0502	0.02462	2.039	0.041410	*
DrivAge(35,70]	0.0776	0.02358	3.291	0.000997	***
DrivAge(70,100]	0.24143	0.04767	5.065	4.09e-07	***
VehAge(7,10]	-0.08609	0.02503	-3.44	0.000582	***
VehAge(10,100]	-0.16591	0.02292	-7.238	4.56e-13	***
VehPower(6,13]	0.05159	0.01948	2.649	0.008069	**
VehPower(13,15]	-0.23251	0.09967	-2.333	0.019658	*
BonusMalus(60,75]	0.17311	0.02782	6.223	4.86e-10	***
BonusMalus(75,100]	0.25976	0.02679	9.695	< 2e-16	***
BonusMalus(100,120]	0.30716	0.06468	4.749	2.05e-06	***
BonusMalus(120,230]	0.47325	0.0945	5.008	5.50e-07	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.2.12 – Modèle de copule de Gauss - Sévérité

3.2.2.3 Modèle à effets aléatoires

Nous avons défini plus tôt, aux équations (2.4.13) et (2.4.14), les fonctions de vraisemblance des modèles que nous allons appliquer. La présence d'intégrales dans leurs expressions pose problème lors de la maximisation. Afin de maximiser ces fonctions, la procédure `NLMIXED` du logiciel `SAS` est utilisée.
Nous avons fait le choix d'utiliser ce logiciel principalement pour deux raisons. Tout d'abord, les performances du logiciel `SAS` sont nettement supérieures en termes de calcul par rapport à ce qui peut être fait dans `R` pour ce type de problèmes. Ensuite, la procédure `NLMIXED` a l'avantage de permettre de personnaliser la fonction de vraisemblance utilisée et de la spécifier manuellement. Ceci est particulièrement adapté pour l'utilisation des effets aléatoires que nous souhaitons faire, celle-ci n'étant, à notre connaissance, pas implémentée telle quelle dans un package `R`.

Le fonctionnement d'une procédure `SAS` pouvant à première vue paraître plus opaque, nous décrivons tout d'abord brièvement son fonctionnement.
Pour ceci, nous nous focalisons sur l'algorithme d'approximation numérique de l'intégrale utilisé. Ce dernier est basé sur la quadrature adaptive Gaussienne (Pinheiro et Bates 1995). Cette méthode remplace le calcul de l'intégrale par une somme pondérée sur certaines valeurs des effets aléatoires :

$$I = \int_{-\infty}^{\infty} g(x)dx \approx \sum_{i=1}^{n} w_i e^{x_i^2} g(x_i)$$

avec w_i = les poids et abscisses standards de Gauss Hermite (tels que décrits dans la Table 25.10 Abramowitz *et* Stegun *1972).*
Le nombre de points de quadrature est choisi automatiquement par la procédure `NLMIXED` en fonction de la variation de la log-vraisemblance entre les itérations et des valeurs initiales données.

La procédure `NLMIXED` propose plusieurs algorithmes d'optimisation. Aucun n'étant optimal pour tous les problèmes, il doit être choisi selon la situation.
Dans notre cas, nous avons opté pour la méthode du gradient conjugué ("Conjugate gradient methods"). Celle-ci est en effet adaptée pour l'optimisation de problèmes avec de nombreux paramètres et une quantité relativement importante de données.
La méthode du gradient conjugué est basée sur un algorithme itératif utilisant la valeur de la fonction de vraisemblance et le vecteur de gradient (dérivées partielles de premier ordre). À chaque itération, la fonction est mise à jour selon la méthode "Automatic Restart Update" développée par Powell 1977.
Par rapport aux autres procédés proposés, qui eux, calculent les dérivées partielles de second ordre par une approximation de la matrice Hessienne, chaque itération est plus rapide à calculer. Cependant, en contrepartie, elle nécessite davantage d'itérations. Par conséquent, le temps de calcul est supérieur et sa fiabilité dépend fortement des valeurs

initiales choisies, qui peuvent amener à tendre vers un optimum local. Tout en restant attentif à ces points, elle reste la méthode la plus adaptée aux dimensions de notre problème.

Une fois familiarisés avec le fonctionnement de cette procédure, nous pouvons l'appliquer sur nos données. Nous résumons l'appel de la procédure ci-dessous.

```
1   proc nlmixed data=[...] tech=congra;
2
3       /*Likelihood ClaimNumber*/
4       y = b0 + b1*Area_modif2 + [...] + s*zeta;
5       lambda = exp(y);
6       ll1 = ClaimNb*y - lambda - log(1-exp(-lambda)) -
           log(fact(ClaimNb));
7
8       /*Likelihood ClaimAmount*/
9       xx = a0 + a1*Area_modif_amount2 + [...] + zeta;
10      mu = exp(xx);
11      ll_Amount1 = 0; ll_Amount2 = 0; ll_Amount3 = 0; ll_Amount4
           = 0;
12
13      if ClaimAmount1 > 0 then do;
14      ll_Amount1 = alpha*(-ClaimAmount1/mu-xx) +
           log(alpha**alpha/ GAMMA(alpha) *
           ClaimAmount1**(alpha-1));    ;
15      end;
16      else if ClaimAmount2 > 0 then do;
17      [...]
18      end;
19
20      L = ll1+ll_Amount1+ll_Amount2+ll_Amount3+ll_Amount4;
21
22      loss = ClaimAmount1+ClaimAmount2+ClaimAmount3+ClaimAmount4;
23
24
25      model loss ~ general(L);
26      random zeta~ normal(0,sigma_zeta) subject=IDPol;
27  run;
```

Cet appel de fonction correspond aux effets aléatoires partagés. Pour les effets aléatoires corrélés, le principal changement se situe au niveau de l'appel de la commande `random` à la ligne 26 :

```
1   [...]
2   random zeta zeta1~ normal([0,0],
       [sigma_zeta,covvar,sigma_zeta1]) subject=IDPol;
3   [...]
```

La maximisation de la fonction de vraisemblance étant sensible aux choix des valeurs initiales lors de l'appel de la procédure, nous devons y être attentif au moment de la calibration du modèle. Nous avons choisi de réaliser une approche par étapes.

Pour cela, nous commençons par estimer les deux modèles GLM marginaux, respectivement de la fréquence et de la sévérité, séparément sans prise en compte d'effets aléatoires. Les paramètres ainsi obtenus sont ensuite utilisés comme valeurs initiales pour le modèle joint. Afin de nous assurer que le modèle converge correctement vers l'optimum, nous appliquons une maximisation par itération.
Nous débutons avec un modèle réduit à une seule variable explicative et nous ajoutons les autres une par une à chaque itération. Chacune de ces dernières peut se décomposer de la manière suivante :

1. Modèle 1 : estimation du modèle joint sans effet aléatoire. Les valeurs initiales sont celles données par les modèles marginaux indépendants.
2. Modèle 2 : estimation du modèle joint avec effets aléatoires. Les valeurs initiales sont :
   - Paramètres de régression de la variable ajoutée (β,α) : estimations du Modèle 1
   - Paramètres des effets aléatoires (σ_ψ,σ_ξ, s, ρ) et autres paramètres de régression (β,α) : estimations de l'itération précédente du Modèle 2

Nous reprenons dans la table 3.2.13 les principaux résultats des modèles estimés à l'aide de cet algorithme.

Modèle joint	**Paramètres d'intérêt**	**AIC**
Modèle sans effet aléatoire		267'153
Modèle à effets aléatoires indépendants	$\sigma^2_\psi = 0.517^{***}$ $\sigma^2_\xi = 0.1082$	255'108
Modèle avec effets aléatoires corrélés	$\sigma^2_\psi = 0.5142^{***}$ $\sigma^2_\xi = 0.1061$ $\rho = 0.1185^{**}$	255'100
Modèle avec effets aléatoires partagés	$\sigma^2_\psi = 0.5127^{**}$ $s = 0.2388^{**}$	255'098

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.2.13 – Résultats des modèles à effets aléatoires

Sur base de ces chiffres, nous pouvons tout d'abord remarquer que l'intégration d'effets aléatoires améliore significativement le modèle. L'AIC passe effectivement de 267'153 à 255'108. Cela semble confirmer l'existence d'une hétérogénéité résiduelle importante provenant de facteurs non observés. Étant donné le nombre limité de variables explicatives mises à disposition, ceci n'est pas vraiment étonnant. Nous remarquons néanmoins que le paramètre des effets aléatoires du côté de la fréquence n'est pas significativement différent de 0. Ceci suggère que l'hétérogénéité est peu importante dans ce modèle marginal. Ceci est sans doute dû aux données considérées, étant donné que nous ne considérons que les polices sinistrées.

