PROMOTION 2011

Mémoire présenté devant

INSTITUT DE STATISTIQUES DE L'UNIVERSITE DE PARIS

23, Avenue d'Italie – 75013 Paris

Pour l'obtention du

Diplôme de Statisticien

Mention Actuariat

Assurance ■ ***Finance*** □

Par Mlle Isa ENNADIFI

Sujet : Provisionnement d'un portefeuille de RC médicale anglais (MDU)

Lieu du stage : SCOR UK (Londres)

Responsables du stage : Damien Pujade-Lauraine

Encadrant I.S.U.P : Olivier Lopez

CONFIDENTIEL

Résumé

Dans le cadre de ses activités de réassurance Scor UK souscrit auprès de la Medical Defence Union LTD (MDU) un contrat couvrant les médecins adhérents contre les négligences se produisant durant la période de couverture du contrat. Dans le cadre de l'évaluation des provisions techniques d'un portefeuille de Responsabilité Civile (RC) médicale, des méthodes classiques telles que Chain Ladder, Bornhuetter Ferguson, ou la méthode fréquence/coûts moyens sont souvent utilisées. Les méthodes utilisées pour une grande part agrègent l'ensemble des données disponibles. L'objectif de cette étude est d'établir des méthodes actuarielles alternatives permettant de tirer partie des données sinistre par sinistre, ainsi que des variables qualitatives disponibles.

Ce mémoire propose trois modèles différents de régression pour déterminer la charge à l'ultime. Les deux premiers utilisent l'information dossier/dossier. Deux techniques sont alors proposées pour étudier ces données : la régression multiple linéaire et généralisée dans un premier temps puis la théorie des modèles de durée et de censure par le biais du modèle de Cox appliqué sur la charge ultime dans un deuxième temps. Enfin, le troisième modèle utilise également la régression linéaire généralisée (GLM) mais sur des données agrégées.

Le premier modèle utilisant la régression linéaire pour les sinistres attritionnels et l'ajustement de la loi de Pareto pour les sinistres graves est le plus pertinent parmi ceux étudiés. En effet, il utilise l'historique dossier/dossier et permet de séparer la modélisation des sinistres selon leur gravité. En outre, il permet une meilleure modélisation des sinistres graves.

L'utilisation de la censure dans le deuxième modèle fournit une estimation plus prudente et contribue à l'utilisation des sinistres ouverts pour paramétrer le modèle. Cependant, la mise en pratique du modèle de Cox n'autorise pas la décomposition du traitement des sinistres selon leur gravité.

Le recours aux GLM sur les triangles de liquidation est un compromis entre le modèle de Chain Ladder et les deux premiers modèles. En effet, ils permettent un traitement des données plus fin que l'utilisation des coefficients multiplicatifs de Chain Ladder, mais reste au niveau de la mise en œuvre plus facilement praticable que les deux premiers modèles.

Abstract

SCOR UK provides the Professional Indemnity coverage in respect of medical negligence only, to the Medical Defence Union LTD (MDU). For this kind of coverage, the usual analytical approaches are Chain Ladder, Bornhuetter Ferguson, or frequency/severity methods. For most of these, the available data is aggregated. The purpose of this study is to establish some alternative actuarial methods which take advantage of additional information through claim-by-claim data, as well as qualitative variables.

This thesis provides three different models based on regression in order to determine the ultimate losses. The first two methods use the information claim-by-claim whereas the third one uses aggregated information in triangle. The first model uses classical linear regression and also Generalized Linear Model (GLM) theory, whereas the second one uses artificial data point theory and censored data through the Cox model. The last one uses also GLM but applied this time on the aggregated triangle and not claim-by-claim data.

The first model using linear regression for small claims and fitting a Pareto distribution for large claims is the more relevant among all the models studied. Indeed, it uses claim-by-claim data, permitting a split between claims according their severity. And moreover it is more better at modelling large claims.

The use of censored data in the second model provides also a more careful estimation of the ultimate amount than the Chain Ladder, and takes advantage of the information from open claims. Unfortunately, the practicalities of the Cox model don't allow a split according to severity.

The use of GLM on triangles is a compromise between Chain Ladder and the first two models. The treatment of the data and the pattern is more precise than the multiplicative coefficients of Chain Ladder, and it is easier to use than the other models.

Remerciements

En préambule de ce mémoire, je souhaite adresser mes remerciements à toutes les personnes qui m'ont apporté leur aide et soutien durant son élaboration.

Je tiens tout d'abord à remercier les deux personnes grâce auxquelles j'ai pu mener ce projet à bien : Damien PUJADE-LAURAINE, manager et tuteur pour son encadrement, sa confiance, ainsi que Olivier LOPEZ, professeur à l'ISUP, pour ses conseils précieux, et sa grande disponibilité.

Je souhaite remercier l'ensemble de mes collègues de SCOR UK pour leur accueil et leur gentillesse, et plus particulièrement, Pape TOP pour son aide et soutien réguliers.

Je souhaite également remercier chaleureusement ma famille et mes amis pour leurs aides, conseils, et lectures.

Sommaire

Introduction

I. *Objectif de l'étude*

Dans le cadre de ses activités de réassurance, Scor UK souscrit auprès de la Medical Defence Union LTD (MDU) un contrat couvrant les médecins adhérents contre les négligences se produisant durant la période de couverture du contrat. Il s'agit d'un contrat d'assurance directe car MDU n'a pas d'agrément d'assurance pour couvrir ces sinistres. Selon les années de souscription, la Scor a entre 50% et 100% du portefeuille.

Ce portefeuille de Responsabilité Civile (RC) médicale présente les difficultés d'un portefeuille de RC classique, à savoir un temps de développement long et une dispersion des montants, ainsi qu'une hétérogénéité des comportements. A cela, s'ajoutent les spécificités du domaine médical. En effet, l'incertitude sur les montants des sinistres est renforcée par le temps de recherche de la responsabilité du ou des médecins. En particulier, plusieurs années peuvent être nécessaires pour déterminer s'il y a eu une négligence à l'origine de l'état d'un patient, notamment dans le cas où une plainte est déposée devant le tribunal pénal. Cela est renforcé par la loi anglaise qui n'impose aucune limite de temps pour des dommages subis au niveau neurologique ou lorsque la victime est mineure.

Par ailleurs, prévoir l'état d'une victime dans dix ans ou vingt ans est extrêmement complexe, l'estimation du coût du sinistre peut donc varier des années après l'incident. De surcroît, MDU, de par sa nature, effectue peu de sélection des risques, ce qui ne permet pas d'éliminer les comportements extrêmes. Par conséquent, la durée de développement des sinistres est rallongée de manière significative ; une plus grande volatilité des coûts est alors constatée. Cela amène MDU à une politique très prudente de clôture des sinistres (sinistres zéro-inflatés) et une ouverture de dossier à chaque plainte recensée (sinistres de types « Precautionnary »). Etant donnés leurs nombres, ces deux types de sinistres ont tendance à perturber l'analyse.

L'inflation est également une problématique sous-jacente de ce type de portefeuille, en raison de la longueur de développement des sinistres. Son impact est important car l'inflation est estimée à 9% sur ce portefeuille. De plus, son évaluation peut s'avérer complexe car l'inflation médicale est volatile en fonction des cycles économiques et de la sur-volatilité des salaires médicaux.

Dans le cadre de l'évaluation des provisions techniques, la SCOR utilise des méthodes classiques telles que Chain Ladder, Bornhuetter Ferguson ou la méthode fréquence/coûts moyens. Même si l'ensemble des sinistres est géré par MDU, la Scor a accès à l'intégralité de la base de données des

sinistres. Nous pouvons ainsi, sur ce portefeuille, avoir une base de données par sinistre avec leur évolution dans le temps depuis 1977 ainsi que des informations qualitatives.

Les méthodes utilisées pour une grande part agrègent l'ensemble des données disponibles pour des raisons de simplification, ce qui implique une perte d'information. L'objectif de ce mémoire est donc d'utiliser des méthodes actuarielles alternatives permettant de tirer partie des données sinistre par sinistre et ainsi mieux appréhender et modéliser les différences de comportements de ce portefeuille Pour cela, l'exploitation approfondie des informations utiles pour le provisionnement de cette branche est une piste envisagée pour le provisionnement de cette branche.

Dans un premier temps, les modèles proposés ne traitant pas directement de l'inflation et des Incurred But Not yet Reporded (IBNyR), ils ont été traites à part. Ce mémoire propose également le traitement des sinistres zéros-inflatés par l'utilisation d'un modèle linaire généralisé binomial.

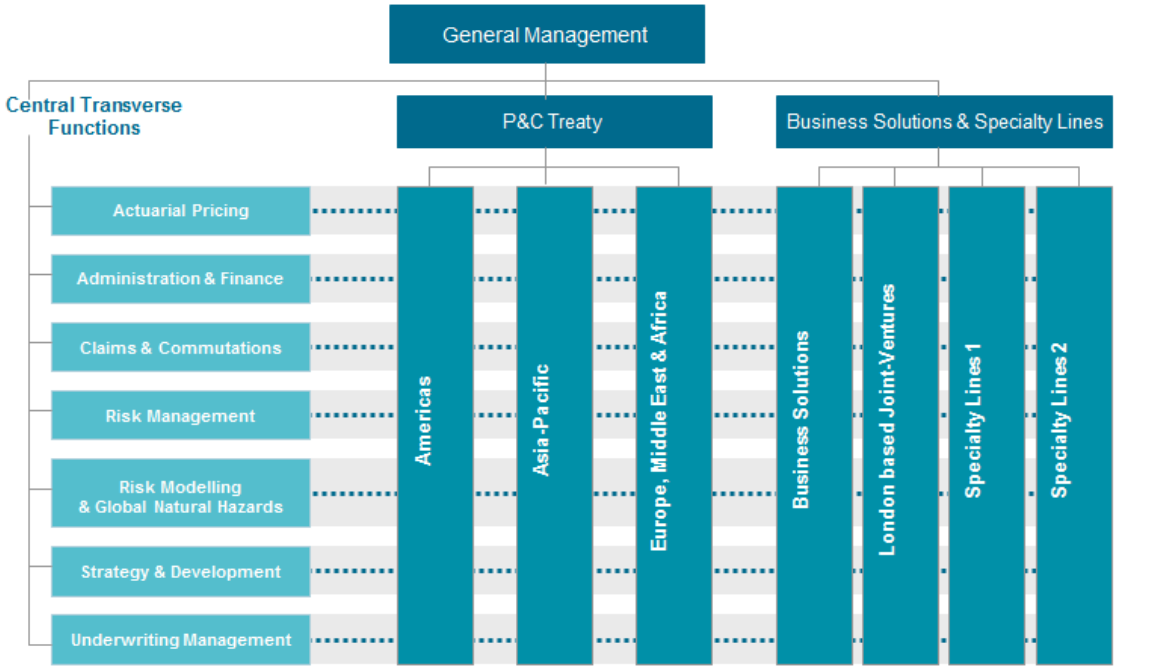
Par la suite, ce mémoire propose donc trois modèles différents pour déterminer la charge ultime. Le premier modèle utilise la régression linéaire multiple sur l'information dossier par dossier en décomposant le modèle selon la sévérité des montants afin de prendre en compte le phénomène de dispersion des montants, cependant il n'utilise pas l'information des sinistres ouverts. C'est pourquoi dans le deuxième modèle, nous cherchons une manière d'utiliser cette information supplémentaire, représentant des données censurées exploitables. La théorie des modèles de durée et de censure, nous permettent l'utilisation du modèle de Cox pour ce deuxième modèle. Ce dernier permet de calculer le montant de charge et de l'appliquer à l'historique sinistre par sinistre. Cependant l'application de ces deux modèles est beaucoup plus complexe qu'une méthode déterministe classique, le troisième modèle est donc la recherche d' un compromis intéressant entre, la mise en œuvre couteuse en temps des deux premiers modèles, et la précision de modélisation souhaitée. Il utilise la régression linéaire généralisée mais cette fois ci sur des données agrégées.

II. Contexte de l'étude

1. SCOR Global P&C – SCOR UK Ltd

Le groupe SCOR a été créé en 1970 à l'initiative des pouvoirs publics français. Le Groupe s'organise aujourd'hui autour de deux activités commerciales, SCOR Global P&C (réassurance Dommages) et SCOR Global Life (réassurance Vie), et d'une activité de gestion d'actifs, SCOR Global Investment. En 2008 et 2009, SCOR a mis en place une organisation structurée autour de pôles d'activités ou hubs : Paris, Zurich, Cologne et Londres pour l'Europe, Singapour pour l'Asie et New York pour les Amériques. Cette structure en hub possède deux atouts majeurs : cela permet une gestion des risques sur le marché local dans leur région géographique ainsi que l'utilisation de fonctions de souscription ou de gestion au niveau Groupe, ce qui convient parfaitement à l'activité internationale de réassurance. De plus, le groupe SCOR a depuis Juillet 2007 le statut de société européenne. Actuellement, SCOR est le 1er réassureur Français et le 5ème mondial.

Voici l'organisation de SCOR Global P&C auquel le hub de Londres appartient :



P&C Treaty: Proportional and Non proportional Property, Casualty, Motor and other non specialty lines of business
Business Solutions: Natural Resources, Industrial & Commercial Risks, Alternative Risk Financing, Cedant Facultative Services
London based Joint-Ventures: London Joint-Ventures (MDU, Lloyd's), Contingency
Specialty Lines 1: Agriculture, Engineering, Marine, Cat driven US mainland nationwide business
Specialty Lines 2: Aviation, Credit & Surety, Inherent Defects Insurance, Space, Structured Risk Transfer

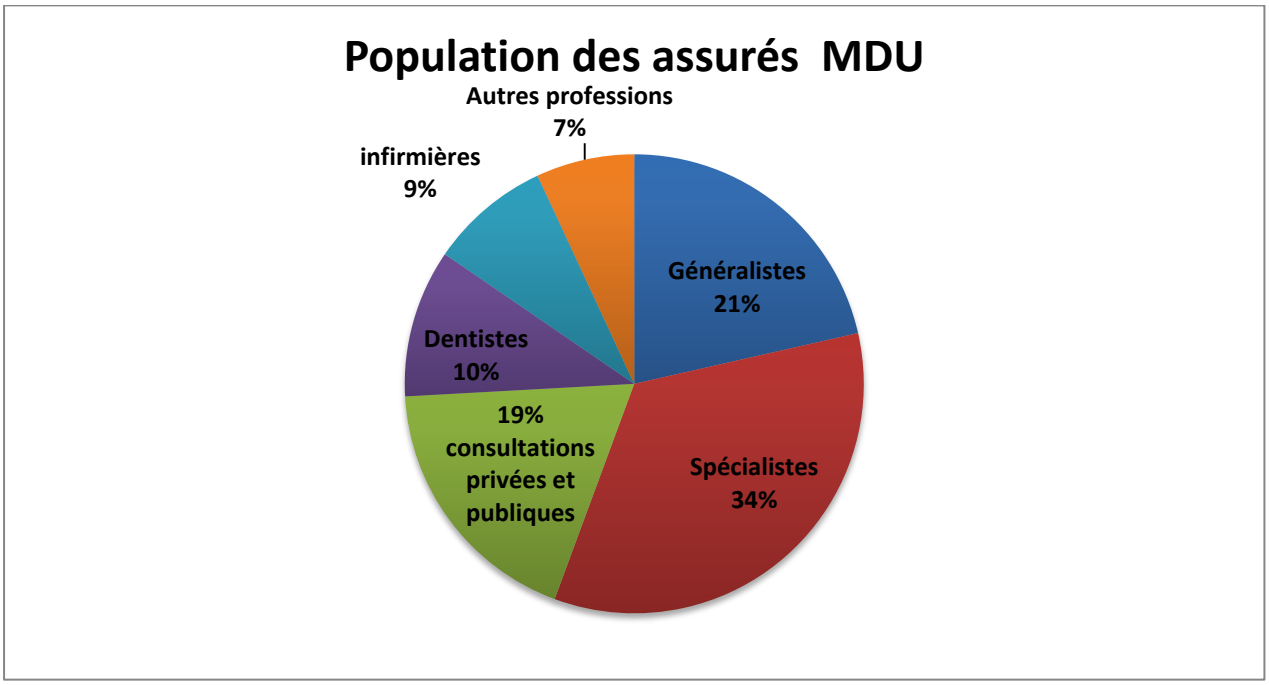
SCOR UK Ltd est une entité légale de SCOR SE, rattachée au hub de Londres et contrôlée par la FSA (Financial Services Authority, autorité de contrôle anglaise). SCOR UK Ltd a l'agrément aussi bien pour l'activité d'assurance que pour celle de réassurance.

A travers cette entité, SCOR SE a signé un accord pour une durée de 10 ans, à compter du 1er Avril 2008 avec Medical Defence Union LTD (MDU) afin de souscrire annuellement les contrats de RC médiale de ses membres.

2. Medical Defence Union LTD (MDU)

Historique

MDU est une organisation de protection médicale au Royaume-Uni comme MPS ou MDDUS. Elle a été créée en 1885. C'est une association à but non lucratif pour les professions médicales. La population des assurés aujourd'hui se compose majoritairement de médecins (généralistes et spécialistes) mais aussi de dentistes, infirmières et autres professionnels de santé. MDU est la plus ancienne mais également la plus importante des organisations de défense médicale présentes au Royaume-Uni. Voici ci-dessous le graphe de la répartition des membres par secteur :



Suite à l'augmentation de la sinistralité de Responsabilité Civile (RC) de ses membres et n'ayant pas l'agrément pour s'auto-assurer, MDU signa à partir de 2000 un partenariat avec Zurich Re. Zurich Re fut ensuite achetée en 2003 par Converium puis par SCOR SE en 2007.

Au Royaume-Uni, les professions médicales sont dans l'obligation d'adhérer à une organisation telle que MDU pour exercer. En revanche, les organisations de protection médicale n'ont aucune obligation d'accepter un membre. MDU est une organisation de médecins dont le but est de protéger leur réputation. Les médecins y adhèrent dès la fin de leurs études et ne présentent donc pas de

mauvais risque a priori, et quittent peu leur organisation. Le portefeuille de souscription est donc relativement stable dans le temps.

Le contrat d'assurance

La police d'assurance de SCOR couvre les membres contre toute négligence ou faute médicale seulement si le médecin est membre de MDU au moment de l'accident et également au moment de la notification de ce dernier.

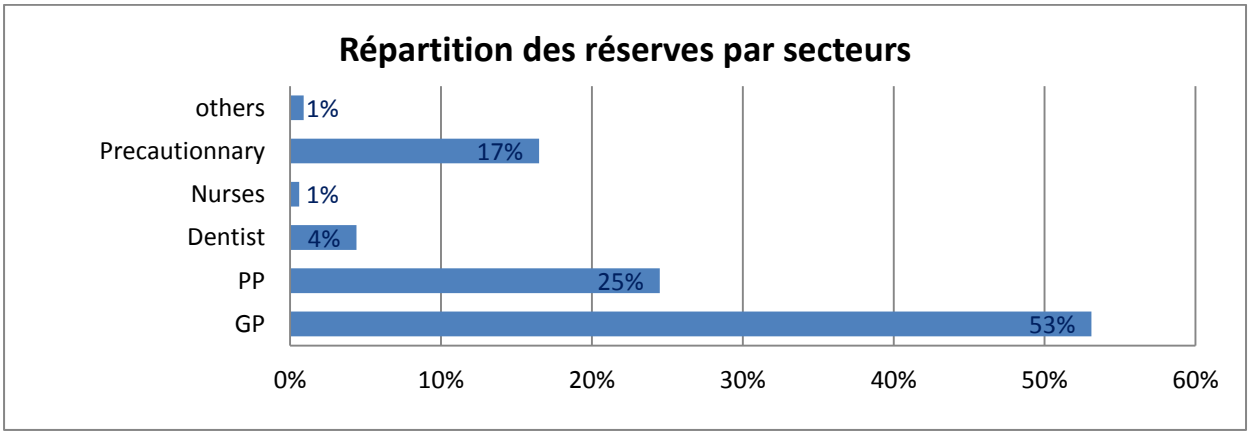
Le contrat exclut les accidents s'étant déroulés lorsque le médecin était un employé NHS (National Health Service), système de santé public. En effet en 1990, le ministère de la santé anglais a mis en place le plan « NHS Indemnity Scheme ». Ce plan a permis la protection des médecins anglais par le gouvernement à partir du 1er Janvier 1990. Depuis, les autorités de santé ont pour obligation de couvrir les médecins qu'elles emploient en cas de plainte. Un accord a par la suite été passé entre les organisations de défense médicale et le ministère de la santé anglais afin qu'il devienne rétroactif.

Pour autant, bon nombre de médecins anglais « consultants » continuent d'être membres auprès de MDU, car ils exercent à la fois dans le secteur privé et public. Ainsi, en conservant leur adhésion à MDU, ils sont couverts pour les actes exercés dans le domaine privé et peuvent ainsi profiter de toute l'aide juridique et légale dispensée par MDU.

Par ailleurs, SCOR couvre les sinistres jusqu'à hauteur de £10M par médecin et par sinistre.

Le provisionnement

Le provisionnement se fait par secteur car le risque est différent d'une catégorie à une autre. Les secteurs les plus importants de MDU sont les généralistes (General Practitionner, GP) et les spécialistes (Private Practitionner, PP). Ils représentent 78% des réserves à eux seuls.



Pour cette étude, il a été choisi d'étudier plus particulièrement la catégorie « PP ». Cette catégorie a l'avantage d'être représentative des deux difficultés principales, à savoir, les sinistres qui ont une grande incertitude sur leur durée de développement (i.e. la durée que prend le sinistre pour se clôturer) mais aussi une grande dispersion au niveau des montants. Cependant, les sinistres ont un comportement relativement plus homogène que les GP.

Partie 1 : Base de données

- Description de la base et problématiques

Description

MDU communique une base de données très complète à SCOR. L'inventaire du portefeuille réalisé au 31/12/2009 a été utilisé.

Elle est composée de l'historique des montants de paiements et réserves depuis 1977 jusqu'à aujourd'hui. Ces montants sont eux-mêmes ventilés entre les frais juridiques et les indemnités. En effet, les négligences médicales induisent la plupart du temps un procès afin de déterminer la responsabilité du médecin et les montants des indemnisations envers la victime. Les frais juridiques ayant une volatilité relativement faible et pouvant aisément être modélisés par des méthodes classiques, l'étude se concentre sur les indemnités. Elle possède également d'autres variables :

- l'année de notification (« yon »)
- l'année de l'accident (« yoi »)
- l'année de clôture (« yos »),
- l'état de la plainte: actif, procès gagné, prescription du sinistre,…(« Case status »)
- le pays de l'accident (« country »)
- le type de plainte : sinistre ouvert ou sinistre potentiel (cf « Precautionnary claim »)

Depuis peu, MDU essaye au niveau de la gestion des sinistres de développer l'analyse qualitative de chaque sinistre afin de pouvoir évaluer plus facilement le comportement futur de la plainte. Les gestionnaires de sinistres renseignent une nouvelle variable qualitative (« Probability ») qui indique dans quelle fourchette le sinistre devrait se situer lors de son règlement : entre 0 et £613k, entre £613k et £3M, ou supérieur à £3M. Cette variable n'a malheureusement pas été utilisée car l'historique nécessaire n'était pas encore disponible. Elle pourra cependant être utilisée dans le futur afin de raffiner les modèles utilisés dans ce mémoire.

La base des « PP » possède 17 297 sinistres référencés. Ces sinistres peuvent avoir des comportements extrêmement variés : ils peuvent être potentiels (i.e. un médecin déclare un incident lors d'une intervention sans qu'une plainte n'ait été encore notifiée à MDU), attritionnels, graves, clos ou non, les clos pouvant s'être conclus avec ou sans paiement.

La complexité est de pouvoir gérer distinctement variétés de situations, tout en ayant la masse de sinistres nécessaires pour obtenir des modèles suffisamment stables. Il s'agit donc de trouver le meilleur compromis possible.

L'état du sinistre

La première action menée a été de trier les données selon l'état des sinistres. En effet, pour prévoir la charge ultime des sinistres ouverts, il a été décidé de calibrer un modèle régressif sur les sinistres clos en raison de leur meilleure qualité prédictive. Il était donc important de définir le tri des sinistres, MDU détaillant le statut de la plainte. Deux difficultés principales ont été rencontrées lors de cet exercice.

La première est liée à la longueur de la branche RC médicale. En effet, il peut être difficile de déterminer si une maladie provient de la négligence d'un médecin lors d'une intervention ou d'un facteur naturel. Lors d'une intervention plusieurs médecins peuvent intervenir. Se pose alors la problématique de la responsabilité entre ces différents médecins. De plus, si la plainte se termine au pénal, la procédure judiciaire peut être longue.

Outre ces questions de responsabilités, il est extrêmement complexe de prévoir l'état d'une victime dans dix ans ou vingt ans. Une dégradation ultérieure de l'état de santé du patient peut être rattachée à une opération. La loi britannique pour des dommages subis au niveau neurologique ou par des enfants d'âge inférieur à 18 ans n'impose aucune limite de temps, tandis que pour les autres sinistres, la victime a trois ans pour déclarer le sinistre. MDU dispose donc d'une politique prudente de clôture des sinistres. Par conséquent, les années de clôture sont renseignées uniquement pour les sinistres clos avec paiement ; la date enregistrée reflète par conséquent le dernier paiement d'indemnité. Ainsi, pour certains sinistres, aucun paiement n'a été effectué et aucun mouvement n'a été observé depuis longtemps. Etant donné leurs nombres, ces sinistres ont tendance à perturber l'analyse.

La deuxième difficulté réside dans l'ouverture d'un dossier à chaque plainte recensée. MDU ouvre un dossier même si l'information n'est pas suffisante pour quantifier une réserve : si la responsabilité du médecin membre n'est pas encore avérée ou si une plainte n'a pas été déposée par la victime. Ce sont les plaintes potentielles : « Precautionnary claim ». Ainsi, 75% des sinistres présentent une charge totale nulle. Comme l'information est insuffisante pour avoir un montant de réserve, il existe trois montants nominaux :

- « 200 » représentant une blessure sur un mineur,

- « 300 » représentant une blessure au niveau du cerveau (adulte ou mineur). Ce sont des sinistres potentiels qui n'ont pas de limite de temps pour porter plainte.
- « 100 » représentant tout autre sinistre.

De plus, l'état d'une plainte peut faire un « va-et-vient » entre un sinistre potentiel ou non. De nouveaux éléments du dossier peuvent la convertir en sinistre, avec un montant de réserve non nominal. Ou bien, un sinistre peut redevenir une plainte potentielle à la suite de nouveaux éléments.

