

INSTITUT STATISTIQUE, BIOSTATISTIQUE ET SCIENCES
ACTUARIELLES



LACTU2910: MÉMOIRE PROJET

Modélisation des frais médicaux en accident du travail - Etude de la fréquence de sinistres ouvriers à l'aide des méthodes GLM, GLMM et Poisson-LogNormale.

Mémorant:
Elhadî LATELIER

Promoteur:
Michel DENUIT
Co-Promoteur:
Thibault MONTFORT

Contents

1	Remerciements	3
2	Introduction	4
3	Préliminaires	5
3.1	Facteurs de risque	5
3.2	Segmentation des risques et choix d'une catégorie de référence	6
3.3	Modèles linéaires et problèmes classiques	7
3.4	La famille exponentielle	8
3.5	Exemples de distributions de la famille exponentielle	8
3.6	Distribution de Poisson Mixte	10
3.7	Modèle de Poisson-LogNormale	11
3.8	Modèle linéaire généralisé (GLM)	12
3.9	Modèle linéaire généralisé mixte (GLMM)	13
3.10	Choix d'un modèle	16
3.10.1	La déviance du modèle GLM	16
3.10.2	l'AIC des différents modèles	17
4	Modélisation	17
4.1	Description de la garantie frais médicaux	17
4.2	Description de la base de données	18
4.3	Récolte des données	20
4.4	Limites de l'analyse	21
4.5	Sélection d'une distribution théorique	22
4.6	Analyse exploratoire	23
4.6.1	Analyse univariée	24
4.6.2	Analyse bivariée	29
4.7	Discussion et choix de l'exposition au risque	33
4.8	Modèles de fréquence	35
4.8.1	GLM Poisson	37
4.8.2	GLMM Poisson	46
4.8.3	Poisson - Lognormal	53
5	Sélection du meilleur modèle	55
6	Conclusion finale	57
7	Annexes	59
7.1	Analyse Anova - Groupements potentiels	59
7.2	Boxplots employés	60
7.3	Analyse univariée des employés	62
7.4	Sur-dispersion	64
7.5	Sur-dispersion	65
7.6	Sur-dispersion	66
7.7	Consistence des estimateurs	66
7.8	Fréquences de variables catégorielles	67

1 Remerciements

Je remercie le Professeur Michel Denuit pour sa disponibilité, son écoute et son encadrement très qualitatif pour la rédaction de ce mémoire.

Je remercie la Banque Nationale de Belgique pour la mise à disposition de tout le matériel nécessaire au développement de ce modèle et au bien-être quotidien dans mes missions professionnelles.

Je remercie particulièrement Valérie Woit et Thibault Monfort pour leur encadrement et leurs disponibilités dans le cadre de ce mémoire. Je tiens à remercier également Monsieur Luc Kaiser qui suite à une présentation d'un premier modèle, m'a proposé de travailler au sein de son équipe.

Je remercie ma mère Rafki Zahra, ma tante Najia Tabrouni et mon père Latelier Fernand pour leurs sacrifices, leur amour et leurs enseignements parfois bien cachés de ce qu'ils savent de la vie.

Je remercie également la personne qui partage actuellement ma vie, Marie Scrève, première personne capable de m'accepter totalement comme je suis. Ayant un mental de gagnante, elle supporte aisément mes longues périodes de travail isolé.

Je remercie également Flavian Iqbal, ancien élève de l'UCL, Actuaire aujourd'hui, qui me fera découvrir l'actuariat en 2018 dans une chambre d'étudiant et m'aidera à remettre en question ma façon d'aborder les mathématiques de manière générale.

Je remercie également mes autres amis et colocataires, capable d'accepter mon indisponibilité constante, tout en me donnant toujours leur soutien.

2 Introduction

Le présent mémoire est le fruit de mon travail en tant que stagiaire à la Banque Nationale de Belgique. La branche Accidents du Travail fait l'objet d'une attention particulière par le régulateur en raison de ses spécificités (marché très compétitif, différentes formes d'indemnisation possible en fonction de la gravité de l'accident, sinistralité dépendante du statut du travailleur, ...)

Ce mémoire présente différentes méthodes pour l'étude de la fréquence de sinistres ouvriers, sur le lieu du travail, spécifiquement pour la garantie frais médicaux. Nous utiliserons les modèles linéaires généralisés (GLM, massivement utilisés dans les applications actuarielles en raison de leurs propriétés permettant de modéliser diverses distributions pertinentes telles que Poisson pour le nombre de sinistres et Gamma pour la sévérité de ceux-ci par exemple) et les modèles linéaires généralisés mixtes (GLMM, qui constituent une extension des modèles GLM classiques).

De manière générale, Les GLM traitent les variables explicatives catégorielles à travers un certain nombre d'effets fixes ($n-1$) découlant de la binarisation des différents niveaux (n) de la variable initiale. Cet aspect de traitement des variables présente l'inconvénient que pour une variable explicative catégorielle à n niveaux, le modèle produira $n-1$ effets fixes distincts pour expliquer la variable réponse, ce qui pourrait entraîner une sur-paramétrisation du modèle et par conséquent augmenter sa variance. En pratique, la ré-étude d'un autre jeu de données avec le même GLM fournira des résultats (coefficients estimés) bien différents. De plus, les effets fixes capturent uniquement l'effet commun à toutes les observations qui appartiennent à un niveau donné d'une variable explicative.

Le GLM fait donc l'hypothèse que les classes d'individus dont nous voulons prédire la sinistralité sont homogènes. Or, il pourrait y avoir des variables non observées expliquant une variabilité résiduelle au sein d'une classe supposée homogène. Les GLMM permettent de résoudre ces deux problèmes (grand nombre de paramètres en raison d'une variable catégorielle avec un grand nombre de niveaux et une hétérogénéité résiduelle dans les classes d'individus) à travers l'introduction d'un ou plusieurs effets aléatoires, assignant à la variable catégorielle une variable aléatoire (et remplaçant donc les effets fixes liés à chaque niveau par un effet aléatoire). Nous pourrions également traiter des variables catégorielles hiérarchiques en traduisant la hiérarchie par des effets aléatoires "emboîtés". Cela nous permettra d'utiliser le code NACE (code à cinq chiffres destiné à classer les activités d'une entreprise où chaque chiffre précise l'activité de l'entreprise au sein du chiffre qui le précède) en exploitant son caractère hiérarchique, grâce à quelques effets aléatoires emboîtés (autant que de niveaux considérés, donc avec un maximum de 5), plutôt que des centaines de niveaux.

Dans un premier temps les méthodes GLM et GLMM seront comparées afin de démontrer qu'il existe deux manières différentes d'estimer les effets d'une variable explicative sur la réponse. Dans cette première analyse, la sinistralité passée de l'entreprise ne sera pas incluse et nous tenterons de réconcilier ces deux modèles. Dans un second temps, nous soulignerons l'utilité pratique du GLMM en ajoutant davantage de positions NACE et en intégrant la sinistralité passée, démontrant ainsi l'indispensabilité de cette méthode dans le cas qui nous intéresse. Enfin, nous challengerons les GLMM à l'aide d'un modèle Poisson-LogNormale qui lui intégrera un effet aléatoire à l'aide d'un second paramètre de distribution, destiné à modifier le paramètre de la distribution classique de Poisson.

La première partie de ce mémoire présentera les concepts théoriques nécessaires aux différentes méthodes de modélisation. Ensuite, une deuxième partie développera les différentes étapes de la modélisation ainsi qu'une analyse détaillée des résultats obtenus. Après discussion, nous terminerons ce mémoire par le choix d'un modèle, considéré comme étant le plus adapté à la problématique étudiée.

3 Préliminaires

3.1 Facteurs de risque

Les facteurs de risque sont des éléments qui influencent le risque de manière directe et indirecte. Parmi ceux-ci, l'assureur n'a bien souvent pas la possibilité de prendre en compte l'entièreté de l'information qui explique la réponse modélisée car il existe en général une part de l'information, dite cachée, à laquelle il ne pourra jamais avoir accès. Cette non accessibilité résulte de facteurs qui peuvent être mesurables mais trop onéreux pour les avoir à sa disposition ou encore des facteurs qui ne peuvent simplement pas se mesurer : En accidents du travail par exemple, le nombre d'accidents ou encore le coût de ceux-ci, peuvent être influencé par des informations telles l'inaptitude particulière d'un travailleur avec certaines machines, l'esprit d'équipe d'un travailleur, le sens des responsabilités et plus encore...

Lors de l'établissement d'un tarif, l'assureur devra alors se contenter des informations qu'il peut avoir à sa disposition. Dans notre cas, ces informations peuvent par exemple être le secteur d'activité de l'entreprise, la taille ou encore la répartition ouvrier-employés.

Bien que l'assureur dispose de ces informations, il n'est aucunement établi qu'elles constituent des variables qui influencent de manière systématique la variable dépendante étudiée. Il se peut que les variables à disposition influencent indirectement la réponse en raison du fait qu'elles sont corrélées avec les véritables variables qui l'influencent, considérées comme trop coûteuses ou non-mesurables.

(Denuit et al., 2019) définit formellement ces variables comme :

- $\mathbf{X}$, un vecteur qui rassemble les facteurs de risque observables et utilisés par l'assureur (ayant ou non une relation causale avec la réponse Y)
- $\mathbf{X}^+$, un vecteur des facteurs de risque cachés qui impactent la réponse Y . Ceux-ci peuvent être corrélés ou non avec le précédent vecteur $\mathbf{X}$.

En conséquence, nous pouvons considérer que la prime calculée par l'assureur sera

$$\mu(\mathbf{X}) = E[Y|\mathbf{X}] \quad (1)$$

où μ est l'espérance conditionnelle de la réponse sachant les informations $\mathbf{X}$ récoltées par l'assureur. En réalité, la prime qui devrait normalement être demandée s'il disposait de toutes les informations est plutôt de la forme

$$\mu^+(\mathbf{X}, \mathbf{X}^+) = E[Y|\mathbf{X}, \mathbf{X}^+] \quad (2)$$

où μ^+ contient cette fois toutes les informations cachées, en plus des informations disponibles, pour estimer Y .

3.2 Segmentation des risques et choix d'une catégorie de référence

La classification des risques est une étape essentielle du processus de modélisation. Elle permet à l'assureur de transformer un portefeuille hétérogène en un ensemble de classes de risques homogènes distinctes, dont la prime est la mesure du risque encouru par un individu appartenant à cette classe. De cette manière, l'assureur augmente la maîtrise de ses prestations et se retrouve en capacité de demander une prime plus juste à chacun des assurés. (Denuit al. ,2019)

Toutefois, une classification parfaite du risque est impossible car l'assureur réalise celle-ci uniquement sur base des informations dont il dispose. En ce sens, les informations cachées permettraient à l'assureur s'il en disposait d'affiner sa segmentation et de la rendre plus adaptée au risque réel de chaque assuré.

La variance de Y peut se décomposer comme suit

$$\text{Var}[Y] = \underbrace{\text{E}[\text{Var}[Y|\mathbf{X}]]}_{\text{Variance supportée par l'assureur}} + \underbrace{\text{Var}[\text{E}[Y|\mathbf{X}]]}_{\text{Variance supportée par l'assuré}} \quad (3)$$

où

- Le premier terme est la partie de la variance absorbée par l'assureur. Après avoir créé des classes de risques homogènes, il subsiste une variabilité au sein de chacune des classes qui sera compensée par l'assureur.
- Le second terme est la variance supportée par l'assuré. Celui-ci paie une prime qui est fonction de son profil de risque.

En décomposant le premier terme de droite, on a

$$\text{Var}[Y] = \underbrace{\text{E}[\text{Var}[Y|\mathbf{X}, \mathbf{X}^+]]}_{\text{Effet de mutualisation}} + \underbrace{\text{E}[\text{Var}[\text{E}[Y|\mathbf{X}, \mathbf{X}^+|\mathbf{X}]]]}_{\text{Effet de classification imparfaite}} + \underbrace{\text{Var}[\text{E}[Y|\mathbf{X}]]}_{\text{Variance supportée par l'assuré}} \quad (4)$$

L'équation (4) permet décomposer la variance supportée par l'assureur en deux mécanismes différents qui permettent de supporter la variabilité de Y.

- L'effet de mutualisation, lorsque les variations de Y sont dues entièrement à de l'aléatoire, tous les assurés paient une prime moyenne identique. Alors la mutualisation de ces primes au sein d'un même portefeuille, permet de compenser cette variabilité.
- La classification imparfaite : Au sein d'un ensemble de classes de risques déterminées par la matrice $\mathbf{X}$, il existe des assurés qui paient une prime supérieure et d'autres qui paient une prime inférieure à la valeur de la prime qu'ils devraient réellement payer. Cette différence provient de la matrice $\mathbf{X}^+$. L'assuré qui paie une prime supérieure à sa propre sinistralité compensera l'assuré qui paie une prime inférieure à sa sinistralité.

3.3 Modèles linéaires et problèmes classiques

(James et al., 2017) discutent des outils statistiques nécessaires à la construction de modèles prédictifs. Parmi eux, les modèles linéaires constituent une approche élémentaire pour modéliser une relation entre une variable dépendante, classiquement surnommée variable "réponse" et des variables indépendantes, appelées variables "explicatives". Ce modèle suppose que la "vraie" relation entre la variable réponse et les variables explicatives est de forme :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \quad (5)$$

où

- β_0 est un paramètre appelé intercept qui modélise la valeur moyenne de la réponse sans tenir compte des variables explicatives
- β_j modélise l'effet de la variable explicative j sur la réponse, pour $j = 1, \dots, p$
- Y représente la variable réponse que l'on cherche à expliquer.
- $\mathbf{x}_{ij}$ les réalisations du vecteur $\mathbf{X}_j$ pour $j = 1, \dots, p$
- ϵ est un terme d'erreur tel que $\epsilon \sim N(0, \sigma^2)$.

Graphiquement, nous pouvons représenter notre modèle linéaire par une surface de dimension p dont le point de départ est un coefficient β_0 traversant un nuage d'observations Y_i , pour $i = 1, \dots, n$. Nous estimons alors notre modèle réel, qui est supposé représenter la véritable relation entre Y et X , par un modèle estimé de forme :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p \quad (6)$$

où

- $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_j$ sont respectivement les estimateurs des paramètres réels $\beta_0, \beta_1, \dots, \beta_j$.
- $\hat{Y}$ représente l'estimation de la réponse Y que l'on cherche à estimer.
- $\mathbf{x}_{ij}$ les réalisations du vecteur $\mathbf{X}_j$ pour $j = 1, \dots, p$

Les modèles linéaires reposent sur les hypothèses suivantes :

1. Les couples (Y_i, X_i) sont iid mais la distribution de probabilité conditionnelle $Y_i | X_i = x$ varie avec les réalisations de x . De plus, pour chaque réalisation de X_i , les différentes réalisations de Y_i (dispersées autour de l'équation de régression du modèle linéaire) sont indépendantes et de loi Gaussienne.
2. En considérant que le terme d'erreur est une variable aléatoire telle que $\epsilon \sim N(0, \sigma^2)$, on a $Y_i \sim N(\beta_0 + \sum_{j=1}^p \beta_j X_{ij}, \sigma^2)$
3. La variance de chaque distribution des Y_i est constante, quelles que soient les valeurs prises par les variables explicatives. En conséquence, la variance du terme d'erreur est également constante (homoscédasticité).
4. Il existe une relation linéaire entre la variable dépendante Y et les variables explicatives X . Par linéaire on entend, linéaire en β : Par exemple, $Y = \beta_1 X_1 + \beta_2 X_1 X_2$ est linéaire mais pas $Y = \beta_1 X_1 + \log(\beta_2 X_2)$
5. Les termes d'erreurs sont indépendants entre eux. Précisément, $\text{cov}(\epsilon_i, \epsilon_j) = 0$, pour i différent de j .

En pratique, les modèles linéaires classiques présentent une limite majeure : On suppose que la réponse suit une loi Normale. Cette hypothèse n'étant pas toujours adaptée pour les applications actuarielles (on a besoin de variable de comptages qui ont un domaine discret), nous avons besoin de nous intéresser à des modèles qui relâchent cette hypothèse. Les sections suivantes nous permettent d'explorer une telle piste.

3.4 La famille exponentielle

La famille exponentielle est une famille de distributions dont la densité de probabilité (cas continu) ou la fonction de masse de probabilité (cas discret) s'exprime comme :

$$f(y; \theta, \phi) = c(y, \phi) \exp \left\{ \frac{y\theta - a(\theta)}{\phi} \right\} \quad (7)$$

avec

- θ le paramètre canonique.
- ϕ le paramètre dispersion.
- $a(\theta)$ est une fonction convexe de θ
- $c(\cdot)$ est une fonction qui ne dépend pas de θ et définie de manière à ce que l'intégrale de f sur son domaine soit égale à 1.

A différents choix de θ et $a(\theta)$ correspondent différentes distributions dont nous donnerons quelques exemples à la section suivante.

Les moments d'ordre 1 et 2 auront la forme générale exprimée par les equations (5) et (6) :

$$E(y) = a'(\theta) \quad (8)$$

$$Var(y) = \phi a''(\theta) \quad (9)$$

où $a'(\theta)$ et $a''(\theta)$ sont respectivement les dérivées première et seconde.

3.5 Exemples de distributions de la famille exponentielle

Nous donnons dans cette section des exemples de distributions de la famille exponentielle pertinentes pour la suite de nos travaux, à savoir, pour la modélisation du nombre de sinistres, la loi de Poisson et la loi Binomiale Négative.

Pour le comptage des sinistres, sous une hypothèse d'équi-dispersion ($E=Var$), on utilise classiquement une loi de Poisson de paramètre μ :

$$Y \sim \text{Poi}(\mu) \quad (10)$$

dont la fonction de probabilité est de la forme :

$$p(y) = \frac{e^{-\mu} \mu^y}{y!}, y = 0, 1, 2 \quad (11)$$

Les moments d'ordre 1 et 2 sont donnés par:

$$\begin{aligned} E(y) &= \mu \\ Var(y) &= \mu \end{aligned}$$

En posant

$$\begin{aligned} \theta &= \ln(\mu) \\ \mu &= e^\theta \\ a(\theta) &= e^\theta \\ \phi &= 1 \end{aligned}$$

On voit que la distribution de Poisson appartient à la famille des distributions exponentielles.

L'une des hypothèses fondamentale d'une distribution de Poisson est l'équi-dispersion des données représentée par l'égalité entre la moyenne et la variance des données. En pratique, cela signifie qu'à mesure où la moyenne du nombre de sinistres augmentera, la variabilité de ce nombre de sinistres augmentera aussi.

Empiriquement, nous ne trouverons jamais une moyenne et une variance strictement égales, toutefois, si nous souhaitons nous assurer que l'hypothèse d'une distribution de Poisson est adaptée à la variable dépendante que l'on cherche à expliquer, nous devons accorder toute notre attention à ce critère et démontrer que la variance empirique est raisonnablement proche de notre moyenne empirique. Plus ce critère d'équi-dispersion sera respecté, plus nous aurons la confirmation que le choix de la distribution de Poisson pour nos données est adapté.

Lorsque que les données ne sont pas équi-dispersées, il est possible que celles-ci soient sur-dispersées. Dans un tel cas, le passage suivant présente une distribution possible adaptée à un tel cas. Dans les cas où on observe une surdispersion des données :

$$E(y) < Var(y) \quad (12)$$

On doit se tourner vers une distribution plus adaptée. (De Jong et Heller, 2008) propose l'utilisation de la loi Binomiale Négative dont la fonction de probabilité a la forme

$$f(y) = \binom{r+y-1}{r-1} \pi^r (1-\pi)^y \quad (13)$$

avec

- π la probabilité de succès
- r un paramètre, entier positif ($r \in \mathbb{N}^+$)
- $y = 0, 1, 2, \dots$

Afin de gagner en flexibilité, nous souhaiterions étendre le paramètre r à toutes les valeurs réelles positives. Pour ce faire, nous allons généraliser le coefficient Binomial à l'aide d'une fonction Gamma :

$$f(y) = \frac{\Gamma(y+r)}{y! \Gamma(r)} \pi^r (1-\pi)^y \quad (14)$$

Enfin, définissons deux nouveaux paramètres qui vont s'avérer utiles pour caractériser notre distribution par la suite :

$$\mu = \frac{r(1-\pi)}{\pi}$$

$$\kappa = \frac{1}{r}$$

En utilisant ces paramètres, la fonction de densité de la loi Binomiale Négative devient

$$f(y) = \frac{\Gamma(y + \frac{1}{\kappa})}{y! \Gamma(\frac{1}{\kappa})} \left(\frac{1}{1 + \kappa\mu} \right)^{1/\kappa} \left(\frac{\kappa\mu}{1 + \kappa\mu} \right)^y \quad (15)$$

et on a

$$E(y) = \mu$$

$$V(y) = \mu(1 + \kappa\mu)$$

où le paramètre κ modélise la sur-dispersion. On note

$$y \sim NB(\mu, \kappa) \quad (16)$$

On peut prouver que

$$Si \kappa \rightarrow 0, NB(\mu, \kappa) \rightarrow Poi(\mu)$$

En posant

$$\begin{aligned} \theta &= \ln \frac{\mu}{(1 + \kappa\mu)} \\ a(\theta) &= -\frac{1}{\kappa} \ln(1 - \kappa e^\theta) \end{aligned}$$

On voit que la distribution Binomiale Négative appartient à la famille exponentielle. Les moments sont alors donnés par :

$$\begin{aligned} E(y) &= \frac{e^\theta}{1 - \kappa e^\theta} \\ Var(y) &= \frac{e^\theta}{(1 - \kappa e^\theta)^2} \end{aligned}$$

ou encore

$$\begin{aligned} E(y) &= \mu \\ Var(y) &= \mu(1 + \kappa\mu) \end{aligned}$$

3.6 Distribution de Poisson Mixte

Dans la section précédente, nous avons pu découvrir la distribution de Poisson, classiquement utilisé pour le traitement des variables de comptages, très pratique dans le cadre d'une modélisation puisqu'elle dispose des propriétés de la famille exponentielle. Toutefois, la distribution de Poisson est régulièrement considérée comme très contraignante en raison de son hypothèse rigide d'équi-dispersion. Or, la réalité assurantielle étant tout autre, la présence d'hétérogénéité au sein des portefeuilles d'assurance est régulièrement rencontrée. Dès lors, la distribution classique de Poisson peut présenter certaines limites en omettant de considérer les aspects individuels de chacun. En accident du travail, nous chercherons donc à considérer les différences individuelles qu'il existe entre les entreprises d'une même section, d'une même division...

Dès lors (Pitrebois, Denuit, Marechal et Walhin ,2007) proposent l'inclusion d'un effet aléatoire positif, noté Θ , multiplicateur de la fréquence de sinistres de Poisson initiale, noté λ en tant que stratégie envisageable pour prendre en considération l'hétérogénéité dont il est question. Nous obtiendrons alors une fréquence de sinistres λ , capable de varier en fonction de Θ . L'objectif étant de considérer que toutes les entreprises du portefeuille n'ont pas la même fréquence de sinistres.

La fréquence du portefeuille conditionnelle au paramètre Θ est donnée par

$$Pr[N = k | \Theta = \theta] = p(k | \lambda\theta) = \exp(-\lambda\theta) \frac{(\lambda\theta)^k}{k!}, k = 0, 1, \dots \quad (17)$$

A l'aide du paramètre θ ci-dessus, nous pouvons aisément ajuster la fréquence de sinistres des assurés :

- $\theta > 1$: Augmentation de la fréquence moyenne λ
- $\theta < 1$: Diminution de la fréquence moyenne λ

en posant $E(\Theta)=1$, on retrouve la fréquence moyenne du portefeuille. En conséquence, il vient que le nombre de sinistres générés par une entreprise prise aléatoirement au sein de notre portefeuille, suivra une distribution mixte de Poisson.

