UNIVERSITE CATHOLIQUE DE LOUVAIN

INSTITUT DES SCIENCES ACTUARIELLES

Segmentation du nombre de sinistres en assurance :
de la loi de Poisson aux modèles gonflés à zéro et à barrière

Membres du jury :

Prof. Michel Denuit *(Promoteur)*
Prof. Pierre Devolder *(Président du Jury)*
Prof. Montserrat Guillén
Prof. Philippe Lambert
Prof. Jean-François Walhin

Thèse présentée en vue de
l'obtention du grade de
Docteur en Sciences
(orientation sciences actuarielles)
par :

Jean-Philippe Boucher

Louvain-La-Neuve, Belgique
Mai 2007

Remerciements

Je tiens à remercier Michel Denuit, mon directeur, pour m'avoir donné autant d'informations pertinentes tant sur la science actuarielle que sur les statistiques. Grâce à ses commentaires, je pense avoir encore de la matière pour plusieurs idées de recherches au cours des prochaines années ! Je te remercie plus particulièrement d'avoir accepté de diriger un Québécois, et d'avoir accepté ses conditions particulières de travail...

Pour m'avoir accepté dans son équipe de travail pendant plusieurs semaines, je tiens aussi à remercier Montserrat Guillén. Pendant ce trop court séjour à Barcelone (qui a tout de même causé de la jalousie dans le département d'actuariat de l'UCL), j'ai pu bénéficier de ses précieux conseils théoriques et de ses remarques judicieuses sur la vie universitaire. Merci beaucoup, j'espère pouvoir avoir la chance de travailler de nouveau avec toi.

Je tiens aussi à souligner l'apport de tous les professeurs que j'ai pu croiser, tant au D.E.A. (maîtrise...) à l'Université de Lyon I, que lors de mon baccalauréat à l'Université Laval à Québec.

Je tiens tout particulièrement à exprimer mon affection à tous mes amis qui ont su m'encourager au cours des dernières années :

- Mes amis de Louvain-la-Neuve, Donatien et Jérôme, avec qui j'ai pu avoir de grandes discussions philosophiques sur l'actuariat, la finance, la recherche en général et la politique. J'ai bien hâte de découvrir votre fameux théorême Hainaut-Barbarin. Merci aussi de m'avoir appris que Bush n'est pas seulement le nom de l'actuel président des États-Unis ;
- Mes amis dijonnais : Sabrina, Pascal, Stéphanie et Valérie avec qui j'ai appris à utiliser différemment plusieurs éléments gastronomiques français (patate, fromage, poulet, etc.). Nos visites de plusieurs endroits particulièrement enivrants de la Bourgogne furent agréables... malgré le froid ;
- A Rudy, qui fut l'un des premiers Français à s'intéresser à nous. Je te remercie d'avoir pris

soin de nous et de nous avoir acceptés dans ton groupe d'amis pour nos magnifiques voyages dans le Pays Basque et en Toscane, ainsi que pour ces fêtes du Jour de l'An ;
– Merci à Mathieu, mon cousin préféré, avec qui j'ai passé de nombreuses heures à discuter de tout, mais surtout de rien sur MSN (tap tap tap) ;
– Je remercie tous mes amis québécois qui ont pris la peine de venir nous visiter à Dijon pendant ces dernières années. Dans les moments plus difficiles, votre présence nous a beaucoup aidés (à moins que ce soit la visite de Beaune ?). Merci donc à Délo, Antoine et Anni, Isabelle et Martin, Sébastien et Bénédicte, Eric, Marie-Andrée, Pascale... et toute la belle-famille ;
– Malgré qu'ils ne soient pas venus nous voir en Europe, je tiens à souligner mon amitié pour mes autres amis québécois qui ont su nous accueillir lors de nos visites annuelles dans la Belle Province ;
– Et un dernier grand merci à tous mes autres amis belges, français, suisses et espagnols.

Bien entendu, je veux exprimer toute mon affection à ma mère Francine qui m'a toujours soutenu et encouragé. Tu as toujours su comment me donner confiance en moi. Je te remercie de m'avoir appuyé dans cette aventure en Europe. Un mot aussi pour mon défunt père Pierre qui a réussi à me donner le goût de me dépasser et à me donner la curiosité intellectuelle nécessaire pour réussir mes études et mes autres projets. J'espère que tu es fier de tout ce que j'accomplis.

Mon amour va évidemment à ma *blonde*, qui a su me donner l'opportunité et surtout la motivation nécessaire pour retourner aux études. Véronique, dans un avenir pas si lointain, de grands projets nous attendent et je suis vraiment heureux d'avoir la chance de les vivre avec toi.