Nous pouvons à présent nous concentrer sur les résultats des modèles intégrant la dépendance. Les deux modèles semblent donner des résultats consistants, ils décrivent en effet tous deux l'existence d'une dépendance positive. En prenant la dépendance en compte, nous observons que l'AIC a diminué par rapport au modèle à effets aléatoires indépendants. Cette dépendance semble donc importante pour expliquer nos données.

Tout d'abord, nous analysons le modèle utilisant des effets aléatoires partagés. Le paramètre d'échelle (s) a une valeur significativement positive à 0.2388. Ceci traduit bien une relation linéaire positive entre la fréquence et la sévérité. Il est logique que ce paramètre soit inférieur à 1, car nous passons de la sévérité à la fréquence, où l'ordre de grandeur n'est évidemment pas le même.
Regardons ensuite les paramètres du modèle à effets aléatoires corrélés. Dans ce modèle, le paramètre de corrélation (ρ) a une valeur de 0.1185 qui est significativement différente de 0. Il existe donc une relation positive entre fréquence et sévérité.

Parmi les deux modèles intégrant la dépendance, il s'agit désormais de choisir le modèle le plus adapté. Le modèle intégrant des effets aléatoires liés par un paramètre d'échelle linéaire a un AIC légèrement inférieur à celui prenant en compte des effets corrélés (respectivement 255'098 et 255'100). Néanmoins, nous observons que le ρ liant les deux effets aléatoires σ_ψ^2 et σ_ξ^2 distribués normalement est éloigné de la valeur 1. Cette valeur indique que la dépendance détectée par le modèle n'est pas linéaire, un simple paramètre de mise à l'échelle semble donc trop limitatif pour ce cas.
De plus, le modèle à effets aléatoires corrélés permet une modélisation de la dépendance plus flexible. Pour ces deux raisons, nous choisissons d'étudier ce modèle plus en détails et de laisser l'autre de côté, bien qu'il ait un AIC légèrement plus faible.

En regardant plus en détails les paramètres de variance estimés, nous observons que la variance des effets aléatoires du modèle de sévérité est plus significative que ceux du modèle de fréquence. L'intégration de la dépendance semble induire plus de modifications dans le modèle marginal de la sévérité, ce qui a déjà pu être observé dans le cadre des autres modèles développés précédemment.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	6.8975	0.03484	197.95	<.0001	***
Area 2	-0.2363	0.06264	-3.77	0.0002	***
VehPower(6,13]	0.03076	0.02183	1.41	0.1588	
VehPower(13,15]	-0.1865	0.1115	-1.67	0.0945	.
VehAge(7,10]	-0.09136	0.02822	-3.24	0.0012	**
VehAge(10,100]	-0.1833	0.02579	-7.11	<.0001	***
DrivAge(35,70]	0.125	0.0265	4.72	<.0001	***
DrivAge(70,100]	0.2912	0.05347	5.45	<.0001	***
VehBrand 2	0.09593	0.02863	3.35	0.0008	***
VehBrand 3	-0.09587	0.02924	-3.28	0.001	***
Region 2	0.1265	0.02963	4.27	<.0001	***
Region 3	0.2198	0.03009	7.3	<.0001	***
Region 4	0.07535	0.02764	2.73	0.0064	**
BonusMalus(60,75]	0.1937	0.03131	6.19	<.0001	***
BonusMalus(75,100]	0.2933	0.02984	9.83	<.0001	***
BonusMalus(100,120]	0.3704	0.07244	5.11	<.0001	***
BonusMalus(120,230]	0.5256	0.1061	4.95	<.0001	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.2.14 – Modèle à effets aléatoires corrélés - Sévérité

Nous reprenons dans les tables 3.2.14 et 3.2.15, les estimations pour les paramètres de chaque modèle marginal.
Si nous regardons plus en détails les coefficients estimés pour la sévérité, nous pouvons voir que les coefficients gagnent globalement en importance, sauf pour la variable `Area` et la puissance du véhicule (`VehPower`) qui perdent en significativité. Par rapport aux modèles en deux parties, la variance des facteurs diminue de façon assez homogène, de l'ordre de -7%. En revanche, l'effet inverse est observé si nous faisons la comparaison par rapport au modèle de copules. Par rapport à ce dernier, nous observons une augmentation de variance de l'ordre de 15%. Cette hausse de variance est souvent compensée par une hausse de la valeur du coefficient, ce qui explique que la significativité des facteurs ne diminue globalement pas.

Concernant la fréquence, la valeur des paramètres estimés n'évolue que de façon marginale. Nous pouvons voir que les variables géographiques `Region` et `Area` perdent légèrement en importance, mais l'effet reste marginal. Ces variables géographiques ne devraient pas influencer directement le risque, elles permettent simplement de capturer en partie les effets d'autres variables non observées. Ces derniers sont désormais capturés, du moins partiellement, par les effets aléatoires. Concernant l'évolution des variances, il y a une augmentation par rapport aux autres modèles, celle-ci reste néanmoins très limitée, de l'ordre de 2%.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	-1.6637	0.1431	-11.63	<.0001	***
Area 2	0.2906	0.07693	3.78	0.0002	***
Area 3	0.1649	0.08764	1.88	0.0598	.
VehBrand 2	0.8445	0.06545	12.9	<.0001	***
Region 2	-0.531	0.09637	-5.51	<.0001	***
Region 3	-0.2033	0.08794	-2.31	0.0208	*
Region 4	0.1504	0.08258	1.82	0.0686	.
Region 5	0.4099	0.1179	3.48	0.0005	***
DrivAge(25,50]	-0.3646	0.09495	-3.84	0.0001	***
DrivAge(50,100]	-0.2097	0.1051	-1.99	0.0462	*
BonusMalus(65,100]	0.5268	0.06668	7.9	<.0001	***
BonusMalus(100,230]	0.8794	0.1196	7.35	<.0001	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.2.15 – Modèle à effets aléatoires corrélés - Fréquence

Comme cela a déjà pu être observé dans les autres modèles, nous pouvons voir que l'intégration de la dépendance a un effet beaucoup plus important sur les facteurs du modèle de sévérité que ceux de la fréquence. Ceci pourrait signifier que la dépendance entre les deux variables de réponse provient d'une influence du nombre de sinistres sur leur montant et non d'un lien inverse.

3.2.2.4 Modèle dépendant en deux parties : distributions mixtes

Comme mentionné précédemment, l'introduction des effets aléatoires améliore fortement le modèle. Ceci semble indiquer la présence d'une hétérogénéité résiduelle importante. Le modèle dépendant en deux parties présenté antérieurement se base quant à lui sur les distributions de Poisson et Gamma. Nous ne pouvons comparer directement ce dernier avec la modélisation de la dépendance utilisant les effets aléatoires. En effet, il ne serait pas possible de conclure si l'amélioration provient de distributions marginales plus adaptées ou alors d'une meilleure manière de modéliser la dépendance.

Pour répondre à ce problème, nous ajoutons dans le modèle dépendant en deux parties deux effets aléatoires indépendants dans les expressions marginales de la fréquence et de la sévérité. Ceci nous permet d'utiliser des distributions Poisson mélange et Gamma mélange. Nous utilisons ainsi la fonction de vraisemblance spécifiée pour le modèle à effets aléatoires corrélés à l'équation (2.4.14) en fixant $\rho = 0$. De plus, dans la régression de la sévérité nous ajoutons les termes prenant en compte l'effet de la fréquence annuelle sur le montant des sinistres. Ces termes prennent la même forme qu'à l'équation (2.4.4).