Dès lors, la fonction de masse de probabilité associée à la densité mixte de Poisson est

$$Pr[N = k] = \mathcal{E}(p(k|\lambda\theta)) = \int_0^\infty \exp(-\lambda\theta) \frac{(\lambda\theta)^k}{k!} dF_\Theta(\theta) \quad (18)$$

avec $F_\Theta(\theta)$, la fonction de distribution de Θ .

avec $F_\Theta(0) = 0$

En développant, il vient que :

$$Pr[N = k] = \mathcal{E}(p(k|\lambda\theta)) = \int_0^\infty \exp(-\lambda\theta) \frac{(\lambda\theta)^k}{k!} f_\Theta(\theta) d\theta \quad (19)$$

En imposant la fonction de densité du paramètre Θ , soit $f_\Theta(\theta)$, comme étant la distribution Poisson-LogNormale, c'est à dire :

$$f_\theta(\Theta) = \frac{1}{\theta\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln \theta + \sigma^2/2)^2}{2\sigma^2}\right), \theta > 0 \quad (20)$$

où $\Theta \sim \mathcal{LN}(\mu = -\sigma^2/2, \sigma^2)$

(Pitrebois, Denuit, Marechal et Walhin ,2007) définissent la fonction de masse de probabilité de la distribution de Poisson-LogNormale comme étant de forme

$$Pr[N = k] = \frac{1}{\sigma\sqrt{2\pi}} \frac{(\lambda d)^k}{k!} \int_0^\infty \exp(-\lambda d\theta) \theta^{k-1} \exp\left(-\frac{(\ln \theta + \sigma^2/2)^2}{2\sigma^2}\right) d\theta \quad (21)$$

avec $E[N] = \lambda$

avec $V[N] = \lambda + \lambda^2(\exp(\sigma^2) - 1)$

3.7 Modèle de Poisson-LogNormale

Selon (Pitrebois, Denuit, Marechal et Walhin ,2007), si nous supposons que les effets aléatoires $\Theta_1, \dots, \Theta_n$ sont des variables aléatoires indépendantes qui suivent une distribution logNormale $\mathcal{LN}(\mu = -\sigma^\epsilon / \epsilon, \sigma^\epsilon)$:

La distribution commune de Θ_i est donné par

$$f_\theta(\Theta) = \frac{1}{\theta\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln \theta + \sigma^2/2)^2}{2\sigma^2}\right), \theta > 0 \quad (22)$$

Sachant que $E[\Theta_i] = 1$ et $V[\Theta_i] = \exp(\sigma^2) - 1$, la probabilité que l'entreprise i réalise k_i sinistres est donné par

$$Pr[N_i = k_i | x_i] = \int_0^\infty \exp(-\lambda_i \theta) \frac{(\lambda_i \theta)^{k_i}}{k_i!} \frac{1}{\theta\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln \theta + \sigma^2/2)^2}{2\sigma^2}\right) d\theta \quad (23)$$

Cette équation de vraisemblance ne possède pas de forme fermée et doit être résolue numériquement.

3.8 Modèle linéaire généralisé (GLM)

(De Jong et Heller, 2008) définissent les modèles linéaires généralisés (GLM) en tant qu'extension des régressions linéaires classiques. En effet, les GLM font l'hypothèse d'une relation linéaire entre une transformation de la moyenne de la réponse, notée μ et permettent de supposer des distributions non plus uniquement Normale mais faisant partie de la famille exponentielle.

Les modèles GLM utilisent une fonction de lien, notée g , qui est une fonction monotone, dérivable, permettant d'établir un lien entre la réponse moyenne et les variables explicatives. Le rôle d'une fonction lien est d'imposer aux valeurs prédites de se conformer à l'échelle des valeurs observées. Par exemple, le nombre de sinistres est positif ou nul. En prenant une fonction de lien log, on transforme cette variable strictement positive, en une variable réelle, ce qui rend plus plausible le fait de la modéliser par une fonction linéaire qui par construction prend toutes les valeurs réelles. Formellement, on a :

$$g(\mu) = \mathbf{x}'\beta \quad (24)$$

$\mathbf{X}'$ est un vecteur transposé et β une matrice. En développant, on obtient :

$$g(\mu) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (25)$$

ou, en inversant:

$$\mu = g^{-1}(\mathbf{x}'\beta) \quad (26)$$

Voici quelques exemples de fonctions lien communément choisies pour diverses distributions:

Distribution	Lien $g(\mu)$
Normal	μ
Binomiale Négative	μ^{-1} ou $\log(\mu)$
Poisson	$\log(\mu)$

Figure 3.1: Fonctions liens associées à différentes distributions possibles de la réponse.

Comme pour la régression linéaire, on veut estimer le modèle. Autrement dit, on veut estimer les β_j , ce qui se fait généralement par maximum de vraisemblance. La fonction de vraisemblance est définie comme le produit des fonctions de densité pour chaque valeur observées :

$$L = L(\theta, \phi) = \prod_{i=1}^n f(y_i; \theta, \phi) \quad (27)$$

La log-vraisemblance est donnée par

$$LL = l(\theta, \phi) = \sum_{i=1}^n \ln f(y_i; \theta, \phi) \quad (28)$$

En considérant une fonction de densité dans la famille exponentielle, nous pouvons préciser la forme de la log-vraisemblance dans ce cas :

$$LL = \sum_{i=1}^n \left\{ \ln c(y_i, \phi) + \frac{y_i \theta - a(\theta)}{\phi} \right\} \quad (29)$$

en dérivant par rapport à θ ,

$$\frac{\partial LL}{\partial \theta} = \frac{\partial \left\{ \sum_{i=1}^n \ln c(y_i, \phi) + \left\{ \frac{y_i \theta - a(\theta)}{\phi} \right\} \right\}}{\partial \theta} = 0 \quad (30)$$

Le premier terme de gauche n'étant pas dépendant de θ disparaît puisqu'il représente une constante, on a alors :

$$\frac{\partial}{\partial \theta} \left\{ \sum_{i=1}^n \frac{y_i \theta - a(\theta)}{\phi} \right\} = \sum_{i=1}^n \frac{y_i - a'(\theta)}{\phi} = \frac{(\sum_{i=1}^n y_i) - n a'(\theta)}{\phi} = 0 \quad (31)$$

En conséquence, on a

$$\implies a'(\theta) = \frac{\sum_{i=1}^n y_i}{n} = \bar{y} \quad (32)$$

En conclusion, l'estimation des paramètres par maximum vraisemblance (MLE) sous l'hypothèse que la distribution de la réponse appartient à la famille exponentielle, permet d'estimer la moyenne d'échantillonnage de nos données par le biais de la moyenne prédite du modèle.

De plus, notons que

- Le GLM se construit comme la somme des caractéristiques systématiques qui interviennent pour expliquer la réponse moyenne (c'est à dire le score) et d'un terme d'erreur aléatoire qui représente les écarts possibles de cette moyenne en fonction des caractéristiques propres à chaque individu. L'assureur considère qu'il compense cette volatilité de la moyenne par la mutualisation des individus au sein d'un même portefeuille (Denuit et al., 2019)
- Les modèles GLM font l'hypothèse que toute l'information nécessaire pour expliquer la réponse est à notre disposition. En réalité, il n'en est rien, il subsistera toujours une part d'information non disponible pour l'actuaire lors de la modélisation. En pratique, l'hypothèse formulée est donc

$$\mathbf{X} = \mathbf{X}^+$$

- Sous un GLM, on fait l'hypothèse que les catégories de risques définies par l'actuaire sont parfaitement homogènes. En réalité, il n'en est rien. Pour cette raison, la section suivante fera l'introduction des modèles mixtes qui utilisent un effet aléatoire pour modéliser l'hétérogénéité résiduelle qui subsiste au sein des différentes classes de risque après homogénéisation de celles-ci.

3.9 Modèle linéaire généralisé mixte (GLMM)

(Denuit et al., 2019) discutent de l'utilité et de la particularité du modèle linéaire généralisé mixte (GLMM). Jusqu'à maintenant, nous avons vu que les GLM font l'hypothèse que les classes de risques sont homogènes. Ce n'est cependant pas toujours le cas : Une partie des caractéristiques non disponibles ($\mathbf{X}^+$) des assurés peut entraîner une différence de profil entre assurés d'une même classe, générant de l'hétérogénéité au sein des classes de risque supposées homogènes.

Les GLMM permettent de tenir compte de cette hétérogénéité à l'aide d'effets aléatoires.

Formellement, (Denuit et al., 2019) définissent les GLMM comme des modèles de la forme:

$$g(\mu_{new}) = \mathbf{x}_i^T \beta + \varepsilon_i \quad (33)$$

où $x_i^T \beta$ a la même signification que dans le cas des GLM et ε_i représente cette fois-ci un effet aléatoire distribué normalement, tel que $\varepsilon_i \sim N(0, \sigma^2)$. En conséquence, la prédiction moyenne sera donné par

$$\mu_{new} = g^{-1}(\mathbf{x}_i^T \beta + \varepsilon_i) \quad (34)$$

L'effet aléatoire ici correspondra à la variable hiérarchique code NACE (traité comme un effet fixe dans le cas des GLM). Le traitement d'une telle variable en tant qu'effet fixe aura pour conséquence d'ajouter un grand nombre de paramètres au modèle, ce qui n'est pas désirable lorsque le nombre de données est limité (risque de sur-paramétrisation du modèle). Dans un tel cas, nous pourrions observer une instabilité dans l'estimation des coefficients assimilable en pratique à des écarts-types respectifs élevés. Cette situation se déclare lorsque le modèle dispose d'un échantillon trop faible d'observations par rapport aux nombres de paramètres qu'il cherche à estimer. En conséquence, plus l'échantillon sera élevé, plus il sera possible de construire un modèle avec un degré de complexité important. L'une des solutions massivement utilisée est le groupement de certains niveaux d'une même variable catégorielle lorsque ceux-ci produisent un effet similaire sur la réponse. Néanmoins, il faudra trouver le meilleur compromis entre la perte d'information potentielle et l'interprétabilité du modèle sur les prédictions (réduction du critère AIC défini à la section 3.8.2). Toutefois, si la formation de certains groupements ont du sens d'un point de vue statistique, il n'est pas certain que ceux-ci aient du sens d'un point de vue business. Par exemple, nous pourrions imaginer que les code NACE 43 et 22 pourraient potentiellement être regroupés car les impacts individuels respectifs de ceux-ci ne diffèrent pas significativement l'un de l'autre mais les entreprises qui appartiennent à ces deux codes appartiennent respectivement à des industries bien différentes... Dans une telle situation, nous devons rester prudents quant au choix des groupements possibles.

Dans ce cas précis, la modélisation du code NACE comme effet aléatoire peut s'avérer être une stratégie très judicieuse qui ne nécessitera aucun groupement ni aucun ajout massif de nombre de paramètres. Enfin, puisque l'effet aléatoire peut se décomposer en une somme d'effets aléatoires relatifs à chaque niveau du code NACE, cela permettra la prise en compte de la structure hiérarchique de cette variable comme suit :

Le Règlement (CE) N° 1893/2006 du Parlement Européen et du Conseil du 20 décembre 2006 établissant la nomenclature des activités économiques belges décrit la structure hiérarchique du code NACE comme suit

- un premier niveau comportant des rubriques identifiées par un code alphabétique (sections),
- un deuxième niveau comportant des rubriques identifiées par un code numérique à deux chiffres (divisions),
- un troisième niveau comportant des rubriques identifiées par un code numérique à trois chiffres (groupes),
- un quatrième niveau comprenant des rubriques identifiées par un code numérique à quatre chiffres (classes).
- un cinquième niveau comprenant des rubriques identifiées par un code numérique à cinq chiffres (sous-classes).

Formellement la décomposition s'établit comme suit :

$$\varepsilon_i = \varepsilon_f + \varepsilon_{fj} + \varepsilon_{fjk} + \varepsilon_{fjkg} + \varepsilon_{fjkgm} \quad (35)$$

où

- ε_f est l'effet aléatoire de la première position de la section à laquelle l'entreprise appartient.

- ε_{fj} est l'effet aléatoire de la deuxième position (division au sein de la section).
 - ε_{fjk} est l'effet aléatoire de la troisième position (groupe au sein de la division).
 - ε_{fjkg} est l'effet aléatoire de la quatrième position (classe au sein du groupe).
 - ε_{fjkgm} est l'effet aléatoire de la cinquième position (Sous-classe au sein de la classe).
- Graphiquement, cela peut se traduire par la figure 3.2

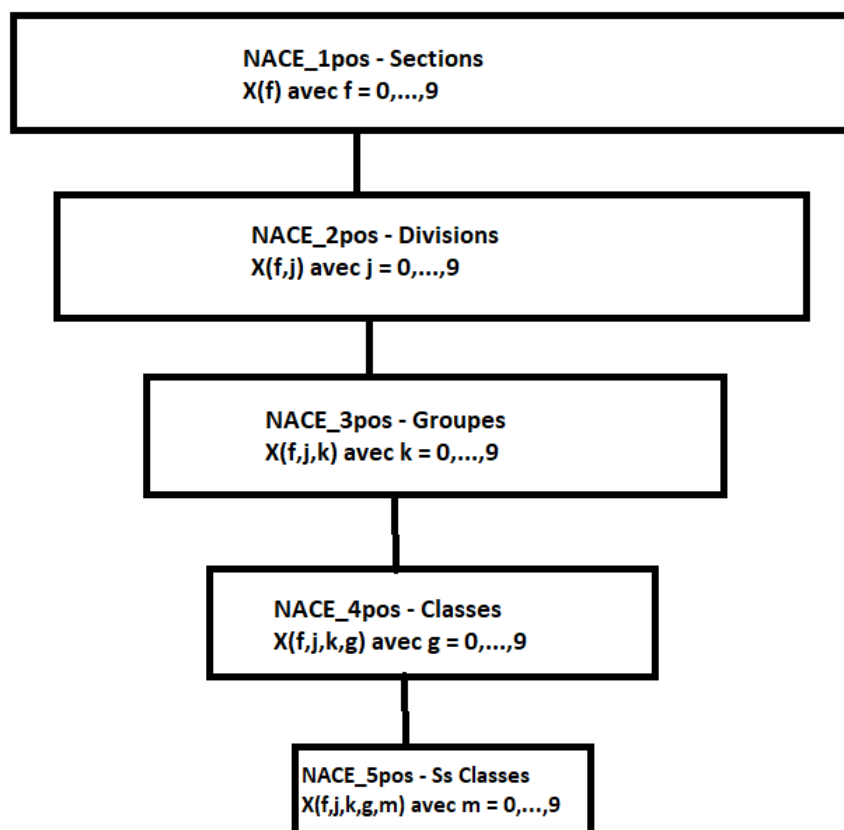


Figure 3.2: Structure du code NACE

3.10 Choix d'un modèle

Dans cette section, nous passons en revue les différents critères que nous utiliserons pour choisir un modèle, en fonction de ses performances et de son interprétabilité. En réalité, un modèle n'est pas meilleur qu'un autre de manière générale, mais bien préféré à un autre sur base de différents critères statistiques (robustesse, qualité des prédictions) et d'utilisation (interprétabilité, simplicité). Ces critères sont notamment présentés par (Denuit et al. (2019))

3.10.1 La déviance du modèle GLM

La déviance est un critère mesurant la qualité d'ajustement d'un modèle sur base d'une comparaison avec un modèle parfaitement ajusté (dont les résidus sont nuls), appelé "modèle saturé". Ce dernier se définit comme le modèle dont le nombre de paramètres est égal au nombre d'observations. Par construction, la valeur prédite par le modèle saturé est donc égale à la valeur observée, pour toutes les observations utilisées pour paramétrer le modèle. Notons cependant que le modèle saturé, qui minimise l'erreur sur les données utilisées pour le paramétrer (training set), est bien entendu sujet à de l'overfitting, impliquant une faible qualité prédictive.

La déviance se calcule en faisant la somme des écarts entre la vraisemblance du modèle étudié et saturé. Mathématiquement on a :

$$\begin{aligned} D(y, \hat{\mu}) &= 2\phi(L_{full} - L(\hat{\beta})) \\ &= 2 \sum_{i=1}^n \nu_i(y_i(\tilde{\theta} - \hat{\theta}) - a(\tilde{\theta}_i) + a(\hat{\theta}_i)) \end{aligned}$$

avec

- $\tilde{\theta}$ l'estimation des paramètres sous le modèle saturé
- $\hat{\theta}$ l'estimation sont les estimations du modèle d'intérêt.

Dès lors, la déviance échelonnée est donnée par :

$$\tilde{D}(y, \hat{\mu}) = \frac{1}{\phi} D(y, \hat{\mu}) \quad (36)$$

Plus le critère de déviance est réduit, moins les écarts entre la log-vraisemblance du modèle étudié et celle du modèle saturé est faible. Cela signifie que plus la déviance est faible, plus le modèle est ajusté aux observations. Toutefois, comme nous l'avons expliqué pour le modèle saturé, le fait de vouloir minimiser la déviance risque de nous pousser à l'overfitting, ce qui nous incite à introduire d'autres mesures qui tendent à limiter le nombre de paramètres, comme l'AIC.

Cette mesure s'appliquera uniquement aux modèles GLM en raison du fait que les coefficients linéaires sont estimés par maximum vraisemblance. De plus, notons que ce critère sera uniquement pertinent pour réaliser des tests de modèles emboîtés, c'est-à-dire des modèles faisant la même hypothèse de distribution et utilisant des ensembles variables explicatives emboîtés les uns dans les autres. Par exemple, deux GLM de Poisson avec un lien logarithmique, utilisant, le premier, la variable X1 et, le deuxième, les variables X1 et X2 sont emboîtés. Par contre, deux GLM de Poisson avec un lien logarithmique, utilisant, le premier, la variable X1 et, le deuxième, les variables X2 et X3 ne sont pas emboîtés.

En revanche, lorsque nous souhaiterons comparer des modèles non emboîtés (c'est-à-dire utilisant des ensembles de variables explicatives qui ne sont pas inclus les uns dans les autres ou faisant des hypothèses de distributions différentes), nous utiliserons d'autres critères, comme l'AIC que nous définissons ci-après.

3.10.2 l'AIC des différents modèles

L'AIC mesure la quantité d'information à laquelle l'actuaire renonce lors de sa modélisation. Naturellement, nous savons que plus nous avons de variables explicatives à disposition, plus nous augmentons l'ajustement aux observations.

L'objectif d'un modèle prédictif est de prédire, avec le plus de précision possible, une réponse pour des données dont on ne dispose pas encore. Pour ce faire, nous sélectionnerons un modèle sur base de critères permettant d'éviter le sur-apprentissage (overfitting), c'est-à-dire le risque que le modèle s'ajuste trop bien aux observations et modélise également une partie du bruit, ce qui nous conduirait à une augmentation de l'erreur de prédiction.

Finalement, l'AIC réalise un compromis entre complexité du modèle (nombre de paramètres) rendant le modèle de plus en plus difficile à interpréter et de plus en plus enclin au sur-apprentissage et le critère de parcimonie statistique et l'interprétabilité du modèle.

Mathématiquement, L'AIC se définit comme

$$AIC = 2k - 2 \ln(\hat{L}) \quad (37)$$

- A mesure que k augmente, le biais d'information entre notre modèle prédictif et le modèle saturé diminue.
- A mesure que la log-vraisemblance augmente, la plausibilité de l'estimation des coefficients de régression augmente selon les données observées.

L'objectif de l'AIC sera de nous aider à nous positionner entre deux extrêmes :

- Un modèle avec trop peu de paramètres, sans capacités prédictives, trop simpliste, ne représentant pas la réalité que l'on cherche à modéliser, avec une déviance extrême.
- Un modèle disposant de trop de paramètres, sujet à de l'overfitting, de déviance nulle.

En conclusion, nous chercherons à minimiser le critère d'AIC car plus nous minimisons cette perte d'information, plus nos prédictions se rapprochent des données observées. D'autre part, l'avantage de l'AIC est qu'il permet la comparaison de modèles ayant des distributions différentes contrairement à la déviance.

4 Modélisation

4.1 Description de la garantie frais médicaux

La loi du 10 avril 1971, régissant les accidents du travail pour les entreprises du secteur privé, définit la garantie frais médicaux telle que la couverture qui permet d'indemniser l'ensemble des frais suivants :

1. médicaux (par exemple, la consultation d'un médecin ou d'un kinésithérapeute)
2. pharmaceutiques (les produits achetés chez le pharmacien)
3. d'hospitalisation (par exemple, une admission à l'hôpital)
4. chirurgicaux (par exemple, une opération)

En conséquence, les entreprises publiques ne feront pas l'objet de notre analyse.

4.2 Description de la base de données

Notre base de données contient 1,710,674 observations. Une observation équivaut à une entreprise au cours d'une année donnée. Nous sélectionnons toutes les entreprises censées être couvertes par un produit accident du travail. Nous traiterons essentiellement le volet ouvriers dans ce mémoire sachant que ces deux volets se traitent de manière distinctes. Pour ce faire, nous associerons à chacun de ces deux volets un training-set et un validation-set. Le training-set contient les années 2009 à 2017 et permettront d'entraîner notre modèle. Ensuite, nous testerons celui-ci sur le validation-set, contenant la sinistralité observée en 2018. La table suivante décrit les variables explicatives que nous avons à disposition.

LDAT2310 - Mémoire projet

Variables	Niveaux	Description
tva	-	Identifiant de l'entreprise
NACE_1pos	0-9 (10 niveaux)	Classification de l'entreprise à 1 degré de précision.
NACE_2pos	01-97 (83 niveaux)	Classification de l'entreprise à 2 degrés de précision. (Sections)
NACE_3pos	011-970 (255 niveaux)	Classification de l'entreprise à 3 degrés de précision.(Divisions au sein des sections)
NACE_4pos	0111-9700 (575 niveaux)	Classification de l'entreprise à 4 degrés de précision.(Groupes au sein des divisions)
NACE_5pos	01110-97000 (871 niveaux)	Classification de l'entreprise à 5 degrés de précision. (Sous groupes au seins des groupes)
c_categ	A1- RF (16 niveau)	Catégorie d'entreprises. "A1", "A2","E13 : types de "grandes sociétés" ; "B1", "B2","B3","C1" ,"C2", "E2" : types de "PME"; "BC", "BL" : types "d'unités TVA"; "H1","H2" ,"H3" ,"H4" : types d'ASBL ; "RF"
Age	-	Age de l'entreprise.
cdy0001323	-	Nombre d'équivalents temps pleins ouvriers.
cdy0001343	-	Nombre d'équivalents temps pleins employés.
cdy0000062	-	Masse salariale de l'entreprise
Nbclaims_lieu	-	Nombre de sinistres sur le lieu du travail par les ouvriers si nous sommes dans le data-set ouvriers et Nombre de sinistres sur le lieu du travail par les employés si nous sommes dans le data-set employés.
Sev_tot_lieu	-	Sévérité totale sur le lieu du travail par les ouvriers si nous sommes dans le data-set ouvriers et sévérité totale sur le lieu du travail par les employés si nous sommes dans le data-set employés
Année	-	Année du fichier

Variables	Description
c_legal	Forme légale de l'entreprise. "003";"006"(S.C. à responsabilité illimité et solidaire (S.C.R.I.S.)); "008"(S.C. à responsabilité limitée (S.C.R.L.)); "009" (S.C.R.L., coopérative de participation (S.C.R.L., C.P.));"010"(SPRL - unipersonnelle (S.P.R.L.U.)) ;"011"(société en nom collectif (S.N.C.));"012"(société en commandite simple (S.C.S.));"013"(société en commandite par actions (S.C.A.)) ;"014"(société anonyme (S.A.)); "015" (société privée à responsabilité limitée (S.P.R.L.)); "016"(société coopérative (ancien statut) (S.C.)); "017"(association sans but lucratif (ASBL)) ;"018" (établissement d'utilité publique);"022"(organisations scientifiques internationales de droit belge);"023"(associations étrangères privées ayant un établissement, une agence, un bureau ou une succursale en Belgique); "025"(SOCIETES AGRICOLES (S. AGR.)); "026"; "027"; "029" ;"030"(ayant un établissement stable en Belgique) ; "060" (groupement d'intérêt économique (G.I.E.) ayant son siège en Belgique) ;"065"(groupement européen d'intérêt économique (G.E.I.E.) ayant son siège en Belgique) ;"108"(S.C. à responsabilité limitée) ; "114"(société anonyme) ;"124"(établissements publics (Loi du 16 mars 1954 - cat.ABCD)); "125"(Association internationale sans but lucratif); "206" ; "208" ;"211"; "212"; "213"; "214"; "215" ;"225" ;"416" ;"506"(S.C. à responsabilité illimité et solidaire (S.C.R.I.S. FS)); "508" (S.C. à responsabilité limitée (S.C.R.L., FS)) ; "513"(société en commandite par actions (S.C.A., FS)); "514"(société anonyme (S.A., FS)); "515" (société privée à responsabilité limitée (S.P.R.L., FS)); "721"

Notons que les différents codes NACE, la catégorie d'entreprise, la forme légale, l'âge de l'entreprise, le nombre de ETP ouvriers/employés et la masse salariale sont les informations dont l'assureur dispose a priori. Afin que l'analyse soit pertinente, nous devons uniquement utiliser des informations disponibles pour l'assureur au moment de la souscription, puisque nous cherchons à établir un modèle de tarification.

