# Tarification automobile sur données de panel

S. PITREBOIS, M. DENUIT et J.-F. WALHIN, Bruxelles, Louvain-la-Neuve

# Tarification automobile sur données de panel

## 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. 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 de contrat. Les différentes étapes de la segmentation peuvent 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 éventuelle

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. Compte tenu de 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 le 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 personnalisation *a posteriori* au moyen d'un système de type bonus-malus ou d'une autre forme "d'experience-rating".

### *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. Nous présentons une méthode simple et performante de segmentation *a priori* et de calcul des fréquences annuelles de sinistre par l'assureur. 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 (2001). Il rejoint à certains endroits PINQUET, GUILLEN & BOLANCÉ (2002) et BOLANCÉ, GUILLÉN & PINQUET (2003), bien que nous nous intéressions 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* (et induisent la dépendance sérielle à l'aide de variables latentes corrélées).

### *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.

### *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, mais 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 sont sans doute corrélées (ce qui est la raison d'être de la tarification *a posteriori*). Nous sommes donc en présence de données en panel.

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 à l'aide des techniques proposées par LIANG & ZEGER (1986) et ZEGER ET AL. (1988).

Les estimateurs des fréquences de sinistres 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 croît). Dès lors, on peut raisonnablement espérer que pour des portefeuilles automobiles de grande taille, l'impact de l'hypothèse simplificatrice d'indépendance sur les estimations ponctuelles soit minime. C'est en effet ce que nous mettrons en évidence dans la partie empirique de notre étude.

### *1.5 Notations*

Comme nous l'avons expliqué plus haut, les compagnies d'assurance utilisent souvent plusieurs périodes d'observation pour construire leur tarif. 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 $i$. En plus des variables explicatives, on peut introduire le temps calendaire en composante de régression afin de prendre en compte certains événements ponctuels ou d'éventuelles tendances dans la sinistralité, dans l'esprit de BESSON & PARTRAT (1992).

Typiquement, nous sommes en présence de données de panel : une même variable est mesurée 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).

### *1.6 Score et codage des variables explicatives*

Le niveau de risque de chaque assuré est reflété dans un score. Dorénavant, nous notons $\eta_{it} = \boldsymbol{\beta}^t \boldsymbol{x}_{it}$ le prédicteur linéaire, à savoir une combinaison linéaire $\beta_0 + \sum_{j=1}^{p} \beta_j x_{itj}$ des variables explicatives $\boldsymbol{x}_{it} = (1, x_{it1}, \ldots, x_{itp})^t$ relatives à l'individu $i$ et à la période $t$. Le prédicteur linéaire $\eta_{it}$ est encore appelé score car il permet de ranger les assurés du moins risqué au plus risqué, en suivant les valeurs croissantes de $\eta_{it}$.

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).

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 (2003). Une variable catégorielle à $k$ facteurs est généralement codée par $k - 1$ variables binaires qui sont toutes nulles pour le niveau de référence. Expliquons la technique de codage à l'aide de l'exemple élémentaire suivant. Considérons 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

$$X_1 = \begin{cases} 0 \text{ si l'assuré est un homme} \\ 1 \text{ si l'assuré est une femme} \end{cases}$$

$$X_2 = \begin{cases} 0 \text{ si le véhicule n'a pas de caractère sportif} \\ 1 \text{ si le véhicule a un caractère sportif} \end{cases}$$

$$X_3 = \begin{cases} 1 \text{ si l'assuré a moins de 30 ans} \\ 0 \text{ sinon} \end{cases}$$

$$X_4 = \begin{cases} 1 \text{ si l'assuré a plus de 65 ans} \\ 0 \text{ sinon.} \end{cases}$$

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 $\beta_0$ 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).

### *1.7 Présentation du jeu de données*

Dans cet article, nous illustrons nos propos sur un portefeuille d'assurance belge comprenant 20 354 polices, observées durant une période de 3 ans. La Figure 1 donne une idée de la durée d'exposition au risque des polices en portefeuille. Un peu plus de 34% des assurés sont restés en portefeuille durant les trois ans. Pour chaque police et pour chaque année sont renseignés le nombre de sinistres et certaines caractéristiques de l'assuré : 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). Sur l'ensemble du portefeuille la fréquence annuelle moyenne est de 18.4% (ce qui est largement supérieur à la moyenne européenne).
Les Figures 2 à 6 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 %) de sinistres.

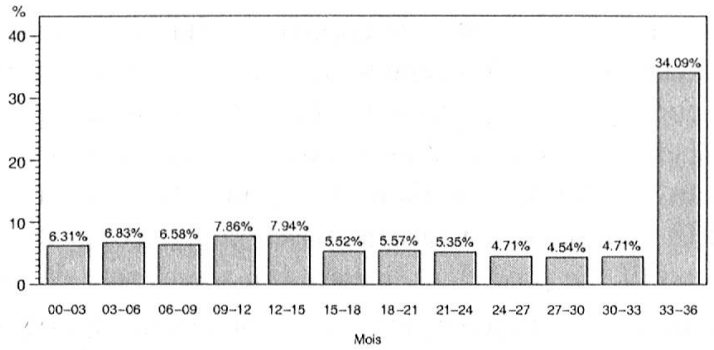


Figure 1: Durée d'exposition au risque (en mois)

