INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0143

PERSONNALISATION DES PRIMES-FRÉQUENCE EN ASSURANCE AUTOMOBILE PAR REGRESSION POISSONIENNE EN PRESENCE DE DONNEES LONGITUDINALES

MICHEL DENUIT, SANDRA PITREBOIS & JEAN-FRANCOIS WALHIN

http://www.stat.ucl.ac.be

PERSONNALISATION DES PRIMES-FRÉQUENCE EN ASSURANCE AUTOMOBILE PAR REGRESSION POISSONIENNE EN PRESENCE DE DONNEES LONGITUDINALES

MICHEL DENUIT[★], SANDRA PITREBOIS[♦†]
et JEAN-FRANÇOIS WALHIN[†‡]

[★]Institut de Statistique
Université Catholique de Louvain
Voie du Roman Pays, 20
B-1348 Louvain-la-Neuve, Belgium

[♦]Facultés universitaires Saint-Louis
Boulevard du Jardin Botanique, 43
B-1000 Bruxelles, Belgium

[†]Secura Belgian Re
Rue Montoyer, 12
B-1000 Bruxelles, Belgium

[‡]Unité de Sciences Actuarielles
Institut d'Administration et de Gestion
Université Catholique de Louvain
Place des Doyens, 1
B-1348 Louvain-la-Neuve, Belgium

October 19, 2001

Résumé

Souvent, les actuaires utilisent plusieurs années d'observation de leur portefeuille automobile afin de personnaliser les primes-fréquence. Ce faisant, ils négligent la dépendance sérielle existant entre les données relatives à un même individu. Cet article examine précisément à l'aide des "Generalized Estimating Equations" (GEE) l'impact de cette méthodologie sur les estimations des fréquences des sinistres. Une illustration est proposée à l'aide du logiciel SAS sur base d'un portefeuille d'assurance automobile belge observé au cours de trois années.

1 Introduction

1.1 Le concept de segmentation et ses implications

Le terme “segmentation” est actuellement considéré comme faisant partie du jargon professionnel de l'assurance. Selon les travaux de la Commission des Assurances belge, on qualifie de segmentation toute technique que l'assureur utilise pour différencier la prime, voire la couverture, en fonction d'un certain nombre de caractéristiques spécifiques du risque à assurer, et ce afin de parvenir à une meilleure concordance entre les coûts qu'une personne déterminée met à charge de la collectivité des preneurs d'assurance et la prime que cette personne doit payer pour la couverture offerte. Dans certains cas, cela peut impliquer que l'assureur refuse purement et simplement de couvrir le risque.

La segmentation ne se limite pas à la différenciation tarifaire, bien connue de tous, mais comporte aussi la sélection du risque à laquelle procède l'assureur lors de la conclusion du contrat (acceptation) ou en cours du contrat (sélection ultérieure). Ceci peut se représenter schématiquement comme suit:

Acceptation du risque
↓
Tarification *a priori*
Imposition de franchises ou de découverts obligatoires
Transformation du risque à assurer
↓
Tarification *a posteriori*
Résiliation

Le principe qui consiste à demander au preneur d'assurance une prime qui correspond au risque individuel qu'il représente ne peut pas être mis en pratique dès la souscription du contrat. Ceci requerrait en effet que tous les facteurs influençant le risque soient connus et que leur impact puisse être établi sans équivoque. Eu égard à l'hétérogénéité encore présente au sein des classes d'assurés créées par l'actuaire, la différence dans les statistiques de sinistres des assurés ne doit pas seulement être attribuée au hasard mais doit être considérée dans une certaine mesure comme un reflet de l'influence des facteurs de risque qui n'ont pas été pris en considération *a priori.* L'intégration de l'historique des sinistres dans la tarification donne lieu à une deuxième forme de segmentation faisant le plus souvent usage d'une personnalisation *a posteriori* au moyen d'un système de type bonus-malus ou “d'experience-rating”.

La personnalisation *a posteriori* constitue ainsi une correction des manquements de la personnalisation *a priori.* Par conséquent, la mesure dans laquelle la statistique sinistres intervient dans la détermination du montant des primes doit dépendre du degré de précision tarifaire qui a été appliqué *a priori.* Dans l'hypothèse théorique où la tarification *a priori* conduirait à des groupes de risques parfaitement homogènes, l'application d'une personnalisation *a posteriori* ne se justifierait pas.

1.2 Portée du travail

Dans ce travail, nous considérons le problème de la segmentation *a priori* sur la composante fréquentielle de la prime pure. Etant donné le degré limité, voire l'inexistence, de la segmentation sur les coûts des sinistre dans bon nombre d'assurance de responsabilité, pour des raisons inhérentes à la nature même de ces produits, la segmentation sur les fréquences de sinistre revêt une importance primordiale.

Nous avons scindé notre étude en deux parties. Dans la première partie, nous présentons une méthode simple et performante de segmentation *a priori* et de calcul des primes-fréquence (c'est-à-dire des anticipations des fréquences de sinistre par l'assureur, à la lumière des informations dont il dispose à propos de ses assurés). La principale originalité de notre démarche est de reconnaître explicitement l'aspect sériel des données servant de base à l'établissement du tarif. A cet égard, le présent travail complète et précise certains aspects de Denuit, Pitrebois & Walhin (2001a). Il est, par certains aspects, très proche de Pinquet, Guillen & Bolancé (2001), à ceci près que nous nous intéressons ici exclusivement à la tarification *a priori* (la dépendance sérielle entre les observations apparaît donc comme une nuisance) alors que ces auteurs se focalisent sur la tarification *a posteriori*.

1.3 Modèles de régression en tarification

De nombreuses techniques statistiques ont été utilisées pour répartir les assurés en classes aussi homogènes que possible. Globalement, on peut distinguer les méthodes relevant de l'analyse des données (notamment les arbres de classification) et celles basées sur les modèles de régression. Cet article est entièrement consacré à cette dernière optique.

Au cours de la dernière décennie, de nombreux actuaires ont fait usage de modèles de régression pour des données non-normales. Parmi ceux-ci, on notera les modèles linéaires généralisés, permettant de modéliser des situations bien plus variées que ne le permet le modèle linéaire classique. Bien que la régression linéaire reste une des techniques statistiques les plus utilisées dans beaucoup de domaines, force est de constater qu'il y a de nombreuses situations où elle ne s'applique pas (ou très mal) en sciences actuarielles. Nous songeons notamment à l'analyse des fréquences des sinistres, ou encore à celle de l'occurrence des sinistres.

Le nombre annuel des sinistres par police étant d'ordinaire relativement faible, la régression logistique fut proposée pour décrire la sinistralité des assurés (les deux modalités étant "0 sinistre", ou "1 sinistre et plus"), sans doute inspiré par le modèle individuel de théorie du risque. Néanmoins, cet outil n'est qu'une approximation, et devient clairement inapproprié lorsque le nombre annuel de sinistre peut prendre de grandes valeurs, comme dans les assurances pour entreprises. De plus, la régression logistique, qui a connu son heure de gloire en assurance automobile, obligeait les actuaires à effectuer une dernière pirouette, dans l'esprit du passage du modèle individuel au modèle collectif en théorie du risque, pour passer de la loi de Bernoulli à la loi de Poisson, modèle incontournable lorsqu'il s'agit de créer un système bonus-malus. Pour plus de détails quant à l'incompatibilité des modèles de Bernoulli et de Poisson dans le contexte de la tarification *a priori*, nous suggérons au lecteur de consulter Le Bailly de Tilleghem & Denuit (2001).

1.4 Tarification sur base de données en panel

Souvent, les actuaires utilisent plusieurs années d'observation afin de construire leur tarif (dans le but d'augmenter la taille de la base de données à analyser, et aussi pour éviter d'accorder trop d'importance à des événements relatifs à une année particulière). Ceci a notamment pour conséquence que certaines des données ne seront plus indépendantes. En effet, les observations réalisées sur un même assuré au cours des différentes périodes considérées ne sont sans doute pas indépendantes (ce qui est la raison d'être de la tarification *a posteriori*). Nous sommes donc en présence de données en panel: nous disposons d'un nombre assez réduit (typiquement moins de 5) d'observations à propos d'un grand nombre d'assurés.

Dans le cadre de la tarification *a priori*, la dépendance existant entre les observations relatives à la même police est considérée comme une nuisance: l'actuaire veut à ce stade déterminer l'impact des facteurs observables sur le risque assuré, et les corrélations existant entre les données l'empêchent de recourir aux techniques statistiques classiques (pour la plupart fondées sur l'hypothèse d'indépendance). Nous montrerons ici comment prendre cette dépendance en compte afin d'améliorer la qualité des estimations; il s'agira là d'une des innovations majeures de notre travail. Pour ce faire, nous aurons recours aux techniques proposées par Liang & Zeger (1986) et Zeger et al. (1988).

Nous verrons que les estimateurs des primes-fréquence obtenus sous l'hypothèse d'indépendance des données individuelles relatives à différentes périodes sont convergents (c'est-à-dire qu'ils tendront en probabilité vers les valeurs population si la taille de l'échantillon s'accroît indéfiniment). Dès lors, on peut raisonnablement espérer que pour des portefeuilles automobiles de grande taille, l'impact de l'hypothèse simplificatrice d'indépendance soit minime. C'est en effet ce que nous mettrons en évidence dans la partie empirique de notre étude, et qui conforte la pratique au sein des compagnies d'assurance.

2 Régression de Poisson pour des données longitudinales

2.1 Modèle

Comme nous l'avons expliqué plus haut, les compagnies d'assurance utilisent souvent plusieurs périodes d'observation pour construire leur tarif. Il n'est pas rare en pratique de travailler avec trois voire cinq années d'observation. Lors de la constitution d'une telle base de données, il est alors essentiel de tenir compte de l'aspect sériel des observations, en suivant un même assuré au cours du temps.

Les observations individuelles sont donc doublement indicées, par la police i et la période t. Dorénavant, N_{it} représente le nombre de sinistres déclarés par l'assuré i durant la période t, $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, T_i$, où T_i désigne le nombre de périodes d'observation pour l'assuré i. Nous noterons d_{it} la durée de la $t$ème période d'observation pour l'individu i. Lors de chaque modification des variables observables, un nouvel intervalle commence, de sorte que d_{it} peut être différent de 1. Nous supposons que nous disposons par ailleurs d'autres variables $\boldsymbol{x}_{it}$, connues au début de la période t, et pouvant servir de facteurs explicatifs pour la sinistralité de l'assuré i. Ces variables peuvent comporter des facteurs purement individuels indépendants du temps, des facteurs purement temporels, des facteurs dépendant des deux indices voire même des valeurs passées $N_{i,t-j}$, $j = 1, 2, \ldots,$ de la variable d'intérêt. En plus des variables explicatives, on peut introduire des indicatrices des différentes années afin de prendre en compte d'événements ponctuels ou d'éventuelles tendances dans la sinistralité, dans l'esprit de Besson & Partrat (1992).

Typiquement, nous sommes en présence de données longitudinales, ou en panel: une même variable est mesurées sur un grand nombre n d'individus au cours du temps, à un nombre $\max_{1 \leq i \leq n} T_i$ relativement faible de reprises. L'asymptotique se fera ici en faisant tendre n vers l'infini, et non pas le nombre d'observations effectuées sur un même individu (comme c'est typiquement le cas dans le cadre de l'analyse des séries chronologiques).

2.2 Première approximation: indépendance temporelle

2.2.1 Description du modèle

En première approximation, on supposera les N_{it} indépendantes pour différentes valeurs de i et de t, comme le suggère le modèle du processus de Poisson. Bien entendu, comme nous le verrons dans la suite, les observations relatives au même individu au cours du temps ne sont en général pas indépendantes du fait de l'omission de certaines variables explicatives importantes.

Nous décrirons plus loin différentes méthodes permettant de tenir compte de cette dépendance sérielle.

Nous supposerons que la loi conditionnelle de N_{it} sachant $\boldsymbol{x}_{it}$ est de Poisson et on spécifiera une moyenne de forme exponentielle linéaire, i.e.

$$N_{it} =_d \text{Poisson}\,\big(d_{it}\exp(\boldsymbol{\beta}^t\boldsymbol{x}_{it})\big),\;\; i = 1, 2, \ldots, n,\;\; t = 1, 2, \ldots, T_i. \tag{1}$$

Dorénavant, nous notons $\eta_{it} = \boldsymbol{\beta}^t\boldsymbol{x}_{it}$ le prédicteur linéaire, encore appelé score car il permet de ranger les assurés du moins risqué au plus risqué, en suivant les valeurs croissantes de η_{it}. La prime-fréquence dont doit s'acquitter l'individu i durant la période t est $\lambda_{it} = d_{it}\exp(\eta_{it})$. La prime pure s'obtient en multipliant la prime-fréquence par le coût moyen d'un sinistre.