Suite à nos calculs, nous remarquons que du côté de la sévérité, l'utilisation de la Gamma mixte améliore sensiblement le modèle. En effet, l'AIC diminue fortement, passant de 258'349 à 246'371. En revanche, dans le modèle de la fréquence nous remarquons que l'AIC augmente légèrement, passant de 8'480.85 à 8'482. Ce résultat semble nous indiquer qu'une distribution de Poisson mixte n'est pas indispensable. Cette observation semble être confirmée par les résultats du modèle à effets aléatoires présentés dans la table 3.2.13. En effet, le paramètre de variance des effets aléatoires concernant le modèle de fréquence n'est pas significativement différent de 0. L'hétérogénéité résiduelle semble être concentrée du côté de la sévérité.

Nous pouvons tout de même noter que ce résultat n'est pas forcément attendu. En effet, il est courant de retrouver de la sur-dispersion du côté de la fréquence. Ceci implique qu'en général, la distribution de Poisson qui impose que la variance soit égale à la moyenne, n'est pas adaptée. En effet, une Poisson mixte lui sera souvent préférée afin de prendre en compte cette hétérogénéité résiduelle. Notre démarche peut néanmoins expliquer ce résultat puisque nous ne considérons que les polices avec au moins un sinistre, toutes celles non sinistrées étant ignorées. La loi de Poisson s'adapte correctement sur la partie positive des sinistres. En revanche, nous pouvons supposer qu'en incluant la partie nulle, un modèle prenant en compte la sur-dispersion sera nécessaire. Nous considérerons ce cas dans la partie suivante.

Pour la suite, nous conservons tout de même ce modèle de Poisson mixte pour la fréquence, afin de faciliter la comparaison avec les modèles à effets aléatoires qui l'utilisent également. Pour la sévérité, où la distribution Gamma mixte est encore plus adaptée, nous la conservons également dans notre modèle. Les coefficients du modèle de fréquence sont présentés dans la table 3.2.16 et ceux de la sévérité dans la table 3.2.17. Les effets aléatoires suivent une distribution normale avec $\sigma^2_\xi = 0.1082$ pour la fréquence et $\sigma^2_\psi = 0.503^{***}$ pour la sévérité.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	-1,6747	0,1431	-11,70	<.0001	***
Area 2	0,2884	0,0770	3,75	0,0002	***
Area 3	0,1657	0,0877	1,89	0,0589	.
VehBrand 2	0,8474	0,0655	12,93	<.0001	***
Region 2	-0,5311	0,0964	-5,51	<.0001	***
Region 3	-0,2040	0,0880	-2,32	0,0204	*
Region 4	0,1531	0,0826	1,85	0,0640	.
Region 5	0,4131	0,1180	3,50	0,0005	***
DrivAge(25,50]	-0,3614	0,0951	-3,80	0,0001	***
DrivAge(50,100]	-0,2039	0,1053	-1,94	0,0528	.
BonusMalus(65,100]	0,5250	0,0667	7,88	<.0001	***
BonusMalus(100,230]	0,8818	0,1196	7,37	<.0001	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.2.16 – Modèle dépendant en deux parties - Poisson mélange

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	6,7006	0,03704	180,88	<.0001	***
Area 2	-0,2365	0,06214	-3,81	0,0001	***
VehPower(6,13]	0,0277	0,02167	1,28	0,2017	
VehPower(13,15]	-0,1888	0,11070	-1,71	0,0882	.
VehAge(7,10]	-0,0987	0,02797	-3,53	0,0004	***
VehAge(10,100]	-0,2045	0,02559	-7,99	<.0001	***
DrivAge(35,70]	0,1548	0,02631	5,89	<.0001	***
DrivAge(70,100]	0,3537	0,05312	6,66	<.0001	***
VehBrand 2	0,0315	0,02868	1,1	0,2722	
VehBrand 3	-0,0990	0,02900	-3,42	0,0006	***
Region 2	0,1127	0,02940	3,83	0,0001	***
Region 3	0,1893	0,02992	6,33	<.0001	***
Region 4	0,0602	0,02743	2,19	0,0282	*
BonusMalus(60,75]	0,1796	0,03107	5,78	<.0001	***
BonusMalus(75,100]	0,2300	0,02982	7,72	<.0001	***
BonusMalus(100,120]	0,3505	0,07181	4,88	<.0001	***
BonusMalus(120,230]	0,5260	0,10540	4,99	<.0001	***
Fréquence moyenne	0,2283	0,02598	8,79	<.0001	***
Fréquence élevée	0,4183	0,02591	16,15	<.0001	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.2.17 – Modèle dépendant en deux parties - Gamma mélange

3.3 Sélection d'un modèle final

Afin de comparer les modèles et de choisir celui le plus adapté à nos données, nous procédons entre autres par simulations. Ces dernières sont utilisées pour comparer la perte totale prédite par chaque modèle avec la perte observée sur le portefeuille. Comme nous ne cherchons pas à obtenir une formule finale pour la construction d'un tarif, nous avons préféré l'approche par simulation plutôt que celle se basant sur les formules de l'espérance, celle ci nous donnant une distribution complète de la perte totale.

Comme nous nous sommes jusqu'à présent concentrés sur la modélisation de la dépendance entre la fréquence et la sévérité et non sur la construction d'un tarif complet, nous n'avons considéré que la partie du portefeuille ayant déjà été sinistrée jusqu'ici. Dorénavant, afin de prendre en compte l'impact sur la tarification, nous allons devoir considérer le portefeuille complet, y compris les polices pour lesquelles aucun sinistre n'a été observé. Nous rappelons que pour la calibration de nos modèles, nous avons utilisé des distributions mixtes de Poisson tronquées en 0 (ZTP) pour la fréquence et de Gamma pour la sévérité. En théorie, le paramètre estimé de la ZTP devrait s'appliquer sur l'ensemble du portefeuille à l'aide de la loi de Poisson complète, incluant donc également les 0.

Afin de valider la distribution de Poisson, nous réalisons quelques simulations initiales à l'aide du modèle en deux parties indépendant. En effectuant ceci, nous nous rendons rapidement compte que la loi de Poisson appliquée à l'ensemble du portefeuille surestime largement la perte totale. Sur l'ensemble du portefeuille, nous trouvons des montants de perte totale jusqu'à dix fois supérieurs aux montants observés. Il semble donc que le modèle de Poisson ne parvienne pas à expliquer la masse d'observations sans sinistre. En utilisant le modèle de Poisson, nous nous attendons à avoir 71'506 polices sinistrées sur les 658'754 qui composent le portefeuille alors que le nombre de polices sinistrées observé est quant à lui de 14'801. Cet écart important est illustré par la Figure 3.3.1 et explique les problèmes rencontrés lors de la prédiction de la perte totale.

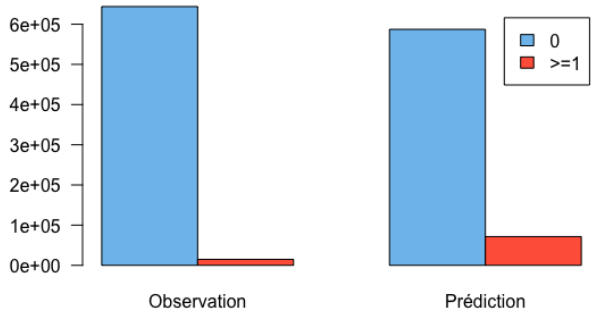


FIGURE 3.3.1 – Prédiction Poisson vs observation

Nous avions précédemment conclu qu'il n'y avait pas de sur-dispersion significative pour le nombre de sinistres quand nous considérions uniquement la partie strictement positive des sinistres. Cette conclusion ne semble pas être valide lorsque que nous considérons l'ensemble du portefeuille, si bien que le modèle Poisson adapté à la partie du portefeuille ayant au moins un sinistre n'est pas capable de capturer la masse d'observations en 0 sur l'ensemble des polices.