Une piste de modélisation des « Precautionnary » pourrait être les chaînes de Markov décrivant les passages suivants entre les deux états :

Sélection des années de déclaration

La base de données a un historique des sinistres déclarés depuis 1977. Cependant pour MDU, il n'est pas pertinent de garder l'ensemble des années de déclaration. En effet, les sinistres les plus anciens n'ont pas un comportement homogène en raison des changements de législations de jurisprudences, ou de prévention.

Les principaux événements concernant l'environnement de la RC médicale au Royaume-Uni sont les suivants :

En 1998, la Chambre des Lords sous l'impulsion de Lord Ogden a imposé un taux d'actualisation aux assureurs anglais pour les sinistres de personnes, afin de protéger les victimes contre le risque d'inflation. Le taux utilisé aujourd'hui est de 2,5%.

La même année, à la suite de nombreux cas de plaintes soulignant les limites du système médical, MDU a introduit de nouvelles initiatives de prévention. En parallèle, d'importants efforts de développement de techniques ou de procédures pour sécuriser les pratiques médicales ont été réalisés afin d'éviter les erreurs médicales. Cela a conduit à une réduction du nombre de plaintes depuis 1998.

En 1999, Lord Woolf a réformé le système juridique afin de rendre plus accessibles les procédures de litiges civils et moins coûteuses.

En 2003 a été mis en place le Court Act 2003, en application depuis 2005. Il a donné autorité à la justice, en cas de litige, de forcer l'assureur à indemniser la victime par une rente, appelée PPO (Periodic Payment Orders). En effet depuis la seconde guerre mondiale, les litiges sont indemnisés par un capital unique, la victime endossant la responsabilité de sa gestion. Et cela pour principalement deux raisons. La première est que les PPO sont nominatives si la victime décède, aucune réversion n'est donc possible pour la famille. La seconde est qu'un capital permet d'acquérir une maison et de l'emménager en fonction des handicaps. Cela n'est pas nécessairement réalisable avec les montants annuels de rentes. Depuis 2005, la SCOR étudie donc de près le risque d'augmentation du nombre de PPO car l'impact pourrait être important. Si jamais les PPO se généralisaient, spécialement après le cas Thompstone (voir ci-dessous), cela reviendrait à une augmentation du coût par sinistre.

Le nombre de PPO reste cependant limité pour la Scor à ce jour en raison de la limite par sinistre de £10m. En cas de PPO et en raison de l'indexation, le montant de £10m serait atteint et la victime ne serait pas couverte au-delà. Les juges privilégient ainsi les paiements par capitaux pour les cas où MDU est concernée. Un seul cas de PPO est recensé à ce jour intervenu en 2011.

En 2008, le cas Thompstone a fait jurisprudence. La justice a choisi un taux de revalorisation lié à l'indice ASHE (Annual Survey of Hours and Earnings, correspondant à l'inflation des salaires médicaux) plutôt que le RPI (Retail Price Index). Or, le taux ASHE sur les dernières années a été supérieur au RPI et est également plus volatile, ce qui a amené une augmentation des coûts.

Description de la base finale

1) Elimination de certains sinistres

Suite à ces considérations, certains sinistres ont été exclus de l'étude :

- les sinistres de type « Precautionnary », i.e. ceux qui ont été « Precautionnary claim » tout au long de leur développement.
- les sinistres n'ayant eu aucun paiement et une réserve nulle entre 1998 et 2009, même si certains sinistres pourraient rouvrir, l'historique montre que les montants en jeux sont faibles et ces sinistres perturbent les résultats des modèles utilisés dans cette étude.
- les sinistres qui ont une réserve négative, car ils représentent des situations où l'on a eu des remboursements supérieurs aux paiements. Ils correspondent à des cas biens particuliers et sont rares. Ils sont au nombre de treize ce qui est inférieur à 0.08% des cas.
- les sinistres notifiés avant 1998 : leurs comportements sont différents des années les plus récentes en raison des événements décrits précédemment (Lord Ogden, nouvelles initiatives).

Suite à ces retraitements de la base de données originale, la nouvelle base de données est composée de 4 072 sinistres parmi lesquels 2 101 ont une charge ultime strictement supérieure à zéro.

2) Détermination de l'année de clôture

Afin d'optimiser les résultats des modèles, une variable nouvellement créée représente l'année de clôture des sinistres sur le principe suivant :

L'année de clôture $\widehat{yos}$ est telle que $\widehat{yos} = \max\ (pyear, ryear)$

Où $pyear$ correspond à la dernière année pendant laquelle il y a eu un paiement et $ryear$ correspond à la dernière année où l'on constate des réserves.

3) Séparation entre sinistres larges et attritionnels

La forte volatilité des montants et des durées nous a amené à séparer les sinistres en deux catégories :

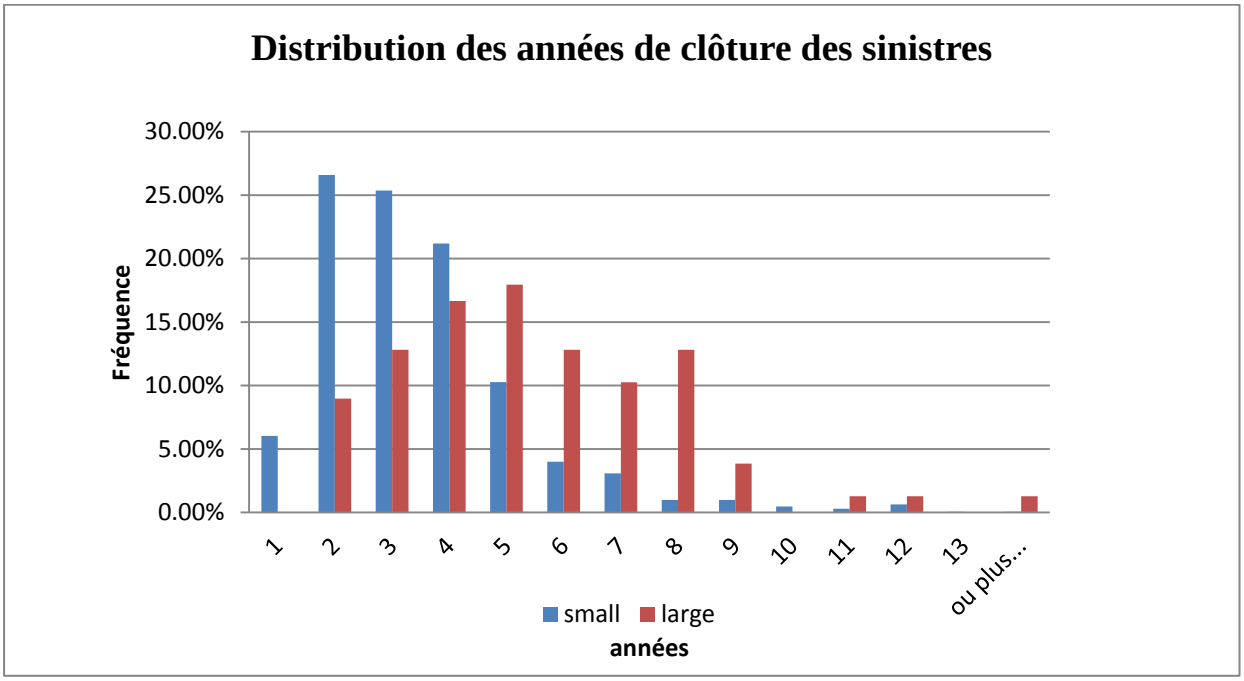
- « Small » correspondant aux sinistres attritionnels dont le risque de fréquence est prédominant.
- « Large » correspondant aux sinistres dont le risque d'intensité est prédominant.

Dans le cadre de la revue de ses réserves, SCOR a déterminé un seuil de £500k pour séparer les « Large » des « Small » en utilisant la théorie des valeurs extrêmes. Dans le cadre de cette étude, il a été décidé de conserver ce seuil afin d'avoir une cohérence avec les résultats déjà obtenus par ailleurs.

Voici les statistiques de base décrivant ces deux catégories ainsi que les graphiques des distributions du nombre d'années de développement ainsi que des montants disponibles fin 2009.

	Développement (années)		Charge fin 2009 (£)	
Type de sinistre	**Small**	**Large**	**Small**	**Large**
Moyenne	3.28	5.46	59,096	1,380,532
Médiane	3.00	5.00	32,000	859,750
Mode	2.00	5.00	900,000	900,000
Écart-type	1.85	2.41	81,344	1,226,418
Variance de l'échantillon	3.41	5.78	6,616,873,349	1,504,101,273,310
Minimum	1.00	2.00	37	505,000
Maximum	16.00	14.00	480,940	5,500,000
Somme	-	-	118,959,435	121,486,846
Nombre d'échantillons	2,013.00	88.00	2,013	88

On peut noter la différence de taille des deux échantillons. Les sinistres « Large » représentent 4% du nombre de sinistres mais plus de 50% de la charge. Voici la distribution de la charge entre les deux catégories :



Nous pouvons également remarquer que ces deux catégories se distinguent au niveau de leur développement. En effet, les sinistres « Small » ont tendance à se clôturer plus rapidement que les « Large ». Par exemple, 89% des sinistres « Small » sont clos en moins de cinq ans alors que les « Large » sont clos en huit ou neuf ans.

Nous allons donc, quand cela sera possible, modéliser chaque catégorie séparément. Cela permet d'obtenir des résultats plus pertinents à partir de comportements plus homogènes, particulièrement pour les sinistres de type « Small ».

Partie 2 : Modèles mathématiques utilisés

I. Méthode de provisionnement déterministe : Chain Ladder

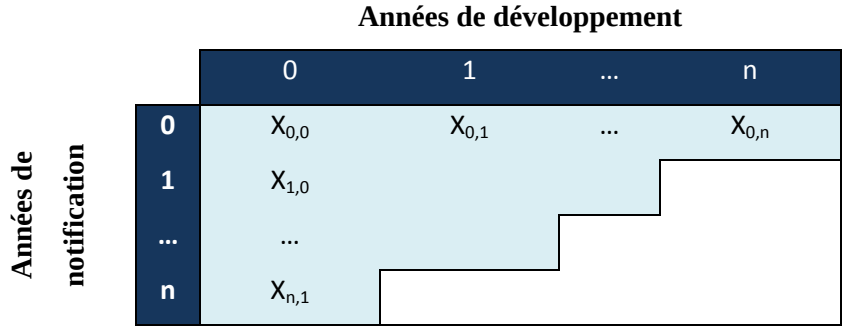
1. Généralités et notations

Il existe de nombreuses méthodes de provisionnement déterministes et stochastiques. La première méthode utilisée dans le cadre de cette étude est la méthode de Chain Ladder. Elle servira de benchmark afin de comparer nos résultats et de souligner les avantages et désavantages des autres méthodes. Chain Ladder est une méthode déterministe liquidative, ce qui signifie qu'elle est basée sur des triangles de liquidation, ainsi qu'un raisonnement utilisant des facteurs de développement.

Le principe du triangle de liquidation est simple. On agrège les données représentant la sinistralité passée selon trois axes de temps : les années de notification (ligne), les années de développement du sinistre (colonne) et la diagonale représentant les années calendaires. On peut utiliser le triangle pour des montants de charges ou bien de règlements et il peut également prendre en compte des incréments ou des montants cumulés.

Soit n l'horizon maximal défini pour les sinistres de ce portefeuille.
Soit i l'indice des années de notification pour $i \in [0, \dots, n]$
Soit j l'indice des années de développement pour $j \in [0, \dots, n]$
Soit $Y_{i,j}$ le montant des sinistres notifiés l'année i et payés au bout de j année. (Incréments)
Soit $C_{i,j}$ le montant cumulé des sinistres notifiés l'année i et payés depuis la première année de notification jusqu'à l'année j.
D'où la relation suivante : $C_{i,j} = \sum_{k=0}^{j} Y_{i,k}$
Nous noterons par la suite $X_{i,j} = \begin{cases} Y_{i,j} \\ C_{i,j} \end{cases}$ qui permet une écriture neutre s'adaptant selon les besoins à des incréments ou des paiements cumulés.

Voici une représentation d'un triangle de liquidation :



Cette méthode permet d'estimer la partie inférieure du triangle en supposant que la liquidation future sera similaire à la liquidation passée. Cela correspond aux paiements ou charges futures. La charge ultime s'obtient pour chaque année de notification, et par conséquent la charge ultime globale et les réserves par ailleurs.

Soient L_i la charge ultime des sinistres notifiés durant l'année i et L la charge ultime globale telles que :

$$L_i = \sum_{j=0}^{n} C_{i,j} \text{ et } L = \sum_{i=0}^{n} L_i$$

Le montant de réserves se déduit alors comme la différence de cette même charge ultime et les montants déjà payés.

2. *Théorie*

La méthode de Chain Ladder se décompose en deux étapes. En premier lieu, elle permet l'estimation des facteurs de développement que l'on utilise dans un second temps pour évaluer les montants futurs. Elle utilise les montants cumulés.

Pour cette étude, les triangles de charge ont été choisis car les comportements sont plus stables. Par ailleurs, étant donnée la longueur de la branche, de nombreux sinistres ne sont pas encore payés à la fin de la triangulation. Ainsi, la pertinence de la projection repose principalement sur le coefficient de queue dont l'évaluation peut être difficile. Les triangles de charge permettent de se reposer sur une évaluation préalable des gestionnaires de sinistres.

2.1. Définition

Le modèle estime les montants inconnus de la partie inférieure du triangle de liquidation $C_{i,j}$ pour $i + j > n$ d'après le modèle suivant :

$$\begin{cases} \exists\,(f_1,\dots,f_{n+1}) \in [1,\infty[^{n+1}\ tels\ que\ \ \forall i \in [0,\dots,n], \forall j \in [0,\dots,n],\quad E\left[C_{i,j+1}\right] = E\left[C_{i,j}\right] * f_{j+1} \\ \forall j \in [0,\dots,n], \sum_{i=0}^{n-j} C_{i,j-1} > 0 \end{cases}$$

Ce qui amène aux hypothèses sous-jacentes.

2.2. Hypothèses

(H1) Les années de notifications sont indépendantes entre elles.

(H2) Les années de développement sont des variables explicatives du comportement des sinistres futurs.

2.3. Estimation des paramètres

Cette méthode suppose que les ratios $\frac{C_{i,j+1}}{C_{i,j}}$ ne dépendent pas de l'année de notification i.

D'où $\forall\, j \in [0,\dots,n],\quad \frac{C_{0,j+1}}{C_{0,j}} = \frac{C_{1,j+1}}{C_{1,j}} = \cdots = \frac{C_{n-j-1,j+1}}{C_{n-j-1,j}}$.

Première étape :

Cela nous amène pour la méthode standard à l'estimation suivante des facteurs de développements $\widehat{f_j}$:

$$\forall\, j \in [1,\dots,n], \widehat{f_j} = \frac{\sum_{i=0}^{n-j} C_{i,j}}{\sum_{i=0}^{n-j-1} C_{i,j-1}}.$$

Deuxième étape :

Les estimateurs de la charge future $\widehat{C_{i,j}}$ sont définis comme l'estimateur de l'espérance $E\left[C_{i,j}\right]$ de l'état $C_{i,j}$, i.e.

$$\forall\, i \in [0,\dots,n], \forall j \in [n-i,\dots,n], \widehat{C_{i,j}} = C_{i,n-i} \prod_{k=n-i+1}^{j} \widehat{f_k}$$

Il est aisé d'en déduire les estimations de R_i le montant de réserve pour l'année d'origine i ainsi que R le montant global des réserves. En effet, ils sont définis par :

$$\widehat{R_i} = \widehat{C_{i,n}} - C_{i,n-i} \qquad \text{et} \qquad \hat{R} = \sum_{i=0}^{n} \widehat{R_i}$$

2.4. Vérification des hypothèses

Une manière de vérifier que la méthode de Chain Ladder est pertinente est de valider l'hypothèse d'indépendance des années de notification. On peut utiliser le triangle des facteurs individuels $f_{i,j}$ définis par :

$$\forall\, i + j \leq n - 1, \quad f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}$$

L'hypothèse d'indépendance est considérée acceptable si pour tout $j \in [0, \dots, n-2]$, les éléments de la j^{ème} colonne de ce même triangle sont relativement constants. Si tel n'est pas le cas, il sera nécessaire de gérer les facteurs atypiques qui peuvent correspondre à des changements de législation ou de politique de souscription. Les coefficients de ces années de survenance peuvent être soit éliminés soit lissés afin de retrouver la stabilité nécessaire à l'application de la méthode Chain Ladder.

II. Modèle de régression multiple

1. Théorie

1.1. Notations et généralités

Le premier objectif est de dégager les facteurs les plus importants afin d'obtenir une modélisation simplifiée de la réalité. En d'autres termes, on cherche à expliquer une variable endogène Y par des variables exogènes X_1, X_2,…,X_p liées par une relation linéaire. Par la suite, ce modèle explicatif va permettre d'obtenir une prévision de notre variable Y à partir d'un échantillon de n observations.

Par la suite, nous allons utiliser une écriture matricielle afin d'alléger les écritures et faciliter l'expression de certains résultats.

Soit Y la variable à expliquer, un vecteur de dimension n : $Y = (Y_1, \dots, Y_n)'$ et $\hat{Y}$ son estimation.

Soit X la matrice des vecteurs explicatifs de dimension (n,p+1) : $X = \left(1, X_1, \dots, X_p\right)$

1 représente le vecteur unité de dimension n. On appelle ce vecteur l'intercept.

Soit A le vecteur des coefficients de taille p+1 : $A = \left(a_0, \dots, a_p\right)'$ et $\hat{A}$ son estimation.

Soit ε le vecteur des résidus de taille n : $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$.

Soit $\hat{\varepsilon}$ le résidu empirique tel que $\hat{\varepsilon} = Y - \hat{Y}$

1.2. Définition

La définition du modèle matriciel est donc :

$$\begin{cases} Y = XA + \varepsilon \\ rg\ X = p + 1 \\ E(\varepsilon) = 0\ , V(\varepsilon) = \sigma^2 I_n \end{cases}$$

Où I_n est la matrice identité de taille n x n

1.3. Hypothèses

(H1) Les observations Y_i sont indépendantes.

(H2) Les X_i sont déterministes mais en réalité on peut supposer ces X_i aléatoires et travailler conditionnellement à leur réalisation.

(H3) Les résidus sont indépendants et identiquement distribués.

(H4) Il y a absence de colinéarité entre les variables explicatives, i.e. la matrice (X'X) est régulière donc de plein rang, ce qui signifie :

$$\begin{cases} \det(X'X)\ existe \\ \det(X'X) \neq 0 \end{cases}$$

(H5) Les erreurs sont linéairement indépendantes des variables exogènes, i.e. :

$$\forall i \neq j, cov(\varepsilon_i, X_j) = 0$$

1.4. Estimation des paramètres

Pour obtenir l'estimation des coefficients du vecteur A, on utilise la méthode des moindres carrés ordinaires (MCO) qui revient à un problème d'optimisation. Il faut alors minimiser la somme des carrés des erreurs, i.e. :

$$\min \varepsilon'\varepsilon = \min\ (Y - XA)'(Y - XA) = \min f(A)$$

Comme $(X'X)$ est inversible d'après les hypothèses, on obtient $\hat{Y}$ comme estimateur de Y tel que

$$\hat{Y} = (X'X)^{-1}X'Y$$

Dans le cas où l'hypothèse de normalité des résidus est validée, la méthode du maximum de vraisemblance peut s'appliquer (cf. GLM). En effet, l'estimateur du maximum de vraisemblance fournit des résultats identiques à ceux de l'estimateur des MCO dans ce cas précis.

1.5. Propriétés

D'après le théorème de Gauss-Markov, cet estimateur est appelé BLUE (Best Linear Unbiased Estimator), car parmi les estimateurs sans biais, il fournit les variances les plus faibles.

$$\begin{cases} E(\hat{Y}) = Y \\ \sigma_{\hat{Y}}^2 = \sigma_\varepsilon^2 (X'X)^{-1} \end{cases}$$

En outre, si les résidus sont gaussiens, alors l'estimateur des MCO est efficace, i.e. que l'estimateur a la plus petite variance possible.

Après un calcul matriciel, il apparait que l'estimateur $\widehat{\sigma_\varepsilon^2}$ sans biais de σ_ε^2 est défini tel que :

$$\widehat{\sigma_\varepsilon^2} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n - (p+1)}$$

En remplaçant la variance de l'erreur par son estimateur, nous obtenons :

$$\sigma_{\hat{Y}}^2 = \widehat{\sigma_\varepsilon^2} (X'X)^{-1}$$

2. Sélection du modèle optimal

Dans notre base de données, nous possédons de nombreuses variables explicatives. L'objectif est donc de sélectionner les variables et les régresseurs les plus pertinents afin d'obtenir le modèle le mieux ajusté possible.

Se distinguent alors les modèles explicatifs des modèles prédictifs. Si le nombre de variables explicatives est trop élevé et le phénomène trop complexe, il devient difficile de déterminer un modèle explicatif. La base de données disponible ayant trop de variables explicatives potentielles, un travail de tri a donc été indispensable. Il a été effectué en 2 étapes.

2.1. Premier tri

- ✓ Test de cohérence

Des variables peuvent être supprimées en discernant celles dont nous savons au préalable qu'elles ne seront pas utiles pour le modèle. C'est le cas par exemple du numéro de membre des médecins concernés par la plainte, qui n'a aucun impact sur la charge ultime.

- ✓ Corrélation partielle

Le coefficient de corrélation partielle mesure la liaison entre deux variables lorsque l'influence d'une ou plusieurs variables explicatives sont retirées. Plus le coefficient de corrélation partielle d'une variable est élevé, plus la contribution de cette variable est importante à l'explication globale du modèle. Il faut choisir les variables les plus fortement corrélées à notre variable d'intérêt et le moins possible entre elles afin de minimiser les conséquences de la colinéarité. Dans notre cas, les variables qui avaient un coefficient de corrélation partielle supérieur à 0,5 avec la variable à expliquer ont été choisies. En effet, en dehors des montants qui étaient très fortement corrélés, les autres variables explicatives ne possédaient pas de coefficient proche de 1. Le seuil de 0,5 s'est donc imposé de lui-même.

- ✓ Test de multi-colinéarité : Farrar et Glauber

Pour déterminer s'il existe de la multi-colinéarité entre les variables explicatives choisies, le test de Farrar et Glauber a été utilisé. Le terme de multi-colinéarité est employé lorsque les séries explicatives sont liées entre elles, i.e. elles possèdent une covariance non nulle entre elles. Si, pour des études théoriques nous pouvons supposer l'orthogonalité entre deux séries, dans la pratique les séries explicatives sont toujours plus ou moins liées entre elles.

Cette multi-colinéarité a pour effets : l'augmentation de la variance estimée de certains coefficients, l'instabilité des estimateurs des coefficients des MCO (de faibles fluctuations concernant les données entraînent de fortes variations des valeurs estimées). Et dans le pire des cas, la matrice X'X est singulière donc il devient impossible d'obtenir des estimations.

Le test de Farrar et Glauber (1967) consiste à calculer le déterminant D de la matrice des coefficients de corrélation entre les variables explicatives. Lorsque la valeur du déterminant D tend vers zéro, le risque de multi-colinéarité est important.

Le test consiste à tester :

$$\begin{cases} H_0 : D = 1 \\ H_1 : D < 1 \end{cases}$$

La variable de décision utilisée est $S = -\left[n - 1 - \frac{1}{6}(2(p+1)+5)\right]\ln(D)$ où n est la taille de l'échantillon et p le nombre de variables explicatives.

Sous H_0, $S \sim \chi^2(\frac{1}{2}p(p+1))$.

La région critique est $[S \geq \chi_{5\%}]$ où $\chi_{5\%}$ est le quantile de la loi du Chi-deux correspondant au seuil de 5%. Dans ce cas, il y a présomption de multi-colinéarité.

2.2. Deuxième tri

- ✓ R^2 et R^2 ajusté

A partir de l'équation fondamentale d'analyse de la variance :

$$(Y - \bar{Y})'(Y - \bar{Y}) = \left(\hat{Y} - \bar{\hat{Y}}\right)'\left(\hat{Y} - \bar{\hat{Y}}\right) + \varepsilon'\varepsilon$$

Elle permet de décomposer la variance totale (SCT) comme la somme de la variance expliquée (SCE) et de la variance des résidus (SCR).

$$SCT = SCE + SCR$$

Cette équation nous permet donc de juger la qualité de l'ajustement du modèle. Cependant, ces valeurs dépendent des unités de mesure. C'est pourquoi le coefficient de détermination R^2 qui est un nombre sans dimension a été utilisé. Il appartient à l'intervalle $[0,1]$, définit comme ceci :

$$R^2 = \frac{SCE}{SCT}$$

Plus ce coefficient est proche de 1, meilleur est l'ajustement. Le problème est que R^2 est une fonction croissante du nombre de régresseurs. Il permet uniquement de comparer deux modèles qui ont le même nombre de variables explicatives.

Le R^2 ajusté est une variante du R^2 classique qui tient compte de la dimension du modèle. Il est défini par :

$$R_a^2 = 1 - \frac{n-1}{n-p-1}\,(1 - R^2)$$

On a donc généralement $R_a^2 < R^2$, cependant sur un grand échantillon, ils sont équivalents.

Le logiciel R nous fournit les deux.

On peut alors choisir le modèle qui maximise R_a^2.

- ✓ Akaike Informative Criterion (AIC)

AIC utilise la vraisemblance. La vraisemblance est d'autant plus grande, que la log-vraisemblance est importante et l'ajustement du modèle est alors de meilleure qualité. Ce critère est largement utilisé. Il permet d'arbitrer entre deux modèles qui ont des degrés de liberté différents ainsi que des modèles non emboités.

Soit L la fonction de log-vraisemblance d'un modèle.

$$L = \log\left\{\prod_{i=1}^{n} P\left(Y = y_i \middle| X_1 = x_1, \dots, X_p = x_p\right)\right\}$$

Si nos résidus sont gaussiens alors $Y \sim \mathcal{N}\left(X\hat{A}, \sigma^2 I_n\right)$

Le critère d'AIC est donc défini par :

$$AIC = -2L + 2p$$

On choisit le modèle qui minimise l'AIC.

- ✓ Significativité des coefficients

Introduisons la notation suivante : $S\left(X_1, \dots, X_p\right) = \sum_{i=1}^{n} (\widehat{Y}_i - \bar{Y})^2$ la somme des carrés expliquée quand on régresse Y par rapport aux variables $\left(X_1, \dots, X_p\right)$.