Le modèle 1 est celui de la régression de Poisson. Outre ses applications en assurance automobile, signalons que la régression de Poisson est maintenant utilisée également en assurance sur la vie, pour l'établissement des tables de mortalité et leur projection; voyez notamment Verrall (1996), Sithole, Haberman & Verrall (2000) et Brouhns & Denuit (2001a). Elle apparaît à différents endroits en sciences actuarielles, comme dans l'évaluation des montants des réserves IBNR (voyez par exemple Kaas et al. (2001) pour une présentation simple et unifiée).

2.2.2 Types de variables explicatives et codage associé

Les variables explicatives composant $\boldsymbol{x}_{it}$ peuvent être de différents types. Certaines d'entre elles peuvent être quantitatives et continues (comme la puissance de la voiture ou l'âge de l'assuré par exemple). D'autres variables explicatives dont l'assureur dispose à propos de ses assurés peuvent être quantitatives discrètes (le nombre d'enfants de l'assuré, par exemple). D'autres encore sont qualitatives ou catégorielles (comme le sexe ou l'état-civil de l'assuré, par exemple).

Il est important de remarquer que la linéarité dont il est question dans les modèles utilisés s'entend exclusivement par rapport aux paramètres. Cela n'exclut donc pas de remplacer une variable explicative x par une fonction $g(x)$ de celle-ci.

En plus des données individuelles dont l'assureur dispose, une stratégie intéressante consiste à construire des indices résumant les caractéristiques de la zone d'habitation de l'assuré à partir de données publiques, comme celles issues d'un recensement. En condensant l'information pertinente à propos de la sinistralité disponible auprès d'un institut national de statistique, on peut ainsi obtenir de nouvelles variables explicatives dont le pouvoir prédictif (reconnu par le modèle) est souvent très important. Les techniques de l'analyse des données permettent de construire de tels indices. Pour plus de détails, voyez Brouhns, Cullus & Denuit (2001). De même, les caractéristiques des véhicules peuvent être résumées dans un, voire deux, indices, dans l'esprit d'Ingenbleeck & Lemaire (1988). Enfin, la prise en compte de la variabilité spatiale des risques se fait en fin d'analyse, en travaillant sur les résidus, comme décrit dans Boskov & Verrall (1994) et Brouhns, Denuit, El Ghelbzouri & Verrall (2001).

Dorénavant, nous supposerons, comme c'est le cas en pratique, que toutes les variables sont catégorielles; pour plus de détails quant au traitement des variables continues, voyez Brouhns & Denuit (2001b). On convient alors de coder tout facteur partitionnant la population en k catégories par les entiers $0, 1, \ldots, k-1$. Certains facteurs peuvent être ordinaux et résulter d'une variable quantitative (comme la puissance d'un véhicule que l'on a codée en différentes classes), soit ordinaux mais sans échelle quantitative (comme le niveau d'études) ou encore être purement qualitatif, sans induire d'ordre (comme le sexe). Une variable catégorielle à k facteurs est généralement codée par $k-1$ variables binaires qui sont toutes nulles pour le niveau de référence.

La plupart du temps, les variables explicatives sont toutes catégorielles dans un tarif commercial. Considérons par exemple une compagnie segmentant selon le sexe, le caractère sportif du véhicule et l'âge de l'assuré (3 classes d'âges, à savoir moins de 30 ans, 30-65 ans et plus de 65 ans). Un assuré sera représenté par un vecteur binaire donnant les valeurs des variables

$$
\begin{aligned}
X_1 &= \begin{cases} 0 \text{ si l'assuré est un homme} \\ 1 \text{ si l'assuré est une femme} \end{cases} \\
X_2 &= \begin{cases} 0 \text{ si le véhicule n'a pas de caractère sportif} \\ 1 \text{ si le véhicule a un caractère sportif} \end{cases} \\
X_3 &= \begin{cases} 1 \text{ si l'assuré a moins de 30 ans} \\ 0 \text{ sinon} \end{cases} \\
X_4 &= \begin{cases} 1 \text{ si l'assuré a plus de 65 ans} \\ 0 \text{ sinon.} \end{cases}
\end{aligned}
$$

On choisit généralement comme niveau de référence (i.e. celui pour lequel toutes les X_i valent 0) les modalités les plus représentées dans le portefeuille; ici, le niveau de référence correspond à un homme dans la tranche d'âges 30-65 ans conduisant un véhicule non sportif. Les résultats s'interprèteront ensuite comme une sur- ou sous-sinistralité par rapport à cette classe de référence. Ainsi, le vecteur (0,1,1,0) représente un assuré masculin de moins de 30 ans conduisant un véhicule sportif. Le prédicteur linéaire (ou score) sera de la forme $\beta_0 + \sum_{j=1}^{4} \beta_j X_j$; l'intercept β_0 représente donc le risque associé à la classe de référence (i.e. celle pour laquelle $X_i = 0$ pour tout i, à savoir les hommes entre 30 et 65 ans dont le véhicule n'a pas de caractère sportif); si $\beta_j > 0$, cela indique que le fait de présenter la modalité traduite par X_j est un facteur aggravant la sinistralité (par rapport à l'individu de référence), au contraire $\beta_j < 0$ indiquera les classes d'assurés moins risqués que les individus de référence.

Chaque assuré est donc représenté par un vecteur $\boldsymbol{x}_{it}$ dont les composantes valent 0 ou 1. Dans ce cas, la prime-fréquence annuelle λ_{it} apparaît comme un produit de coefficients de majoration ou de réduction par rapport à la prime-fréquence de l'individu de référence du portefeuille. Plus précisément,

$$\lambda_{it} = \exp(\beta_0) \prod_{j=1}^{p} \exp(\beta_j x_{itj}),$$

où p désigne le nombre de variables binaires résultant du codage expliqué ci-dessus; $\exp(\beta_0)$ est la prime dont doit s'acquitter l'individu de référence du portefeuille tandis que chacun des facteurs $\exp(\beta_j x_{itj})$ traduit l'influence d'un critère de segmentation (si $\beta_j > 0$, les individus présentant cette caractéristique subiront une majoration de prime par rapport à la prime de référence $\exp(\beta_0)$, tandis que $\beta_j < 0$ indiquera une réduction de prime).

2.2.3 Estimation par maximum de vraisemblance

Notons n_{it} le nombre de sinistres déclarés par l'assuré i durant la période t; nous considérons en première approximation les n_{it}, $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, T_i$, comme des réalisations de variables aléatoires N_{it} indépendantes de loi de Poisson de moyennes respectives λ_{it}. La vraisemblance associée à ces observations vaut alors

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} \exp\{-\lambda_{it}\} \frac{\{\lambda_{it}\}^{n_{it}}}{n_{it}!};$$

il s'agit de la probabilité d'obtenir les observations réalisées au sein du portefeuille dans le modèle considéré (notez que $\mathcal{L}$ est une fonction des paramètres, les observations étant supposées connues). L'estimation de $\boldsymbol{\beta}$ par la méthode du maximum de vraisemblance consiste à déterminer

$\widehat{\boldsymbol{\beta}}$ en maximisant $\mathcal{L}(\boldsymbol{\beta})$. Afin de faciliter l'obtention du maximum, on passe souvent à la log-vraisemblance, laquelle est donnée par

$$L(\boldsymbol{\beta}) = \ln \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \Big\{ -\ln n_{it}! + n_{it}(\eta_{it} + \ln d_{it}) - \lambda_{it} \Big\}.$$

Comme L est une fonction concave en le paramètre $\boldsymbol{\beta}$, les conditions du premier ordre sont nécessaires et suffisantes pour caractériser l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$. Cette concavité rend également plus facile l'application des procédures numériques d'optimisation de la log-vraisemblance. Les conditions au premier ordre sont

$$\frac{\partial}{\partial \beta_0} L(\boldsymbol{\beta}) = 0 \Leftrightarrow \sum_{i=1}^{n} \sum_{t=1}^{T_i} n_{it} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \lambda_{it} \tag{2}$$

et pour $j = 1, 2, \ldots, p$,

$$\begin{aligned} \frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}) = 0 \quad &\Leftrightarrow \quad \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{itj} \Big\{ n_{it} - \lambda_{it} \Big\} = 0 \\ &\Leftrightarrow \quad \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{itj} \Big\{ n_{it} - \mathbb{E}[N_{it} | \boldsymbol{x}_{it}] \Big\} = 0. \end{aligned} \tag{3}$$

Si on définit le résidu-fréquence $nres_{it}$ relatif à l'individu i et à la période t comme

$$nres_{it} = n_{it} - \lambda_{it} = n_{it} - \mathbb{E}[N_{it} | \boldsymbol{x}_{it}],$$

on peut interpréter les équation de vraisemblance (3) comme une relation d'orthogonalité entre les variables explicatives $\boldsymbol{x}_{it}$ et les résidus d'estimation $nres_{it}$. Cette orthogonalité peut s'interpréter comme une “indépendance” entre les résidus d'estimation et les variables explicatives, signifiant que les variables explicatives n'ont aucun pouvoir prédictif des résidus $nres_{it}$.

2.2.4 Signification tarifaire des équations de vraisemblance

Comme les variables explicatives sont les indicatrices des niveaux des facteurs de risque, les équations de vraisemblance (3) ont une signification tarifaire très importante. Elles garantissent que pour chaque sous-portefeuille correspondant à un niveau d'un des facteurs de risque, le nombre total des sinistres observés est égal à son homologue théorique. En effet, supposons par exemple que $x_{it1} = 1$ si l'individu i est un homme, et 0 sinon; (3) garantit alors pour $j = 1$ que

$$\sum_{\text{hommes}} n_{it} = \sum_{\text{hommes}} \widehat{\lambda}_{it},$$

c'est-à-dire que les primes-fréquences supportées par les hommes compensent exactement les sinistres causés par ceux-ci. Il n'y a donc pas de subsidiation entre les différentes catégories d'assurés dans le tarif multiplicatif ainsi obtenu.

De plus, en vertu de (2) la somme des primes-fréquence est égale au nombre total de sinistres déclarés, puisque

$$\sum_{i=1}^{n} \sum_{t=1}^{T_i} \widehat{\lambda}_{it} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} n_{it}$$

pour autant qu'un intercept β_0 soit inclus dans le score η_{it} (c'est-à-dire pour autant que les primes-fréquence soient exprimées par rapport à une prime de référence $\exp(\beta_0)$).

2.2.5 Variance asymptotique des estimateurs

La variance asymptotique de $\widehat{\boldsymbol{\beta}}$ peut être calculée comme l'inverse de la matrice d'information de Fisher $\mathcal{I}$ donnée par

$$\mathcal{I} = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \boldsymbol{x}_{it} \boldsymbol{x}_{it}^t \lambda_{it}$$

de sorte que la matrice variance-covariance $\boldsymbol{\Sigma}$ de l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$ peut être estimée par

$$\widehat{\boldsymbol{\Sigma}} = \left\{ \sum_{i=1}^{n} \sum_{t=1}^{T_i} \boldsymbol{x}_{it} \boldsymbol{x}_{it}^t \widehat{\lambda_{it}} \right\}^{-1}.$$

Cette expression n'est valable que si les données sont indépendantes.

En vertu de la théorie asymptotique de la méthode du maximum de vraisemblance, $\widehat{\boldsymbol{\beta}}$ est approximativement de loi normale de moyenne la vraie valeur du paramètre et de matrice variance-covariance $\widehat{\boldsymbol{\Sigma}}$; ceci permet d'obtenir des intervalles et des zones de confiance pour les paramètres.

2.2.6 Méthodes de sélection des variables explicatives

Trois approches existent dans la plupart des logiciels: forward, backward et stepwise. La procédure forward part d'un modèle sans variables explicatives (mais comportant uniquement un intercept β_0, donc supposant les observations identiquement distribuées) et incorpore un à un les facteurs de risque jugés les plus pertinents sur base de la comparaison des log-vraisemblances. Aucune variable n'est plus introduite lorsque la prise en compte de celles-ci ne rend pas le modèle significativement meilleur. Cette première approche est donc fort semblable à celle de la segmentation du portefeuille selon un arbre de classification. Le portefeuille est éclaté en sous-portefeuilles à mesure que de nouvelles variables sont incorporées au modèle.