Pour cette raison, nous utiliserons un modèle à obstacle. Tout d'abord, afin de prédire la masse d'observations en 0 et de prendre en compte la sur-dispersion, nous calibrons un modèle binomial afin de prédire si la police est sinistrée ou non. Pour les polices ainsi sinistrées, nous utilisons les modèles de fréquence et de sévérité présentés précédemment. Plus exactement, nous prédisons le nombre de sinistres à l'aide d'une distribution conditionnelle ZTP mixte et leur montant à l'aide d'une distribution Gamma mixte. Nous effectuons la calibration du modèle binomial à l'aide des variables utilisées pour le modèle de fréquence. Plus précisément, nous avons :

```
table_totale_Pol$ClaimNb_dummy=0
table_totale_Pol$ClaimNb_dummy[table_totale_Pol$ClaimNb > 0] = 1
glm_binom <- glm(ClaimNb_dummy ~ Area_modif+VehBrand_modif+[...]+
   offset(Exposure),data=table_totale_Pol,family=binomial(link="log
   "))
```

Les coefficients obtenus sont donnés dans la table 3.3.1. Ce modèle permet de nous rapprocher du nombre de polices sinistrées observé. Nous nous attendons ainsi à en observer 8'292 avec au moins un sinistre. Ce nombre est sensiblement inférieur au nombre effectivement observé. Cette différence provient probablement de la courte exposition de certaines polices sinistrées. En effet, un certain nombre de polices avec au moins un sinistre ont une très courte durée d'exposition. Pour des observations de ce type, notre modèle a tendance à leur attribuer une sinistralité nulle.
Cette première observation semble soutenir l'existence d'une hétérogénéité résiduelle, qui est une hypothèse sous-jacente à l'utilisation d'effets aléatoires dans les distribution marginales. Pour rappel, ceux-ci sont notamment utilisés dans le modèle à effets aléatoires corrélés afin de modéliser la dépendance.

Avant d'effectuer les simulations pour le modèle dépendant en deux parties, nous devons tout d'abord faire un choix. En effet pour un assuré, sa prime est calculée en fonction de sa fréquence annuelle observée. Dès lors, la question de la tarification d'un nouvel entrant se pose. Nous choisissons d'utiliser la fréquence annuelle moyenne prédite de la classe de risque à laquelle l'assuré est rattaché. En effet, celle-ci est censée représenter son niveau de risque. Pour la simulation sur le portefeuille nous utilisons la fréquence prédite par le modèle de Poisson.

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	-4.456272	0.040174	-110.923	< 2e-16	***
Area 2	0.269088	0.021075	12.768	< 2e-16	***
Area 3	0.089304	0.022849	3.908	9.29e-05	***
VehBrand 2	-0.509011	0.024203	-21.031	< 2e-16	***
Region 2	-0.008458	0.025151	-0.336	0.7367	
Region 3	0.239456	0.025213	9.497	< 2e-16	***
Region 4	0.253120	0.026032	9.724	< 2e-16	***
Region 5	0.113835	0.041510	2.742	0.0061	**
DrivAge(25,50]	-0.228026	0.031272	-7.292	3.06e-13	***
DrivAge(50,100]	-0.260459	0.034715	-7.503	6.25e-14	***
BonusMalus(65,100]	0.400960	0.020231	19.819	< 2e-16	***
BonusMalus(100,230]	1.226460	0.047045	26.070	< 2e-16	***

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 3.3.1 – Modèle à obstacle - Binomiale

Nous avons ensuite effectué 5'000 simulations pour chaque modèle afin d'obtenir la distribution de la perte totale prédite par chacun d'entre eux. Ci-dessous, nous reprenons à titre d'exemple l'algorithme de simulation utilisé sur R pour le modèle en deux parties indépendant.

```
val_simul_indep <- c()
for(i in 1:5000) {
  #Simule polices avec sinistres
  u1 <- runif(nrow(table_totaleMerge))
  val_binom <- qbinom(u1,1,(table_totaleMerge$Binom_claim*
     table_totaleMerge$Exposure))

  #Simule nombre de sinistre pour chacune
  u2 <- runif(sum(val_binom))
  u4 <- runif(sum(val_binom))
  y1 <- rnorm(u4,mean=0,sd=sigma_nb) #Simule effets aleatoires pour
      frequence

  lambda = table_totaleMerge$Freq_indep[which(val_binom>0)]*
     table_totaleMerge$Exposure[which(val_binom>0)]*exp(y1)
  val_ztp <- qztpois(u2,lambda)

  #Pour chaque sinistre: simule la valeur
  u5 <- runif(sum(val_binom))
  y2 <- rnorm(u5,mean=0,sd=sigma_amount) #Simule effet alea pour
     severite
```

```
18
19    mu = table_totaleMerge$Sev_indep[which(val_binom>0)]*exp(y2)
20    rep_sev <- rep(mu,val_ztp)
21    u3 <- runif(length(rep_sev))
22    val_gamma <- qgam(u3,rep_sev,1/phi)
23
24    #Garde la somme de tous les sinistres
25    val_tot = sum(val_gamma)
26    val_simul_indep = c(val_simul_indep,val_tot)
27  }
```

Sur la figure 3.3.2 ainsi que dans la table 3.3.2, nous reprenons les différentes pertes estimées exprimées relativement à la perte totale observée.

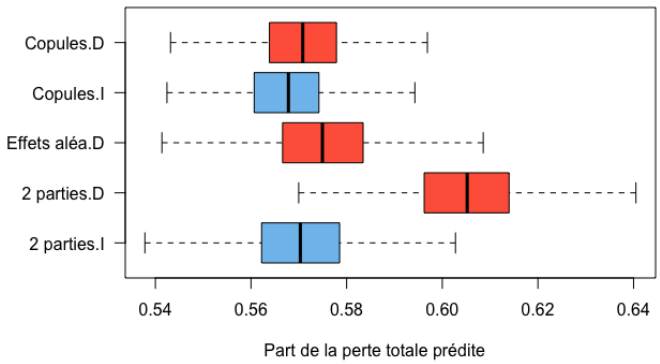


FIGURE 3.3.2 – Perte totale modèles indépendants (I) et dépendants (D)

Modèle	Indépendant	Dépendant
En deux parties	14'473'966	15'358'263
	(307'833)	*(331'071)*
Copules	14'405'191	14'485'804
	(245'913)	*(254'873)*
Effets aléatoires	14'473'966	14'596'015
	(307'833)	*(315'909)*

TABLE 3.3.2 – Perte totale estimée et écart-type de prédiction (5'000 simulations)

Tout d'abord, nous remarquons que pour chacun des modèles, la perte totale estimée est sensiblement inférieure à la perte réelle de 25'376'331€. Trois explications peuvent être données pour cela. Premièrement, étant donné le nombre limité de variables explicatives disponibles, il est probable que le modèle ne soit pas capable de décrire toute la nuance des

données. Deuxièmement, il y a un grand nombre de polices avec une petite exposition ayant pour certaines un sinistre. Comme observé lors de l'introduction du modèle binomial, les modèles ont plus de mal à prédire un ou plusieurs sinistres pour ces cas avec une exposition très faible. Enfin, nous avons retiré les sinistres extrêmes lors de notre modélisation (> 20'000 €). Ces derniers contiennent probablement des informations impactant à la hausse les estimations pour les autres polices.
Il est également intéressant de noter que l'introduction de la dépendance permet pour chaque modèle de rapprocher la prédiction et la perte totale effectivement observée. Cet effet est le plus marqué pour le modèle en deux parties.

Nous remarquons tout de même que le modèle dépendant en deux parties donne la valeur totale prédite la plus proche de la perte effectivement observée. Ce modèle prédit en effet environ 60% de la perte totale du portefeuille. Nous vérifions sur la figure 3.3.3 si cette observation est valide à un niveau plus granulaire. Pour cela, nous comparons les valeurs prédites avec celles observées pour chaque classe de risque, pour les trois modèles en deux parties qui sont relativement proches. À quelques classes près, le modèle dépendant en deux parties produit une estimation plus proche des montants observés. La différence par rapport aux autres modèles est d'autant plus marquée que la perte totale observée de la classe est importante.