Une autre manière de mesurer la qualité de l'ajustement du modèle est de tester la significativité globale des coefficients par le test suivant:

$$\begin{cases} H_0 : \theta_1 = \theta_2 = \cdots = \theta_p = 0 \\ H_1 : \exists i \epsilon \{1, \dots, p\} \; tel \; que \; \theta_i \neq 0 \end{cases}$$

En effet, si H_0 n'est pas rejetée, cela signifie alors que Y ne dépend pas des variables $(X_1, \dots, X_p)$. La variable de décision utilisée est $F = \frac{\|\hat{Y} - \bar{Y}\|^2 / p}{\|\hat{\varepsilon}\|^2 / (n-p-1)} = \frac{S(X_1, \dots, X_p)/p}{\|\hat{\varepsilon}\|^2 / (n-p-1)}$ (où $\| \; \|^2$ correspond à la norme euclidienne au carré)

Sous H_0, $F \sim \mathcal{F}(p, n-p-1)$.

La région critique est $[F \geq f_{5\%}]$ où $f_{5\%}$ est le quantile de la loi de Fisher correspondant au seuil de 5%.

La significativité marginale de chaque variable i peut aussi être testée. Le test deviendra alors simplement :

$$\begin{cases} H_0 : \theta_i = 0 \\ H_1 : \theta_i \neq 0 \end{cases}$$

La variable de décision utilisée sera ${T_j}^2 = \frac{S(X_1, \dots, X_p) - S(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p)}{\|\hat{\varepsilon}\|^2 / (n-p-1)}$

Sous H_0, $T_j \sim \tau(n-p-1)$.

La région critique est $\left[\left|T_j\right| \geq t_{5\%}\right]$ où $t_{5\%}$ est le quantile de la loi de Student correspondant au seuil 5%.

Le logiciel R effectue ces tests et renseigne sur le résultat sous forme de la p-value p. La p-value est la probabilité de rejeter H_0 lorsque H_0 est vrai. Si p est faible, on rejette H_0 car il y a peu de chance que H_0 soit vraie.

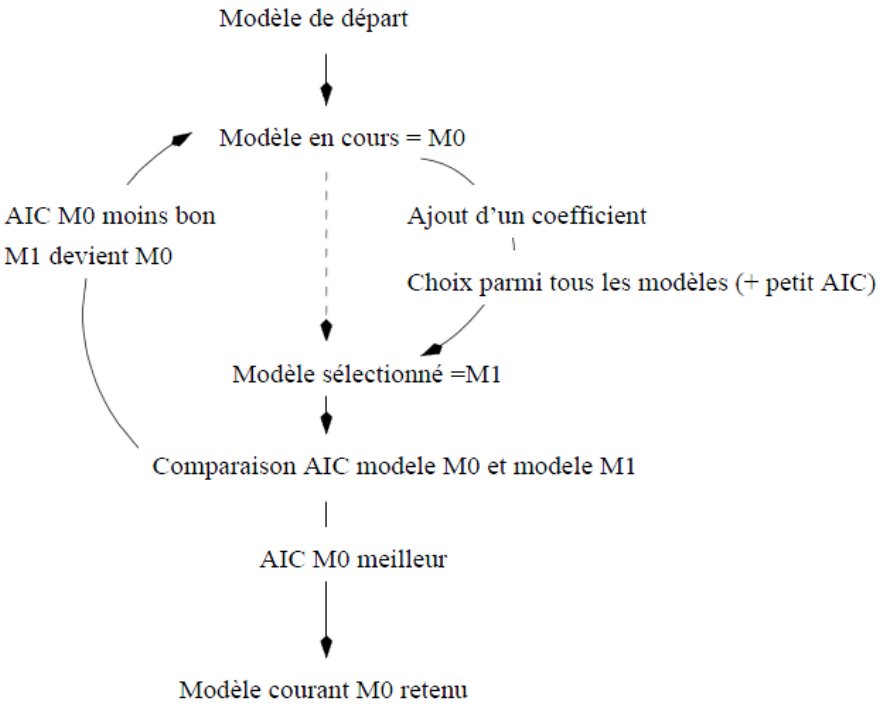
Dans notre cas, on a choisi le seuil de 5% donc le modèle ou la variable sont significatifs si $p \leq 5\%$.

- ✓ Méthodes de sélection

Les procédures décrites permettent de sélectionner un modèle à partir de familles de modèles donnés. En particulier, lorsque le nombre de variables explicatives est important, il devient nécessaire de disposer de méthodes de sélection automatique des variables. En effet, lorsque l'on a p variables explicatives, le nombre de sous-ensembles est 2^p. En raison du long temps de calcul, il est préférable d'utiliser des méthodes de recherche pas à pas.

Les méthodes les plus utilisées sont la régression pas à pas ascendante « forward selection », descendante « backward selection» et la méthode progressive « stepwise selection ».

Dans cette étude, la méthode progressive a été utilisée. Elle est identique à la méthode ascendante au détail près que l'on peut éliminer des variables déjà introduites. La méthode ascendante sélectionne dans une première étape la variable explicative dont le coefficient de corrélation simple est le plus élevé avec la variable Y. Par la suite, à chaque pas une variable est ajoutée au modèle utilisant le critère de minimisation de l'AIC. Le schéma décrit la méthode ascendante :



Finalement, pour obtenir le modèle optimal à la suite du deuxième tri, le processus a été le suivant :

1. Création d'une famille de modèles potentiels :
    a. Le premier modèle est le modèle saturé.
    b. Utilisation de la méthode progressive qui nous donne un deuxième modèle potentiel.
    c. Utilisation de la méthode progressive où l'on ajoute les variables d'interaction parmi les variables candidates permettant d'obtenir un deuxième modèle.
    d. Essais éventuels de modèles qui auraient plus de sens en terme actuariel.

2. Etude de la significativité des variables dans chaque modèle

    Si les variables ne sont pas toutes significatives, celles non significatives sont retirées pas à pas. Le critère permettant de déterminer la variable à supprimer est de retenir la variable ayant la plus grande p-value. On obtient une nouvelle famille de modèles potentiels.

3. On choisit donc celui qui a le plus grand R^2 ajusté et/ ou le plus petit AIC.

3. Validation du modèle

Une fois le modèle final choisi, une étude des résidus est faite afin de détecter des observations aberrantes et de valider les hypothèses du modèle linéaire.

L'étude des résidus bruts est envisageable, mais comme ils ne possèdent pas la même variance, leur comparaison s'avère difficile. Il a donc été choisi d'analyser les résidus de Pearson r définis par :

$$r_{=} \frac{Y - \hat{Y}}{\sqrt{V(Y)}}$$

Ces résidus doivent répondre à trois propriétés :

- Espérance nulle
- Homoscédasticité (constance de la variance de l'erreur)
- Adéquation des résidus à une loi normale

L’analyse des résidus se fait de manière graphique :

- Graphique prédiction linéaire/résidus

Ce graphe représente les résidus en fonction des prédictions. Il permet de tester les deux premières propriétés des résidus. Il ne doit pas apparaitre de structuration particulière, auquel cas il sera peut-être adéquat d’ajouter une nouvelle variable afin de prendre en compte cette structuration.

- Q-Q plot

Ce graphique permet de valider l’hypothèse de normalité des résidus. Le graphique représente les résidus en fonction de leur espérance théorique respective.

- Graphique des prévisions linéaires en fonction de la racine carrée des résidus standardisés. C’est le même principe que le premier graphique à une échelle différente.

- Graphique Résidus/poids

Ce graphique montre les résidus en fonction de leur poids. Il permet de remarquer les points influents et éventuellement de les enlever.

III. Modèle de régression linéaire généralisé : GLM

1. Théorie

1.1. Familles exponentielles

Le modèle linéaire généralisé est une généralisation du modèle linéaire que l'on vient de présenter. La variable aléatoire à expliquer Y se doit d'appartenir à la famille exponentielle.

On dit qu'une loi de probabilité appartient à la famille exponentielle si sa fonction de densité s'écrit de la manière suivante :

$$f_{\theta,\phi}(y) = exp\left\{\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right\}$$

Où

- $\theta \in \mathbb{R}$ est appelé paramètre naturel.
- $\phi \in \mathbb{R}$ est appelé paramètre de dispersion.
- $a, b\ et\ c$ sont des fonctions connues et dérivables, b est de plus trois fois dérivables et sa dérivée première b' est inversible.

Pour les lois de la famille exponentielle, les liens entre l'espérance, la variance et les paramètres θ et ϕ sont :

$$\begin{cases} E(Y) = b'(\theta) = m \\ Var(Y) = b"(\theta)a(\phi) = V(m) \end{cases}$$

Remarques :

- La moyenne est donc liée à la variance par cette fonction V.
- En général $a(\phi)$ est de type $a(\phi) = \frac{a\prime(\phi)}{d}$ où d est un poids a priori sur une observation. Cette notion de poids a priori est généralement inutile donc nous le choisissons égal à 1.
- La fonction b' étant par définition inversible, on obtient $\theta = b'^{-1}(m)$ et la fonction b'^{-1} est appelé lien canonique.

Voici un tableau récapitulant quelques exemples classiques :

Loi	Notation	Densité	ϕ	b(.)	c(.)	$E(Y) = m$	V(m)
Normale	$\mathcal{N}(m, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{(y-m)^2}{2\sigma^2}\right\}$	σ^2	$\frac{\theta^2}{2}$	$\frac{-\left(\frac{y^2}{\phi}+\ln(2\pi\phi)\right)}{2}$	θ	1
Poisson	$P(m)$	$P(Y=y) = \exp(-m)\frac{m^y}{y!}$	1	e^θ	$-\ln(y!)$	e^θ	m
Gamma	$\Gamma(m,s)$	$\frac{\left(\frac{s}{m}\right)^s y^{s-1} exp\left(-\frac{s}{m}y\right)}{\Gamma(s)}$	$\frac{1}{s}$	$-\ln(-\theta)$	$s\ln(sy) - \ln(y) - \ln(\Gamma(s))$	$-\frac{1}{\theta}$	m^2
Binomiale	$\frac{\beta(n,m)}{n}$	$P(Y=y) = \binom{n}{ny} m^{ny}(1-m)^{n-ny}$	$\frac{1}{n}$	$ln(1+e^\theta)$	$\ln\left(\binom{n}{ny}\right)$	$\frac{e^\theta}{1+e^\theta}$	$m(1-m)$

1.2. Définitions

Le modèle GLM est un modèle qui tente de relier des variables explicatives $X_1, \dots, X_p$ à une variable Y à expliquer. Il est défini selon trois composantes :

- La composante aléatoire

La composante aléatoire est la variable Y à expliquer. Sa distribution doit appartenir à la famille exponentielle.

- La composante systématique

La composante systématique représente la partie explicative du modèle.

Soit $x = (x_1, \dots, x_p)$ l'observation correspondant aux variables $X_1, \dots, X_p$ explicatives du modèle. On construit une combinaison linéaire des X_i, i.e. un modèle de régression.

$$\eta(x) = \sum_{i=1}^{p} x_i \beta_i$$

Où la variable η peut être vue comme une variable synthétique, un résumé linéaire des variables explicatives.

- La fonction de lien

L'espérance E(Y) dépend de $\eta(X)$ au travers d'une fonction g appelée fonction de lien. Cette fonction est inversible.

$$g\big(E(Y)\big) = g(m) = \eta(X)$$

Un choix particulier qui simplifie les calculs est le choix de la fonction de lien canonique :

$$g \equiv b'^{-1}$$

Parmi les fonctions de lien canonique les plus couramment utilisées, on trouve les fonctions :

Nom du lien	Fonction du lien	Loi
Lien identité	$g(m) = m$	Normale
Lien log	$g(m) = \ln(m)$	Poisson
Lien réciproque	$g(m) = -\frac{1}{m}$	Gamma
Lien logit	$g(m) = log\left(\frac{m}{1-m}\right)$	Binomiale

Le logiciel R a les lois usuelles implémentées. Il permet de sélectionner la fonction de lien ainsi que le paramètre de dispersion.

1.3. Méthodologie

Pour choisir un modèle GLM, il faut donc choisir :

- La loi de Y dans la famille exponentielle.
- Les variables explicatives pertinentes.
- La fonction de lien qui rattache notre variable Y aux variables explicatives préalablement choisies.

Une fois ces paramètres fixés, on peut estimer le paramètre de dispersion ϕ qui ne dépend pas des coefficients β_i. Dans notre cas, nous avons fixé ce paramètre à 1 donc cette étape est inutile.

Par la suite, il faut estimer les coefficients β_i. Cela permet d'obtenir la quantité $\eta(x)$, grâce à la fonction de lien on obtient $m = g^{-1}\big(\eta(x)\big)$ qui correspond à la moyenne, donc à la prévision du modèle.

Remarque : le modèle linéaire classique est juste un cas particulier de GLM, utilisant la distribution gaussienne et la fonction de lien identité.

1.4. Estimation des paramètres

L'estimation des coefficients β_i se fait par la méthode du maximum de vraisemblance. Il s'agit de trouver le paramètre $\hat{\beta}$ qui maximise la fonction l de log-vraisemblance. On observe pour les n sinistres les réalisations de la variable Y, ainsi que pour chacune des p variables.

Le modèle GLM est supposé choisi, donc la distribution de Y est connue et la fonction de lien également. Puisque la densité $f_{\theta,\phi}$ est connue, on a :

$$l((y,\theta,\phi) = \frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)$$

Lorsque ϕ n'est pas fixé, il est d'usage de le considérer fixé afin de l'estimer de manière séparée a posteriori. Le paramètre $\beta = (\beta_1, \dots, \beta_p)$ n'apparait pas directement dans l'équation de la log-vraisemblance mais de la manière suivante $\begin{cases} m = g^{-1}(\eta) = g^{-1}(X'\beta) \\ b'(\theta) = m \end{cases} \Leftrightarrow \theta = b'^{-1}(g^{-1}(X'\beta))$.

Une condition nécessaire pour calculer le maximum consiste à annuler la dérivée partielle du premier ordre par rapport à β. Cependant plutôt que de chercher à faire apparaître β, il est plus judicieux d'utiliser une décomposition. Voici les notations utilisées :

Soit l_i la fonction de log-vraisemblance pour la i^{ème} observation.
Soit m_i l'espérance de Y conditionnellement pour la i^{ème} observation.
Soit l_i la fonction de log-vraisemblance pour la i^{ème} observation.
Soit θ_i la valeur du paramètre θ au point X_i .
Soit η_i le prédicteur linéaire pour la i^{ème} observation.

$$\frac{\partial l_i}{\partial \beta_i} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial m_i}\frac{\partial m_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_i}$$

Finalement, l'estimation par maximum de vraisemblance nous amène à résoudre :

$$\frac{\partial l_i}{\partial \beta_i} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i - m_i)x_{ij}}{V(m_i)g'(m_i)} = 0$$

Il n'est pas possible de résoudre cette équation et de trouver une formule explicite pour β_i. Il faut donc appliquer une méthode numérique itérative pour obtenir ces estimations, comme par exemple le « Fisher Scoring ».

2. *Sélection du modèle optimal*

Avant l'optimisation, il faut tout d'abord déterminer un modèle, i.e. la densité de Y et la fonction de lien. Puis par la suite, optimiser ce modèle en partant de ces bases.

2.1. Choix de la densité de Y

- ✓ Le choix

L'un des inconvénients principaux de la mise en pratique est qu'il faut choisir parmi les densités disponibles, à savoir généralement les lois classiques présentées plus haut.

On peut choisir la densité en fonction du type de variable : si la variable est binaire, la densité sera de la forme binomiale, si la variable est un comptage ou une survie, la densité choisie sera la loi de Poisson ; enfin s'il s'agit d'une variable continue, le choix sera entre la loi Gamma et la loi Normale en fonction de l'existence de valeurs négatives et nulles.

Le choix peut être aussi en fonction de connaissances a priori. Ou encore, on peut choisir en fonction de la variance et de son adéquation à la fonction de variance. Cela revient à tracer $Var(y|X = x)$ en fonction de $m = g^{-1}(\eta)$. Le problème est que nous ne connaissons pas η. Il faudrait donc l'estimer, or elle dépend de la fonction de lien et de la loi que l'on cherche. Cette méthode qui serait sans doute la plus précise est très peu praticable.

- ✓ La validation

Une fois la densité sélectionnée, elle peut être confirmée ou infirmée à l'aide d'un test d'ajustement comme le test du Chi-deux ou Kolmogorov Smirnov.

Le test de Kolmogorov Smirnov ne convient pas dans cette étude, car il y a de nombreuses répétitions dans les données.

Voici le principe du test du chi-deux au seuil α:

On possède un échantillon d'observations de taille n que nous regroupons en k classes. Et soient μ la distribution empirique de l'échantillon et μ_0 la distribution théorique recherchée. Le vecteur $(F_1, \dots, F_k)$ est le vecteur des fréquences empiriques. Le vecteur $(p_1, \dots, p_k)$ est le vecteur des probabilités sous l'hypothèse H_0. Le problème à tester est :

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

La variable de décision utilisée est :

$$D = n \sum_{i=1}^{k} \frac{(F_i - p_i)^2}{p_i}$$

Sous H_0, la variable de décision suit la loi $D \sim \chi^2_{k-1}$.

La région critique est $[D \geq \chi_{5\%}]$ où $\chi_{5\%}$ est le quantile de la loi du Chi-deux correspondant au seuil de 5%.

2.2. Choix de la fonction de lien

Le choix de la fonction de lien se fait également en fonction du type de variables à expliquer. Si Y est un comptage borné ou binaire, alors seule la proportion est intéressante. Dans ce cas, le lien classique est le lien logistique, mais il est possible d'utiliser d'autres liens comme par exemple le lien probit ($g(m) = \Phi^{-1}(m)$ où ϕ est la fonction de répartition d'une loi normale centrée réduite) ou cloglog ($g(m) = \ln(-\ln(1-m))$). Si Y est une variable de comptage, alors le lien classique est le lien log. Par contre, si Y est continue, un choix classique est le lien canonique des lois concernées. Si l'on choisit le lien canonique d'une loi, alors cela revient à ne faire qu'un seul choix, celui de la loi.

En pratique, R propose plusieurs fonctions de lien pour chaque loi. Ce qui permet de choisir la fonction qui minimise le critère AIC.

2.3. Optimisation du modèle

Une fois, la loi de Y choisie et la fonction de lien, la procédure est réitérée à l'identique que pour le modèle linéaire classique au détail près que le critère du coefficient de détermination R^2 n'a pas de sens ici. Il faut alors chercher seulement à minimiser le critère AIC.

Pour la validation finale du modèle, les résidus de Pearson peuvent être analysés de façon identique à celle expliquée dans la régression linéaire multiple, afin de pouvoir valider les trois mêmes propriétés.

.

IV. Modèle de données censurées (semi-paramétrique) : Modèle de Cox

1. Notations et généralités

Il existe différents modèles de régression de données censurées. Le modèle de Cox (Cox 1972) a été appliqué dans notre étude. Ce modèle est souvent utilisé dans ce genre de cas. Notamment en raison de sa facilité de mise en œuvre et d'interprétation, mais également en raison de la présence de censure à droite.

Le principe des modèles de durée est de s'intéresser principalement à la fonction de survie, qui comme la fonction de répartition ou de densité, caractérise complètement la loi.

Initialement, dans les modèles de durées, comme son nom l'indique, on s'intéresse à une variable aléatoire X décrivant des durées. Bien que par la suite, nous verrons que ces modèles ont des applications pour d'autres problématiques. La variable aléatoire étudiée est donc dans $[0; \infty[$.

Soit $F(t)$ la fonction de répartition de X et $f(t)$ une fonction de densité de X.

Soit $S(t)$ la *fonction de survie* de X. Il s'agit d'une fonction décroissante telle que :

$$S(t) = 1 - F(t) = P(X > t)$$
$$\lim_{t \to \infty} S(t) = 0$$
$$S(0) = 1$$

La *fonction de survie conditionnelle* est, par définition, la survie d'un élément après une durée t sachant qu'il a déjà survécu jusqu'en u. Elle est définie par :

$$S(t|u) = P(X > t + u | X > u) = \frac{P(X > t + u)}{P(X > u)} = \frac{S(t + u)}{S(u)}$$

La *fonction de risque* est par définition la survie d'un élément après une durée t sachant qu'il a déjà survécu jusqu'en u. Elle est définie par :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dx} \ln S(t)$$

La *fonction de risque cumulé* est définie telle que :

$$H(t) = \int_0^t h(s) ds$$

Il en résulte directement que la fonction de risque détermine entièrement la loi de *X*. Si l'on veut modéliser une variable aléatoire, on peut choisir sa fonction de risque plutôt que sa densité. Nous obtenons la relation suivante entre la fonction de risque et la fonction de survie :

$$S(t) = \exp(-\int_0^t h(s)ds) = \exp(-H(t))$$

La *censure* à laquelle nous sommes confrontés est une censure à droite de type I, i.e. de type fixe. Le principe est le suivant :

Soit un échantillon de durées de survie $(X_1, \dots, X_n)$ et $C >$ fixé; on dit qu'il y a censure à droite pour cet échantillon si au lieu d'observer $(X_1, \dots, X_n)$ directement on observe $(T_1, D_1), \dots, (T_n, D_n)$ avec :

$$T_i = X_i \wedge \mathrm{C} \qquad \text{et} \qquad D_i = \begin{cases} 1 & si\ X_i \leq \mathrm{C} \\ 0 & si\ X_i > C \end{cases}$$

La censure que l'on souhaite traiter dans notre étude correspond à la question suivante : est-ce que le sinistre est clos à fin 2009 ? L'événement $[X_i \leq \mathrm{C}]$ représente le cas où le sinistre est clos. La clôture d'un sinistre s'observe uniquement si elle a lieu avant fin 2009, date de notre inventaire.

La méthode du maximum de vraisemblance sert par la suite pour l'estimation des coefficients. La vraisemblance du modèle associée aux observations $(T_1, D_1), \dots, (T_n, D_n)$ possède une composante continue et une composante discrète, elle s'écrit :

$$L(\theta) = \prod_{i=1}^{n} f_\theta(T_i)^{D_i}\ S_\theta(C)^{1-D_i}$$

En d'autres termes, lorsque l'on a observé la sortie avant la censure, c'est le terme de densité qui intervient dans la vraisemblance, et dans le cas contraire on retrouve le terme discret, avec comme valeur la fonction de survie à la date de censure. La distribution est donc continue par rapport à T_i et discrète par rapport à D_i.

2. *Théorie*

2.1. Définition

Le modèle de Cox est un modèle de régression qui spécifie la fonction de risque de la manière suivante :

$$\ln \frac{h(t|Z,\theta)}{h_0(t)} = -Z\acute{}\theta$$

où Z est un vecteur de longueur p de covariable que l'on n'autorise pas ici à dépendre du temps. Le vecteur θ est un vecteur de paramètres. h_0 est la fonction de risque instantanée dite de base. Cette fonction est inconnue dans cette étude. Elle devient donc un paramètre de nuisance. C'est pourquoi le modèle de Cox est un modèle semi-paramétrique par la présence du paramètre θ et la fonction de base inconnue. C'est finalement un modèle linéaire généralisé pour la probabilité de survie.

2.2. Hypothèses

(H1) Hypothèse des risques proportionnels

Les rapports des risques instantanés pour deux sinistres i et j ne doit pas dépendre du temps :

$$\frac{h(t|Z_i)}{h(t|Z_j)} = \frac{\exp(Z_i'\theta)}{\exp(Z_j'\theta)}$$

Cette hypothèse est la plus importante à vérifier lorsque l'on souhaite valider ce modèle.

(H2) Hypothèse de log-linéarité

Dans le modèle de Cox, la relation entre la fonction de risque et les covariables est log-linéaire, c'est-à-dire :

$$\log(h(t|Z) = \log(h_0(t) + \theta^{(1)}Z^{(1)} + \cdots + \theta^{(p)}Z^{(p)})$$

Cette hypothèse est rarement vérifiée de manière générale. D'une part, c'est une hypothèse forte et, d'autre part, elle est difficilement vérifiable. Il existe cependant des méthodes graphiques, comme celle utilisant les résidus martingales en fonction des variables. Mais cette méthode n'est pas Nécessairement fiable.

2.3. Estimation des paramètres

Dans notre étude, nous avons deux spécificités par rapport à un modèle de Cox classique : la censure et le problème du traitement des ex-æquo.

- En effet, dans le cas où la fonction de risque est connue, les coefficients sont estimés par la méthode classique du maximum de vraisemblance. Cependant lorsque ce n'est pas le cas, les coefficients de régression peuvent être estimés par la méthode du maximum de vraisemblance partielle. Cox propose une décomposition de la fonction de vraisemblance :

$$L(\theta, h_0) = \prod_{i=1}^{n} [h_0(t_i) \exp(Z_i'\theta) \exp(-H_0(t_i) \exp(Z_i'\theta))]^{D_i} [\exp(-H_0(t_i) \exp(Z_i'\theta))]^{1-Di}$$

On remarque facilement que la fonction de risque de base intervient de deux manières, directement et à travers la fonction de risque cumulé.

Cox propose donc une décomposition de la vraisemblance en deux termes. Dans l'un des termes, l'incidence de la fonction de risque de base est isolée ; elle est négligée par la suite. D'où l'obtention d'une vraisemblance partielle de la forme :

$$L_{part}(\theta) = \prod_{i=1}^{n} \left[\frac{\exp\left(Z_i \acute{} \theta\right)}{\sum_{j=1}^{n} \exp\left(Z_j \acute{} \theta\right) 1_{[T_i \leq T_j]}} \right]^{D_i}$$

Soit r_i le « risk score » pour le sinistre i tel que :

$$r_i = \exp(Z_i'\theta)$$

Cette notation nous permet de rendre plus lisibles les équations.

- Dans la réalité, les durées de vie sont des variables continues. Il ne devrait donc pas y avoir d'ex-æquo. Mais, pour des raisons pratiques, les observations sont arrondies à l'unité, ce qui amène à de nombreux cas dans notre jeu de données.
  Deux méthodes existent pour pallier à cette difficulté. Elles s'ajoutent à la méthode dite « exacte » correspondant à la vraisemblance partielle décrite au-dessus: Efron et Breslow. Elles sont équivalentes sur des données distinctes.