La procédure backward quant à elle part du portefeuille le plus segmenté et regroupe les classes définies à partir des facteurs les moins pertinents. Ainsi, si le facteur "présence d'airbags" est jugé non pertinent (car son omission ne détériore pas significativement le modèle), on regroupera les classes correspondant aux différentes modalités de ce facteur.

L'approche stepwise conjugue l'esprit des deux algorithmes précédents (elle peut se voir comme une procédure forward pour laquelle, après chaque inclusion de variable explicative, on se demande si une des variables entrées dans le modèle ne pourrait pas être supprimée).

La procédure `GENMOD` de SAS, qui permet d'effectuer la régression de Poisson, n'offre pas les procédures décrites ci-dessus (au contraire de `LOGISTIC` et `GLM`, par exemple), mais bien des analyses de types 1 et 3. L'analyse de type 1 introduit une à une les variables dans le modèle, dans l'ordre dans lequel elles ont été spécifiées dans `MODEL`. Les résultats de cette analyse dépendent donc de cet ordre, somme toute arbitraire. Un test du rapport de vraisemblance est effectué entre deux modèles successifs emboîtés; cela permet de se faire une idée de la pertinence de la dernière variable introduite, compte tenu de celles déjà incorporées au modèles. L'analyse de type 1 diffère de la procédure forward en ce que les variables sont introduites dans l'ordre dans lequel elles ont été spécifiées par l'utilisateur, et pas en fonction de leur pouvoir prédictif.

On préférera donc l'analyse de type 3 à son homoloque de type 1. Cette analyse comparera le modèle complet (c'est-à-dire comprenant toutes les variables spécifiées dans `MODEL`) avec les différents modèles obtenus en supprimant une des variables. Ceci permet de tester la pertinence de chacune des variables explicatives, compte tenu des autres. Il s'agit donc de l'optique backward de sélection des variables tarifaires: à chaque étape, on excluera la variable possédant la p-valeur la plus élevée, jusqu'à ce qu'aucune variable ne puisse plus être exclue (i.e. jusqu'à ce que toutes les p-valeurs soient inférieures à un seuil choisi par l'utilisateur, en général 5%). Il convient ici d'insister sur le fait que l'analyse de type 3 travaille avec les variables, et pas avec les différents niveaux de celles-ci. Ainsi, une variable jugée pertinente à l'issue de l'analyse de type 3 pourrait comporter certains niveaux non-significatifs.

2.2.7 Qualité de l'ajustement

Une fois le modèle ajusté (i.e. les variables explicatives pertinentes sélectionnées et l'estimation du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ obtenue), il est primordial d'en évaluer la qualité, c'est-à-dire son habileté à décrire le nombre des sinistres touchant les différents assurés du portefeuille. Cette évaluation peut se faire à l'aide de statistiques mesurant la qualité de l'ajustement global fourni par le modèle.

Notons $L(\widehat{\boldsymbol{\lambda}})$ la log-vraisemblance du modèle ajusté, où $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_{11}, \widehat{\lambda}_{12}, \ldots, \widehat{\lambda}_{nT_n})$. La log-vraisemblance maximale qu'il est possible d'obtenir dans le modèle spécifiant que les N_{it} sont des variables indépendantes de loi de Poisson s'obtient lorsqu'il y a autant de paramètres que d'observations; notons $L(\boldsymbol{n})$ la log-vraisemblance de ce modèle (qui prédira n_{it} pour la $i$ème observation au cours de l'année t). La déviance est alors définie comme

$$D(\boldsymbol{n}, \widehat{\boldsymbol{\lambda}}) = 2\Big\{L(\boldsymbol{n}) - L(\widehat{\boldsymbol{\lambda}})\Big\},$$

soit comme deux fois la différence entre la log-vraisemblance maximale et celle du modèle considéré. Dans notre cas,

$$\begin{aligned} D(\boldsymbol{n}, \widehat{\boldsymbol{\lambda}}) &= 2\ln\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\exp(-n_{it})\frac{n_{it}^{n_{it}}}{n_{it}!}\right\} - 2\ln\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\exp(-\widehat{\lambda}_{it})\frac{\widehat{\lambda}_{it}^{n_{it}}}{n_{it}!}\right\} \\ &= 2\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left\{n_{it}\ln\frac{n_{it}}{\widehat{\lambda}_{it}} - (n_{it} - \widehat{\lambda}_{it})\right\} \end{aligned}$$

où l'on a posé $y \ln y = 0$ lorsque $y = 0$. Puisque l'inclusion d'un intercept β_0 garantit que (2) est valable, la déviance s'écrit dans ce cas

$$D(\boldsymbol{n}, \widehat{\boldsymbol{\lambda}}) = 2\sum_{i=1}^{n}\sum_{t=1}^{T_i} n_{it}\ln\frac{n_{it}}{\widehat{\lambda}_{it}}.$$

La déviance noue des liens étroits avec la statistique du rapport de vraisemblance: ainsi, lorsqu'il s'agit de comparer deux modèles, la différence des déviances qui leur sont associées donnera la valeur de la statistique du rapport de vraisemblance.

L'analyse des résidus permet de détecter les observations pour lesquelles le modèle ne fournit pas une prédiction satisfaisante, i.e. celles pour lesquelles la valeur observée n_{it} de N_{it} et sa valeur prédite $\widehat{\lambda}_{it}$ diffèrent trop. L'analyse des résidus permet aussi de détecter les défauts du modèle et suggère souvent la façon d'y remédier.

Les résidus de déviance se définissent à partir de la différence entre la log-vraisemblance évaluée en l'observation n_{it} et celle évaluée en la valeur prédite $\widehat{\lambda}_{it}$. Plus précisément,

$$r_{it}^{D} = \text{signe}(n_{it} - \widehat{\lambda}_{it})\sqrt{2\left\{n_{it}\ln\frac{n_{it}}{\widehat{\lambda}_{it}} - (n_{it} - \widehat{\lambda}_{it})\right\}}$$

où $y \ln y = 0$ lorsque $y = 0$. Ainsi, $\{r_{it}^{D}\}^2$ est la contribution de l'observation à la statistique de déviance D.

On commencera par représenter les résidus en fonction des valeurs prédites $\widehat{\lambda}_{it}$. Ceci permettra de vérifier si l'ajustement est raisonnable pour les grandes (ou petites) valeurs de $\widehat{\lambda}_{it}$;

souvent en effet, un ajustement est raisonnable pour les valeurs centrales des $\widehat{\lambda}_{it}$, mais devient inadéquat pour les valeurs extrêmes des $\widehat{\lambda}_{it}$.

On peut également représenter les résidus en fonction des variables explicatives non incluses au modèle (afin de vérifier la pertinence de leur inclusion) ainsi qu'en fonction de variables explicatives retenues dans le modèle afin de détecter si une éventuelle transformation de ces variables s'avère nécessaire. Notons que ceci s'applique surtout aux variables continues.

La normalité approximative des résidus confortera le modèle. Ainsi, on ordonnera les résidus par ordre croissant

$$r^*_{(1)} \leq r^*_{(2)} \leq \ldots \leq r^*_{(T_1+\ldots+T_n)}$$

et on portera en graphique $r^*_{(j)}$ en fonction de

$$rnorm_j = \overline{r} + s_r \Phi^{-1}\left(\frac{j-0.5}{T_1+\ldots+T_n}\right), \;\; j=1,2,\ldots,T_1+\ldots+T_n,$$

où $\overline{r}$ est la moyenne des $r^*_{(j)}$ et s_r leur écart-type, et où Φ est la fonction de répartition associée à la loi normale centrée-réduite. Si les résidus sont effectivement normaux, le graphique doit laisser apparaître une ligne droite.

2.2.8 Prédiction des primes-fréquence anuelles

Pour l'assuré i et la période t, caractérisés par un vecteur de variables explicatives $\boldsymbol{x}_{it}$, la prime fréquence annuelle prédite est $\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})$. Ceci sera aussi le cas pour les nouveaux assurés présentant les mêmes caractéristiques (l'hypothèse implicite étant que les nouvelles polices sont conclues par des individus s'identifiant parfaitement aux assurés qui sont à la base de la construction du tarif; cela suppose notamment que la compagnie maîtrise parfaitement l'antisélection).

On peut également obtenir un intervalle de confiance pour la prime-fréquence annuelle. Ceci permettra d'avoir une idée quant à la précision de l'estimation de celle-ci, et guidera le choix du taux de chargement de sécurité. Partons de la variance du prédicteur linéaire $\widehat{\eta}_{it} = \boldsymbol{x}_{it}^t\widehat{\boldsymbol{\beta}}$, donnée par

$$\mathbb{V}\mathrm{ar}[\widehat{\eta}_{it}] = \boldsymbol{x}_{it}^t\widehat{\boldsymbol{\Sigma}}\boldsymbol{x}_{it}.$$

Un intervalle de confiance approximatif au niveau de confiance $1-\alpha$ pour la prime-fréquence annuelle est alors fourni par

$$\left[\exp\left(\boldsymbol{x}_{it}^t\widehat{\boldsymbol{\beta}} \pm z_{\alpha/2}\sqrt{\boldsymbol{x}_{it}^t\widehat{\boldsymbol{\Sigma}}\boldsymbol{x}_{it}}\right)\right].$$

2.2.9 Convergence des estimateurs du maximum de vraisemblance en présence de dépendance sérielle

Réécrivons sous forme matricielle les équations de vraisemblance (2)-(3):

$$\sum_{i=1}^n \boldsymbol{X}_i\big(\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]\big) = 0 \tag{4}$$

où $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iT_i})^t$. La solution $\widehat{\boldsymbol{\beta}}$ de l'équation (4) est l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$ si les N_{it} étaient indépendantes. Notons $\boldsymbol{A}_i$ la matrice diagonale

$$\boldsymbol{A}_i = \begin{pmatrix} \mathbb{V}\mathrm{ar}[N_{i1}] & 0 & \cdots & 0 \\ 0 & \mathbb{V}\mathrm{ar}[N_{i2}] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{V}\mathrm{ar}[N_{iT_i}] \end{pmatrix}.$$

On peut montrer que $\widehat{\boldsymbol{\beta}}$ est convergent pour $\boldsymbol{\beta}$ et $\sqrt{n}\big(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\big)$ est asymptotiquement de loi normale multivariée, de moyenne $\mathbf{0}$ et de matrice variance-covariance estimée par

$$\left(\sum_{i=1}^{n}\boldsymbol{X}_i^t\boldsymbol{A}_i\boldsymbol{X}_i\right)^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\left(\sum_{i=1}^{n}\boldsymbol{X}_i^t\big(\boldsymbol{n}_i-\mathbb{E}[\boldsymbol{N}_i]\big)\big(\boldsymbol{n}_i-\mathbb{E}[\boldsymbol{N}_i]\big)^t\boldsymbol{X}_i\right)_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\left(\sum_{i=1}^{n}\boldsymbol{X}_i^t\boldsymbol{A}_i\boldsymbol{X}_i\right)^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}.$$

Ainsi, la corrélation existant entre les variables $N_{i1},\ldots,N_{iT_i}$ oblige à modifier le calcul de la variance asymptotique de l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$, et donc aussi les intervalles de confiance sur les primes-fréquence.

2.3 Détection de l'aspect sériel

Afin d'avoir une première idée du type de dépendance existant entre les N_{it}, il est conseillé de considérer les observations N_{it}, $t=2,\ldots,T_i$, $i=1,\ldots,n$, et d'effectuer une régression de celles-ci sur les variables explicatives $\boldsymbol{x}_{it}$ correspondantes ainsi que le nombre $N_{i,t-1}$ de sinistres observés au cours de la période de couverture précédente. Ceci permettra de voir l'effet de l'inclusion de valeurs passées de la variable d'intérêt dans les variables explicatives.