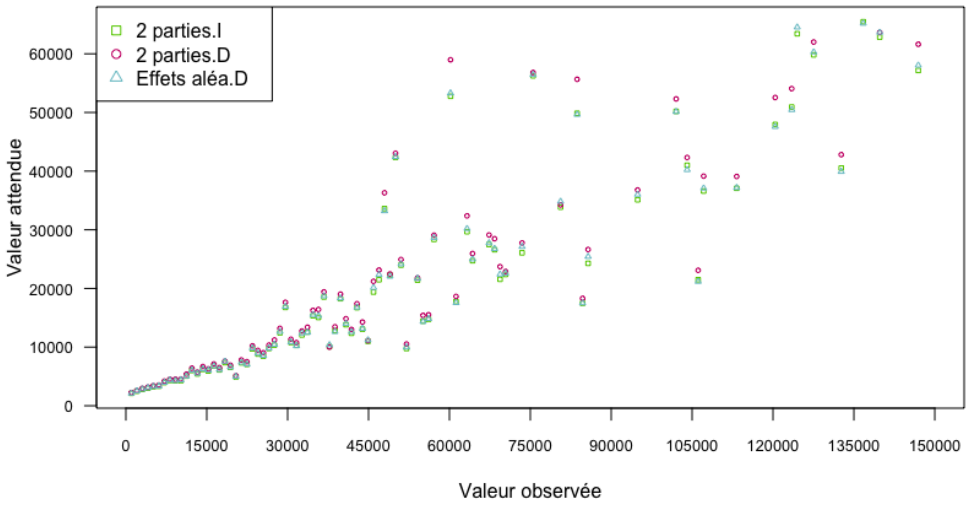


FIGURE 3.3.3 – Perte observée vs attendue

L'introduction de la dépendance a également un effet sur la volatilité des observations. Dans les modèles incluant la dépendance, la valeur finale de la perte totale est légèrement plus variable. Celle-ci reste néanmoins raisonnable et n'impacte pas excessivement la qualité des résultats de la simulation.

Un autre critère qu'il est intéressant de comparer afin de choisir un modèle est l'AIC. Ce dernier nous donne une indication sur le modèle le plus adéquat pour nos données. Nous reprenons dans la table 3.3.3 les différents AIC obtenus.

Modèle	Indépendant	Dépendant
En 2 parties	255'108	254'853
Copules	257'691	255'986
Effets aléatoires	255'108	255'100

TABLE 3.3.3 – AIC obtenus pour chaque modèle

Tout d'abord, comme observé précédemment, nous remarquons que l'introduction de la dépendance dans les modèles a systématiquement comme effet de faire diminuer l'AIC.
Le modèle de copules utilisant des données de sévérité moyenne, nous ne pouvons pas comparer directement ses valeurs avec celles des autres modèles.
Comme noté dans la partie précédente l'ajout des effets aléatoires a un impact très important sur l'AIC. Cette présence d'hétérogénéité résiduelle nous a amené à considérer des distributions mixtes pour la sévérité et la fréquence. Leur inclusion a permis d'abaisser l'AIC de 12'044 par rapport au modèle indépendant en deux parties utilisant des distributions simples. Cet AIC diminue à nouveau lorsque nous considérons la dépendance, pour atteindre 254'853. Cette valeur est la plus faible observée à travers tous les modèles.
Selon ce critère, il semble que le modèle dépendant en deux parties soit le plus adapté.

Néanmoins, la sélection d'un modèle ne s'effectue pas uniquement sur la valeur d'indicateurs chiffrés. Il est intéressant d'étudier les propriétés mêmes des modèles.

Pour le modèle dépendant en deux parties, une des principales limitations que nous avons identifiée provient de la structure de la dépendance. En effet, cette méthode permet de capturer une dépendance linéaire, ce qui peut s'avérer trop restrictif dans certains cas. En revanche, étant très semblable au modèle indépendant en deux parties classique, l'avantage de ce modèle est qu'il reste facile à appliquer. De plus, ce modèle est particulièrement intéressant lors de l'étude de l'impact de la dépendance sur la tarification. En effet, ce type de modèle permet d'adapter de manière intuitive la prime de l'assuré en fonction de son expérience. Ainsi, l'expérience relative à la fréquence propre de l'assuré aura un impact sur la prédiction de la sévérité à l'aide des paramètres liant nombres et montants des sinistres dans la régression.

Concernant le modèle de copules, celui-ci utilise des données agrégées. Comme nous avons à disposition la valeur de chacun des sinistres, cette perte d'information diminue l'intérêt du modèle. Bien que l'utilisation de données agrégées n'ait pas d'impact sur la valeur estimée des paramètres lors de l'utilisation d'une Gamma, l'estimation du paramètre

de dispersion s'en trouve altérée. Ce dernier est entre autres utilisé lors des simulations. Nous préférons donc les modèles utilisant toutes les informations à disposition. De plus, afin de réaliser des calculs et simulations, ce modèle nécessite beaucoup de temps par rapport aux autres modèles. L'avantage de l'utilisation des copules est que ces dernières permettent la modélisation de structures de dépendance complexes.

Dans le modèle à effets aléatoires corrélés, nous pouvons modéliser une structure de dépendance plus complexe qu'une relation simplement linéaire. Celle-ci reste tout de même plus limitée que ce que permettent les copules. Par rapport aux autres modèles, celui-ci permet d'isoler une dépendance qui ne serait qu'apparente et due aux variables ignorées lors de sa construction. Ceci pourrait être adapté dans notre cas, étant donné le peu de variables explicatives à notre disposition. Concernant la fréquence, nous avons néanmoins pu observer qu'il n'y avait pas d'évidence d'une hétérogénéité résiduelle significative. Ce type de modèle semble moins pertinent dans ce cas là, la dépendance étant fonction de paramètres non significativement différents de zéro.

Le modèle dépendant en deux parties est celui qui permet de prédire la perte totale la plus proche de celle observée. C'est également celui-ci qui a l'AIC le plus faible. Pour ces raisons, ainsi que pour sa facilité d'utilisation et ses propriétés concernant la tarification a posteriori, nous choisissons de conserver le modèle dépendant en deux parties afin d'analyser l'impact de la dépendance sur la tarification du produit d'assurance.
En effet, il semble que la structure de dépendance permise par celui-ci, bien que limitée, soit suffisante dans notre cas. Notons que la structure de dépendance choisie peut être étendue. Il serait possible, par exemple, de choisir plus de catégories pour la fréquence annuelle ou alors d'utiliser des GAM afin de décrire une relation plus linéaire.

Le fait que ce modèle soit préféré à un modèle décrivant la dépendance à l'aide des effets aléatoires corrélés a également un impact sur l'interprétation de celle-ci. En effet, si la dépendance n'avait été qu'apparente, due à l'hétérogénéité résiduelle causée par des variables non considérées lors de l'étude, nous pourrions nous attendre à ce que le modèle à effets aléatoires corrélés donne les meilleurs résultats. Dans le portefeuille considéré, ce n'est pas le cas. Il semble donc y avoir une dépendance réelle entre les deux grandeurs.

Chapitre 4

Impact sur la tarification

4.1 Tarification a priori

Au-delà de la sélection du modèle, il est intéressant de relever que chaque modèle décrit l'existence d'une dépendance positive entre la sévérité et la fréquence. Nous avons vu précédemment que lors de l'inclusion de la dépendance dans le modèle, ce sont essentiellement les coefficients concernant la sévérité qui sont impactés. L'interprétation de la dépendance se fait donc probablement dans le sens de l'impact de la fréquence sur la sévérité. Ainsi, un conducteur qui cause davantage d'accidents aura, en moyenne, des sinistres plus coûteux.

Cette dépendance positive implique donc que les montants des sinistres augmentent en moyenne avec le nombre annuel moyen de sinistres de la police.
Au terme de leur recherche, Czado et al. 2012 fournissent quelques éléments pouvant expliquer l'existence d'une telle dépendance. Tout d'abord, celle-ci peut provenir de l'existence d'une franchise. En effet, si le contrat est assortit d'une franchise annuelle, il est probable qu'elle ne s'applique que sur les premiers sinistres, la capacité étant épuisée pour les sinistres suivants. L'assureur payera donc pour les premiers sinistres un montant en moyenne inférieur à celui payé pour les suivants. La seconde explication possible est plutôt d'ordre comportemental. Nous pouvons effectivement supposer qu'un assuré ayant de multiples sinistres a, en général, une conduite plus risquée. Ce comportement peut également avoir un impact à la hausse sur la sévérité, les accidents occasionnés étant plus graves.
Park et al. 2018 étendent quant à eux leur recherche en incorporant une explication d'ordre comportementale basée sur le système de Bonus-Malus (BMS). Ils montrent que la dépendance entre les deux variables d'intérêt peut provenir du comportement du conducteur expliqué par sa "soif de Bonus". Celle-ci pouvant inclure une dépendance positive en fonction de sa position sur l'échelle de Bonus-Malus. En effet, des assurés situés dans les niveaux inférieurs de celle ci, ont tendance à déclarer des sinistres de plus en plus importants.