  Nous avons choisi de garder la méthode d'Efron qui est plus précise et moins gourmande en temps de calcul que la méthode de Breslow ou la méthode « exacte ». Prenons l'exemple suivant : sur un ensemble de quatre sinistres, les deux premiers sinistres se clôturent en deux ans. Si le pas de temps

des données était plus fin (unité en mois, jour ou heure), on pourrait savoir si les deux premiers termes de la vraisemblance sont :

$$L_{1,2} = \frac{r_1}{r_1+r_2+r_3+r_4}\frac{r_2}{r_2+r_3+r_4} \quad \text{ou} \quad L_{1,2} = \frac{r_2}{r_1+r_2+r_3+r_4}\frac{r_1}{r_1+r_3+r_4}$$

On remarque que le numérateur du produit est inchangé mais pas le dénominateur.
Dans le cas de la méthode de Breslow, l'approximation est la suivante :

$$L_{1,2} \approx \frac{r_1 r_2}{\left(\sum_{i=1}^{4} r_i\right)^2}$$

Alors que pour la méthode Efron :

$$L_{1,2} = \frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)\left(\frac{r_1}{2} + \frac{r_2}{2} + r_3 + r_4\right)}$$

Cette méthode peut être considérée comme plus précise car r_1 et r_2 ont la même probabilité d'être la bonne valeur.

3. Validation du modèle

Afin d'évaluer si l'ajustement du modèle de Cox choisi est pertinent, certains critères doivent être vérifiés :

3.1. Significativité des coefficients

La significativité des coefficients est étudiée de la même manière que pour un modèle de régression classique; à savoir, test de manière globale grâce à la statistique de Fisher et de manière marginale grâce la statistique de Student et utilisation de la p-value.

3.2. L'hypothèse des risques proportionnels

Sous R, nous avons deux possibilités de vérifier que les paramètres ne dépendent pas du temps, grâce au test du chi-deux et la méthode graphique.

3.2.1. Test du chi-deux

Le test marginal du Chi-deux analyse le problème suivant pour chaque covariable j, i.e.:

$$\begin{cases} H_0 : \theta^{(j)}(t) = \theta^{(j)} \\ H_1 : \theta^{(j)}(t) \neq \theta^{(j)} \end{cases}$$

Tandis que le test global, lui teste directement :

$$\begin{cases} H_0 : \theta(t) = \theta \\ H_1 : \theta(t) \neq \theta \end{cases}$$

Où $\theta(t) = (\theta^{(1)}(t), \dots, \theta^{(p)}(t))$ le vecteur des paramètres.

3.2.2. Analyse graphique

On peut procéder à l'étude des graphes représentant les résidus de type « Schoenfeld » en fonction du temps. Contrairement aux résidus rencontrés habituellement en régression, ceux-ci ne sont pas des écarts entre les valeurs observées et prédites de la variable réponse. Les résidus portent sur les covariables et mesurent des écarts par rapport au profil moyen en termes de ces variables. Plus précisément, ils mesurent la différence entre la valeur des covariables d'un cas connaissant l'événement en t et une moyenne pondérée des valeurs de ces mêmes covariables pour les cas exposés en t, i.e. ceux qui ont une durée de survie supérieure à t.

Ils servent à évaluer la tendance au cours du temps et sont particulièrement utiles pour juger de la pertinence de l'hypothèse de proportionnalité. Cependant ils ne donnent pas d'indicateurs sur la qualité de l'ajustement. Le résidu de Schoenfeld correspondant à la covariable j et à la $i^{ème}$ durée non-censurée est donné par :

$$\widehat{\mathrm{Scho}_i}^{(j)} = Z_i^{(j)} - \bar{Z}^{(j)}(\hat{\theta}, T_i)$$

Où T_i correspond à la durée observée pour le $i^{ème}$ sinistre et $Z_i = \left(Z_i^{(1)}, \dots, Z_i^{(p)}\right)'$ est le vecteur de covariable du $i^{ème}$ sinistre (p=4). La $j^{ème}$ coordonnée du vecteur de moyenne pondérée à l'instant t des vecteurs de covariables des sinistres à risque à l'instant t est donnée par :

$$\bar{Z}^{(j)}(\hat{\theta}, t) = \sum_{k=1}^{n} \frac{1(T_k \geq t)\exp(\theta' Z_k)}{\sum_{l=1}^{n} 1(T_l \geq T_k)\exp(\theta' Z_l)} Z_k^{(j)}$$

La droite que l'on pourra étudier indique la tendance. L'absence de pente valide l'hypothèse de proportionnalité, i.e qu'il n'y a pas de tendance en fonction du temps. En effet, une pente positive indiquerait une croissance de l'effet de la covariable en fonction du temps alors que dans le cas d'une pente négative, une diminution de l'effet de cette covariable serait constatée. Cependant dans la pratique, l'analyse graphique ne se révèle pas entièrement fiable. Néanmoins elle permet visuellement d'observer pour chaque variable si l'hypothèse n'est pas trop éloignée de la réalité et d'en déduire de nouvelles pistes le cas échéant.

Partie 3 : Applications et résultats

III. Préliminaires

1. Etude de l'inflation

L'inflation est un phénomène persistant qui fait monter l'ensemble des prix. Sur un portefeuille à développement long comme la RC médicale, elle a un impact important. Il est donc indispensable de l'évaluer correctement.

Sa variation n'est pas directement mesurable. Elle s'évalue à partir des variations des prix à la consommation des biens et services, mesurée à quantité et qualité égales, ou des salaires. Au Royaume-Uni, l'inflation est évaluée au moyen du Retail Price Index (RPI ~ 4-5%) ou du National Average Earnings (NAE~3-4%). Pour cette étude, son évaluation se fera à partir des coûts moyens des sinistres, car pour la RC médicale, les indemnités comprennent différents montants qui ne subissent pas de la même manière l'inflation tels que des frais médicaux, frais juridiques ou les salaires des personnels de santé.

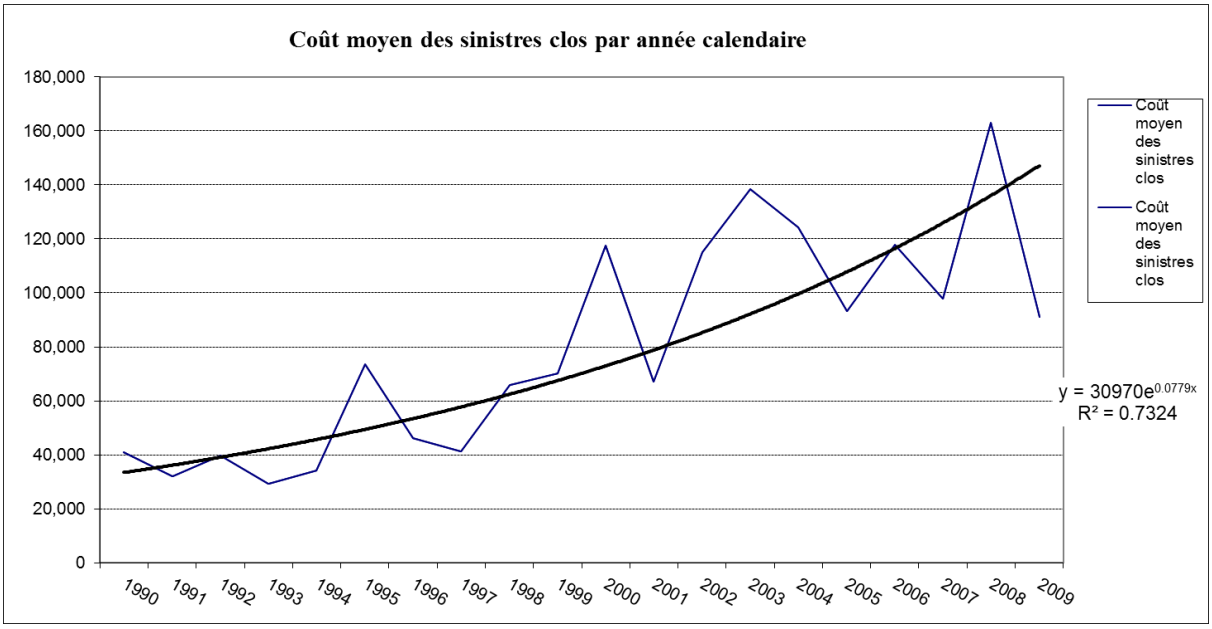
L'inflation est supposée la même quel que soit le type de sinistre « Small » ou « Large ». L'approche pour évaluer l'impact de l'inflation sur les plaintes est la création de triangles de liquidation incrémentaux par année de notification (ligne) et par année calendaire (colonne) concernant le nombre de sinistres, et également les montants d'indemnités. Les triangles de liquidation contiennent uniquement le nombre de plaintes ou des montants pour l'année calendaire de la clôture du sinistre. Par exemple pour un sinistre notifié en 2009 et clos en 2010, la valeur pour l'année de notification 2009 est enregistrée dans l'année calendaire 2010. Cela permet de déterminer les coûts moyens des indemnités pour chaque année calendaire.

L'étape suivante est de trouver une tendance à la courbe des montants. Le meilleur ajustement pour les coûts moyens est donc une courbe exponentielle de la forme :

$$y = b\, e^{ax}$$

Le paramètre $\hat{a}$ a été estimé sur des historiques différents compris entre quinze et vingt années. Différentes estimations de l'inflation $\hat{\imath}$ sont alors définies par $\hat{\imath} = e^{\hat{a}} - 1$.

Voici l'ajustement fait pour l'historique depuis 1990 à 2009 :



Le taux d'inflation obtenu est de 9%, ce qui est bien au-dessus du RPI ou NAE. Cependant, une récente étude « IUA Bodily Injury Study » a estimé le taux d'inflation concernant le secteur de RC médicale britannique autour de 10% pour les dix prochaines années. Le taux de 9% estimé apparait ainsi pertinent par rapport à cette étude.

2. *Estimation de la variable explicative de développement*

Pour les modèles de régression, la variable explicative « dev » est importante. Elle décrit le nombre d'années de développement pour un sinistre. Dans le cas des sinistres clos, le calcul est:

$$dev = yos - yon + 1$$

Où yos correspond à l'année de clôture et yon l'année de notification.

Mais les sinistres ouverts ne possèdent pas d'année de clôture ce qui complexifie le calcul. Le nombre d'années nécessaires pour clôturer le sinistre ne peut être déterminé de façon certaine, Par conséquent, il doit être estimé. Nous sommes dans l'incapacité d'observer la réalisation de cette variable pour tous les sinistres. Seul le nombre d'années écoulées jusqu'à fin 2009 est observable.

$$T = 2009 - yon + 1$$

Ceci est l'exemple d'une censure de type I à droite. Le modèle de Cox permet d'estimer la variable « dev » pour tous les sinistres.

La variable de censure D_i est la suivante :

$$D_i = \begin{cases} 1 & si\ le\ sinistre\ est\ clos\ fin\ 2009 \\ 0 & si\ le\ sinistre\ est\ ouvert\ fin\ 2009 \end{cases}$$

Le modèle optimal sélectionné régresse la fonction de survie de la variable *dev* en fonction de quatre variables explicatives :

- *Yon* indique l'année de notification du sinistre.
- *Prec* indique si le sinistre a été un « Precautionnary claim » auparavant.
- *Inc2009* indique le montant de la charge à fin 2009.
- *Zero-paid* indique si le sinistre n'a eu aucun paiement jusqu'à maintenant.

Voici la sortie R obtenue à titre d'exemple. La première colonne décrit les coefficients de la régression. La deuxième colonne « se » est l'écart-type des coefficients, la troisième colonne « z » est le ratio des coefficients sur leurs écart-types et la dernière colonne correspond à la p-value des tests de significativité de chaque coefficient.

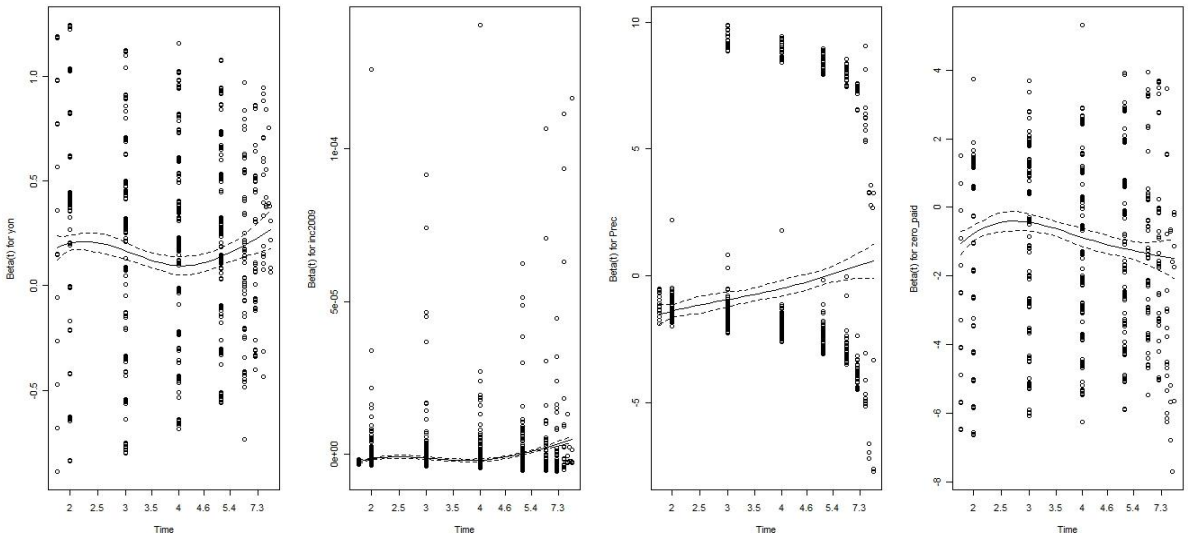
```
> summary(coxdev)
Call:
coxph(formula = Surv(data$dev, censure) ~ yon + Prec + inc2009 +
    zero_paid, data = data)

  n= 1801, number of events= 1671

                coef  exp(coef)   se(coef)       z Pr(>|z|)
yon         1.651e-01  1.180e+00  1.116e-02  14.800   <2e-16 ***
Prec       -8.021e-01  4.484e-01  8.115e-02  -9.884   <2e-16 ***
inc2009    -8.952e-07  1.000e+00  1.255e-07  -7.130    1e-12 ***
zero_paid -8.230e-01  4.391e-01  6.834e-02 -12.043   <2e-16 ***
```

Validation du modèle :

- Au seuil 5%, les quatre variables explicatives sont significatives.
- Pour l'hypothèse de proportionnalité, la p-value « p » du test du chi-deux nous permet d'accepter l'hypothèse de risque proportionnel pour chaque variable ainsi que pour le modèle global.

```
> print(test)
              rho  chisq        p
yon       -0.0319   1.86 1.72e-01
Prec       0.1798  50.89 9.79e-13
inc2009    0.1124  71.30 0.00e+00
zero_paid -0.0708   6.76 9.33e-03
GLOBAL         NA 136.12 0.00e+00
```

Une analyse graphique des résidus de Schoenfeld a été effectuée:



La conclusion semble bien moins évidente que par le test du Chi-deux. Les droites lissées ne semblent pas avoir une pente nulle. Pour les montants de l'année 2009 et l'année de notification, il est possible de considérer une droite horizontale, donc la validation de l'hypothèse de proportionnalité. En ce qui concerne la variable *Prec*, la présence d'une pente positive permet de conclure à une augmentation de son effet en fonction du temps alors que pour la variable *zero_paid*, l'effet serait décroissant.

Comme l'analyse graphique des résidus ne porte que sur la validation marginale variable par variable, la conclusion du test du Chi-deux global prime sur l'analyse des résidus. En effet, la validation de l'hypothèse pour le modèle global est avant tout recherchée. Cependant, si l'hypothèse est validée marginalement, les résultats ne peuvent en être que meilleurs.

Ce modèle semble donc avoir un ajustement satisfaisant puisque les variables sont toutes significatives et le test du Chi-deux ne rejette pas l'hypothèse de proportionnalité pour le modèle global ainsi que pour chaque variable.

Ce modèle va donc être utilisé afin de fournir des prévisions de durées de développement des sinistres encore ouverts. R calcule directement les prévisions du modèle linéaire, cependant il faut effectuer quelques transformations manuelles avant d'obtenir la durée de développement finale.

Soit p_i la prévision linéaire fournie sous R pour le sinistre i.

Afin d'obtenir le taux de survie, nous devons appliquer la fonction suivante :

$f(t) = \exp(-\exp(t))$.

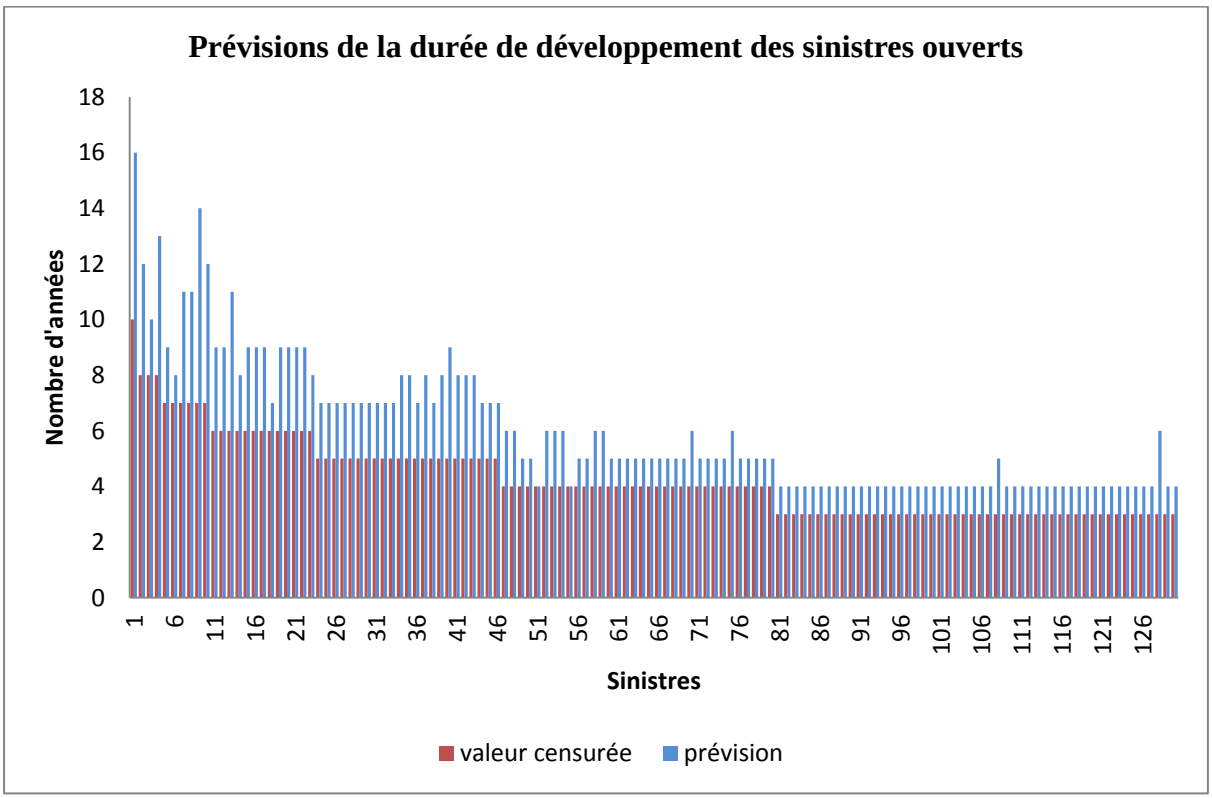
En effet, la fonction de risque est définie par :

$$\ln \frac{h(t|Z,\theta)}{h_0(t)} = -Z'\theta$$

Ceci permet de retrouver la fonction $f(t)$ afin d'avoir effectivement les prévisions de la quantité $h(t|Z,\hat{\theta})$.

La prédiction de la durée de développement des sinistres ouverts s'est faite à l'aide d'un échantillon composé de 2101 sinistres (censurés et non censurés).

Voici les résultats sous forme graphique :



Les prévisions obtenues semblent raisonnables et cohérentes avec la distribution des durées de développement des sinistres fermés. En effet, la moyenne est légèrement inférieure car elle est égale à 3.18 au lieu de 3.28, le minimum et le maximum ainsi que la médiane sont égaux. On retrouve avec les estimations des sinistres ouverts une distribution similaire aux sinistres fermés. Ces estimations sont donc exploitables afin d'éviter un problème de double censure des données.

3. *Méthodologie des IBNyR*

Les provisions pour IBNyR (Incurred But Not yet Reported) sont une estimation du coût ultime des sinistres survenus mais non encore déclarés à la date de clôture. Les IBNyR peuvent être calculés en utilisant une méthode déterministe de fréquence/sévérité. Comme par la suite les modèles sont divisés entre « Small » et « Large », il faut calculer les IBNyR pour les deux catégories de manière séparée. En effet, le risque principal des IBNyR est basé sur l'incertitude des sinistres « Large ».

Pour la modélisation de la fréquence, la méthode de Chain Ladder classique sera utilisée sur des triangles de liquidation incrémentaux concernant le nombre de sinistres en fonction de l'année de notification (ligne) et de l'année calendaire (colonne). Les triangles de liquidation contiennent uniquement le nombre de cas ouverts pour l'année calendaire d'ouverture effective du sinistre. Ce qui permet d'obtenir le nombre de sinistres à l'ultime pour chaque année de notification.

Pour la modélisation de la sévérité, la méthode est différente pour chaque catégorie :

- Pour les sinistres « Small », le coût moyen est calculé de manière similaire à l'inflation. En effet, sur ce type de sinistre le risque de volatilité des montants est faible au contraire des sinistres « Large ». Le coût moyen reflète de manière assez fiable la sévérité pour ces sinistres.
- Pour les sinistres de type « Large », on ne peut pas procéder ainsi en raison de la volatilité des montants. L'objectif est de trouver la loi la mieux ajustée aux montants. La sévérité sera donc l'espérance de cette loi avec les paramètres estimés. Trois lois ont été retenues car elles sont classiques pour ce type de sinistre, à savoir la loi lognormale, la loi de Pareto et la loi de Weibull. La loi du Ch-deux a été retenue en raison sa plus petite valeur statistique, i.e. la loi de Pareto. En effet, cela signifie que l'écart entre les probabilités empiriques et théoriques est le plus faible.

On obtient finalement les résultats suivants :

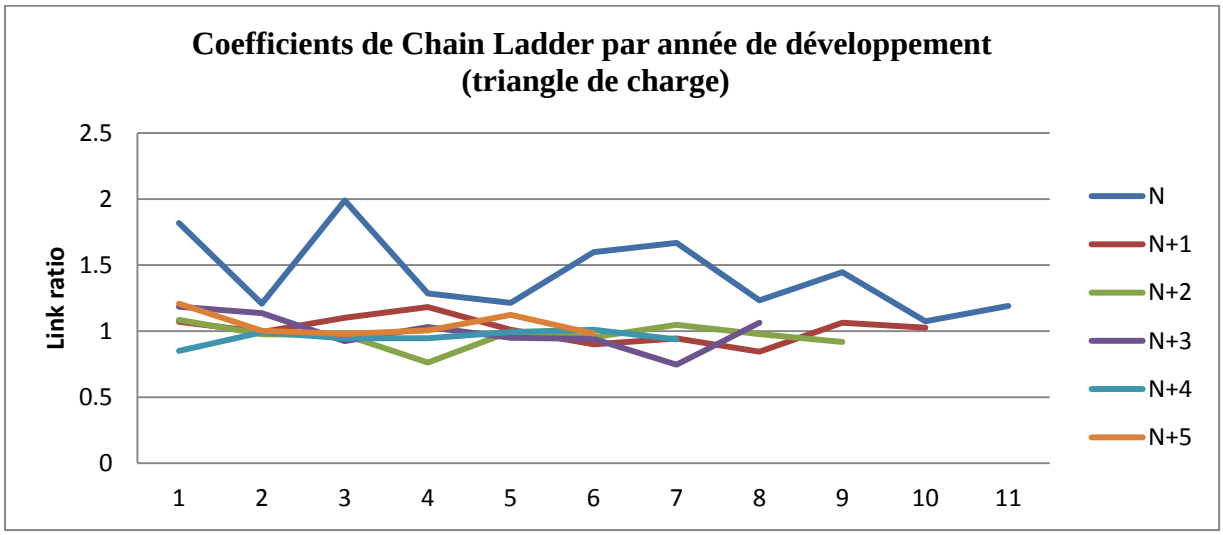
IBNyR	Montant (£)	Proportion (%)
Small	478 352	2%
large	23 771 101	98%
Total	24 249 454	100%

IV. Modèle de référence

Le modèle de Chain-Ladder sert de benchmark pour cette étude, i.e. il va être utilisé comme référence afin de comparer les résultats des méthodes actuarielles alternatives. Un montant de charge ultime est associé à chaque type de sinistres. Ce modèle prend en compte de manière implicite l'inflation et les IBNyR. Le logiciel ResQ a été utilisé pour les calculs de la méthode Chain Ladder.

1. Modélisation à partir du triangle de charge

En premier lieu, il faut vérifier l'hypothèse d'indépendance des années de notification pour que cette méthode soit applicable. Le graphe ci-dessous représente les six premières années de développement.



La première année ne présente pas un comportement stable au cours du temps. Cependant on peut considérer que les facteurs individuels sont sensiblement constants pour les autres années et donc valider l'hypothèse d'indépendance des années de notifications pour ces années-là. Cela souligne le fait que, malgré l'attrait universel envers de la méthode Chain Ladder en raison de sa simplicité, elle n'est pas toujours la mieux adaptée. C'est pourquoi, il est intéressant dans cette étude de présenter de nouvelles méthodes plus adéquates aux spécificités des données. Dans la réalité, une alternative possible à la méthode de Chain Ladder, parmi les méthodes déterministes qui utilise les triangles également, dans le cas où l'hypothèse d'indépendance des années de notification n'est pas validée pour plusieurs années, est la méthode de Bornhuetter Ferguson. C'est une méthode composite qui utilise le principe de la crédibilité. Elle suppose que l'on dispose d'une information externe sur la valeur probable finale de la charge totale des sinistres.

Cette méthode se formule alors ainsi :

$$L = D\frac{1}{LDF} + A\left(1 - \frac{1}{LDF}\right)$$

Où L le coût total estimé par cette méthode.

D le coût total estimé en fonction des sinistres connus.

A le coût total (connus, tardifs) attendu a priori.

LDF la proportion de la liquidation déjà constatée.

Lorsque la méthode de Chain Ladder n'est pas très performante en début de développement car l'information connue est relativement faible. La méthode de Bornhuetter Ferguson permet donc de contourner ce problème en donnant un poids à un loss ratio a priori.

