Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Pierre Hénin

Sujet : Un modèle de provisionnement ligne à ligne en assurance Responsabilité Civile

Confidentialité : ☑ NON ☐OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.

Membres présents du jury de l'Institut des Actuaires

Membres présents du jury de la filière

Entreprise :

Nom : Gras Savoye

Signature :

Directeur de mémoire en entreprise :
Nom : Lionel MORAIS ALVES

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

Secrétariat

Bibliothèque :

Résumé

Mots-clés : Provisionnement, modèle individuel, GLM, méthodes stochastiques

L'activité d'assurance de caractérise par son cycle de production inversé. Cependant, les comptes de l'assureur doivent représenter le coût ultime de ses sinistres, c'est-à-dire les règlements passés et les règlements futurs des sinistres survenus jusqu'à la date d'établissement des comptes. Il doit donc tenir compte des sinistres connus ainsi que des sinistres tardifs.

Pour calculer le coût total des sinistres, les méthodes les plus utilisées se basent sur l'agrégation des règlements antérieurs par année de survenance et par année de développement, sous forme de triangle de liquidation. La méthode déterministe la plus courante, la méthode de Chain Ladder, permet d'étudier les cadences de règlement des sinistres d'une année de développement à l'autre. Elle est cependant sensible à des données exogènes, comme l'évolution de la politique de gestion des sinistres ou une évolution jurisprudentielle. Pour cela, des méthodes stochastiques ont été développées afin d'estimer l'incertitude causée par ces variables extérieures. Nous citerons le modèle de Mack, qui dispose d'une formule fermée pour estimer l'erreur de prédiction, et la méthode du bootstrap, qui permet d'obtenir une distribution des réserves.

Le point faible de ces deux méthodes est qu'elles sont toutes deux basées sur la méthode de Chain Ladder. Or cette dernière est sensible aux caractéristiques des règlements des sinistres (durée de règlement, volatilité des montants réglés). De plus, ses hypothèses sont très restrictives.

Nous nous sommes alors tournés vers l'étude des données individuelles de chaque sinistre, afin de tenir compte des caractéristiques individuelles complexes de ces sinistres. Pour cela, nous avons dû déterminer les moments clés de la vie d'un sinistre : sa date de survenance, sa date de déclaration, sa date de clôture et les règlements effectués entre temps.

Le modèle ici présenté étudie les sinistres connus par l'assureur au moment du calcul des provisions. Nous avons travaillé sur un portefeuille de responsabilité civile (RC). Pour étudier la chronique de règlement des sinistres, nous avons dû modéliser trois phénomènes : la date de clôture des sinistres, les dates des flux de règlements et les montants associés.
L'étude des dates de clôture s'est faite à l'aide des modèles de survie, afin de prendre en compte les sinistres encore en gestion au moment de l'étude.
L'étude des flux de règlement consistait à modéliser les dates de flux futurs à l'aide de facteurs explicatifs ou de lois paramétriques, et à modéliser les montants de règlement associés à l'aide d'un mélange de lois de probabilités.

Pour bien comprendre les effets du modèle individuel, nous avons comparé les résul-

tats obtenus avec les résultats des méthodes standards. Pour faciliter cette comparaison, nous n'avons pas introduit de facteur de queue de développement ni de retraitement des sinistres graves.
Les méthodes standard donnent des résultats comparables, mais l'incertitude générée par ces modèles est très élevée, et rend les résultats peu intéressants.
Le modèle individuel fournit une estimation de la charge totale des sinistres inférieure à celle des méthodes traditionnelles, et l'incertitude associée est beaucoup mieux contrôlée. Si ce modèle pourrait sous-estimer légèrement la charge ultime des sinistres, les résultats obtenus sont toutefois satisfaisants et permettent un meilleur contrôle du risque de provisionnement.
Le modèle individuel propose donc une alternative crédible et intéressante aux méthodes standards.

Abstract

Keywords : Claims reserving, individual model, GLM, stochastic methods

The insurance's main characteristic is a reversed economic cycle. However the ultimate cost of the insurer's liabilities - its past payments and its futur payments - has to appear in its accounts. So the actuaries have to anticipate the futur payments for reported losses and not yet reported claims.

The most widespread methods to assess the ultimate cost of claims are based on the aggregation of past payments into run-off triangles. The standard determinist method is the Chain Ladder method and is based on the claims' settlement pattern. Yet external factors - such as changes in the claim handling procedures or changes in legislation - may have some influence on its outcomes. Stochastic methods - such as the Mack model or the bootstrap theory - have thus been developed in order to deal with this bias. Such methods can be used to obtain a confidence interval for the insurer's outstanding liabilities.

But both of this methods use the Chain Ladder method - which implies they have the same restrictive hypothesis and the same sensibilities to both external factors and the claims' characteristics.

In order to take into account the claims individual characteristics, we can work on detailed data too. So we will have to identify the main characteristics of a claim management process - its occurence time, its date of report, its date of settlement and its payments.

Our model describes the claim process. Our data come from a professional legal liability portfolio. The loss development is described by a settlement date and a sequence of cash flows.
Survival analysis is adapted for modeling the settlement date.
We used GLMs to described the dates of payments - except the first payment's date which was studied separately. The amount of the associated cash-flow has been modeled by a parametric distribution.

In order to fully understand the effects of our model, we compare its outcomes with the outcomes of standard methods - without any tail factors.
The standard model have similar results with a high degree of uncertainty - but the underlying assumptions of these models are not fully met.
Our model provides better results. The outstanding liabilities is lower and less volatile with our individual model than with the traditional reserving methods.
So the detailed model gives a consistent alternative to classical reserving method.

Remerciements

Ma reconnaissance s'adresse tout d'abord à Philippe MORILHAT, pour m'avoir accueilli au sein de son équipe et pour tout ce qu'il a pu m'apporter.

Je voudrais particulièrement remercier mon tuteur professionnel, Lionel MORAIS ALVES, et l'équipe de la RAC, Lydie PHY et Sophie GARCIN, pour m'avoir permis d'effectuer mon stage dans une bonne ambiance et pour m'avoir formé, tant sur le monde professionnel que sur les tâches actuarielles.

Je voudrais aussi adresser ma gratitude à mon tuteur académique, Jean-Marie NESSI, pour ses conseils avisés et sa disponibilité.

Je remercie enfin les équipes de Gras Savoye, le corps professoral de l'ISUP et son administration, ainsi que tous ceux qui ont contribué de près ou de loin à la réalisation de ce mémoire.

Table des matières

Introduction générale

La spécificité de l'activité d'assurance est son cycle de production inversé : le coût de production (l'indemnisation des assurés) est inconnu lors de la détermination du coût de vente (le montant de la prime demandée à l'assuré). Pour être en mesure de faire face à ses engagements futurs, l'assureur doit donc identifier les risques couverts et provisionner au plus juste ses charges futures.

Les provisions techniques en assurance IARD sont de différentes natures. Nous nous intéressons dans ce mémoire au calcul de la provision pour sinistres à payer (PSAP), qui correspond aux charges des sinistres déjà survenus mais non encore entièrement réglés. Les provisions pour sinistres non déclarés ne sont pas prises en compte ici.

Les méthodes usuellement employées pour le calcul de cette provision utilisent des triangles de liquidation. Ces derniers agrègent les paiements par année de survenance et par année de développement.
Cependant, ces méthodes présentent plusieurs défauts. Notamment, l'agrégation des paiements entraîne une perte d'information. De plus, les résultats obtenus sont moins robustes en cas d'irrégularité des cadences de paiement, d'une forte volatilité des montants ou si la fréquence des sinistres est trop faible. Il est aussi difficile d'introduire des contraintes particulières (comme de la réassurance non proportionnelle), et des sinistres extraordinaires peuvent biaiser les résultats obtenus.
Il peut alors être intéressant de revenir aux données individuelles.

L'objet de ce mémoire est de présenter une méthode de calcul des provisions sinistre par sinistre, pour les sinistres ayant déjà été déclarés à l'assureur. Pour ce faire, chaque sinistre est considéré individuellement : les données ne sont pas regroupées par années de survenance et de développement. Chaque sinistre se caractérise alors par une date de survenance, un processus de règlement et un processus d'état (clos ou en cours de règlement). Néanmoins, nous nous servons des résultats obtenus avec les modèles usuels de provisionnement pour analyser le résultat du modèle individuel.

La première partie consistera à rappeler l'environnement règlementaire. Elle servira aussi à détailler les différentes étapes de la vie d'un sinistre, afin de se familiariser avec les données dont nous disposons pour faire l'étude. Enfin, nous présenterons les méthodes sur données agrégées qui nous serviront de référence, à savoir la méthode de Chain Ladder et le modèle de Mack.
Dans un deuxième temps, nous présenterons le cadre retenu pour l'étude des sinistres dans notre approche ligne à ligne. Nous introduirons aussi les modèles utilisés pour

l'évaluation des provisions.
Enfin, dans une dernière partie, nous appliquerons notre modèle individuel et les méthodes standard de provisionnement à un portefeuille de responsabilité civile. Nous pourrons alors comparer les résultats obtenus et essayer d'expliquer les différences.

Première partie

Les provisions pour sinistres à payer en assurance dommage

Cette première partie doit nous servir à définir le cadre de l'étude. Nous nous intéressons au coût ultime des sinistres et à leur cadence de règlement.

Elle nous permettra donc de faire un rappel des bases de l'assurance non-vie, et notamment de l'enjeu des provisions techniques. Nous présenterons alors les différentes méthodes pour calculer ces provisions. L'analyse théorique de ces méthodes mettra en évidence leurs limites, ce qui nous amènera à présenter les avantages d'un modèle individuel. Enfin, avant toute modélisation, nous procéderons à une analyse des données dont nous disposons, afin de nous familiariser avec ces données.

1 Pratique de l'assurance non-vie

Ce chapitre doit nous permettre de présenter succinctement le déroulement d'un sinistre en assurance non-vie, et de définir la notion et l'importance des provisions techniques.

1.1 Généralités sur l'assurance non-vie

1.1.1 Introduction

Le contrat d'assurance est un accord passé entre deux parties, une compagnie d'assurance et un souscripteur (ou preneur d'assurance), fixant à l'avance et pour une période déterminée des échanges financiers en fonction d'un ensemble bien défini d'éléments aléatoires : « Le contrat aléatoire est une convention réciproque dont les effets, quant aux avantages et aux pertes, soit pour toutes les parties, soit pour l'une ou plusieurs d'entre elles, dépendent d'un évènement incertain »(article 1964 du Code Civil).
Il s'agit d'un contrat aléatoire. En effet, les engagements des deux parties ne sont pas fixés dans la police (l'écrit qui constate la formation d'un contrat d'assurance). En ce qui concerne l'assuré, seul le mode de calcul de la prime peut être décrit dans le contrat. La prestation de l'assureur est bien évidemment inconnue à la signature du contrat, mais les modalités de sa détermination doivent être spécifiées dans la police.

Le preneur d'assurance verse des primes (ou cotisations) pour bénéficier de la garantie de l'assureur.
La prestation de l'assureur a principalement pour objet une somme d'argent, mais elle peut aussi consister en un service à fournir.
En assurance dommages ou responsabilité, l'assureur part d'une évaluation des dommages subis lors du sinistre ou des réclamations des tiers ayant subi un préjudice, après une phase contradictoire qui se terminera par une transaction à l'amiable ou l'intervention d'un jugement. Le règlement final de l'assureur sera égal à cette évaluation, éventuellement diminuée d'une franchise ou limitée par l'application d'un plafond.
Lorsque la prestation ne consiste pas en une somme d'argent, l'assureur peut promettre des services personnels. Dans les assurances de responsabilité, il se réserve la direction du procès contre le tiers lésé. Dans les assurances de frais de justice, il promet à l'assuré de l'assister de ses conseils dans des cas déterminés.
Lorsque la prestation consiste en une somme d'argent, deux types de réparation peuvent avoir lieu :

— une réparation indemnitaire : le montant répare le préjudice subi par l'assuré ou par le tiers lésé. Le montant est donc inconnu a priori. La police d'assurance prévoit

alors le mode d'évaluation du dommage (indemnisation à neuf, en valeur d'usage . . .) ;
— une réparation forfaitaire : le montant est alors prévu au contrat. Il n'y a alors aucun aléa concernant la dépense en cas de sinistre (l'assurance en cas de perte de revenus de l'assuré est un exemple).

Certaines formes d'assurance comprennent à la fois un volet forfaitaire et un volet indemnitaire (par exemple les assurances frais soins de santé : le volet indemnitaire correspondant aux frais d'hôpital, à l'achat de médicaments,. . .et un forfait par jour d'hospitalisation est prévu).

1.1.2 La dynamique de la vie des sinistres

Pour un type de risque donné (RC automobile, santé, marine,. . .), les sinistres sont constatés (plus ou moins longtemps après la date de survenance), puis payés avec là aussi un délai entre la déclaration et la date de paiement. Le paiement d'un sinistre peut s'étaler sur plusieurs années, selon le type de sinistre, la gravité du sinistre, l'évolution de la réglementation. . .Dans certains cas, les cadences de règlement obligent à distinguer l'exercice de survenance, l'exercice de déclaration du sinistre, et l'exercice de règlement du sinistre (année où le sinistre a été définitivement réglé par l'assureur). Les exercices séparant l'année de survenance et l'année de règlement sont appelés années de développement. Ces années de développement ont des caractéristiques propres à la branche considérée. Les différents aspects de la vie des sinistres peuvent être visualisés sur la figure ci-dessous, analogue des diagrammes de Lexis utilisés en assurance vie.

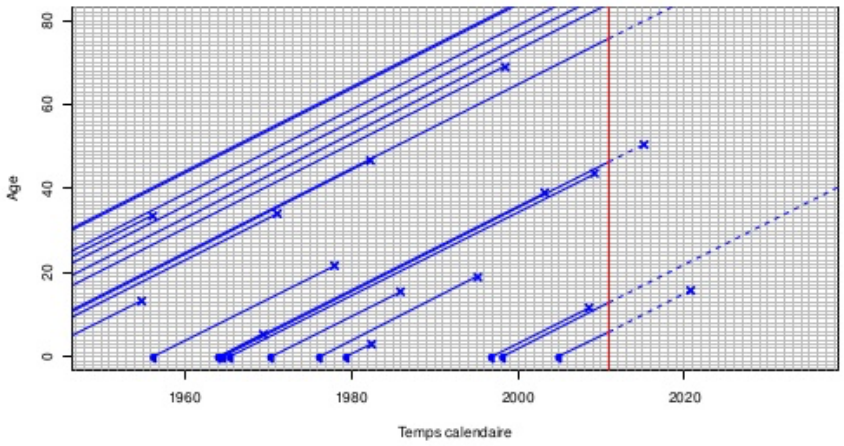


FIGURE 1.1 – Évolution de la vie des sinistres (Source : Denuit et Charpentier, 2005 [2])

Le graphique s'interprète de la manière suivante : en abscisse est représenté le temps calendaire et en ordonnée l'âge des sinistres ; les sinistres surviennent à la date •, sont

déclarés à l'assureur à la date +, et clôturés à la date x.

Le déroulement d'un sinistre dépend fortement de la branche considérée. Le tableau suivant donne une idée des cadences de règlement pour quelques branches :

Branche considérée	n	n+1	n+2	n+3	n+4
Multirisque habitation	55%	90%	94%	95%	96%
Automobile	55%	79%	84%	89%	90%
Responsabilité civile	10%	25%	35%	40%	45%

TABLE 1.1 – Proportion des sinistres clos survenus l'année n par année de développement pour différentes branches (Source : Denuit et Charpentier, 2005 [2])

Pendant tout ce temps, le bilan de l'assureur doit refléter le coût probable des sinistres. C'est ici qu'interviennent les provisions techniques. Du fait des résiliations, l'assureur ne peut pas compter sur des primes futures pour honorer ses engagements actuels. Comme l'année de paiement des sinistres peut différer de l'année d'encaissement des primes, les compagnies doivent constituer des réserves ou provisions techniques à l'aide des primes relatives à un exercice, afin de régler les sinistres survenus au cours de cet exercice une fois que leurs montants seront connus. À tout instant, l'assureur doit disposer de provisions suffisantes pour lui permettre, à supposer que la souscription s'arrête à cet instant, d'indemniser intégralement les sinistres ayant affecté ou devant affecter les contrats souscrits jusqu'alors.

1.2 Les provisions pour sinistres

Les provisions techniques doivent donc permettre à l'assureur d'honorer ses engagements envers les assurés et les bénéficiaires des contrats. Elles représentent en moyenne en assurance non-vie 75% du total du bilan de l'assureur. Il existe différents types de provisions techniques.

La charge ultime d'une année de survenance donnée vues à la date d'inventaire peut se décomposer comme suit :

- le montant des règlements déjà effectués par l'assureur ;
- les provisions dossier-dossier, constituées par le gestionnaire sinistre en charge du dossier pour chaque sinistre déclaré. Constituée uniquement d'une provision pour sinistre à payer (PSAP) lors de l'ouverture du dossier, elle est révisée plus ou moins régulièrement, et diminuée du montant des règlements effectués au fur et à mesure du développement du sinistre ;
- la provision IBNR (*Incurred But Not Reported*). Elle-même se compose de deux provisions distinctes :
  (i) la provision IBNeR (I*Incurred But Not Enough Reported*), qui concerne les sinistres mal provisionnés (trop ou pas assez), et peut donc être positive ou néga-

tive. Calculée au niveau du portefeuille, elle est censée refléter des changements au niveau global (changement du processus de gestion juridique, évolution juridique du sinistre. . .) ;

(ii) la provision IBNyR (*Incurred But Not yet Reported*), qui concerne les sinistres survenus mais non encore déclarés.