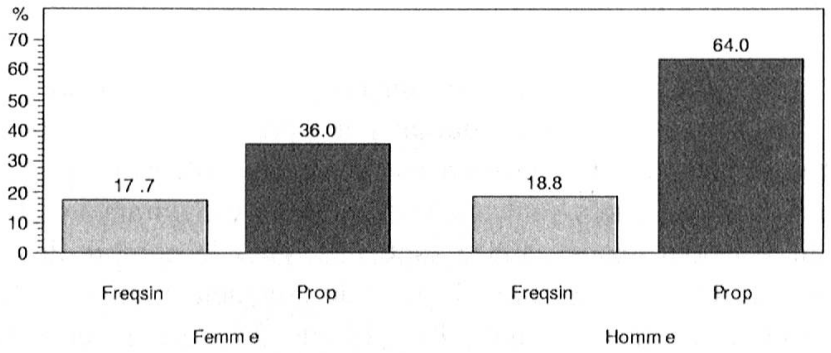


Figure 2: Répartition et fréquence de sinistres selon le sexe du conducteur

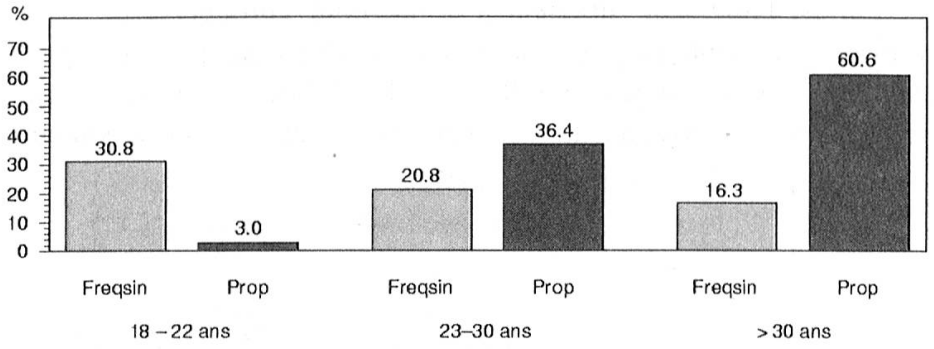


Figure 3: Répartition et fréquence de sinistres selon l'âge du conducteur

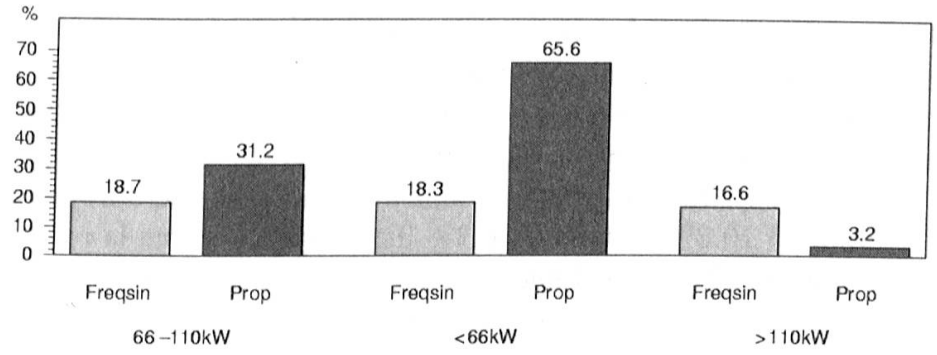


Figure 4: Répartition et fréquence de sinistres selon la puissance du véhicule

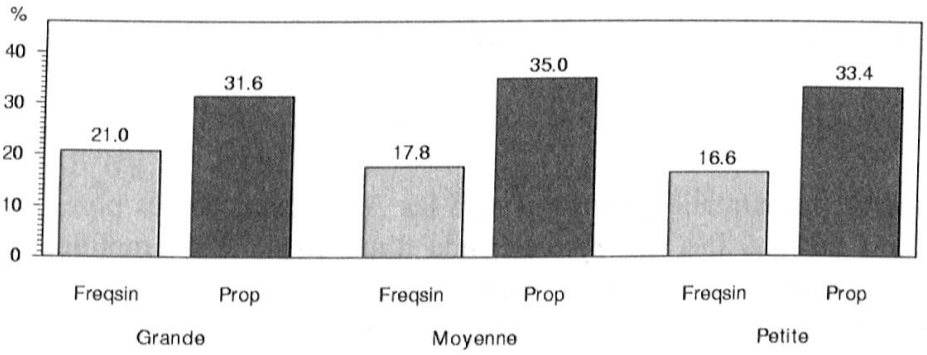


Figure 5: Répartition et fréquence de sinistres selon la taille de la ville

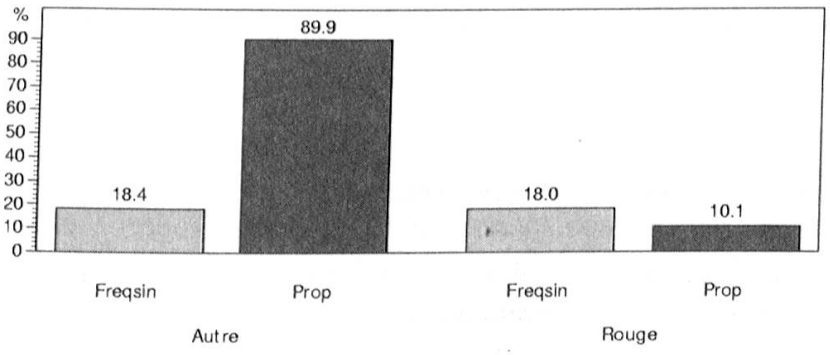


Figure 6: Répartition et fréquence de sinistres selon la couleur du véhicule

Ces histogrammes appellent les quelques commentaires suivants. On constate à la Figure 2 une légère sous-sinistralité pour les femmes (17.7% contre 18.8%), qui représentent 36% des assurés du portefeuille. La sur-sinistralité des jeunes conducteurs ressort clairement de la Figure 3 (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 à la Figure 4 une sous-sinistralité pour les grosses cylindrées. L'examen de la Figure 5 révèle que la fréquence des sinistres est plus élevée dans les grandes agglomérations. La sinistralité semble décroître avec la taille de l'agglomération. Enfin, on constate à la Figure 6 que la couleur rouge ne semble pas être un facteur aggravant.