Introduction

Contexte historique

Données de comptage

Malgré que le titre de la thèse fasse référence au nombre de réclamations à l'assureur, il est clair qu'avant toute chose, ce travail touche aux données de comptage, c'est-à-dire aux données prenant comme valeurs les nombres entiers $0, 1, 2, ...$ Il existe plusieurs exemples de ces événements aléatoires. Nommons ainsi le nombre de fautes de frappe par page dans un document, le nombre de brevets déposés annuellement par des compagnies d'un secteur donné ou le nombre annuel de grèves dans le secteur manufacturier.

A la base de ce type d'analyse se trouve la loi de Poisson, dérivée par Siméon-Denis Poisson (1781-1840) lui même, dans son ouvrage de 1837 intitulé *Recherches sur la probabilité des jugements en matière criminelle et matière civile*. L'auteur a pu montrer que la loi de Poisson peut être représentée comme une limite de la loi binomiale. En effet, en notant $Y_{n,\pi}$ le nombre total de succès de n essais indépendants avec π représentant la probabilité de succès pour chaque essai, alors :

$$\Pr[Y_{n,\pi} = k] = \binom{n}{k}\pi^k(1-\pi)^{n-k}, \quad k = 0, 1, 2, ...$$

Dans le cas où $n \to \infty$, $\pi \to 0$ et $n\pi = \mu > 0$, signifiant que la moyenne μ demeure constante alors que $n \to \infty$, nous obtenons :

$$\lim_{n\to\infty}\left[\binom{n}{k}\left(\frac{\mu}{n}\right)^k\left(1-\frac{\mu}{n}\right)^{n-k}\right] = \frac{\mu^k e^{-\mu}}{k!}$$

Ce qui représente la distribution de probabilité de la loi de Poisson. Parce qu'elle est dérivée de la limite d'une loi binomiale lorsque la probabilité du succès est petite par rapport au nombre d'essais, il arrive fréquemment que la loi de Poisson soit appelée *loi des petits nombres*. Parmi les premières études empiriques de la loi de Poisson, notons la célèbre étude du nombre de soldats tués par ruade dans 10 corps de l'armée prussienne (von Bortkewisch (1898)).

Régression

Historiquement, afin de considérer l'effet de certaines variables explicatives x_i sur une variable aléatoire, l'estimation par les moindres carrés était utilisée. Néanmoins, parce que cette évaluation suppose une loi Gaussienne, elle n'est pas appropriée pour les données de comptage. Conséquemment, l'idée est de modéliser le paramètre de moyenne de la loi de Poisson à l'aide de

régresseurs, afin de créer un modèle de régression de Poisson. Typiquement, ces régresseurs x_i seront introduits dans le modèle par la fonction de score $\mu_i = \exp(x_i'\beta)$ créant ainsi diverses classes de risque.

Une découverte importante pour l'analyse des données de comptage eu lieu en 1972 grâce à Nelder et Wedderburn. Les auteurs ont développé la théorie des modèles linéaires généralisés (GLM), dont la loi de Poisson fait partie. Ces modèles peuvent être vus comme une généralisation de la régression linéaire classique et expriment les équations de vraisemblance des paramètres d'une régression pour les lois de probabilité provenant de la famille exponentielle. Quelques années plus tard, les papiers de Gouriéroux, Monfort et Trognon (1984a,b) ont apporté un argument important pour justifier l'utilisation de ces modèles. En effet, les auteurs ont prouvé que lorsque la moyenne de la distribution est bien définie, il y a convergence des estimateurs des GLM peu importe si la distribution a été correctement sélectionnée. Notons, toutefois, que même si ces paramètres de moyennes restent convergents, la justesse de l'écart-type de ces derniers n'est aucunement vérifiée dans le cas d'une erreur sur la vraie distribution.

Au niveau pratique, le choix de la loi de Poisson pour modéliser les paramètres de régression n'est pas toujours idéal. En effet, dans de nombreuses situations, malgré l'introduction de régresseurs, il existe des différences dans chaque classe de risque, probablement causées par l'absence de régresseurs importants. Puisque la loi de Poisson n'est efficace que pour la modélisation de classes de risque homogènes, l'apparition d'hétérogénéité peut être problématique.

La modélisation de l'hétérogénéité par l'introduction d'un terme aléatoire dans la moyenne est une façon de corriger le modèle. Lorsque le terme aléatoire ajouté est distribué selon une loi gamma, la loi binomiale négative peut être déduite. Deux exemples célèbres de la modélisation de cette loi sont pertinents : Greenwood et Yule (1920), qui ont montré que cette distribution suppose une contagion apparente entre les réalisations, et Eggenberger et Polya (1923) qui ont montré que la même distribution peut aussi impliquer une vraie contagion. En conséquence, en généralisant, on peut conclure qu'on ne peut pas connaître le type de contagion supposée par un modèle de comptage avec hétérogénéité (le théorême d'impossibilité de Neyman de 1965).