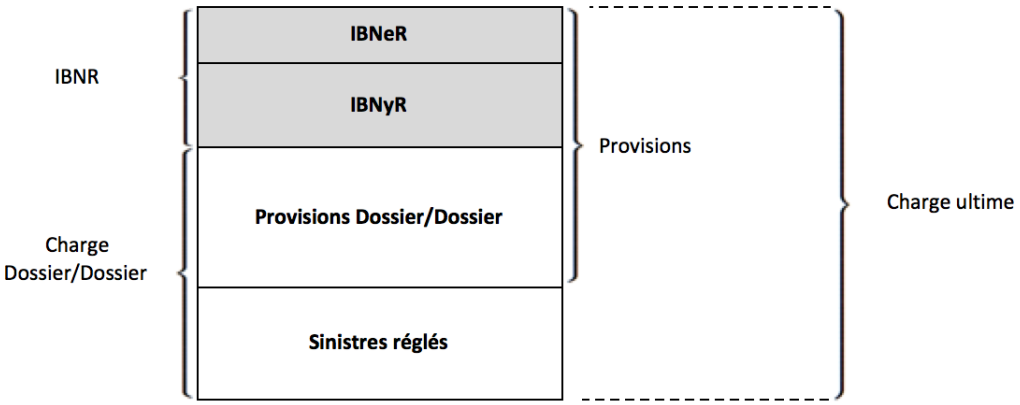


FIGURE 1.2 – Décomposition de la charge ultime entre les différents postes

Comme annoncé précédemment, nous nous intéressons dans ce mémoire uniquement au calcul de la provision pour sinistres à payer, qui sert à régler les sinistres. Elle est estimée par des méthodes statistiques, ce qui conduit à un écart avec les montants réellement versés. Cet écart peut être positif et fait augmenter les fonds propres de la compagnie (on parle de bonus), ou négatif et fait diminuer ces mêmes fonds propres (on parle alors de malus).

1.3 Les triangles de liquidation

Les méthodes standards de provisionnement, présentées dans le chapitre suivant, sont toutes basées sur des triangles de liquidation. Ces derniers sont l'information des règlements de tous les sinistres survenus dans le portefeuille considéré, regroupés par années de survenance et par années de développement. Ils reflètent la dynamique globale des sinistres, et permettent d'avoir une vision agrégée de ceux-ci.

Les notations sont les suivantes :

— i correspond à l'indice des années de survenance des sinistres, avec i=1,. . .,n ;

— j correspond à l'indice des années de développement des sinistres, avec j=1,. . .,n (on ne considère pas dans ce mémoire de facteurs de queue) ;

— $Y_{i,j}$ correspond au montant des règlements effectuées l'année j pour les sinistres survenus l'année i (ou montant incrémental) ;

— $C_{i,j}$ correspond à la somme des règlements effectuées de l'année i à l'année j pour les sinistres survenus l'année i (ou montant cumulé) ; $C_{i,j} = Y_{i,1} + \cdots + Y_{i,j}$.

Avec ces notations, le triangle de liquidation des sinistres prend la forme suivante (nous ne représentons ici que le triangle des paiements cumulés) :

$C_{1,1}$	$C_{1,2}$	$\dots$	$C_{1,j}$	$\dots$	$C_{1,n-1}$	$C_{1,n}$
$C_{2,1}$	$C_{2,2}$	$\dots$	$C_{2,j}$	$\dots$	$C_{2,n-1}$	
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\ddots$		
$C_{i,1}$	$C_{i,2}$		$C_{i,j}$			
$\vdots$	$\vdots$	$\ddots$				
$C_{n-1,1}$	$C_{n-1,2}$					
$C_{n,1}$						

TABLE 1.2 – Triangle des paiements cumulés

La lecture de ces triangles peut se faire par année de survenance (par ligne i), par année de développement (par colonne j), ou par année calendaire (par diagonale i+j).

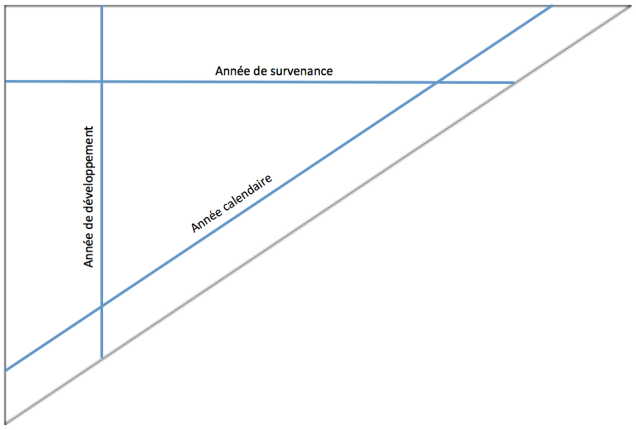


FIGURE 1.3 – Lecture d'un triangle de liquidation des sinistres

Nous travaillons désormais avec le triangle des paiements cumulés $C_{i,j}$. L'information disponible apportée par une ligne du triangle est notée $\mathscr{H}_i = \{C_{i,j} | 0 \leqslant j \leqslant n-i\}$.
Toute l'information disponible dans le triangle est notée $\mathscr{H}_n = \{C_{i,j} | 0 \leqslant i+j \leqslant n\}$.

Nous avons rappelé dans ce chapitre le fonctionnement de l'assurance non vie, l'importance du calcul des provisions techniques, ainsi que les données utilisées pour ce

calcul par les méthodes de provisionnement standard.
Nous allons maintenant présenter les méthodes traditionnelles utilisées pour l'estimation de ces provisions.

2 Les méthodes classiques de provisionnement sur données agrégées, leurs limites, et les apports d'un modèle de provisionnement individuel

Dans ce chapitre, nous présentons les méthodes usuelles de provisionnement, à savoir la méthode de Chain Ladder, le modèle de Mack et la théorie du bootstrap.. Les résultats obtenus par ces méthodes nous serviront de référence pour comparer les provisions calculées dans le modèle individuel.

2.1 Chain Ladder, une méthode déterministe pour obtenir le montant des réserves à l'ultime

Cette méthode est très répandue car facile à comprendre et à mettre en oeuvre. Elle se base sur un triangle des paiements cumulés. Nous notons dans cette partie i pour l'année de survenance des sinistres et j pour l'année de développement, avec (i,j) à valeurs dans $[\![1,\dots,n]\!]$.

2.1.1 Hypothèses et estimation des réserves

Les hypothèses sous-jacentes sont les suivantes :
(H1) les années de survenance sont indépendantes entre elles
(H2) la cadence de règlement dépend uniquement des années de développement

Formellement, en notant i l'année de survenance, j l'année de développement des sinistres et $C_{i,j}$ le montant cumulé pour les années de survenance et de développement respectives, l'hypothèse (H2) s'écrit :

$$\exists(\lambda_1,\dots,\lambda_{n-1}), \text{ tels que } C_{i,j+1} = \lambda_j \cdot C_{i,j}$$

Les coefficients λ_i sont appelés *link-ratios*.

Ces *link-ratios* dépendant uniquement des années de développement, un estimateur

naturel est :

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}} = \sum_{i=0}^{n-j} \frac{C_{i,j}}{\sum_{k=1}^{n-j} C_{k,j}} \cdot \frac{C_{i,j+1}}{C_{i,j}} \tag{2.1}$$

Ainsi, au lieu de faire une simple moyenne des facteurs de développement individuels, nous réalisons une moyenne pondérée de ces facteurs avec les poids $\frac{C_{i,j}}{\sum_{k=1}^{n-j} C_{k,j}}$. Cette pondération permet d'accorder plus d'importance aux années avec des paiements plus importants.

À partir de ces estimateurs, l'estimation des coûts futurs :

$$\hat{C}_{i,j} = \hat{\lambda}_{n+1-i} \dots \hat{\lambda}_{j-1} \cdot C_{i,n+1-i} \tag{2.2}$$

Nous en déduisons le montant des provisions à l'ultime pour chaque année de survenance, que l'on somme sur toutes les lignes :

$$\hat{R}_i^u = \hat{C}_{i,n+2-i} - C_{i,n+1-i} \qquad \hat{R}^u = \sum_{i=1}^{n} \hat{R}_i^u \tag{2.3}$$

DY / SY	1	2	...	j	...	n-1	n	$\hat{R}_i^u$
1	$C_{1,1}$	$C_{1,2}$	...	$C_{1,j}$	...	$C_{1,n-1}$	$C_{1,n}$	0
2	$C_{2,1}$	$C_{2,2}$	...	$C_{2,j}$	...	$C_{2,n-1}$	$\hat{C}_{2,n}$	$\hat{C}_{2,n} - C_{2,n-1}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	$C_{i,1}$	$C_{i,2}$	...	$C_{i,j}$	...	$\hat{C}_{i,n-1}$	$\hat{C}_{i,n}$	$\hat{C}_{i,n} - C_{i,j}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n-1	$C_{n-1,1}$	$C_{n-1,2}$	...	$\hat{C}_{n-1,j}$	...	$\hat{C}_{n-1,n-1}$	$\hat{C}_{n-1,n}$	$\hat{C}_{n-1,n} - C_{n-1,2}$
n	$C_{n,1}$	$\hat{C}_{n,2}$	...	$\hat{C}_{n,j}$	...	$\hat{C}_{n,n-1}$	$\hat{C}_{n,n}$	$\hat{C}_{n,n} - C_{n,1}$
		λ_1	...	λ_{j-1}	...	λ_{n-2}	λ_{n-1}	

TABLE 2.1 – Triangle complété grâce à ses coefficients de développement

2.1.2 Critique de la méthode

Son principal avantage est sa simplicité d'utilisation. De plus, elle est toujours fonctionnelle en cas d'incréments négatifs (à la différence des régressions sur les log incréments par exemple). Enfin, il est possible d'ajouter des facteurs de queue lorsque l'historique disponible ne permet pas d'évaluer l'ensemble des cadences de règlement.

Elle présente toutefois des inconvénients. Le premier est l'hypothèse d'indépendance du coefficient de développement aux années de survenance. Pour que cette hypothèse soit vérifiée, il faut que le passé soit régulier (pas de changement de jurisprudence par exemple), peu volatil (des valeurs extraordinaires se répercutent sur tout le triangle), et le futur doit suivre la même cadence que le passé (par exemple au niveau de l'inflation, qui est prise en compte dans les coefficients de développement). De plus, pour les années de développement tardives pour lesquelles nous disposons de peu d'observations, l'incertitude autour du coefficient de développement est élevée. Enfin, la méthode de Chain Ladder ne permet pas d'obtenir une mesure de l'erreur d'estimation.

2.2 Le modèle de Mack, une méthode stochastique pour obtenir l'erreur d'estimation des réserves

Le modèle de Mack est un modèle non paramétrique permettant d'estimer les erreurs commises lors de l'évaluation des réserves. Il constitue le pendant stochastique de Chain Ladder dans le sens où les hypothèses se font sur l'espérance des paiements et non les paiements eux-mêmes.

2.2.1 Hypothèses et estimations

Les hypothèses sous jacentes au modèle sont les suivantes :

(H1) les années de survenance sont indépendantes entre elles

(H2) l'espérance conditionnelle de $C_{i,j+1}$ sachant le passé $C_{i,1}$ à $C_{i,j}$ est proportionnel à $C_{i,j}$:

$$\mathbb{E}\left(C_{i,j+1}|\mathscr{H}_i\right) = \mathbb{E}\left(C_{i,j+1}|C_{i,1}\dots C_{i,j}\right) = \lambda_j \cdot C_{i,j}$$

(H3) un hypothèse sur la volatilité des facteurs de développement :

$$\mathbb{V}\left(C_{i,j+1}|\mathscr{H}_i\right) = \mathbb{V}\left(C_{i,j+1}|C_{i,1}\dots C_{i,j}\right) = \sigma_j^2 \cdot C_{i,j}$$

Sous les hypothèses (H1) et (H2), nous pouvons montrer que les estimateurs de Chain Ladder $\hat{\lambda}_j = \dfrac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}$ sont sans biais et non corrélés (cf en annexe).

Cette absence de corrélation s'écrit $\mathbb{E}\left(\hat{\lambda}_j \dots \hat{\lambda}_k\right) = \lambda_j \dots \lambda_k$, ce qui signifie que l'estimateur des réserves de Chain Ladder $\hat{R}_i = \hat{C}_{i,n} - C_{i,j}$ est sans biais.

Un estimateur de la variance des coefficients de développement est donné par :

$$\begin{cases} \hat{\sigma}_j^2 = \dfrac{1}{n-j-1} \cdot \sum\limits_{i=1}^{n-j} C_{i,j} \cdot \left(\dfrac{C_{i,j+1}}{C_{i,j}} - \hat{\lambda}_j \right)^2 \text{ pour j=1,. . .,n-2} \\ \hat{\sigma}_{n-1}^2 = min \left\{ \dfrac{\hat{\sigma}_{n-2}^2}{\hat{\sigma}_{n-3}^2}, min\left(\hat{\sigma}_{n-3}^2, \hat{\sigma}_{n-2}^2\right) \right\} \text{ pour j=n-1} \end{cases} \tag{2.4}$$

2.2.2 L'erreur de prévision

À partir de ces estimateurs, il est possible d'estimer l'erreur de prévision, c'est-à-dire la distance moyenne entre l'estimateur des réserves de chaque année de survenance $\hat{R}_i$ et la véritable valeur R_i.

L'erreur quadratique moyenne, ou *mean squared error* (mse), du montant de provisions pour l'année i est définie par

$$mse\left(\hat{R}_i\right) = \mathbb{E}\left[\left(\hat{R}_i - R_i\right)^2 | \mathscr{H}_i \right]$$

et est estimée par

$$\hat{mse}\left(\hat{R}_i\right) = \hat{C}_{i,n}^2 \sum_{k=n-i+1}^{n-1} \frac{\hat{\sigma}_k^2}{\hat{\lambda}_k^2} \cdot \left(\frac{1}{\hat{C}_{i,k}} + \frac{1}{\sum\limits_{j=1}^{n-k} C_{j,k}} \right) \tag{2.5}$$

L'erreur quadratique moyenne du montant total des provisions est alors estimée par

$$\hat{mse}\left(\hat{R}\right) = \sum_{i=2}^{n} \left(\hat{mse}\left(\hat{R}_i\right) + \hat{C}_{i,n} \cdot \left(\sum_{j=i+1}^{n} \hat{C}_{j,n} \right) \cdot \sum_{k=n-i+1}^{n-1} \frac{\dfrac{2 \cdot \hat{\sigma}_k^2}{\hat{\lambda}_k^2}}{\sum\limits_{j=1}^{n-k} C_{j,k}} \right) \tag{2.6}$$

2.2.3 Critiques de la méthode

Les inconvénients du modèle de Mack, dans la mesure où ce modèle reprend les hypothèses de Chain Ladder, sont les mêmes que pour cette dernière.

En revanche, contrairement à Chain Ladder, le modèle de Mack permet d'estimer l'erreur de prévision des réserves, et donc d'obtenir un intervalle de confiance. Elle ne permet toutefois pas d'obtenir la distribution des réserves sans hypothèse supplémentaire.

2.3 Le bootstrap, un méthode stochastique pour obtenir une distribution des réserves à l'ultime

Il s'agit d'une méthode de provisionnement non paramétrique, basée sur le ré-échantillonnage du triangle des paiements et la méthode de Chain Ladder. Elle consiste à reproduire un

triangle de paiements par un tirage aléatoire sans remise des résidus.

2.3.1 Détermination des résidus de Pearson

La première partie du processus consiste à déterminer les facteurs de développement du triangle. À partir de ces facteurs de développement, nous pouvons estimer les valeurs du triangle initial :

$$\hat{C}_{i,j+1}^{estimé} = \lambda_j \cdot C_{i,j}$$

DY / SY	1	2	...	j	...	n-1	n
1	$C_{1,1}$	$\hat{C}_{1,2}^{estimé}$	...	$\hat{C}_{1,j}^{estimé}$	...	$\hat{C}_{1,n-1}^{estimé}$	$\hat{C}_{1,n}^{estimé}$
2	$C_{2,1}$	$\hat{C}_{2,2}^{estimé}$	...	$\hat{C}_{2,j}^{estimé}$	...	$\hat{C}_{2,n-1}^{estimé}$	
⋮	...	...	...	...	...	...	...
i	$C_{i,1}$	$\hat{C}_{i,2}^{estimé}$	...	$\hat{C}_{i,j}^{estimé}$	...		
⋮	...	...	...	...	...	...	...
n-1	$C_{n-1,1}$	$\hat{C}_{n-1,2}^{estimé}$					
n	$C_{n,1}$						
		λ_1	...	λ_{j-1}	...	λ_{n-2}	λ_{n-1}

TABLE 2.2 – Triangle estimé à partir des coefficients de développement

Nous calculons alors les résidus de Pearson comme suit :

$$\hat{r}_{i,j}^{P} = \frac{C_{i,j} - \hat{C}_{i,j}^{estimé}}{\sqrt{\hat{C}_{i,j}^{estimé}}}$$

2.3.2 Ré-échantillonnage du triangle des résidus

Nous effectuons ensuite N ré-échantillonnage du triangle des résidus, avec ou sans remise. Ceci nous permet d'obtenir N triangles de résidus, desquels nous déduisons N triangle des paiements :

$$C_{i,j}^{bootstrapé} = \hat{C}_{i,j} + \sqrt{\hat{C}_{i,j}} \cdot \hat{r}_{i,j}^{P}$$

SY \ DY	1	2	...	j	...	n-1	n
1	$C_{1,1}$	$C_{1,2}^{bootstrapé}$	...	$C_{1,j}^{bootstrapé}$	...	$C_{1,n-1}^{bootstrapé}$	$C_{1,n}^{bootstrapé}$
2	$C_{2,1}$	$C_{2,2}^{bootstrapé}$	...	$C_{2,j}^{bootstrapé}$	...	$C_{2,n-1}^{bootstrapé}$	$\hat{C}_{2,n}^{bootstrapé}$
⋮	...	...	...	...	...	...	...
i	$C_{i,1}$	$C_{i,2}^{bootstrapé}$	...	$C_{i,j}^{bootstrapé}$	...		
⋮	...	...	...	...	...	...	...
n-1	$C_{n-1,1}$	$C_{n-1,2}^{bootstrapé}$					
n	$C_{n,1}$						

TABLE 2.3 – Un triangle obtenu par ré-échantillonnage des résidus

Nous appliquons alors la méthode de Chain Ladder à chacun de ces N triangles de paiements, en recalculant les facteurs de développement respectifs et le montant des réserves associé à chaque triangle. Nous obtenons alors la distribution des réserves, au lieu des simples moments d'ordre 1 et 2 obtenus avec le modèle de Mack.