## 2 Régression de Poisson en supposant l'indépendance temporelle

### *2.1 Modélisation*

En première approximation, on supposera les $N_{it}$ indépendantes pour différentes valeurs de $i$ et de $t$. Il s'agit bien entendu d'une hypothèse simplificatrice forte dont nous évaluerons l'impact en comparant les résultats obtenus à ceux fournis par différentes méthodes permettant de tenir compte de cette dépendance sérielle. Nous supposons que la loi conditionnelle de $N_{it}$ sachant $\boldsymbol{x}_{it}$ est de Poisson et nous spécifions une moyenne de forme exponentielle linéaire, i.e.

$$N_{it} =_d \text{Poisson}\left(d_{it}\exp(\eta_{it})\right), \quad i = 1,2,\ldots,n, \quad t = 1,2,\ldots,T_i\,. \tag{1}$$

La fréquence de sinistre relative à l'individu $i$ durant la période $t$ est $\lambda_{it} = d_{it}\exp(\eta_{it})$.

### *2.2 Estimation par maximum de vraisemblance*

Notons $n_{it}$ le nombre de sinistres déclarés par l'assuré $i$ durant la période $t$. 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 $\boldsymbol{\beta}$, 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 $\mathcal{L}(\boldsymbol{\beta})$ : $\widehat{\boldsymbol{\beta}}$ est donc la valeur du paramètre rendant les observations recueillies par l'actuaire les plus probables. 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 $j = 1, 2, \ldots, p$,

$$\frac{\partial}{\partial \beta_j} L(\boldsymbol{\beta}) = 0 \Leftrightarrow \sum_{i=1}^{n} \sum_{t=1}^{T_i} x_{itj} \Big\{ n_{it} - \lambda_{it} \Big\} = 0 . \tag{3}$$

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 $nres_{it}$.

### *2.3 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

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

En supposant les coûts des sinistres constamment égaux à 1, ceci garantit donc que les primes supportées par les hommes compensent exactement les sinistres causés par ceux-ci. Il n'y a donc pas de transfert de primes entre hommes et femmes induit par le tarif appliqué par la compagnie.
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 $\beta_0$ soit inclus dans le score $\eta_{it}$ (c'est-à-dire pour autant que les fréquences soient exprimées par rapport à une fréquence annuelle de référence $\exp(\beta_0)$). Le modèle reconstitue donc sans erreur le nombre total de sinistres observés.

### 2.4 *Variance asymptotique des estimateurs*

La matrice variance-covariance $\boldsymbol{\Sigma}$ de l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$ est l'inverse de la matrice d'information de Fisher $\mathcal{I}$. Elle peut être estimée par

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

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.

### 2.5 *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 (compor-

tant uniquement un intercept $\beta_0$, 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 modèle. 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 homologue 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 exclura 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.

## 2.6 *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 la statistique de déviance 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 $\beta_0$ garantit que (2) est valable, la déviance s'écrit dans ce cas

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

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. Dans le cadre de la régression de Poisson, les résidus se définissent à partir de la contribution de chaque observation à la statistique de déviance $D$. Plus précisément, le résidu associé à l'observation $i$ durant l'année $t$ est donné

par

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

avec la même convention que ci-dessus, à savoir $y \ln y = 0$ lorsque $y = 0$.
Dans la plupart des applications actuarielles, l'analyse des résidus individuels $r_{it}^D$ n'apprend pas grand chose quant à la qualité de l'ajustement. En effet, ces résidus présentent une structure forte induite par le petit nombre de valeurs observées pour les $N_{it}$ (rarement plus de 4). Si on veut juger de la qualité du modèle, il vaut mieux grouper les assurés en classes et calculer les résidus au niveau des classes (en remplaçant $N_{it}$ par le nombre de sinistres observés pour la classe et $\widehat{\lambda}_{it}$ par l'anticipation au niveau de la classe dans la formule du résidu de déviance donnée plus haut). Contrairement au modèle linéaire classique, la loi des résidus est difficile à appréhender dans la régression de Poisson. On se contente donc souvent en pratique de vérifier que les résidus ne laissent plus apparaître de structure.

## *2.7 Prédiction des fréquences annuelles de sinistres*

Pour l'assuré $i$ et la période $t$, caractérisés par un vecteur de variables explicatives $\boldsymbol{x}_{it}$, la prime fréquence annuelle prédite est $\exp(\boldsymbol{x}_{it}^t \widehat{\boldsymbol{\beta}})$. 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

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

Comme l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$ est approximativement gaussien en grand échantillon, $\widehat{\eta}_{it}$ l'est également et un intervalle de confiance approximatif au niveau de confiance $1 - \alpha$ pour la prime-fréquence annuelle est alors fourni par

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

## 2.8 *Illustration numérique*

### 2.8.1 Ajustement du modèle

La procédure GENMOD de SAS permet de réaliser la régression de Poisson du nombre de sinistres sur les 5 variables explicatives présentées à la Section 1.7. Les résultats sont présentés dans le Tableau 1.

| Variable | Level | Coeff $\beta$ | 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 − 110kW | 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.