4.3 Récolte des données

Les données dont nous disposons consistent en deux fichiers par année calendrier, de 2009 à 2018:

1. Un fichier sinistres contenant l'ensemble des montants versés par accident déclarés et acceptés par l'assureur.
2. Un fichier entreprise qui contient un échantillon représentatif du marché belge avec principalement un nombre d'équivalents temps plein inférieur à 200.

Nous croisons ensuite ces fichiers de manière à disposer d'un fichier unique (ou plutôt deux : un training set et un validation set) dont chaque ligne représente une entreprise, lors d'une année calendrier, avec toutes ses caractéristiques et le nombre de sinistres (éventuellement nul) observés pendant l'année en question.

4.4 Limites de l'analyse

La détermination d'un champ d'application en statistique permet de rationaliser les résultats obtenus à une réalité donnée. Sans cette délimitation, nous pourrions penser à tort que notre analyse est généralisable à une infinité de situations. En pratique, le nettoyage des données représente parfois 80% du travail de l'actuaire et lui permet à la fois de limiter et affiner son analyse. A cette fin, les manipulations suivantes ont été effectuées :

1. Nous commençons par retirer toutes les lignes doublons qui existent au sein de nos fichiers.
2. Nous séparons les sinistres sur le lieu du travail des sinistres sur le chemin du travail. Bien que les sinistres sur le chemin du travail appartiennent à la branche accident du travail, il est préférable de traiter distinctement ces deux catégories de sinistres. En effet, nous pensons que les accidents sur le chemin du travail dépendent des conditions de transport de l'individu et sont donc assimilables à des accidents de la route. En conséquence, ils ne feront pas l'objet de modélisation dans le présent document.
3. Nous remarquons que les entreprises qui ne déposent pas de bilan à la BNB ont des données de moins bonne qualité. La prise en compte de celles-ci pourraient venir biaiser nos prédictions futures, nous préférons donc y renoncer. A ce stade, nous renonçons à plus ou moins 35 % des données de chaque année. Nous remarquons que d'une année à l'autre, ce taux reste stable dans le temps, ce qui est une bonne chose. Dans le cas contraire, notre analyse serait soumise à un biais d'échantillonnage pouvant détériorer la qualité de notre analyse.
4. Nous constatons la présence d'entreprises n'ayant pas de personnel au cours de l'année. Nous faisons le choix de retirer ces entreprises car elles ne possèdent aucune obligation de souscrire une assurance accident du travail. En moyenne, renoncer à ces entreprises revient à retirer +/- 5% des entreprises présentes.
5. Dans nos données, nous ne disposons pas des filiales des entreprises. En revanche, nous disposons des maisons mères avec comme information, qu'elles disposent ou non de filiales. Le problème étant que ces entreprises déclarent l'ensemble des sinistres au sein de la maison mère (avec peu de personnel à l'intérieur) mais pas au sein de chaque filiale (avec souvent la majorité du personnel). Par conséquent, nous obtenons par exemple des entreprises déclarant plus d'une centaine de sinistres sur l'année avec 1 ouvrier et 0 employé. Dans ce genre de situation, la quantité de personnel actif au sein de l'entreprise n'est pas le reflet de la sinistralité observée. Pour cette raison, nous conservons uniquement les entreprises qui ne disposent pas de filiales. En moyenne, renoncer à celles-ci, revient à se séparer +/- 4% d'entreprises supplémentaires.
6. Au sein des données restantes, nous constatons deux dernières anomalies : D'une part, il existe un certain nombre d'entreprises qui déclarent une masse salariale positive sans avoir de personnel à sa charge. D'autre part, il existe des entreprises qui déclarent du personnel avec une masse salariale nulle. En conséquence, nous choisissons de garder toutes les entreprises ayant une masse salariale positive et un nombre de ETP positif.

La table suivante résume les différentes limites présentées ci-dessus.

Limites	Description
Limite 1	Toutes les entreprises du secteur privé, du marché belge de moins de 200 membres du personnel.
Limite 2	Toutes les entreprises qui ont l'obligation de souscrire une assurance accident du travail.
Limite 3	Toutes les entreprises qui ne possèdent pas de filiales
Limite 4	Toutes les entreprises qui déposent un bilan à la Banque Nationale.
Limite 5	Toutes les entreprises ayant une masse salariale positive et un nombre de FTE positif
Limite 6	Les frais médicaux uniquement

4.5 Sélection d'une distribution théorique

Dans cette partie, nous souhaitons déterminer dans quelle mesure supposer une distribution de Poisson (distribution classique utilisée pour modéliser une variable de comptage) pour le nombre de sinistres est une hypothèse raisonnable. Celle-ci s'avère raisonnable lorsque la moyenne et la variance (conditionnellement aux variables explicatives) sont suffisamment proches. Dans un tel cas, on dit qu'il y a équi-dispersion des données et la loi de Poisson est supposée adaptée pour modéliser le comportement général de la variable réponse.

En pratique, il est courant d'observer de la sur-dispersion dans les données. Celle-ci provenant d'une part de caractéristiques non observables des assurés, traduisant une certaine hétérogénéité des comportements individuels (variables explicatives qui devraient être prise en compte pour expliquer la réponse mais non disponible pour diverses raisons) et d'autre part, de la dépendance en série : lorsque l'on observe un même preneur d'assurance, plusieurs fois au cours du temps, il est raisonnable de penser que la réalisation observée à un instant $t+s$ est partiellement dépendante de l'observation observée à l'instant t .

Le danger principal de la sur-dispersion porte sur la sous estimation des écarts-types qui pourrait nous amener à considérer comme significatives des variables qui ne le sont en fait pas.

Premièrement, nous comparons les paramètres de variance et espérance avec et sans prise en compte des variables explicatives. Les tables suivantes nous montre respectivement l'espérance et la variance du nombre de sinistres lorsque nous omettons de considérer nos variables explicatives :

Table 4.1: Paramètres nombre de sinistres ouvriers (sans prise en compte des co-variables)

Moyenne 0.06715621
Variance 10.25619
Phi 152.7214

Notre premier constat est que lorsque nous omettons les variables explicatives, il y a présence de sur-dispersion dans nos données (cf. paramètre ϕ). Ce paramètre est le ratio de la variance/moyenne. Naturellement, pour une distribution de Poisson, ce paramètre vaut 1. Notons qu'il n'est pas exclu que cette sur-dispersion disparaisse lors de la prise en compte des variables explicatives. En effet, un choix judicieux de variables explicatives et de groupements des individus en classes raisonnablement homogènes devrait faire disparaître la partie de la sur-dispersion induite par l'hétérogénéité des observations. Par la table 4.1, nous avons cherché à établir l'égalité non conditionnelle suivante :

$$E(Y) = V(Y) \quad (38)$$

Or, dans le cadre d'une modélisation GLM, nous cherchons l'égalité conditionnelle aux classes d'observation

définies à travers les différentes variables explicatives, c'est à dire

$$E(Y|X) = V(Y|X) \quad (39)$$

C'est à dire pour la variable code NACE

NACE_2pos	Moyenne	Variance	Phi
01	0.04733111	0.09144063	1.931935
02	0.09779074	0.1395004	1.42652
03	0.0281827	0.0754565	2.677405
08	0.06819403	0.3021735	4.431085
09	0.06122449	0.04431685	0.7238419
10	0.05577845	0.2533108	4.541375

En prenant en compte la variable code NACE comme exemple, la sur-dispersion a considérablement diminué sans pour autant disparaître totalement. (cf. nouveaux paramètres phi). Les nouveaux paramètres bien que bien plus faibles que 152, restent toutefois bien plus grand que 1.

Notons que l'exemple de la variable NACE_2pos est représentatif de la situation générale des données puisque, le constat est le même dans le cas des autres variables catégorielles (Voir annexes).

Dans le cadre des prochaines modélisations, nous proposerons deux alternatives pour relâcher l'hypothèse d'équi-dispersion: La première est le passage d'un GLM à un GLMM, qui permet l'inclusion d'effets aléatoires (distribué selon une loi Gaussienne) en plus des effets fixes classiques, destiné à modéliser l'hétérogénéité résiduelle (non capturable par les effets fixes d'un GLM) et dont la conséquence est la présence de sur-dispersion dans nos données. La seconde est de modéliser le nombre de sinistres à l'aide d'une distribution mixte, dans notre cas, la Poisson-logNormale sera choisie. Cette distribution considère (contrairement à la distribution de Poisson classique) que toutes les observations n'ont pas le même paramètre d'intensité λ . Au contraire, elle considère que ce paramètre varie d'une observation à l'aide d'un paramètre multiplicatif θ destiné à ajuster positivement ou négativement et distribué selon une loi LogNormale.

Dans les prochaines sections, nous discuterons notamment des résultats de ces différents modèles.

4.6 Analyse exploratoire

Dans cette section, nous allons prendre le temps d'explorer l'ensemble des données dont nous disposons. L'objectif de cette section est de donner une description des différentes variables et de montrer les liens qui peuvent exister entre les variables explicatives (entre autres, le code NACE, la forme légale,...) et la variable réponse (nombre de sinistres sur le lieu du travail). Cette étape est essentielle pour le développement d'un modèle car elle permettra d'avoir une vision critique de nos résultats. En pratique, il est courant d'effectuer des aller-retour entre modélisation et analyse exploratoire en raison du fait, notamment, que certains problèmes de modélisation peuvent être résolus par des groupements judicieux des données ou des modifications de ces groupements, ces derniers pouvant être suggérés par des analyses visuelles..

Nous prenons la décision de diviser notre base de données allant de 2009 à 2018 en deux parties:

- Les données de 2009 à 2017 feront l'objet d'un sous-jeu de données, appelé " training set". Nous utiliserons ce sous-jeu afin d'entraîner les différents modèles candidats.

- Les données de 2018 constitueront un second sous-jeu, appelé " validation-set". Le rôle de ce dernier est de tester les capacités prédictives de chacun des meilleurs modèles ,préalablement sélectionnés lors de l'entraînement,sur le "training-set".
- Les volets ouvriers et employés seront analysés séparément. Nous aurons donc un training-set et un validation-set pour chacune de ces deux études.

Cette séparation nous semble pertinente dans la mesure où elle permet de disposer d'un grand nombre de données pour entraîner le modèle ainsi que d'un nombre conséquent de données pour tester ses capacités prédictives. De plus, le fait de séparer une année du reste des données disponibles permet d'éviter un biais de sélection, dans la mesure où, le marché étant relativement stable, sa composition est sensiblement similaire entre les différentes années; la répartition des profils de risques devrait donc être comparable entre le training set et le validation set.

4.6.1 Analyse univariée

La première analyse est une analyse univariée dont l'objectif est d'analyser le comportement individuel de chaque variable. En considérant que la plupart de nos variables sont catégorielles, nous porterons entre autres notre attention sur la répartition des données au sein de chacun des niveaux d'une variable donnée. Nous observons cette répartition pour les entreprises présentes dans notre portefeuille entre 2009 et 2017.

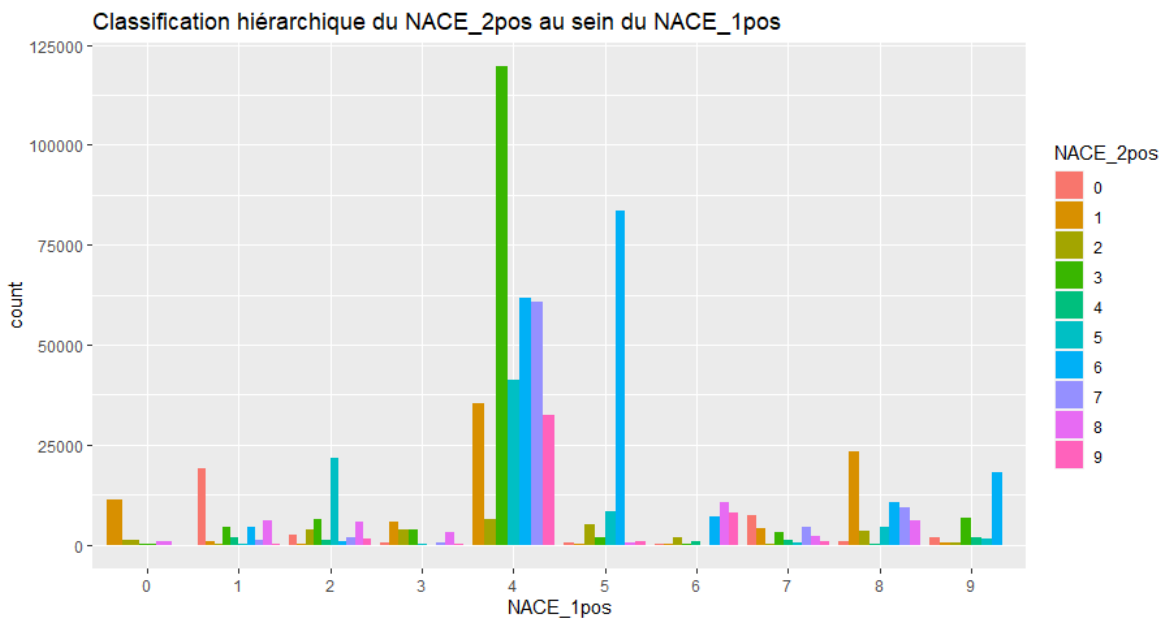


Figure 4.1: Analyse hiérarchique du code NACE ouvriers

La table suivante nous montre la répartition en pourcentage du code NACE_2pos (axe horizontal) au sein du code NACE_1pos (axe vertical).

LDAT2310 - Mémoire projet

	0	1	2	3	4	5	6	7	8	9	Total
0	0.00	1.58	0.17	0.04	0.00	0.00	0.00	0.00	0.10	0.00	1.89
1	2.66	0.11	0.02	0.60	0.26	0.04	0.62	0.18	0.84	0.01	5.33
2	0.34	0.05	0.55	0.88	0.17	3.02	0.14	0.26	0.79	0.20	6.39
3	0.06	0.80	0.51	0.53	0.00	0.02	0.00	0.07	0.43	0.02	2.45
4	0.00	4.96	0.91	16.70	0.00	5.78	8.63	8.48	0.00	4.54	50.00
5	0.07	0.01	0.69	0.27	0.00	1.17	11.66	0.00	0.09	0.12	14.09
6	0.01	0.04	0.25	0.03	0.11	0.00	0.99	0.00	1.50	1.12	4.06
7	1.01	0.59	0.05	0.45	0.17	0.09	0.00	0.60	0.30	0.10	3.36
8	0.12	3.27	0.49	0.00	0.02	0.61	1.46	1.32	0.83	0.00	8.12
9	0.25	0.07	0.08	0.92	0.25	0.19	2.55	0.00	0.00	0.00	4.31
Total	4.51	11.48	3.72	20.42	0.98	10.92	26.07	10.92	4.88	6.10	100.00

De manière générale, on observe que le code NACE_1pos "4" représente à lui seul 50 % de notre base de données, ce qui signifie qu'une entreprise sur deux de notre base de données appartient à l'un des codes suivants : 40,41,...,49. Au sein de cette première position "4", c'est la deuxième position "3" qui est la plus représentée (16,7%). Au delà de cela, les entreprises appartenant au NACE_1pos "5" sont les deuxièmes entreprises les plus représentées (14,09%).

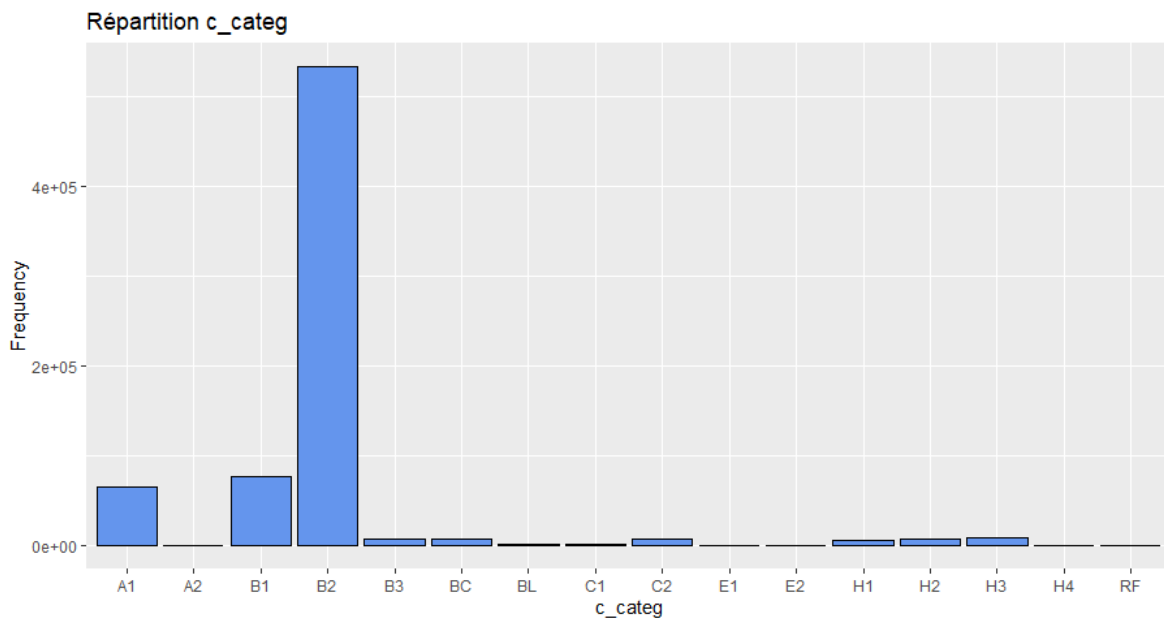


Figure 4.2: Analyse univariée de la variable c_catég_

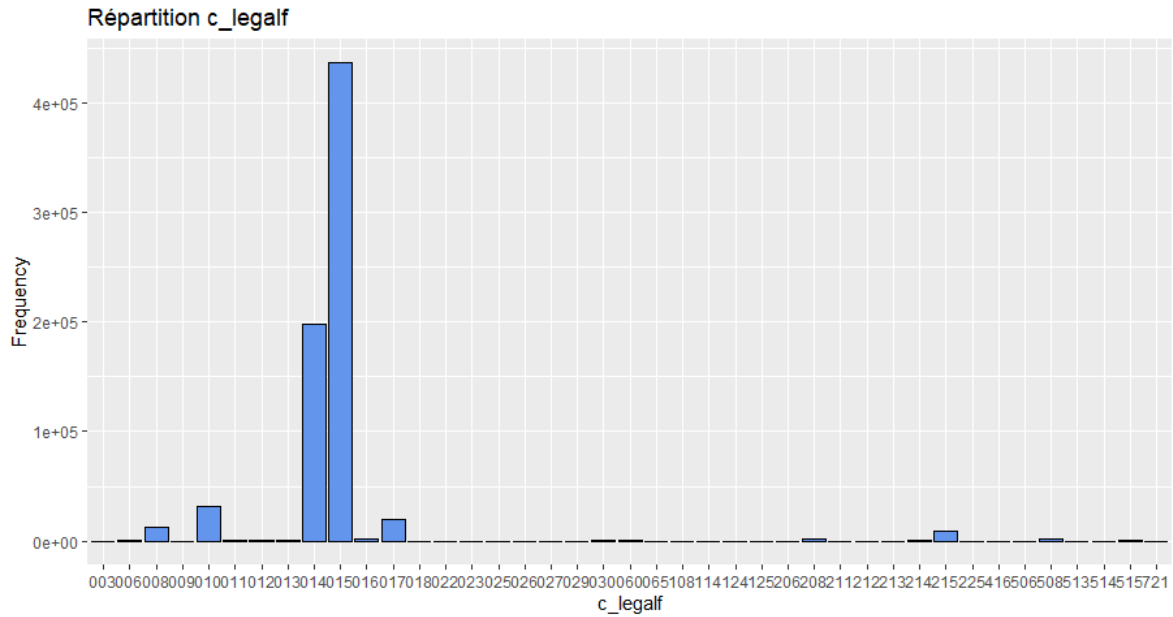


Figure 4.3: Analyse univariée de la variable c_legal

c_legal	
215	0.01
008	0.02
017	0.03
010	0.04
014	0.28
015	0.61

La forme légale 015 est la forme la plus représentée au sein des données ouvriers. En effet, 61% des entreprises présentes dans nos données ont une forme légale "015". A l'inverse, les formes 215 et 008 ne sont pratiquement pas représentées dans nos données.

	c_catég_
RF	0.00
E1	0.00
E2	0.00
H4	0.00
A2	0.00
BL	0.00
C1	0.00
H1	0.01
H2	0.01
BC	0.01
B3	0.01
C2	0.01
H3	0.01
A1	0.09
B1	0.11
B2	0.74

74% des entreprises appartiennent à la catégorie d'entreprise B2. D'autre part, les catégories d'entreprises RF, E1 et E2 sont les moins représentées dans nos données. Afin d'annuler cet effet, nous prendrons le temps d'analyser l'impact de chaque variable explicatives sur la fréquence des sinistres à l'aide d'une analyse bi-variée. La fréquence sera représentée par le nombre de sinistres divisé par le nombre d'équivalents temps plein.

Les figures suivantes montrent la répartition des différentes variables continues. Parmi ces variables, nous avons le nombre et le coût des sinistres, l'âge de l'entreprise, la masse salariale et le nombre d'ouvriers de l'entreprise.