2.4 Les avantages d'un modèle de provisionnement individuel

Comme les méthodes stochastiques, un modèle individuel permet d'obtenir un intervalle de confiance pour l'estimation de la réserve ultime.

Les résultats d'un modèle individuel sont additifs, c'est-à-dire que le montant des réserves pour un portefeuille donné est égal à la somme des montants des réserves pour toute partition de ce portefeuille.

Nous nous intéressons ici à la dynamique de règlement des sinistres. Les méthodes ne modélisant que le coût final des sinistres ne nous intéresse donc pas.

Enfin, les méthodes classiques montrent leurs limites quand les montants réglés sont volatils ou quand les sinistres ont une longue durée de vie. Notamment les hypothèses sous-jacentes ne sont souvent plus vérifiées, ce qui remet en cause la validité des résultats obtenus avec ces méthodes.
Un modèle individuel permet lui de considérer l'ensemble des caractéristiques importantes des sinistres. Avec une modélisation plus fine et avec moins de perte d'information due à l'agrégation des données, nous pouvons espérer obtenir de meilleurs résultats, en terme de montant mais surtout de volatilité des réserves.

L'approche est toutefois très différente de celle des méthodes classiques. Aussi est-il intéressant, avant de passer à la modélisation, d'explorer notre base de données pour connaître un peu mieux le portefeuille étudié et ainsi disposer d'un oeil critique sur

notre modèle et ses résultats.

3 L'étude d'un portefeuille de Responsabilité Civile

Par souci de confidentialité, toutes les statistiques et les montants de règlements liées à ce portefeuille ont été changés d'échelle.

Nous travaillons dans ce mémoire sur un portefeuille de Responsabilité Civile, pour les professionnels de l'immobilier. Les modèles sur données agrégées ne fonctionnent pas aussi bien sur ces sinistres que sur des sinistres à durée de vie plus courte comme les sinistres automobile. C'est pourquoi nous avons voulu revenir à une étude ligne à ligne.

3.1 Les données disponibles

Nous disposons de données sur 5 110 RC professionnelle survenus entre le 1^{er} janvier 2011 et le 31 décembre 2015, et déclarés avant le 31 décembre 2015, dont 1 941 ont engendré des paiements avant le 31 décembre 2015.

Pour chacun de ces sinistres, nous disposons de :
- une date de survenance du sinistre ;
- une date de déclaration du sinistre ;
- une date d'ouverture du dossier ;
- un statut du sinistre (clos ou en gestion) ;
- l'activité concernée par le sinistre ;
- la cause du sinistre ;

Nous possédons de plus des informations suivantes sur les flux de sinistres (pour un total de 3 593 cash flow) :
- la date du déboursement ;
- le montant déboursé par l'assureur ;
- le montant déjà déboursé par l'assureur lors de la survenance d'un flux : 0 s'il s'agit du premier flux de règlement, la somme des flux précédents sinon.

Un premier traitement des données a consisté à annuler tous les flux négatifs, qui correspondent à l'annulation d'un paiement effectué le jour même ou peu de temps auparavant. Nous les avons donc supprimées en considérant qu'il s'agissait d'erreurs de gestion.
De même, nous supprimons les sinistres clos le jour de leur survenance, considérant

qu'il s'agit d'erreurs de gestion.

3.2 L'analyse exploratoire des données

Nous cherchons à savoir quels facteurs peuvent influer sur le développement d'un sinistre.

Notons tout d'abord que sur les 5 110 sinistres survenus depuis 2011, 2 040 sinistres sont clos (soit 40%).
Pour les activités exercées par les assurés sinistrés, trois activités sont réellement importantes, car elles sont associées à 4 572 sinistres (soit 90% du total des sinistres), ce qui signifie que les autres activités sont en nombre très faible et risquent de ne pas donner de résultats significatifs.

3.2.1 Sur la date de clôture des sinistres

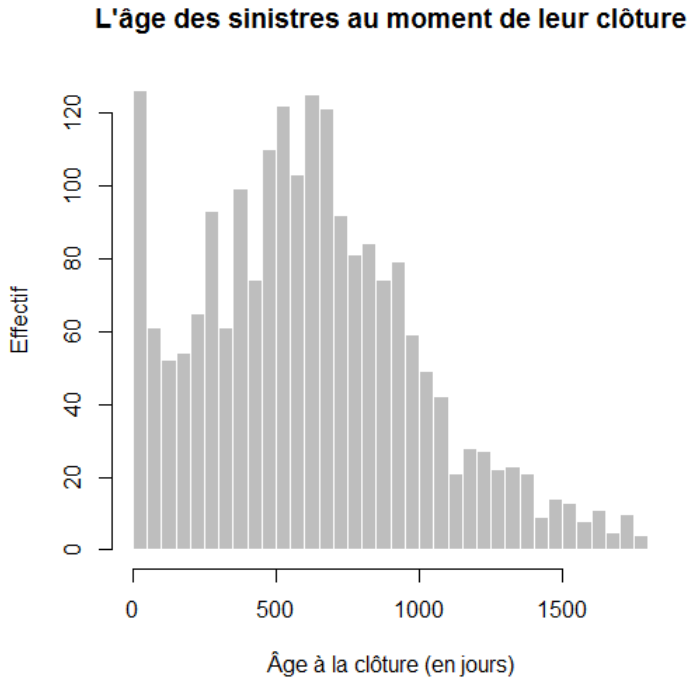


FIGURE 3.1 – L'âge des sinistres clos au moment de la fermeture du dossier

Un nombre important de sinistres sont clos dans les 50 jours suivant leur survenance : il s'agit probablement de sinistres refusés suite à la déclaration du sinistre. Sinon, les

sinistres ont une durée de vie moyenne de 623 jours (soit un peu moins de 2 ans). Attention toutefois à ne pas tirer de conclusions hâtives : comme dit précédemment, moins de la moitié des sinistres sont clos. Plus précisément, encore 37% des sinistres survenus en 2011 et 2012 sont encore en gestion. L'historique des règlements étant relativement court (5 ans), les résultats concernant la date de clôture des sinistres sont biaisés par la sur-représentation des sinistres courts.

De plus, les sinistres réglés suite à une procédure judiciaire ont une durée de vie généralement plus longue que les sinistres réglés à l'amiable : les sinistres réglés à l'amiable le sont après 518 jours en moyenne (1 an et 5 mois), contre 745 jours en moyenne pour les sinistres suivant une procédure judiciaire (2 ans). Encore une fois, ce résultat est à relativiser par l'absence d'observation des sinistres au déroulement long.

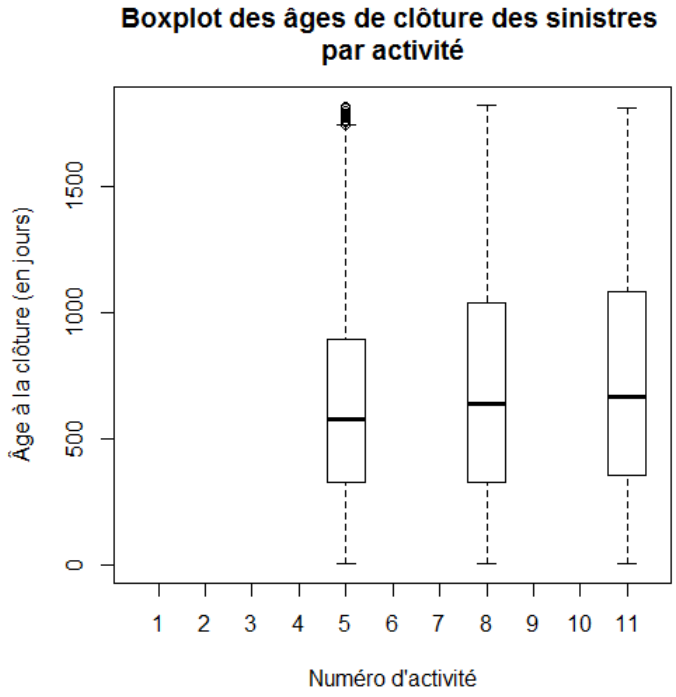


FIGURE 3.2 – L'âge des sinistres clos au moment de la fermeture du dossier

Enfin, nous avons représenté les boxplot de l'âge des sinistres à la clôture pour les trois activités les plus représentées. Nous n'observons pas de différence significative pour ces trois activités, et les autres activités sont sous-représentées pour que les résultats soient exploitables.

3.2.2 Sur les dates des flux de règlement

Nous nous demandons tout d'abord si tous les flux surviennent après un même laps de temps. En effet, nous pouvons supposer que le premier règlement ne survient pas immédiatement,à cause du délai de déclaration, de la volonté de l'assureur d'attendre un peu pour régler en une seule fois les petits sinistres ou à cause du délai causé par une procédure judiciaire. Au contraire, il est possible qu'une fois le premier règlement effectué, les flux s'enchaînent à des dates plus rapprochées.

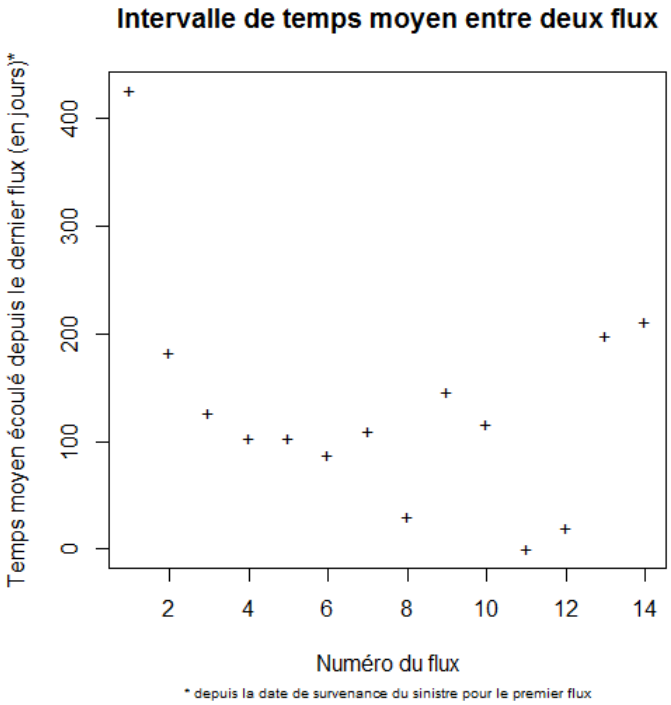


FIGURE 3.3 – L'âge des sinistres clos au moment de la fermeture du dossier

Effectivement, comme nous le montre le graphique ci-dessus, le premier sinistre survient en moyenne plus d'un an après la survenance du sinistre, quand les autres flux s'enchaînent plus rapidement (tous les 100 jours en moyenne). La représentation graphique du dixième flux et plus n'est pas représentative, car elle ne concerne qu'un seul sinistre qui a connu plus de 10 flux de règlement.

De plus, le délai entre la date de survenance du sinistre et le règlement d'un flux est positivement corrélé au délai entre la réalisation de deux règlements :

Délai depuis la déclaration (en jours)	Délai moyen entre deux flux (en jours)
≤ 210	102
211 - 419	225
420 - 689	305
≥ 690	436

TABLE 3.1 – Délai moyen entre deux flux selon le temps écoulé depuis la survenance du sinistre

En effet, pour une date de déclaration plus ancienne, il est possible que ce temps d'attente soit dû à une procédure judiciaire, auquel cas des recours peuvent avoir lieu ce qui peut reculer la date du règlement suivant.

3.2.3 Sur les montants des flux de règlement

Le tableau suivant détaille les caractéristiques principales des flux de règlement selon l'activité exercée par le client.

Activité	Nombre de flux	Flux moyen (en €)
Autre	3	2 199,76
Baux commerciaux	82	1 816,23
Caisse de garantie	121	1 139,11
Divers	47	1 309,68
Gestion habitation professionnelle	1 426	1 906,47
Immobilier d'entreprise	1	2 852,13
Location saisonnière	26	962,28
Syndic de copropriété	679	1 582,61
Transaction location habitation	57	1 934,32
Transaction sur fonds de commerce	77	1 374,85
Transaction sur immeuble	1 073	1 291,34

TABLE 3.2 – Statistiques des flux par activité

Nous observons bien la sur-représentation de trois branches d'activité : Gestion habitation professionnelle, Syndic de copropriété et Transaction sur immeuble.
Sur l'ensemble des flux, les activités Caisse de garantie et Immobilier d'entreprise se caractérisent par un montant moyen par flux significativement différent des autres activités. Cependant, le faible nombre de sinistres correspondant à ces activités (1 sinistre pour l'activité Immobilier d'entreprise, 81 pour l'activité Caisse de garantie) ne permettent pas de considérer ces montants moyens comme vraiment représentatifs. L'activité ne semble donc pas être une variable très discriminante dans le montant moyen des flux de règlement des sinistres.

Deuxième partie

Le modèle de provisionnement ligne à ligne

Nous cherchons maintenant à établir un modèle de provisionnement utilisant des données détaillées. Nous nous restreignons à l'étude des sinistres déjà survenus ; les IBNR ne sont pas analysés ici.

Nous avons décidé d'étudier les flux de règlement futurs. Nous avons donc trois paramètres à modéliser :

— la date de clôture de chacun des sinistres ;
— les dates de survenance des règlements ;
— les montants associés à ces dates.

Nous disposons de relativement peu de sinistres clos, et certains sinistres anciens ne le sont pas encore : nous observons donc surtout des sinistres clos rapidement, pas représentatifs de la population. Nous disposons donc de données censurées à droite. L'utilisation des techniques de l'analyse de survie nous a donc semblé appropriée.

Nous considérons d'autre part que la survenance des flux se déroule selon un modèle de Poisson. Toutefois, nous avons vu précédemment que le premier flux se distinguait des autres flux par le délai écoulé avant sa survenance (500 jours en moyenne après la survenance du sinistre, quand l'intervalle de temps entre les autres flux est d'approximativement 100 jours). En effet, une fois que le sinistre a donné lieu à une indemnisation, il est possible que les paiements s'enchaînent relativement rapidement jusqu'à la clôture du sinistre, un paiement en entrainant un autre.

4 La modélisation des dates de clôture à l'aide des modèles de survie

Le premier paramètre à modéliser est la date de clôture des sinistres.

Les modèles de durée sont utilisés pour modéliser et estimer des lois décrivant le temps qui s'écoule entre deux évènements (durée de vie d'un individu, durée d'un épisode de chômage...). Mais les données traitées ne sont pas toujours complètes : censures et troncatures sont les perturbations les plus connues, et typiques des données de survie.
Nous ne disposons que de peu de sinistres clos ; de plus, l'historique des sinistres dont nous disposons étant relativement cours (5 ans d'historique), les sinistres courts sont sur-représentés et les sinistres plus longs sous-représentés. C'est pourquoi nous avons choisi d'utiliser les techniques d'analyse de survie pour étudier la date de clôture des sinistres.

4.1 Les principes de l'analyse de survie

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenance d'un évènement précis. Cet évènement est le passage irréversible d'un état de « vivant »à l'état « décès ». L'évènement terminal n'est toutefois pas forcément la mort : ce peut être l'apparition d'une maladie, une guérison, une panne d'appareil...L'analyse des données de survie est l'étude du délai avant la survenance de cet évènement. Dans notre cas, il s'agit du délai avant la clôture d'un sinistre.

Pour ce type d'étude, nous devons disposer de quatre données minimum :
- la date d'origine correspond à l'origine de la durée étudiée, et peut être différente pour chaque individu observé. Dans notre mémoire, il s'agit de la date d'ouverture du sinistre ;
- la date d'observation, ou date de point, est la date à laquelle les données sont observées et l'étude réalisée ;
- la date des dernières nouvelles est la date la plus récente où des informations sur un sujet ont été recueillies. Elle n'est pas forcément identique à la date d'observation, dans le cas d'un sinistre clos avant la fin de la période d'observation par exemple ;
- la variable d'état, qui est l'état du sujet aux dernières nouvelles. Dans notre cas, le sinistre peut être en gestion ou ouvert ; la variable d'état est donc binaire.

Deux situations se présentent à nous avec nos données :
- soit le sinistre a été clos au cours de la période d'observation : la durée de survie est donc le délai entre la date d'ouverture du sinistre et la date de clôture du sinistre ;
- soit le sinistre est encore en gestion à la date d'observation : la donnée est alors censurée à droite, nous savons uniquement que la durée de survie du sinistre est supérieure au délai entre la date d'ouverture du sinistre et la date d'observation.