2.4 Prise en compte de la dépendance temporelle par l'inclusion d'un effet aléatoire invariant dans le temps

2.4.1 Description du modèle

Afin de tenir compte de la dépendance naturelle qui existe entre les observations relatives à différentes périodes d'observation, nous considérerons ici le modèle suivant:

$$[N_{it}|\boldsymbol{x}_{it},\Theta_i]=_d \text{Poisson }(\lambda_{it}\Theta_i),\;\; t=1,2,\ldots,T_i,$$

où Θ_i est une variable aléatoire positive et $\lambda_{it}=d_{it}\exp(\eta_{it})$. Au niveau du portefeuille, les effets aléatoires Θ_i, $i=1,2,\ldots,n$ sont indépendants et identiquement distribués. L'effet aléatoire Θ_i permet de rendre compte de la surdispersion des données empiriques tout en générant la dépendance sérielle entre les N_{it} à i fixé. Comme l'ignorance de variables explicatives est la raison d'être de l'introduction de Θ_i dans le modèle, nous considérons donc que la dépendance entre nombres annuels de sinistres causés par un assuré est apparente, et due à notre manque d'information à propos de celui-ci. Voyez Purcaru & Denuit (2001a) pour une étude précise de la dépendance générée par le modèle classique de théorie de la crédibilité. De plus, nous supposerons que N_{it} ne dépend que de $\boldsymbol{x}_{it}$ et pas des valeurs passées $\boldsymbol{x}_{i,t-j}$, $j=1,2,\ldots$, des variables explicatives, i.e.

$$\mathbb{E}[N_{it}|\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\ldots,\boldsymbol{x}_{it},\Theta_i]=\mathbb{E}[N_{it}|\boldsymbol{x}_{it},\Theta_i]=\Theta_i\lambda_{it}.$$

On précise le plus souvent la forme de la loi conditionnelle de ces variables omises ou non observables. La loi Gamma est souvent utilisée dans ce cadre, en raison des facilités techniques résultant de son application de concert avec la loi de Poisson. On peut espérer que les approximations successives $\widehat{\Theta}_{it}=\mathbb{E}[\Theta_i|N_{i1}=n_{i1},\ldots,N_{it}=n_{it}]$ en fonction de l'information disponible vont tendre vers Θ_i lorsque l'information s'accroît ($t\to+\infty$). Au cours du temps, on aura alors une adaptation de la prime due à cet apprentissage du facteur Θ_i sous-jacent. C'est ce phénomène d'apprentissage de Θ_i qui est à la base du mécanisme du Bonus-Malus en assurance RC auto.

2.4.2 Loi *a priori* de Θ_i et des N_{it}

Le modèle classique suppose que conditionnellement aux caractéristiques observables $\boldsymbol{x}_{it}$, les effets aléatoires Θ_i, $i = 1, 2, \ldots, n$, sont indépendants et de même loi Gamma de moyenne 1 et de variance $1/a$; la densité de Θ_i est donc donnée par

$$f_\Theta(\theta) = \frac{1}{\Gamma(a)} a^a \theta^{a-1} \exp(-a\theta), \;\; \theta \in \mathbb{R}^+.$$

Le choix de la densité f_Θ se justifie par des considérations purement analytiques (la loi Gamma étant la loi conjuguée à la loi de Poisson). Conditionnellement aux variables observables $\boldsymbol{x}_{it}$ et à l'effet aléatoire Θ_i, la loi de N_{it} est donnée par

$$\mathbb{P}[N_{it} = n_{it} | \boldsymbol{x}_{it}, \Theta_i = \theta_i] = \exp\Big(- \theta_i d_{it} \exp(\eta_{it}) \Big) \frac{\big\{\theta_i d_{it} \exp(\eta_{it})\big\}^{n_{it}}}{n_{it}!}, \;\; n_{it} \in \mathbb{N}.$$

Non conditionnellement, N_{it} est de loi Binomiale Négative, et les probabilités sont

$$\begin{aligned} \mathbb{P}[N_{it} = n_{it} | \boldsymbol{x}_{it}] &= \int_{\theta \in \mathbb{R}^+} \mathbb{P}[N_{it} = n_{it} | \boldsymbol{x}_{it}, \Theta_i = \theta] f_\Theta(\theta) d\theta \\ &= \binom{a + n_{it} - 1}{n_{it}} \left(\frac{d_{it} \exp(\eta_{it})}{a + d_{it} \exp(\eta_{it})} \right)^{n_{it}} \left(\frac{a}{a + d_{it} \exp(\eta_{it})} \right)^{a}, \end{aligned}$$

pour $n_{it} \in \mathbb{N}$.

2.4.3 Loi *a posteriori* de Θ_i et des N_{it}

La loi de probabilité du vecteur aléatoire $\boldsymbol{N}_i = (N_{i1}, N_{i2}, \ldots, N_{iT_i})$ est décrite par

$$\begin{aligned} & \mathbb{P}[N_{i1} = n_{i1}, N_{i2} = n_{i2}, \ldots, N_{iT_i} = n_{iT_i}] \\ = & \int_{\theta \in \mathbb{R}^+} \mathbb{P}[N_{i1} = n_{i1}, N_{i2} = n_{i2}, \ldots, N_{iT_i} = n_{iT_i} | \Theta_i = \theta] f_\Theta(\theta) d\theta \\ = & \int_{\theta \in \mathbb{R}^+} \left\{ \prod_{t=1}^{T_i} \mathbb{P}[N_{it} = n_{it} | \Theta_i = \theta] \right\} f_\Theta(\theta) d\theta \\ = & \int_{\theta \in \mathbb{R}^+} \left\{ \prod_{t=1}^{T_i} \exp(-\theta \lambda_{it}) \frac{(\theta \lambda_{it})^{n_{it}}}{n_{it}!} \right\} f_\Theta(\theta) d\theta \\ = & \left\{ \prod_{t=1}^{T_i} \frac{\lambda_{it}^{n_{it}}}{n_{it}!} \right\} \left(\frac{a}{a + \sum_{t=1}^{T_i} \lambda_{it}} \right)^a \left(a + \sum_{t=1}^{T_i} \lambda_{it} \right)^{-\sum_{t=1}^{T_i} n_{it}} \frac{\Gamma\left(a + \sum_{t=1}^{T_i} n_{it}\right)}{\Gamma(a)}. \end{aligned} \tag{5}$$

La densité jointe des variables N_{it}, $t = 1, 2, \ldots, T_i$ et Θ_i est donnée par

$$\begin{aligned} & \prod_{t=1}^{T_i} \exp\Big(- \theta_i d_{it} \exp(\eta_{it}) \Big) \frac{\big\{\theta_i d_{it} \exp(\eta_{it})\big\}^{n_{it}}}{n_{it}!} \frac{1}{\Gamma(a)} a^a \theta_i^{a-1} \exp(-a\theta_i) \\ \propto & \;\; \exp\left\{ -\theta_i \sum_{t=1}^{T_i} d_{it} \exp(\eta_{it}) \right\} \theta_i^{\sum_{t=1}^{T_i} n_{it} + a - 1} \exp(-a\theta_i). \end{aligned} \tag{6}$$

La loi conditionnelle de Θ_i sachant les sinistres passés de l'assuré, c'est-à-dire sachant $N_{it} = n_{it}$,

$t=1,2,\dots,T_i$, est obtenue en faisant le rapport de (5) et (6), ce qui donne

$$
\begin{aligned}
&\frac{\exp\left\{-\theta_i\left(a+\sum_{t=1}^{T_i} d_{it}\exp(\eta_{it})\right)\right\}\theta_i^{a+\sum_{t=1}^{T_i} n_{it}-1}}{\int_{\xi\in\mathbb{R}^+}\exp\left\{-\xi\left(a+\sum_{t=1}^{T_i} d_{it}\exp(\eta_{it})\right)\right\}\xi^{a+\sum_{t=1}^{T_i} n_{it}-1}d\xi}\\
= \quad &\exp\left\{-\theta_i\left(a+\sum_{t=1}^{T_i} d_{it}\exp(\eta_{it})\right)\right\}\theta_i^{a+\sum_{t=1}^{T_i} n_{it}-1}\frac{\left(a+\sum_{t=1}^{T_i} d_{it}\exp(\eta_{it})\right)^{a+\sum_{t=1}^{T_i} n_{it}}}{\Gamma\left(a+\sum_{t=1}^{T_i} n_{it}\right)}.
\end{aligned}
$$

Conditionnellement aux sinistres causés durant les T_i périodes passées par l'assuré i dans le portefeuille, Θ_i suit donc une loi Gamma dont les deux premiers moments sont

$$
\begin{aligned}
\mathbb{E}[\Theta_i|N_{it}=n_{it},t=1,2,\dots,T_i] &= \frac{a+\sum_{t=1}^{T_i} n_{it}}{a+\sum_{t=1}^{T_i} d_{it}\exp(\eta_{it})}\\
\mathbb{V}\mathrm{ar}[\Theta_i|N_{it}=n_{it},t=1,2,\dots,T_i] &= \frac{a+\sum_{t=1}^{T_i} n_{it}}{\left(a+\sum_{t=1}^{T_i} d_{it}\exp(\eta_{it})\right)^2}.
\end{aligned}
$$

La première de ces deux quantités multipliée par λ_{i,T_i+1} est donc l'anticipation de la fréquence annuelle de sinistres pour la période T_i+1.

2.4.4 Estimation des paramètres par la méthode des moments

Il est assez simple d'obtenir une estimation de $\sigma^2=\mathbb{V}\mathrm{ar}[\Theta_i]$ à l'aide de la méthode des moments. Pour ce faire, remarquons que

$$
\mathbb{E}[N_{it}] = d_{it}\exp(\eta_{it})
$$

et

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}[N_{it}] &= \mathbb{E}\Big[\mathbb{V}\mathrm{ar}[N_{it}|\Theta_i]\Big]+\mathbb{V}\mathrm{ar}\Big[\mathbb{E}[N_{it}|\Theta_i]\Big]\\
&= \mathbb{E}\Big[\Theta_i d_{it}\exp(\eta_{it})\Big]+\mathbb{V}\mathrm{ar}\Big[\Theta_i d_{it}\exp(\eta_{it})]\Big]\\
&= d_{it}\exp(\eta_{it})+\Big\{d_{it}\exp(\eta_{it})\Big\}^2\sigma^2\\
&= \mathbb{E}[N_{it}]+\Big\{d_{it}\exp(\eta_{it})\Big\}^2\sigma^2 && (7)\\
& && (8)
\end{aligned}
$$

L'équation

$$
\sum_{i=1}^{n}\sum_{t=1}^{T_i}\Big\{n_{it}-d_{it}\exp(\eta_{it})\Big\}\boldsymbol{x}_{it}=0
$$

dont l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ est solution peut se voir comme l'équivalent empirique de l'équation

$$
\mathbb{E}\left[\sum_{i=1}^{n}\sum_{t=1}^{T_i}\{N_{it}-d_{it}\exp(\eta_{it})\}\boldsymbol{x}_{it}\right]=0, \tag{9}
$$

laquelle est valable à la fois dans le modèle *a priori* et dans le modèle avec effet aléatoire. Afin d'estimer σ^2, écrivons une relation semblable à (9) pour la variance, à savoir

$$
\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left\{\Big(n_{it}-d_{it}\exp(\eta_{it})\Big)^2-n_{it}-\Big(d_{it}\exp(\eta_{it})\Big)^2\sigma^2\right\}=0. \tag{10}
$$

Les estimateurs $\widehat{\boldsymbol{\beta}}$ et $\widehat{\sigma}$ sont solutions du système formé par (9) et (10); ils sont convergents dans le modèle à effet aléatoire. Dès lors, l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$, solution de (9), est convergent dans le modèle à effet aléatoire. L'estimateur de σ^2 est quant à lui donné par

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left\{\Big(n_{it}-d_{it}\exp(\widehat{\eta}_{it})\Big)^2-n_{it}\right\}}{\sum_{i=1}^{n}\sum_{t=1}^{T_i}\Big\{d_{it}\exp(\widehat{\eta}_{it})\Big\}^2}$$

où $\widehat{\eta}_{it}=\widehat{\boldsymbol{\beta}}^t\boldsymbol{x}_{it}$, $\widehat{\boldsymbol{\beta}}$ étant l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$ dans le modèle sans effet aléatoire.