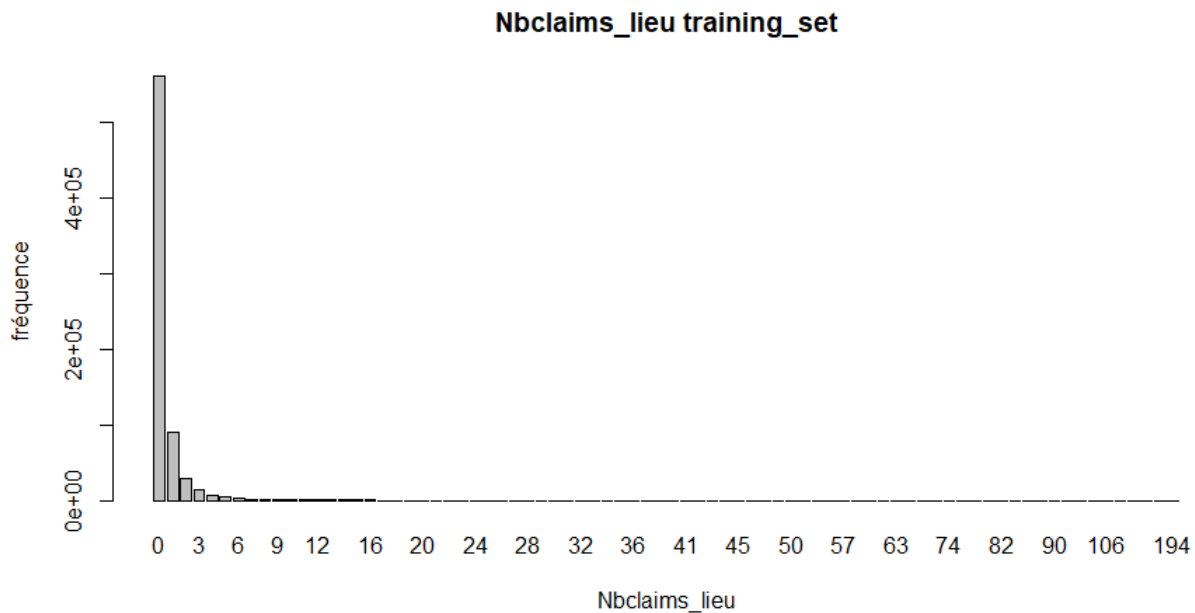


Figure 4.4: Disparité des variables continues ouvriers

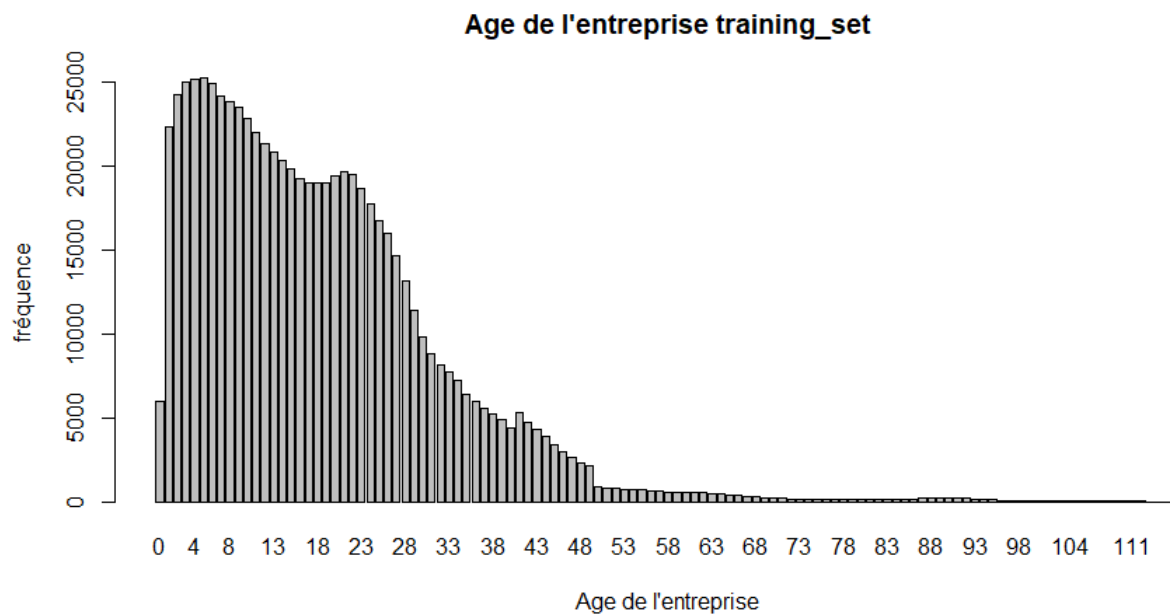


Figure 4.5: Disparité des variables continues ouvriers

Nbclaims_lieu	
6	0.00
5	0.01
4	0.01
3	0.02
2	0.04
1	0.13
0	0.78

Min	1Q	Médiane	Moyenne	3Q	Max.
0	0	0	0.4631	0	194

Nous constatons que nous disposons de 78% d'observations pour lesquelles aucun sinistre n'est enregistré alors que 22 % des observations ont au moins produit un sinistre. De plus, une observation en moyenne produit 0,46 sinistres.

Age	
8 ans	0.033
7 ans	0.034
2 ans	0.034
6 ans	0.035
3 ans	0.035
4 ans	0.035
5 ans	0.035

Les entreprises d'âge 3, 4 et 5 sont les entreprises qui sont les plus représentées dans nos données ouvriers. Les entreprises plus âgées sont systématiquement moins représentée au sein de nos données.

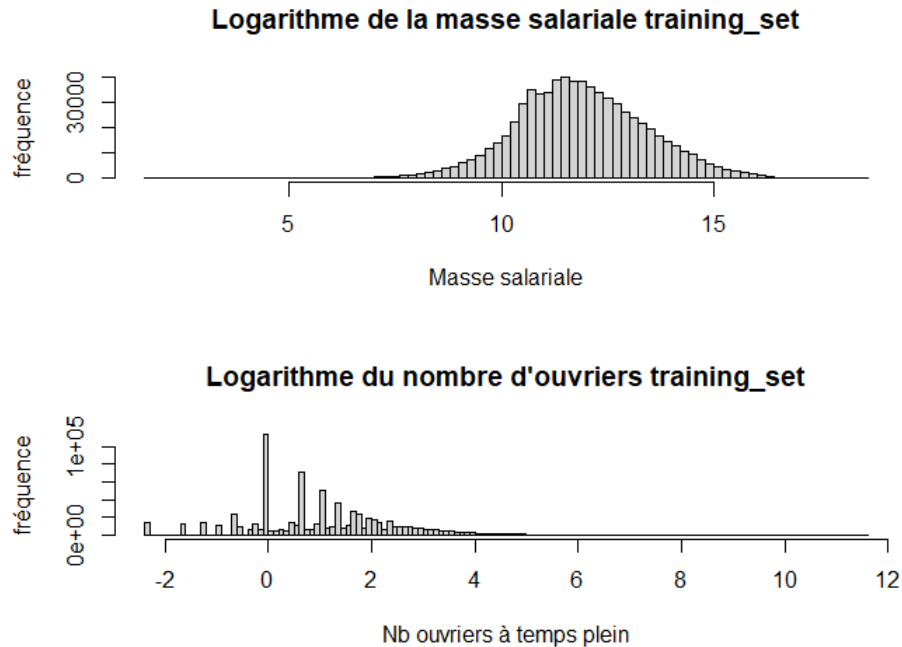


Figure 4.6: Répartition de la masse salariale et du nombre d'ouvriers à temps pleins

Les tables suivantes présentent les statistiques descriptives respectivement pour la masse salariale et le nombre d'équivalents temps pleins ouvriers.

Min	1Q	Médiane	Moyenne	3Q	Max.
6	50585	132321	462087	382134	100659673

La masse salariale moyenne d'une entreprise dans les données dont nous disposons est de 462087,00€. De plus, 50 % de ces entreprises possèdent une masse salariale inférieure à 50585,00€.

Min	1Q	Médiane	Moyenne	3Q	Max.
0.1	1.0	2.6	6.9	6.3	200

En moyenne, les entreprises disposent de 6.9 ouvriers à temps plein par entreprise. Au minimum, elles engagent un ouvrier en support qui travaille l'équivalent 1/10 de temps dans l'entreprise. Au maximum, 200 travailleurs.

Nous trouverons les graphiques associés à nos statistiques descriptives employés en annexes.

4.6.2 Analyse bivariée

Dans la section précédente, nous avons focalisé notre attention sur l'aspect individuel de chaque variable au sein de notre jeu de données. À présent, nous nous interrogeons sur l'impact des co-variables sur les réponses respectives. Notre analyse aura deux objectifs principaux :

1. Mettre en évidence des niveaux de variables explicatives ayant une fréquence moyenne de sinistres bien distinctes. Une fois identifiés, les niveaux ayant une réponse moyenne similaire pourront faire l'objet de groupements afin de créer des classes de risques homogènes. Toutefois, cette démarche devra être accompagnée d'une analyse anova permettant une approche plus rigoureuse et non arbitraire.

2. Sélectionner les variables explicatives les plus pertinentes pour expliquer une réponse donnée.

Pour ce faire, les boîtes à moustaches sont des outils qui permettent de visualiser certaines statistiques descriptives d'une variable :

- La barre noire à l'intérieur de la boîte représente la médiane.
- Le bas de la boîte représente le premier quartile.
- Le haut de la boîte représente le troisième quartile
- La taille de la boîte représente l'étendue.
- Les points les plus hauts et les plus bas à l'extérieur de la boîte sont respectivement le premier et le dernier percentile.

Les figures 4.7 à 4.10 nous montrent respectivement les impacts marginaux des variables explicatives catégorielles sur la fréquence des sinistres pour les ouvriers. Il est à noter que le nombre de sinistres est ici divisé par l'exposition au risque des entreprises (le nombre d'équivalents temps pleins ouvriers), afin de comparer toutes les entreprises sur une même échelle. La figure 4.8 nous donne les impacts marginaux combiné de la position 1 et 2 du code NACE alors que la figure 4.9 et 4.10 nous montre l'impact de chaque position du code NACE sans prise en compte du niveau parent auquel il appartient.

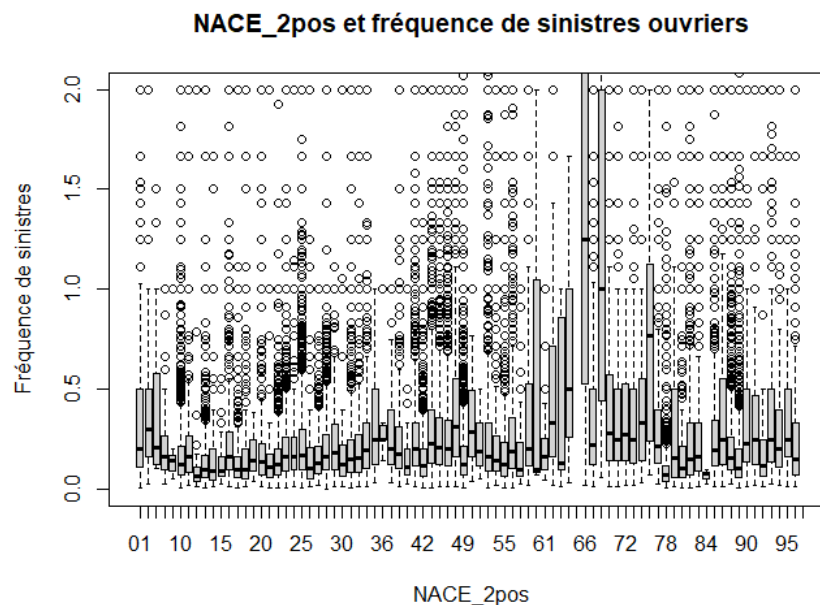


Figure 4.7: Impact du code NACE sur la fréquence des sinistres ouvriers

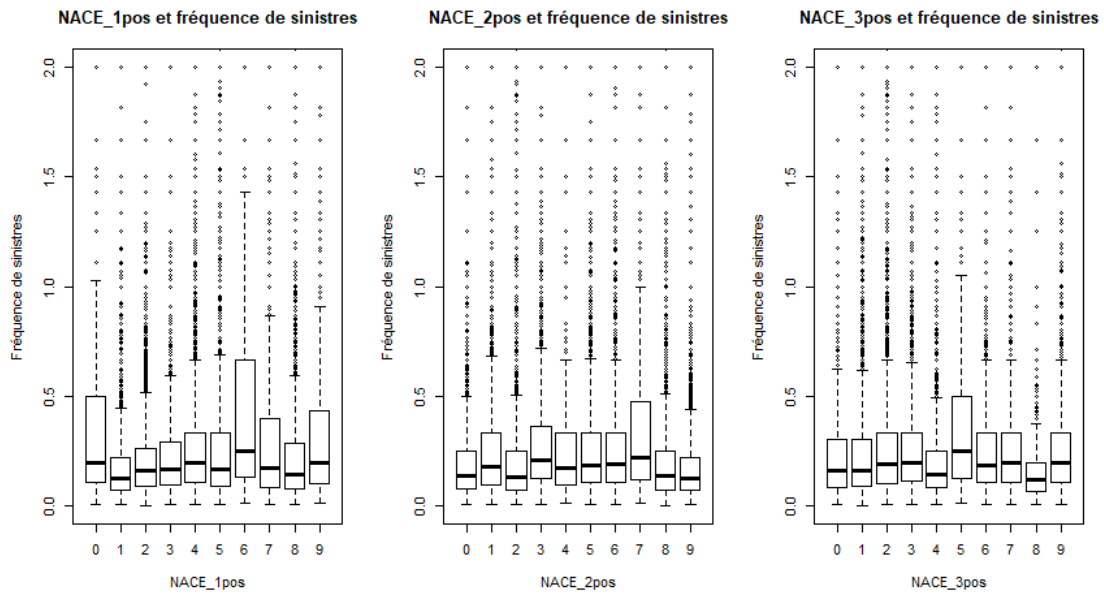


Figure 4.8: Impact du code NACE sur la fréquence des sinistres ouvriers

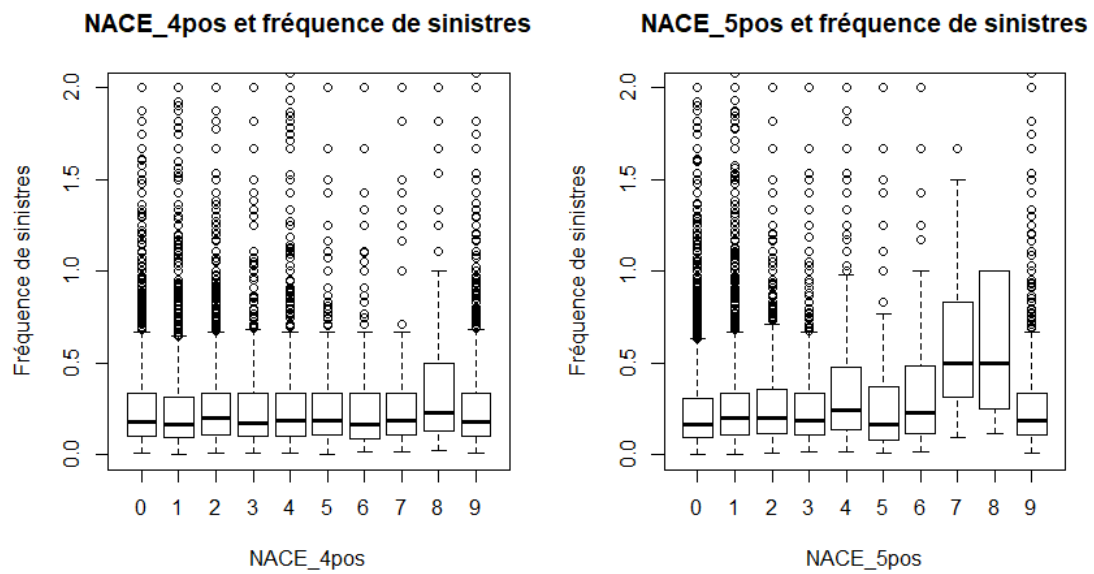


Figure 4.9: Impact du code NACE sur la fréquence des sinistres ouvriers

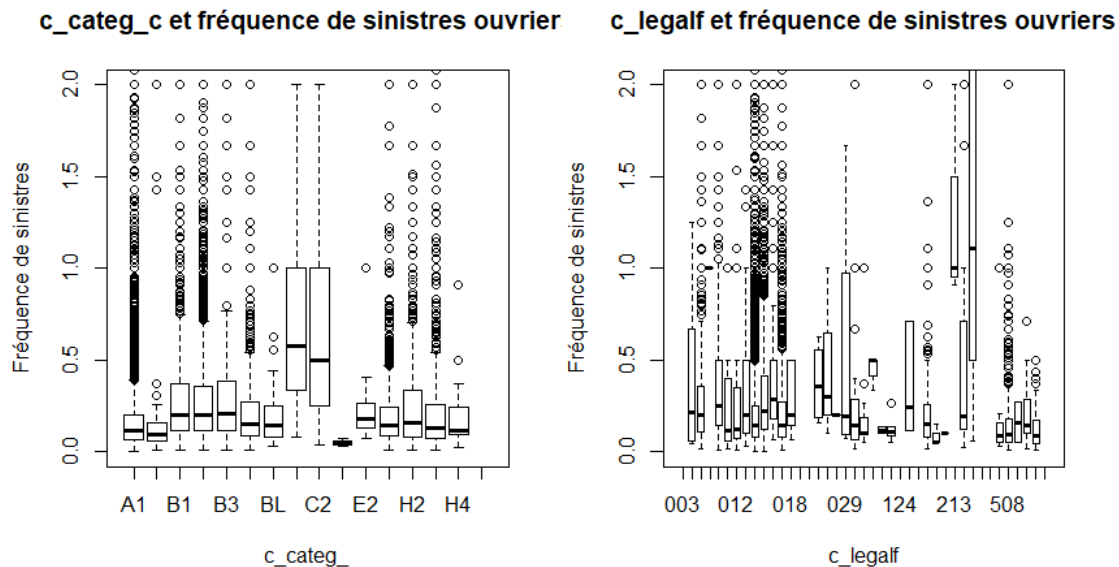


Figure 4.10: Impact de c_catég et c_legal sur la fréquence des sinistres ouvriers

En analysant les différents boxplots, on constate que l'ensemble des variables présentées ci-dessus :

- Sur la figure 4.7, les deux premières positions du code NACE combinées présentent des statistiques descriptives différentes d'une boîte à l'autre. Cela signifie que la distribution de la réponse, fréquence de sinistres, varie en fonction la double position à laquelle une entreprise donnée appartient. En conséquence, on suppose que cette variable combinée aura un pouvoir explicatif sur notre réponse.
- Les figures 4.8 et 4.9 présente les statistiques descriptives des différentes positions du code NACE sélectionnées individuelle. On observe que les positions 1,2,3 et 5 influencent massivement la réponse puisque la distribution de celle-ci varie en conséquence. En revanche, ce n'est pas le cas de la position 4, pour laquelle la majorité des boîtes et des statistiques associées sont similaires quelle que soit la 4ème position.
- Finalement, la figure 4.10 montre également des différences de distribution de la réponse selon les niveaux respectifs de la catégorie d'entreprise et de la forme légale.

En conclusion, plusieurs variables ci-dessus semblent, a priori, avoir un pouvoir explicatif sur la réponse. Les niveaux de certaines variables explicatives devront faire l'objet de groupements judicieux afin d'améliorer leur caractère significatif sur la réponse. Pour déterminer ces groupements, nous réaliserons une analyse ANOVA qui permettra de comparer les effets des différents niveaux d'une variable donnée avec ses autres niveaux sur la réponse.

Dès lors, lorsque deux niveaux d'une même variable ont des moyennes non significativement différentes, nous les grouperons ensemble. Cette approche s'appliquera facilement pour les variables "catégorie d'entreprise" et "forme légale". En revanche, elle sera davantage discutable pour la variable NACE, qui disposant d'un caractère hiérarchique, devra remettre en question les groupements statistiques proposés par l'analyse ANOVA. Bien que les groupements ANOVA puissent être statistiquement valide, il est possible qu'il ne le soit pas dans la réalité. Par exemple, ANOVA peut nous proposer un groupement entre le 43 et 21 qui statistiquement est judicieux, mais dans la réalité, 43 et 21 appartiennent à des industries respectives très différentes et il est dès lors très douteux de les grouper ensemble...

4.7 Discussion et choix de l'exposition au risque

Le choix d'une exposition au risque ("Exposure" de son terme en Anglais), permet d'obtenir la mesure d'un risque de la forme la plus unitaire possible. De plus, en cas de sur-représentation ou sous-représentation du niveau d'une variable explicative, l'exposition au risque représente une mesure de volume qui vient corriger notre variable réponse afin qu'il devienne possible de comparer sur une échelle unique les différentes réalisations de la réponse. (cf.Denuit al.,2019)

En accidents du travail, il semble judicieux de considérer la taille de l'entreprise comme exposition au risque pertinente. En effet, à taille d'entreprise plus importante, se voit attribué un nombre de sinistres plus important également. Nous mesurons la taille de l'entreprise de deux manières :

1. Un nombre d'équivalents temps pleins (ouvriers ou employés) plus important signifie une entreprise de taille plus importante ;
2. A salaires moyens constants, une masse salariale plus importante reflète une entreprise de taille plus importante ;

Toutefois, est-il possible de déterminer qu'une exposition au risque est plus adaptée qu'une autre ? Nous proposons une réflexion sur ce point ci-dessous.

Premièrement, les montants qu'un assureur peut verser sous forme de rentes d'invalidités sont fonction du salaire annuel perçu par l'assuré. Par exemple, un assuré dont le salaire annuel est de 100K et dont l'incapacité est évaluée à 70 %, pourra prétendre à une rente d'invalidité de 70K annuellement, à durée déterminée ou indéterminée.

En conséquence, nous estimons que la masse salariale est une mesure de volume adéquate pour les produits de rentes. Il est donc logique pour l'entreprise de payer une prime proportionnelle à la masse salariale totale de l'entreprise.

Toutefois, les frais médicaux n'étant pas dépendants du salaire annuel du travailleur, il paraît dérisoire de penser que le prix d'une jambe cassée puisse coûter deux fois plus cher lorsque le salaire annuel double également.

De plus, lorsque l'assureur comptabilise un nombre de sinistres qui génère des frais médicaux, il comptabilise l'ensemble des frais générés par les rendez-vous chez le médecin, le nombre d'hospitalisations ou encore le nombre d'interventions chirurgicales.

Sous les hypothèses suivantes :

1. Pour un nombre d'accidents donnés, un équivalent temps plein (ETP) peut se rendre une ou plusieurs fois chez le médecin. Il peut également être responsable d'une ou plusieurs interventions chirurgicales. En général, un ETP correspondra à un sinistre, mais il est possible qu'un ETP corresponde à plus d'un accident.
2. En considérant que deux individus ne se blessent jamais totalement manière identique lors d'un accident du travail, il apparaît évident le nombre d'interventions, le nombre de visites chez le médecin et/ou à l'hôpital varient fortement d'un individu à l'autre.
3. En considérant que deux individus ne réagissent jamais parfaitement de la même manière à la suite d'une intervention, la durée de l'hospitalisation, la dose de médicament prescrite, les potentielles rééducations ne sont jamais parfaitement équivalentes d'un individu à l'autre.

Nous pensons que le nombre d'ETP, qui représentent l'ensemble des travailleurs de l'entreprise disposant un contrat de 38h semaine en moyenne, est la mesure de volume la plus granulaire dont nous disposons pour quantifier l'effet volume associé au nombre d'accidents sur le lieu du travail.

De plus, la masse salariale n'a pas d'effets directs sur les frais médicaux alors que le nombre d'ETP a un effet direct sur le nombre de jambes cassées et d'interventions médicales qui en découlent. Plus encore, nous pouvons considérer que le nombre d'ETP sujets à un accident générant des frais médicaux comme étant une proportion du nombre d'ETP total de l'entreprise.