Au-delà de l'observation sinistre par sinistre, ce qui intéresse particulièrement l'assureur, c'est l'impact de cette dépendance lorsque le portefeuille entier est considéré. Comme nous pouvons le voir sur la figure 3.3.2, cette dépendance positive augmente la perte prédite ainsi que la variabilité de cette dernière. Il semble donc dangereux pour l'assureur de l'ignorer, puisque ceci le mènerait à sous-estimer la perte de son portefeuille et le risque que celle-ci représente. En effet, une police avec plusieurs sinistres subit une double augmentation de la sinistralité. D'une part, celle qui est évidente, provenant de l'accumulation de plusieurs sinistres, et d'autre part, celle provenant de l'augmentation de la sinistralité moyenne. L'augmentation décrite par le modèle en deux parties sur l'ensemble du portefeuille est relativement importante. L'ignorer pourrait amener l'assureur à faire une perte de 886'297€ sur son tarif technique. Ce qui représente 6% du tarif technique donné par le modèle indépendant. Nous comprenons dès lors que la non-prise en compte de la dépendance par l'assureur peut avoir un impact important sur son résultat et produire d'importantes déviations par rapport à ses prévisions.

Cette déviation déjà relativement importante au niveau du portefeuille global, n'est pas forcément répartie de façon homogène entre les différentes classes de risques. Il est en effet possible que nous observions des compensations entre différentes classes et des écarts plus sévères pour certaines. Afin de pouvoir analyser les variations à un niveau plus granulaire, nous comparons pour chacun des profils de risque la prime pure donnée par le modèle indépendant à celle donnée par le modèle dépendant.

Afin d'obtenir des tendances plus claires nous étudions ces écarts en fonction du niveau de risque de chaque profil. Pour estimer ce dernier, nous nous basons sur la fréquence et la sévérité prédite par le modèle indépendant. Nous représentons sur la Heatmap à la figure 4.1.1 la valeur des écarts en fonction de la sévérité et la fréquence prédite. La table 4.1.1 reprend la valeur des écarts représentés dans la Heatmap. Ceci nous permet d'avoir une vision claire de l'impact de l'hypothèse d'indépendance en fonction des niveaux de risque.

	Freq1	Freq2	Freq3	Freq4	Freq5
Sev5	-2%	3%	5%	6%	13%
Sev4	-2%	5%	6%	7%	12%
Sev3	-1%	4%	6%	8%	12%
Sev2	-1%	4%	6%	8%	13%
Sev1	-3%	3%	7%	11%	16%

TABLE 4.1.1 – Impact en tenant compte de la dépendance (données Heatmap)

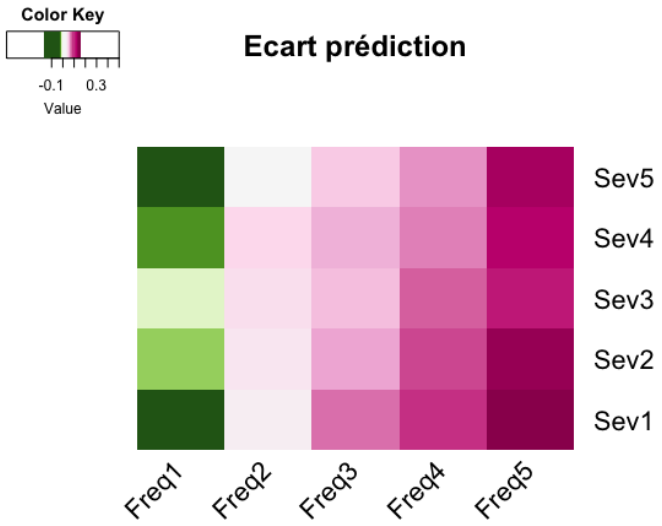


FIGURE 4.1.1 – Comparaison des prédictions du modèle indépendant vs dépendant

L'axe des abscisses reprend les fréquences annuelles prédites par le modèle en deux parties indépendant, allant de la moins élevée (Freq1) à la plus élevée (Freq5). Afin de construire les cinq classes, nous utilisons les quantiles de la distribution empirique de la fréquence. Ainsi, nous nous assurons que chaque case de la Heatmap contient suffisamment d'observations. L'axe des ordonnées reprend, quant à lui, la moyenne des sévérités prédites par le modèle en deux parties indépendant, de la moins élevée (Sev1) à la plus élevée (Sev5). La répartition entre les classes se fait également selon les quantiles de la distribution.

Lorsque la case est de couleur verte, le modèle dépendant prédit en moyenne une perte totale inférieure à celle prédite par le modèle indépendant. Au contraire, la couleur rose indique les cas pour lesquels le modèle dépendant prédit une perte totale plus élevée.

Nous remarquons immédiatement que l'inclusion de la dépendance entraîne une augmentation de la prime pure pour la majorité des classes de risque. Ceci est en ligne avec l'observation faite au niveau du portefeuille. La Heatmap illustre également que cette augmentation ne se fait pas de manière uniforme à travers tout le portefeuille. Alors que l'étude globale révélait une augmentation de prime de +6%, nous voyons grâce à celle-ci que la variation de prime est comprise entre -3% et +16% selon les profils.

Cette hétérogénéité apporte également un facteur d'aggravation pour l'assureur. En effet, plus les profils sont risqués, plus l'augmentation nécessaire de la prime est importante. Cet effet est particulièrement visible lorsque la fréquence espérée de la classe de risque augmente. Il y a un effet clair d'augmentation de la prime pour chaque niveau de fréquence. Cette augmentation est particulièrement importante sur les sévérités les plus faibles et reste relativement uniforme pour les autres. La construction du modèle en deux parties

dépendant peut expliquer cette tendance claire. En effet, dans cette paramétrisation, la sévérité espérée dépend directement et positivement de la fréquence annuelle prédite.

Nous avons étudié dans cette partie l'impact de l'existence de la dépendance sur la tarification a priori. Cependant, il serait trop réducteur de se limiter à ce point de vue. En effet, cette dépendance entraîne des changements tant pour la prime d'un nouvel entrant que pour la prime a posteriori ajustée sur la base de la sinistralité personnelle.

4.2 Tarification a posteriori

4.2.1 Système de Bonus-Malus classique

En assurance automobile, la norme afin d'adapter la prime d'un assuré en fonction de son expérience est de se baser sur le système de Bonus-Malus (BMS). Ce processus de tarification a posteriori est utilisé pour fixer la prime de la prochaine période du contrat en fonction de l'historique des sinistres de l'assuré. Ainsi, la prime pure déterminée par la classe de risques à laquelle il appartient est ajustée à l'aide d'un Bonus ou d'un Malus en fonction de la position de l'assuré sur l'échelle du système.

En France, la prime pure correspond au coefficient 100 sur l'échelle. Celle-ci varie ensuite entre 50 (le bonus maximum) et 350 (le malus maximum). Les règles de changement de niveau sur l'échelle sont définies exclusivement en fonction du nombre de sinistres observés sur l'année.

De ce fait, le but du système de Bonus-Malus est de diminuer le montant des primes en l'absence d'accident (le bonus) afin de récompenser les bons conducteurs. Inversement, en cas de sinistres, l'assureur augmente le montant des primes (le malus) pour pénaliser les mauvais conducteurs.

Ce système ne se base que sur la fréquence afin de faire évoluer la prime de l'assuré. Il suppose ainsi implicitement l'indépendance entre la fréquence et la sévérité des sinistres, en prenant comme hypothèse que tous les sinistres sont en moyenne de même taille. Nous avons mis en évidence, pour le portefeuille considéré, l'existence d'une dépendance positive entre fréquence et sévérité. Si celle-ci n'est pas prise en compte, cela pourrait mener à des ajustements de primes non représentatifs du risque réel, en pénalisant, par exemple, par un coefficient trop faible les polices multi-sinistrées.