En plus de la présence d'hétérogénéité dans les données, d'autres facteurs peuvent justifier l'analyse et la recherche d'autres distributions de comptage. Notons ainsi, entre autres, la trop grande

restriction supposée par l'unique paramètre de la loi de Poisson, la censure ou la troncation des données, le grand nombre de valeurs à zéro ou l'échec de l'hypothèse d'indépendance conditionnelle entre les sujets analysés.

Données de panel

Les données de panel, ou données longitudinales, réfèrent à des observations dans lesquelles des unités, telles que des personnes, des assurés, des voitures ou des ménages cumulent plusieurs observations individuelles au cours du temps. La structure d'analyse de ce type de données peut s'inspirer de l'analyse multivariée ou des séries temporelles. Dans ce genre de modèles, une dépendance peut être supposée entre les observations d'une même unité, alors que chaque unité est supposée indépendante l'une de l'autre. Un avantage important de ce genre d'analyse, par rapport aux modèles précédents qui supposaient une indépendance entre chaque observation, est qu'il permet une meilleure modélisation de l'hétérogénéité individuelle. De plus, comme noté par Neyman (1965), ce genre d'analyse peut permettre de distinguer la vraie contagion de l'apparente.

Le papier de Hausman, Hall et Griliches (1984) a introduit une nouvelle façon de travailler avec ces données de panel. Dans ce modèle, un facteur individuel commun à toutes les réalisations individuelles, permet d'introduire une dépendance entre celles-ci. Ce nouveau paramètre peut être aléatoire et suivre une distribution appropriée quelconque, ou encore être fixe et déterministe. Notons aussi le papier de Liang et Zeger (1986) qui ont utilisé une autre approche pour estimer les paramètres du modèle à effets aléatoires. En effet, au lieu de supposer un modèle complètement paramétrique pour les données de panel, les auteurs ont généralisé la méthode d'estimation des GLM afin d'inclure la dépendance introduite par l'effet aléatoire individuel.

Conditionnellement à l'observation de certaines données individuelles, il est donc possible de déduire la distribution prédictive d'une nouvelle observation. La dépendance entre les observations passées et les futures observations est déjà construite dans les modèles construits sous l'angle des modèles multivariés, des séries temporelles ou du modèle à effets fixes de Hausman, Hall et Griliches (1984). A l'opposé, le modèle à effets aléatoires de Hausman, Hall et Griliches (1984) nécessite plutôt une approche Bayésienne pour la prédiction. En effet, le terme aléatoire ajouté à la distribution de fréquence est mis à jour à chaque période selon les observations recueillies. L'analyse *a posteriori* de ce terme aléatoire permet un ajustement révélant ainsi partiellement l'effet des caractéristiques non observées sur la distribution du comptage.

Tarification a priori

Les distributions pour données de comptage sont évidemment appropriées pour modéliser le nombre de réclamations à l'assureur. Bien que ce document se spécialise pour les données provenant d'assurance automobile, il peut être facilement généralisé pour d'autres types de couvertures telles que l'assurance habitation, l'assurance commerciale, l'assurance responsabilité professionnelle ou autres. Cette modélisation du nombre de réclamations est une tâche importante pour l'actuaire car elle lui permet de calculer la prime à charger aux assurés.

Cette prime d'assurance est souvent calculée sous base annuelle. Cette unité de calcul est cohérente puisqu'elle représente bien l'exposition au risque, même si dans certaines situations, cette unité d'exposition aurait être plus juste. Pensons, par exemple, au nombre de kilomètres parcourus pour l'assurance automobile, qui modélise encore mieux l'exposition au risque d'accidents que la durée de couverture. Néanmoins, parce que les coûts administratifs associés à une telle mesure sont souvent trop élevés, les assureurs préfèrent normalement l'utilisation du temps de couverture.

Bien évidemment, en dehors du nombre de réclamations, le calcul de la prime annuelle d'un portefeuille d'assurance inclus le montant des réclamations. En effet, l'analyse de la sinistralité du portefeuille d'assurance, dans le but d'en calculer une prime annuelle, s'effectue en séparant le nombre des coûts moyens. Cette séparation permet une analyse et une compréhension plus en profondeur de la sinistralité tout en favorisant l'analyse statistique puisque plusieurs recherches scientifiques se développent autour des données de comptage, alors qu'une analyse commune compliquerait la tâche (loi de Tweedie).