4.2 L'estimation non paramétrique avec l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t, c'est être en vie juste avant t et ne pas mourir au temps t.

4.2.1 Définition et principales propriétés de l'estimateur de Kaplan-Meier

Nous utilisons les notations suivantes :
- $n_i$le nombre de sinistres ouverts en t_i^- ;
- d_i le nombre de clôtures au temps t_i.

L'estimateur de Kaplan-Meier s'écrit alors :

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{4.1}$$

Nous pouvons estimer la variance de cet estimateur par la formule de Greenwood :

$$\hat{\mathbb{V}}_{KM}(t) = \left(\hat{S}_{KM}(t)\right)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{4.2}$$

L'estimateur de Kaplan-Meier est cohérent, c'est-à-dire que la probabilité de décéder au-delà de la date t est la somme de :
- la probabilité de n'être ni décédé, ni censuré à la date t ;
- la probabilité d'avoir été censuré avant la date t et d'être toujours en vie à la date t.

De plus, l'estimateur de Kaplan-Meier est asymptotiquement gaussien, ce qui permet de définir un intervalle de confiance asymptotique.

Enfin, nous pouvons montrer que l'estimateur de Kaplan-Meier est un estimateur non-paramétrique du maximum de vraisemblance.

Critique de l'estimateur de Kaplan-Meier

Les principaux avantages de cet estimateur sont sa simplicité de mise en oeuvre et l'absence d'hypothèse faite sur la distribution des décès. Il est toutefois peu efficace dans le cas où beaucoup d'entrées (censures) et sorties (censures et décès) se produisent au même moment.

Pour un intervalle de confiance à 95%, nous obtenons le graphique suivant :

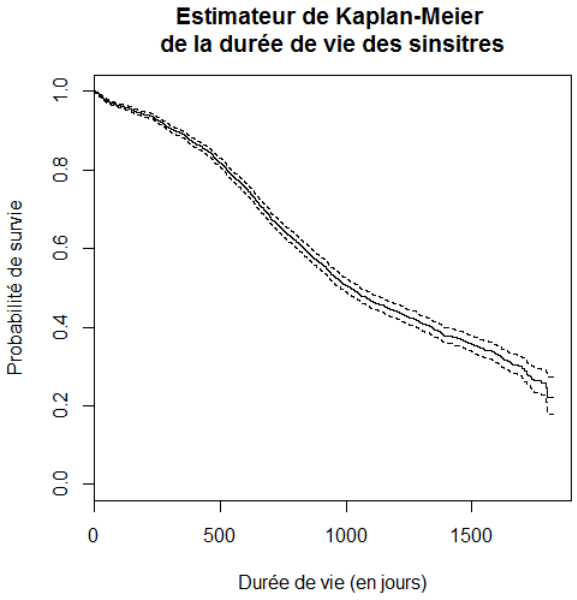


FIGURE 4.1 – Représentation de la fonction de survie

4.3 La prise en compte de variables exogènes : le modèle semi-paramétrique de Cox

L'hétérogénéité d'une population correspond en fait à un mélange de populations avec des caractéristiques différentes. Il existe différents modèles qui prennent en compte ce phénomène : le modèle additif d'Aalen, . . .De par leur simplicité de mise en oeuvre et d'interprétation, ainsi que par la prise en compte des censures et troncatures, le modèle à hasard proportionnel de Cox et le modèle d'Aalen sont les plus courants.

Nous nous focalisons sur le modèle de Cox, qui permet de prendre en compte l'effet de variables exogènes sur la durée de vie.

4.3.1 Les modèles à hasards proportionnels

Ces modèles expriment un effet multiplicatif des diverses covariables sur la fonction de hasard, définie par $\lambda(t) = \frac{f(t)}{S(t)}$, avec f(t) la densité de la variable et S(t) sa fonction de survie.
Les modèles à hasards proportionnels introduisent une fonction de hasard de base λ_0, qui donne la forme générale du hasard, valable pour tous les individus. Les modèles à hasards proportionnels se caractérisent alors par la relation suivante : $\forall t > 0, \forall z$

$$\lambda(t|z) = \lambda_0(t)h(z,\beta) \tag{4.3}$$

avec z le vecteur des variables exogènes, β le paramètre d'intérêt et h une fonction positive. En général, l'effet des covariables est supposé se résumer à une quantité réelle $\beta^t Z$, appelée index :

$$\lambda(t|z) = \lambda_0(t)h(\beta^t Z)$$

Ce modèle est dit à hasards proportionnels car, quelque soient deux individus i et j ayant pour covariables Z_i et Z_j, le rapport des fonctions de hasard ne varie pas au cours du temps :

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = \frac{h(\beta^t Z_i)}{h(\beta^t Z_j)}$$

L'interprétation du modèle est la suivante : toutes choses égales par ailleurs, une covariable qui modifie h par rapport au niveau de référence induit un effet multiplicatif de même ampleur sur le hasard à toute date t.

4.3.2 Le modèle de Cox

Il s'agit d'un cas particulier des modèles à hasards proportionnels, qui suppose que la fonction h est la fonction exponentielle :

$$\lambda(t|z) = \lambda_0(t)exp(z,\beta) \tag{4.4}$$

L'utilisation de la fonction exponentielle est très appréciée car ses valeurs sont toujours positives et exp(0)=1.

Nous avons utilisé diverses variables, telles que l'année de déclaration, le délai de déclaration, l'activité ou la cause des sinistres, mais les résultats n'étaient pas significatifs. Nous avons alors décidé de nous tourner vers l'ajustement paramétrique de la loi de survie.

4.4 La modélisation paramétrique de la date de clôture de nos sinistres

Nous voulons ajuster la loi de manière paramétrique. Nous allons pour cela comparer les résultats obtenus avec les lois classiques de l'analyse de survie, à savoir la loi exponentielle, la loi de Weibull et la loi log-normale, en tenant compte de la censure à

droite des données (cf Zhang and Wie [1] pour l'estimation des paramètres de la loi de Weibull dans le cas censuré, à adapter aux autres lois). La loi exponentielle donnait des résultats très éloignés de la fonction de survie, mais nous avons comparé les fonctions de distributions des deux autres lois :

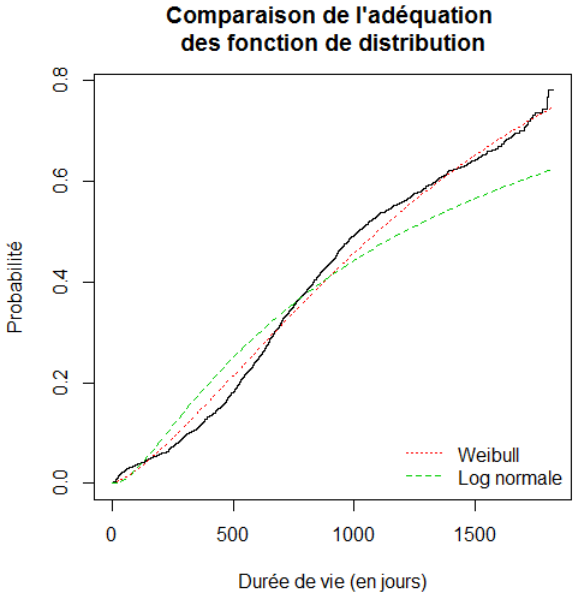


FIGURE 4.2 – Comparaison des fonctions de répartition

La loi Weibull semble être la plus adaptée à nos données : la loi log-normale surestime en effet la probabilité de décès aux grands âges.

Nous voulons donc valider cette estimation par un test statistique : le test du log-rank. Ce test est un test non paramétrique qui permet de comparer la fonction de survie de deux échantillons : ici la fonction de survie empirique (l'estimateur de Kaplan-Meier) et la fonction de survie théorique (la loi Weibull).
Le principe de ce test est de comparer chacune des valeurs des deux distributions, et de comparer la somme normalisée de ces écarts à un chi-deux à un degré de liberté.

```
> with(Data.cloture,survdiff(Surv(Age,Status==0)~offset(1-prob0),rho=0))
Call:
survdiff(formula = Surv(Age, Status == 0) ~ offset(1 - prob0),
    rho = 0)

 Observed  Expected         Z         p
2042.0000 2040.1777   -0.0403    0.9680
```

FIGURE 4.3 – Le test du log-rank pour comparer deux fonctions de survie

Nous ne rejetons alors pas l'hypothèse d'égalité des fonctions de survie ; le modèle est

bien ajusté avec une loi de Weibull.

L'ajustement de la loi de Weibull à notre fonction de survie donne le graphique suivant :

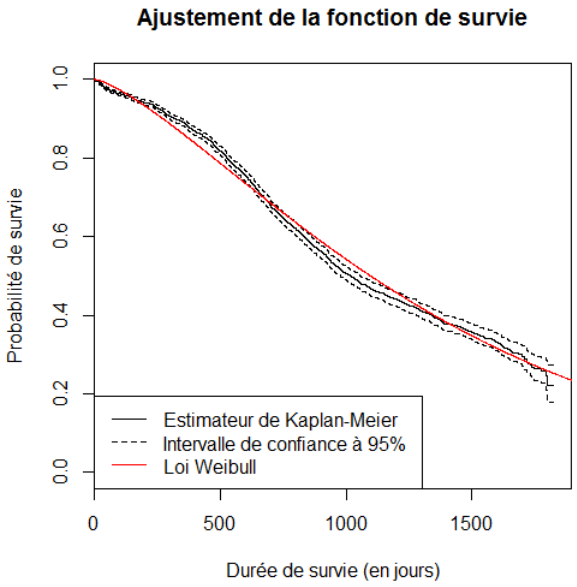


FIGURE 4.4 – L'adéquation de la loi Weibull à notre fonction de survie

Le choix de la date de clôture des sinistres dépend de la politique de gestion des sinistres. En effet, certains sinistres n'ayant pas donné lieu à des flux de règlement depuis un certain temps peuvent rester ouverts à la discretion du gestionnaire, par mesure de prudence.
Nous pouvons regarder le délai depuis le dernier flux de règlement. Nous manquons toutefois d'historique pour déterminer à partir de quel seuil un sinistre ne donnera plus lieu à un règlement. Nous étudions une branche longue, et il n'est pas anormal de ne pas observer de règlement depuis plus d'un an pour un sinistre. Nous n'avons donc pas procédé à la clôture de certains dossiers avant d'effectuer nos simulations.

5 Les modèles utilisés pour l'étude des flux de règlement

Après avoir modélisé la date de cloture des sinistres, nous voulons étudier les flux de règlement survenant entre la date de l'étude et le date de clôture des sinistres en gestion.
Ce chapitre doit nous permettre d'exposer la théorie des modèles utilisés pour l'étude des dates de règlement et des montants versés. Nous cherchons dans un premier temps à modéliser l'effet de variables exogènes sur nos variables réponse. Pour cela, nous présentons la théorie des modèles linéaires généralisés, puis celles des modèles additifs généralisés.

5.1 Une première approche : les Modèles Linéaire Généralisés (GLM)

Dans le modèle linéaire standard, la variable réponse est supposée suivre une loi normale dont la moyenne est expliquée par une combinaison linéaire des variables explicatives :

$$Y \sim \mathcal{N}(\mu, \sigma^2) \tag{5.1}$$

où

$$\mu = \alpha + \sum_{i=1}^{n} \beta_i x_i \tag{5.2}$$

Le modèle linéaire généralisé permet de définir un type de modèle plus étendu. Il permet de considérer d'autres lois de probabilité pour la variable réponse, de conserver la structure linéaire du score, et considère que l'espérance de Y est une transformation de cette combinaison linéaire.

La famille des lois de probabilité de la variable réponse

L'hypothèse sous-jacente au GLM sur la loi de probabilité suivie par la variable réponse est son appartenance à la famille exponentielle. Autrement dit, sa densité peut se mettre sous la forme suivante :

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\} \tag{5.3}$$

avec $y \in \mathscr{S}$, où $\mathscr{S}$ est un sous-ensemble de $\mathbb{N}$ ou de $\mathbb{R}$. Le paramètre θ est appelé paramètre naturel, et ϕ est le paramètre de dispersion. Quand une pondération est nécessaire, ϕ est rapporté à un poids connu ω et le paramètre de dispersion vaut alors $\frac{\phi}{\omega}$.

Parmi cette famille de lois de probabilité, nous retrouvons notamment la loi normale, la loi de poisson, la loi binomiale, la loi gamma...Toutes les lois ne possèdent pas un paramètre de dispersion (par exemple, pour la loi de poisson, $\phi = 1$). Pour les lois possédant un paramètre de dispersion ϕ différent de 1, ce dernier contrôle la variance.

Pour une variable aléatoire Y dont la densité peut se mettre sous la forme 5.3, nous pouvons exprimer ses deux premiers moments :

$$\mathbb{E}(Y) = b'(\theta) \tag{5.4}$$

$$\mathbb{V}(Y) = \phi b''(\theta) \tag{5.5}$$

Le modèle de régression

Pour le modèle de régression, nous considérons des variables aléatoires indépendantes mais non identiquement distribuées $Y_1, \ldots, Y_n$, dont la densité est de la forme (5.3). Plus précisément, nous supposons que la densité de la variable Y_i est de la forme

$$f(y_i|\theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\} \tag{5.6}$$

Avec le GLM, nous voulons expliquer la moyenne μ_i comme une fonction simple des prédicteurs :

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \tag{5.7}$$

où la fonction monotone et dérivable g est appelée fonction de lien, les $x_{i,j}$ sont les valeurs prises par les variables explicatives relatives à l'individu i, et les β_i sont les valeurs des paramètres relatifs à chaque individu i.

Un modèle linéaire généralisé se compose donc de trois éléments :
(i) des variables réponses $Y_1, \ldots, Y_n$ dont les densités sont de la forme (5.6) ;
(ii) d'un ensemble de paramètres $(\beta_0, \ldots, \beta_p)$ appartenant à un ouvert non vide de $\mathbb{R}^{p+1}$ et de variables explicatives $X_1, \ldots, X_n$;
(iii) d'une fonction de lien g telle que $g(\mu_i) = x_i^t\beta$, où $\mu_i = \mathbb{E}(Y_i)$.

5.1.1 La mesure de la qualité d'ajustement du modèle

Mesurer la qualité d'ajustement du modèle revient à se demander si les écarts entre les valeurs prédites par le modèle $\hat{y}_i$ et les observations initiales y_i traduisent une mauvaise qualité du modèle ou peuvent être attribués au hasard. Pour ce faire, deux principales statistiques sont utilisées : la déviance et les résidus de Pearson.

La déviance

La déviance est une généralisation du R^2 des moindres carrés ordinaires (MCO) du modèle linéaire classique. Elle se définit en comparant le GLM au modèle saturé : ce dernier est un modèle où à chaque observation est associé un paramètre, et qui fournit donc une description parfaite des données. La déviance réduite se calcule comme deux fois la différence entre la vraisemblance du modèle saturé et la vraisemblance du modèle linéaire généralisé :

$$D = -2\left(ln(\mathscr{L}(y|y)) - ln(\mathscr{L}(\hat{y}|y))\right) \tag{5.8}$$

La déviance non réduite est elle donnée par $D^* = \phi D$.

La qualité d'ajustement du modèle est généralement évaluée par la déviance. Une petite valeur de la déviance indique un bon ajustement du modèle aux données, puisque la vraisemblance du modèle est proche de celle du modèle saturé. Au contraire, une valeur élevée de la déviance traduit un piètre ajustement du modèle aux données.

Concrètement, une première approche consiste à comparer la déviance normée (le rapport entre la déviance et le paramètre de dispersion) divisée par son degré de liberté à 1.
De plus, si le modèle décrit bien les données, la déviance suit une loi du khi-deux à n-p-1 degrés de liberté : $D \sim \chi^2_{n-p-1}$.

En pratique, nous jugeons de la qualité du modèle en comparant la valeur de la déviance au quantile d'ordre $1-\alpha$ de la loi du khi-deux à n-p-1 degrés de liberté : nous considérons que le modèle est de mauvaise qualité si

$$D > q_{\chi^2_{n-p-1}}(1-\alpha) \tag{5.9}$$

La statistique de Pearson

La statistique khi-carré de Pearson, définit par :

$$\chi^2_{Pearson} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mathbb{V}(Y_i)} \tag{5.10}$$

permet aussi de mesurer la qualité d'ajustement du modèle aux données. La statistique χ^2 suit, comme la déviance, une loi du khi-deux (résultat exact dans le cas gaussien, résultat asymptotique sinon). La même comparaison au quantile de la loi du khi-deux permet de déterminer la qualité de l'adéquation du modèle aux données.

L'analyse des résidus

Les mesures étudiées ci-dessus, déviance et statistique de Pearson, permettent de juger l'adéquation globale du modèle aux données. L'analyse des résidus permet de découvrir l'origine du décalage entre le modèle et les données, et ainsi éventuellement d'améliorer le modèle initial.
Deux situations d'inadéquation globale du modèle peuvent se produire : soit un petit nombre d'observations sont mal décrites par le modèle, soit l'ensemble des observations présentent un écart systématique par rapport au modèle. Une représentation graphique des résidus permet alors de détecter de quel type erreur il s'agit, et d'agir en conséquence (par exemple retirer les données extrêmes qui biaisent le modèle, si l'inadéquation du modèle ne concerne que quelques observations).

Les résidus de Pearson, ou les résidus de déviance

Deux types de résidus sont couramment utilisés dans le cadre des modèles linéaires généralisés : les résidus de Pearson et les résidus de déviance.