Les estimations ponctuelles des $\beta_j$ sont fournies dans la troisième colonne du Tableau 1, les deux premières permettant d'identifier le niveau auquel le coefficient de régression se rapporte. Les lignes où apparaissent des 0 correspondent aux niveaux de référence des différentes variables tarifaires. La colonne "Wald 95% Conf Limit" reprend les bornes inférieure et supérieure des intervalles 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 $\widehat{\Sigma}$.
Les colonnes "Chi-Sq" et "Pr>ChiSq", qui est la $p$-valeur associée, permettent de tester si le coefficient $\beta_j$ 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 $\beta_j$ 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 modèle.

La valeur de $L(\widehat{\beta})$ pour le modèle reprenant les 5 variables explicatives est -19282.6. 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 – 110kW" 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.

| Variable | Level | Coeff $\beta$ | 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 | ≥ 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.

La log-vraisemblance vaut -19 283.2 et l'analyse de Type 3 fournit les résultats présentés au Tableau 4. A l'exception de la variable puissance, toutes les variables sont statistiquement significatives et l'omission d'une quelconque d'entre elles détériore significativement le modèle (au seuil de 5%). Nous décidons cependant de garder la variable puissance en raison de son importance dans les tarifs pratiqués par les compagnies d'assurances et du faible dépassement du seuil (0.93%, seulement). La log-vraisemblance du modèle final est à peine moins bonne que celle du modèle non contraint (à savoir, -19 282.6).

| 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.

### 2.8.2 Qualité de l'ajustement

La Figure 7 décrit les résidus de déviance individuels. On peut y constater la structure reflétant les quelques valeurs observées pour les $N_{it}$. On ne peut donc juger de la qualité du modèle sur base de la Figure 7. Si on recalcule les résidus par classes, on obtient la Figure 8. On n'y constate aucune structure particulière, mais des valeurs assez élevées de certains résidus, qui mettent en question la justesse du modèle.

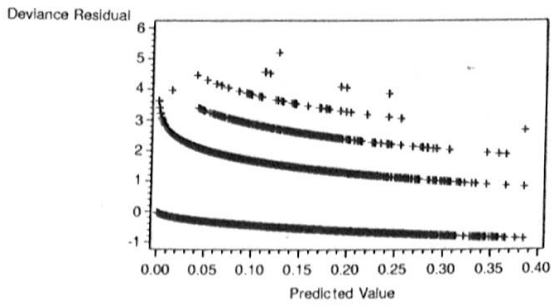


Figure 7: Graphe des résidus individuels en fonction des valeurs prédites

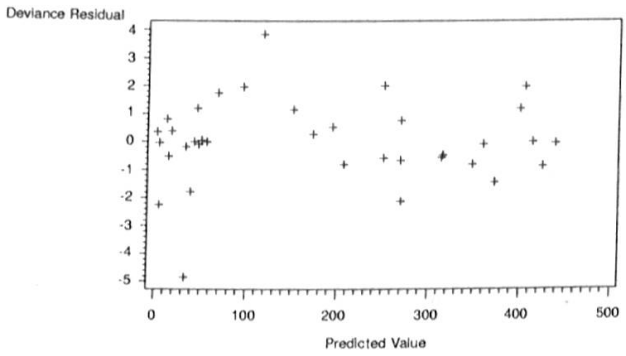


Figure 8: Graphe des résidus par classe en fonction des valeurs prédites

### 2.8.3 Surdispersion

Le modèle de Poisson impose des contraintes assez fortes sur les deux premiers moments de la variable de comptage $N_{it}$ compte tenu des facteurs de risque $\boldsymbol{x}_{it}$,

puisque

$$\mathbb{E}[N_{it}|\boldsymbol{x}_{it}] = \mathbb{V}\text{ar}[N_{it}|\boldsymbol{x}_{it}] = \lambda_{it}\,. \tag{4}$$

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 (4) est rarement satisfaite en pratique, ce qui met en doute le modèle de Poisson
En pratique, afin de vérifier la validité de (4), on calcule pour chaque classe de risque la moyenne et la variance empirique des nombres des sinistres, $\widehat{m}_k$ et $\widehat{\sigma}^2_k$, disons, et on porte le nuage de points $\{(\widehat{m}_k, \widehat{\sigma}^2_k),\ k = 1, 2, \ldots\}$ 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 valide (4) et conforte le modèle de Poisson. Dans le cas contraire, on observe souvent un phénomène de surdispersion, c'est-à-dire des classes pour lesquelles $\widehat{\sigma}^2_k > \widehat{m}_k$. Ce phénomène est dû la plupart du temps à des variables omises.

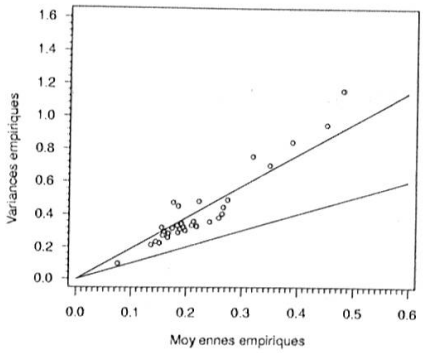


Figure 9: Vérification de la validité de (4) sur les données

On peut en effet donner une interprétation simple de la surdispersion. Pour ce faire, considérons deux classes de risque $C_1$ et $C_2$ sans effet de surdispersion ($\widehat{\sigma}^2_1 = \widehat{m}_1$ 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

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

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.