Finalement, la critique majeure adressée au choix de la masse salariale comme exposition au risque est qu'elle ne considère pas le secteur dans lequel l'entreprise se trouve. Nous pourrions penser à tort qu'une entreprise ayant une masse salariale relativement élevée emploie un volume plus important de travailleurs alors que dans la réalité, le salaire moyen proposé par cette entreprise est simplement plus élevé. Toutefois, la masse salariale pourrait constituer une bonne exposition au risque pour le traitement du voler sévérité, dans le cadre des rentes d'invalidités.

La figure 4.11 montre la relation linéaire entre le nombre de ETP (FTE en anglais) et le nombre de sinistres :

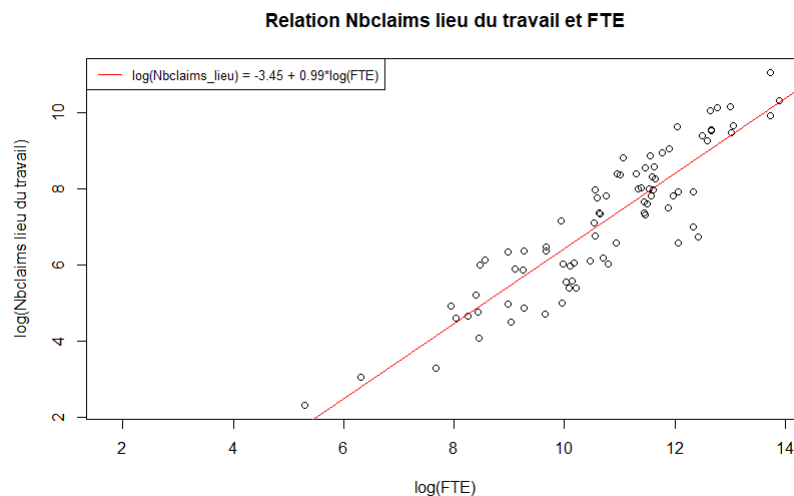


Figure 4.11: Recherche d'une relation linéaire entre le nombre de ETP et le nombre de sinistres

La prise en compte d'une exposition au risque s'effectue selon deux méthodes possibles :

1. Une prise en compte en tant que "Offset"
2. Une prise en compte en tant que "Weights"

Lorsque nous utilisons la méthode de l'offset, nous faisons le choix d'ajouter notre mesure de volume comme un terme additif à notre modèle dont le coefficient de régression est égal à 1. Lors de l'estimation de notre variable réponse, notre terme offset viendra se soustraire à notre variable réponse afin de la corriger tandis que l'utilisation d'un "weights" divise directement notre réponse par la mesure de volume. Dans notre cas, nous choisirons d'utiliser un offset afin de prendre en compte notre mesure de volume dans notre réponse.

L'objectif étant de venir pondérer chaque réponse par notre mesure de volume afin qu'elles deviennent de nouveaux comparables entre elles. Cette méthode aura pour impact de transformer un nombre d'accidents en fréquence moyenne de sinistres. Il ne s'agit plus d'estimer un nombre de sinistres par entreprises, mais bien

un nombre de sinistres moyens par entreprise.

Finalement, dans le cas d'une fonction lien logarithmique, la prise en compte de l'exposition au risque selon une méthode ou l'autre (Offset ou Weights) ne fait aucune différence d'un point de vue méthodologique. (Denuit al.,2019)

4.8 Modèles de fréquence

Dans la présente section, nous allons comparer l'ajustement des trois modèles suivants aux caractéristiques bien différentes:

- Une régression GLM de Poisson
- Une régression Poisson-LogNormal
- Une régression GLMM de Poisson

Le GLM de Poisson sera notre premier candidat modèle disposant des propriétés des familles exponentielles. Nous comparerons celui-ci avec son extension, le GLMM Poisson, qui en plus des effets fixes classiques (co-variables autres que le code NACE), permet l'introduction d'un effet aléatoire pour le traitement des variables hiérarchique, ayant un grand nombre de niveaux. Cet effet aléatoire sera supposé distribué normalement. A ce deuxième modèle nous comparons le modèle de Poisson LogNormal, étant un GLM multivarié, un modèle distribution mixte qui est la combinaison d'une distribution Poisson et lognormal. Cette distribution permettra également l'introduction d'un effet aléatoire en supposant cette fois-ci qu'il suit une distribution LogNormale. Pour les différents modèles nous testerons le caractère significatif des variables suivantes :

Sélection des variables			
Variables	Type	Groupements	Commentaires
tva	Catégorielle	Non	Représente l'ID de l'assuré et se traite comme effet aléatoire dans les modèles mixtes
NACE_1pos	Catégorielle	Non	Pas de groupements nécessaire afin de conserver au mieux le caractère hiérarchique.
NACE_2pos	Catégorielle	Non	Pas de groupements nécessaire afin de conserver au mieux le caractère hiérarchique.
c_categ_	Catégorielle	Oui	Groupements ANOVA (figure 4.10)
c_legalif	Catégorielle	Oui	Groupements ANOVA (figure 4.11)
age	Continue	Oui	Groupements au niveau des quartiles 0-25-75-100
cdy0001323	Continue	Non	Utilisée comme un offset pour corriger le nombre de sinistres
Nbclaims_lieu	Continue	Non	Variable réponse du modèle

Les tableaux suivants montrent respectivement les groupements réalisés pour les variables **c_categ_**, **c_legalif** et **age**. Pour les variables **c_categ_** et **c_legalif**, nous utilisons une analyse ANOVA. Cette

LDAT2310 - Mémoire projet

méthode permet de réaliser des groupements de plusieurs niveaux d'une même variable en détermine lorsque ceux-ci n'ont pas un effet significativement différent l'un de l'autre sur la réponse. Dès lors, les groupements qui ressortent pour nos deux variables sont les suivants :

Groupements de la variable c_categ_	
Niveau	Groupes
A1	A1,RF
A2	A2
B	B1,B2,B3
BC	BC,BL
C	C1,C2
E	E1,E2
H	H1,H2,H3
H4	H4

Groupements de la variable c_legalf	
Niveau	Groupes
00	003,006,008,009
01	011,012,013,015,016,018
014	014
017	017
02	022,023,025,026,027,029
030	030
06	060,065
11	108,114
124	124
125	125
20	206,208,211,212,213,214,215
50	416,506,508,513,514,515,721

La variable age étant une variable continue, nous n'utiliserons pas d'analyse ANOVA pour découper cette variable. En revanche, nous voulons diviser celle-ci en un nombre de niveaux suffisamment grand (plus de deux niveaux) afin de prendre en considération une suite de tranches d'âge la plus réaliste que possible. Pour ce faire, nous utilisons les quantiles 25-50-75 et 100 % de cette variable.

Quantiles training set	Age
0	0
25	8
50	16
75	25
100	116

Afin d'obtenir un nombre d'observations le plus équitablement réparti entre les différents âges, nous choisissons d'effectuer le découpage de notre variable sur base de ses quantiles.

Groupes variable age
[-1-8[
[8-16[
[17-25[
[25-118[

Nous choisissons volontairement de ne pas effectuer de groupements sur le code NACE afin de conserver son caractère hiérarchique. Dès lors, le présent modèle utilisera la variable **NACE_1pos** comme étant le premier chiffre du code à 5 positions et la variable **NACE_2pos** comme le deuxième chiffre de ce même code à 5 positions. En procédant de cette manière, nous conservons le caractère hiérarchique de la variable et nous évitons de multiplier inutilement le nombre de paramètres.

Nous pourrions traiter le code NACE à une et deux positions comme une seule et même variable, non-hiérarchique, reprenant 99 positions. Cette approche n'est pas favorisée, car en présence d'un jeu de données non suffisamment riche, nous devrions faire face à des problèmes de sur-paramétrisation du modèle ("over fitting") ce qui conduirait notre modèle à trop s'ajuster aux données d'entraînement et pas suffisamment à un autre jeu de données inconnu disposant des mêmes variables. Ce modèle verrait alors ses capacités prédictives diminuées.

Notons ici que le jeu de données étant suffisamment riche, ce problème ne se présentera pas forcément. Toutefois, si nous voulons que ce modèle puisse potentiellement être adapté à un jeu de données moins important dans le futur, il est préférable de choisir notre première approche.

Pour pallier à ce type de problèmes, nous aurions pu effectuer des groupements (toujours à l'aide d'une analyse Anova) en considérant cette approche comme discutable pour la variable NACE en raison du fait que la méthodologie Anova propose des groupements uniquement sur base du critère de significativité statistique, omettant la réalité business, qui peut être totalement différente. Par exemple, l'analyse pourrait considérer que les codes 43 et 22 peuvent être groupés alors que les industries 4 et 2 sont bien trop différentes dans la réalité. En conséquence, le code NACE comme une variable hiérarchique (2 variables à 10 niveaux chacun), nous évitons l'ensemble des problèmes explicités ci-dessus.

4.8.1 GLM Poisson

Dans les sections 3.5 et 3.6, nous avons défini la méthodologie GLM et les différentes distributions qui peuvent être choisies pour la réponse. Ce premier modèle traite l'ensemble de nos variables comme des effets fixes. En ce sens, il mesure les effets globaux de chaque variables explicatives sur la réponse, sans prise en compte des différences qu'il peut exister entre les individus d'une même groupe.

Dans un premier temps, à l'aide de la commande `glm()`, nous régressons le nombre de sinistres sur cinq variables explicatives retenues comme pertinentes pour notre modèle.

```
glm_Poisson <- glm(Nbclaims_lieu ~ NACE_1pos + NACE_2pos + [...],
  offset = log(cdy0001323) ,data = training_set_ouvriers ,
  family = poisson(link = "log"))
```

A présent, nous définissons la catégorie de référence du modèle de fréquence. La catégorie de référence est celle qui contient le plus d'observations. Dans notre cas, il s'agit des entreprises dont

- le code NACE à 1 position est 4.
- le code NACE à 2 position est 3.
- la catégorie d'entreprise est B.
- la forme légale est 01.

On recense 38391 observations qui appartiennent à l'ensemble de ces niveaux simultanément, soit 5,36% des observations totales. La table suivante montre un aperçu du recensement des observations par combinaison des différents niveaux des variables explicatives.

Table 4.2: Catégorie de référence du modèle de fréquence des sinistres ouvriers

	NACE_1pos	NACE_2pos	Groupe_c_cat	Groupe_c_legal	Groupe_age	Observations
208	4	3	B	01	(-1,8]	38931.00
230	5	6	B	01	(-1,8]	35675.00
1043	4	3	B	01	(8,17]	29426.00
1065	5	6	B	01	(8,17]	21438.00
1872	4	3	B	01	(17,28]	19279.00
237	4	7	B	01	(-1,8]	13083.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Notons que le logiciel R détermine automatiquement le niveau de référence par ordre alphabétique pour les lettres et par ordre croissant pour les chiffres. Dès lors, nous utiliserons la commande "relevel" afin de déterminer manuellement la catégorie de référence, représenté par l'intercept β_0 dans notre modèle.

Si nous choisissons une catégorie de référence avec trop peu d'observations, notre modèle est sujet à des problèmes de consistance ("Consistency") dans l'estimation de ses paramètres. Autrement dit, l'estimateur $\hat{\beta}_0$ ne convergera pas suffisamment autour de sa vraie valeur β_0 et son intervalle de confiance sera très large, en raison d'écarts-types relativement gonflés (voir en annexes).

Intuitivement, un problème de consistance de nos estimateurs génère donc une détérioration de la stabilité de performance des calculs des coefficients, c'est-à-dire, que l'ajout d'une observation supplémentaire peut changer sensiblement la valeur de cet intercept. Le paramètre $\hat{\beta}_0$ représente la moyenne estimée de la variable réponse lorsque l'entreprise partage toutes les caractéristiques de la catégorie de référence.

Finalement, la table suivante fournit un test de modèle emboîtés pour le nombre de sinistres. Nous mesurons l'impact du retrait de chaque variables explicatives du modèle complet et nous en observons les conséquences sur la déviance résiduelle du modèle. Nous constatons que celle-ci diminue de plus en plus à l'ajout de chaque variable supplémentaire. Cela signifie que chaque variable supplémentaire augmente l'ajustement du modèle aux observations du training-set et par conséquent, diminue à chaque paramètre supplémentaire, la déviance résiduelle. De plus, les p-valeurs du test de chi-2 sont toutes inférieures au seuil de significativité de 5 %. Une p-valeur inférieure au seuil de significativité signifie que l'ajout de la variable correspondante améliore significativement l'ajustement du modèle. Dans le cas inverse, nous n'avons donc aucun intérêt à rajouter cette variable au modèle.

LDAT2310 - Mémoire projet

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			715652	752266.40	
NACE_1pos	9	13180.84	715643	739085.56	0.0000
NACE_2pos	9	10751.73	715634	728333.83	0.0000
Groupe_c_categ	7	1111.97	715627	727221.86	0.0000
Groupe_c_legalf	11	3341.71	715616	723880.15	0.0000
Group_age	3	520.54	715613	723359.62	0.0000

En conclusion, il semble pertinent d'utiliser l'ensemble des variables explicatives pour expliquer le nombre de sinistres ouvriers dans le cas d'un modèle de Poisson.

La table suivante montre l'output de régression du GLM de Poisson. Celui-ci présente la particularité de traiter les deux premières positions du code NACE en tenant compte de l'aspect hiérarchique, c'est à dire, deux blocs de 10 positions NACE.

LDAT2310 - Mémoire projet

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.5424	0.0055	-460.03	0.0000
NACE_1pos0	-0.4453	0.0175	-25.49	0.0000
NACE_1pos1	-0.3309	0.0087	-38.04	0.0000
NACE_1pos2	0.0418	0.0060	7.02	0.0000
NACE_1pos3	-0.0860	0.0096	-8.95	0.0000
NACE_1pos5	-0.3280	0.0079	-41.35	0.0000
NACE_1pos6	-0.0640	0.0195	-3.28	0.0011
NACE_1pos7	-0.3182	0.0128	-24.84	0.0000
NACE_1pos8	-0.4676	0.0087	-53.59	0.0000
NACE_1pos9	-0.5193	0.0143	-36.32	0.0000
NACE_2pos0	-0.2817	0.0100	-28.09	0.0000
NACE_2pos1	-0.1118	0.0065	-17.30	0.0000
NACE_2pos2	-0.1467	0.0077	-19.17	0.0000
NACE_2pos4	-0.4716	0.0183	-25.78	0.0000
NACE_2pos5	-0.1828	0.0067	-27.48	0.0000
NACE_2pos6	-0.4368	0.0068	-64.32	0.0000
NACE_2pos7	-0.6642	0.0085	-77.88	0.0000
NACE_2pos8	-0.3347	0.0091	-36.88	0.0000
NACE_2pos9	-0.3730	0.0071	-52.89	0.0000
Groupe_c_categA1	-0.1312	0.0044	-29.53	0.0000
Groupe_c_categA2	-0.4876	0.0527	-9.25	0.0000
Groupe_c_categBC	0.0462	0.0161	2.87	0.0041
Groupe_c_categC	-0.3367	0.0430	-7.83	0.0000
Groupe_c_categE	-0.0944	0.1302	-0.72	0.4686
Groupe_c_categH	0.8469	0.1022	8.28	0.0000
Groupe_c_categH4	0.5061	0.1817	2.78	0.0054
Groupe_c_legal00	0.2420	0.0161	15.05	0.0000
Groupe_c_legal014	0.2231	0.0045	49.71	0.0000
Groupe_c_legal017	-0.4652	0.1028	-4.53	0.0000
Groupe_c_legal02	-1.5583	0.2759	-5.65	0.0000
Groupe_c_legal030	0.4065	0.1017	4.00	0.0001
Groupe_c_legal06	0.5456	0.0951	5.74	0.0000
Groupe_c_legal11	-0.3514	0.2136	-1.65	0.0999
Groupe_c_legal124	-9.7169	17.5919	-0.55	0.5807
Groupe_c_legal125	0.1671	0.2960	0.56	0.5724
Groupe_c_legal20	0.2811	0.0357	7.87	0.0000
Groupe_c_legal50	0.2724	0.0205	13.30	0.0000
Group_age(8,17]	0.1331	0.0059	22.73	0.0000
Group_age(17,28]	0.0821	0.0058	14.13	0.0000
Group_age(28,118]	0.0784	0.0059	13.23	0.0000

LDAT2310 - Mémoire projet	
AIC	1095781.66
BIC	1096240.90
Log Likelihood	-547850.83
Deviance	723359.62
Num. obs.	715653
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$	

Table 4.3: Statistical models

La première colonne des paramètres représente l'estimation des effets marginaux d'une variable explicative sur la réponse. Nous pouvons notamment dire que

- L'intercept représente la probabilité d'accident sur le lieu du travail d'un ouvrier dont l'entreprise appartient à la classe de référence. On voit qu'une telle entreprise a une probabilité de $\exp(-2.5424)$, soit 7,88% d'avoir un tel accident.
- Tous les autres coefficients positifs et négatifs représentent respectivement une augmentation et une diminution de la probabilité de sinistre ouvriers par rapport à la catégorie de référence.
- Par exemple, toute autre chose étant égales par ailleurs, si l'entreprise passe du code NACE 43 au code NACE 51, elle voit sa probabilité de sinistre multipliée par $\exp^{-0.3280-0.1118}$, c'est à dire passer de 7,88% à 5,07%. Cela signifie donc que si nos paramètres sont correctement estimés, travailler dans une entreprise du code NACE 51 est moins risqué pour les travailleurs ouvriers que de travailler dans une entreprise au code NACE 43. De plus, cette méthode de calcul s'applique aux autres variables autres que le code NACE.

La deuxième colonne nous indique les écarts-types relatifs à chaque coefficient. Ils représentent une mesure de variabilité des coefficients. Au plus ces écarts-types sont élevés, au moins nous avons confiance en l'estimation du paramètre, au plus la probabilité que la variable soit considérée par notre modèle est non significative augmente.

La troisième colonne représente le ratio entre le coefficient de régression et l'écart-type (puisque le beta réel est nul, sous l'hypothèse nulle).

La quatrième colonne est un test de significativité pour chaque niveau de nos variables explicatives. Pour chaque niveau, R nous indique s'il a un impact significativement différent de la catégorie de référence ou non. Si tel est le cas, il nous indique 1,2 ou 3 étoiles à côté du coefficient de régression.

Finalement, la table suivante donne un aperçu de la déviance nulle et résiduelle du modèle. La déviance nulle nous donne la déviance du modèle sans aucune variable explicative pour expliquer la réponse alors que la déviance résiduelle est la déviance restante après l'ajout de toutes les variables explicatives actuellement présentes dans le modèle.

Null deviance	Degrees of freedom
752266	715652
Residual deviance	Degrees of freedom
723360	715613

Table 4.4: Deviance et degrés de liberté du GLM Poisson ouvriers - fréquence

Nous constatons une déviance résiduelle de 723360, cette valeur devra être comparée au prochain GLM (sans hiérarchisation de la variable NACE). À cet instant, nous pouvons simplement constater que le ratio de cette

résidual deviance sur les degrés de libertés est très proche de 1. Celui-ci constitue une estimation du paramètre ϕ de sur-dispersion. Une valeur égale à un est synonyme d'équi-dispersion. Donc, plus notre paramètre sera proche de 1, plus cela nous confortera dans le choix d'une hypothèse de distribution de Poisson pour la variable réponse. De plus, le paramètre AIC équivaut à 1095781.66. Nous cherchons à minimiser cette valeur. Elle sera utile pour comparer tous les modèles entre eux quelle que soit la méthodologie utilisée, ce qui n'est pas le cas avec la déviance. À présent, l'output de régression suivant montre les résultats du même modèle, sans prise en compte de l'aspect hiérarchique de code NACE.

LDAT2310 - Mémoire projet

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.4708	0.0057	-434.56	0.0000
NACE2_second01	-0.7114	0.0201	-35.31	0.0000
NACE2_second02	-0.0287	0.0507	-0.57	0.5713
NACE2_second03	-1.3839	0.2042	-6.78	0.0000
NACE2_second08	-0.3629	0.0419	-8.65	0.0000
NACE2_second09	-0.5184	0.3334	-1.56	0.1199
NACE2_second10	-0.5522	0.0101	-54.62	0.0000
NACE2_second11	-0.1356	0.0324	-4.18	0.0000
NACE2_second12	-0.9714	0.1012	-9.60	0.0000
NACE2_second13	-0.7995	0.0192	-41.57	0.0000
NACE2_second14	-1.9715	0.0701	-28.11	0.0000
NACE2_second15	-1.2432	0.1099	-11.32	0.0000
NACE2_second16	-0.0392	0.0161	-2.43	0.0150
NACE2_second17	-0.8592	0.0304	-28.30	0.0000
NACE2_second18	-1.1338	0.0241	-47.05	0.0000
NACE2_second19	0.2643	0.0997	2.65	0.0080
NACE2_second20	-0.6550	0.0213	-30.72	0.0000
NACE2_second21	-1.1104	0.0652	-17.04	0.0000
NACE2_second22	-0.4178	0.0151	-27.62	0.0000
NACE2_second23	-0.0868	0.0129	-6.73	0.0000
NACE2_second24	-0.2951	0.0214	-13.80	0.0000
NACE2_second25	-0.0205	0.0079	-2.60	0.0094
NACE2_second26	-1.4873	0.0564	-26.39	0.0000
NACE2_second27	-0.5516	0.0270	-20.45	0.0000
NACE2_second28	-0.0398	0.0126	-3.15	0.0016
NACE2_second29	-0.2081	0.0217	-9.60	0.0000
NACE2_second30	-0.4961	0.0559	-8.87	0.0000
NACE2_second31	-0.3711	0.0161	-23.07	0.0000
NACE2_second32	-1.1702	0.0380	-30.78	0.0000
NACE2_second33	0.0773	0.0164	4.72	0.0000
NACE2_second35	-0.0958	0.1314	-0.73	0.4662
NACE2_second36	-0.0828	0.3334	-0.25	0.8038
NACE2_second37	0.2843	0.0499	5.70	0.0000
NACE2_second38	-0.0046	0.0198	-0.23	0.8180
NACE2_second39	-0.3988	0.0805	-4.96	0.0000
NACE2_second41	-0.0625	0.0075	-8.31	0.0000
NACE2_second42	-0.3539	0.0118	-29.96	0.0000
NACE2_second45	-0.3317	0.0098	-33.67	0.0000
NACE2_second46	-0.3030	0.0079	-38.16	0.0000
NACE2_second47	-0.9749	0.0116	-84.28	0.0000
NACE2_second49	-0.3924	0.0076	-51.93	0.0000
NACE2_second50	0.8239	0.0429	19.20	0.0000