Les résidus de Pearson sont définis par :

$$r_i^{Pearson} = \frac{\sqrt{\omega_i}(y_i - \mu_i)}{\sqrt{\mathbb{V}(\mu_i)}} \tag{5.11}$$

Ils peuvent en effet être vus comme la racine carrée de la contribution de la $i^{ème}$ observation à la statistique de Pearson :

$$\chi^2_{Pearson} = \sum_{i=1}^{n} (r_i^{Pearson})^2 \tag{5.12}$$

De même, les résidus de déviance sont définis comme la racine carrée de la contribution d_i de la $i^{ème}$ observation à la déviance D, affectée du signe du résidu brut :

$$r_i^D = signe(y_i - \mu_i)\sqrt{d_i} \tag{5.13}$$

d'où

$$D = \sum_{i=1}^{n} (r_i^D)^2 \tag{5.14}$$

La représentation graphique des résidus peut se faire selon le numéro d'observation, ce qui permet d'identifier les observations conduisant à de grands résidus. Elle peut aussi se faire en fonction des valeurs prédites $\hat{\mu}_i$ ou des prédicteurs linéaires $\hat{\eta}_i$, ou en fonction de chacune des variables explicatives.

Les observations influentes

Une observation y_i est dite influente lorsqu'une petite variation de cette variable ou son omission conduit à des estimations fort différentes des paramètres du modèle. Une telle observation n'est cependant pas obligatoirement un *outlier* : elle peut se trouver près des autres observations et avoir une forte influence.

La distance de Cook est utilisée pour déterminer les observations qui influencent l'estimation des paramètres. En notant $\hat{\beta}$ l'estimateur du maximum de vraisemblance du modèle complet, et $\hat{\beta}_i$ l'estimateur du maximum de vraisemblance du modèle sans la variable i, la distance de Cook représente la distance entre ces deux estimateurs. Si cette distance est grande, cela indique que l'observation i a une forte influence sur l'estimation des paramètres du modèle.

5.2 Un seconde approche : les Modèles Additifs Généralisés (GAM)

D'une manière analogue au GLM, le GAM dispose d'une version simplifiée : le modèle additif. Dans le modèle additif standard, la variable réponse est supposée suivre une loi normale dont la moyenne est expliquée par une combinaison non forcément linéaire des variables explicatives :

$$Y \sim \mathcal{N}(\mu, \sigma^2) \tag{5.15}$$

où

$$\mu = \alpha + \sum_{i=1}^{n} f_i(x_i) \tag{5.16}$$

Les fonctions f_i sont des fonctions quelconques d'une ou plusieurs variable(s). Elle peuvent être paramétriques ou non.

Les modèles additifs généralisés permettent de prendre en compte des effets non linéaires des variables explicatives, sans avoir à en spécifier la forme a priori. Ils sont une extension non paramétrique des GLM.

La famille des lois de probabilité de la variable réponse

Comme pour le GLM, la loi de probabilité suivie par la variable réponse doit être de la famille exponentielle (donc de la forme (5.3)).

Le modèle de régression

Avec les mêmes notations que pour le GLM, nous exprimons la moyenne μ_i avec une fonction non linéaire des prédicteurs :

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} f_j(x_{i,j}) \tag{5.17}$$

La seule différence avec le GLM réside donc dans les fonctions composant le prédicteur η.

Les fonctions du prédicteur

Le prédicteur peut contenir des fonctions trigonométriques, des fonctions polynomiales,. . .ainsi que d'autres fonctions qui n'entrent pas dans ces catégories. Les plus utilisées d'entre elles sont les fonctions splines et les fonctions loess. Elles permettent en effet un ajustement « souple »aux données et d'exercer un lissage de celles-ci. Nous n'utiliserons toutefois pas ces dernières, qui sont plus adaptées à la modélisation de séries temporelles.

6 La modélisation des flux de règlement futurs

Après avoir présenté la théorie des modèles linéaires généralisés et des modèles additifs généralisés au chapitre précédent, nous utilisons ces modèles pour l'étude des dates et des montants des flux de règlement des sinistres.

6.1 La modélisation des dates des flux de règlement

Nous voulons dans un premier temps modéliser le temps écoulé entre la date de survenance d'un sinistre et la date d'un flux à l'aide des variables exogènes suivantes :

- le numéro du flux : le premier flux se distingue en effet des autres flux par un délai plus long ;
- le montant déjà payé lors de la survenance d'un flux, à savoir 0 s'il s'agit du premier flux, et la somme des flux précédents sinon ;
- le temps écoulé entre la survenance d'un sinistre et la déclaration de ce sinistre ;
- le temps écoulé entre la date de survenance d'un sinistre et la date du flux précédent (s'il en existe un) ;
- l'activité du sinistré : la typologie des sinistres et leur cadence de règlement peut varier selon l'activité exercée par l'assuré (par exemple une activité peut être plus sujette à des règlements devant un tribunal, ce qui rallonge le délai entre la date du sinistre et la date du flux) ;
- la cause du sinistre ;
- l'évaluation du montant total du sinistre lors de l'ouverture du dossier ;
- l'année de survenance.

6.1.1 Une première approche par les GLM sur l'ensemble des flux

Le premier type de modèle utilisé pour cette modélisation est un modèle linéaire généralisé (GLM).

Nous supposons dans un premier temps qu'il s'agit d'un modèle de Poisson avec lien logarithmique : l'hypothèse sous-jacente est que le délai entre la date de survenance des sinistres et la date de versement d'un flux de règlement suit une loi de Poisson, dépendant des variables explicatives choisies. Autrement dit :

$$T \sim P\left(\mathbb{E}(T)\right)$$

avec :

$$log\mathbb{E}(T_{i,j}) = \alpha + \sum_{i=1}^{n} \beta_i X_i \tag{6.1}$$

L'analyse des résultats du GLM nous montre que toutes les variables sont significatives dans la détermination du modèle, à l'exception de certaines modalités des variables activité et cause du sinistre.
La déviance calculée dans notre modèle est de 408 200, contre 804 969 sans l'introduction de variables explicatives. La rapport de la déviance sur son degré de liberté est de 115, ce qui traduit un mauvais ajustement du modèle.

En choisissant comme hypothèse de loi de distribution une loi de Poisson, nous avons implicitement imposé un paramètre de dispersion égal à 1. Or nous n'avons aucune raison de penser que c'est le cas. Nous avons donc testé cette hypothèse :

```
        Overdispersion test

data:  glm.global
z = 41.091, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
    112.27
```

FIGURE 6.1 – Tests de l'hypothèse de sur-dispersion du GLM global

Nous observons ici que l'hypothèse d'équi-dispersion est rejetée, avec une sur-dispersion estimée à 112. Nous avons alors décidé de travailler avec un modèle de Poisson sur-dispersé, c'est-à-dire que le paramètre de dispersion ϕ n'est plus fixé à 1 mais est à estimer avec les autres paramètres du modèle.

Les paramètres associés aux variables exogènes ne sont cette fois plus tous significativement non nuls. Le nouveau modèle est formé du numéro du flux, , du temps écoulé entre la date de survenance et la date de déclaration du sinistre, et de l'année de survenance.

Dans ce modèle, l'introduction de variables exogènes ne permet de réduire la déviance que de 46% par rapport au modèle sans variables explicatives. Ceci n'est pas forcément très surprenant, dans la mesure où nous disposons de variables exogènes très différentes et qu'à première vue, leurs effets ne semblaient pas découler de transformation simples. Autrement, le rapport de la déviance normée sur son nombre de degrés de liberté est de 1,015. Cependant, l'analyse graphique des valeurs prédites par le GLM montre que ce dernier n'est pas adapté à nos données (cf figure 6.3 page suivante).

```
Call:
glm(formula = Delai.surv.flux ~ No.flux + Montant.deja.paye +
    Delai.surv.decla + Annee.surv, family = quasipoisson(link = log),
    data = Data.flux)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-31.716   -8.976   -1.426    6.682   33.653

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.335e+00  1.889e-02  335.32  < 2e-16 ***
No.flux           1.035e-01  6.477e-03   15.98  < 2e-16 ***
Montant.deja.paye 1.083e-05  2.966e-06    3.65 0.000266 ***
Delai.surv.decla  6.508e-04  5.840e-05   11.14  < 2e-16 ***
Annee.surv2012   -1.593e-01  1.881e-02   -8.47  < 2e-16 ***
Annee.surv2013   -4.481e-01  2.330e-02  -19.23  < 2e-16 ***
Annee.surv2014   -8.421e-01  2.734e-02  -30.80  < 2e-16 ***
Annee.surv2015   -1.474e+00  6.313e-02  -23.35  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 119.2067)

    Null deviance: 804969  on 3592  degrees of freedom
Residual deviance: 433907  on 3585  degrees of freedom
AIC: NA
```

FIGURE 6.2 – Estimation du GLM au global avec hypothèse de loi de Poisson surdispersée

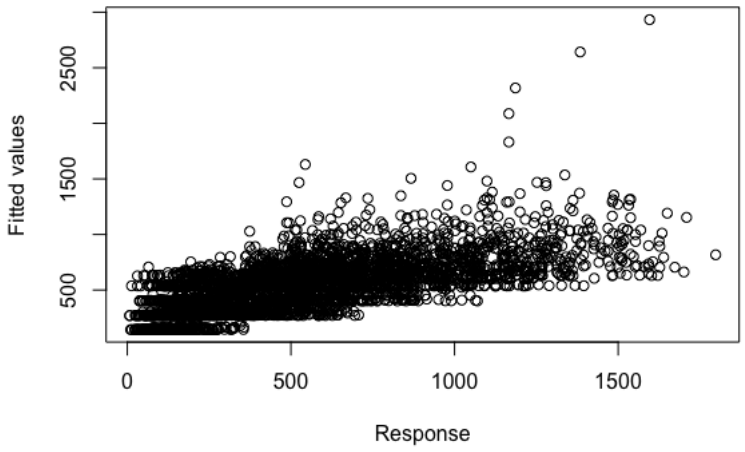


FIGURE 6.3 – Comparaison des valeurs prédites avec le GLM contre valeurs réelles

Nous avons aussi utilisé une loi gamma pour faire la régression. Les résultats obtenus étaient alors similaires à ceux obtenus avec la loi de Poisson. En effet, pour simuler une loi de poisson, England utilise une loi gamma en faisant correspondre moyenne et

variance.

Nous nous sommes alors tournés vers la théorie des modèles additifs généralisés (GAM), qui permettent des transformations plus complexes des variables explicatives.

6.1.2 L'utilisation des GAM sur l'ensemble des flux

La modélisation par un GLM n'ayant pas donné de résultats concluants, nous nous sommes dit que l'effet des variables explicatives n'était pas forcément linéaire. Nous nous sommes donc intéressés aux modèles additifs généralisés (GAM). En effet, si les GLM se basent sur des courbes de réponse paramétriques de premier ou deuxième ordre essentiellement, limitant ainsi la réponse à une droite ou une parabole, les modèles additifs sont une extension non paramétrique des GLM qui introduisent des courbes de réponse lissées à partir des données d'observation.

Nous utilisons de nouveau une loi de Poisson avec un lien logarithmique. Cinq paramètres sont alors significativement non nuls : le numéro du flux, le montant déjà payé, le temps écoulé entre la date de survenance et la date de déclaration du sinistre, le temps écoulé entre la date de déclaration et la date du règlement précédent, et l'évaluation du coût total du sinistre à l'ouverture du dossier.

```
Family: poisson
Link function: log

Formula:
Delai.surv.flux ~ s(No.flux) + s(Montant.deja.paye) + s(Delai.surv.decla) +
    s(Delai.decla.fluxprec) + s(Evaluation)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 6.203380   0.000813    7630   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df   Chi.sq p-value
s(No.flux)             8.944  8.998   4594.7  <2e-16 ***
s(Montant.deja.paye)   8.898  8.997    584.6  <2e-16 ***
s(Delai.surv.decla)    8.933  8.999  54459.1  <2e-16 ***
s(Delai.decla.fluxprec) 8.955  8.999 103335.6  <2e-16 ***
s(Evaluation)          8.983  9.000  20521.8  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.489   Deviance explained = 44.4%
UBRE = 125.32  Scale est. = 1         n = 3263
```

FIGURE 6.4 – Estimation du GAM au global avec hypothèse de loi de Poisson

Nous pouvons légèrement augmenter les résultats de la modélisation par le GAM en croisant la variable délai entre la date de survenance et la date de déclaration avec la variable délai entre la date de déclaration et la date du flux précédent.

Nous testons de nouveau un modèle de Poisson sur-dispersé : tous les paramètres ne sont alors plus significativement non nuls ; la variable du montant déjà payé n'est

pas significative dans le modèle. Le modèle peut là encore être amélioré en croisant les mêmes variables que précédemment, et nous obtenons un R2 ajusté de 0,498.

En nous intéressant aux fonctions de lissage (figure 6.5), nous observons toutefois un comportement différent selon qu'il s'agit du premier flux de règlement ou d'un autre flux.

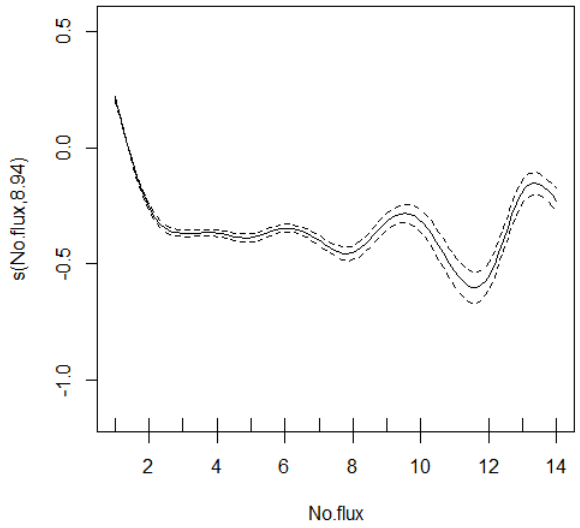


FIGURE 6.5 – Fonction de lissage du numéro de paiement associé au GAM

Ceci confirme donc les conclusions de l'analyse exploratoire des données, à savoir qu'il nous faut étudier le premier flux à part. En effet, le premier flux met généralement plus d'un an à arriver, alors que les flux suivants se renouvellent plus fréquemment, un paiement en appelant généralement un autre. Nous étudions une branche longue. Le premier paiement arrive souvent après une procédure judiciaire ou des avis d'experts, ce qui peut expliquer le délai plus important.

6.1.3 La séparation du premier flux des autres flux

L'étude précédente nous a montré qu'il nous fallait distinguer le premier flux de règlement des autres flux dans notre modélisation.

Le délai entre la date de survenance et la date du premier flux de règlement

Pour l'analyse du délai entre la date de survenance est la date du premier flux, nous disposons des variables suivantes : le délai entre la survenance et la déclaration du sinistre, l'activité, la cause du sinistre, l'évaluation du sinistre lors de l'ouverture du

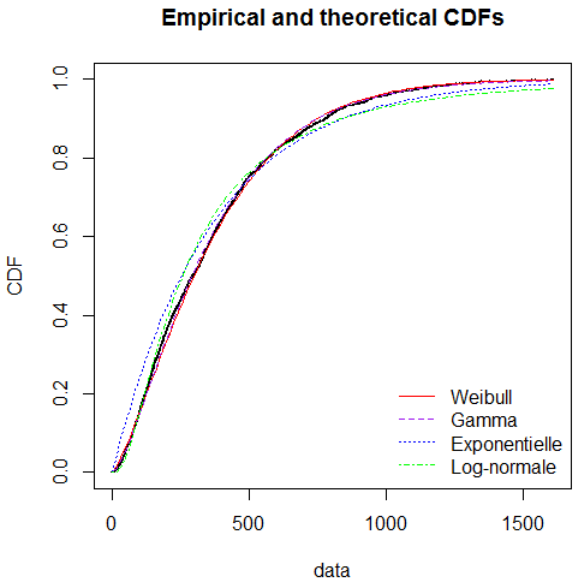
dossier et l'année de survenance.

Le résultat le plus concluant est obtenu avec la loi de Poisson sur-dispersée et son lien canonique (lien logarithmique). Les seules variables significatives sont l'année de survenance et le délai entre la survenance et la déclaration du sinistre. L'apport des variables explicatives ne permet pas de réduire significativement la déviance (255 034 contre 390 496 pour le modèle sans variables explicatives), et les données simulées avec le GLM ne concordent pas avec les données réelles.

Nous avons aussi essayé d'ajuster un GAM pour la date du premier flux de règlement. Les résultats ne sont pas plus concluants qu'avec le GLM.

Nous avons alors décidé d'ajuster une loi paramétrique pour décrire le délai entre la survenance du sinistre et le premier flux de règlement. Il s'agit d'une variable positive et continue. Nous testerons donc des lois à support sur $\mathbb{R}_+$: la loi gamma, la loi de Weibull, la loi exponentielle et la loi log-normale.