Ici la méthode de Chain Ladder reste satisfaisante, en voici les résultats :

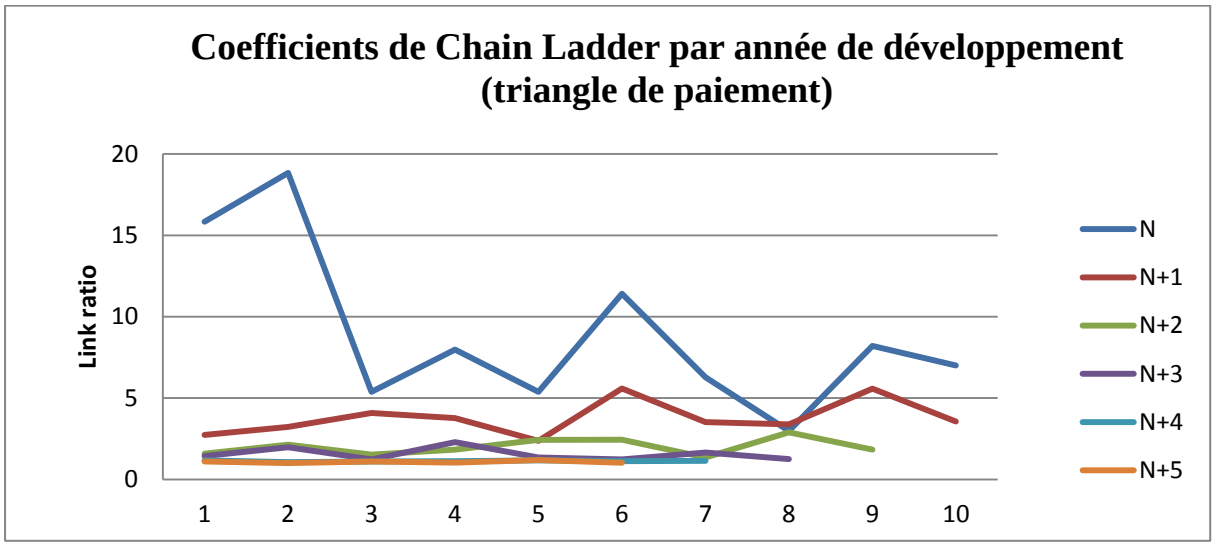
L'unité est £1 000.

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	Ultime
1998	6 544	11 899	12 738	13 819	16 367	15 596	18 847	17 991	19 788	20 502	20 807	20 787	24 611
1999	11 693	14 127	14 025	13 699	15 561	14 458	14 507	14 058	13 974	13 974	13 974	13 961	16 529
2000	8 746	17 398	19 163	18 606	17 186	14 653	14 347	14 184	14 126	14 050	14 174	14 161	16 766
2001	11 459	14 727	17 421	13 266	13 676	13 313	13 389	13 431	13 376	13 554	13 674	13 661	16 174
2002	10 015	12 159	12 283	12 240	11 626	13 288	14 925	15 015	15 417	15 623	15 761	15 746	18 642
2003	12 703	20 300	18 271	17 422	16 397	14 702	14 421	14 167	14 547	14 741	14 871	14 857	17 590
2004	11 495	19 172	18 120	18 992	14 183	14 577	15 327	15 058	15 462	15 668	15 806	15 791	18 696
2005	18 766	23 133	19 535	19 102	20 310	19 457	20 458	20 099	20 638	20 913	21 098	21 077	24 955
2006	16 334	23 626	25 131	23 070	22 736	21 781	22 902	22 499	23 103	23 411	23 618	23 595	27 936
2007	25 770	27 682	28 391	27 919	27 515	26 359	27 715	27 228	27 958	28 331	28 582	28 554	33 807
2008	28 220	33 589	34 517	33 943	33 452	32 047	33 696	33 104	33 991	34 445	34 749	34 716	41 102
2009	28 926	38 952	40 029	39 363	38 794	37 164	39 076	38 389	39 419	39 945	40 298	40 259	47 666

La charge ultime correspond à la somme de la dernière colonne car une queue de distribution a été ajoutée en raison du développement de la branche qui est plus long pour certains sinistres. En effet, il existe un risque de dégradation des sinistres au-delà de l'horizon de onze ans, ce phénomène pris en compte avec cette queue de distribution. La distribution Inverse Normale a été retenue car elle présentait le meilleure R^2. La charge ultime obtenue est de **£ 304 475 182.**

2. *Modélisation à partir du triangle de paiement*

De la même manière, l'hypothèse d'indépendance des années de notifications pour la méthode de Chain Ladder appliquée au triangle des paiements est testée.



La même instabilité est constatée sur la première année de développement. Mais de manière similaire au Chain Ladder appliqué aux montants de charge, la méthode reste satisfaisante puisque pour les autres années l'hypothèse est validée.

Voici le résultat de la méthode Chain Ladder :

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	Ultime
1998	9	1,414	3,864	6,142	8,857	10,417	11,538	15,947	16,291	17,026	20,566	20,586	20,679
1999	59	929	3,000	6,395	12,647	13,422	13,551	13,729	13,771	13,771	13,771	13,837	13,900
2000	57	1,495	6,103	9,334	11,458	12,657	13,922	14,027	14,049	14,049	15,664	15,739	15,810
2001	38	714	2,690	4,920	11,286	12,667	13,127	13,230	13,230	13,450	14,996	15,069	15,137
2002	32	984	2,333	5,674	7,660	8,996	10,767	11,220	11,301	11,489	12,809	12,871	12,929
2003	109	588	3,287	8,031	9,889	11,060	11,376	12,325	12,413	12,620	14,070	14,138	14,202
2004	111	886	3,118	4,211	6,955	7,952	8,533	9,246	9,312	9,467	10,555	10,606	10,654
2005	136	733	2,481	7,173	8,965	10,063	10,799	11,700	11,784	11,980	13,357	13,422	13,482
2006	58	659	3,682	6,742	10,100	11,337	12,166	13,181	13,275	13,496	15,047	15,120	15,188
2007	154	967	3,445	6,609	9,900	11,113	11,925	12,920	13,013	13,229	14,750	14,821	14,888
2008	511	1,520	5,517	10,584	15,855	17,797	19,098	20,692	20,840	21,187	23,622	23,736	23,844
2009	101	863	3,132	6,008	9,000	10,102	10,840	11,745	11,829	12,026	13,408	13,473	13,534

Une queue de distribution a été ajoutée pour la même raison que précédemment. La loi de Weibull a été retenue car elle présentait le meilleure R^2. La charge ultime s'élève à **£ 184 247 394**.

On remarque tout de même un grand écart de la charge ultime estimée à partir de la charge ou des paiements. En effet, sur une branche à développement long, les paiements sont beaucoup moins stables pour évaluer la charge ultime comme nous le verrons dans le troisième modèle.

V. Modèle 1 : modélisation dossier/dossier et problématique des « zéro-inflatés »

Ce modèle est composé de trois sous modèles afin de s'adapter à chaque problématique de manière appropriée. L'intérêt fondamental de ce modèle est l'utilisation de l'historique dossier/dossier du passé et non pas une agrégation des montants à travers les triangles de liquidations. En effet, l'agrégation des données simplifie la modélisation, mais l'information sinistre par sinistre se perd.

Un modèle GLM binomial permet de traiter la problématique des nombreux sinistres ayant une charge ultime nulle. Une fois ces sinistres traités, les prévisions des sinistres « Small » et « Large » sont séparées en utilisant deux modèles différents.

1. *GLM binomial*

L'une des particularités de ce portefeuille est l'ouverture d'un dossier par les gestionnaires à toute déclaration de problème par un hôpital lors d'une intervention, et ce, avant toute mise en cause de la part du patient. Cela conduit à avoir 75% de sinistres sans aucun paiement à la clôture dans la base de données. Par conséquent, ces sinistres ne peuvent pas être utilisés directement lors de la mise en place des modèles de prévision. En effet, ils biaisent de manière importante les résultats et amènent à sous-estimer les réserves. C'est la problématique des modèles zero-inflatés. La suppression de ces sinistres, sans les prendre en considération, aurait été possible. Cela est bien le cas dans la méthode de Chain Ladder, puisqu'un montant nul ne fait en rien varier la somme totale. Cependant l'objectif est de raffiner l'étude du comportement de ce portefeuille et donc de réussir à en appréhender au mieux ses caractéristiques. C'est pourquoi l'explication de ces sinistres clos sans aucun paiement est intéressante. Et cela permet par la suite, de déterminer la probabilité du sinistre d'avoir, ou non, une charge ultime nulle.

Le modèle GLM binomial semble le plus adéquat.

Soit Y_0 la variable aléatoire à expliquer telle que $Y_0 = \begin{cases} 1 \; si \; la \; charge \; ultime \; > 0 \\ 0 \; sinon \end{cases}$.

Y_0 est donc la réalisation d'une variable aléatoire de Bernoulli $Y_0 \sim \beta(\alpha)$ où α est la probabilité qu'un sinistre ait une charge ultime strictement supérieure à 0.

L'estimation des coefficients est faite à partir des 3581 sinistres clos. Dans le modèle optimal, Y_0 est expliqué à l'aide de trois variables explicatives et de la fonction de lien Cauchit $(g(m) = F_C^{-1}(m)$, où F_c^{-1} est l'inverse de la fonction de répartition d'une loi de Cauchit)

```
glm(formula = Y0 ~ dev_est + yon + zero_paid, family = binomial("cauchit"),
    data = data)
```

Ce modèle permet d'estimer $\hat{\alpha}$ et de décider si le sinistre aura une charge nulle : si $\hat{\alpha} \leq \frac{1}{2}$, alors le sinistre aura une charge ultime nulle.

Finalement, toutes les probabilités prédites sont supérieures à 0.9, donc aucun des sinistres ouverts ne devraient se clore sans paiement.

2. Modélisation des « Small »

Au départ, la piste des GLM a été testée en supposant que la variable à expliquer, la charge ultime, suivait une loi Gamma. Les tests d'adéquation à la loi ne rejetaient pas la loi. Cependant en raison de la dispersion des montants, le modèle ne convergeait pas. Une transformation logarithmique des montants a été essayée par la suite. Le modèle était satisfaisant et les prévisions correctes. Mais, les erreurs de prévisions qui étaient minimes à l'échelle logarithmique devenaient importantes à l'échelle normale. Les résultats n'étaient donc pas satisfaisants. Les autres lois disponibles dans le logiciel R ne correspondaient pas à la variable à expliquer. Ainsi, l'approche des GLM n'a pas été retenue pour un traitement sinistre par sinistre des données.

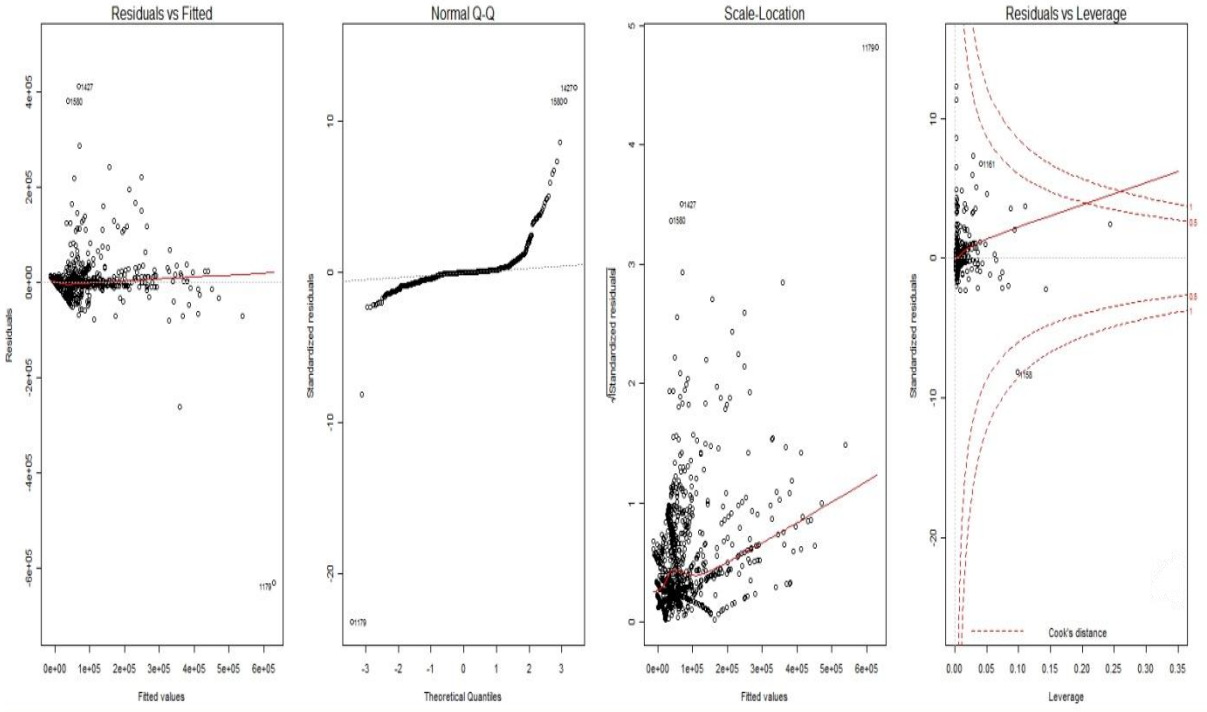
Le modèle linéaire multiple s'est donc imposé. En effet, même si le GLM semblait plus intéressant au premier abord car le choix de la loi et de la fonction de lien permettent de mieux tenir compte des spécificités ; la simplicité du modèle linéaire permet de prédire directement les montants et d'éviter l'écueil du passage logarithme-exponentiel.

L'estimation des coefficients est faite à partir des 1 636 sinistres clos afin de pouvoir l'appliquer sur les 377 sinistres ouverts. Le modèle linéaire optimal explique donc le montant de la charge ultime en fonction de sept variables explicatives :

```
lm(formula = inc2009 ~ zero_paid + inc2008 + zero_paid:inc2008 +
    Prec:inc2008 + dev_est:inc2008 + dev_est:Prec + yon:dev_est,
    data = clossmall)
```

L'ajustement du modèle semble satisfaisant a priori car on obtient un R^2 ajusté de 81,1%, et toutes les variables sont significatives.

Pour confirmer le bon ajustement du modèle, l'analyse graphique des résidus de Pearson est effectuée :



Le graphique de prédiction linéaire/résidus (1er) et le graphique des prévisions linéaires en fonction de la racine carrée des résidus standardisés (3ème) ne présentent pas de structuration particulière. Par conséquent, les hypothèses d'homoscédasticité et d'espérance nulle des résidus sont validées. Et le graphique résidus/poids (4ème) n'indique pas la présence de valeurs atypiques dans nos données. Par contre l'analyse du Q-Q plot (2ème) ne valide pas l'hypothèse de normalité des résidus, ce qui invalide l'efficacité des estimations. En effet, la normalité des résidus implique que l'estimation par les MCO est équivalente à celle du maximum de vraisemblance et donc les estimateurs ont les mêmes propriétés, dont l'efficacité. Comme la méthode des MCO est utilisée, il y a perte de l'efficacité dans ce modèle. Nous pourrions utiliser une distribution plus adéquate parmi les distributions à queue lourde, telle que Weibull, Pareto ou Lognormale.

Pour autant, le modèle apparait globalement satisfaisant. Il ne prend cependant pas en compte l'inflation, qui sera alors appliquée par la suite sur le nombre d'années de vie restantes estimées de chaque sinistre. Il faut également ajouter à ce montant l'estimation des IBNyR, car pour ce modèle les estimations ne sont calculées que sur les sinistres clos.

Voici les résultats :

	Montants (£)	Proportions (%)
Charge ultime des sinistres clos	89 097 711	73,2%
Prévisions de la charge ultime des ouverts	25 569 728	21,0%
Inflation	6 573 718	5,4%
IBNyR	478,352	0.4%
Charge ultime totale	**121 719 509**	100.0%

3. Modélisation des « Large »

Les sinistres « Large » sont plus délicats à modéliser car l'échantillon est de taille réduite et leurs montants varient entre £500k et £5,5M.

La piste des GLM a aussi été rejetée pour le traitement des sinistres « Large ». En effet, il y avait des problèmes de convergences similaires aux sinistres « Small ». Même la transformation logarithmique n'était pas utilisable. Le nombre de données était trop faible par rapport au nombre de variables explicatives pour que le modèle puisse converger.

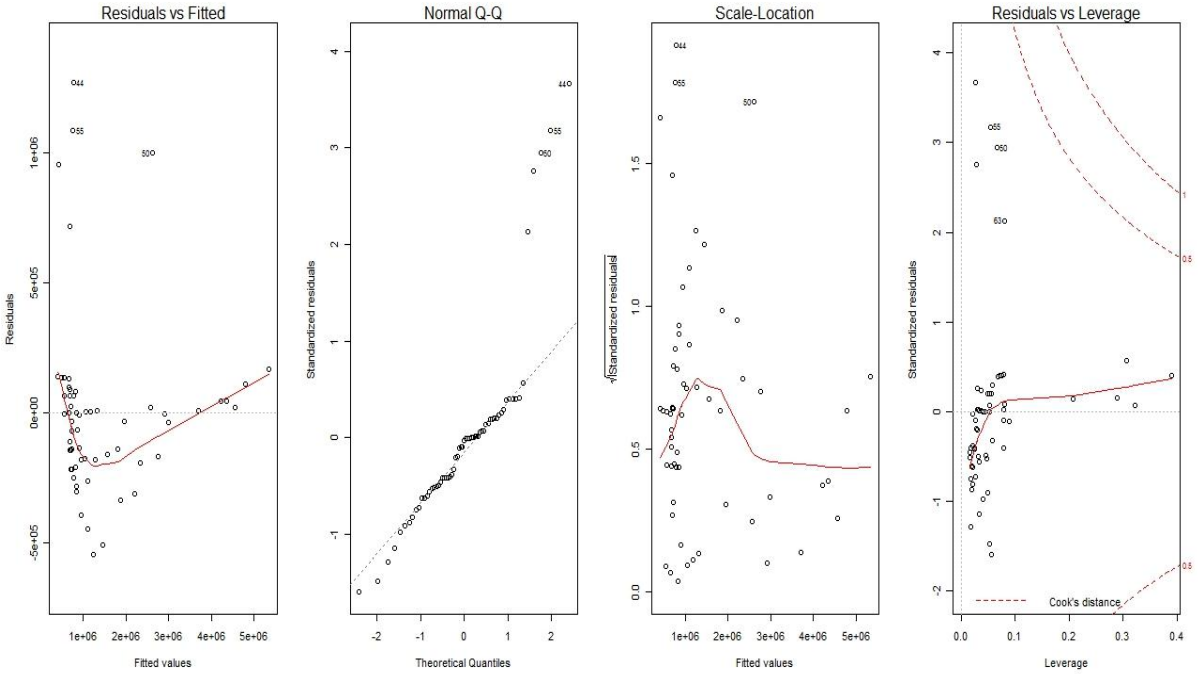
3.1. Modèle linéaire multiple

L'approche du modèle linéaire multiple était donc mieux adaptée.

L'estimation des coefficients est faite à partir des sinistres clos afin de pouvoir l'appliquer sur les sinistres ouverts. Le modèle linéaire optimal explique le montant de la charge ultime en fonction de trois variables explicatives :

```
Call:
lm(formula = inc2009 ~ yon2 + inc2008 + yon2:inc2008, data = closlargecarre)
```

On obtient un R^2 ajusté de 92,07% et toutes les variables sont significatives. L'ajustement semble bon. Il faut valider les propriétés sur les résidus pour confirmer que l'ajustement est acceptable.



Le graphique de prédiction linéaire/résidus (1er) et le graphique des prévisions linéaires en fonction de la racine carrée des résidus standardisés (3ème) ne présentent aucune structuration. Ils permettent de valider les hypothèses d'homoscédasticité et d'espérance nulle des résidus. L'analyse du Q-Q plot (2ème) représente bien une droite d'angle 45°, la normalité des résidus n'est pas rejetée. Les résidus ont les bonnes propriétés. Enfin, le graphique résidus/poids (4ème) n'indique pas la présence de valeurs aberrantes. Le modèle final est donc pertinent.

De la même manière que pour le modèle des sinistres « Small », l'estimation de l'inflation et les IBNyR sont à ajouter pour obtenir la charge ultime. Les IBNyR représentent un risque important pour les sinistres « Large », comme on peut le constater ils représentent 16% de la charge totale.

Voici les résultats :

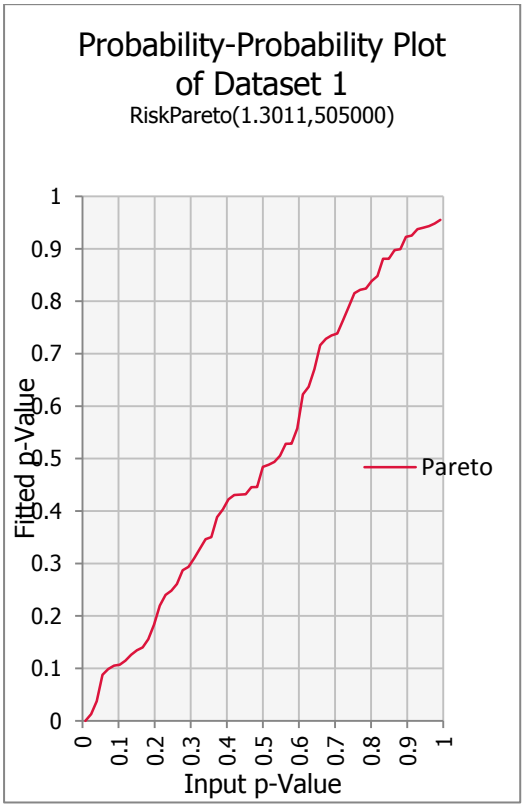
	Montants (£)	Proportions (%)
Charge ultime des sinistres clos	89 995 948	53,0%
Prévisions de la charge ultime des ouverts	38 595 236	22,7%
Inflation	17 551 993	10,3%
IBNyR	23 771 101	14,0%
Charge ultime totale	**169 914 279**	100.0%

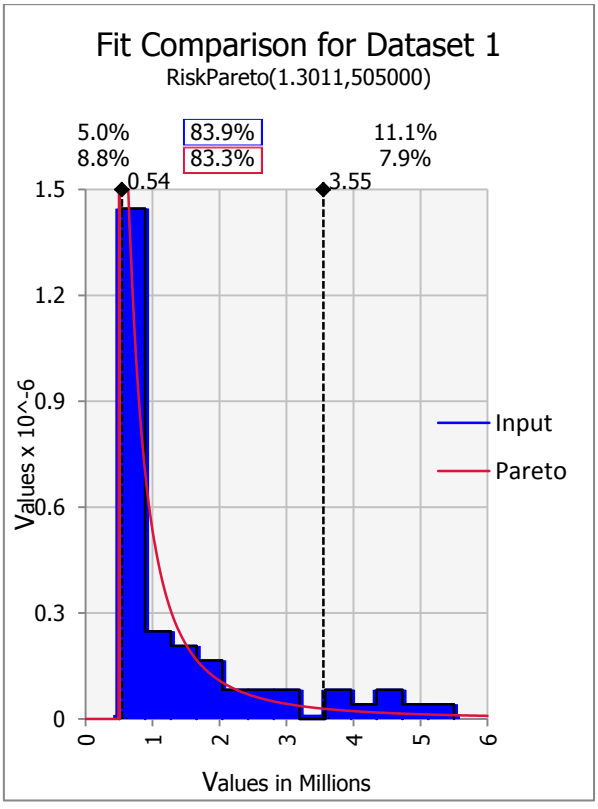
3.2. Adéquation de loi

L'ajustement d'une loi pour les montants est une alternative intéressante au modèle linéaire multiple pour les sinistres « Large ». Ils représentent 4% du nombre de sinistres et représentent 60% de la charge. Les lois de Pareto, Lognormale ou encore Weibull sont des distributions appropriées aux sinistres « Large » en raison de leurs queues épaisses. La charge ultime est représentée par l'espérance de chaque sinistre. Différentes lois ont été étudiées à l'aide du logiciel @Risk, qui procède à plusieurs tests, notamment le test du Chi-deux.

La loi de Pareto est la loi qui s'ajuste le mieux aux données. Le PP-plot, similaire au QQ-plot, représente les probabilités empiriques en fonction des probabilités théoriques au lieu des quantiles. Il permet de confirmer la conclusion du test du chi-deux, i.e. l'ajustement est satisfaisant.

Ci-dessous, le PP-Plot ainsi que l'ajustement de la loi de Pareto estimée par rapport à l'histogramme des données.





Par ailleurs, cette loi possède une propriété intéressante : si la variable X suit une loi de Pareto $P(m,a)$, alors la loi conditionnelle au dépassement d'un seuil $u > m$ notée $X \mid X > u$ est encore une loi de Pareto, de paramètres(u,a) , i.e.

Si $X \sim P(m,a)$ Alors $X|X > c_i \sim P(c_i,a)$

Ceci permet de garder le paramètre $\hat{a}$, estimé par @risk et de faire la translation par rapport au dernier montant de charge c_i connu pour le sinistre i.