2.4.5 Estimation des paramètres par la méthode du maximum de vraisemblance

Si on retient la loi gamma pour Θ_i, la vraisemblance s'écrit

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) &= \prod_{i=1}^{n}\Pr[N_{i1}=n_{i1},N_{i2}=n_{i2},\ldots,N_{iT_i}=n_{iT_i}]\\ &= \prod_{i=1}^{n}\left\{\prod_{t=1}^{T_i}\frac{\lambda_{it}^{n_{it}}}{n_t!}\right\}\left(\frac{a}{a+\sum_{t=1}^{T_i}\lambda_{it}}\right)^a\left(a+\sum_{t=1}^{T_i}\lambda_{it}\right)^{-\sum_{t=1}^{T_i}n_{it}}\frac{\Gamma\left(a+\sum_{t=1}^{T_i}n_{it}\right)}{\Gamma(a)}.\end{aligned}$$

Les estimateurs du maximum de vraisemblance des paramètres $\boldsymbol{\beta}$ et de a sont solutions de l'équation

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}\boldsymbol{x}_{it}\left(n_{it}-\lambda_{it}\frac{a+\sum_{t=1}^{T_i}n_{it}}{a+\sum_{t=1}^{T_i}\lambda_{it}}\right)=0 \tag{11}$$

qui peut également se voir comme une relation d'orthogonalité entre les variables explicatives et les résidus qui sont obtenus en faisant la différence entre l'observation n_{it} relative à l'assuré i durant la période t et son anticipation sachant $N_{it}=n_{it}$, $t=1,2,\ldots,T_i$, à savoir

$$\lambda_{it}\mathbb{E}[\Theta_i|N_{it}=n_{it},\;\; t=1,2,\ldots,T_i]=\lambda_{it}\frac{a+\sum_{t=1}^{T_i}n_{it}}{a+\sum_{t=1}^{T_i}\lambda_{it}},$$

basée sur l'ensemble des observations relatives à cet assuré.

Il y a une différence fondamentale entre les équations (2)-(3) et (11). En effet, (2)-(3) sont basées sur l'hypothèse d'indépendance temporelle: le nombre de sinistres causés par l'assuré au cours d'une période d'observation n'apprend donc rien sur ce qui se passe durant une autre. Au contraire, (11) reconnaît la dépendance sérielle entre les données relatives à un même assuré. Dès lors, ces données sont combinées pour obtenir une réévaluation de la fréquence de sinistre, laquelle intervient dans la relation “d'orthogonalité”.

Les équations de vraisemblance (11) n'admettent pas de solution explicite; on aura donc recours à des méthodes de résolution numérique, en utilisant les estimations fournies par la méthode des moments comme valeurs initiales.

2.5 Prise en compte de la dépendance temporelle: les GEE

2.5.1 Le principe

La dépendance existant entre les observations individuelles relatives à différentes années doit être considérée comme une nuisance. En effet, l'actuaire n'est pas réellement intéressé par l'évolution de la sinistralité au cours du temps, mais plutôt par l'influence des variables explicatives $\boldsymbol{x}_{it}$ sur les nombres de sinistres N_{it}. Lorsque $T_i \geq 2$, la dépendance existant entre $N_{i1}, N_{i2}, \ldots$ ne peut en général pas être négligée. La méthode des GEE (pour l'anglais "Generalized Estimating Equation") proposée par LIANG & ZEGER (1986) ne spécifie pas la loi de probabilité jointe du vecteur $\boldsymbol{N}_i$, mais retient une modélisation linéaire généralisée (grâce à la loi de Poisson dans notre cas) pour chacune des composantes N_{it}. Les estimateurs fournis par cette méthode sont convergents; on espère donc que les estimations ainsi obtenues seront de bonne qualité vu le grand nombre d'observations dont dispose en général l'actuaire.

L'approche du maximum de vraisemblance décrite plus haut avec un effet aléatoire de loi gamma, bien que théoriquement satisfaisante, est difficile à mettre en pratique (pour des raisons numériques). D'autant plus difficile en fait qu'on complique le modèle, en substituant d'autres lois de probabilité à la loi gamma, ou en permettant aux effets aléatoires d'évoluer au cours du temps. La méthode décrite dans cette section remédie à ces inconvénients. Elle allie simplicité et très grande flexibilité dans la modélisation.

Notons $\boldsymbol{V}_i$ la matrice variance-covariance du vecteur $\boldsymbol{N}_i$. Cette matrice ne sera sans doute pas diagonale, comme dans le cas où les N_{it} sont indépendantes. L'adaptation de l'équation (4) au cas sériel est

$$\sum_{i=1}^{n} \boldsymbol{X}_i^t \boldsymbol{A}_i^t \boldsymbol{V}_i^{-1} \big(\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]\big) = 0 \tag{12}$$

Il est important de comprendre qu'on reconnaît ici la dépendance entre les N_{it} sans spécifier complètement la loi de $\boldsymbol{N}_i$. Il n'y a donc plus de fonction de vraisemblance à maximiser, et cette méthode est donc très différente de celle suivie jusqu'ici.

L'estimateur $\widehat{\boldsymbol{\beta}}^{\text{GEE}}$ solution de (12) peut s'obtenir à partir du schéma itératif

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{j+1}^{\text{GEE}} &= \widehat{\boldsymbol{\beta}}_{j}^{\text{GEE}} - \left\{\sum_{i=1}^{n} \boldsymbol{D}_i^t\left(\widehat{\boldsymbol{\beta}}_{j}^{\text{GEE}}\right)\boldsymbol{V}_i^{-1}\left(\widehat{\boldsymbol{\beta}}_{j}^{\text{GEE}}\right)\boldsymbol{D}_i\left(\widehat{\boldsymbol{\beta}}_{j}^{\text{GEE}}\right)\right\}^{-1} \\ &\qquad \left\{\sum_{i=1}^{n} \boldsymbol{D}_i^t\left(\widehat{\boldsymbol{\beta}}_{j}^{\text{GEE}}\right)\boldsymbol{V}_i^{-1}\left(\widehat{\boldsymbol{\beta}}_{j}^{\text{GEE}}\right)\boldsymbol{S}_i\left(\widehat{\boldsymbol{\beta}}_{j}^{\text{GEE}}\right)\right\}\end{aligned}$$

où

$$\boldsymbol{D}_i = \boldsymbol{A}_i \boldsymbol{X}_i \text{ et } \boldsymbol{S}_i = \boldsymbol{N}_i - \mathbb{E}[\boldsymbol{N}_i].$$

Cette méthode peut être interprétée comme des moindres carrés itératifs.

2.5.2 Prise en compte de la dépendance temporelle: Effets aléatoires évoluant dans le temps

Considérons les processus stationnaires $\boldsymbol{\Theta}_i = \{\Theta_{i1}, \Theta_{i2}, \ldots\}$, $i = 1, 2, \ldots, n$, de variables aléatoires positives; $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots$ sont supposés indépendants et identiquement distribués. Conditionnellement à $\boldsymbol{\Theta}_i$, les variables aléatoires N_{it}, $t = 1, 2, \ldots, T_i$, représentant la suite des nombres annuels des sinistres déclarés par l'assuré i, sont supposées indépendantes et admettent pour moments conditionnels

$$\mathbb{E}[N_{it}|\boldsymbol{\Theta}_i, \boldsymbol{x}_{it}] = \lambda_{it}\Theta_{it} \text{ et } \mathbb{V}\text{ar}[N_{it}|\boldsymbol{\Theta}_i, \boldsymbol{x}_{it}] = \lambda_{it}\Theta_{it}.$$

Comme l'ont fort bien montré PINQUET, GUILLEN & BOLANCÉ (2001), l'intérêt des effets aléatoires évoluant avec le temps dans la construction d'un tarif d'expérience est de tenir compte de l'ancienneté des sinistres. On sent bien intuitivement que le pouvoir prédictif d'un sinistre doit décroître avec son ancienneté; ceci n'est pas le cas dans les systèmes basés sur les modèles de crédibilité avec un effet aléatoire invariant dans le temps.

Dans notre cadre de travail (tarification *a priori*), l'invariance des effets aléatoires dans le temps peut paraître restrictive. En effet, ceux-ci rendent compte de l'information cachée relative aux assurés. Tout comme les caractéristiques observables, les variables cachées peuvent évoluer dans le temps. Ceci est d'autant plus vrai que l'effet aléatoire modélise entre autre les actions entreprises par l'assuré afin de diminuer sa sinistralité; l'étendue de telles actions sera évidemment variable au cours du temps.

Supposons en outre que

$$\mathbb{E}[\Theta_{it}] = 1 \text{ quel que soit } t$$

et

$$\mathbb{C}\text{ov}[\Theta_{it}, \Theta_{i,t+\tau}] = \sigma^2 \rho_\Theta(\tau)$$

où σ^2 et $\rho_\Theta(\cdot)$ désignent respectivement la variance commune des Θ_{it} et la fonction autocorrélation des processus $\boldsymbol{\Theta}_i$. On obtient alors

$$\begin{aligned}
\mathbb{E}[N_{it}] &= \mathbb{E}[\lambda_{it}\Theta_{it}] = \lambda_{it} \\
\mathbb{V}\text{ar}[N_{it}] &= \mathbb{E}\Big[\mathbb{V}\text{ar}[N_{it}|\boldsymbol{\Theta}_i]\Big] + \mathbb{V}\text{ar}\Big[\mathbb{E}[N_{it}|\boldsymbol{\Theta}_i]\Big] \\
&= \mathbb{E}[\lambda_{it}\Theta_{it}] + \mathbb{V}\text{ar}[\lambda_{it}\Theta_{it}] = \lambda_{it}(1+\sigma^2\lambda_{it}) \\
\mathbb{C}\text{ov}[N_{it}, N_{i,t+\tau}] &= \mathbb{E}\Big[\mathbb{C}\text{ov}[N_{it}, N_{i,t+\tau}|\boldsymbol{\Theta}_i]\Big] + \mathbb{C}\text{ov}\Big[\mathbb{E}[N_{it}|\boldsymbol{\Theta}_i], \mathbb{E}[N_{i,t+\tau}|\boldsymbol{\Theta}_i]\Big] \\
&= \mathbb{C}\text{ov}[\lambda_{it}\Theta_{it}, \lambda_{i,t+\tau}\Theta_{i,t+\tau}] = \lambda_{it}\lambda_{i,t+\tau}\sigma^2\rho_\Theta(\tau) \\
\mathbb{C}\text{orr}[N_{it}, N_{i,t+\tau}] &= \frac{\lambda_{it}\lambda_{i,t+\tau}\sigma^2\rho_\Theta(\tau)}{\sqrt{\lambda_{it}(1+\sigma^2\lambda_{it})\lambda_{i,t+\tau}(1+\sigma^2\lambda_{i,t+\tau})}} \\
&= \frac{\rho_\Theta(\tau)}{\sqrt{\left(1+\frac{1}{\sigma^2\lambda_{it}}\right)\left(1+\frac{1}{\sigma^2\lambda_{i,t+\tau}}\right)}}
\end{aligned}$$

Il est intéressant de remarquer que

$$\mathbb{C}\text{orr}[N_{it}, N_{i,t+\tau}] \leq \mathbb{C}\text{orr}[\Theta_{it}, \Theta_{i,t+\tau}]$$

de sorte que l'autocorrélation dans $\boldsymbol{N}_i$ est toujours plus faible que celle dans $\boldsymbol{\Theta}_i$. De plus, la dépendance entre les composantes de $\boldsymbol{N}_i$ décroît par rapport à celle existant entre les $\boldsymbol{\Theta}_i$ lorsque les λ_{it} ou σ^2 décroissent. Les processus $\boldsymbol{\Theta}_i$ induisent à la fois la surdispersion et la dépendance sérielle entre les composantes de $\boldsymbol{N}_i$. Voyez Purcaru & Denuit (2001b) pour une étude précise de la dépendance entre les composantes de $\boldsymbol{N}_i$ générées par $\boldsymbol{\Theta}_i$.

2.5.3 Modélisation de la dépendance à l'aide de la "working correlation matrix"

Comme nous l'avons compris à la lecture de ce qui précède, c'est la matrice $\boldsymbol{V}_i$ qui tient compte de la dépendance entre les observations relatives à un même assuré. La matrice $T_i \times T_i$ $\boldsymbol{V}_i$ est modélisée comme

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$$

où $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \mathbb{C}\text{orr}[\boldsymbol{N}_i]$ est appelée "working correlation matrix". Il s'agit d'une matrice de corrélation de forme spécifiée dépendant d'un certain nombre de paramètres repris dans le vecteur $\boldsymbol{\alpha}$. En identifiant l'expression obtenue plus haut pour $\mathbb{C}\text{ov}[N_{it}, N_{i,t+\tau}]$ avec la décomposition de $\boldsymbol{V}_i$, on voit que $\boldsymbol{R}_i(\boldsymbol{\alpha})$ est la matrice variance-covariance associée au vecteur $\boldsymbol{\Theta}_i$.