En ajustant ces lois par maximum de vraisemblance, nous obtenons que les lois gamma et Weibull s'ajustent bien mieux aux données que la loi log-normale et la loi exponentielle.



s

FIGURE 6.6 – Comparaison des fonctions de distribution théoriques pour la date du premier flux

Pour choisir entre la loi de Weibull et la loi gamma, nous nous basons sur l'étude du

qqplot :

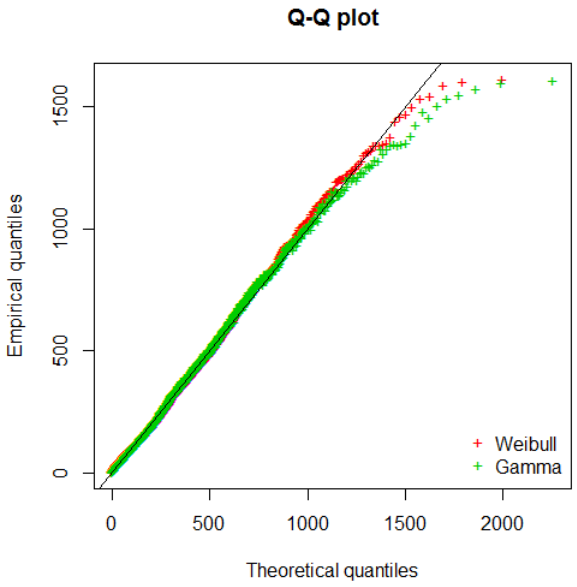


FIGURE 6.7 – Comparaison des qqplot pour la date du premier flux

La loi de Weibull sur-estime moins les quantiles élevés que la loi gamma. Notons que jusqu'à trois ans, nous avons une forte probabilité d'avoir un nouveau flux. Aux alentours de trois ans, la probabilité d'avoir un nouveau paiement est légèrement sous-estimée. A contrario, au-delà de trois ans, plus le temps passe, plus la modélisation sur-estime la probabilité de survenance d'un nouveau flux.
De plus, en considérant toutes les années de survenance, nous considérons un échantillon où les courts délais de déclaration sont sur-représentés. Pour corriger ce biais, nous restreignons l'étude des premiers flux de règlement aux sinistres survenus en 2011.

Avec ce nouvel échantillon de 443 observations, la loi la plus en adéquation avec les données est une loi de Weibull. L'ajustement est meilleur que précédemment, notamment pour les valeurs élevées. La moyenne théorique est alors supérieure à la moyenne théorique du précédent ajustement.

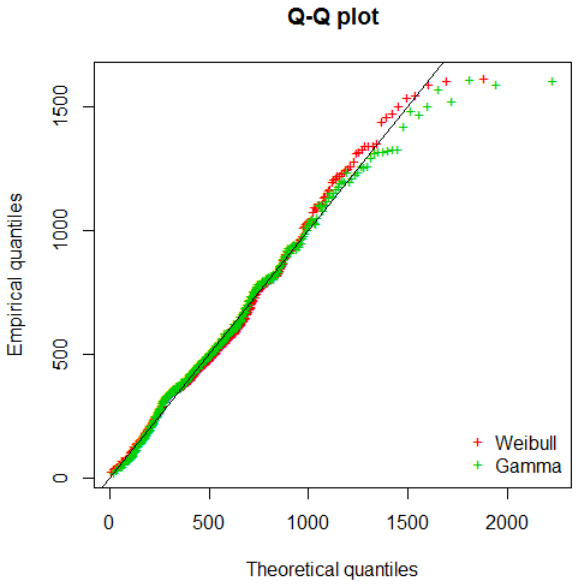


FIGURE 6.8 – Comparaison des qqplot pour la date du premier flux sur les sinistres survenus en 2011

La dernière étape consiste à valider le choix de la loi par un test statistique. En effet, aussi utiles qu'elles soient et bien que l'ajustement soit ici de bonne qualité, les méthodes d'ajustement graphique ne constituent pas une réponse mathématiquement rigoureuse au problème de l'adéquation des lois.
Il existe deux principaux tests pour vérifier l'adéquation d'une loi à des données : le test de Kolmogorov-Smirnov pour tester une loi continue, et le test du khi-deux pour tester une loi discrète.
Seul le test de Kolmogorov nous intéresse dans ce cas. Ce test est construit sur la distance entre la fonction de répartition théorique (loi testée) et la fonction de répartition empirique (loi des données) pour chaque observation. La statistique de ce test est :

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

avec F_n la fonction de répartition empirique et F la fonction de répartition théorique. Sa loi est donnée par la table de Kolmogorov.
Dans notre cas, le test d'adéquation ne rejette pas l'hypothèse d'égalité des distributions pour la loi de Weibull.

Nous acceptons donc la modélisation de la date du premier flux de règlement par une loi de Weibull.

L'analyse des temps incrémentaux entre deux flux de règlement

Pour les flux autres que le premier règlement, nous considérons de plus le délai entre la déclaration du sinistre et la survenance du dernier flux de règlement.

Nous avons d'abord essayé d'ajuster un GLM avec une loi de Poisson sur-dispersée, mais le modèle ne s'ajustait pas suffisamment bien à nos données.

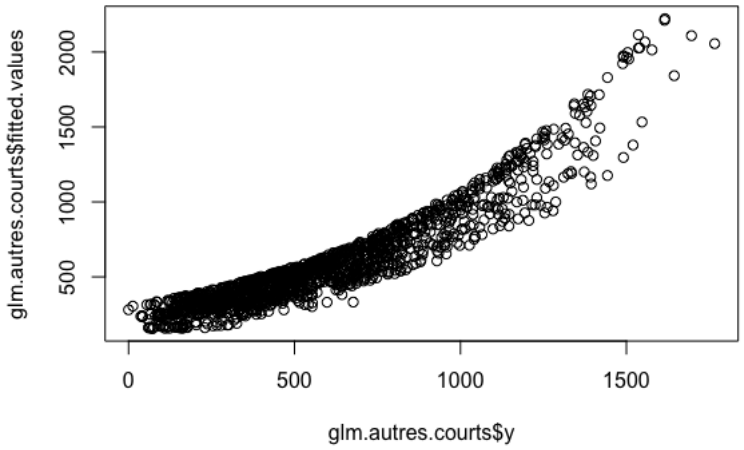


FIGURE 6.9 – GLM pour une loi gamma pour les flux autres que le premier règlement

De plus, en procédant aux première simulations, nous nous sommes rendu compte que trop de paiements s'enchainent avec ce modèle. En réalité, le paiement des sinistres s'effectue par vagues successives. En effet, l'évaluation du sinistre entraîne une première vague de paiements. Ensuite, soit le sinistre est clos, soit il apparaît un contentieux entre les parties, ce qui implique le temps d'attente du jugement et éventuellement une deuxième vague de paiements.

Nous avons donc voulu séparer les sinistres selon leur état : un état court pour les sinistres clos rapidement, et un état long pour les sinistres avec un plus long développement. Notre première idée a été d'utiliser le modèle logit pour expliquer l'état du sinistre à l'aide de variables exogènes, mais aucune ne s'est révélée significativement non nulle. Nous modélisons donc l'état du sinistre par une loi de Bernoulli, avec une probabilité de 0,15 que ce soit un sinistre long.

Avec cette nouvelle indicatrice, le meilleur résultat est obtenu avec une loi gamma avec comme fonction de lien l'identité, fonction non utilisée précédemment car l'algorithme ne convergeait pas en considérant tous les sinistres.

Les seules variables significatives dans ce modèle sont l'année de survenance, le délai entre la survenance du sinistre et le dernier flux de règlement, l'évaluation initiale du montant total du sinistre.

```
Call:
glm(formula = Delai.surv.flux ~ Annee.surv + Delai.surv.fluxprec +
    Evaluation + Court, family = Gamma(link = identity), data = tempsautres)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.06594  -0.13712  -0.03095   0.09918   0.88712

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.952e+02  1.553e+01  38.326  < 2e-16 ***
Annee.surv2012      -1.522e+01  9.749e+00  -1.561 0.118656
Annee.surv2013      -3.685e+01  9.717e+00  -3.792 0.000156 ***
Annee.surv2014      -7.950e+01  1.013e+01  -7.844 8.95e-15 ***
Annee.surv2015      -1.124e+02  1.112e+01 -10.105  < 2e-16 ***
Delai.surv.fluxprec  9.134e-01  1.436e-02  63.617  < 2e-16 ***
Evaluation           3.521e-03  8.587e-04   4.100 4.39e-05 ***
Court1              -4.093e+02  1.344e+01 -30.454  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.04120391)

    Null deviance: 398.932  on 1322  degrees of freedom
Residual deviance:  52.503  on 1315  degrees of freedom
AIC: 16456
```

FIGURE 6.10 – GLM pour une loi gamma pour les flux autres que le premier règlement

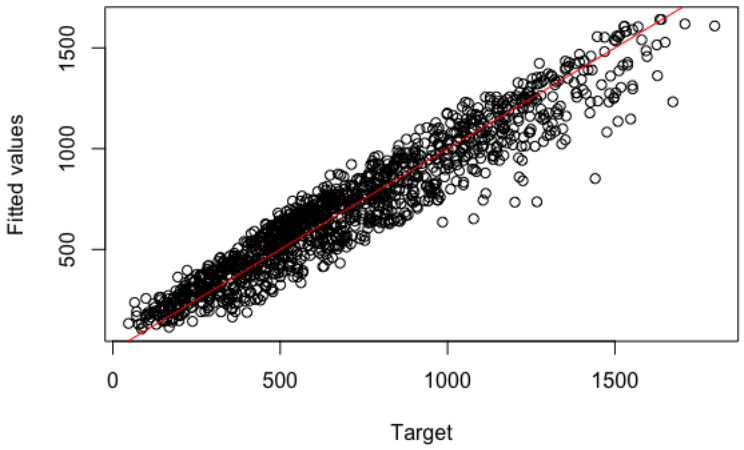


FIGURE 6.11 – GLM pour une loi de Poisson avec sur-dispersion pour les flux autres que le premier règlement

L'absence de la variable du numéro du flux nous montre que cette dernière n'était

significative que pour séparer le premier flux des flux suivants.

Nous avons donc établit un modèle pour les dates des flux de règlement :
- la date du premier flux de règlement est modélisée par une loi de Weibull ;
- la date des flux de règlement suivants est fonction de cinq facteurs explicatifs : le temps entre la survenance et le précédent flux, l'année de survenance, l'évaluation du sinistre à l'origine et une indicatrice permettant de distinguer les sinistres réglés rapidement des sinistres à plus long développement.

6.2 La modélisation des montants de règlement

6.2.1 L'utilisation des variables exogènes

Pour modéliser les montants des règlements, notre première idée fut d'utiliser de nouveau les GLM et les GAM. Nous avons testé ces modèles avec les variables suivantes :
- le numéro du flux ;
- le montant déjà payé ;
- le temps écoulé depuis le dernier règlement (ou depuis la date de déclaration pour les premiers flux de règlements) ;
- l'activité ;
- la cause ;
- l'évaluation initiale du sinistre ;
- l'année de survenance.

Nous faisons l'hypothèse d'une loi gamma avec son lien canonique (lien logarithmique).
Seuls le nombre de flux déjà effectués, le temps écoulé depuis le dernier flux et l'évaluation initiale sont significativement non nuls. Les résultats graphiques ne sont cependant pas satisfaisants.

Nous avons ensuite travaillé avec les GAM. Encore une fois, les résultats ne sont pas concluants. L'étude des fonctions de lissage ne nous a pas apporté d'explications.

C'est pourquoi, comme pour les dates de règlements, nous avons travaillé sans facteurs explicatifs.

6.2.2 L'ajustement paramétrique des montants de règlement

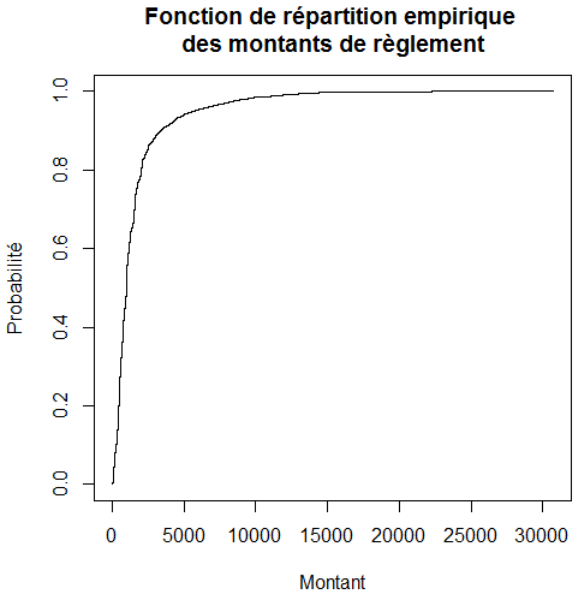


FIGURE 6.12 – Fonction de distribution empirique des montants de règlement

Les sinistres de responsabilité civile sont souvent des sinistres à queue épaisse. Nous devrons faire attention aux valeurs extrêmes, bien que la fonction de distribution empirique ne révèle pas d'irrégularité particulière.

Nous essayons d'ajuster différentes lois de probabilités par maximum de vraisemblance ; l'algorithme ne converge toutefois pas pour l'ajustement d'une loi de Pareto. Les résultats obtenus avec une loi log-normale, une loi exponentielle et une loi gamma sont les suivants :

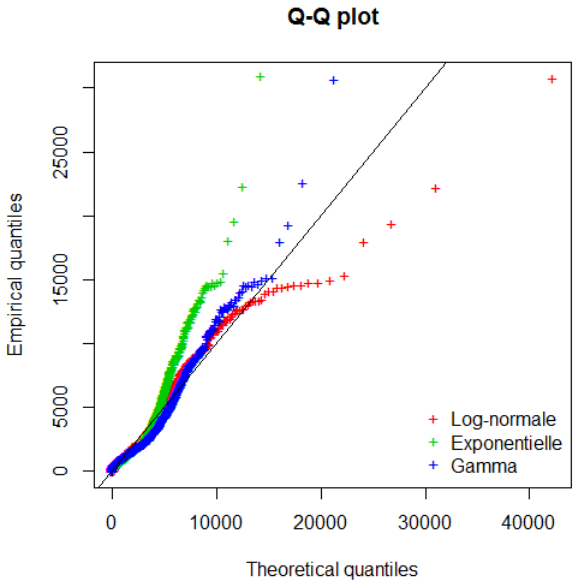


FIGURE 6.13 – Ajustement de différentes lois aux montants de règlement

Si la loi gamma semble être relativement bien adaptée aux données jusqu'à 10 000€, au-delà de ce seuil les valeurs des quantiles théoriques divergent des valeurs des quantiles empiriques.

Nous séparons donc notre échantillon en deux échantillons : les règlements de moins de 10 000€ d'une part, et les règlements supérieurs à 10 000€ d'autre part.

L'ajustement de la queue de distribution

Pour le choix de la loi de la queue de distribution, nous avons comparé l'ajustement d'une loi gamma, d'une loi log-normale, d'une loi de Burr, d'une loi de Pareto et d'une loi de Weibull. L'algorithme pour l'ajustement de la loi de Pareto n'a pas convergé, et la loi de Weibull n'a pas donné de résultats intéressants. Nous avons utilisé les fonctions de distribution pour déterminer la loi la plus adaptée à nos montants :

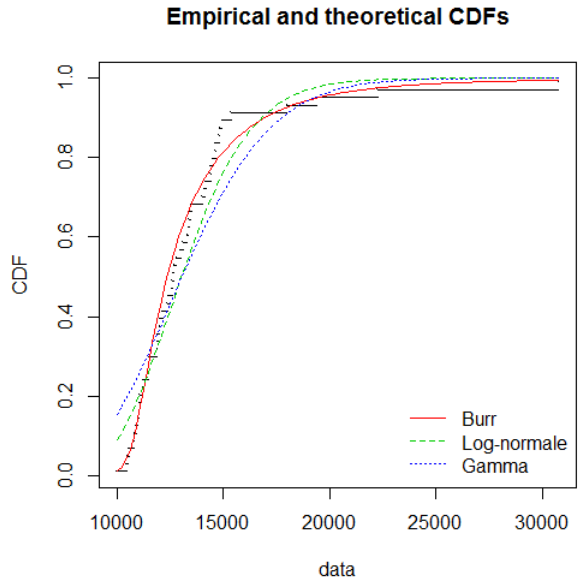


FIGURE 6.14 – Ajustement de différentes lois aux valeurs extrêmes des montants de règlement

À la vue du graphique des fonctions de distribution, la loi de Burr est la mieux adaptée pour les règlements d'un montant supérieur à 10 000€.
Le test de Kolmogorov-Smirnov ne rejette pas cette hypothèse.

L'ajustement global d'une loi mélange aux montants de règlements

En dehors de la queue de distribution, nous voulons ajuster nos données à une loi plus classique (loi gamma, log-normale notamment). Les meilleurs résultats sont obtenus avec une loi log-normale.

Ayant déterminé les lois du mélange, nous n'avons plus qu'à en estimer les paramètres. Nous nous basons pour cela sur les résultats suivants : en notant X et Y deux variables aléatoires réelles, et Z la variable aléatoire mélangée telle que $\mathbb{P}(Z = X) = p$

—

$$F_z(t) = p \cdot F_X(t) + (1-p) \cdot F_Y(t) \tag{6.2}$$

où $F_X(t)$ représente la fonction de densité de la variable X ;

—

$$\mathbb{E}(Z) = p \cdot \mathbb{E}(X) + (1-p) \cdot \mathbb{E}(Y) \tag{6.3}$$

—

$$\mathbb{V}(Z) = p \cdot \mathbb{V}(X) + (1-p) \cdot \mathbb{V}(Y) + p \cdot (1-p) \cdot (\mathbb{E}(X) - \mathbb{E}(Y))^2 \tag{6.4}$$

L'ajustement global à nos données est donc l'ajustement du mélange d'une loi log-normale pour le coeur de la distribution et d'une loi de Burr pour les valeurs extrêmes. Le résultat en terme de fonction de distribution est alors le suivant :

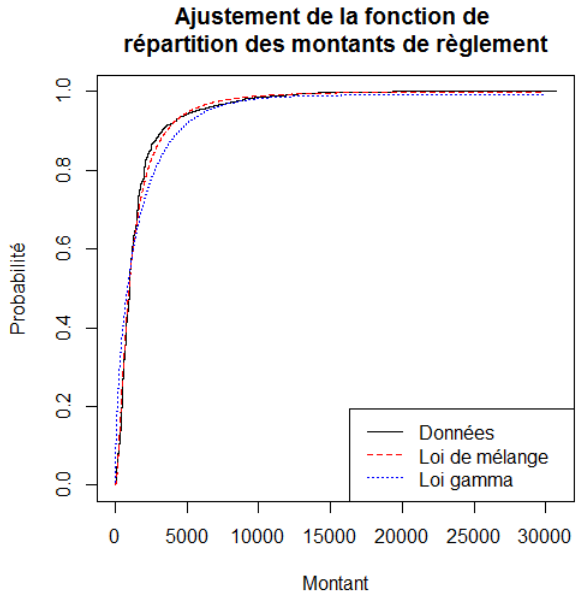


FIGURE 6.15 – Ajustement de de la loi mélange aux montants de règlement

Cet ajustement permet d'obtenir une modélisation satisfaisante des montants de règlements.Si la modélisation par une loi gamma aurait pu convenir, le mélange de lois semble graphiquement plus adapté. Enfin, le test de Kolmogorov-Smirnov ne rejette pas l'hypothèse d'adéquation de la loi mélange.