Ainsi, en associant d'une quelconque façon la prime pour la fréquence de réclamation à celle obtenue par un exercice similaire pour la sévérité, l'actuaire peut obtenir une prime pure, représentant la prime à charger à un assuré pour couvrir ses réclamations annuelles espérées. Le profit, les frais fixes, les frais variables, les taxes, le coût de la réassurance ainsi qu'une charge supplémentaire pour couvrir les variations aléatoires sont ajoutés à la prime pure afin d'obtenir la prime réelle. Des considérations au niveau du marketing peuvent aussi venir influencer la valeur de la prime, alors que d'autres considérations économiques favorisant l'équilibre entre les demandeurs d'assurance et les assureurs peuvent aussi avoir un rôle à jouer.

Puisque les assurés ne sont pas tous égaux devant le risque (certains sont plus risqués que d'autres), il convient aussi de considérer les caractéristiques du risque dans la tarification de la prime d'assurance. Ces caractéristiques du risque assuré utilisées dans la structure de tarification aide à partager de manière équitable les coûts d'assurance entre tous les assurés. Le but de l'exercice est de séparer les contrats (et les assurés) en plusieurs classes de risque, de façon à ce qu'à l'intérieur d'une catégorie, les risques puissent être considérés comme *équivalents*. Puisque l'historique de réclamations est exclu de l'analyse, cet exercice est fréquemment appelé *tarification a priori*.

Non seulement cet exercice de segmentation est important au niveau de l'équité entre assuré, mais aussi pour la santé financière de l'assureur qui réduit ainsi son risque d'anti-sélection. En effet, en ne considérant pas un critère de sélection que les autres assureurs utilisent dans leur tarification, un assureur risque de ne pas bien identifier et de sous-tarifer certains mauvais risques. Par opposition, puisque surtarifés, les meilleurs risques iront s'assurer ailleurs. En conséquence, l'assureur affichera une expérience de sinistre pire que prévue, puisque seulement la partie la plus risquée de sa clientèle cible sera intéressée aux prix de notre assureur.

Notons par ailleurs que ce travail de segmentation des risques était déjà effectué au 18^e siècle, par la célèbre Lloyd's de Londres pour couvrir les risques en assurance maritime. En effet, celle-ci basait sa tarification et son type de protection d'assurance selon les caractéristiques individuelles du risque. En assurances personnelles, l'intérêt plus particulier de ce travail de segmentation des différents assurés selon leurs caractéristiques est assez récent puisque les actuaires ne s'y sont intéressés sérieusement qu'à partir des années 1980.

Bien que le choix des caractéristiques des assurés pouvant influencer le nombre de réclamations peut paraître être un problème exclusivement statistique, diverses considérations politiques et sociologiques influencent ce choix. En effet, il pourrait être socialement impossible d'utiliser certaines caractéristiques du risque (religion, sexe, classe sociale, etc.), malgré le fait que leur impact sur la fréquence de réclamation soit indéniable. D'autres critères utilisés pour sélectionner les variables de classification ont une influence. Notons ainsi :

- la mesurabilité, signifiant qu'une variable de classification se doit d'être facilement évaluable. La rapidité des réflexes ou l'agressivité au volant, par exemple, ne peuvent pas être facilement utilisables ;
- le critère opérationnel, signifiant que les variables de classification doivent être définies de

manière objectives et que les coûts administratifs d'utilisation soient minimaux ;
– le critère légal, selon lequel les variables utilisées peuvent être assujetties à des dispositions spécifiques dans la constitution ou dans les diverses législations.

En assurance automobile, ces caractéristiques du risque utilisées pour partitionner les assurés, appelés variables *a priori*, sont habituellement l'âge du conducteur, le sexe du conducteur, le type, le modèle et l'âge du véhicule ou encore la ville de résidence, pour ne citer que quelques exemples.

Tarification a posteriori et crédibilité

Comme noté par Lemaire (1995), si les assureurs n'étaient autorisés à utiliser qu'une seule variable de tarification, celle-ci devrait être sous la forme d'un système de mérite. En effet, le meilleur prédicateur du nombre de réclamations n'est pas l'âge ou le sexe du conducteur, mais bien son expérience passée de sinistres.