Il pourrait être possible de prendre en compte dans le système de Bonus-Malus actuel la dépendance en calibrant les coefficients de manière à ce qu'ils prennent en compte l'impact de cette dépendance. Il existe également d'autres méthodes permettant de prendre celle-ci en compte.

4.2.2 Méthodes prenant la dépendance en compte

Afin de proposer le tarif le plus juste possible, étant donné l'existence de ce lien positif entre les deux composantes de la sinistralité, l'assureur devrait prendre en compte à la fois le nombre de sinistres et leur taille.

Après avoir démontré l'existence d'une dépendance, plusieurs études proposent des méthodes pour adapter le processus de tarification a posteriori en assurance automobile.
Par exemple, dans leur papier, Oh, Shi et al. 2019 proposent un ajustement du système BMS classique. Tout d'abord, ils utilisent un modèle à effets aléatoires bivariés basé sur les copules, afin de tenir compte à la fois de la dépendance entre le nombre et la taille des sinistres ainsi que de l'hétérogénéité résiduelle. En ajustant leur modèle sur la sévérité moyenne puis individuelle, les résultats montrent l'existence d'une dépendance négative.
Les auteurs proposent ensuite d'ajuster les relativités optimales du système, qui définissent les règles de passage d'un niveau à l'autre, afin de prendre en compte la dépendance entre la fréquence des sinistres et la sévérité. Ils fournissent ainsi des formules fermées permettant d'appliquer ce modèle relativement flexible. Ils observent ainsi que la relativité optimale augmente lorsque la dépendance entre la fréquence et la sévérité des sinistres augmente.

Une autre possibilité pour ajuster la prime a posteriori est de se baser sur la crédibilité, qui est une généralisation du système de Bonus-Malus. Celle-ci redéfinit les paramètres de la distribution de la fréquence et de la sévérité en fonction de l'expérience de l'assuré.
Dans ce cadre là, Hernández-Bastida et al. 2009 étudient par exemple l'impact de l'existence d'une dépendance sur la prime collective ainsi que sur la prime de Bayes, utilisée pour la tarification a posteriori. Comme nous pouvons nous y attendre, cette prime sera impactée par la valeur de la corrélation entre la fréquence et sévérité.
Les auteurs trouvent ainsi, que même des niveaux de dépendance faibles entraînent une forte variation de la prime de Bayes. Ainsi, dans le cas d'une dépendance positive, dès que la police est sinistrée, celle-ci augmente sensiblement la valeur de la prime a posteriori. Ignorer cette dépendance semble donc à nouveau être dangereux pour l'assureur, car ceci pourrait l'amener à sous-évaluer encore davantage les primes futures.

Un système de tarification a posteriori semblable pourrait être appliqué dans le cadre du modèle dépendant en deux parties que nous avons sélectionné. Tout d'abord, la fréquence annuelle historique de l'individu est prise en compte à l'aide des paramètres concernant la fréquence présents dans le modèle de sévérité. Il est ensuite possible d'ajuster les distributions respectives de la fréquence et de la sévérité. Pour cela, un ajustement des effets aléatoires présents dans les distributions marginales mixtes respectives peut être envisagé.

Néanmoins, le développement d'une tarification a posteriori complète dépassant le cadre de notre mémoire, nous nous concentrons ici sur l'impact de l'existence de la dépendance positive sur la tarification a priori.

Conclusion

Alors qu'en général, les modèles utilisés lors de la tarification font l'hypothèse d'indépendance entre fréquence et sévérité, nous avons étudié dans le cadre de cette recherche plusieurs modèles permettant de relâcher cette hypothèse en incorporant de la dépendance entre les deux variables. Tout d'abord, nous avons commencé par étendre le modèle en deux parties en incorporant la fréquence annuelle comme variable explicative de la sévérité. Afin de permettre une structure de dépendance plus complexe, nous avons ensuite présenté un modèle liant les deux distributions marginales à l'aide de la copule de Gauss. Pour finir, et afin de prendre en compte la dépendance apparente induite par l'hétérogénéité résiduelle, nous avons construit un modèle prenant en compte la dépendance à l'aide d'effets aléatoires corrélés entre les deux composantes de la sinistralité.

Bien que tous les modèles décrivent une dépendance positive, nous avons conclu que le modèle dépendant en deux parties correspondait le mieux à nos données. Grâce à ce modèle, nous avons pu observer l'impact de cette dépendance pour l'assureur. Si ce dernier ne la prend pas en compte, cela peut avoir des effets néfastes sur son tarif. Tout d'abord, cela l'amène à sous-estimer la perte attendue ainsi que son risque. Cette sous-estimation est d'autant plus problématique, qu'elle semble être plus grave pour les cas les plus risqués. De plus, lors du suivi d'un assuré cela l'amènera sans doute à sous-estimer l'ajustement nécessaire de la prime a posteriori.

Ces résultats sont toutefois à nuancer. Tout d'abord, du fait que nous ne possédons que peu de polices sinistrées, la validation des modèles s'est effectuée sur les mêmes données utilisées pour le construire. L'idéal aurait été de construire un test set et un validation set distincts, ce qui n'a pu être fait ici par manque d'observations. De plus, dans la littérature nous avons pu observer que les résultats étaient très variables selon la base de données utilisée. Il serait ainsi intéressant de calibrer nos modèles sur une nouvelle base de données. Le modèle ainsi sélectionné et le signe de la dépendance pourrait varier.

Avec des données de plusieurs années consécutives disponibles, nous trouvons qu'il serait intéressant d'étendre notre recherche afin d'inclure une dimension longitudinale dans la dépendance. Cela permettrait, par exemple, d'étendre nos résultats en développant un modèle plus abouti de tarification a posteriori. Par exemple, une fois le constat sur la dépendance fait, certaines publications s'intéressent plus en détails à son impact sur la

tarification a posteriori des polices. En assurance automobile, la norme est l'application d'un système de bonus-malus (BMS) basé sur le nombre de sinistres déclarés par l'assuré et ne prenant pas en compte une éventuelle dépendance. Des articles comme celui de Oh, Shi et al. 2019 proposent donc une transformation du système de bonus-malus classique permettant d'intégrer cette dimension. Pour cela, ils proposent l'intégration d'effets aléatoires corrélés permettant de tenir compte de l'hétérogénéité résiduelle entre les assurés, ce qui impacte les relativités optimales dans le système de bonus-malus proposé. Leur méthode a l'avantage de présenter un modèle relativement simple tout en intégrant un degré de flexibilité plus élevé.
Le but de notre mémoire n'était pas de développer un modèle de tarif complet, nous n'avons donc pas appliqué cette méthode. Il resterait tout de même intéressant de savoir quel pourrait être l'impact d'une telle dépendance sur la vie des contrats en pratique.

Annexe A

Regroupements des variables explicatives

Nom variable	Détail composition	Nb observations
`Area 1`	A, B, F	3'677
`Area 2`	E, D	7'191
`Area 3`	C	3'933
`VehBrand 1`	B1, , B2, B3, B4, B5, B6, B10, B11, B13, B14	12'678
`VehBrand 2`	B12	2'123
`Region 1`	Alsace, Aquitaine, Franche-Comté, Haute-Normandie, Nord-Pas-de-Calais, Midi-Pyrenées, Limousin, Champagne-Ardenne, Poitou-Charentes, Bourgogne, Auvergne	2'613
`Region 2`	Bretagne , Centre , Basse-Normandie	4'363
`Region 3`	Pays-de-la-Loire, Picardie, Basse-Normandie, Rhone-Alpes	3'758
`Region 4`	Provence-Alpes-Cotes-D'Azur, Ile-de-France	3'337
`Region 5`	Languedoc-Roussillon, Corse	730
`DrivAge(18,25]`		1'353
`DrivAge(25,50]`		8'666
`DrivAge(50,100]`		4'482
`BonusMalus(50,65]`		9'948
`BonusMalus(65,100]`		4'382
`BonusMalus(100,230]`		471