Synthèse des résultats des modèles

Finalement, nous avons retenu trois lois paramétriques ajustées à nos données, les modèles introduisant des variables exogènes ne fournissant par de bons résultats. Nos trois modèles retenus sont alors les suivants :

- une loi de Weibull pour la modélisation des dates de cloture ;
- une loi de Weibull pour la modélisation de la date du premier règlement, et quatre variables exogènes pour la modélisation de la date des flux de règlement suivants ;
- le mélange d'une loi log-normale et d'une loi de Burr pour les montants réglés.

Troisième partie

Le calcul des réserves et la comparaison des différents modèles

Nous avons déterminé un modèle sur les données détaillées. Nous voulons maintenant comparer les résultats obtenus avec ce modèle avec les résultats obtenus avec les méthodes standards, et tenter d'expliquer les différences.

Les méthodes standards les plus répandues et qui nous serviront de référence sont la méthode de Chain Ladder, le modèle de Mack et la méthode du bootstrap. Ces trois méthodes supposent la reproductibilité de la sinistralité passée pour déterminer le futur. La méthode de Chain Ladder est une méthode déterministe qui permet d'obtenir la réserve à l'ultime ; les deux autres méthodes sont des méthodes stochastiques qui permettent de plus d'estimer l'incertitude de ces provisions ultimes, et plus précisément fournit une distribution des réserves dans le cas du bootstrap.

Nous n'avons pas introduit de facteur de queue afin de faciliter la comparaison des modèles. Nous faisons donc l'hypothèse qu'il n'y aura pas de règlement après 5 années de développement, bien que cette hypothèse soit contraire à la réalité. Ceci nous permettra de comparer plus facilement les modèles entre eux et notamment les résultats obtenus. Lorsque nous parlerons d'ultime, nous désignerons donc les paiements cumulés au bout de 5 années de développement, et les réserves sont les réserves associées à ces années.

7 Les résultats obtenus avec les différentes méthodes de provisionnement

Nous avons vu dans la première partie que les méthodes sur données agrégées travaillent sur des triangles de règlements cumulés. Le triangle formé à partir de nos données est le suivant :

SY \ DY	1	2	3	4	5
2011	42 033	438 434	1 068 900	1464 325	1 729 272
2012	73 009	743 187	1 193 820	1 547 185	
2013	137 406	671 602	1 111 338		
2014	214 934	1 107 144			
2015	284 617				

TABLE 7.1 – Triangle des paiements cumulés

7.1 Les résultats obtenus avec la méthode de Chain Ladder

En appliquant la méthode de Chain Ladder décrite au chapitre 2, nous obtenons le triangle de règlements complété suivant :

SY \ DY	1	2	3	4	5	Ultime	Réserves
2011	42 033	438 434	1 068 900	1464 325	1 729 272	1 729 272	0
2012	73 009	743 187	1 193 820	1 547 185	1 827 124	1 827 124	279 939
2013	137 406	671 602	1 111 338	1 479 107	1 746 729	1 746 729	635 391
2014	214 934	1 107 144	2 015 714	2 682 764	3 168 169	3 168 169	2 061 025
2015	284 617	1 802 745	3 282 156	4 368 303	5 158 680	5 158 680	4 874 063
						13 629 974	7 850 418

λ_k		6.3339	1.8206	1.3309	1.1809

TABLE 7.2 – Application de Chain Ladder

Nous obtenons donc une réserve à l'ultime de 7.85M€. Pour l'année 2011, le règlement des sinistres a eu lieu plus tardivement que pour les autres années.

La question de la qualité des données peut se poser pour les années 2012 et 2013. Ces deux années ont en effet des charges ultimes équivalentes à celle de 2011, alors que les règlements de la première année de développement sont largement supérieurs. L'évolution entre l'année de développement 1 et 2 pour 2013 et l'année de développement 2 et 3 pour 2012 semblent faibles. La charge ultime des ces années pourrait donc être sous-estimée. De même, à règlements de la première année de développement initial presque égaux, l'année 2015 a une charge ultime presque deux fois supérieure à celle de l'année 2014. IL pourrait s'agir d'une sous-estimation des règlements de 2014, surtout entre la première et la seconde année de développement, ou d'une sur-estimation des montants pour les sinistres survenus en 2015.
De fait, en vérifiant les hypothèses sous-jacentes à la méthode de Chain Ladder, nous observons que nous ne pouvons pas l'appliquer à nos données :

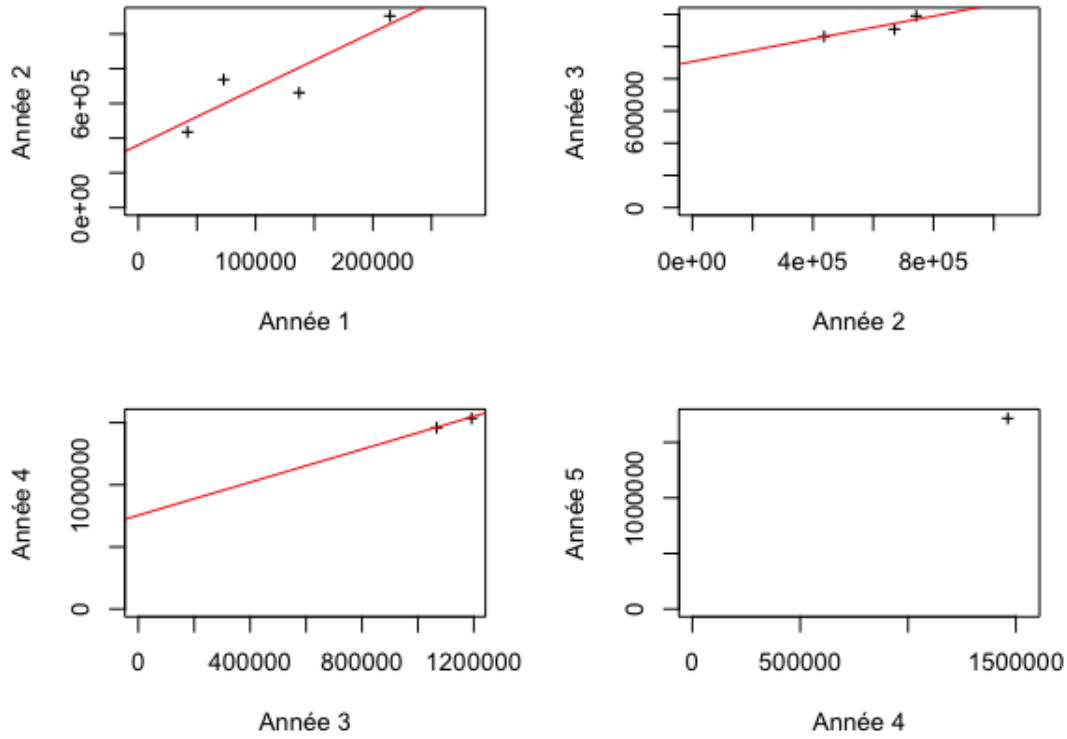


FIGURE 7.1 – Vérification de l'hypothèse de linéarité de Chain Ladder

Malgré le faible nombre de données, les points semblent relativement bien alignés. Cependant, les droites de régression ne passent pas toutes pas l'origine, ce qui contredit l'hypothèse (H2) de Chain Ladder. Au contraire, il semblerait qu'il existe un terme positif à rajouter. Nous avons alors pensé à la méthode de London-Chain, qui est une variante de Chain Ladder. Elle suppose en effet l'existence d'un facteur additif (une constante)

relatif à chaque année de survenance, supposé nul avec Chain Ladder :

$$C_{i,j} = \lambda_j C_{i,j-1} + \alpha_j \tag{7.1}$$

Nous nous retrouvons donc avec un système linéaire à résoudre. La résolution de ce système se fait par les moindres carrés ordinaires :

$$(\hat{\lambda}_k, \hat{\alpha}_k) = \text{argmin} \left\{ \sum_{i=1}^{n-k} \left(C_{i,k+1} - \alpha_k - \lambda_k C_{i,k} \right)^2 \right\} \tag{7.2}$$

Les résultats obtenus sont équivalents à ceux obtenus par la méthode de Chain Ladder, et même un peu supérieurs, et cette méthode n'est pas le sujet du mémoire, donc nous n'approfondissons pas l'analyse de ces résultats. De plus, l'analyse de la régression à la figure 7.1 montre que le terme additif n'est pas le même pour toutes les années.

7.2 L'application du modèle de Mack

Le modèle de Mack est l'équivalent stochastique de la méthode de Chain Ladder. Si les résultats sont égaux en moyenne, le modèle de Mack permet aussi d'obtenir une estimation de l'erreur de prédiction. Cette erreur de prévision s'élève à 2,26M€ dans notre cas, ce qui est relativement élevé en comparant aux 7,85M€ de provisions.

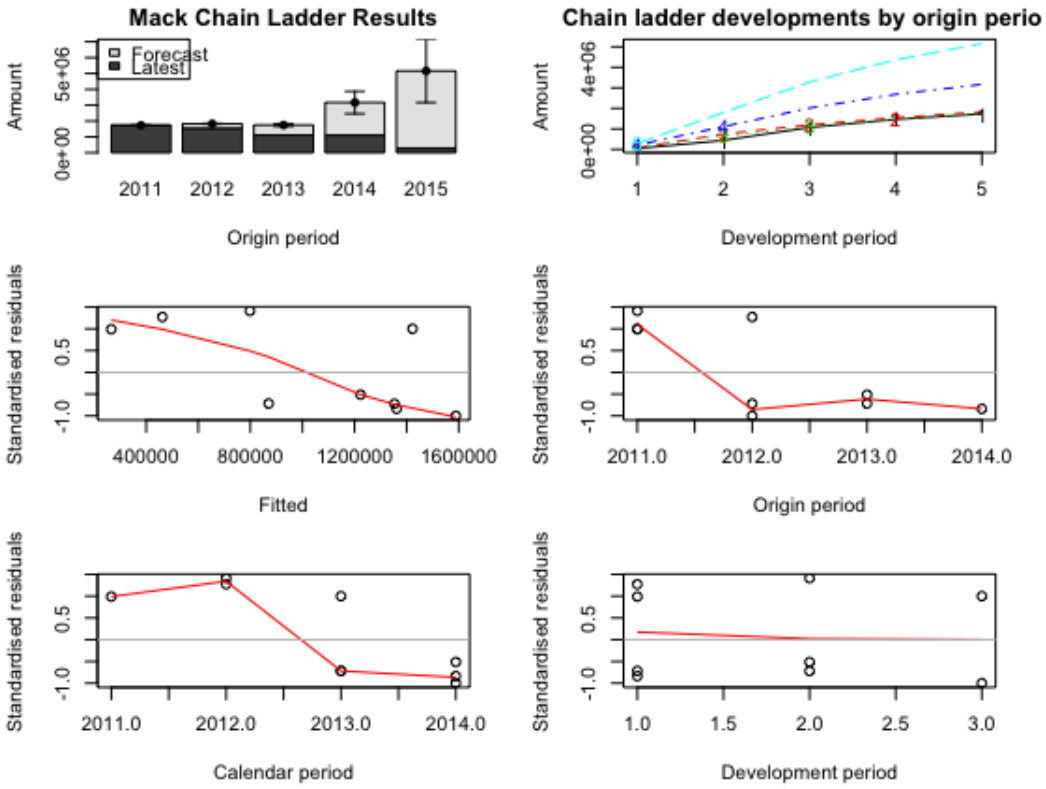


FIGURE 7.2 – Le modèle de Mack

Comme le montre le premier graphique, pour les dernières années, la prévision est supérieure aux règlements effectués (ce qui semble logique), mais l'erreur de prédiction atteint près de 50% de la prévision pour la dernière année. Si nous avions représenté l'estimation des coefficients de passage pour chaque année de développement, nous aurions vu qu'en effet l'intervalle de confiance explose pour les années les plus récentes. L'analyse des résidus nous montre que le modèle n'est pas adapté aux données : les résidus devraient avoir une répartition aléatoire entre -2 et 2, sans motif apparent.

Ceci nous montre donc que les résultats obtenus avec la méthode de Chain Ladder souffrent d'une grande incertitude, et doivent être considérés avec précaution.

7.3 La méthode du bootstrap

La dernière méthode de référence qui nous intéresse est la méthode du bootstrap. C'est en effet une méthode stochastique, basée sur la méthode de Chain Ladder, qui permet l'estimation de la distribution des réserves. L'idée de la méthode est de ré-échantilloner les résidus obtenus avec la méthode de Chain Ladder un certain nombre de fois, ce qui permet d'appliquer Chain Ladder sur un certain nombre de triangles différents et d'obtenir la distribution des réserves.

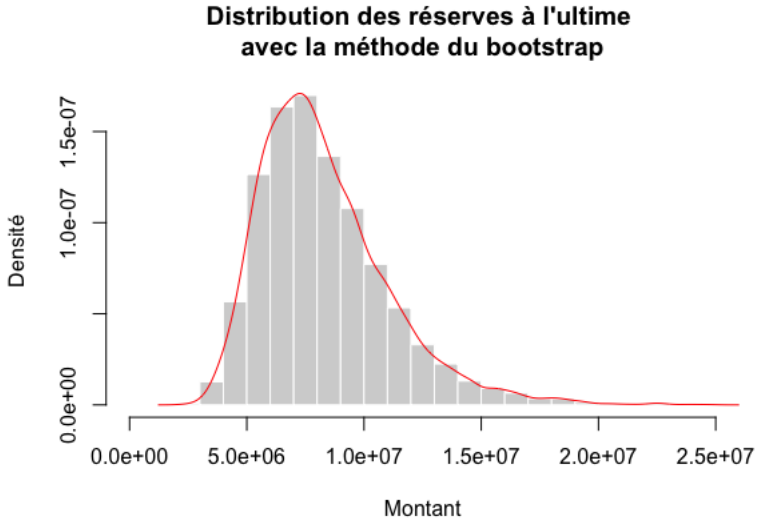


FIGURE 7.3 – La distribution des réserves obtenue avec la méthode du bootstrap comparée à une loi log-normale

Nous pouvons observer une petit disymétrie dans la distribution. Mais le point le plus important à noter est la forte dispersion des réserves.

Nous pouvons regarder un peu plus en détail les résultats obtenus et les intervalles de confiance avec un bootstrap de 10 000 simulations :

```
BootChainLadder(Triangle = cum.triangle, R = 10000, process.distr = "od.pois")

        Latest Mean Ultimate Mean IBNR  IBNR.S.E  IBNR 75%   IBNR 95%
2011 1,729,272     1,729,272         0         0         0          0
2012 1,547,185     1,833,004   285,819   149,291   373,857    568,526
2013 1,111,338     1,755,972   644,634   225,953   779,452  1,050,863
2014 1,107,144     3,202,404 2,095,260   589,861 2,449,155  3,170,449
2015   284,617     5,581,713 5,297,096 2,587,385 6,541,288 10,163,815

                    Totals
Latest:          5,779,556
Mean Ultimate:  14,102,364
Mean IBNR:       8,322,808
IBNR.S.E         2,827,085
Total IBNR 75%:  9,725,406
Total IBNR 95%: 13,542,676
```

FIGURE 7.4 – Les résultats du bootstrap

Aussi bien l'estimation moyenne que la volatilité des réserves sont supérieures à celles du modèle de Mack. L'écart d'estimation le plus important est pour l'année 2015. L'intervalle de confiance obtenu n'est pas intéressant, notamment pour l'année 2015 où les bornes à 95% varient de 1,9M€ à 11,7M€.

Finalement, ce que nous pouvons retenir est une grande volatilité des résultats obtenus. Ceci est à nuancer par le fait que ces modèles standards ne peuvent rigoureusement pas être appliqués tels quels. Le point commun de ces trois méthodes est Chain Ladder, qui est inadaptée à nos données.
Il existe certes d'autres méthodes de provisionnement qui ne reposent pas sur les hypothèses de Chain Ladder, et qui pourraient obtenir de meilleurs résultats. Mais celles-ci ne sont pas le sujet du mémoire donc nous ne nous attarderons pas dessus.
Nous allons maintenant étudier les résultats obtenus par notre modèle individuel.

7.4 L'estimation des réserves avec le modèle ligne à ligne

Après avoir utilisé les méthodes standards afin d'avoir des résultats de référence, nous appliquons notre modèle individuel à nos données.