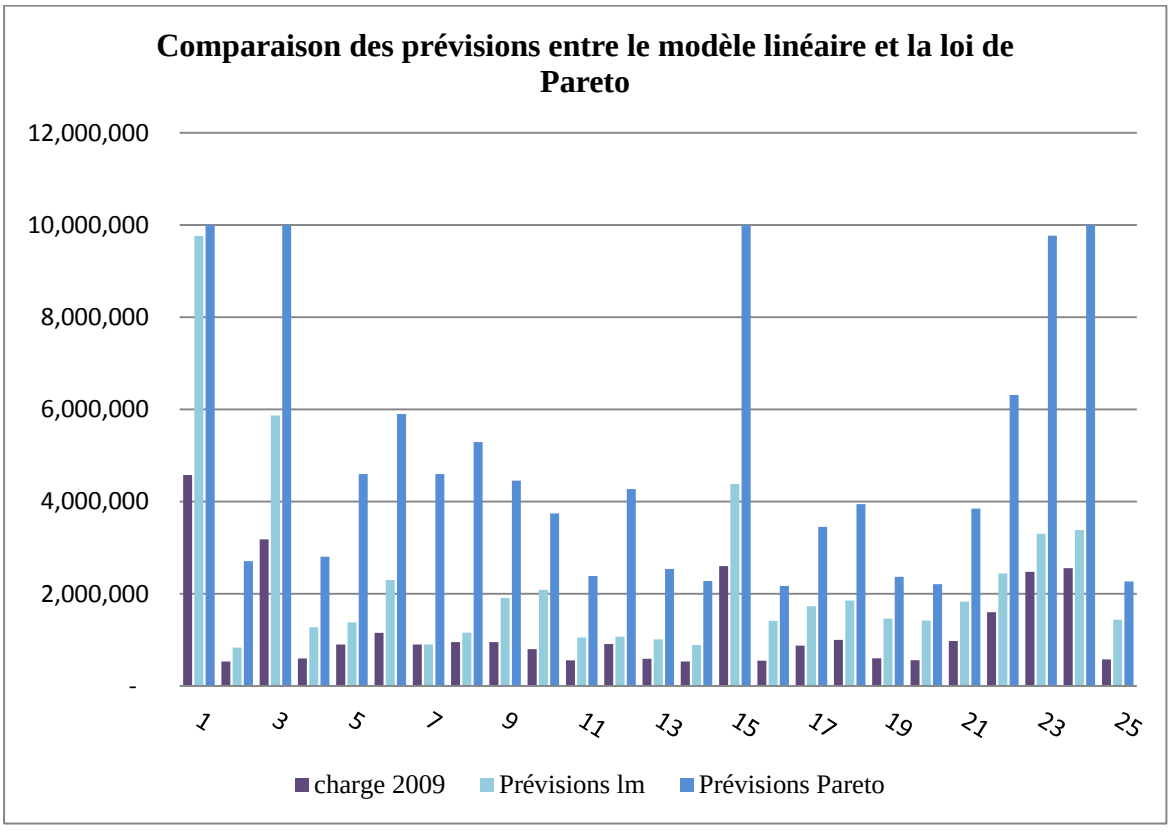
L'espérance pour chaque sinistre est ensuite calculée :

$$E(X) = a\frac{c_i}{(a-1)}$$

D'où les résultats suivants :

	Montants (£)	Proportions (%)
Charge ultime des sinistres clos	89,995,948	38,2%
Prévisions de la charge ultime des ouverts	104 586 177	44,4%
Inflation	17 330 318	7,4%
IBNyR	23,771,101	10,1%
Charge ultime totale	235 683 545	100.0%

Ci-dessous, la présentation graphique des prévisions selon les deux modélisations :



De manière globale, les résultats de la charge ultime montre un degré supérieur de prudence avec l'ajustement par la loi de Pareto, i.e. 28%. Ce graphique montre que l'estimation de la charge ultime

pour chaque sinistre est supérieure. Cette méthode donne des estimations plus prudentes que la modélisation linéaire d'un point de vue global comme individuel.

4. Résumé et comparaison

Nous obtenons finalement deux estimations de la charge ultime globale. Elles ont l'avantage de prendre en considération l'information dossier/dossier. Ces estimations sont plus précises que la méthode Chain Ladder.

Le *modèle 1 (lm)* est le modèle utilisant le modèle linéaire multiple pour les sinistres « Small » et « Large ». Et le *modèle 1 (Pareto)* est le modèle utilisant le modèle linéaire multiple pour les sinistres « Small » et la loi de Pareto pour les sinistres « Large ».

	Modèle 1 (lm)	Ecart	Modèle 1 (Pareto)	Ecart	Chain Ladder
Total	291 633 787	**-3,0%**	357 403 054	**12,4%**	304 475 182

Les deux modèles se distinguent par leurs écarts par rapport au benchmark ou leur prudence. Comme attendu, le modèle de Pareto est plus prudent que les deux autres modèles. Cependant, une analyse plus poussée montre qu'une méthode de Chain-Ladder simple sous-estime pour nos données la première année de notification des sinistres en particulier sur les larges. Ainsi, l'estimation par la méthode de Pareto, bien que prudente, serait plus proche d'une estimation Best Estimate.

VI. Modèle 2 : modélisation dossier/dossier et phénomène de censure

Nous utilisons de nouveau l'historique dossier/dossier du passé et non pas une agrégation des montants à travers les triangles de liquidations. Ce deuxième modèle est également un modèle de régression, dont l'originalité est la prise en considération du phénomène de censure existante, grâce au modèle de Cox. En effet, le dernier montant de charge des sinistres ouverts est connu. Cette information n'a donc pas encore été utilisée lors du premier modèle puisque nous utilisions uniquement les sinistres clos pour estimer les coefficients. Pourtant, le montant de la charge 2009 peut être considéré comme une donnée censurée dans le cas où le sinistre est ouvert. Les sinistres ouverts apportent une information supplémentaire pour estimer leur propre charge ultime.

Les sinistres « Small » et « Large » sont modélisés de manière séparée. Puisque dans le premier modèle, les charges ultimes des sinistres ouverts étaient prédites comme non nulles, nous les avons

éliminées directement car l'application est la même. Une régression est alors appliquée à la fonction de survie des montants 2009 qui n'est pas une variable décrivant le temps mais le principe est exactement le même et la variable de censure D_i est la suivante :

$$D_i = \begin{cases} 1 & si\ le\ sinistre\ est\ clos\ fin\ 2009 \\ 0 & si\ le\ sinistre\ est\ ouvert\ fin\ 2009 \end{cases}$$

1. *Modélisation des « Small »*

L'échantillon est composé de 2013 sinistres dont 377 sinistres censurés. Le modèle sélectionné finalement régresse la fonction de survie de la variable *inc2009* en fonction de trois variables explicatives :

```
coxph(formula = Surv(coxsmall$inc2009, censures, type = "right") ~
    yon + dev_est + inc2008, data = coxsmall)
```

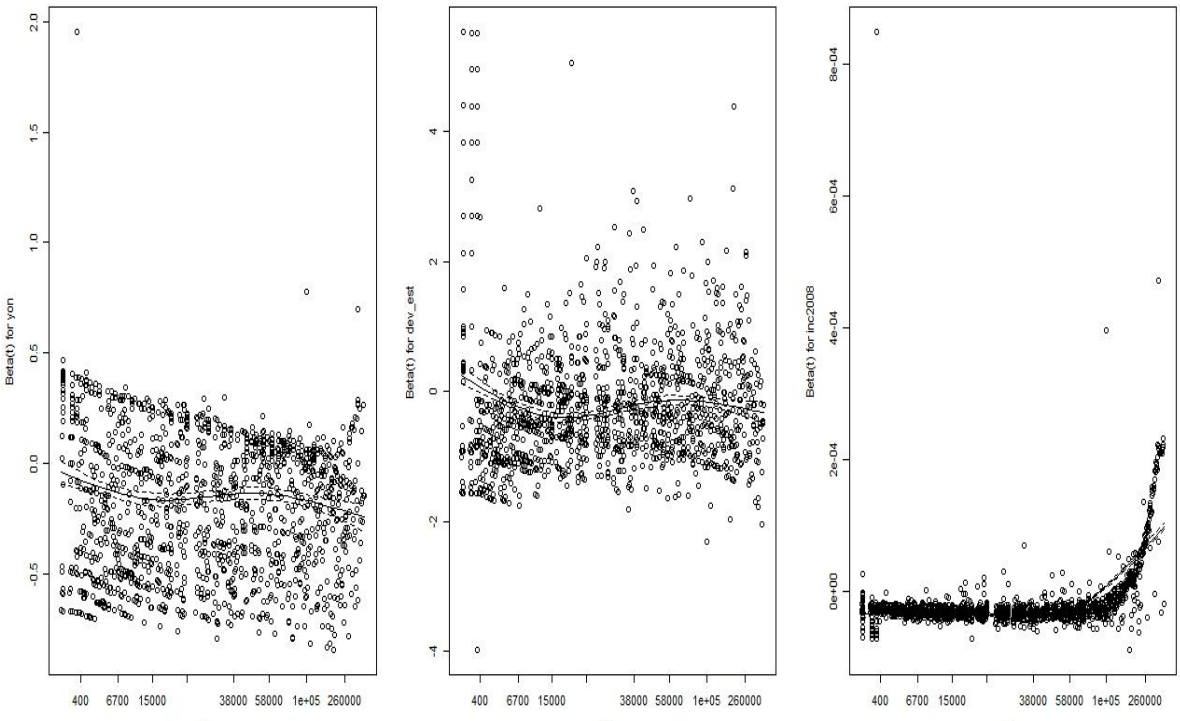
Les tests présentés auparavant sont donc effectués afin de valider le modèle :

- Au seuil 5%, les trois variables explicatives sont significatives.
- On teste l'hypothèse de proportionnalité par un test du chi-deux :

```
> tests=cox.zph(coxs,transform="km",global=TRUE)
> print(tests)
             rho  chisq        p
yon      -0.0888  12.88 0.000333
dev_est  -0.0368   3.32 0.068238
inc2008   0.3822 776.52 0.000000
GLOBAL        NA 881.09 0.000000
```

Au niveau marginal, la p-value « p » du test du chi-deux nous permet d'accepter l'hypothèse de risque proportionnel pour les deux variables explicatives *yon* et *inc2008* mais pas pour celle décrivant la durée de développement des sinistres *dev_est* ($0{,}068 > 0{,}05$). Cependant, l'hypothèse est acceptée pour le modèle global, ce qui est le plus important. L'hypothèse de proportionnalité est donc considérée comme validée pour ce modèle.

- L'analyse des résidus de Schoenfeld est utile pour étudier la validation de l'hypothèse de proportionnalité pour chaque variable marginale. Comme pour les modèles linéaires, la conclusion du test du Chi-deux global prime sur cette analyse, mais ellepermet tout de même d'étudier de manière plus fine les variables choisies.



En raison d'un grand nombre de résidus, la lecture graphique n'est pas évidente. Cependant, on peut considérer que les droites lissées ont une pente nulle pour les variables *yon* (l'année de notification) et *dev_est* (le nombre d'années de développement). Pour la variable décrivant les montants de l'année 2008, *inc2008*, on observe une pente nulle sur les plus faibles montants. Mais à partir de £100 000, on peut observer une cassure et la droite devient strictement croissante possédant une pente importante. La variable aurait donc un effet croissant sur les montants importants, cela parait cohérent de penser que plus la charge 2008 est grande, plus son influence est importante. Ce graphe nous guiderait sur la piste d'un nouveau seuil autour de £100 000 pour améliorer ce modèle.

En conclusion, les coefficients sont significatifs au seuil 5%, le test du Chi-deux pour le modèle global ne rejette pas l'hypothèse de risque proportionnel, il est donc considéré comme un modèle acceptable.

L'inflation n'est pas prise en compte implicitement dans ce modèle donc elle est appliquée de manière identique au premier modèle ainsi que l'estimation globale des IBNyR.

Voici les résultats :

	Montants (£)	Proportion (%)
Charge ultime des sinistres clos	89 097 711	56,4%
Prévisions de la charge ultime des ouverts	52 854 369	33,5%
Inflation	15 532 140	9,8%
IBNyR	478 352	0.3%
Charge ultime totale	157 484 220	100,0%

2. *Modélisation des « Large »*

L'échantillon est composé de 78 sinistres dont 15 sinistres censurés. La taille de l'échantillon est très faible, nous sommes donc limités dans le nombre de variables explicatives à tester. On obtient comme modèle optimal la fonction de survie de la variable *inc2009* en fonction de trois variables explicatives :

```
coxph(formula = Surv(coxlarge$inc2009, censurel, type = "right") ~
    yon + dev_est + inc2008, data = coxlarge)
```

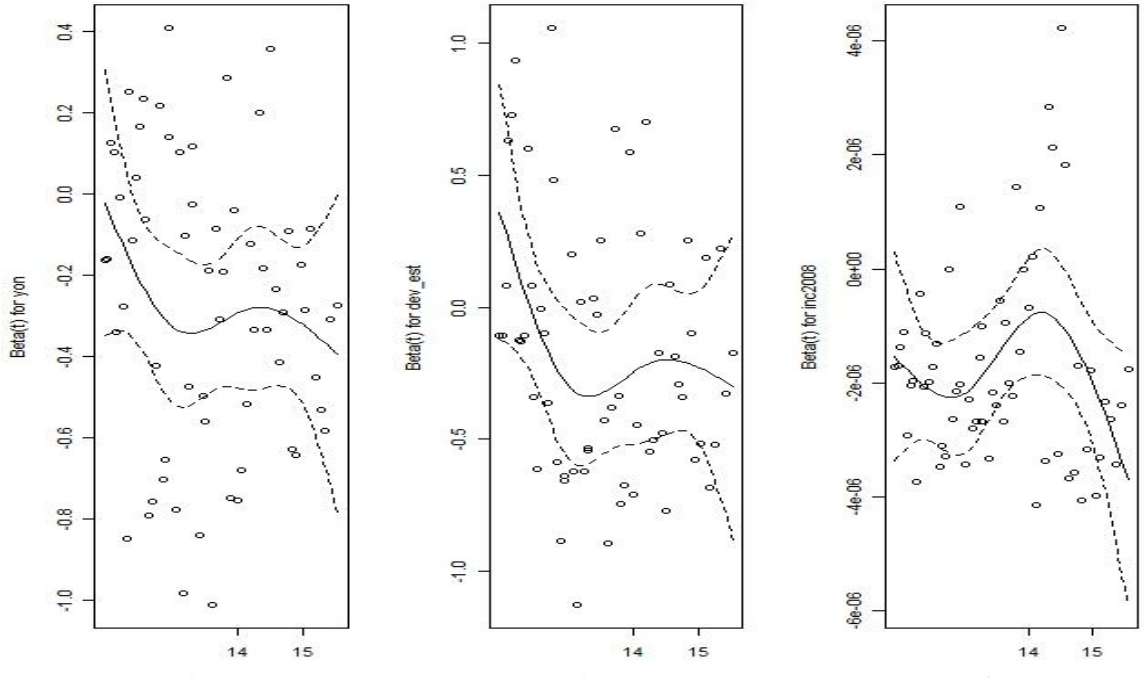
Les tests usuels sont effectués pour confirmer ou infirmer que le modèle est satisfaisant.

- Au seuil 5%, les trois variables explicatives sont significatives.
- On teste l'hypothèse de proportionnalité par un test du chi-deux :

```
> print(testl)
             rho    chisq     p
yon      -0.16701 1.31902 0.251
dev_est  -0.22029 1.97764 0.160
inc2008   0.00813 0.00242 0.961
GLOBAL         NA 2.61277 0.455
```

L'hypothèse de proportionnalité est rejetée que ce soit au niveau marginal ou global.

- L'analyse graphique des résidus de Schoenfeld confirme ce rejet, car les courbes lissées ne représentent aucunement des droites de pente nulle.



Par conséquent, le modèle de Cox sur les sinistres « Large » ne peut pas être utilisé. Nous sommes donc obligés d'agréger les deux catégories « Small » et « Large » afin d'avoir un échantillon de taille suffisante. Finalement, nous avons opté pour l'application du modèle de Cox sur l'ensemble des données.

3. *Modèle global*

On procède de manière identique pour traiter la base de données globale. L'échantillon est composé de 2 101 sinistres dont 402 sinistres censurés. Le modèle optimal sélectionné régresse la fonction de survie de la variable *inc2009* en fonction de quatre variables explicatives :

```
coxph(formula = Surv(cox$inc2009, censure, type = "right") ~
    yon + dev_est + inc2008 + large, data = cox)
```

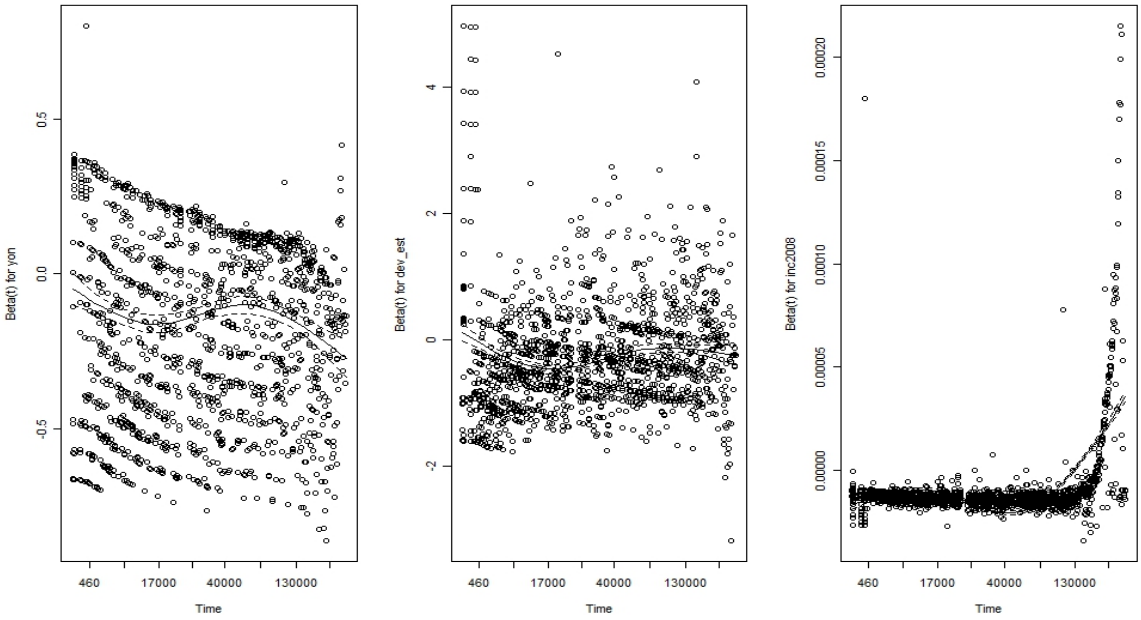
Les tests présentés auparavant sont donc effectués afin de valider le modèle :

- Au seuil 5%, les quatre variables explicatives sont significatives.
- On teste l'hypothèse de proportionnalité par un test du chi-deux :

```
> print(test)
             rho    chisq      p
yon      -0.0716    8.072 0.0045
dev_est   0.0186    0.821 0.3648
inc2008   0.3202  457.239 0.0000
large         NA      NaN    NaN
GLOBAL        NA  517.541 0.0000
```

L'hypothèse de risque proportionnel n'est pas acceptée pour toutes, mais n'est pas rejetée au niveau global.

- L'analyse graphique des résidus de Schoenfeld :



Dû au grand nombre de résidus, l'analyse graphique n'est pas évidente comme précedemment pour le modèle sur les « Small ». C'est pourquoi, il est porté plus de crédit au test du Chi-deux qu'à l'analyse graphique. On retrouve, par ailleurs les mêmes tendances, et la même piste d'utiliser un seuil supplémentaire mais cette fois autour de £130 000 .

En conclusion, les coefficients sont significatifs, le test du Chi-deux ne rejette pas l'hypothèse de risque proportionnel pour le modèle global ; ce modèle a donc un ajustement acceptable.

4. Résumé et comparaison

Similairement au modèle sur « Small » et sur les « Large », l'inflation est à rajouter ainsi que les IBNyR.

	Montants (£)	Proportion (%)
Charge ultime des sinistres clos	179 093 659	52,8%
Prévisions de la charge ultime des ouverts	109 378 608	32,3%
Inflation	26 308 352	7,8%
IBNyR	24 249 453	7.2%
Charge ultime totale	**339 030 072**	100,0%

La prévision du modèle de Cox est supérieure de 11% à la charge ultime globale du Benchmark. Son utilisation est donc plus prudente qu'un modèle de Chain Ladder puisqu'on émet l'hypothèse que la charge ultime est toujours supérieure à la charge déjà connue. Cette hypothèse est erronée dans la réalité, car la charge peut diminuer au cours du temps. Pour autant, ce biais représente une prudence supplémentaire, ce qui est un avantage. Cette méthode possède également un degré de précision supérieur au premier en terme d'historique puisqu'elle prend en considération les informations des sinistres ouverts grâce à la censure.

Malheureusement, la pratique de ce modèle a rendu impossible la mise en place de la décomposition d'un modèle pour les « Small » et pour les « Large ». Il est donc envisageable, pour améliorer ces résultats et rendre possible cette décomposition, l'application d'un modèle différent prenant en compte la censure. De plus, l'analyse des résidus incite à essayer dans le futur la piste d'un seuil inférieur à celui utilisé lors de cette étude, ce qui permettrait d'obtenir une population de « Large » supérieure et donc un modèle satisfaisant sur les « Large ».

VII. Modèle 3 : GLM sur triangle de liquidations

L'approche initiale du premier modèle était l'application d'un modèle GLM sur l'historique dossier/dossier des sinistres au lieu d'agréger les données en triangles. En pratique, il n'a pas été possible d'utiliser les GLM gamma, poisson ou gaussien sur les données non agrégées pour des problèmes de convergences ou de support de loi. Nous avons donc déjà présenté deux méthodes qui tiraient avantage de l'utilisation de l'historique dossier/dossier. Elles ont pour avantage de traiter les spécificités de manière dissociée, cependant la mise en œuvre est beaucoup plus complexe que les méthodes déterministes classiques. Ce troisième modèle prend donc une direction différente. Il conserve la simplicité des triangles de liquidation, mais l'intérêt est de décrire la cadence de liquidation de manière plus précise que l'utilisation des coefficients multiplicateurs de la méthode Chain Ladder. Nous utiliserons donc les modèles régressifs GLM.

Dans ce contexte, les GLM s'appliquent sur des triangles d'incréments. Nous avons travaillé sur des triangles de paiements et de charges. Jusqu'à maintenant, nous avions uniquement travaillé avec les montants de charges car ils sont plus stables. Mais, les incréments de charges peuvent être négatifs sur certaines années. Cela est dû principalement à la politique prudente de provisionnement de MDU. C'est pourquoi les paiements nous garantissent des incréments positifs.

L'utilisation du GLM sur les triangles revient à expliquer la sinistralité selon les années de notification et les années de développement. Il n'y a pas de travail de tri sur les variables explicatives. Pour autant, il faut choisir la loi et la fonction de lien. Afin de pouvoir régresser la charge en fonction

des vecteurs explicatifs : $yon = (yon_1, \dots, yon_{11})$ et $dvpt = (dvpt_1, \dots, dvpt_{11})$, les triangles de liquidations sont transformés de la manière suivante:

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11
1998	9	1 406	2 450	2 278	2 715	1 560	1 121	4 409	344	734	3 540	20
1999	59	870	2 071	3 395	6 252	775	129	178	42	-	-	
2000	57	1 438	4 608	3 230	2 124	1 199	1 265	105	21	-		
2001	38	676	1 976	2 230	6 365	1 381	461	102	-			
2002	32	953	1 349	3 341	1 986	1 336	1 770	454				
2003	109	479	2 699	4 743	1 858	1 171	316					
2004	111	775	2 232	1 093	2 744	997						
2005	136	597	1 748	4 693	1 792							
2006	58	602	3 022	3 061								
2007	154	813	2 478									
2008	511	1 010										
2009	101											

yon	Dvpt	Loss
yon0	Dvpt0	9
yon1	Dvpt0	59
yon2	Dvpt0	57
yon3	Dvpt0	38
yon4	Dvpt0	32
yon5	Dvpt0	109
yon6	Dvpt0	111
yon7	Dvpt0	136
yon8	Dvpt0	58
yon9	Dvpt0	154
yon10	Dvpt0	511
yon11	Dvpt0	-
…	…	…
Yon11	Dvpt11	101

Les résultats des méthodes appliquées au triangle total issu de l'ensemble des sinistres sont ensuite présentés. Nous comparons de la même manière nos résultats à ceux d'une méthode de Chain Ladder classique appliquée toujours au même triangle total, la nature diffère : soit paiement, soit charge pour des soucis de cohérence. L'unité de tous les tableaux de cette partie est en millier de Livres Sterling.

1. *Modélisation à partir de triangle de paiement*

Puisque les incréments sont positifs, nous pouvons tester des lois telles que la loi de Poisson, ou Gamma, variantes d'une loi normale.

Il y a des incréments nuls donc la loi Gamma est incompatible. La loi Normale n'est pas la plus indiquée sur un support positif et plus encore peut prédire des paiements négatifs. Un paiement négatif n'a pas de sens à moins de prendre en compte l'existence de remboursement, ce qui est anecdotique pour ce portefeuille. En minimisant le critère d'AIC, la loi de Poisson est sélectionnée avec un coefficient de dispersion égal à 1 et la fonction de lien logarithmique.

```
glm(formula = Loss ~ yon + Dvpt, family = poisson(link = "log"),
    data = data)
```

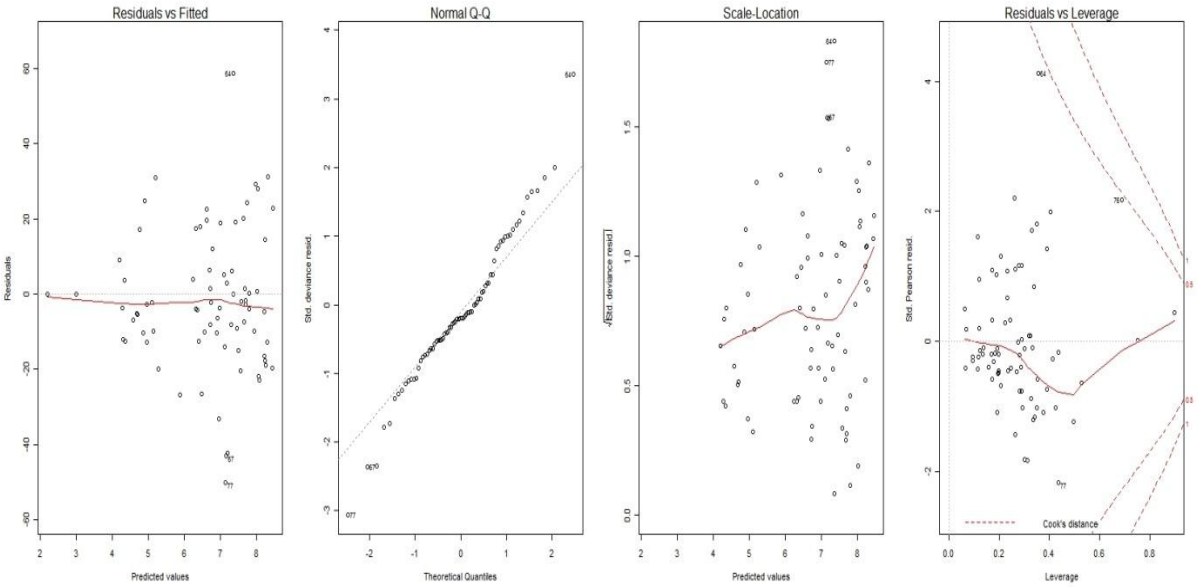
Pour valider ce modèle, il faut tester la significativité des résultats, ainsi que les propriétés des résidus.