Epinglons quelques choix intéressants pour $\boldsymbol{R}_i(\boldsymbol{\alpha})$:

1. Si $\boldsymbol{R}_i(\boldsymbol{\alpha})$ =Identité (c'est-à-dire que les composantes de $\boldsymbol{\Theta}_i$ sont non-corrélées), (12) devient
$$\sum_{i=1}^{n} \boldsymbol{X}_i^t \underbrace{\boldsymbol{A}_i^t \boldsymbol{A}_i^{-1}}_{\text{Identité}} \big(\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]\big) = 0$$
et on retrouve exactement les équations de vraisemblance (4) sous l'hypothèse d'indépendance.

2. Si les composantes de $\boldsymbol{\Theta}_i$ sont identiques, i.e. $\boldsymbol{\Theta}_i = (\Theta_i, \Theta_i, \dots)$, on a $\rho_\Theta(\tau) = 1$ et
$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}.$$
Dans ce cas, (12) fournit une méthode d'estimation alternative à la résolution numérique de (11).

3. Si les composantes de $\boldsymbol{\Theta}_i$ sont échangeables,
$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}.$$

4. Si la structure de dépendance de $\boldsymbol{\Theta}_i$ est inconnue,
$$\boldsymbol{R}_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1T_i} \\ \alpha_{12} & 1 & \cdots & \alpha_{2T_i} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1T_i} & \alpha_{2T_i} & \cdots & 1 \end{pmatrix}.$$

Les composantes du vecteur $\boldsymbol{\alpha}$ paramétrant la matrice $\boldsymbol{R}_i(\alpha)$ décrivant le type de dépendance entre les données sont à estimer sur base des observations.

3 Illustration numérique

3.1 Présentation du jeu de données

Dans cette section, nous illustrons les méthodes expliquées plus haut sur un portefeuille d'assurance belge comprenant 20 354 polices, observées durant une période de 3 ans. Pour chacune d'elles et pour chaque année sont renseignés certaines caractéristiques de l'assuré et le nombre de sinistres. Sur l'ensemble du portefeuille la fréquence annuelle moyenne est de 18.4% (ce qui est largement supérieur à la moyenne européenne).

Les variables explicatives retenues pour l'étude sont les suivantes : le sexe du conducteur (homme-femme), l'âge du conducteur (trois classes d'âge : $18-22$ ans, $23-30$ ans et >30 ans), la puissance du véhicule (trois classes de puissance : <66kW, $66-110$kW et >110kW), la taille de la ville de résidence du conducteur (grande, moyenne ou petite, en fonction du nombre d'habitants) et la couleur du véhicule (rouge ou autre). Nous avons volontairement limité le nombre de critères de segmentation afin de simplifier notre exposé.

Les Figures 1 à 5 montrent des histogrammes décrivant, pour chaque variable explicative, la répartition du portefeuille entre les différents niveaux de la variable et, pour chacun de ces niveaux, la fréquence moyenne (en %) du nombre total de sinistres.

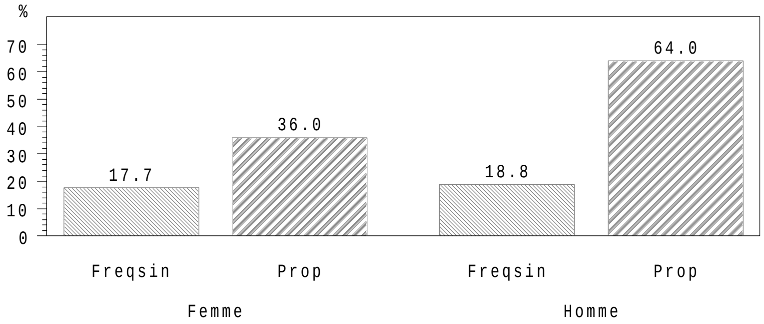

Figure 1: Répartition et fréquence de sinistres selon le sexe du conducteur

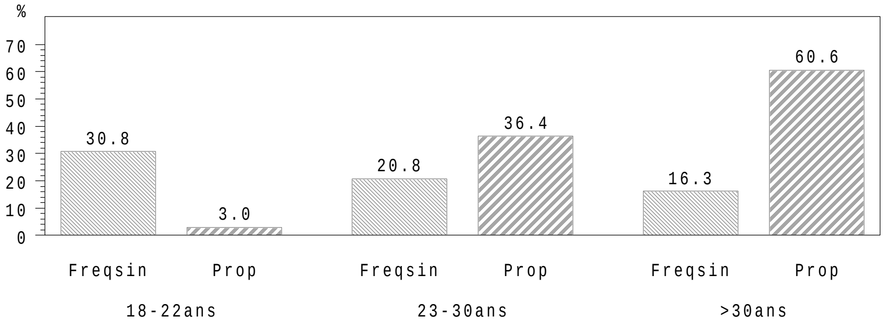

Figure 2: Répartition et fréquence de sinistres selon l'âge du conducteur

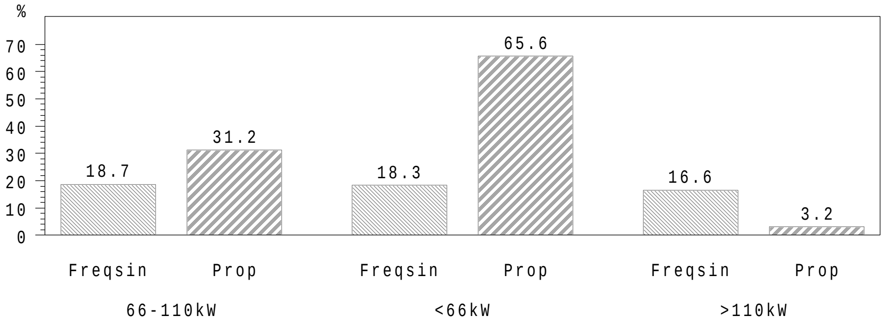

Figure 3: Répartition et fréquence de sinistres selon la puissance du véhicule

Ces histogrammes appellent les quelques commentaires suivants. On constate une légère sous-sinistralité pour les femmes (17.7% contre 18.8%), qui ne représentent que 36% des assurés du portefeuille. La sur-sinistralité des jeunes conducteurs est évidente (mais ils sont sous-représentés dans le portefeuille). Les fréquences de sinistres semblent décroître avec l'âge, passant de 30.8% à 20.8% et enfin à 16.3%. En ce qui concerne la puissance du véhicule, on constate une sous-sinistralité pour les grosses cylindrées. La fréquence des sinistres est plus élevée dans les grandes agglomérations, et la sinistralité semble décroître avec la taille de l'agglomération. Enfin, la

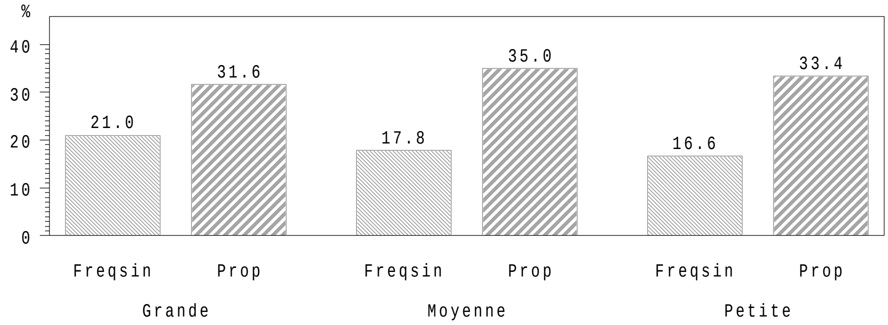

Figure 4: Répartition et fréquence de sinistres selon la taille de la ville

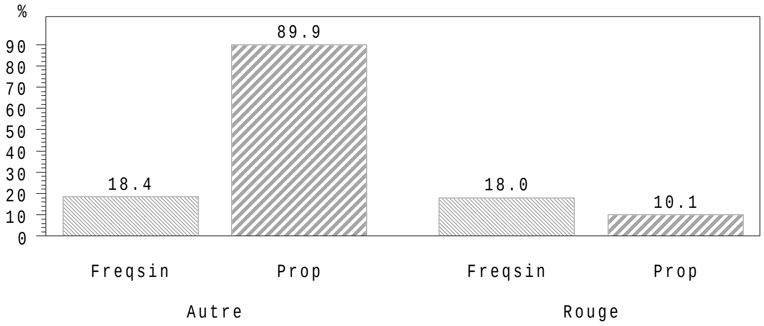

Figure 5: Répartition et fréquence de sinistres selon la couleur du véhicule

couleur rouge ne semble pas être un facteur aggravant.

3.2 Régression de Poisson en supposant l'indépendance temporelle

La procédure GENMOD de SAS permet de réaliser la régression de Poisson du nombre total de sinistres sur les 5 variables explicatives présentées plus haut. Les résultats sont présentés dans le Tableau 1.

Variable	Level	Coeff β	Std Error	Wald 95% Conf Limit		Chi-Sq	Pr>ChiSq
Intercept		-1.9242	0.0302	-1.9833	-1.8650	4063.54	$< .0001$
Sexe	Femme	-0.0581	0.0265	-0.1100	-0.0063	4.82	0.0281
Sexe	Homme	0	0	0	0	.	.
Age	$17-22$	0.6651	0.0583	0.5508	0.7793	130.23	$< .0001$
Age	$23-30$	0.2525	0.0261	0.2015	0.3036	93.87	$< .0001$
Age	> 30	0	0	0	0	.	.
Puissance	> 110kW	-0.0116	0.0750	-0.1586	0.1353	0.02	0.8769
Puissance	$66-110$kW	0.0563	0.0275	0.0024	0.1102	4.19	0.0406
Puissance	< 66kW	0	0	0	0	.	.
Ville	Grande	0.2549	0.0306	0.1949	0.3150	69.27	$< .0001$
Ville	Moyenne	0.0756	0.0311	0.0147	0.1364	5.92	0.0150
Ville	Petite	0	0	0	0	.	.
Couleur	Rouge	-0.0236	0.0416	-0.1052	0.0580	0.32	0.5710
Couleur	Autre	0	0	0	0	.	.

Table 1: Résultats de la régression de Poisson pour le modèle avec les 5 variables.

La colonne "Wald 95% Conf Limit" reprend les bornes inférieure et supérieure de confiance pour les paramètres au niveau 95%, calculées à l'aide de la formule

$$\text{Coeff}\,\beta_j \pm 1.96\,\text{Std Error}\,\beta_j,$$

où 1.96 est le quantile d'ordre 97.5% de la loi normale centrée réduite et Std Error est la racine du $j$ème élément diagonal de $\widehat{\boldsymbol{\Sigma}}$.

Les colonnes "Chi-Sq" et "Pr>ChiSq", qui est la p-valeur associée, permettent de tester si le coefficient β_j correspondant est significativement différent de 0. Ce test est effectué grâce à la statistique de Wald

$$\frac{(\text{Coeff}\,\beta_j)^2}{(\text{Std Error}\,\beta_j)^2},$$

qui obéit approximativement à la loi Chi-carrée à 1 degré de liberté. On rejettera la nullité de β_j lorsque la p-valeur est inférieure à 5%. L'examen de la dernière colonne du Tableau 1 indique clairement que certaines variables explicatives pourraient être omises sans nuire à la qualité du système.

La valeur de $L(\widehat{\boldsymbol{\beta}})$ pour le modèle reprenant les 5 variables explicatives est -19 282.6. L'analyse de Type 3 fournit les résultats présentés au Tableau 2. L'analyse de Type 3 permet d'examiner la contribution de chacune des variables par rapport à un modèle ne la contenant pas. Dans la colonne "ChiSquare" est calculée, pour chaque variable, 2 fois la différence entre la log-vraisemblance obtenue par le modèle contenant toutes les variables et la log-vraisemblance du modèle sans la variable en question. Cette statistique est asymptotiquement distribuée comme une Chi-carrée avec DF degrés de liberté, où DF est le nombre de paramètres associés à la variable explicative examinée. La dernière colonne nous fournit la p-valeur associée au test du rapport de vraisemblance; cela permet d'apprécier la contribution de cette variable explicative à la modélisation du phénomène étudié.