Au cours des années, les assureurs ont réussi à accumuler des informations longitudinales sur leurs assurés, motivant ainsi l'utilisation des modèles pour données de panel et les analyses *a posteriori*. Conséquemment, lorsqu'un assuré accumule un historique de perte, sa prime annuelle est modifiée en fonction de son expérience. Contrairement aux idées populaires, il ne s'agit pas d'un mécanisme visant à ce que l'assuré rembourse ce que sa compagnie a dû payer pour ses sinistres passés, mais bien d'une façon de mieux évaluer le risque couvert par la compagnie d'assurance. La révision de la prime se justifie de plusieurs manières : l'expérience passée donne une meilleure idée du risque réel de l'assuré, mais aussi parce qu'un accident d'automobile peut engendrer une contagion en modifiant les perceptions de conduite.

Cette analyse des primes prédictives est fortement liée à la théorie de la crédibilité qui fascine depuis longtemps les actuaires. Rappelons, par exemple, le cas de la société General Motors qui, en 1910, demandait une réduction de prime pour son assurance contre les accidents de travail. La compagnie clamait que son expérience de sinistres était meilleure que le reste de l'industrie, et qu'elle était suffisamment grande pour justifier de ne se baser que sur son expérience individuelle. Les actuaires ont depuis longtemps dû faire face à ce genre de dilemme où se confrontent la tarification basée sur l'expérience du groupe d'assurance à celle de l'individu.

En 1967, Bühlmann se penche sur le problème et développe un modèle de crédibilité où la prime d'assurance d'un assuré est exprimée linéairement en fonction d'une moyenne pondérée entre

son expérience passée et sa prime *a priori*. D'autres travaux ont fait évoluer cette théorie depuis, tels que les papiers de Bühlmann et Straub (1970), Hachemeister (1975) ou encore Jewell (1975). De nos jours, cette approximation linéaire de la prime prédictive n'est plus toujours nécessaire puisque l'analyse Bayésienne (empirique) exacte, utilisant diverses fonctions de pertes, est populaire. Notons, à ce sujet, le papier de Frees, Young et Luo (1999) qui explique le lien entre les données de panel et les modèles de crédibilité.

Nombre de réclamations

Evidemment, le point de départ de l'analyse du nombre de réclamation est souvent la loi de Poisson, avec laquelle le paramètre de moyenne est modélisé pour incorporer les caractéristiques de l'assuré. L'utilisation de la loi de Poisson pour modéliser le nombre de réclamations est intéressante car elle réfère au processus de Poisson.

En effet, la loi de Poisson suppose que les temps d'attente entre deux réclamations sont distribués exponentiellement. Ceci signifie que le nombre de réclamations survenant dans l'intervalle de temps $[0, t]$ est distribué selon une loi de Poisson ayant un paramètre de moyenne λt, c'est-à-dire :

$$P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}.$$

Parce qu'il est bien connu que la loi exponentielle n'a pas de mémoire, cette propriété de la loi de Poisson explique que le paramètre de moyenne varie au prorata de l'exposition au risque. Concrètement, cela signifie que lorsque la prime d'assurance d'un assuré est calculée selon une loi de Poisson, celle-ci sera proportionnelle à son exposition au risque.

Parce qu'il existe une hétérogénéité dans les données d'assurance, de nombreuses généralisations de la loi de Poisson, basées sur diverses distribution d'hétérogénéité, ont été testées pour le nombre de réclamations, telles que la binomiale négative (Gossiaux et Lemaire (1981), Willmot (1987) et Besson et Partrat (1992)), la distribution Poisson-inverse Gaussienne dans Willmot (1987), Besson et Partrat (1992) et Tremblay (1992), la Poisson-Pascal généralisée dans Consul (1989) et la Poisson-Goncharov dans Denuit (1997).

Les distributions de comptage peuvent être construites différemment que par le choix d'une distribution de l'hétérogénéité. Nommons, par exemple, la distribution géométrique généralisée dans Gossiaux et Lemaire (1981) ou la distribution de Consul dans Islam et Consul (1992). Notons aussi

Keiding et Andersen (1995) qui se sont intéressé au temps d'attente entre les sinistres et ont utilisé le modèle semiparamétrique de Cox. Walhin (2001) ainsi que Yip et Yau (2005) se sont, quant à eux, intéressés au modèle Poisson gonflés à zéro.

Afin de modéliser une dépendance temporelle entre les contrats d'un même assuré, notons les papiers de Gouriéroux et Jasiak (2004) qui ont introduit en sciences actuarielles le modèle INAR(1), celui de Denuit et Lang (2004) qui utilisent des modèles additifs généralisés ou encore Pinquet (1997) qui généralise Hausman, Hall et Griliches (1984) en proposant d'utiliser des effets aléatoires dynamiques pour les données de panel.