La Figure 9 montre les points $(\widehat{m}_k, \widehat{\sigma}_k^2)$ correspondant aux classes de risque du modèle final décrit au Tableau 3. On y constate une forte surdispersion et ce pour toutes les catégories d'assurés. Les points $(\widehat{m}_k, \widehat{\sigma}_k^2)$ sont en effet situés au-dessus de la première bissectrice du quadrillage. Ceci nous conduit également à considérer que le modèle de Poisson avec indépendance temporelle n'est pas adapté.

Il est possible de tenir compte de la surdispersion constatée dans les données, sans reconnaître l'éventuelle dépendance sérielle. A cette fin, on recourt soit à un modèle de Poisson mélange, soit à une approche de quasi-vraisemblance en spécifiant

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

Afin d'éprouver graphiquement la validité de cette dernière relation, nous avons ajusté le nuage de points de la Figure 9 à l'aide d'une droite passant par l'origine (donc d'équation $y = \phi x$). Ceci donne un paramètre de dispersion $\phi$ estimé à 1.9122 et un coefficient de détermination $R^2 = 86.17\%$ (ce qui signifie que la droite explique plus de 86% de la variabilité du nuage de points). A titre de comparaison, si nous avions tenté un ajustement à l'aide d'une courbe du second degré (du type $y = x + \gamma x^2$, caractéristique du lien moyenne-variance dans un modèle de Poisson mélange), on aurait obtenu $y = x + 2.9545x^2$ avec $R^2 = 90.90\%$. Un mélange de Poisson (la loi binomiale négative, par exemple) aurait donc pu également être considéré. Nous privilégions cependant dans cet article une approche de quasi-vraisemblance. Cela consiste à déterminer $\widehat{\boldsymbol{\beta}}$ en résolvant le système (2)-(3). Ensuite, $\widehat{\phi}$ est obtenu en divisant soit la déviance, soit la statistique de Pearson par le nombre de degrés de liberté. La valeur estimée de $\phi$ sur nos données est 1.35, ce qui traduit bien la surdispersion des données.

L'introduction du paramètre de surdispersion $\phi$ gonfle les variances et les covariances des $\widehat{\beta}_j$ (lesquelles sont multipliées par $\widehat{\phi}$). Ceci a pour effet de réduire la valeur des statistiques de test utilisées pour éprouver la nullité des $\beta_j$ ou la pertinence de l'inclusion de certaines variables dans le modèle. La prise en compte de la surdispersion peut donc mener à l'exclusion de variables tarifaires qui auraient été conservées dans le modèle de Poisson pur. On observe un phénomène de ce type sur notre jeu de données, la $p$-valeur de la variable puissance dans l'analyse de type 3 passant à 10.44%.

## 3 Prise en compte de la dépendance temporelle

### *3.1 Détection de l'aspect sériel*

Afin d'avoir une première idée du type de dépendance existant entre les $N_{it}$, on peut par exemple considérer les observations $N_{it}$, $t = 2, \ldots, T_i$, $i = 1, \ldots, n$, et 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 également de voir l'effet de l'inclusion de valeurs passées de la variable d'intérêt sur les variables explicatives. 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 $\beta$ | 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 − 110kW | 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 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" (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 $\beta$ | 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 en figeant l'influence des variables explicatives.

| 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 en figeant l'influence des variables explicatives.

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.

Les données suggèrent donc une dépendance sérielle. Cela invalide les résultats obtenus à la section précédente, lesquels se fondent notamment sur l'hypothèse que les $N_{it}$ sont indépendantes pour différentes valeurs de $i$ et de $t$. Théoriquement, on peut cependant montrer que l'estimateur du maximum de vraisemblance $\widehat{\beta}$ calculé sous l'hypothèse d'indépendance sérielle (donc avec erreur de spécification) est convergent. Si la taille du portefeuille est suffisamment grande, on s'attend donc à peu d'impact sur les estimations ponctuelles des différents $\beta_j$. Par contre, la variance de $\widehat{\beta}$ ne peut plus être calculée comme décrit plus haut, et est quant à elle affectée par la dépendance sérielle.

### *3.2 Estimation des paramètres à l'aide de la technique GEE*

En présence de dépendance sérielle, on pourrait songer à garder l'estimateur du maximum de vraisemblance dans le modèle de Poisson avec indépendance temporelle (donc solution de (2)-(3)), choix qui se justifie par le caractère convergent de celui-ci. Comme l'ont montré LIANG & ZEGER (1986), il est possible d'améliorer cette approche (i.e. d'obtenir des estimateurs dont la variance asymptotique sera plus faible que celle de ceux que nous venons de décrire). Il s'agit de la méthode des GEE (pour l'anglais "Generalized Estimating Equation") proposée par LIANG & ZEGER (1986). 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'idée est simple : retenir l'estimateur du maximum de vraisemblance $\widehat{\beta}$ solution de (2)-(3) pour estimer $\beta$ dans le modèle avec effet aléatoire n'est certainement pas optimal puisqu'on ne tient pas compte de la structure de corrélation des $N_{it}$. Réécrivons le système (2)-(3) sous forme vectorielle :

$$\sum_{i=1}^{n} \boldsymbol{X}_i^t(\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]) = \boldsymbol{0} \text{ où } \boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT_i})^t. \tag{5}$$