Source	DF	ChiSquare	Pr > ChiSq
Sexe	1	4.85	0.0276
Age	2	173.56	< .0001
Puissance	2	4.38	0.1120
Ville	2	74.10	< .0001
Couleur	1	0.32	0.5698

Table 2: Résultats de l'analyse de Type 3 pour le modèle avec les 5 variables.

L'examen des résultats des Tableaux 1 et 2 nous permet de diminuer le nombre de variables explicatives. Nous constatons en effet que la variable "couleur du véhicule" n'est pas significative. Son omission n'affecte pas le modèle, comme en témoigne la p-valeur de 56.98% du Tableau 2. Nous l'éliminons donc du modèle. Nous résumons ci-après les conclusions obtenues en poursuivant l'analyse statistique (sans fournir les résultats numériques). Dans une deuxième étape nous regroupons les niveaux de puissance "$66-110$kW" et ">110kW" en une seule classe étant donné que le niveau ">110kW" n'est pas significatif. A chaque fois, ces modifications n'affectent pas la qualité du modèle. Nous en arrivons alors au modèle retenu, lequel est décrit au Tableau 3.

La log-vraisemblance vaut -19 283.2 et l'analyse de Type 3 fournit les résultats présentés au Tableau 4. Toutes les variables retenues sont statistiquement significatives et l'omission d'une quelconque d'entre elles détériore significativement le modèle. Par ailleurs, la log-vraisemblance du modèle final est à peine moins bonne que celle du modèle non contraint (à savoir, -19 282.6).

Variable	Level	Coeff β	Std Error	Wald 95% Conf Limit		Chi-Sq	Pr>ChiSq
Intercept		-1.9277	0.0299	-1.9862	-1.8692	4165.69	$< .0001$
Sexe	Femme	-0.0575	0.0265	-0.1093	-0.0056	4.72	0.0299
Sexe	Homme	0	0	0	0	.	.
Age	$17-22$	0.6668	0.0582	0.5526	0.7809	131.02	$< .0001$
Age	$23-30$	0.2547	0.0260	0.2038	0.3056	96.09	$< .0001$
Age	> 30	0	0	0	0	.	.
Puissance	$\geq$ 66kW	0.0508	0.0269	-0.0019	0.1034	3.57	0.0587
Puissance	$<$ 66kW	0	0	0	0	.	.
Ville	Grande	0.2545	0.0306	0.1944	0.3145	69.03	$< .0001$
Ville	Moyenne	0.0757	0.0311	0.0148	0.1365	5.93	0.0148
Ville	Petite	0	0	0	0	.	.

Table 3: Résultats de la régression de Poisson pour le modèle final.

Source	DF	ChiSquare	Pr > ChiSq
Sexe	1	4.74	0.0294
Age	2	176.07	$< .0001$
Puissance	1	3.56	0.0593
Ville	2	73.82	$< .0001$

Table 4: Résultats de l'analyse de Type 3 pour le modèle final.

3.3 Qualité de l'ajustement

La Figure 6 représente les résidus ordonnés par ordre croissant $r^*_{(j)}$ en fonction de $rnorm_j$ et permet de vérifier la normalité des résidus. On est loin d'observer une ligne droite, comme ce devrait être le cas si le modèle était approprié. Ceci nous amène donc à devoir rejeter le modèle, pour les raisons expliquées plus loin.

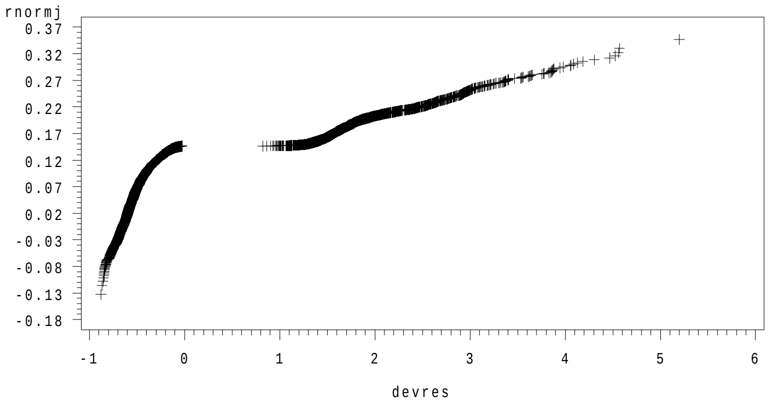

Figure 6: Vérification de la normalité des résidus

3.4 Surdispersion

Le modèle de Poisson impose des contraintes assez fortes sur la dépendance entre la variable de comptage N_{it} et les facteurs de risque $\boldsymbol{x}_{it}$, puisque

$$\mathbb{E}[N_{it}|\boldsymbol{x}_{it}] = \mathbb{V}\mathrm{ar}[N_{it}|\boldsymbol{x}_{it}] = \lambda_{it}. \tag{13}$$

Ceci revient donc à supposer l'égalité entre le nombre moyen de sinistre et la variabilité de ce nombre au sein de chaque classe de risque. L'équidispersion (13) est rarement satisfaite en pratique, ce qui met en doute le modèle de Poisson

En pratique, afin de vérifier la validité de (13), on calcule pour chaque classe de risque la moyenne et la variance empirique des nombres des sinistres, $\widehat{m}_k$ et $\widehat{\sigma}_k^2$, disons, et on porte le nuage de points $\{(\widehat{m}_k, \widehat{\sigma}_k^2),\ \ k = 1, 2, \dots\}$ en graphique. Ceci permet de voir comment la variance évolue en fonction de la moyenne. Lorsque les points sont autour de la première bissectrice du quadrillage, on peut considérer que les deux premiers moments conditionnels sont égaux, ce qui conforte le modèle de Poisson. Dans le cas contraire, on observe souvent un phénomène de surdispersion, c'est-à-dire des classes pour lesquelles $\widehat{\sigma}_k^2 > \widehat{m}_k$. Ce phénomène est dû la plupart du temps à des variables omises.

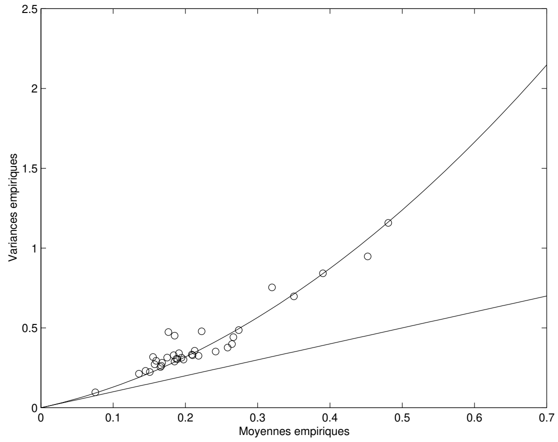

Figure 7: Vérification de la validité de (13) sur les données

On peut donner une interprétation simple de la surdispersion. Pour ce faire, considérons deux classes de risque C_1 et C_2 sans effet de surdispersion ($\widehat{\sigma}_1^2 = \widehat{m}_1$ et $\widehat{\sigma}_2^2 = \widehat{m}_2$), mais que l'on aurait omis de séparer. Dans la classe $C_1 \cup C_2$, la moyenne vaut $\widehat{m} = p_1\widehat{m}_1 + p_2\widehat{m}_2$ où p_1 et p_2 désignent les poids relatifs de C_1 et C_2, respectivement. La variance quant à elle passe à

$$\widehat{\sigma}^2 = p_1\widehat{\sigma}_1^2 + p_2\widehat{\sigma}_2^2 + p_1(\widehat{m}_1 - \widehat{m})^2 + p_2(\widehat{m}_2 - \widehat{m})^2.$$

On constate donc une surdispersion dans $C_1 \cup C_2$ puisque $\widehat{\sigma}^2 > \widehat{m}$, l'égalité n'étant possible que si $\widehat{m}_1 = \widehat{m}_2$. On comprend donc aisément que l'oubli de variables explicatives importantes puisse conduire à une surdispersion des observations au sein des classes de risque.

Sur la Figure 7, on constate une forte surdispersion et ce pour toutes les catégories d'assurés. Ceci nous conduit également à considérer que le modèle de Poisson avec indépendance temporelle n'est pas adapté. Afin d'éprouver graphiquement l'hypothèse d'un mélange de Poisson, nous avons ajusté le nuage de points de la Figure 7 à l'aide d'une courbe du second degré (du type $y = x + \gamma x^2$; voyez (7)); ceci fournit la courbe

$$y = x + 2.9545x^2.$$

La qualité de l'ajustement est reflétée par la valeur élevée du coefficient de détermination, $R^2 = 0.7753$. Ceci conforte l'hypothèse du mélange de Poisson pour les N_{it}.

3.5 Mise en évidence de la dépendance temporelle

Afin de mettre cette dépendance en évidence, nous travaillons avec les observations des deux dernières années à notre disposition. Nous considérons donc les observations N_{it}, $t = 2, 3$,

$i = 1, \ldots, n$, et nous effectuons une régression de celles-ci sur les variables explicatives $\boldsymbol{x}_{it}$ correspondantes auxquelles nous ajoutons la variable $N_{i,t-1}$, i.e. le nombre de sinistres observés au cours de la période précédente. Nous partons d'un modèle contenant les 5 variables explicatives déjà présentées et nous l'affinons, comme précédemment, par étapes successives, grâce à l'analyse de Type 3. Nous commençons par éliminer la variable "couleur du véhicule" qui a une p-valeur de de 27.37% et dans une deuxième étape nous éliminons la variable "sexe du conducteur" dont la p-valeur est devenue 21.10%. Nous obtenons alors le modèle dont les résultats sont présentés dans les Tableaux 5 et 6. Le coefficient de régression obtenu pour le nombre passé de sinistres est hautement significatif, ce qui indique une dépendance sérielle.

Variable	Level	Coeff β	Std Error	Wald 95% Conf Limit		Chi-Sq	Pr>ChiSq
Intercept		-2.0405	0.0370	-2.1131	-1.9680	3041.80	$< .0001$
Age	$17-22$	0.5841	0.0983	0.3914	0.7767	35.31	$< .0001$
Age	$23-30$	0.1822	0.0348	0.1140	0.2503	27.41	$< .0001$
Age	> 30	0	0	0	0	.	.
Puissance	> 110kW	-0.0745	0.1035	-2.2773	0.1283	0.52	0.4716
Puissance	$66-110$kW	0.0933	0.0357	0.0233	0.1633	6.83	0.0090
Puissance	< 66kW	0	0	0	0	.	.
Ville	Grande	0.2201	0.0412	0.1394	0.3009	28.54	$< .0001$
Ville	Moyenne	0.1050	0.0413	0.0242	0.1859	6.48	0.0109
Ville	Petite	0	0	0	0	.	.
N_{t-1}		0.3113	0.0371	0.2387	0.3839	70.59	$< .0001$

Table 5: Résultats de la régression pour le modèle tenant compte de la sinistralité passée.

Source	DF	ChiSquare	Pr > ChiSq
Age	2	50.58	$< .0001$
Puissance	2	7.94	0.0188
Ville	2	28.68	$< .0001$
N_{t-1}	1	63.38	$< .0001$

Table 6: Résultats de l'analyse de Type 3 pour le modèle tenant compte de la sinistralité passée.

Dans une deuxième approche nous repartons de la prime fréquence obtenue sous l'hypothèse d'indépendance et sans ajout du nombre de sinistres de l'année précédente comme variable explicative. Cette prime est alors corrigée par un facteur multiplicatif, obtenu par une régression de Poisson sur la seule variable "nombre de sinistres de l'année précédente", considérée comme continue (en mettant la prime fréquence obtenue sous l'hypothèse d'indépendance en offset). Les résultats de cette régression se trouvent dans les Tableaux 7 et 8.

Variable	Coeff β	Std Error	Wald 95% Conf Limit		Chi-Sq	Pr>ChiSq
Intercept	-0.1147	0.0180	-0.1500	-0.0793	40.42	$< .0001$
N_{t-1}	0.3040	0.0370	0.2316	0.3765	67.65	$< .0001$

Table 7: Résultats de la régression pour le modèle tenant compte de la sinistralité passée.

Source	DF	ChiSquare	Pr > ChiSq
N_{t-1}	1	60.84	$< .0001$

Table 8: Résultats de l'analyse de Type 3 pour le modèle tenant compte de la sinistralité passée.