LDAT2310 - Mémoire projet

	Estimate	Std. Error	z value	Pr(> z)
NACE2_second51	-0.7029	0.2133	-3.30	0.0010
NACE2_second52	0.3283	0.0117	28.10	0.0000
NACE2_second53	-0.6822	0.0417	-16.36	0.0000
NACE2_second55	-1.0259	0.0210	-48.75	0.0000
NACE2_second56	-1.0826	0.0105	-102.98	0.0000
NACE2_second58	-1.2777	0.1133	-11.27	0.0000
NACE2_second59	-1.0379	0.0929	-11.17	0.0000
NACE2_second60	-1.0598	0.5774	-1.84	0.0664
NACE2_second61	-0.4086	0.1099	-3.72	0.0002
NACE2_second62	-0.7181	0.0729	-9.86	0.0000
NACE2_second63	-1.1308	0.1645	-6.88	0.0000
NACE2_second64	-0.3619	0.0982	-3.69	0.0002
NACE2_second65	-6.7686	99.6821	-0.07	0.9459
NACE2_second66	-1.2481	0.1049	-11.89	0.0000
NACE2_second68	-0.2310	0.0214	-10.81	0.0000
NACE2_second69	-1.1220	0.0814	-13.79	0.0000
NACE2_second70	-0.5336	0.0295	-18.10	0.0000
NACE2_second71	-0.1432	0.0287	-4.99	0.0000
NACE2_second72	-1.0136	0.1319	-7.69	0.0000
NACE2_second73	-0.6095	0.0396	-15.37	0.0000
NACE2_second74	-0.7391	0.0721	-10.25	0.0000
NACE2_second75	-0.4278	0.1611	-2.66	0.0079
NACE2_second77	-0.1848	0.0242	-7.64	0.0000
NACE2_second78	-1.0678	0.0233	-45.92	0.0000
NACE2_second79	-1.0543	0.1006	-10.48	0.0000
NACE2_second80	-0.9326	0.0431	-21.64	0.0000
NACE2_second81	-0.6884	0.0105	-65.28	0.0000
NACE2_second82	-0.6428	0.0302	-21.27	0.0000
NACE2_second84	-0.9072	0.2586	-3.51	0.0005
NACE2_second85	-0.4087	0.0258	-15.85	0.0000
NACE2_second86	-0.9162	0.0416	-22.05	0.0000
NACE2_second87	-0.8315	0.0190	-43.70	0.0000
NACE2_second88	-0.9803	0.0180	-54.39	0.0000
NACE2_second90	-0.6219	0.0471	-13.21	0.0000
NACE2_second91	-0.5428	0.0605	-8.97	0.0000
NACE2_second92	-1.6153	0.1196	-13.51	0.0000
NACE2_second93	-0.1190	0.0218	-5.45	0.0000
NACE2_second94	-0.7017	0.0488	-14.38	0.0000
NACE2_second95	-0.4674	0.0597	-7.83	0.0000
NACE2_second96	-1.4728	0.0252	-58.47	0.0000
NACE2_second97	-10.9600	172.6545	-0.06	0.9494

LDAT2310 - Mémoire projet

	Estimate	Std. Error	z value	Pr(> z)
Groupe_c_categA1	-0.1579	0.0045	-35.14	0.0000
Groupe_c_categA2	-0.5234	0.0527	-9.93	0.0000
Groupe_c_categBC	0.0207	0.0161	1.28	0.2000
Groupe_c_categC	-0.3475	0.0430	-8.08	0.0000
Groupe_c_categE	-0.0336	0.1303	-0.26	0.7963
Groupe_c_categH	0.8696	0.1028	8.45	0.0000
Groupe_c_categH4	0.4528	0.1823	2.48	0.0130
Groupe_c_legalf00	0.1863	0.0162	11.47	0.0000
Groupe_c_legalf014	0.1882	0.0045	41.81	0.0000
Groupe_c_legalf017	-0.4981	0.1034	-4.82	0.0000
Groupe_c_legalf02	-1.8451	0.2765	-6.67	0.0000
Groupe_c_legalf030	0.3660	0.1017	3.60	0.0003
Groupe_c_legalf06	0.3801	0.0954	3.99	0.0001
Groupe_c_legalf11	-0.5778	0.2137	-2.70	0.0068
Groupe_c_legalf124	-10.9118	21.6974	-0.50	0.6150
Groupe_c_legalf125	-0.1082	0.2980	-0.36	0.7166
Groupe_c_legalf20	0.2434	0.0374	6.50	0.0000
Groupe_c_legalf50	0.3324	0.0212	15.70	0.0000
Group_age(8,17]	0.0987	0.0059	16.82	0.0000
Group_age(17,28]	0.0556	0.0058	9.55	0.0000
Group_age(28,118]	0.0531	0.0060	8.91	0.0000

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Glm_Poisson	715613	723359.62			
Glm_Poisson_NH	715549	702079.14	64	21280.48	0.0000

Nous constatons que le modèle qui ne prend pas en compte l'aspect hiérarchique du code NACE fournit de meilleurs résultats. L'AIC de ce second GLM (sans hiérarchie du NACE) est inférieur à celui du GLM qui considère l'aspect hiérarchique ($1074629.18 < 1095781.66$). De plus, le test du modèle emboîté nous montre que lorsque l'on passe du modèle avec NACE hiérarchisé à celui sans hiérarchisation du NACE, il découle une diminution significative de la déviance du modèle (- 21280.48 points de déviance).

Finalement, nous comprenons que si nous arrêtons à deux positions, le modèle sans hiérarchisation du NACE est meilleur. Cependant, il est très probable que ce ne soit plus le cas si à l'avenir nous décidions d'étendre notre modèle à un niveau NACE plus profond (incorporation du niveau NACE 3 par exemple). Dans une telle situation, nous ferions face à plus de 200 paramètres pour le modèle sans hiérarchisation contre 30 pour le modèle avec hiérarchisation de la variable NACE. De plus, d'un point de vue de la réalité, il est plus logique de prendre en considération l'aspect hiérarchique.

4.8.2 GLMM Poisson

Dans la section précédente, nous avons traité les co-variables comme des effets fixes sur la réponse. En d'autres termes, cela signifie que les paramètres du modèles étaient identiques pour tous les individus d'une même classe. Les modèles GLMs font l'hypothèse que les variables explicatives à disposition, traité comme des effets fixes, capturent la totalité de l'information nécessaire pour expliquer la réponse (d'une part l'effet direct des variables à disposition de l'assureur, d'autre part, l'effet indirect d'autres variables cachées, non mesurables mais corrélées aux variables à disposition de l'assureur). De plus, le traitement de variables ayant un nombre de niveaux important révélera les limites du GLM (Faiblesse du calcul informatique, problème de convergence des estimateurs,...) lorsque la taille d'une classe donnée présente un nombre d'observations trop faible. La principale limite adressée aux GLMs dans le cadre des données qui nous concernent est son incapacité à modéliser les différences existantes d'une entreprise à l'autre au sein d'une même classe.

Désormais, le GLMM entend solutionner cet aspect tenant compte des différences d'une entreprise à l'autre au sein d'une même classe, d'un même groupe (code NACE), etc... Il propose une combinaison d'effets fixes et aléatoires, permettant de traiter à la variables classiques traitées dans un GLM et les variables avec un grand nombre de niveaux, parfois subdivisées hiérarchiquement comme pour le code NACE dans notre cas. Ces effets aléatoires vont permettre de tenir compte simultanément de plusieurs sources de variabilité contrairement au GLM classique. De plus, cette approche permet la réduction importante du nombre de paramètres ce qui limite l'accroissement de l'AIC et évite les problèmes de sur-paramétrisation.

Finalement, la moyenne prédite sera définie comme la combinaison entre des effets fixes communs à toutes les entreprises d'une même classe de risque et la réalisation particulière d'une variable aléatoire Gaussienne qui apporte de la variabilité existante entre deux code NACE d'un même parent (Par exemple les code NACE 43 et 45 appartiennent au même parent "4") mais également de la variabilité d'une entreprise à une autre au sein d'un même code NACE. (Par exemple, l'entreprise "TOI" et l'entreprise "MOI", toutes les deux au sein du code 40).

Le logiciel R prévoit la commande `glmer()` issue du package "lme4" pour traiter ce genre de modèle.

```
glmm_Poi_basic <- glmer(Nbclaims_lieu ~ (1|NACE_1pos) +
  (1|NACE_1pos:NACE_2pos) + [...],
  family = poisson, nAGQ = 0, start = list(coef(glm_Poisson)))
```

Nous construisons un modèle GLMM de Poisson, dont le lien est logarithmique et dont l'approximation des coefficients se fait à l'aide de l'algorithme "Adaptative Gauss Hermitte Quadrature".

GLMM Poisson Basic	
(Intercept)	−3.07*** (0.06)
Groupe_c_categA1	−0.16*** (0.00)
Groupe_c_categA2	−0.52*** (0.05)
Groupe_c_categBC	0.02 (0.02)
Groupe_c_categC	−0.35*** (0.04)
Groupe_c_categE	−0.03 (0.13)
Groupe_c_categH	0.87*** (0.10)
Groupe_c_categH4	0.45* (0.18)
Groupe_c_legal00	0.19*** (0.02)
Groupe_c_legal014	0.19*** (0.00)
Groupe_c_legal017	−0.50*** (0.10)
Groupe_c_legal02	−1.85*** (0.28)
Groupe_c_legal030	0.37*** (0.10)
Groupe_c_legal06	0.38*** (0.10)
Groupe_c_legal11	−0.58** (0.21)
Groupe_c_legal124	−10.91 (21.67)
Groupe_c_legal125	−0.11 (0.30)
Groupe_c_legal20	0.24*** (0.04)
Groupe_c_legal50	0.33*** (0.02)
Group_age(8,17]	0.10*** (0.01)
Group_age(17,28]	0.06*** (0.01)
Group_age(28,118]	0.05*** (0.01)

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 4.5: Effets fixes

Le rapport des effets fixes présenté par la table 4.5 fournit les coefficients estimés pour toutes les entreprises, quelque soit le code NACE 1 et 2 ème position auquel elle appartient. L'intercept (ordonnée à l'origine) nous indique que toutes les entreprises qui font partie de la catégorie de référence mentionnée à la table 4.2 ont au moins une probabilité de $e^{-3.07}$, c'est-à-dire 4,64 % qu'un ouvrier subisse un accident sur le lieu du travail. De plus, cette probabilité sera de $e^{-3.07-0.16}$, soit 3,96 % si l'entreprise fait partie de la catégorie A1 au lieu de la catégorie B. il convient donc de comprendre qu'appartenir à A1 au lieu de B est moins risqué pour une entreprise, toutes autres choses étant égales par ailleurs.

Au titre de comparaison, nous constatons que le modèle GLMM considère que certains coefficients ont un effet non significativement différent de la catégorie de référence pour expliquer le nombre de sinistres. Pour la variable forme légale, les niveaux 124 et 125 ont un effet non significatif pour expliquer la réponse alors que pour la catégorie d'entreprise seul les niveaux BC et E ont un effet non significatif. En conséquence, ces niveaux peuvent être groupés avec l'intercept.

Effets aléatoires à priori			
Effet aléatoire	Variance	Ecart-type	Taille du groupe
NACE_1pos	0.002748	0.05242	10 niveaux
NACE_1pos : NACE_2pos	0.238871	0.48874	83 niveaux

Le rapport des effets aléatoires a priori le niveau de variabilité générale expliqué par les deux premières positions du code NACE. On voit que la plus grande partie de cette variabilité est issue de la précision de la deuxième position, au sein de la première ($0.238871 < 0.002748$). De plus, puisque notre rapport met en lumière cette variabilité, cela signifie que la prise en compte de l'hétérogénéité résiduelle à l'aide de la variable NACE, est cohérente pour notre modélisation.

AIC	1074971.85
BIC	1075247.40
Log Likelihood	-537461.93
Num. obs.	715653

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 4.6: Critères statistiques

Nous constatons un AIC de 1074972, ce qui est inférieur à l'AIC de notre modèle GLM classique. Il semblerait donc que le traitement de notre variable hiérarchique comme effet aléatoire plutôt que fixe ait du sens permet de prendre en compte une partie de l'hétérogénéité résiduelle puisque le GLMM de Poisson Basic est légèrement plus performant que le GLM classique sur base du critère AIC ($1074972 < 1095782$).

La table suivante propose les effets aléatoires à postériori.

NACE_1pos	(Intercept)	Groupe_c_catégA1	Groupe_c_catégA2	Groupe_c_catégBC
4	-3.05	-0.16	-0.52	0.02
0	-3.07	-0.16	-0.52	0.02
1	-3.08	-0.16	-0.52	0.02
2	-3.05	-0.16	-0.52	0.02
3	-3.04	-0.16	-0.52	0.02
5	-3.06	-0.16	-0.52	0.02
⋮	⋮	⋮	⋮	⋮

LDAT2310 - Mémoire projet

NACE_1pos:NACE_2pos	(Intercept)	Groupe_c_categA1	Groupe_c_categA2	Groupe_c_categBC
4:3	-2.49	-0.16	-0.52	0.02
4:1	-2.55	-0.16	-0.52	0.02
4:2	-2.84	-0.16	-0.52	0.02
4:5	-2.82	-0.16	-0.52	0.02
4:6	-2.79	-0.16	-0.52	0.02
⋮	⋮	⋮	⋮	⋮

On observe que les nouveaux intercepts ont très peu variés sur à l'intégration de **NACE_1pos** comme effet aléatoire. Ce résultat ne nous étonne pas puisque la variance expliquée à priori par **NACE_1pos** est très faible. En revanche, les intercepts deviennent beaucoup plus sensibles lorsque nous précisons la seconde position au sein de la première, c'est-à-dire, **NACE_1pos : NACE_2pos**.

Puisque ce modèle ne tient pas compte de l'historique sinistre de l'entreprise, les effets à postériori permettent de réconcilier la présente analyse avec celle du GLM de Poisson classique. Nous comprenons alors que nous avons simplement deux méthodes différentes pour estimer la même chose. Pour cette raison, les divers coefficients restent relativement proches d'un output à l'autre.

Histogramme des effets aléatoires à postériori

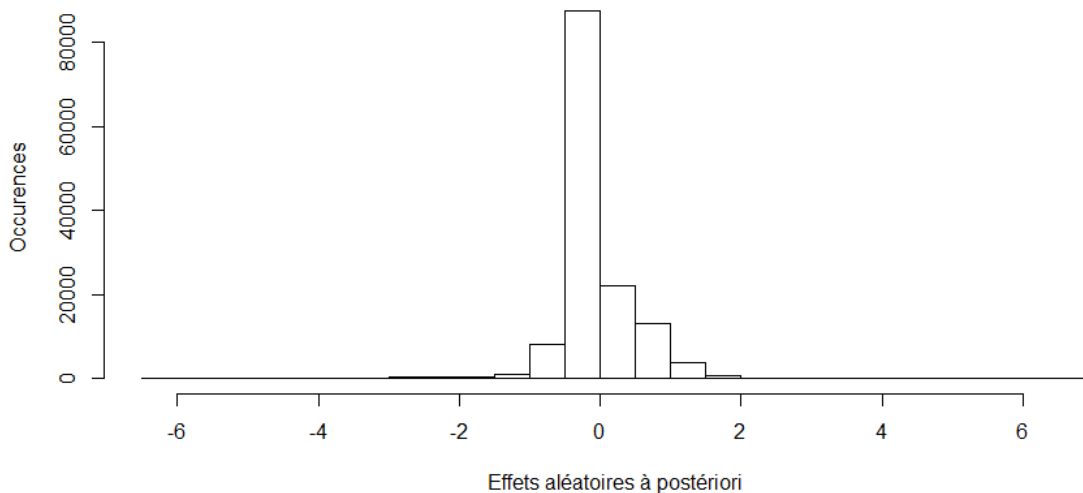


Figure 4.12: Effets aléatoires à postériori

La figure 4.12 montre la répartition des effets aléatoires à postériori. On observe que la plus grande partie de ces effets viennent diminuer la probabilité de sinistre, ce qui correspond à la plus grande barre de notre histogramme. D'autre part, une autre partie des effets aléatoires à postériori vient augmenter la probabilité de sinistres (ceci étant visible à l'aide les 3 barres positives à droite de la plus grande, au-delà de la valeur 0).

LDAT2310 - Mémoire projet

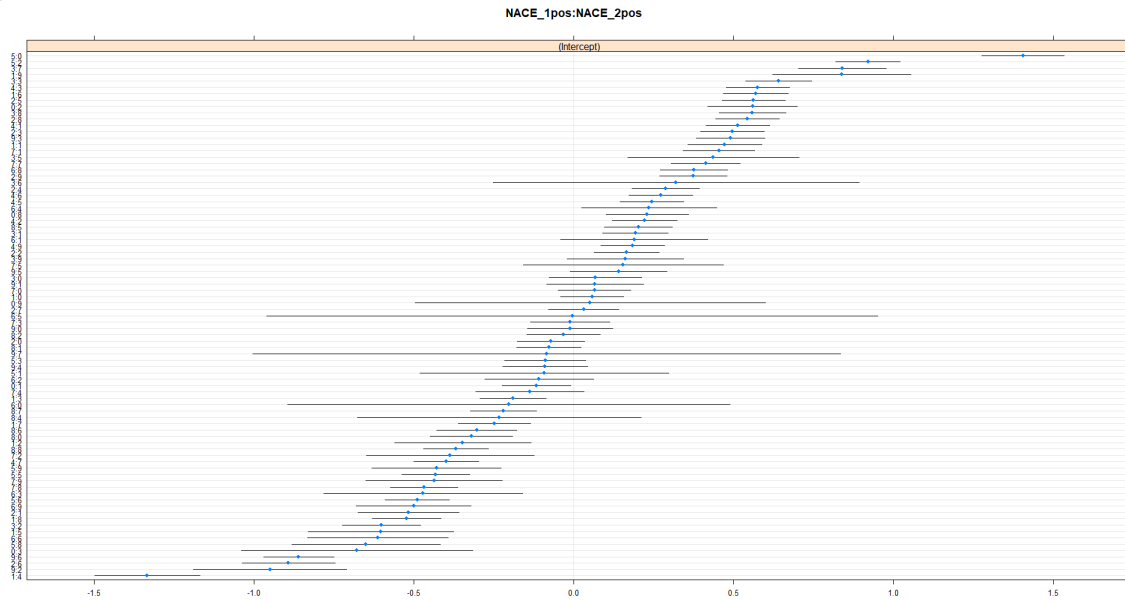


Figure 4.13: dotplot NACE_2pos au sein du NACE_1pos

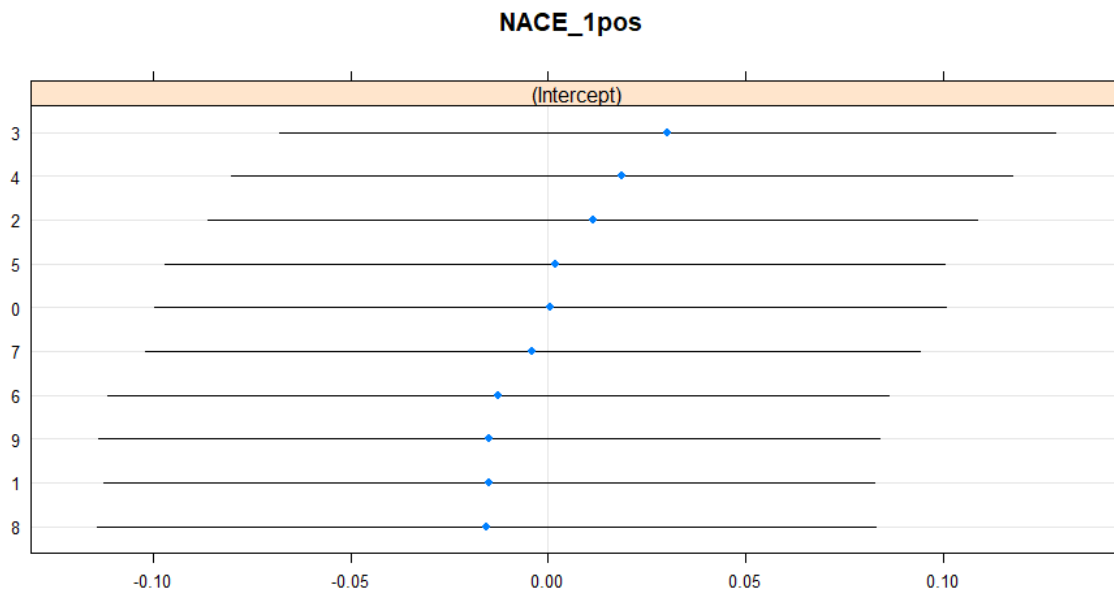


Figure 4.14: dotplot NACE_1pos

Nous avons mentionné dans la partie introductive de cette section que le GLMM était notamment capable d'ajuster la moyenne prédite en fonction de la sinistralité passée d'une entreprise. Cet aspect est très important car les différentes entreprises d'une même classe pourraient avoir des caractéristiques non observables ayant un fort impact sur leur sinistralité, expliquant d'importantes disparités dans les profils de risque des entreprises d'une même classe. En considérant le passé sinistre d'une entreprise, il sera possible pour nous de distinguer les entreprises ayant une plus forte probabilité d'accident des autres. Un second avantage est que contrairement au modèle GLM classique, le GLMM permet l'intégration de davantage de positions du code NACE (ajout des positions 3 et 4 en plus des deux premières par exemple) ce qui a le potentiel d'augmenter significativement la précision du modèle.

Pour ce faire, nous ajustons un second GLMM, qui cette fois-ci, intègre les quatres premières positions du code NACE ainsi que la sinistralité de l'entreprise.

```
glmm_Poisson_final <- glmer(Nbclaims_lieu ~ (1|NACE_1pos) +
  (1|NACE_1pos:NACE_2pos)+
  (1|NACE_1pos:NACE_2pos:NACE_3pos)+
  (1|NACE_1pos:NACE_2pos:NACE_3pos:NACE_4pos)+
  (1|tva)+ [...])
```

Effets aléatoires à priori			
Effet aléatoire	Variance	Ecart-type	Taille du groupe
NACE_1pos	0.02132	0.1460	10 niveaux
NACE_1pos:NACE_2pos	0.09920	0.3150	83 niveaux
NACE_1pos:NACE_2pos:NACE_3pos	0.03283	0.1812	255 niveaux
NACE_1pos:NACE_2pos:NACE_3pos:NACE_4pos	0.05661	0.2379	575 niveaux
tva	0.67195	0.8197	137657 niveaux

AIC	894276.78
BIC	894586.77
Log Likelihood	-447111.39

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 4.7: Statistical models

Les effets aléatoires à priori présente cette fois-ci une nouvelle configuration :

- **NACE_1pos** explique 2,42 % de la variabilité totale.
- **NACE_1pos:NACE_2pos** explique 11,25 % de la variabilité totale.
- **NACE_1pos:NACE_2pos:NACE_3pos** explique 3,72 % de la variabilité totale.
- **NACE_1pos:NACE_2pos:NACE_3pos:NACE_4pos** explique 6,42 % de la variabilité totale.
- **tva** explique 76,19 % de la variabilité totale.

LDAT2310 - Mémoire projet

Nous pouvons donc conclure que depuis l'intégration de nouvelles positions, le pouvoir explicatif des deux premières positions a augmenté comparativement au premier GLMM. Les troisièmes et quatrièmes position explique une partie de la variabilité. En conséquence, l'introduction de ces deux niveaux supplémentaires est cohérent pour notre analyse. Finalement, l'introduction de la sinistralité passée de l'entreprise à l'aide du numéro de TVA explique 76,19 % de la variabilité totale, soit la plus grande partie.

À présent, voici l'évolution des coefficients de régression suite à la prise en compte des effets aléatoires à postériori.

NACE_1pos	(Intercept)	Groupe_c_categA1	Groupe_c_categA2	Groupe_c_categBC
4	-2.95	-0.04	-0.06	-0.08
0	-2.98	-0.04	-0.06	-0.08
1	-3.11	-0.04	-0.06	-0.08
2	-2.97	-0.04	-0.06	-0.08
3	-2.92	-0.04	-0.06	-0.08
5	-3.17	-0.04	-0.06	-0.08
⋮	⋮	⋮	⋮	⋮

NACE_1pos: NACE_2pos	(Intercept)	Groupe_c_categA1	Groupe_c_categA2	Groupe_c_categBC
4:3	-2.73	-0.04	-0.06	-0.08
4:1	-2.89	-0.04	-0.06	-0.08
4:2	-2.80	-0.04	-0.06	-0.08
4:5	-3.11	-0.04	-0.06	-0.08
4:6	-2.98	-0.04	-0.06	-0.08
4:7	-3.22	-0.04	-0.06	-0.08
⋮	⋮	⋮	⋮	⋮

En comparant les coefficients **NACE_1pos** et **NACE_1pos:NACE_2pos** de ce nouveau modèle, nous voyons que l'intercept a évolué. Le **NACE_1pos** "4" enregistre désormais une probabilité plus grande de sinistres $\exp(-2.95) > \exp(-3.05)$ ce qui augmente la probabilité de sinistres pour les entreprises dont **NACE_1pos** est "4" de 4,74% à 5,23 %. Nous pouvons observer un effet similaire pour la position "0" du **NACE_1pos** également.