- On effectue toujours le test du Chi-deux déterminant la significativité des variables explicatives.

```
> anova(glm_pois, test="Chisq")

     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    77     115551
yon  11    20415        66      95136 < 2.2e-16 ***
Dvpt 11    68805        55      26331 < 2.2e-16 ***
```

Les deux variables sont donc significatives au seuil 5%.

- Analyse graphique des résidus de Pearson



Le graphique de prédiction linéaire/résidus (1er) et le graphique des prévisions linéaires en fonction de la racine carrée des résidus standardisés (3ème) ne présentent aucune structuration. Ceci implique la validation des hypothèses d'homoscédasticité et de nullité de l'espérance des résidus. L'analyse du Q-Q plot (2ème) représente bien une droite d'angle 45°, ce qui valide donc l'hypothèse de normalité des résidus. D'après l'analyse graphique, les résidus ont l'air d'avoir les propriétés souhaitées. Pour finir, le graphique résidus/poids (4ème) n'indique pas la présence de valeurs aberrantes. Le modèle final est donc satisfaisant au vu de tous les critères.

On obtient les charges futures suivantes dans la partie inférieure du triangle :

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	Utimate
1998	9	1,406	2,450	2,278	2,715	1,560	1,121	4,409	344	734	3,540	20	20,679
1999	59	870	2,071	3,395	6,252	775	129	178	42	-	-	46	13,879
2000	57	1,438	4,608	3,230	2,124	1,199	1,265	105	21	-	1,856	29	16,007
2001	38	676	1,976	2,230	6,365	1,381	461	102	-	262	1,675	27	15,262
2002	32	953	1,349	3,341	1,986	1,336	1,770	454	138	346	2,209	35	14,010
2003	109	479	2,699	4,743	1,858	1,171	316	1,363	151	379	2,420	38	15,798
2004	111	775	2,232	1,093	2,744	997	620	733	81	204	1,301	21	10,961
2005	136	597	1,748	4,693	1,792	1,794	1,325	1,565	173	435	2,778	44	17,156
2006	58	602	3,022	3,061	3,131	1,193	881	1,041	115	289	1,847	29	15,337
2007	154	813	2,478	1,906	1,948	742	548	647	72	180	1,149	18	10,703
2008	511	1,010	2,986	3,776	3,858	1,470	1,085	1,283	142	356	2,276	36	18,874
2009	101	561	1,658	2,096	2,142	816	603	712	79	198	1,264	20	10,297

Pour les mêmes raisons que dans le modèle de Chain Ladder, un facteur de queue de distribution est appliqué. Les résultats sont donc homogènes. La charge ultime prédite est donc de **£ 184 247 394,** soit un écart de 2.9% avec le benchmark. Ce modèle est, par contre, moins prudent que le modèle de référence.

Par ailleurs, nous remarquons que la charge ultime prédite à partir des paiements ou des charges présentent un grand écart. La méthode de Chain Ladder appliquée sur le triangle total des charges prévoit une charge ultime globale de £304M contre seulement £184M pour les paiements, soit un écart de 39%. En théorie, les méthodes de provisionnement, qu'elles soient appliquées sur des montants de paiements ou de charges, devraient donner la même estimation de la charge ultime globale. Sur les branches à développement court comme le dommage auto, les écarts ne sont pas significatifs mais peuvent être plus importants sur les branches à développement long telles que MDU. En effet, les sinistres les plus larges sont réglés plus lentement et peuvent entraîner une sous-estimation de l'ultime si le facteur de queue projeté est insuffisant.

2. *Modélisation à partir de triangle de charge*

Comme il a été expliqué précédemment, le triangle de charge possède des incréments négatifs. Il est donc nécessaire d'ajuster une loi continue à support dans $\mathbb{R}$. Le modèle final sélectionné, minimisant l'AIC, est défini par une loi Normale et la fonction de lien identité. C'est donc l'application d'un modèle linéaire multiple classique.

```
glm(formula = Loss ~ yon + Dvpt, family = gaussian(link = "identity"),
    data = data)
```

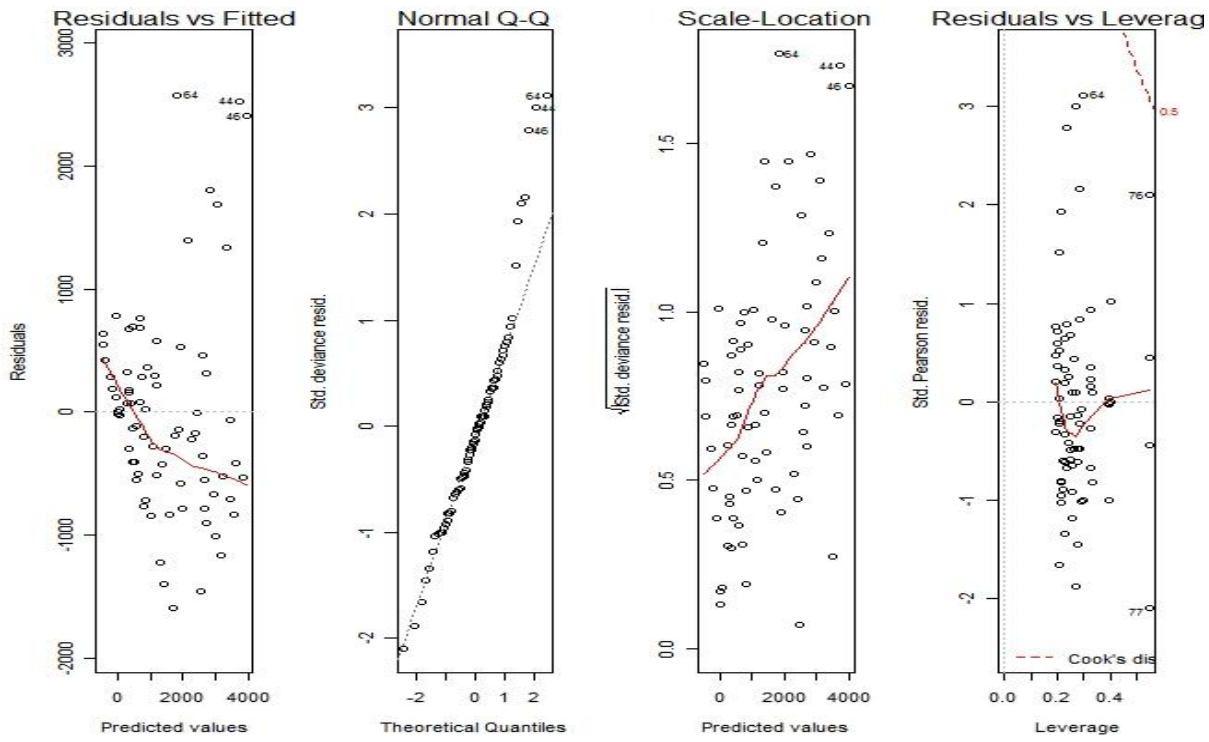
Pour valider ce modèle, on teste la significativité des résultats, ainsi que les propriétés des résidus.

- On effectue toujours le test du chi-deux déterminant la significativité des variables explicatives :

```
> anova(glm_gaus, test="Chisq")

     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    77  172019479
yon  11 28234155        66  143785324  0.002231 **
Dvpt 11 90327876        55   53457448 4.426e-15 ***
```

Les deux variables sont donc significatives au seuil 5%.

- Nous vérifions que l'ajustement du modèle est satisfaisant par l'analyse des résidus



L'analyse usuelle des graphiques permet de conclure à l'espérance nulle et l'homoscédasticité (aucune structuration) ainsi que la normalité des résidus (bissectrice du QQ-plot)

Les charges futures sont obtenues dans la partie inférieure du triangle :

	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	Ultimate
1998	6,544	5,355	839	1,082	2,548	-771	3,250	-856	1,797	714	305	-20	24,611
1999	11,693	2,434	-101	-326	1,862	-1,104	49	-449	-84	-	-	-152	16,545
2000	8,746	8,652	1,766	-557	-1,420	-2,534	-305	-163	-58	-76	-2,258	-1,461	16,635
2001	11,459	3,268	2,694	-4,155	410	-363	76	42	-55	420	883	1,680	19,369
2002	10,015	2,144	124	-44	-614	1,663	1,636	90	-139	-381	83	879	18,916
2003	12,703	7,597	-2,029	-850	-1,024	-1,696	-281	-531	174	-68	395	1,192	19,159
2004	11,495	7,678	-1,053	873	-4,810	395	176	-942	-237	-479	-15	781	18,392
2005	18,766	4,367	-3,598	-433	1,208	-346	1,000	-117	587	345	809	1,606	29,194
2006	16,334	7,292	1,505	-2,061	-242	-698	648	-469	236	-6	457	1,254	30,388
2007	25,770	1,912	709	489	1,024	569	1,914	797	1,502	1,260	1,724	2,520	47,584
2008	28,220	5,369	969	-113	422	-33	1,313	196	900	658	1,122	1,919	48,646
2009	28,926	4,150	-970	-2,051	-1,517	-1,972	-626	-1,743	-1,038	-1,280	-817	-20	39,160

De manière similaire, un facteur de queue de distribution a été utilisé. Finalement une estimation de la charge ultime de **£ 328 600 571** est obtenue, soit un écart avec la prévisions du Benchmark de 7,9%. Ce modèle est donc plus prudent que la méthode de Chain Ladder.

3. Résumé et comparaison

Ce modèle a été proposé dans l'optique d'améliorer la précision d'un modèle de Chain Ladder tout en étant plus facile à mettre en œuvre. Il qui perd donc l'avantage du traitement dossier/dossier des sinistres, mais permet une meilleure modélisation des coefficients de passage par rapport à de simple facteurs multiplicatifs. Le modèle des paiements présentent de meilleurs résultats que le Chain Ladder associé au triangle des paiements, mais l'application sur les paiements n'est pas stable pour ce type de branche. Ceci nous a conduit à utiliser cette méthode sur le triangle de charge malgré les incréments négatifs. Finalement, cette méthode reste une piste intéressante à approfondir, car elle représente clairement un bon compromis entre précision et complexité. C'est un modèle intermédiaire. On peut imaginer l'application d'une autre loi à support réel ou un lissage des valeurs négatives prédites, si elles sont trop importantes.

Partie 4 : Analyse et conclusion

L'intérêt fondamental de ce mémoire est de proposer des méthodes alternatives aux méthodes classiques utilisées pour une branche à développement long telle que la RC médicale afin de mieux en appréhender ses spécificités et difficultés.

- Une non-homogénéité des comportements du portefeuille. En effet, le temps de développement des sinistres est long. Plus long est le développement, plus les sinistres sont susceptibles d'avoir des comportements particuliers. Le caractère obligatoire de la RC n'améliore pas non plus l'homogénéité des comportements, car cela symbolise une non-sélection du risque et donc implique une non-homogénéité supplémentaire. C'est donc la difficulté principale à modéliser. L'importance de mieux prendre en compte toute l'information disponible est donc facilement compréhensible car au sein du même portefeuille le comportement des sinistres peut varier et toute information est à prendre en considération.
- La dispersion des montants. Dans le portefeuille, un écart supérieur à £5M est observé entre le sinistre le plus faible et le plus élevé. Hormis la difficulté de déterminer des seuils pertinents, nous sommes limités par la faible proportion de sinistres graves. Certaines décompositions peuvent entraîner un volume trop faible sur les tranches supérieures. Un échantillon de taille convenable est nécessaire pour déterminer un modèle stable.
- Les sinistres zéro-inflatés. La gestion prudente de MDU amène un historique extrêmement précis avec de nombreux sinistres clos sans aucun paiement. Cette problématique a été traitée à l'aide d'un GLM binomial.

En conclusion, l'enjeu est d'arriver à extraire de manière la plus précise possible une ou des tendances qui permettraient de mieux estimer les réserves et d'obtenir une meilleure adéquation avec les données historiques. L'assureur doit décider à quel niveau il souhaite se placer entre la précision de l'estimation obtenue et la praticabilité de la mise en œuvre. En effet, des modèles plus sophistiqués comme ceux présentés dans cette étude ont des avantages comme celui de la précision mais ils ont l'inconvénient principal d'avoir une mise en œuvre plus compliquée et plus couteuse en temps.

Dans ce mémoire, les estimations sont comparées à celles de la méthode de Chain Ladder car elle est très utilisée par les assureurs du fait de sa simplicité de mise en œuvre. Elle permet également de traiter de manière implicite les IBNyR et l'inflation. Toutefois, elle n'est pas nécessairement la méthode la mieux adaptée à ce portefeuille. Par exemple, l'hypothèse de stabilité des cadences est très forte et dans le cas particulier où les sinistres peuvent mettre plus de vingt ans à se clôturer, des changements de jurisprudence ou de gestion de sinistres peuvent l'invalider complètement. Alors que ces changements sont implicitement pris en compte dans la variation des coefficients de la régression.

De plus, l'incertitude augmente avec les années de survenance. En effet, le coefficient de la dernière année est le produit des précédents facteurs de développement. L'incertitude est d'autant plus grande lorsque l'on se situe sur des branches à développement long, cet écueil est également évité avec les modèles régressifs car les coefficients ne dépendent pas des uns des autres

L'agrégation des données dans les triangles de liquidation représente une perte en information importante pour ce type de branche. Les modèles 1 et 2 y pallient en utilisant un historique dossier/ dossier. Ils sont plus précis dans leurs estimations qu'un Chain Ladder.

Le premier modèle a l'avantage de traiter séparément les sinistres « Small » et « Large ». La difficulté de ce dernier se situe au niveau de la modélisation des « Large » et le degré de prudence que l'on souhaite leur appliquer. En effet, leur dispersion et leur petit nombre en volume imposent des compromis sur la manière de les traiter et sur la qualité des résultats obtenus. L'inconvénient est que les « Large » représentent le risque principal dans ce genre de branche du fait de leur montant et de leur volatilité. C'est pourquoi ce modèle se différencie, il permet de les traiter à part et de manière satisfaisante. Le modèle linéaire semble sous-estimer un peu les « Large » alors que la modélisation par la loi de Pareto semble meilleure car plus prudente.

Le deuxième modèle possède un degré de précision supérieur au premier en terme d'historique puisqu'il prend en considération les informations des sinistres ouverts grâce à la censure. Cependant, le modèle de Cox possède deux inconvénients. Il implique que la charge future sera supérieure à celle déjà connue, ce qui n'est pas le cas nécessairement. Cette hypothèse inflige un biais aux estimations. Pour autant ce biais peut être assimilé à un degré de prudence supplémentaire. Par contre, la pratique a rendu impossible la mise en place de la décomposition d'un modèle pour les « Small » et pour les « Large ». Peut-être que l'application d'un autre modèle prenant en compte la censure permettra cette décomposition qui semble primordiale sur ce portefeuille. De plus, explorer la piste d'un seuil inférieur à celui utilisé lors de cette étude pourrait éventuellement permettre d'obtenir une population de « Large » supérieure et donc un modèle adéquat.

Finalement, ces deux modèles ont rempli leur objectif à savoir être plus précis qu'un modèle de Chain Ladder. Cependant, au niveau de la mise en œuvre, ils sont plus complexes et plus couteux en temps. En effet, la sélection du modèle optimal et sa validation ne sont pas des étapes évidentes à automatiser.

C'est pourquoi nous avons proposé la piste du troisième modèle. Il perd l'avantage du traitement dossier/dossier des sinistres, mais permet une meilleure modélisation des coefficients de passage par rapport à de simple facteurs multiplicatifs. Nous obtenons sur le triangle des paiements et des charges des résultats intéressants. En effet, le modèle des paiements présentent des meilleurs résultats que le Chain Ladder appliqués au triangle des paiements, mais l'application sur les paiements n'est pas stable pour ce type de branche. C'est pourquoi nous avons essayé le traitement sur le triangle de

charge. Les incréments négatifs obligent à choisir la loi normale et donc à prévoir des incréments négatifs. Malgré tout, cette méthode reste une piste intéressante à approfondir, car elle aboutit à un bon compromis entre précision et complexité. C'est un modèle intermédiaire entre la méthode de Chain Ladder et nos deux premiers modèles. On peut imaginer l'application d'une autre loi à support réel ou un lissage des valeurs négatives prédites, si elles sont trop importantes.

Le tableau suivant résume l'ensemble des résultats des modèles retenus dans l'étude.

Modèles	Historique	Décomposition small/Large	Charge Ultime	Ecart
Modèle 1				
LM+Pareto	**Dossier/Dossier**	**oui**	357 403 054	12.4%
Benchmark (charge)	**triangle de liquidation**	**non**	304 475 182	
Modèle 2				
Cox	**Dossier/Dossier**	**non**	339 030 072	11.3%
Benchmark (charge)	**triangle de liquidation**	**non**	304 475 182	
Modèle 3				
GLM Gaussien	**triangle de liquidation**	**non**	328 600 571	7.9%
Benchmark(charge)	**triangle de liquidation**	**non**	304 475 182	

Pour conclure, l'utilisation de l'historique sinistre par sinistre et l'utilisation de variables qualitatives sur les montants de charges amènent à une estimation plus prudente que la méthode Chain Ladder. Ce qui indique que cette information supplémentaire non traitée initialement conduit à une prudence supplémentaire. Il en est de même pour le dernier modèle. L'utilisation de modèles régressifs sur ce type de portefeuille a un intérêt réel, qui sera accru par l'exploitation des nouvelles variables mise en place par les gestionnaires de sinistres.

Partie 5 : Bibliographie

Ouvrages :

Michel DENUIT, Arthur CHARPENTIER (2005), *Mathématiques de l'assurance non-vie*, Economica

Christian PARTRAT, Eric LECOEUR, Jean-Marie NESSI, Ecaterina NISIPASU, Olivier REIZ (2007), *Provisionnement Technique en assurance non-vie*, Economica

Régis BOURDAIS (2011), *Econométrie*, Dunod

Frédéric PLANCHET, Pierre THEROND (2005), *Modèle de durée : applications actuarielles*, Economica

Nelder McCULLAGH, J. A. Nelder , *Generalized Linear Models* ,Chapman and hall (second edition)

Articles :

Greg TAYLOR, Grainne McGUIRE, "Adaptive Reserving using Bayesian Revision for the Exponential Dispersion Family"

Greg TAYLOR, Grainne McGUIRE (2004), "Loss Reserving with GLMs: a case study"

James GUSZCZA (2008), "Hierarchical Models for loss reserving"

Sergio PEZZULLI, Simon MARGETTS (2008), "Granular Loss Modelling", ERNST &YOUNG

Artur CHARPENTIER (2009),"Provisionnement stochastique en assurance non-vie", CARITAT Recherche & Formation

Cédric HEUCHENNE (2007), *Strong uniform consistency results of the weighted average of conditional artificial data points*, Institut de Statistique de l'Université catholique de Louvain

Mémoires d'actuariat :

Noémie ROSE, «Provisionnement en assurance non-vie : Utilisation de modèles paramétriques censurés », ISUP 2009

Morgane FORT, « Méthodes de provisionnement en assurance non-vie et extrapolation des triangles », EURIA 2010

Lise HE, « Méthodes de provisionnement et analyse de la solvabilité d'un entreprise d'assurance non-vie », ENSAE 2004

Partie 6 : Annexe

- **Exemple de Bornhuetter Ferguson**

Supposons que le tarif ait été établi en anticipant une sinistralité totale de 80, que la proportion des sinistres connus au premier bilan soit habituellement de 30% et que la sinistralité constatée au premier bilan soit de 30.

Si l'on considère que le tarif était correctement établi, il n'y a pas lieu de remettre en cause la sinistralité totale et le montant des réserves à provisionner est de 80-30 = 50.

Si l'on considère que la proportion de sinistres connus au premier bilan est un indicateur fiable, il y a lieu de considérer que la sinistralité totale sera de 30/30% = 100 et de provisionner 70 de tardifs.

L'application de la méthode Bornhuetter-Ferguson donne un résultat intermédiaire calculé ainsi :

L = 100 x 30%, 80 x 70% = 86, soit un volume de tardifs attendu de 56.

- **Résultats finaux dans R**

La mise en pratique des modèles a été faite grâce au logiciel R. En raison de la confidentialité de ce mémoire, nous ne mettrons pas de partie de code mais les sorties principales obtenues pour donner au lecteur une vision plus précise du travail effectué.

Voici les variables explicatives potentielles que nous possédions et leurs notations:

- *Yon* indique l'année de notification.
- *dev_est* indique nombre d'années réelles ou estimé que le sinistre a mis pour se clore.
- *zero_paid* indique s'il n'y a eu aucun paiement auparavant.
- *Prec* indique si le sinistre a été au moins une fois Precautionnary dans son développement.
- *Clos_N-1* indique si le sinistre était clos l'année précédente.
- *Status* indique l'état de la plainte.
- *Spec* indique la spécialité du médecin.
- *incN* est le montant de la charge du sinistre pour l'année calendaire N.

A partir de ces variables, nous obtenons un modèle saturé, i.e. un modèle qui utilise le maximum de variables explicatives potentielles. Par la suite, à partir de ce modèle saturé, on obtient grâce à nos critères de choix, le modèle optimal.

I. Modèle 1

A. *GLM binomial*

Le modèle final, qui permet de traiter les sinistres zéro-inflatés, regrésse la variable binomiale Y_0 en fonction des variables explicatives suivantes :*yon ;dev_est ;zero_paid*

```
> summary(glm_binom3)

Call:
glm(formula = Y0 ~ dev_est + yon + zero_paid, family = binomial("cauchit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4672  -0.0147  -0.0107   0.0147   0.0538

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1326929.1  1357238.9  -0.978   <2e-16 ***
dev_est           679.5      721.5   0.942   <2e-16 ***
yon               663.9      678.9   0.978   <2e-16 ***
zero_paid       -7149.3     7170.3  -0.997   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4948.357  on 3580  degrees of freedom
Residual deviance:   21.955  on 3577  degrees of freedom
AIC: 29.955
```

B. *Modèle linéaire pour la modélisation des « Small »*

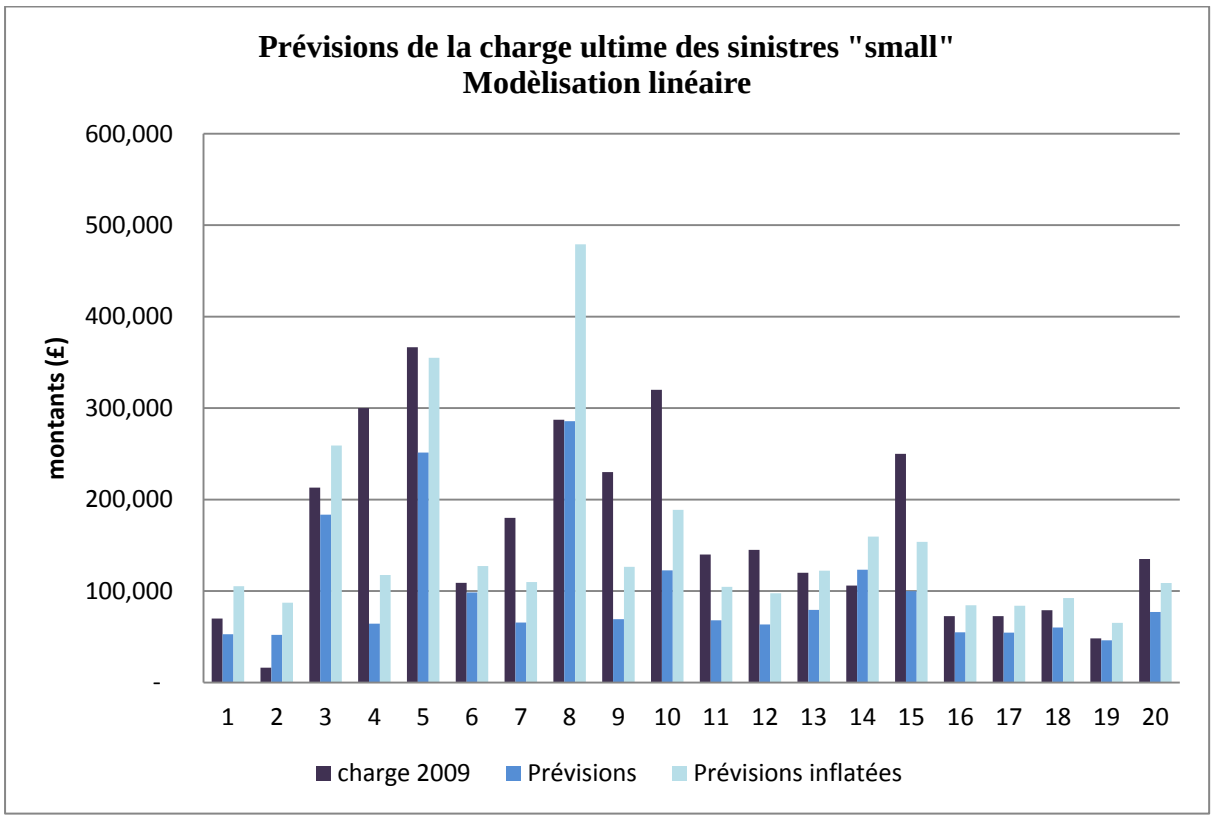
Le modèle saturé est composé des variables explicatives suivantes *: yon, clos_N_1, dev_est, zero_paid, Prec, inc2000, inc2001, inc2002, inc2003, inc2004, inc2005, inc2006, inc2007, inc2008.*

```
> summary(lms)

Call:
lm(formula = inc2009 ~ yon + clos_N_1 + dev_est + zero_paid +
    Prec + inc2000 + inc2001 + inc2002 + inc2003 + inc2004 +
    inc2005 + inc2006 + inc2007 + inc2008, data = clossmall)

Residuals:
    Min      1Q  Median      3Q     Max
-734560   -9437   -1297    3332  412229

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.915e+06  9.480e+05  -4.130 3.81e-05 ***
yon          1.968e+03  4.721e+02   4.168 3.23e-05 ***
clos_N_1    -1.788e+04  6.229e+03  -2.871 0.004148 **
dev_est     -5.738e+02  7.452e+02  -0.770 0.441387
zero_paid   -1.290e+04  5.899e+03  -2.186 0.028936 *
Prec         2.264e+03  3.253e+03   0.696 0.486423
inc2000      1.392e-02  4.064e-02   0.343 0.731895
inc2001      7.940e-02  3.761e-02   2.111 0.034896 *
inc2002     -9.599e-03  2.484e-02  -0.386 0.699248
inc2003      1.323e-02  2.825e-02   0.468 0.639674
inc2004      6.180e-02  3.060e-02   2.019 0.043623 *
inc2005      3.684e-01  3.327e-02  11.074  < 2e-16 ***
inc2006     -2.601e-01  3.583e-02  -7.259 6.08e-13 ***
inc2007      1.138e-01  3.063e-02   3.714 0.000211 ***
inc2008      6.256e-01  2.391e-02  26.167  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36480 on 1593 degrees of freedom
Multiple R-squared: 0.7793,     Adjusted R-squared: 0.7774
F-statistic: 401.9 on 14 and 1593 DF,  p-value: < 2.2e-16
```