La matrice de covariance des $N_{it}$ dans le modèle de Poisson avec indépendance sérielle est

$$\boldsymbol{A}_i = \begin{pmatrix} \lambda_{i1} & 0 & \cdots & 0 \\ 0 & \lambda_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{iT_i} \end{pmatrix}.$$

Cette matrice ne rend donc compte ni de la surdispersion, ni de la dépendance sérielle présente dans les données. Si on fait apparaître explicitement la matrice $\boldsymbol{A}_i$ dans (5), on obtient

$$\sum_{i=1}^{n} \left( \frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}[\boldsymbol{N}_i] \right)^t \boldsymbol{A}_i^{-1} (\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]) = \mathbf{0} \tag{6}$$

puisque

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}[\boldsymbol{N}_i] = \boldsymbol{A}_i \boldsymbol{X}_i .$$

Le principe des GEE consiste à substituer à $\boldsymbol{A}_i$ dans (6) un candidat plus raisonnable pour la matrice variance-covariance de $\boldsymbol{N}_i$, plus raisonnable signifiant ici rendant compte de la surdispersion et de la corrélation temporelle. Spécifions à présent une forme plausible pour la matrice de covariance de $\boldsymbol{N}_i$ : on pourrait penser à

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

où la matrice de corrélation $\boldsymbol{R}_i(\boldsymbol{\alpha})$ rend compte de la dépendance sérielle existant entre les composantes de $\boldsymbol{N}_i$ et dépend d'un certain nombre de paramètres $\boldsymbol{\alpha}$. La matrice $\boldsymbol{R}_i$ est une sous-matrice carrée de dimension $T_i \times T_i$ d'une matrice $\boldsymbol{R}$ de dimension $T_{\max} \times T_{\max}$ dont les éléments ne dépendent pas des caractéristiques $\boldsymbol{x}_{it}$ de l'individu $i$. La surdispersion est quant à elle prise en compte puisque $\mathbb{V}\text{ar}[N_{it}] = \phi \lambda_{it}$. Notez que la matrice $\boldsymbol{V}_i$ ainsi définie n'est la matrice de covariance de $\boldsymbol{N}_i$ que si $\boldsymbol{R}_i$ est la matrice de corrélation de $\boldsymbol{N}_i$, ce qui n'est pas nécessairement le cas.
Comme annoncé ci-dessus, l'idée consiste alors à substituer la matrice $\boldsymbol{V}_i$ à $\boldsymbol{A}_i$ dans (6), et de retenir comme estimation de $\boldsymbol{\beta}$ la solution de

$$\sum_{i=1}^{n} \left( \frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}[\boldsymbol{N}_i] \right)^t \boldsymbol{V}_i^{-1} (\boldsymbol{n}_i - \mathbb{E}[\boldsymbol{N}_i]) = \mathbf{0} \tag{7}$$

Cette dernière relation exprime également une orthogonalité entre les résidus de régression et les variables explicatives. Les estimateurs ainsi obtenus sont convergents quel que soit le choix de la matrice $\boldsymbol{R}_i(\boldsymbol{\alpha})$. On sent évidemment bien qu'ils seront d'autant plus précis que $\boldsymbol{R}_i(\boldsymbol{\alpha})$ est proche de la véritable matrice de corrélation de $\boldsymbol{N}_i$.

### *3.3 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 de corrélation $\boldsymbol{R}_i$ qui tient compte de la dépendance entre les observations relatives à un même assuré. Cette matrice de dimension $T_i \times T_i$ 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 $\boldsymbol{\alpha}$.

Si $\boldsymbol{R}_i(\boldsymbol{\alpha}) =$ Identité, (7) donne exactement les équations de vraisemblance (5) sous l'hypothèse d'indépendance.

En général, on spécifie dans le cadre de la tarification a priori une matrice $\boldsymbol{R}_i(\boldsymbol{\alpha})$ traduisant une structure de type autorégressive. Ainsi, les éléments diagonaux de $\boldsymbol{R}_i$ valent 1 et hors diagonale, l'élément $jk$ vaut $\alpha_{|j-k|}$ pour $|j-k| \leq m$ et 0 pour $|j-k| > m$. On prendra $m = T_{\max} - 1$. 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.

### *3.4 Obtention des estimations*

L'équation (7) est généralement résolue à l'aide d'une méthode du score de Fisher modifiée pour $\boldsymbol{\beta}$ et une estimation des moments pour $\boldsymbol{\alpha}$ (nous renvoyons le lecteur à LIANG & ZEGER (1986) pour une description complète de la méthode). Spécifiquement, partant d'une valeur initiale $\widehat{\boldsymbol{\beta}}^{(0)}$ solution du système (2)-(3), nous calculons

$$\widehat{\boldsymbol{\beta}}^{(j+1)} = \widehat{\boldsymbol{\beta}}^{(j)} + \left\{ \sum_{i=1}^{n} \boldsymbol{D}_i^t(\widehat{\boldsymbol{\beta}}^{(j)}) \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{\alpha}(\widehat{\boldsymbol{\beta}}^{(j)})) \boldsymbol{D}_i(\widehat{\boldsymbol{\beta}}^{(j)}) \right\}^{-1}$$
$$\left\{ \sum_{i=1}^{n} \boldsymbol{D}_i^t(\widehat{\boldsymbol{\beta}}^{(j)}) \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{\alpha}(\widehat{\boldsymbol{\beta}}^{(j)})) \boldsymbol{S}_i(\widehat{\boldsymbol{\beta}}^{(j)}) \right\}$$