Nature du problème

A la suite des récents développements dans les domaine de la biostatistique, de l'économétrie et même de l'actuariat, il semble pertinent d'analyser plus en détails les nombreuses possibilités pour modéliser le nombre de réclamations, tant pour les données avec indépendance temporelle que pour les données de panel. En dehors du besoin évident d'explorer ces autres types de modèles, la justification intuitive de ceux-ci est intéressante afin de tenir compte des particularités du domaine de l'assurance. Ainsi, l'effet des rabais annuels donnés aux assurés qui ne réclament pas (appelé *soif du bonus*) ou l'effet du changement de perception du conducteur après un accident, entre autres, ont certainement un impact sur les statistiques d'assurance.

Selon les différents modèles analysés, il est intéressant de voir l'impact des régresseurs sur la prime d'assurance. En effet, non seulement l'estimation de ces paramètres n'est pas la même selon le modèle, mais l'écart-type peut aussi varier substantiellement. Ainsi, dans certaines distributions, l'effet d'un régresseur pourrait être significatif alors qu'il serait exclu de l'analyse pour certains autres modèles.

Puisque de nombreux modèles statistiques sont possibles pour la modélisation du nombre de réclamations, le développement de divers moyens de comparaison de l'ajustement entre ces modèles devient aussi un exercice essentiel. En effet, il est primordial de pouvoir donner les moyens aux actuaires de choisir le meilleur modèle possible selon des procédures statistiques théoriquement valides.

Pour ces nouveaux modèles introduits en sciences actuarielles, l'analyse de la distribution prédictive, de même que le calcul des primes prédictives deviennent également nécessaires. Dans les

situations où le développement de ces distributions est ardu, des techniques numériques se doivent d'être proposées et développées. La comparaison avec les primes résultant des modèles plus classiques en actuariat est aussi un exercice important puisqu'il permet de voir les divergences et les erreurs potentielles dans les primes prédictives lors d'un mauvais choix de modèle. Dans ces situations où ces primes prédictives nécessitent un développement plus difficile, l'utilisation de la théorie de la crédibilité linéaire est un exercice intéressant puisque bien connu du milieu actuariel, et plus particulièrement en Amérique du Nord.

Puisque l'analyse des données de comptage est fréquente dans plusieurs domaines, bien que dirigées en fonction des besoins de la science actuarielle, les recherches effectuées au cours de la thèse peuvent aussi correspondre aux besoins de certains autres domaines de recherche.

Remarques générales

Cette thèse est construite sous le modèle d'une série d'articles indépendants. Afin d'aider la lecture, il n'y a qu'une bibliographie commune et le format des différents chapitres est le même. Selon les chapitres, certaines illustrations numériques se sont effectuées à l'aide d'une base de données d'assurance belge, alors que certaines autres utilisent des données espagnoles. Les applications numériques ont souvent été effectuées à l'aide du logiciel statistique SAS (procédures IML ou NLMIXED, par exemple).

Bien que les papiers soient indépendants, il peut être justifié de se questionner sur l'utilisation de techniques présentes dans certains chapitres alors qu'elles ne sont pas utilisées dans d'autres. Dans la majorité des cas, cela réfère à une situation d'approximation numérique ou certains modèles pouvaient être évalués numériquement alors que les techniques informatiques actuelles ne permettaient pas une telle évaluation pour certains autres modèles.

Plan des travaux de thèse

Modèles avec indépendance temporelle

Publications :

1. Boucher, J.-P., Denuit, M. et Guillén, M., **Risk Classification for Claim Counts : A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models**, *North American Actuarial Journal*,
2. Boucher, J.-P. et Guillén, M., **Semiparametric Models for Claim Counts : Various Analysis of the Gamma Heterogeneity**, *soumis à ASTIN Bulletin*,

Résumé :

La première partie de la thèse s'intéresse aux modèles de classification du nombre de réclamations sous l'hypothèse que toutes les observations analysées dans la base de données sont indépendantes. En introduisant les caractéristique du risque dans la moyenne par une fonction de score, le développement de ce type de modèles permet de mieux comprendre l'impact de certaines caractéristiques sur la probabilité de réclamer à l'assureur. De plus, en utilisant des modèles qui diffèrent de la loi de Poisson, ces travaux permettent une autre interprétation du mécanisme de réclamation à l'assureur. L'effet du bonus-malus, des déductibles ou du changement de perception du conducteur après un accident sont toutes des explications possibles à la divergence entre le nombre de réclamation et la loi de Poisson.