Il est intéressant de noter au passage que cette manière de procéder fournit immédiatement des coefficients bonus-malus "à la française". En effet, le Tableau 7 nous apprend que les assurés qui n'ont déclaré aucun sinistre sur l'année verront leur prime multipliée par

$$\exp(-0.1147) = 0.8916$$

alors que ceux ayant déclaré k sinistres subiront une majoration de prime valant

$$\exp(-0.1147 + k \times 0.3040) = 0.8916 \times (1.3553)^k.$$

Il est toujours intéressant de comparer ces coefficients à ceux produits par un modèle plus orthodoxe formulé en termes de variables latentes.

3.6 Prise en compte de la dépendance temporelle

La dépendance sérielle des N_{it} à i fixé ayant clairement été mise en évidence, il importe de mesurer l'impact de l'hypothèse d'indépendance souvent formulée par les praticiens sur l'estimation des primes-fréquence. Pour ce faire, nous allons recourir aux GEE.

L'approche GEE peut être réalisée par la procédure GENMOD de SAS. Une sélection des variables, basée comme précédemment sur l'analyse de Type 3, nous conduit à retenir les mêmes variables que pour le modèle où l'on supposait l'indépendance. Les résultats se trouvent dans les Tableaux 9 et 10. La "working correlation matrix" de structure non spécifiée à l'avance, est décrite au Tableau 11.

Variable	Level	Coeff β	Std Error	95% Conf Limit		Z	Pr> $\vert Z\vert$
Intercept		-1.9234	0.0319	-1.9859	-1.8609	-60.32	$< .0001$
Sexe	Femme	-0.0580	0.0289	-0.1147	-0.0013	-2.01	0.0449
Sexe	Homme	0	0	0	0	.	.
Age	$17-22$	0.6589	0.0617	0.5379	0.7799	10.67	$< .0001$
Age	$23-30$	0.2556	0.0281	0.2005	0.3107	9.10	$< .0001$
Age	> 30	0	0	0	0	.	.
Puissance	$\geq$ 66kW	0.0532	0.0292	-0.0040	0.1105	1.82	0.0684
Puissance	$<$ 66kW	0	0	0	0	.	.
Ville	Grande	0.2542	0.0332	0.1892	0.3192	7.67	$< .0001$
Ville	Moyenne	0.0720	0.0336	0.0060	0.1379	2.14	0.0324
Ville	Petite	0	0	0	0	.	.

Table 9: Résultats de la régression de Poisson avec approche GEE.

Source	DF	ChiSquare	Pr > ChiSq
Sexe	1	4.08	0.0435
Age	2	128.78	$< .0001$
Puissance	1	3.28	0.0700
Ville	2	60.70	$< .0001$

Table 10: Résultats de l'analyse de Type 3 pour le modèle avec approche GEE.

$$\begin{pmatrix} 1 & 0.0512 & 0.0462 \\ 0.0512 & 1 & 0.0464 \\ 0.0462 & 0.0464 & 1 \end{pmatrix}$$

Table 11: Working correlation matrix de structure non spécifiée.

Si on compare les $\widehat{\beta}_j$ des Tableaux 3 (sous l'hypothèse d'indépendance) et 9 (reconnaissant la dépendance sérielle), on constate des différences modestes. Les erreurs-standards sont systématiquement plus élevées dans l'approche GEE (la dépendance sérielle augmentant la surdispersion).

Pour terminer, comparons les montants des primes-fréquences obtenues en supposant l'indépendance sérielle ou en reconnaissant explicitement la dépendance temporelle; ceux-ci sont fournis aux Tableaux 12 et 13. On constate des différences au niveau des estimations des fréquences annuelles de sinistre associées aux classes de risque, mais ces différences restent limitées (elles seront néanmoins exacerbées par la multiplication par le coût moyen d'un sinistre et par les chargements de sécurité et commerciaux). La prise en considération de la dépendance sérielle a également un impact sur les intervalles de confiance pour les primes-fréquence, lesquels sont plus larges dans l'approche GEE. Ceci indique qu'en présence de dépendance sérielle dans les données utilisées pour ajuster les primes-fréquences, le chargement de sécurité à inclure dans la prime nette sera plus élevé que celui calculé sur base de l'hypothèse d'indépendance temporelle.

4 En guise de conclusion...

L'enseignement principal qui nous semble devoir être tiré de la présente étude est que l'impact de l'hypothèse d'indépendance temporelle des données individuelles relatives à différentes périodes d'observation, communément utilisée au sein des compagnies d'assurance lorsqu'il s'agit de construire une tarification *a priori* en s'appuyant sur plusieurs années d'observation, ne semble pas avoir d'impact conséquent sur l'estimation des primes-fréquence. Théoriquement, cela s'explique par le caractère convergent des estimateurs obtenus sous cette hypothèse d'indépendance, et par la taille souvent considérable des portefeuilles d'assurance automobile. Cependant, la reconnaissance de l'aspect sériel des données augmente la variance des estimateurs, et, partant, la largeur des intervalles de confiance sur les primes-fréquence. Ceci doit être pris en considération lors de la fixation de la hauteur du chargement de sécurité.

Classes de risque				Primes-fréquence		
Sexe	Age	Puissance	Ville	Inf	Prime	Sup
Homme	17-22	< 66	Petite	0.25251	0.28339	0.31084
			Moyenne	0.27221	0.30566	0.34322
			Grande	0.32535	0.3655	0.41061
		≥ 66	Petite	0.26395	0.29815	0.33678
			Moyenne	0.28464	0.32158	0.36332
			Grande	0.34027	0.38454	0.43457
	23-30	< 66	Petite	0.17708	0.18768	0.19892
			Moyenne	0.19135	0.20243	0.21416
			Grande	0.22878	0.24206	0.25612
		≥ 66	Petite	0.1848	0.19746	0.21099
			Moyenne	0.19976	0.21298	0.22707
			Grande	0.23893	0.25467	0.27146
	> 30	< 66	Petite	0.13721	0.14548	0.15425
			Moyenne	0.1483	0.15692	0.16604
			Grande	0.17754	0.18764	0.19831
		≥ 66	Petite	0.14413	0.15306	0.16254
			Moyenne	0.15588	0.16509	0.17485
			Grande	0.18668	0.19741	0.20876
Femme	17-22	< 66	Petite	0.23779	0.26756	0.30106
			Moyenne	0.25631	0.28859	0.32494
			Grande	0.30646	0.34509	0.38859
		≥ 66	Petite	0.24766	0.2815	0.31996
			Moyenne	0.26704	0.30362	0.34523
			Grande	0.31935	0.36307	0.41277
	23-30	< 66	Petite	0.16643	0.1772	0.18867
			Moyenne	0.17975	0.19113	0.20322
			Grande	0.21509	0.22855	0.24285
		≥ 66	Petite	0.17265	0.18643	0.20131
			Moyenne	0.18652	0.20108	0.21679
			Grande	0.22322	0.24045	0.25901
	> 30	< 66	Petite	0.12906	0.13736	0.14619
			Moyenne	0.13942	0.14816	0.15743
			Grande	0.16703	0.17716	0.1879
		≥ 66	Petite	0.13462	0.14451	0.15513
			Moyenne	0.14549	0.15587	0.167
			Grande	0.17431	0.18639	0.1993

Table 12: Estimations des primes-fréquences des différentes classes de risque sous l'hypothèse d'indépendance.

Classes de risque				Primes-fréquence		
Sexe	Age	Puissance	Ville	Inf	Prime	Sup
Homme	17-22	< 66	Petite	0.24957	0.28236	0.31946
			Moyenne	0.26809	0.30344	0.34345
			Grande	0.32141	0.36409	0.41244
		≥ 66	Petite	0.26139	0.2978	0.33929
			Moyenne	0.2811	0.32003	0.36436
			Grande	0.33676	0.384	0.43787
	23-30	< 66	Petite	0.17722	0.18866	0.20083
			Moyenne	0.19057	0.20274	0.21569
			Grande	0.22861	0.24326	0.25885
		≥ 66	Petite	0.18529	0.19897	0.21367
			Moyenne	0.19963	0.21382	0.22902
			Grande	0.23915	0.25656	0.27525
	> 30	< 66	Petite	0.13725	0.14611	0.15553
			Moyenne	0.14758	0.15701	0.16704
			Grande	0.17746	0.18839	0.20001
		≥ 66	Petite	0.14454	0.15409	0.16428
			Moyenne	0.15578	0.1656	0.17603
			Grande	0.187	0.19869	0.21112
Femme	17-22	< 66	Petite	0.23537	0.26645	0.30162
			Moyenne	0.25301	0.28633	0.32405
			Grande	0.30365	0.34356	0.38872
		≥ 66	Petite	0.24524	0.28101	0.32201
			Moyenne	0.26389	0.30199	0.34559
			Grande	0.31647	0.36235	0.41488
	23-30	< 66	Petite	0.16642	0.17802	0.19043
			Moyenne	0.17918	0.19131	0.20426
			Grande	0.21539	0.22955	0.24463
		≥ 66	Petite	0.17528	0.18775	0.20426
			Moyenne	0.18606	0.20177	0.2188
			Grande	0.22331	0.2421	0.26247
	> 30	< 66	Petite	0.12866	0.13787	0.14773
			Moyenne	0.13851	0.14816	0.15848
			Grande	0.16687	0.17777	0.18939
		≥ 66	Petite	0.13426	0.145451	0.15747
			Moyenne	0.14478	0.15626	0.16865
			Grande	0.17409	0.18749	0.20193

Table 13: Estimations des primes-fréquences des différentes classes de risque dans l'approche GEE.

Bibliographie

1. Besson, J.-L., & Partrat, Ch. (1992). Trend et systèmes de Bonus-Malus. *ASTIN Bulletin* **22**, 11-31.

2. Boskov, M., & Verrall, R.J. (1994). Premium rating by geographical area using spatial models. *ASTIN Bulletin* **24**, 131-143.

3. Brouhns, N., Cullus, P., & Denuit, M. (2001). Intégration d'éléments de statistiques officielles dans une tarification IARD. En préparation

4. Brouhns, N., & Denuit, M. (2001a). Construction de tables de mortalité par régression poissonienne. Discussion Paper, Institut de Statistique, Université catholique de Louvain, Belgique.

5. Brouhns, N., & Denuit, M. (2001b). Tarification automobile et modèles généralisés additifs. Discussion Paper, Institut de Statistique, Université catholique de Louvain, Belgique.

6. Brouhns, N., Denuit, M., El Ghelbzouri, A., & Verrall, R.J. (2001). Hierarchical Bayes GLM's for the analysis of spatial variation in claim frequencies and severities. Discussion Paper, Institut de Statistique, Université catholique de Louvain, Belgique.

7. Denuit, M., Pitrebois, S., & Walhin, J.-F. (2001). Méthodes de construction de systèmes bonus-malus en RC auto. *ACTU-L* **1**, 7-38

8. Ingenbleeck, J.-F., & Lemaire, J. (1988). What is a sports car? *ASTIN Bulletin* **18**, 175-188

9. Kaas, R., Goovaerts, M.J., Dhaene, J., & Denuit, M. (2001). *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, forthcoming.

10. Le Bailly de Tilleghem, C., & Denuit, M. (2001). Logistic or Poisson regression: what impact on ratemaking? Discussion Paper, Institut de Statistique, Université catholique de Louvain, Belgique.

11. Liang, K.Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

12. Pinquet, J., Guillen, M., & Bolancé, C. (2001). Allowance for the age of claims in bonus-malus systems. *ASTIN Bulletin*, to appear.

13. Purcaru, O., & Denuit, M. (2001a). On the dependence induced by frequency credibility models. I. Time-invariant random effects. Discussion Paper, Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium.

14. Purcaru, O., & Denuit, M. (2001b). On the dependence induced by frequency credibility models. II. Time-dependent random effects. Discussion Paper, Institut de Statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium.

15. Sithole, T.Z., Haberman, S., & Verrall, R.J. (2000). An investigation into parametric models for mortality projections, with applications to immediate annuitants and life office pensioners' data. *Insurance: Mathematics & Economics* **27**, 285-312.

16. Verrall, R. (1996). A unified framework for graduation. Actuarial Research Paper 91, Department of Actuarial Science and Statistics, City University, London.

17. Zeger, S.L., Liang, K.Y., & Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049-1060.