où $\boldsymbol{D}_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}[\boldsymbol{N}_i]$ et $\boldsymbol{S}_i(\boldsymbol{\beta}) = \boldsymbol{N}_i - \mathbb{E}[\boldsymbol{N}_i]$. A chaque étape, $\phi$ et $\boldsymbol{\alpha}$ sont réestimés à partir des résidus de Pearson

$$r_{it}^P = \frac{n_{it} - \lambda_{it}}{\sqrt{\lambda_{it}}}$$

grâce aux formules

$$\widehat{\phi} = \frac{1}{\sum_{i=1}^n T_i - p} \sum_{i=1}^n \sum_{t=1}^{T_i} \{r_{it}^P\}^2$$

et

$$\widehat{\alpha}_\tau = \frac{1}{\widehat{\phi}\left(\sum_{i|T_i>\tau}(T_i - \tau) - p\right)} \sum_{i|T_i>\tau} \sum_{t=1}^{T_i-\tau} r_{it}^P r_{it+\tau}^P .$$

### *3.5 Illustration numérique*

La dépendance sérielle des $N_{it}$ à $i$ fixé ayant clairement été mise en évidence à la Section 3.1, il importe de mesurer l'impact de l'hypothèse d'indépendance sur l'estimation des fréquences. 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. L'estimation de la "working correlation matrix" de structure autorégressive d'ordre 2 (i.e. d'ordre $T_{\max} - 1$) donne

$$\begin{pmatrix} 1 & 0.0493 & 0.0462 \\ 0.0493 & 1 & 0.0493 \\ 0.0462 & 0.0493 & 1 \end{pmatrix}$$

et $\widehat{\phi} = 1.3437$.

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).

| Variable | Level | Coeff $\beta$ | Std Error | 95% Conf Limit | | Z | Pr> $\|Z\|$ |
|---|---|---|---|---|---|---|---|
| Intercept | | −1.9233 | 0.0319 | −1.9858 | −1.8608 | −60.32 | < .0001 |
| Sexe | Femme | −0.0581 | 0.0289 | −0.1148 | −0.0014 | −2.01 | 0.0446 |
| Sexe | Homme | 0 | 0 | 0 | 0 | . | . |
| Age | 17 − 22 | 0.6586 | 0.0617 | 0.5376 | 0.7797 | 10.67 | < .0001 |
| Age | 23 − 30 | 0.2557 | 0.0281 | 0.2006 | 0.3107 | 9.10 | < .0001 |
| Age | > 30 | 0 | 0 | 0 | 0 | . | . |
| Puissance | ≥ 66kW | 0.0532 | 0.0292 | −0.0041 | 0.1105 | 1.82 | 0.0686 |
| Puissance | < 66kW | 0 | 0 | 0 | 0 | . | . |
| Ville | Grande | 0.2542 | 0.0332 | 0.1892 | 0.3192 | 7.67 | < .0001 |
| Ville | Moyenne | 0.0719 | 0.0336 | 0.0060 | 0.1379 | 2.14 | 0.0326 |
| Ville | Petite | 0 | 0 | 0 | 0 | . | . |

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

| Source | DF | ChiSquare | Pr > ChiSq |
|---|---|---|---|
| Sexe | 1 | 4.09 | 0.0431 |
| Age | 2 | 128.79 | < .0001 |
| Puissance | 1 | 3.28 | 0.0701 |
| Ville | 2 | 60.71 | < .0001 |

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

### *3.6 Impact sur les fréquences*

Pour terminer, comparons les fréquences obtenues en supposant l'indépendance sérielle ou en reconnaissant explicitement la dépendance temporelle; celles-ci sont fournies aux Tableaux 11 et 12. 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 fréquences, lesquels sont plus larges dans l'approche GEE.

| Classes de risque | | | | Fréquences | | |
|---|---|---|---|---|---|---|
| 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 |
| | | $\geq 66$ | 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 |
| | | $\geq 66$ | 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 |
| | | $\geq 66$ | 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 |
| | | $\geq 66$ | 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 |
| | | $\geq 66$ | 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 |
| | | $\geq 66$ | Petite | 0.13462 | 0.14451 | 0.15513 |
| | | | Moyenne | 0.14549 | 0.15587 | 0.1670 |
| | | | Grande | 0.17431 | 0.18639 | 0.1993 |

Table 11: Estimations des fréquences des différentes classes de risque sous l'hypothèse d'indépendance.