De plus, **NACE_1pos:NACE_2pos** la probabilité de sinistre augmente également lorsque nous précisons qu'une entreprise appartient au code "43" (précision du "3" au sein du "4") de 6,52% à 8,29 % ($\exp(-2.73)$) L'intégration de nouveaux effets aléatoires a également un impact global sur le calcul des effets fixes, à savoir l'effet la catégorie d'entreprise, la forme légale et l'âge de l'entreprise. Ces effets sont recalculés pour être augmentés ($\exp(-0.16)$)

NACE_1pos:—: NACE_4pos	(Intercept)	Groupe_c_categA1	Groupe_c_categA2	Groupe_c_categBC
4:3:1:1	-3.07	-0.04	-0.06	-0.08
4:3:1:2	-3.20	-0.04	-0.06	-0.08
4:3:1:3	-2.92	-0.04	-0.06	-0.08
4:3:2:1	-3.16	-0.04	-0.06	-0.08
4:3:2:2	-2.94	-0.04	-0.06	-0.08
4:3:2:9	-3.03	-0.04	-0.06	-0.08
⋮	⋮	⋮	⋮	⋮

LDAT2310 - Mémoire projet

TVA	(Intercept)	Groupe_c_catgA1	Groupe_c_catgA2	Groupe_c_catgBC
Entreprise 1	-5.28	-0.04	-0.06	-0.08
Entreprise 2	-3.07	-0.04	-0.06	-0.08
Entreprise 3	-3.19	-0.04	-0.06	-0.08
Entreprise 4	-3.08	-0.04	-0.06	-0.08
Entreprise 5	-3.11	-0.04	-0.06	-0.08
Entreprise 6	-2.77	-0.04	-0.06	-0.08
:	:	:	:	:

Dans le rapport des effets aléatoires a priori, il a été mentionné que la sinistralité passée de l'entreprise (**tva**) était l'effet aléatoire qui expliquait la plus grande part de variance. Ceci est confirmé dans ce dernier tableau de régression puisque l'on observe que l'intercept change radicalement d'une entreprise à une autre. On observe que certaines entreprises comme l'entreprise 1 ont une probabilité de sinistre bien plus faible que la moyenne, cela est visible par un intercept égal à -5,28. Pour une telle entreprise, l'intégration de sa sinistralité passée dans notre modèle peut lui permettre d'éviter une sur-prime en raison de la méconnaissance de son réel profil de risque.

En conclusion, l'intégration des nouvelles positions NACE ainsi que celle de la sinistralité passée de l'entreprise nous permet désormais d'augmenter la qualité d'ajustement de notre modèle. A présent, nos coefficients de régression sont bien différents de ceux calculés dans les premiers GLMM et GLM. Les nouveaux effets aléatoires décrits ci-dessus permettent d'augmenter le degré de personnalisation de nos prédictions de sorte que nous puissions désormais distinguer les entreprises ayant une plus grande/petite probabilité de sinistre par rapport à la moyenne du marché calculée par les deux précédents modèles (GLM_Poisson et GLMM_Poi_Basic).

4.8.3 Poisson - Lognormal

Le modèle Poisson-Lognormal peut être ajusté sur le logiciel R à l'aide de la fonction `PLN()`, issu du package "PLNModels". Dans cette modélisation, l'effet aléatoire destiné à modéliser l'hétérogénéité résiduelle, est distribuée selon une loi Log-Normale.

```
fitPLN <- PLN(Nbclaims_lieu ~ NACE_1pos + NACE_2pos + [...]+
  offset(log(cdy0001323)), data = datas ,
  control = list(covariance="full", trace = 0, algorithm="LBFGS"))
```

LDAT2310 - Mémoire projet

	Estimate	Std. Error
(Intercept)	-2.998303	0.005440823
NACE_1pos0	-0.4411729	0.01713111
NACE_1pos1	-0.341588	0.008769979
NACE_1pos2	0.1409555	0.006056319
NACE_1pos3	-0.08301158	0.009622919
NACE_1pos5	-0.5769359	0.007909226
NACE_1pos6	-0.241486	0.02005763
NACE_1pos7	-0.2911058	0.01259698
NACE_1pos8	-0.4184326	0.008680123
NACE_1pos9	-0.7386794	0.0143955
NACE_2pos0	-0.3315312	0.01024781
NACE_2pos1	-0.06802782	0.00640275
NACE_2pos2	-0.2126026	0.007978
NACE_2pos4	-0.5866841	0.01828999
NACE_2pos5	-0.2927419	0.006773254
NACE_2pos6	-0.4616417	0.006609722
NACE_2pos7	-0.480767	0.008646671
NACE_2pos8	-0.307958	0.008946409
NACE_2pos9	-0.3538084	0.007034329
Groupe_c_categA1	0.0540114	0.004410802
Groupe_c_categA2	-0.08323962	0.04910111
Groupe_c_categBC	0.08147672	0.01610444
Groupe_c_categC	-0.3834153	0.04209427
Groupe_c_categE	0.04236154	0.1308673
Groupe_c_categH	0.9141789	0.1031405
Groupe_c_categH4	0.4978557	0.1813718
Groupe_c_legal00	0.1588717	0.0163719
Groupe_c_legal014	0.2213528	0.00442625
Groupe_c_legal017	-0.4049935	0.1035716
Groupe_c_legal02	-1.42777	0.2673364
Groupe_c_legal030	0.07390365	0.1016501
Groupe_c_legal06	0.3265705	0.1065321
Groupe_c_legal11	-0.4650277	0.2111866
Groupe_c_legal124	-5.29151	3.223227
Groupe_c_legal125	0.3854041	0.2914479
Groupe_c_legal20	0.1335854	0.03520406
Groupe_c_legal50	0.3243662	0.02067822
Group_age(8,17]	0.04780739	0.005811364
Group_age(17,28]	-0.003636618	0.005730151
Group_age(28,118]	0.02860611	0.005791674

La catégorie de référence propose une probabilité de sinistres de 4,99% ($\exp(-2.998303)$) ce qui est raisonnablement proche des précédents intercepts. De plus, appartenir au code "45" au lieu de "43" (catégorie de référence) viendra diminuer la probabilité de sinistre d'avoir un accident ($\exp(-2.9983-0.5769) = -3.5752$) tout comme pour le GLMM Poisson Basic qui fait passer cette même probabilité de $\exp(-2.73)$ à $\exp(-3.11)$.

Nous constatons que nos deux modèles accordent leur diagnostic sur le fait qu'il est moins risqué pour entreprise d'être dans le secteur "45" plutôt que "43". En revanche, les deux modèles divergent lorsqu'il

s'agit de la catégorie d'entreprise A1 : Le PLN considère qu'appartenir à une telle catégorie va augmenter la probabilité de commettre un accident (0.054) alors que le GLMM Poisson basic estimera que cette catégorie d'entreprise vient diminuer la probabilité de sinistres(-0.04). Bien que leurs avis divergent, cela peut s'expliquer par le fait que l'estimation de leurs intercepts respectifs est différent. En effet, l'intercept du PLN étant plus petit (-2.99), la catégorie A1 viendra compenser cet effet pour faire en sorte d'augmenter la probabilité de sinistre totale alors que l'intercept du GLMM étant plus grand (-2.73), il viendra diminuer cet effet (-0.06).

5 Sélection du meilleur modèle

Le critère AIC est le critère classiquement utilisé pour comparer la plupart des modèles entre eux. Comme précédemment expliqué, il constitue un compromis entre le nombre de paramètres du modèle (complexité de celui-ci) et la log-vraisemblance des observations. En minimisant notre AIC, on réalise le meilleur compromis possible entre biais de variables omise (volonté de disposer d'un nombre de variables suffisant pour expliquer la réponse) et parcimonie (volonté d'expliquer cette même réponse avec le moins de paramètres possibles).

À l'instar de l'AIC, le critère BIC possède un esprit identique à la différence qu'il accentue le critère de parcimonie en pénalisant davantage le nombre de paramètres. Pour rappel, l'AIC étant de forme :

$$AIC = 2k - 2 \ln(\hat{L}) \quad (40)$$

On obtient alors le BIC en remplaçant le multiplicateur 2 du nombre de paramètres, par le logarithme népérien du nombre d'observations. En conséquence, le BIC aura tendance à être davantage plus élevé que l'AIC.

$$BIC = \ln(n)k - 2 \ln(\hat{L}) \quad (41)$$

Sélection du modèle final			
Modèle	AIC	BIC	LL
GLM Poisson	1095781.66	1096240.90	547851
GLM Poisson NH	1074629	1075823	547851
GLMM Poisson_Basic	1074972	1075247	537462
GLMM Poisson_final	894277	894587	447111
Poisson-LN	/	479700	479424

Selon notre premier critère AIC, le modèle GLMM Poisson_final (incluant les position 3,4 du code NACE ainsi que le passé sinistre de l'entreprise) est le meilleur modèle. En comparant le GLMM Poisson_Basic et celui-ci, on comprends que l'ajout de ces trois éléments supplémentaire est très qualitatif puisque notre AIC se réduit considérablement. De plus, il est à noter sur base de ce même critère que le GLMM sur-performe par rapport au GLM classique. Ceci ne nous étonne pas puisque pour estimer le nombre de sinistres attendu, la méthodologie GLM doit considérablement augmenter son nombre de paramètres (traitement de toutes les variables explicatives comme des effets fixes) alors que le GLMM n'augmentera pas excessivement son nombre de paramètres puisqu'il traite une partie des variables comme des effets aléatoires. Nous constatons également, si on s'arrête aux deux premières positions du code NACE, le modèle GLM_NH qui ne considère pas l'aspect hiérarchique de cette variable aura tendance à être meilleur. Ce résultat n'est valable que si la taille des échantillons reste suffisant pour estimer la valeur des paramètres sans produire de biais. En revanche, si nous devons passer à la position suivante, cette approche ne serait plus possible.

Finalement, sur base du critère BIC, nous observons que le meilleur modèle est le Poisson-logNormal. Ce modèle considère que l'effet aléatoire suit une distribution LogNormal contrairement au GLMM poisson qui

considère que l'effet aléatoire est distribué selon une loi Gaussienne classique. Nous comprenons alors que la formulation de cette hypothèse apporte sa plus value puisque ce modèle, sur base du critère BIC, sur-performe même le meilleur GLMM Poisson que nous avons à disposition.

i	Obs_claims	GLM	GLM_NH	GLMM_Basic	GLMM_final	PLN
55992	0.00	0.32	0.35	0.35	0.51	0.35
66596	0.00	0.08	0.06	0.06	0.03	0.06
60906	2.00	0.63	0.68	0.68	1.61	0.72
46468	0.00	0.01	0.01	0.01	0.01	0.01
1204	5.00	1.73	1.17	1.17	5.27	2.37
53272	0.00	0.02	0.02	0.02	0.03	0.03
24111	0.00	0.02	0.02	0.02	0.01	0.02
43694	1.00	0.31	0.34	0.34	0.16	0.36
3984	0.00	0.15	0.13	0.13	0.23	0.20
7880	0.00	0.15	0.13	0.13	0.16	0.17

En termes de prédictions, on remarque de manière générale que nos prédictions GLM, GLMM_Basic et PLN sont très proches l'une de l'autre pour les 0 sinistres. Le modèle GLMM_final est soit plus prudent en proposant une fréquence plus grande que les autres modèles, soit plus proche de l'observation du 0 sinistres. C'est notamment le cas pour la ligne 24111 (plus précis que les autres modèles) ou encore 3984 (plus prudent). Pour la ligne 43694, nous observons 1 sinistre, mais aucun modèle ne parvient à estimer correctement cette ligne. Cela signifie que les prédictions à certains égards peuvent encore être améliorées. Pour ce faire, le traitement des résidus est une tâche difficile qui incombe à l'actuaire pour écarter les observations incohérentes, présentant des erreurs et autres. Une autre méthode est l'ajout de nouvelles variables explicatives, qui ont un impact direct sur la réponse. À qualité des données équivalente, des variables explicatives plus adaptées améliorent la qualité des prédictions. La ligne 1204 montre l'avantage comparatif de la méthodologie GLMM. Pour une observation de 5 sinistres, le GLMM_final, qui considère le passé sinistre de l'entreprise, est le modèle qui fournit la meilleure prédiction. En revanche, tous les autres modèles sous-estiment le nombre de sinistre attendu pour cette entreprise.

6 Conclusion finale

À travers cette étude, nous avons modélisé la fréquence de sinistre des travailleurs ouvriers à l'aide des méthodologies suivantes : GLM de Poisson, GLMM de Poisson et enfin Poisson-LogNormale. Nous avons pu montrer que l'utilisation d'un GLM de Poisson était assez appropriée lorsqu'il s'agissait d'estimer une fréquence moyenne de marché, mais que ce dernier souffrait de certaines faiblesses lorsqu'il s'agissait de d'évaluer le réel profil de risque d'une entreprise. En plus de l'hypothèse d'équi-dispersion qui n'est pas toujours respectée, le GLM fait l'hypothèse d'une homogénéité des classes, qui ne se vérifie pas toujours, en raison de caractéristiques personnelles propres à chaque observation introduisant ainsi une hétérogénéité au sein des classes, totalement omise par la méthodologie GLM.

En conséquence, nous avons sélectionné d'autres méthodes qui améliorent notre modélisation, à travers la prise en compte de cette hétérogénéité par l'intégration d'un effet aléatoire, soit à travers un modèle de Poisson mixte (modèle de Poisson, dont le paramètre est lui-même aléatoire ; ici, ce paramètre suivait une loi log-normale), soit à travers un GLMM de Poisson, traitant le code NACE comme un effet aléatoire hiérarchique et intégrant la sinistralité passée de l'entreprise.

Plus en détails, les modèles linéaires mixtes généralisés supposent que l'effet aléatoire est distribué selon une loi Gaussienne alors que les modèles Poisson-LogNormale, qui appartiennent à la famille des modèles mixtes non linéaires considèrent que ce même effet aléatoire suit une distribution log-Normale. Qu'il s'agisse du GLMM ou du PLN, l'intégration d'effets aléatoires pour le traitement des variables ayant un grand nombre de niveaux a permis d'améliorer considérablement la qualité du modèle. Cela est visible notamment par une baisse importante des critères AIC et BIC.

Nous avons alors pu démontrer que lorsque l'on s'arrête aux deux premières positions du code NACE et que l'on ne cherche pas à évaluer le profil de risque d'une entreprise en particulier, le GLM peut être un modèle compétent pour une telle tâche. En revanche, lorsque nous souhaitons augmenter son niveau de précision par l'ajout de niveaux NACE supplémentaires, il en est incapable. Il devient alors obligatoire de passer aux GLMM qui nous permettront à la fois d'augmenter le niveau de précision du modèle par l'ajout de variables supplémentaires, mais également d'estimer le profil de risque particulier d'une entreprise donnée en intégrant sa sinistralité passée.

Dans le cadre de cette analyse, nous avons focalisé notre attention sur la garantie frais médicaux. La construction d'un premier modèle sur l'une des garanties en accident du travail n'est que la première étape du présent projet. Par la suite, des développements annexes de cette analyse, permettront d'étendre les connaissances acquises dans la garantie frais médicaux au traitement d'autres garanties, telles que les rentes d'invalidités. De plus, l'aspect sévérité n'étant pas traitée dans cette analyse, il sera nécessaire de la constituer. Cet aspect n'a pas pu être présenté en raison d'une qualité des données insuffisante pour traiter le volet sévérité. La récolte de celles-ci devra être revue afin d'avoir à disposition des montants cohérents avec les variables explicatives. Pour l'entièreté de la recherche, de nouvelles variables explicatives seront nécessaire afin de remplacer les variables catégorie d'entreprise et forme légale. À ce jour, nous pensons fortement que celles-ci sont corrélés avec d'autres variables qui influencent directement le risque mais que nous n'avons pas encore à disposition pour le moment.

Contrairement à la fréquence, la sévérité représente une variable continue, strictement positive conditionnellement disponible à la réalisation d'un sinistre et classiquement modélisée selon une distribution de Gamma.

D'autre part, la sinistralité des travailleurs ouvriers étant ce qu'elle est, nous nous attendons à une masse de probabilité en 0 plus importante pour le volet employé, en raison d'une tendance aux travaux de bureaux plus importante. En conséquence, nous nous attendons à observer une plus proportion de sinistres sur le lieu du

travail pour les ouvriers et une plus grande proportion de sinistres sur le chemin du travail pour les employés. De plus, l'analyse du volet chemin du travail nécessitera une analyse géographique.

7 Annexes

7.1 Analyse Anova - Groupements potentiels

Variable	Composition
NACE 01	02,03
NACE 08	09
NACE 10	12,13
NACE 11	.
NACE 14	15,18
NACE 16	17
NACE 19	.
NACE 20	21,23,25,27
NACE 22	28
NACE 24	.
NACE 26	.
NACE 29	.
NACE 30	31,33,34
NACE 32	35,36,39
NACE 38	.
NACE 41	.
NACE 42	.
NACE 43	.
NACE 45	.
NACE 46	.
NACE 47	.
NACE 49	.
NACE 50	51,52
NACE 53	55,58,59
NACE 56	.
NACE 60	61,62,63,64,65,66
NACE 68	.
NACE 69	.
NACE 70	72,73,74,75,79
NACE 71	.
NACE 77	.
NACE 78	.
NACE 80	81,84,85,87
NACE 82	.
NACE 86	.
NACE 88	.
NACE 90	91,92,93,94
NACE 95	96,97

Figure 7.1: Groupements code NACE_2pos fréquence ouvriers

LDAT2310 - Mémoire projet

	term	group1	group2	null.value	estimate	conf.low	conf.high	p.adj	p.adj.signif
1	NACE_2pos	01	02	0.00	0.10	-0.07	0.28	1.00	ns
2	NACE_2pos	01	03	0.00	-0.12	-0.49	0.24	1.00	ns
3	NACE_2pos	01	08	0.00	0.58	0.35	0.80	0.00	***
4	NACE_2pos	01	09	0.00	0.26	-1.10	1.63	1.00	ns
5	NACE_2pos	01	10	0.00	0.43	0.36	0.50	0.00	***
6	NACE_2pos	01	11	0.00	1.03	0.81	1.24	0.00	***

..... height

7.2 Boxplots employés

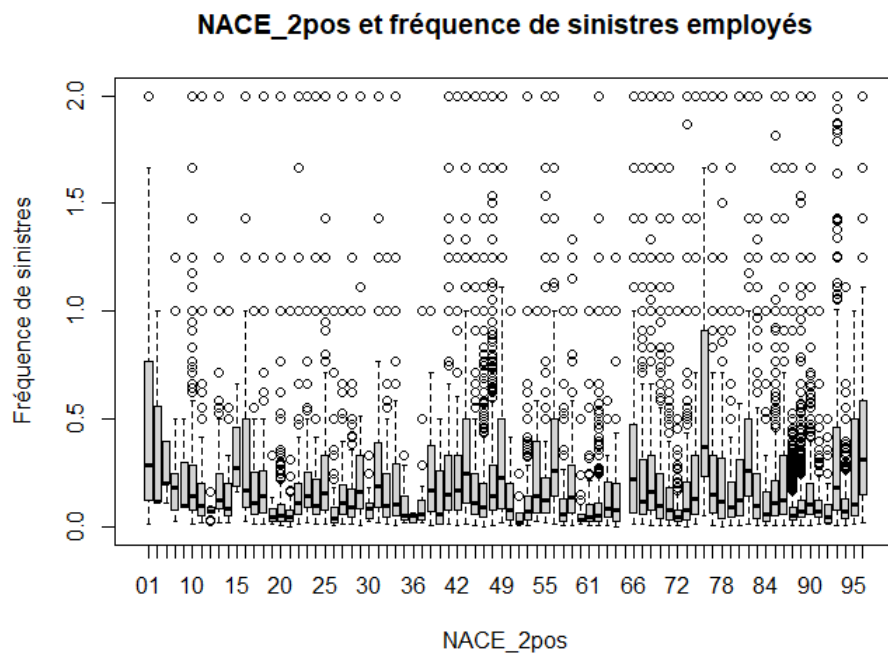


Figure 7.2: Impact du code NACE sur la fréquence de sinistre employés

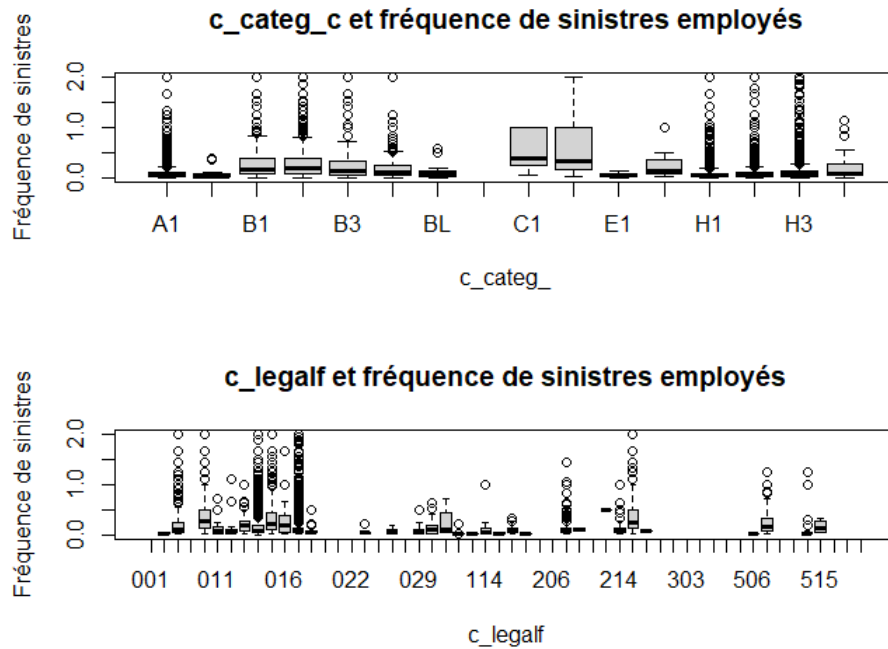


Figure 7.3: Impact de la catégorie d'entreprise et de la forme légale sur la fréquence de sinistre employés