Dans la majorité des situations en assurance, les caractéristiques du risque ne sont pas toutes utilisées dans la tarification, soit parce qu'elles ne sont pas mesurables, soit parce qu'il serait socialement inacceptable de les utiliser. Ainsi, un facteur d'hétérogénéité permettant de capturer ces caractéristiques inobservables est ajouté au paramètre de moyenne de la loi de Poisson. Ce facteur additionnel, appelé hétérogénéité, est modélisé de façon paramétrique ou encore de manière non-paramétrique. Plusieurs méthodes peuvent être envisagées pour déterminer la forme non-paramétrique de cette hétérogénéité. Nous avons analysé ces diverses méthodes d'évaluations et comparé les résultats avec des méthodes complètement paramétriques.

Puisque de nombreux modèles pour données de comptage sont utilisés, des méthodes de comparaisons de modèles ont été développées. Certaines sont basées sur les distributions d'estimation des paramètres alors que certaines autres se basent sur la distribution du maximum de vraisemblance ou les critères d'information. Suite à ces analyses et ces comparaisons, nous pouvons voir que la

distribution de l'hétérogénéité de la loi de Poisson ne peut pas être directement ajustée avec une distribution paramétrique simple. De la même manière, nous avons remarqué qu'une distribution plus complexe que la loi de Poisson, telles que la distribution à barrière, la binomiale négative X ou les modèles gonflés à zéro génèrent non-seulement un meilleur ajustement que les distributions de Poisson avec hétérogénéité, mais que leur utilisation se justifie intuitivement pour les données d'assurance. Une grande partie de cette thèse se consacre justement à l'analyse de ces nouvelles distributions en assurance.

Modèles avec données de panel

Publications :

1. Boucher, J.-P. et Denuit, M., **Fixed versus Random Effects in Poisson Regression Models for Claim Counts : Case Study with Motor Insurance**, *ASTIN Bulletin 36 : 285-301, 2006*,
2. Boucher, J.-P., Denuit, M. et Guillén, M., **Models of Insurance Claim Counts with Time Dependence Based on Generalisations of Poisson and Negative Binomial Distributions**, *Variance*,
3. Boucher, J.-P. et Denuit, M., **Duration Dependence Models for Claim Counts**, *Blatter Deutsche Gesellschaft fur Versicherungsmathematik (German Actuarial Bulletin), sous presse*,

Résumé :

De manière similaire aux travaux effectués selon l'hypothèse que tous les contrats d'assurance sont indépendants, cette section de la thèse se consacre à l'analyse de modèles avec données de panel, supposant donc une dépendance entre les contrats d'un même assuré. Intuitivement, cette dépendance se justifie en considérant l'effet des variables de classification inconnues qui touchent tous les contrats d'un même assuré, ou encore en considérant l'impact d'un accident sur les habitudes de conduite d'un assuré.

Ainsi, en se basant sur les lois de Poisson et binomiale négative, divers modèles créant une dépendance temporelle entre les contrats d'un même assuré sont étudiés. De ces distributions sont aussi calculées les primes prédictives, c'est-à-dire les primes conditionnelles à un historique de sinistres.

Tout comme la partie précédente de la thèse, diverses interprétations des modèles sont proposées. De manière assez claire, un seul modèle sort du lot : le modèle où un effet aléatoire individuel affecte toutes les observations d'un même assuré. Néanmoins, avant d'utiliser ce modèle, une hypothèse essentielle d'indépendance entre les caractéristiques du risque et ce facteur aléatoire doit être vérifiée. Pour les données d'assurance, cette hypothèse n'est pas vérifiée en pratique et il est important d'en comprendre l'impact sur l'analyse actuarielle.

Nous montrons numériquement que cette dépendance entre les caractéristiques du risque et l'effet aléatoire biaise l'espérance du modèle, et donc la prime *a priori*. Toutefois, nous montrons que ce biais n'a que très peu d'importance en assurance puisque les paramètres obtenus avec cette dépendance représentent l'effet apparent des caractéristiques sur le risque de réclamer, ce qui est justement l'intérêt de l'analyse actuarielle dans le cas où certaines caractéristiques sont absentes de l'étude.

Nous tentons également d'utiliser le modèle à effets aléatoires avec une toute nouvelle sorte de distribution de comptage construite à l'aide des temps d'attente entre deux sinistres. Il est connu que la loi de Poisson suppose un temps exponentiel entre chaque sinistre. Ainsi, nous avons voulu voir ce que serait l'impact d'un choix différent pour cette distribution. Parce que les nouveaux assurés et les assurés qui quittent la compagnie d'assurance en cours de couverture ont des expériences de sinistres particulières, il arrive souvent que ceux-ci soient étudiés séparément des autres assurés lors de la tarification. Avec ces modèles de temps d'attente permettant de tarifer ces assurés différemment qu'au prorata de leur exposition au risque, nous introduisons une nouvelle façon de considérer l'analyse des différents contrats d'assurance.