TABLE A.0.1 – Regroupements - Modèle de fréquence

Nom variable	Détail composition	Nb observations
Area 1	A, B, C, D, E	14'348
Area 2	F	453
VehPower(4,6]		8'656
VehPower(6,13]		6'015
VehPower(13,15]		130
VehAge(0,7]		8'342
VehAge(7,10]		2'754
VehAge(10,100]		3'705
DrivAge(18,35]		4'472
DrivAge(35,70]		9'629
DrivAge(70,100]		700
VehBrand 1	B1, B2, B3, B11, B14	9'576
VehBrand 2	B10, B12, B13	2'880
VehBrand 3	B4, B5, B6	2'345
Region 1	Alsace, Basse-Normandie, Limousin, Auvergne, Bourgogne, Franche-Comté, Pays-de-la-Loire, Midi-Pyrenées, Centre, Haute-Normandie	5'322
Region 2	Picardie, Poitou-Charentes, Bretagne, Champagne-Ardenne, Languedoc-Roussillon, Aquitaine	2'790
Region 3	Corse, Provence-Alpes-Cotes-D'Azur, Nord-Pas-de-Calais	2'708
Region 4	Ile-de-France, Rhone-Alpes	3'981
BonusMalus(50,60]		9'025
BonusMalus(60,75]		2'119
BonusMalus(75,100]		3'186
BonusMalus(100,120]		325
BonusMalus(120,230]		146

TABLE A.0.2 – Regroupements - Modèle de sévérité

Annexe B

Modèle de copule indépendante (Gauss)

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	-1,62032	0,12974	-12,489	< 2e-16	***
Area 2	0,28819	0,07652	3,766	0,000332	***
Area 3	0,16704	0,08719	1,916	0,063683	.
VehBrand 2	0,84400	0,06474	13,037	< 2e-16	***
Region 2	-0,53140	0,09598	-5,536	8,82E-08	***
Region 3	-0,20434	0,08745	-2,337	0,026026	*
Region 4	0,15328	0,08200	1,869	0,069540	.
Region 5	0,40870	0,11666	3,503	0,000862	***
DrivAge(25,50]	-0,35895	0,09410	-3,814	0,000276	***
DrivAge(50,100]	-0,20053	0,10416	-1,925	0,062519	.
BonusMalus(65,100]	0,52420	0,06618	7,921	9,49e-15	***
BonusMalus(100,230]	0,88179	0,11796	7,475	2,93e-13	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

TABLE B.0.1 – Modèle de copule indépendante de Gauss - Fréquence

Coefficients :	Estimate	Std. Error	z value	Pr(>|z|)	Sign.
(Intercept)	7,20029	0,03623	198,761	< 2e-16	***
Area 2	-0,30910	0,07002	-4,414	2,34e-05	***
VehPower(6,13]	0,04831	0,03158	1,530	0,12377	
VehPower(13,15]	-0,10087	0,03230	-3,123	0,00304	**
VehAge(7,10]	0,11748	0,03273	3,589	0,00064	***
VehAge(10,100]	0,23842	0,03324	7,173	2,68e-12	***
DrivAge(35,70]	0,05101	0,03054	1,671	0,09883	.
DrivAge(70,100]	0,07800	0,02916	2,675	0,01114	*
VehBrand 2	0,24357	0,05910	4,121	8,19e-05	***
VehBrand 3	-0,08700	0,03117	-2,791	0,00812	**
Region 2	-0,16665	0,02843	-5,862	1,38e-08	***
Region 3	0,05111	0,02408	2,123	0,04192	*
Region 4	-0,23326	0,12374	-1,885	0,06750	.
BonusMalus(60,75]	0,17411	0,03467	5,022	1,33e-06	***
BonusMalus(75,100]	0,26056	0,03290	7,920	9,54e-15	***
BonusMalus(100,120]	0,30394	0,08031	3,784	0,00031	***
BonusMalus(120,230]	0,48044	0,11765	4,084	9,55e-05	***

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

TABLE B.0.2 – Modèle de copule indépendante de Gauss - Sévérité

Annexe C

Code R

cfr. le fichier "NGUYEN_08091501_SPOORENBERG_62401800_2020_Annexe-C.zip"

Annexe D

Code SAS

cfr. le fichier "NGUYEN_08091501_SPOORENBERG_62401800_2020_Annexe-D.zip"

Bibliographie

Abramowitz, Milton et Irene A Stegun (1972). « Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing. » In :

Baumgartner, Carolin (2014). « Comparison of Multivariate Mixed and Copula Models for the Estimation of Losses ». In :

Baumgartner, Carolin et al. (2015). « Bayesian total loss estimation using shared random effects ». In : *Insurance : Mathematics and Economics* 62, p. 194-201.

Boudreault, Mathieu et al. (2006). « On a risk model with dependence between interclaim arrivals and claim sizes ». In : *Scandinavian Actuarial Journal* 2006.5, p. 265-285.

Cossette, Hélène et al. (2019). « Collective risk models with dependence ». In : *Insurance : Mathematics and Economics* 87, p. 153-168.

Czado, Claudia et al. (2012). « A mixed copula model for insurance claims and claim sizes ». In : *Scandinavian Actuarial Journal* 2012.4, p. 278-305.

Denuit, Michel et al. (2019). « Concordance-based predictive measures in regression models for discrete responses ». In : *Scandinavian Actuarial Journal* 2019.10, p. 824-836.

Frees, Edward et al. (2016). « Multivariate frequency-severity regression models in insurance ». In : *Risks* 4.1, p. 4.

Garrido, José et al. (2016). « Generalized linear models for dependent frequency and severity of insurance claims ». In : *Insurance : Mathematics and Economics* 70, p. 205-215.

Gschlössl, Susanne et Claudia Czado (2007). « Spatial modelling of claim frequency and claim size in non-life insurance ». In : *Scandinavian Actuarial Journal* 2007.3, p. 202-225.

Hernández-Bastida, A et al. (2009). « The net Bayes premium with dependence between the risk profiles ». In : *Insurance : Mathematics and Economics* 45.2, p. 247-254.

Jeong, Himchan et al. (2017). « Generalized linear mixed models for dependent compound risk models ». In : *Available at SSRN 3045360*.

Krämer, Nicole et al. (2013). « Total loss estimation using copula-based regression models ». In : *Insurance : Mathematics and Economics* 53.3, p. 829-839.

Lee, Gee Y et Peng Shi (2019). « A dependent frequency–severity approach to modeling longitudinal insurance claims ». In : *Insurance : Mathematics and Economics* 87, p. 115-129.

Lee, Woojoo et al. (2019). « Investigating dependence between frequency and severity via simple generalized linear models ». In : *Journal of the Korean Statistical Society* 48.1, p. 13-28.

Lundberg, Filip (1903). *I. Approximerad framstallning af sannolikhetsfunktionen : II. Aterforsakring af kollektivrisker*. Almqvist & Wiksell.

Oflaz, Zarina Nukeshtayeva et al. (2019). « Aggregate Claim Estimation Using Bivariate Hidden Markov Model ». In : *ASTIN Bulletin : The Journal of the IAA* 49.1, p. 189-215.

Oh, Rosy, Jae Youn Ahn et al. (2019). « On copula-based collective risk models ». In : *arXiv preprint arXiv :1906.03604*.

Oh, Rosy, Peng Shi et al. (2019). « Bonus-Malus premiums under the dependent frequency-severity modeling ». In : *Scandinavian Actuarial Journal*, p. 1-24.

Park, Sojung C et al. (2018). « Does hunger for bonuses drive the dependence between claim frequency and severity ? » In : *Insurance : Mathematics and Economics* 83, p. 32-46.

Pinheiro, José C et Douglas M Bates (1995). « Approximations to the log-likelihood function in the nonlinear mixed-effects model ». In : *Journal of computational and Graphical Statistics* 4.1, p. 12-35.

Powell, Michael James David (1977). « Restart procedures for the conjugate gradient method ». In : *Mathematical programming* 12.1, p. 241-254.

Schulz, Juliana (2013). « Generalized linear models for a dependent aggregate claims model ». Thèse de doct. Concordia University.

Shi, Peng et al. (2015). « Dependent frequency–severity modeling of insurance claims ». In : *Insurance : Mathematics and Economics* 64, p. 417-428.