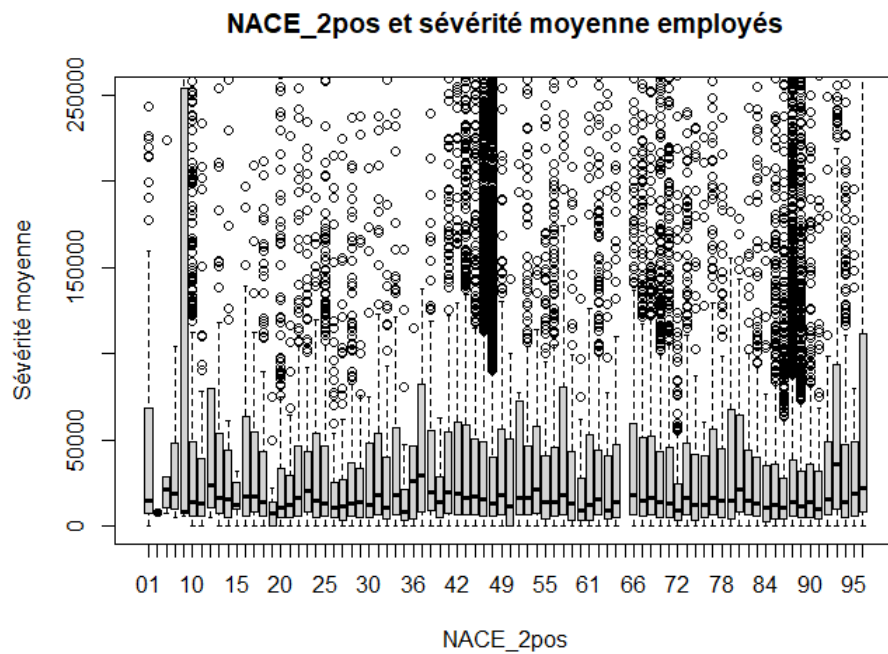


Figure 7.4: Impact du code NACE sur la sévérité moyenne employés

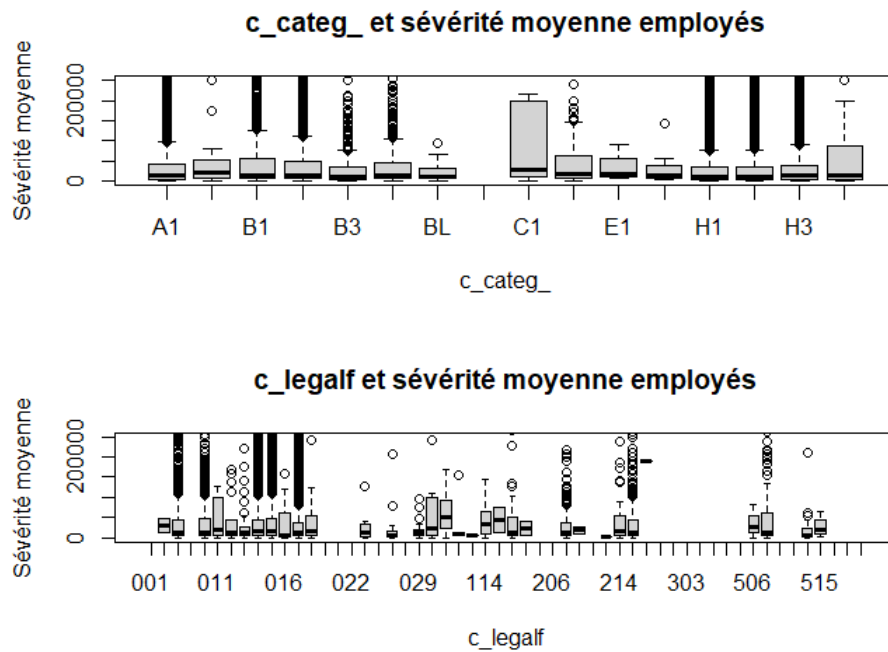


Figure 7.5: Impact de la catégorie d'entreprise et de la forme légale sur la sévérité moyenne employés

7.3 Analyse univariée des employés

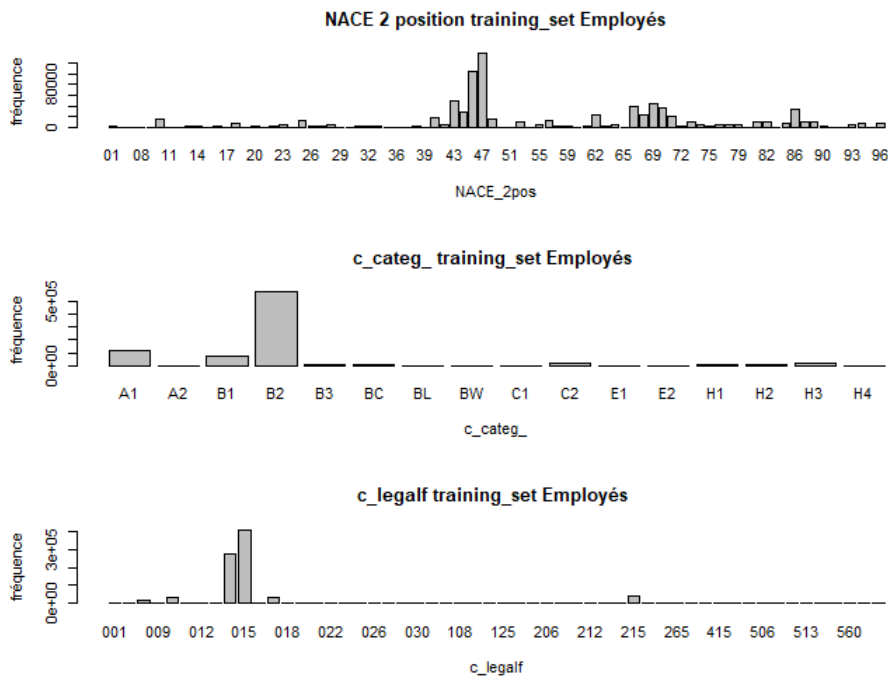


Figure 7.6: Variables catégorielles employés

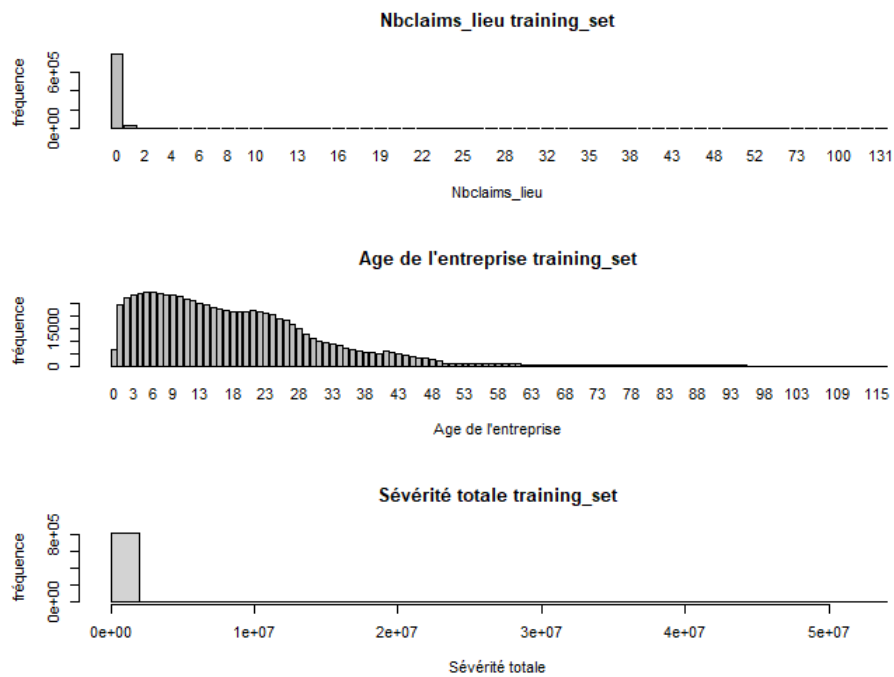


Figure 7.7: Variables continues employés

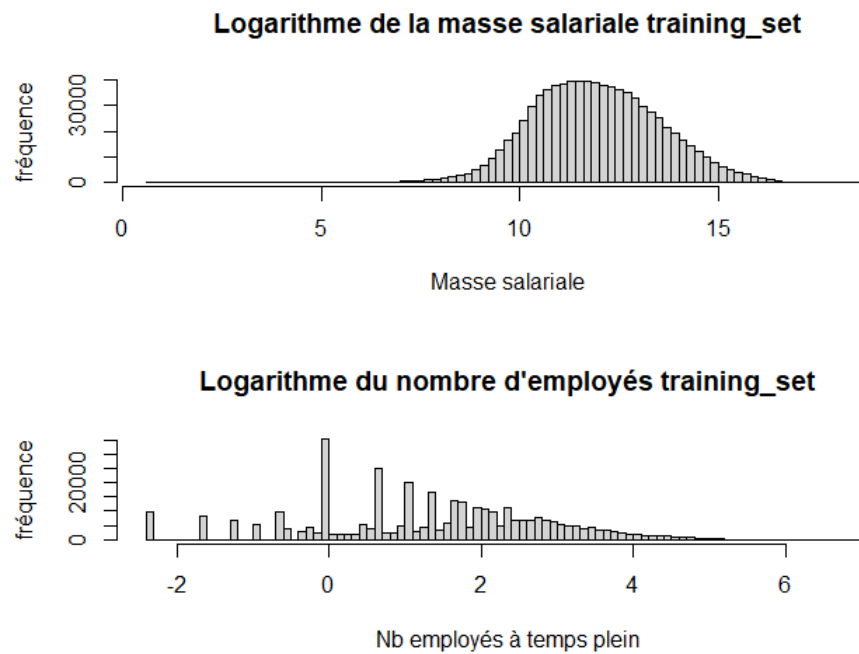


Figure 7.8: Variables continues employés

7.4 Sur-dispersion

NACE_2pos, niveau 01	average = 0.04733111	variance = 0.09144063	phi= 1.931935
NACE_2pos, niveau 02	average = 0.09779074	variance = 0.1395004	phi= 1.42652
NACE_2pos, niveau 03	average = 0.0281827	variance = 0.0754565	phi= 2.677405
NACE_2pos, niveau 08	average = 0.06819403	variance = 0.3021735	phi= 4.431085
NACE_2pos, niveau 09	average = 0.06122449	variance = 0.04431685	phi= 0.7238419
NACE_2pos, niveau 10	average = 0.05577845	variance = 0.2533108	phi= 4.541375
NACE_2pos, niveau 11	average = 0.08440768	variance = 0.3835447	phi= 4.543956
NACE_2pos, niveau 12	average = 0.03821429	variance = 0.0609138	phi= 1.594006
NACE_2pos, niveau 13	average = 0.04369845	variance = 0.1694383	phi= 3.877444
NACE_2pos, niveau 14	average = 0.01408395	variance = 0.01783457	phi= 1.266305
NACE_2pos, niveau 15	average = 0.02891826	variance = 0.06692035	phi= 2.314121
NACE_2pos, niveau 16	average = 0.09583446	variance = 0.3030951	phi= 3.162694
NACE_2pos, niveau 17	average = 0.04065214	variance = 0.1587783	phi= 3.90578
NACE_2pos, niveau 18	average = 0.03187841	variance = 0.08231867	phi= 2.58227
NACE_2pos, niveau 19	average = 0.1200274	variance = 0.493386	phi= 4.11061
NACE_2pos, niveau 20	average = 0.0483122	variance = 0.2251255	phi= 4.659808
NACE_2pos, niveau 21	average = 0.03195683	variance = 0.1400875	phi= 4.383649
NACE_2pos, niveau 22	average = 0.063253	variance = 0.2630764	phi= 4.159113
NACE_2pos, niveau 23	average = 0.0883531	variance = 0.3933387	phi= 4.451894
NACE_2pos, niveau 24	average = 0.06992655	variance = 0.6393806	phi= 9.143604
NACE_2pos, niveau 25	average = 0.09632526	variance = 0.3438861	phi= 3.570051
NACE_2pos, niveau 26	average = 0.0215198	variance = 0.05485442	phi= 2.54902
NACE_2pos, niveau 27	average = 0.05645058	variance = 0.2494707	phi= 4.419275
NACE_2pos, niveau 28	average = 0.09339959	variance = 0.4481147	phi= 4.797824
NACE_2pos, niveau 29	average = 0.07717676	variance = 0.6664196	phi= 8.634978
NACE_2pos, niveau 30	average = 0.05585826	variance = 0.2034269	phi= 3.641842
NACE_2pos, niveau 31	average = 0.06830771	variance = 0.2116122	phi= 3.097926
NACE_2pos, niveau 32	average = 0.0296519	variance = 0.07294451	phi= 2.460028
NACE_2pos, niveau 33	average = 0.1052707	variance = 0.4611072	phi= 4.380205
NACE_2pos, niveau 35	average = 0.08655591	variance = 0.1526067	phi= 1.7631
NACE_2pos, niveau 36	average = 0.09483667	variance = 0.05997618	phi= 0.6324155
NACE_2pos, niveau 37	average = 0.1315913	variance = 0.3381214	phi= 2.569481
NACE_2pos, niveau 38	average = 0.1014252	variance = 0.22871	phi= 2.254964
NACE_2pos, niveau 39	average = 0.06435311	variance = 0.2192511	phi= 3.407001
NACE_2pos, niveau 41	average = 0.08941151	variance = 0.3037393	phi= 3.397094
NACE_2pos, niveau 42	average = 0.06782736	variance = 0.2961558	phi= 4.366318
NACE_2pos, niveau 43	average = 0.09640787	variance = 0.2290205	phi= 2.375537
NACE_2pos, niveau 45	average = 0.07097914	variance = 0.1277515	phi= 1.799845
NACE_2pos, niveau 46	average = 0.07272757	variance = 0.1617557	phi= 2.224131
NACE_2pos, niveau 47	average = 0.03618327	variance = 56.49989	phi= 1561.492
NACE_2pos, niveau 49	average = 0.06575733	variance = 0.1997094	phi= 3.037067

Figure 7.9: Paramètre phi pour les 49 premiers codes NACE

7.5 Sur-dispersion

NACE_2pos, niveau 01	average = 0.04733111	variance = 0.09144063	phi= 1.931935
NACE_2pos, niveau 02	average = 0.09779074	variance = 0.1395004	phi= 1.42652
NACE_2pos, niveau 03	average = 0.0281827	variance = 0.0754565	phi= 2.677405
NACE_2pos, niveau 08	average = 0.06819403	variance = 0.3021735	phi= 4.431085
NACE_2pos, niveau 09	average = 0.06122449	variance = 0.04431685	phi= 0.7238419
NACE_2pos, niveau 10	average = 0.05577845	variance = 0.2533108	phi= 4.541375
NACE_2pos, niveau 11	average = 0.08440768	variance = 0.3835447	phi= 4.543956
NACE_2pos, niveau 12	average = 0.03821429	variance = 0.0609138	phi= 1.594006
NACE_2pos, niveau 13	average = 0.04369845	variance = 0.1694383	phi= 3.877444
NACE_2pos, niveau 14	average = 0.01408395	variance = 0.01783457	phi= 1.266305
NACE_2pos, niveau 15	average = 0.02891826	variance = 0.06692035	phi= 2.314121
NACE_2pos, niveau 16	average = 0.09583446	variance = 0.3030951	phi= 3.162694
NACE_2pos, niveau 17	average = 0.04065214	variance = 0.1587783	phi= 3.90578
NACE_2pos, niveau 18	average = 0.03187841	variance = 0.08231867	phi= 2.58227
NACE_2pos, niveau 19	average = 0.1200274	variance = 0.493386	phi= 4.11061
NACE_2pos, niveau 20	average = 0.0483122	variance = 0.2251255	phi= 4.659808
NACE_2pos, niveau 21	average = 0.03195683	variance = 0.1400875	phi= 4.383649
NACE_2pos, niveau 22	average = 0.063253	variance = 0.2630764	phi= 4.159113
NACE_2pos, niveau 23	average = 0.0883531	variance = 0.3933387	phi= 4.451894
NACE_2pos, niveau 24	average = 0.06992655	variance = 0.6393806	phi= 9.143604
NACE_2pos, niveau 25	average = 0.09632526	variance = 0.3438861	phi= 3.570051
NACE_2pos, niveau 26	average = 0.0215198	variance = 0.05485442	phi= 2.54902
NACE_2pos, niveau 27	average = 0.05645058	variance = 0.2494707	phi= 4.419275
NACE_2pos, niveau 28	average = 0.09339959	variance = 0.4481147	phi= 4.797824
NACE_2pos, niveau 29	average = 0.07717676	variance = 0.6664196	phi= 8.634978
NACE_2pos, niveau 30	average = 0.05585826	variance = 0.2034269	phi= 3.641842
NACE_2pos, niveau 31	average = 0.06830771	variance = 0.2116122	phi= 3.097926
NACE_2pos, niveau 32	average = 0.0296519	variance = 0.07294451	phi= 2.460028
NACE_2pos, niveau 33	average = 0.1052707	variance = 0.4611072	phi= 4.380205
NACE_2pos, niveau 35	average = 0.08655591	variance = 0.1526067	phi= 1.7631
NACE_2pos, niveau 36	average = 0.09483667	variance = 0.05997618	phi= 0.6324155
NACE_2pos, niveau 37	average = 0.1315913	variance = 0.3381214	phi= 2.569481
NACE_2pos, niveau 38	average = 0.1014252	variance = 0.22871	phi= 2.254964
NACE_2pos, niveau 39	average = 0.06435311	variance = 0.2192511	phi= 3.407001
NACE_2pos, niveau 41	average = 0.08941151	variance = 0.3037393	phi= 3.397094
NACE_2pos, niveau 42	average = 0.06782736	variance = 0.2961558	phi= 4.366318
NACE_2pos, niveau 43	average = 0.09640787	variance = 0.2290205	phi= 2.375537
NACE_2pos, niveau 45	average = 0.07097914	variance = 0.1277515	phi= 1.799845
NACE_2pos, niveau 46	average = 0.07272757	variance = 0.1617557	phi= 2.224131
NACE_2pos, niveau 47	average = 0.03618327	variance = 56.49989	phi= 1561.492
NACE_2pos, niveau 49	average = 0.06575733	variance = 0.1997094	phi= 3.037067

Figure 7.10: Paramètre phi pour la catégorie d'entreprise

7.6 Sur-dispersion

c_legalf, niveau 003	average = 0	variance = 0	phi= NaN
c_legalf, niveau 006	average = 0.0194091	variance = 0.02486989	phi= 1.281352
c_legalf, niveau 008	average = 0.07237465	variance = 0.2258307	phi= 3.120301
c_legalf, niveau 009	average = 0.173913	variance = 0.1954007	phi= 1.123554
c_legalf, niveau 010	average = 0.07115375	variance = 0.1186358	phi= 1.667316
c_legalf, niveau 011	average = 0.02980222	variance = 0.108829	phi= 3.651709
c_legalf, niveau 012	average = 0.03721532	variance = 0.1377831	phi= 3.702321
c_legalf, niveau 013	average = 0.05069387	variance = 0.1487753	phi= 2.934779
c_legalf, niveau 014	average = 0.07445381	variance = 0.440291	phi= 5.913613
c_legalf, niveau 015	average = 0.05977855	variance = 19.20033	phi= 321.1909
c_legalf, niveau 016	average = 0.06767273	variance = 0.09702242	phi= 1.4337
c_legalf, niveau 017	average = 0.05888414	variance = 0.2917905	phi= 4.955333
c_legalf, niveau 018	average = 0.07216495	variance = 0.08027951	phi= 1.112445
c_legalf, niveau 022	average = 0	variance = 0	phi= NaN
c_legalf, niveau 023	average = 0	variance = 0	phi= NaN
c_legalf, niveau 025	average = 0.08143322	variance = 0.06250466	phi= 0.7675572
c_legalf, niveau 026	average = 0.0397878	variance = 0.04055862	phi= 1.019373
c_legalf, niveau 027	average = 0.05882353	variance = 0.03704458	phi= 0.6297578
c_legalf, niveau 029	average = 0.01143438	variance = 0.0501022	phi= 4.381715
c_legalf, niveau 030	average = 0.07935297	variance = 0.2012324	phi= 2.535916
c_legalf, niveau 060	average = 0.07546386	variance = 0.3520975	phi= 4.665776
c_legalf, niveau 065	average = 0.2727273	variance = 0.1487603	phi= 0.5454545
c_legalf, niveau 108	average = 0.02142475	variance = 0.07527891	phi= 3.513643
c_legalf, niveau 114	average = 0.04415011	variance = 0.07937571	phi= 1.79786
c_legalf, niveau 124	average = 0	variance = 0	phi= NaN
c_legalf, niveau 125	average = 0.09936125	variance = 0.1354255	phi= 1.362961
c_legalf, niveau 206	average = 0	variance = 0	phi= NaN
c_legalf, niveau 208	average = 0.09801363	variance = 0.2345547	phi= 2.393082
c_legalf, niveau 211	average = 0.05063291	variance = 0.04989312	phi= 0.9853891
c_legalf, niveau 212	average = 0.04291845	variance = 0.02087224	phi= 0.4863232
c_legalf, niveau 213	average = 0.2112676	variance = 0.1682262	phi= 0.7962706
c_legalf, niveau 214	average = 0.05603034	variance = 0.1457317	phi= 2.600942
c_legalf, niveau 215	average = 0.03081149	variance = 0.05341329	phi= 1.733551
c_legalf, niveau 225	average = 0	variance = 0	phi= NaN
c_legalf, niveau 416	average = 0	variance = 0	phi= NaN
c_legalf, niveau 506	average = 0.06266126	variance = 0.07433453	phi= 1.186292
c_legalf, niveau 508	average = 0.05867906	variance = 0.1910492	phi= 3.255833
c_legalf, niveau 513	average = 0.06	variance = 0.01630816	phi= 0.2718027
c_legalf, niveau 514	average = 0.07422068	variance = 0.1917931	phi= 2.584092
c_legalf, niveau 515	average = 0.03674196	variance = 0.1399746	phi= 3.809666
c_legalf, niveau 721	average = 0	variance = 0	phi= NaN

Figure 7.11: Paramètre phi pour la forme légale

7.7 Consistence des estimateurs

Soit $\hat{\theta}_n$ l'estimateur obtenu d'un échantillon de n observations, notées $Y_1, \dots, Y_n$. L'estimateur $\hat{\theta}_n = \hat{\theta}(Y_1, \dots, Y_n)$, fonction des n observations, est défini comme consistant pour le paramètre θ si

$$\lim_{n \rightarrow \infty} \mathcal{P}[|\hat{\theta} - \theta| < \epsilon] = 1 \text{ pour tout } \epsilon > 0 \quad (42)$$

Cette propriété assure que plus la taille de notre échantillon augmente (en tendant vers l'infini), plus la valeur estimée de notre estimateur gravite autour de sa vraie valeur. (cf. Denuit et al., 2019)

7.8 Fréquences de variables catégorielles

NACE_2pos	Fréquences
01	0.0117
10	0.0181
25	0.0187
41	0.0344
43	0.1074
45	0.0391
46	0.0998
47	0.1306
49	0.0287
56	0.0764
62	0.0208
66	0.0325
68	0.0258
69	0.0377
70	0.0351
81	0.0218
86	0.0324

Figure 7.12: Fréquences des NACE_2pos

Categ	Fréquences
A1	0.0951
A2	0.0001
B1	0.0954
B2	0.7343
B3	0.0093
BC	0.0139
BL	0.0003
BW	0.0000
C1	0.0026
C2	0.0177
E1	0.0000
E2	0.0001
H1	0.0063
H2	0.0104
H3	0.0142
H4	0.0003
RF	0.0000

Figure 7.13: Fréquences des catégories

legalf	Fréquences
008	0.0207
010	0.0396
014	0.2641
015	0.5936
016	0.0017
017	0.0297
214	0.0012
215	0.0355

Figure 7.14: Fréquences des formes légales

8 Bibliography

JAMES, Gareth, WITTEN, Danielle, HASTIE, Trevor et TIBSHIRANI, Robert (2017) ."An Introduction to Statistical Learning"

DE JONG, Piet, HELLER, Gillian Z.(2008) ."Generalized Linear Models for Insurance Data"

WINKELMANN, Rainer(2008)." Econometric Analysis of Count Data"

DENUIT, Michel, HAINAUT, Donatien, TRUFIN, Julien(2019)." Effective Statistical Learning Methods for Actuaries I"

DENUIT, Michel, MERACHAL, Xavier, PITREBOIS, Sandra, WALHIN, Jean-François (2007)." Actuarial Modelling of Claim Counts : Risk Classification, Credibility and Bonus-Malus Systems"