| Classes de risque | | | | Fréquences | | |
|---|---|---|---|---|---|---|
| Sexe | Age | Puissance | Ville | Inf | Prime | Sup |
| Homme | 17-22 | $< 66$ | Petite | 0.24954 | 0.28233 | 0.31942 |
| | | | Moyenne | 0.26804 | 0.30348 | 0.34338 |
| | | | Grande | 0.32137 | 0.36405 | 0.41240 |
| | | $\geq 66$ | Petite | 0.26135 | 0.29776 | 0.33924 |
| | | | Moyenne | 0.28104 | 0.31996 | 0.36428 |
| | | | Grande | 0.33671 | 0.38395 | 0.43781 |
| | 23-30 | $< 66$ | Petite | 0.17725 | 0.18869 | 0.20086 |
| | | | Moyenne | 0.19059 | 0.20276 | 0.21571 |
| | | | Grande | 0.22865 | 0.24331 | 0.25890 |
| | | $\geq 66$ | Petite | 0.18531 | 0.19899 | 0.21369 |
| | | | Moyenne | 0.19965 | 0.21384 | 0.22904 |
| | | | Grande | 0.23918 | 0.25660 | 0.27529 |
| | $> 30$ | $< 66$ | Petite | 0.13727 | 0.14612 | 0.15554 |
| | | | Moyenne | 0.14759 | 0.15701 | 0.16704 |
| | | | Grande | 0.17748 | 0.18842 | 0.20003 |
| | | $\geq 66$ | Petite | 0.14455 | 0.15410 | 0.16429 |
| | | | Moyenne | 0.15578 | 0.16560 | 0.17603 |
| | | | Grande | 0.18701 | 0.19871 | 0.21114 |
| Femme | 17-22 | $< 66$ | Petite | 0.23532 | 0.26634 | 0.30156 |
| | | | Moyenne | 0.25294 | 0.28625 | 0.32396 |
| | | | Grande | 0.30359 | 0.34350 | 0.38865 |
| | | $\geq 66$ | Petite | 0.24518 | 0.28095 | 0.32193 |
| | | | Moyenne | 0.26381 | 0.30190 | 0.34550 |
| | | | Grande | 0.31640 | 0.36227 | 0.41479 |
| | 23-30 | $< 66$ | Petite | 0.16643 | 0.17804 | 0.19045 |
| | | | Moyenne | 0.17918 | 0.19131 | 0.20427 |
| | | | Grande | 0.21541 | 0.22957 | 0.24466 |
| | | $\geq 66$ | Petite | 0.17254 | 0.18776 | 0.20427 |
| | | | Moyenne | 0.18606 | 0.20177 | 0.21880 |
| | | | Grande | 0.22333 | 0.24210 | 0.26248 |
| | $> 30$ | $< 66$ | Petite | 0.12867 | 0.13787 | 0.14773 |
| | | | Moyenne | 0.13851 | 0.14815 | 0.15847 |
| | | | Grande | 0.16687 | 0.17778 | 0.18940 |
| | | $\geq 66$ | Petite | 0.13426 | 0.14540 | 0.15747 |
| | | | Moyenne | 0.14477 | 0.15625 | 0.16864 |
| | | | Grande | 0.17409 | 0.18750 | 0.20193 |

Table 12: Estimations des fréquences des différentes classes de risque dans l'approche GEE.

## 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 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 fréquences de sinistres. 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 fréquence. Ceci doit être pris en considération lors de la fixation de la hauteur du chargement de sécurité. Dans des cas extrêmes, la prise en compte de la dépendance temporelle peut même conduire à l'exclusion de certaines variables tarifaires.

**Remerciements**

Nous tenons à remercier les deux arbitres pour leur lecture extrêmement attentive du manuscrit et leurs commentaires détaillés et pertinents. Leurs rapports ont permis des améliorations substantielles du travail.

## Bibliographie


1. BESSON, J.-L. & PARTRAT, CH. (1992). Trend et systèmes de Bonus-Malus. *ASTIN Bulletin* **22**, 11–31.
2. BOLANCÉ, C., GUILLÉN, M. & PINQUET, J. (2003). Time-varying credibility for frequency risk models: estimation and tests for autoregressive specification on the random effect. *Insurance: Mathematics & Economics*, to appear.
3. BROUHNS, N. & DENUIT, M. (2003). Tarification automobile et modèles généralisés additifs. Discussion Paper, Institut de Statistique, Université catholique de Louvain, Belgique.
4. DENUIT, M., PITREBOIS, S. & WALHIN, J.-F. (2001). Méthodes de construction de systèmes bonus-malus en RC auto. *ACTU-L* **1**, 7–38
5. LIANG, K.Y. & ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
6. PINQUET, J. & GUILLEN, M. & BOLANCÉ, C. (2001). Allowance for the age of claims in bonus-malus systems. *ASTIN Bulletin* **31**, 337–348.
7. ZEGER, S.L., LIANG, K.Y. & ALBERT, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.



Sandra Pitrebois
Jean-François Walhin
Secura Belgian Re
Avenue des Nerviens 9-31 boîte 6
B-1040 Bruxelles
Belgium

Michel Denuit
Jean-François Walhin
Université Catholique de Louvain
B-1348 Louvain-la-Neuve
Belgium

## Résumé

Souvent, les actuaires utilisent plusieurs années d'observation de leur portefeuille automobile afin de personnaliser les fréquences de sinistres. Ils négligent cependant la plupart du temps 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 approche sur les estimations des fréquences de 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.

## Summary

In third party liability automobile insurance, actuaries often pool several observation periods to determine the price list. The serial dependence arising from the fact that the same individuals are followed and produce correlated claim numbers is nevertheless almost always neglected in practice. This paper examines with the help of the Generalized Estimating Equations (GEE) technique the impact of this approach on the estimation of claim frequencies. A numerical illustration on a Belgian portfolio observed during three years is performed with the statistical software SAS.

## Zusammenfassung

Oft verwenden die Aktuare mehrere Beobachtungsjahre ihres Autohaftpflicht-Portefeuilles um Schadenfrequenzen zu bestimmen. Meistens wird jedoch die serielle Abhängigkeit zwischen den Daten des gleichen Individuums vernachlässigt. Mit Hilfe von "Generalized Estimating Equations" (GEE) wird im Artikel die Auswirkung dieses Vorgehens auf die Schadenfrequenzen untersucht. Ein numerisches Beispiel, basierend auf einem belgischen Portefeuille, welches drei Beobachtungsjahre umfasst, verdeutlicht diesen Zugang. Dabei wurde die Statistiksoftware SAS eingesetzt.