Généralisations des modèle gonflés à zéro et à barrière

Publications :

1. Boucher, J.-P., Denuit, M. and Guillén, M., **Modeling of Insurance Claim Count with Hurdle Distribution for Panel Data**, *soumis pour publication dans un livre suivant l'International Conference on Mathematical and Statistical Modeling in Honor of Enrique Castillo, June 2006*,
2. Boucher, J.-P., Denuit, M. et Guillén, M., **Independent and Correlated Random Effects for Hurdle Models Applied to Panel Data Count**, *soumis à Econometrics Journal*,

3. Boucher, J.-P., Denuit, M. et Guillén, M., **Zero-Inflated Poisson Models for Panel Data**, *soumis à Journal of Applied Statistics*,

Résumé :

Parce que les modèles gonflés à zéro et les modèles à barrière étaient très intéressants intuitivement et parce que leur ajustement pour données avec indépendance temporelle était particulièrement bon, il semblait évident qu'une analyse de ce modèle pour données de panel devenait intéressante. Parmi les modèles avec dépendance temporelle, le modèle à effets aléatoires était celui qui offrait les résultats les plus prometteurs. Ainsi, dans cette partie de la thèse, nous avons pensé associer ce modèle pour données de panel aux modèles gonflés à zéro et à barrière, et ainsi créer diverses généralisations des modèles pour données de panel.

Pour le modèle à barrière, la distribution suppose que deux processus indépendants déterminent le nombre total de réclamations. Un premier processus indique si l'assuré a réclamé au moins une fois, alors que le second processus détermine le nombre total de réclamations, conditionnellement au fait que l'assuré ait réclamé à son assureur. Des effets aléatoires sont donc ajoutés sur chaque processus de cette distribution, alors que diverses copulas sont sélectionnées afin de modéliser la dépendance entre les effets aléatoires. Nous montrons que la distribution prédictive du nombre de réclamations ne dépend pas seulement du nombre de réclamations passées, mais aussi du nombre de périodes dans lesquelles l'assuré a réclamé. La distribution prédictive ne pouvant pas s'exprimer de manière close pour toutes les distributions jointes des effets aléatoires, des simulations Monte Carlo par chaînes de Markov sont utilisées pour calculer les primes prédictives.

A l'opposé du modèle à barrière, le modèle gonflé à zéro peut n'utiliser qu'un seul effet aléatoire. De plus, la distribution prédictive peut s'exprimer directement, ce qui facilite la comparaison entre modèles et le calcul des diverses statistiques. A l'instar du modèle à barrière, nous montrons que le nombre de réclamations passées n'est pas une statistique exhaustive pour la distribution prédictive puisqu'elle dépend aussi du nombre de périodes dans lesquelles l'assuré a réclamé au moins une fois.

Crédibilité et analyse a posteriori

Publications :

1. Boucher, J.-P. et Denuit, M., **Credibility Premiums for the Zero-Inflated Poisson Model and New Hunger for Bonus Interpretation**, *soumis à Insurance : Mathematics*

and Economics,

2. Boucher, J.-P. et Denuit, M., **Crédibilité linéaire multivariée utilisant le nombre de périodes avec réclamations : modèles de Poisson, modèles à barrière et modèles gonflés à zéro**, *soumis à Assurances et gestion des risques*.

Résumé :

Profitant du fait que le modèle gonflé à zéro avec effets aléatoires gamma peut s'exprimer de manière fermée, diverses primes de crédibilité, utilisant une fonction de perte quadratique ou une fonction de perte exponentielle sont étudiées. Un point de vue totalement différent est aussi étudié, où de la distribution du nombre de réclamation est tirée la distribution du nombre de sinistres. L'analyse sous l'angle de la *soif du bonus* est donc possible.

Afin d'obtenir des primes prédictives plus facilent calculables pour certains modèles à barrière et modèles gonflés à zéro, nous utilisons la théorie de la crédibilité multivariée, basée sur une erreur quadratique. Nous comparons ensuite le résultat de ces primes de crédibilité au vraies résultats obtenus par simulations. A l'aide d'un exemple numérique, nous montrons que la structure de dépendance supposée par diverses distributions jointes des effets aléatoires n'est pas bien modélisée par une approximation linéaire. Néanmoins, dans le cas où un seul effet aléatoire est utilisé ou dans la situation où les effets aléatoires sont supposés indépendants, nous réussissons à montrer qu'un modèle de crédibilité linéaire spécifique génère une approximation satisfaisante.

Table des matières