Le modèle optimal sélectionné est le suivant :

```
> summary(lms2_int_2)

Call:
lm(formula = inc2009 ~ zero_paid + inc2008 + zero_paid:inc2008 +
    Prec:inc2008 + dev_est:inc2008 + dev_est:Prec + yon:dev_est,
    data = clossmall)

Residuals:
    Min      1Q  Median      3Q     Max
-629758   -5114   -1230    1678  411764

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.052e+03  2.527e+03  -2.394  0.01676 *
zero_paid          2.939e+04  2.145e+03  13.701  < 2e-16 ***
inc2008            1.170e+00  3.449e-02  33.915  < 2e-16 ***
zero_paid:inc2008 -5.238e-01  2.102e-02 -24.923  < 2e-16 ***
inc2008:Prec       2.391e-01  3.195e-02   7.484 1.18e-13 ***
inc2008:dev_est   -4.644e-02  7.336e-03  -6.331 3.16e-10 ***
Prec:dev_est      -3.820e+03  6.809e+02  -5.611 2.37e-08 ***
dev_est:yon        1.088e+00  3.854e-01   2.822  0.00483 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33620 on 1600 degrees of freedom
Multiple R-squared: 0.8118,     Adjusted R-squared: 0.811
F-statistic:   986 on 7 and 1600 DF,  p-value: < 2.2e-16
```

Ci-dessous à titre d'exemple, une représentation graphique de quelques prévisions de sinistres « Small » :



C. *Modèle linéaire pour la modélisation des « Large »*

1. Modèle linéaire

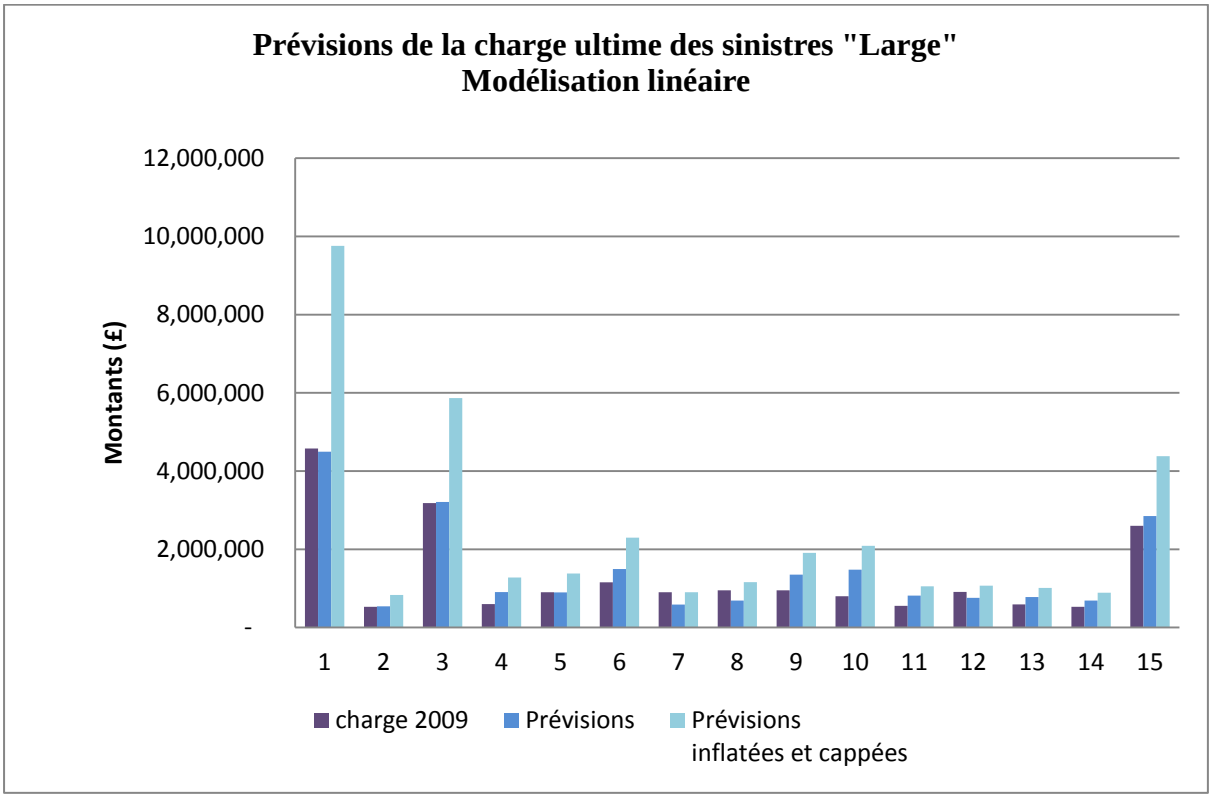
Le modèle saturé est composé des variables explicatives suivantes *: yon, clos_N_1, dev_est, zero_paid, Prec, inc2000, inc2001, inc2002, inc2003, inc2004, inc2005, inc2006, inc2007, inc2008*

```
> summary(lm1)

Call:
lm(formula = inc2009 ~ yon + clos_N_1 + dev_est + zero_paid +
    Prec + inc2000 + inc2001 + inc2002 + inc2003 + inc2004 +
    inc2005 + inc2006 + inc2007 + inc2008, data = closlarge)

Residuals:
    Min      1Q  Median      3Q     Max
-528500 -112177  -32105   66661 1033316

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.087e+07  8.701e+07  -0.125   0.9011
yon          5.430e+03  4.333e+04   0.125   0.9008
clos_N_1     2.326e+04  2.258e+05   0.103   0.9184
dev_est      1.812e+04  4.413e+04   0.411   0.6831
zero_paid    5.589e+05  1.991e+05   2.807   0.0072 **
Prec        -2.025e+05  1.449e+05  -1.398   0.1686
inc2000      1.888e-02  2.959e-01   0.064   0.9494
inc2001     -1.361e-01  1.999e-01  -0.681   0.4991
inc2002      2.766e-02  2.571e-01   0.108   0.9147
inc2003      6.367e-02  1.560e-01   0.408   0.6850
inc2004      1.101e-01  1.142e-01   0.964   0.3398
inc2005     -1.217e-01  1.381e-01  -0.882   0.3822
inc2006      5.139e-02  1.160e-01   0.443   0.6598
inc2007      2.982e-02  8.458e-02   0.353   0.7260
inc2008      8.509e-01  6.438e-02  13.216   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 349300 on 48 degrees of freedom
Multiple R-squared: 0.939,      Adjusted R-squared: 0.9212
F-statistic: 52.76 on 14 and 48 DF,  p-value: < 2.2e-16
```

Le modèle optimal obtenu utilise les variables explicatives suivantes :*yon, inc2008, yon:inc2008.*

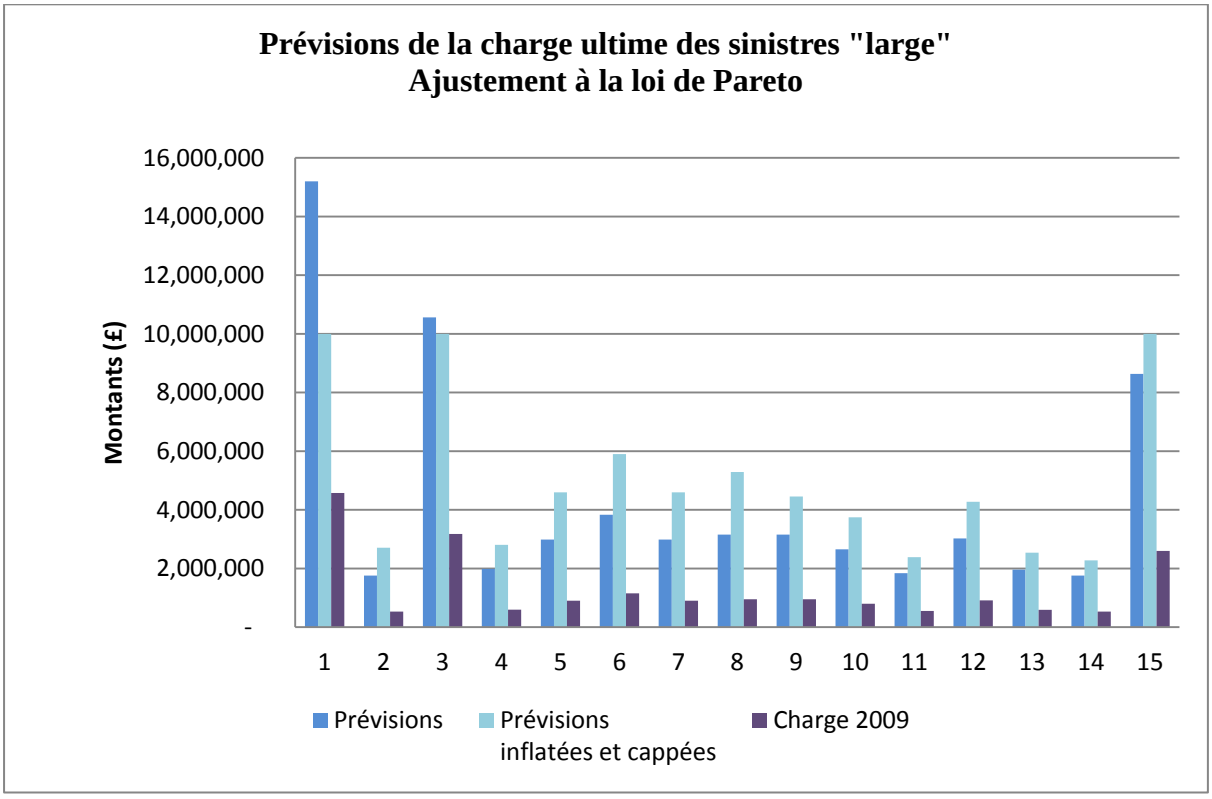
```
> summary(lm13_int)

Call:
lm(formula = inc2009 ~ yon + inc2008 + yon:inc2008, data = closlarge)

Residuals:
    Min      1Q  Median      3Q     Max
-544268 -176721   -9648   66866 1267668

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.610e+08  3.671e+07  -4.387 4.81e-05 ***
yon          8.052e+04  1.832e+04   4.395 4.68e-05 ***
inc2008      3.706e+01  1.782e+01   2.080   0.0419 *
yon:inc2008 -1.803e-02  8.895e-03  -2.027   0.0471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 350400 on 59 degrees of freedom
Multiple R-squared: 0.9245,     Adjusted R-squared: 0.9207
F-statistic: 240.9 on 3 and 59 DF,  p-value: < 2.2e-16
```

Ci-dessous la représentation graphique des prévisions obtenues grâce au modèle linéaire :



2. Ajustement de loi

Voici la représentation graphique des prévisions grâce à l'ajustement d'une loi Pareto:



Voici la sortie du logiciel @Risk, qui nous a permis de conclure que la loi de Pareto avait le meilleur ajustement :

@RISK Fit Results

Performed By: ENNADIFI Isa
Date: 20 July 2011 17:29:31

	Input	Pareto	Lognorm	InvGauss
Fit				
Function		RiskPareto(1.3011,505000)	RiskLognorm(1138857.2,3378665.8,RiskShift(497628))	RiskInvgauss(962261,280174,RiskShift(466245))
Distribution Statistics				
Minimum	505 000,00	505 000,00	497 628,73	466 245,47
Maximum	5 500 000,00	+Infinity	+Infinity	+Infinity
Mean	1 428 507,11	2 182 098,68	1 636 485,90	1 428 507,11
Mode	546329.0000 [est]	505 000,00	534 742,65	558 773,55
Median	840 000,00	860 308,93	861 397,26	835 681,86
Std. Deviation	1 244 215,84	+Infinity	3 378 665,79	1 783 303,48
Skewness	1,78	+Infinity	35,01	5,56
Kurtosis	5,31	+Infinity	11 397,36	54,52
Percentiles				
5%	541 987,00	525 306,02	527 938,43	531 088,16
10%	550 904,00	547 594,75	550 104,20	554 903,84
15%	564 250,00	572 186,98	573 624,00	578 036,29
20%	590 000,00	599 478,54	599 630,88	602 427,26
25%	628 750,00	629 963,98	628 930,28	629 139,96
30%	659 534,00	664 269,94	662 349,52	659 100,40
35%	703 500,00	703 203,18	700 866,06	693 316,37
40%	770 000,00	747 821,57	745 710,94	733 011,07
45%	780 000,00	799 541,84	798 495,70	779 760,91
50%	840 000,00	860 308,93	861 397,26	835 681,86
55%	867 500,00	932 872,58	937 449,49	903 714,11
60%	945 000,00	1 021 260,88	1 031 030,70	988 088,59
65%	1 184 845,00	1 131 638,33	1 148 727,31	1 095 139,00
70%	1 415 000,00	1 273 975,39	1 300 973,20	1 234 811,37
75%	1 850 000,00	1 465 606,83	1 505 442,69	1 423 707,23
80%	2 050 000,00	1 739 801,68	1 794 930,22	1 691 911,50
85%	2 595 000,00	2 170 321,58	2 238 888,95	2 100 803,67
90%	3 615 000,00	2 963 894,89	3 019 331,53	2 801 756,58
95%	4 395 000,00	5 049 238,08	4 863 475,75	4 335 600,85
Chi-Squared Test				
Chi-Sq Statistic		4,14	5,14	6,86
P-Value		0,84	0,74	0,55
Cr. Value @ 0.750		5,07	5,07	5,07
Cr. Value @ 0.500		7,34	7,34	7,34
Cr. Value @ 0.250		10,22	10,22	10,22
Cr. Value @ 0.150		12,03	12,03	12,03
Cr. Value @ 0.100		13,36	13,36	13,36
Cr. Value @ 0.050		15,51	15,51	15,51
Cr. Value @ 0.025		17,53	17,53	17,53
Cr. Value @ 0.010		20,09	20,09	20,09
Cr. Value @ 0.005		21,96	21,96	21,96
Cr. Value @ 0.001		26,12	26,12	26,12

II. Modèle 2

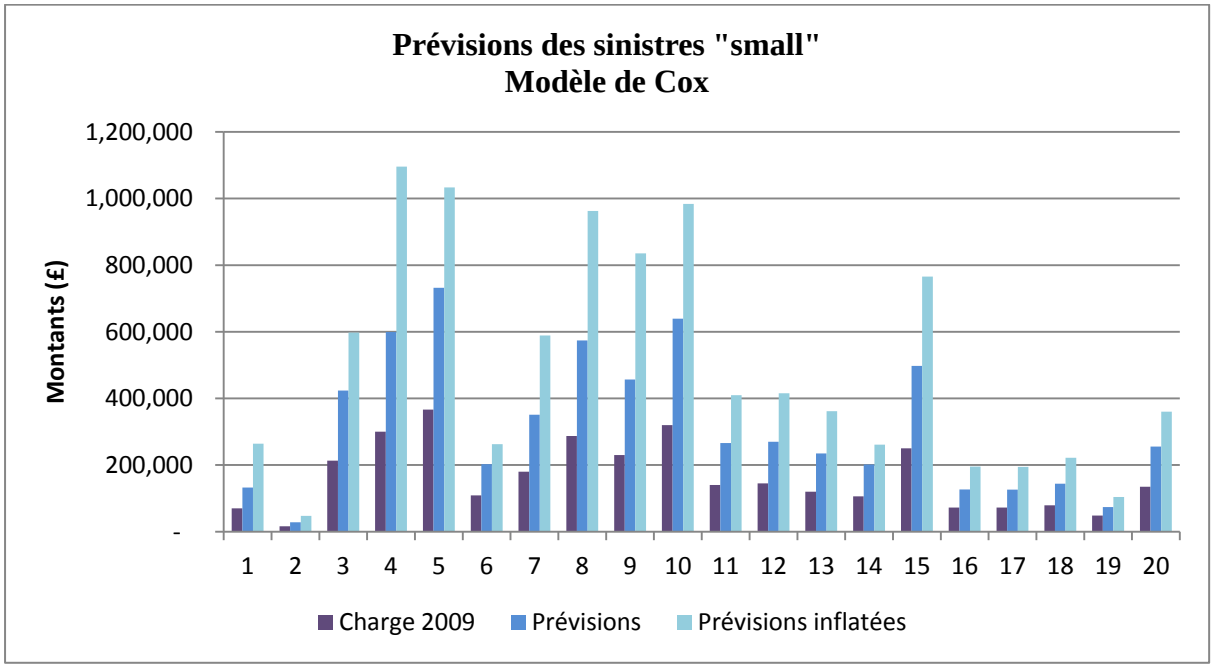
Pour les modèles de Cox, les modèles saturés ne sont pas composés de nombreuses variables explicatives donc nous présentons directement les modèles optimaux.

A. *Modèle sur les « Small »*

Voici les résultats pour le modèle optimal composé des variables explicatives suivantes *: yon ; dev_est ; inc2008.*

```
> summary(coxs)
Call:
coxph(formula = Surv(coxsmall$inc2009, censures, type = "right") ~
    yon + dev_est + inc2008, data = coxsmall)

  n= 1723, number of events= 1608

               coef  exp(coef)   se(coef)       z Pr(>|z|)
yon      -1.446e-01  8.654e-01  7.826e-03 -18.47   <2e-16 ***
dev_est  -2.087e-01  8.117e-01  2.000e-02 -10.43   <2e-16 ***
inc2008  -1.762e-05  1.000e+00  6.481e-07 -27.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
yon         0.8654       1.156     0.8522     0.8788
dev_est     0.8117       1.232     0.7805     0.8441
inc2008     1.0000       1.000     1.0000     1.0000

Concordance= 0.832   (se = 0.008 )
Rsquare= 0.637    (max possible= 1 )
Likelihood ratio test= 1746  on 3 df,   p=0
Wald test            = 1078  on 3 df,   p=0
Score (logrank) test = 912.4  on 3 df,   p=0
```

Voici la représentation graphique à titre d'exemple des charges ultimes prédites pour quelques sinistres « SMALL » :

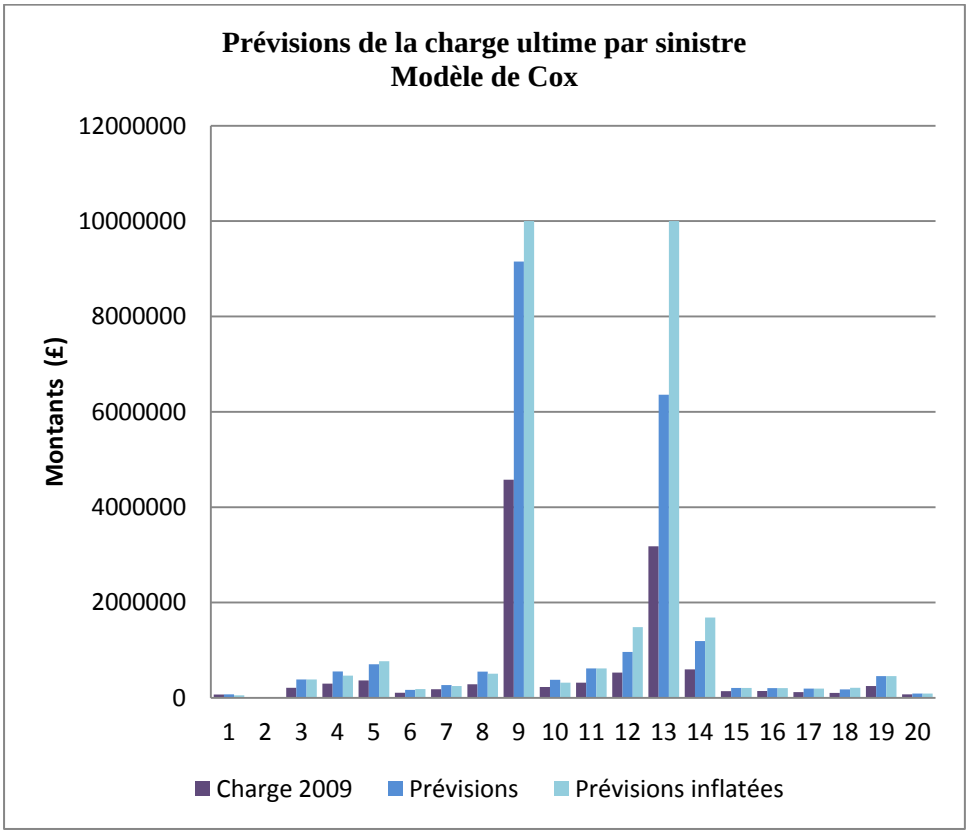


B. Modèle global

Le modèle optimal appliqué à l'ensemble de la base de données est expliqué en fonction des variables explicatives suivantes : *yon ;dev_est ;inc2008 ;large.*

```
> summary(coxgen)
Call:
coxph(formula = Surv(cox$inc2009, censure, type = "right") ~
    yon + dev_est + inc2008 + large, data = cox)

  n= 1801, number of events= 1671

              coef  exp(coef)   se(coef)       z Pr(>|z|)
yon      -1.376e-01  8.715e-01  7.551e-03 -18.22   <2e-16 ***
dev_est  -2.564e-01  7.738e-01  1.890e-02 -13.57   <2e-16 ***
inc2008  -9.663e-06  1.000e+00  2.984e-07 -32.38   <2e-16 ***
large    -2.616e+00  7.311e-02  0.000e+00   -Inf   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
yon        0.87148      1.147   0.85867   0.88447
dev_est    0.77382      1.292   0.74568   0.80302
inc2008    0.99999      1.000   0.99999   0.99999
large      0.07311     13.679   0.07311   0.07311

Concordance= 0.81   (se = 0.008 )
Rsquare= 0.622    (max possible= 1 )
Likelihood ratio test= 1754  on 4 df,    p=0
Wald test            = 1463  on 4 df,    p=0
Score (logrank) test = 628.3  on 4 df,    p=0
```

Voici la représentation graphique des charges ultimes finales prédites pour quelques sinistres :



III. Modèle 3

A. *Modèle à partir de triangle de paiement*

Voici les coefficients du GLM de Poisson mis en place sur le triangle de paiements :

```
> summary(glm_pois)

Call:
glm(formula = Loss ~ yon + Dvpt, family = poisson(link = "log"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-50.115  -12.390   -3.778    6.299   58.616

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.19722    0.33333   6.592 4.35e-11 ***
yonyon1      2.94016    0.33540   8.766  < 2e-16 ***
yonyon10     2.69080    0.33462   8.041 8.89e-16 ***
yonyon11     2.09248    0.33469   6.252 4.05e-10 ***
yonyon2      2.49014    0.33494   7.435 1.05e-13 ***
yonyon3      2.38565    0.33477   7.126 1.03e-12 ***
yonyon4      2.66177    0.33466   7.954 1.81e-15 ***
yonyon5      2.75281    0.33464   8.226  < 2e-16 ***
yonyon6      2.13195    0.33471   6.369 1.90e-10 ***
yonyon7      2.89052    0.33461   8.638  < 2e-16 ***
yonyon8      2.48243    0.33465   7.418 1.19e-13 ***
yonyon9      2.00774    0.33472   5.998 1.99e-09 ***
DvptDvpt1    2.02867    0.02991  67.831  < 2e-16 ***
DvptDvpt10   2.84578    0.03354  84.847  < 2e-16 ***
DvptDvpt11  -1.29397    0.22562  -5.735 9.75e-09 ***
DvptDvpt2    3.11358    0.02908 107.085  < 2e-16 ***
DvptDvpt3    3.34842    0.02905 115.280  < 2e-16 ***
DvptDvpt4    3.37014    0.02913 115.689  < 2e-16 ***
DvptDvpt5    2.40562    0.03051  78.848  < 2e-16 ***
DvptDvpt6    2.10255    0.03182  66.073  < 2e-16 ***
DvptDvpt7    2.26958    0.03174  71.495  < 2e-16 ***
DvptDvpt8    0.06762    0.05728   1.181    0.238
DvptDvpt9    0.99031    0.04685  21.138  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 115551  on 77  degrees of freedom
Residual deviance:  26331  on 55  degrees of freedom
AIC: 26992
```

B. Modèle à partir de triangle de charge

Voici les coefficients du GLM gaussien appliquée sur les triangles de liquidations des charges :

```
>  summary(glm_gaus)

Call:
glm(formula = Loss ~ yon + Dvpt, family = gaussian(link = "identity"),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1585.46   -531.22    -43.21    321.42   2572.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.00     985.88   0.009   0.9927
yonyon1       343.77    1225.61   0.280   0.7802
yonyon10      322.05    1076.62   0.299   0.7660
yonyon11     -430.59    1076.62  -0.400   0.6907
yonyon2        72.82    1164.61   0.063   0.9504
yonyon3       -92.12    1133.28  -0.081   0.9355
yonyon4       421.64    1114.36   0.378   0.7066
yonyon5       603.25    1101.83   0.547   0.5863
yonyon6      -449.39    1093.04  -0.411   0.6826
yonyon7       840.23    1086.67   0.773   0.4427
yonyon8        23.06    1082.00   0.021   0.9831
yonyon9      -488.72    1078.67  -0.453   0.6523
DvptDvpt1     759.46     420.38   1.807   0.0763 .
DvptDvpt10   1815.27     793.07   2.289   0.0259 *
DvptDvpt11    441.59    1076.62   0.410   0.6833
DvptDvpt2    2372.08     435.86   5.442 1.26e-06 ***
DvptDvpt3    3025.95     452.07   6.693 1.20e-08 ***
DvptDvpt4    3115.31     470.37   6.623 1.56e-08 ***
DvptDvpt5    1133.73     492.01   2.304   0.0250 *
DvptDvpt6     865.23     518.72   1.668   0.1010
DvptDvpt7     987.39     553.26   1.785   0.0798 .
DvptDvpt8     236.30     600.64   0.393   0.6955
DvptDvpt9     434.75     671.39   0.648   0.5200
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 971953.6)

    Null deviance: 172019479  on 77  degrees of freedom
Residual deviance:  53457448  on 55  degrees of freedom
AIC: 1317.5

Number of Fisher Scoring iterations: 2
```