Notre modèle est basé sur des simulations. Chaque simulation a un déroulement identique :

(1) simuler la date de clôture des sinistres à partir de la fonction paramétrique déterminée au chapitre 6 ;
(2) simuler une date de paiements entre le 31/12/2015 et la date de clôture. Si le sinistre n'a encore donné lieu à aucun paiement, alors la date du premier règlement

est distribuée selon une loi de Weibull ; s'il y a déjà eu des paiements d'effectués, la date de règlement est fonction des quatre facteurs explicatifs retenus aux chapitre précédent ;

(3) simuler un montant de règlement associé à la date de paiement simulée ; ce montant est distribué selon un mélange d'une loi log-normale et d'une loi de Burr ;

(4) répéter les opération (2) et (3) tant que le sinistre n'est pas clos.

Nous avons effectué 1000 simulations de ce modèle. Afin de faciliter la comparaison avec nos méthodes agrégées, nous avons agrégé ces simulations dans des triangles de liquidation.
Voici le triangle moyen que nous obtenons :

SY \ DY	1	2	3	4	5	Ultime	Réserves
2011	42 033	438 434	1 068 900	1464 325	1 729 272	1 729 272	0
2012	73 009	743 187	1 193 820	1 547 185	1 621 989	1 621 989	74 804
2013	137 406	671 602	1 111 338	1 260 122	1 461 678	1 461 678	350 340
2014	214 934	1 107 144	1 498 226	2 062 023	2 393 063	2 393 063	1 285 919
2015	284 617	876 483	1 624 387	2 097 724	2 381 734	2 381 734	2 097 117
						9 587 736	3 808 180

TABLE 7.3 – Triangle moyen obtenu avec le modèle individuel

Nous obtenons aussi une distribution du montant des réserves avec notre modèle individuel :

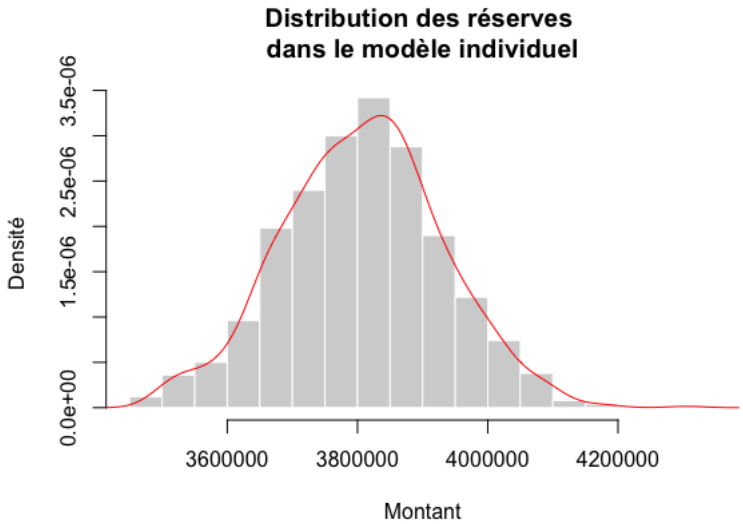


FIGURE 7.5 – Distribution des réserves ultimes obtenue avec le modèle individuel

Nous pouvons observer que la distribution est légèrement asymétrique. Les résultats du modèle individuel sont intéressants, tant sur le montant moyen des réserves que sur la volatilité du résultat obtenu. Nous représentons maintenant le développement des sinistres selon notre modèle individuel, afin de chercher d'où provient l'écart entre nos résultats.

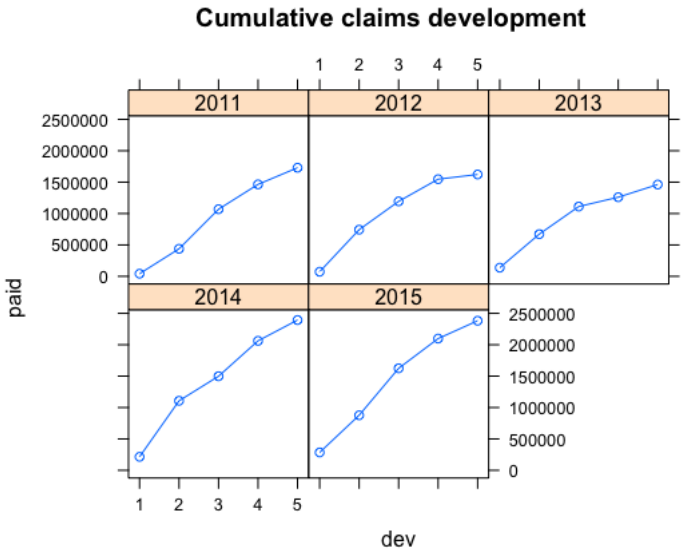


FIGURE 7.6 – Le développement des sinistres selon le modèle individuel

Le graphique ci-dessus permet d'étudier le développement des sinistres par année de survenance. Il peut être croisé avec le tableau des coefficients de développement pour chaque année :

SY \ DY	1	2	3	4
2011	10.43	2.44	1.37	1.18
2012	10.18	1.61	1.30	1.05
2013	4.89	1.65	1.13	1.16
2014	5.15	1.35	1.38	1.16
2015	3.08	1.85	1.29	1.14

TABLE 7.4 – Coefficients de développement par année de survenance du modèle individuell

Le modèle individuel semble sous-estimer légèrement les facteurs de développement des sinistres. Cet écart provient peut-être de la modélisation des flux futurs par un GLM : en effet, la représentation graphique était globalement correcte, mais les valeurs prédites étaient assez volatiles. Le modèle pourrait donc avoir tendance à allonger les temps entre les règlements, et donc à sous-estimer le montant total des sinistres. Cette

critique est à relativiser tout de même : nous ne disposons que d'un faible historique, ce qui ne permet pas d'avoir assez de recul pour bien connaître le développement des sinistres.
De plus, la figure 7.6 montre que si la charge ultime des sinistres des années 2012 et 2013 pourrait être sous-estimée, les années 2014 et 2015 semblent elles bien modélisées. Cette interprétation nous est confirmée par la figure 7.7 ci-dessous, qui permet une comparaison du développement des sinistres par année de survenance.

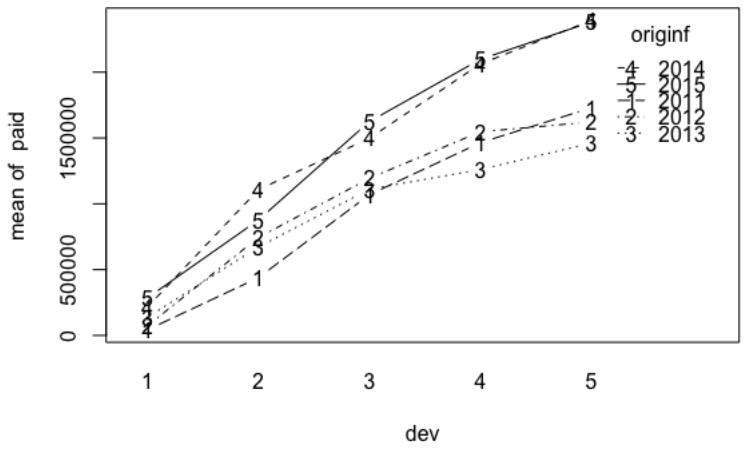


FIGURE 7.7 – Comparaison du développement des sinistres

7.5 La comparaison des résultats du modèle individuel avec ceux des modèles standards

En comparant les résultats du modèle individuel à ceux obtenus avec Chain Ladder, nous retombons sur les analyses précédentes. D'une manière générale, la méthode de Chain Ladder a tendance à régler plus rapidement que notre modèle individuel au départ, notamment la première année. Au global, le modèle individuel estime une provision inférieure de 4M€ à celle estimée par les méthodes traditionnelles. Cependant, nous avons constaté que le modèle individuel sous-estime peut être légèrement la charge ultime des sinistres ; l'écart réel devrait donc être moins important.

Nous pouvons aussi comparer les intervalles de confiance pour les provisions obtenus avec respectivement le modèle de Mack, la méthode bootstrap et notre modèle individuel :

	Année	2012	2013	2014	2015	Global
Mack	moyenne	279 939	635 391	2 061 025	4 874 063	7 850 418
	borne inf	263 345	549 466	1 356 047	2 882 182	5 051 040
	borne sup	296 533	721 316	2 766 003	6 865 944	10 649 796
Bootstrap	moyenne	285 819	644 634	2 095 260	5 297 096	8 332 808
	borne inf	68 750	277 420	1 130 575	1 892 779	4 339 911
	borne sup	631 034	1 148 686	3 446 136	11 709 430	15 281 411
Mod. ind.	moyenne	100 505	384 755	1 275 387	2 047 072	3 807 719
	borne inf	65 835	315 747	1 140 666	1 871 200	3 393 448
	borne sup	143 984	464 465	1 416 556	2 218 291	4 243 296

TABLE 7.5 – Comparaison des intervalles de confiance à 95% pour les réserves

Les bornes de l'intervalle de confiance à 95% sont inférieures dans le modèle individuel aux mêmes bornes obtenues avec les méthodes traditionnelles, comme cela était prévisible. Il est toutefois intéressant de noter que la longueur de l'intervalle de confiance est plus faible avec le modèle de Mack pour la première année de développement, alors que les rôles s'inversent par la suite. La méthode bootstrap ne fournit elle pas de résultats intéressants, ce qui s'illustre en représentant la distribution des réserves obtenue dans nos différents modèles.

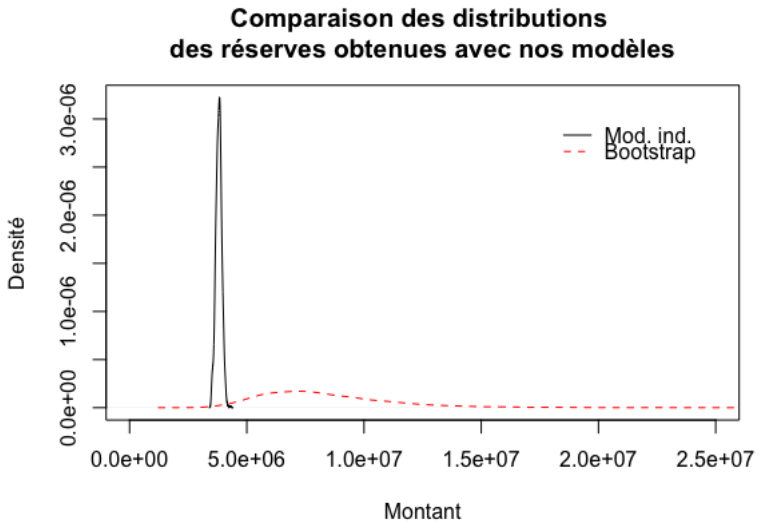


FIGURE 7.8 – Comparaison de la distribution des réserves obtenue avec nos différents modèles

8 La prise en compte de l'inflation

Le but de ce chapitre est d'étudier l'impact de l'inflation sur le calcul des provisions techniques dans le cas de notre modèle individuel et sur une branche longue. Pour pouvoir quantifier l'impact sur notre modèle, nous étudierons aussi l'effet de l'inflation sur le montant des réserves calculées avec les méthodes standard.

Nous allons donc essayer de modéliser l'évolution de l'inflation au cours du temps pour pouvoir corriger les montants réglés. Nous réinjecterons ensuite nos prévisions d'inflation dans les règlements futurs pour observer son influence sur le montant des réserves calculées.

8.1 Les enjeux et conséquences de l'inflation

L'inflation ne concerne que le montant total des sinistres, et non leur fréquence. Dans l'utilisation des méthodes de provisionnement standard, les règlements effectués sont supposés intégrer l'information de l'inflation passée, et la prédiction des flux futurs est supposée conserver ce type d'information et donc intégrer l'information de l'inflation future. Derrière ce raccourci se trouve l'idée que l'inflation future sera similaire à l'inflation passée. Cette hypothèse est remise en cause par le fait que la branche responsabilité civile soit une branche longue, et le règlement des sinistres peut s'étaler sur des périodes relativement différentes d'un sinistre à l'autre.

Lorsqu'un sinistre est ouvert, le gestionnaire sinistre ouvre aussi une provision dossier/dossier qui est une estimation du coût final du sinistre. Ce montant sera réglé sur les années de développement du sinistre. Sa bonne estimation est donc cruciale pour l'assureur. Les règlements pouvant s'étaler sur plusieurs années, le montant de chaque règlement (et donc le montant final) dépendra du niveau d'inflation pendant la période de règlement. En effet, une inflation supposée inférieure à l'inflation réellement observée conduira à un sous-provisionnement et donc un risque de faillite plus important. A contrario, une hypothèse d'inflation plus élevée que l'inflation réelle mènera à un sur-provisionnement et une perte de rentabilité.

8.2 La correction de l'inflation passée et la projection de l'inflation future

En travaillant avec nos règlements initiaux, nous avons fait l'hypothèse que l'inflation future sera la même que l'inflation passée. Afin de corriger ce biais, nous voulons

travailler sur des règlements nets d'inflation. Nous pouvons ensuite estimer le montant des réserves sur ces nouveaux flux, pour enfin réinjecter l'inflation future estimée dans les flux de règlement futurs.

8.2.1 La correction de l'inflation sur nos règlements passés

Nous voulons considérer des règlements en *as if* en base 2011. Nous les corrigeons donc de l'inflation accumulée depuis 2011, obtenus d'après l'indice INSEE correspondant [9]. Notons au passage les trois conditions nécessaires pour qu'un indice soit jugé pertinent :

- sa facilité d'élaboration ;
- sa facile compréhension ;
- sa validation par des représentants professionnels.

Après avoir déflaté les règlements par l'indice cumulé d'inflation selon les années calendaires, nous pouvons appliquer nos modèles (méthodes standard et modèle individuel). L'inflation étant alors nulle, les règlements futurs simulés le sont avec une inflation constante et égale à 0 ; il ne nous reste plus qu'à réinjecter l'inflation future.

8.2.2 L'estimation de l'inflation future

Pour réinjecter l'inflation dans nos règlements futurs, nous devons prédire les taux d'inflation annuels.

Nous avons tout d'abord essayé de la modéliser par un modèle de régression d'ordre 2 en fonction du temps uniquement : en notant X_t l'inflation,

$$X_t = \alpha + \beta_1 t + \beta_2 t^2 + \epsilon_t \tag{8.1}$$

où ϵ_t est le terme d'erreur, gaussien.
L'estimation des paramètres de régression se fait par la méthode des moindres carrés.

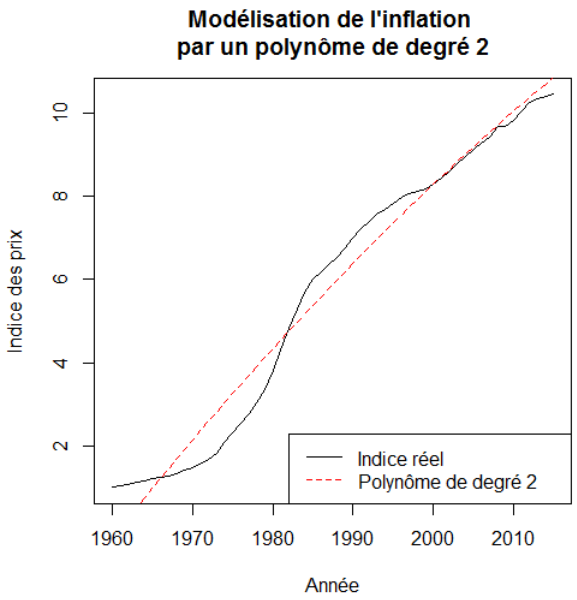


FIGURE 8.1 – Modélisation de l'indice des prix par un polynôme de degré 2

L'analyse des résidus (figure 8.2) montre que ces derniers ne sont pas gaussiens ; plus précisément, il semble exister un terme d'autocorrélation non modélisé dans la régression polynomiale.

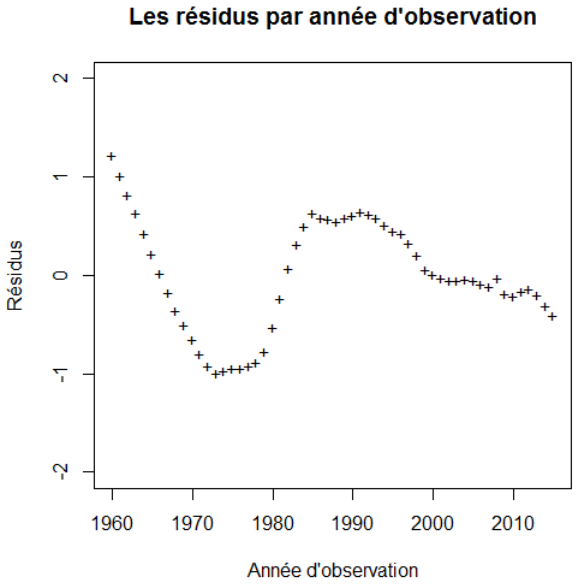


FIGURE 8.2 – Représentation graphique des résidus par année d'observation

Nous avons alors voulu modéliser l'indice des prix avec un processus SARIMA simple. La première étape consiste à centrer le processus.
Comme nous pouvons l'observer sur le graphique 8.1, le processus n'est clairement pas stationnaire. Nous le différencions une première fois.
Nous avons alors pu modéliser ce nouveau processus par un processus AR(1), c'est-à-dire de la forme :

$$X_t = \phi_1 X_{t-1} + \epsilon_t \tag{8.2}$$

où ϵ_t est un bruit blanc faible.

Le modèle décrivant l'évolution de l'indice des prix est alors le suivant :

$$Y_t = \phi Y_{t-1} + \mu\left(1-\phi\right) + \epsilon \tag{8.3}$$

avec $\mu = \mathbb{E}\left[(Y_t)\right]$. Le modèle est alors plus adapté aux données :

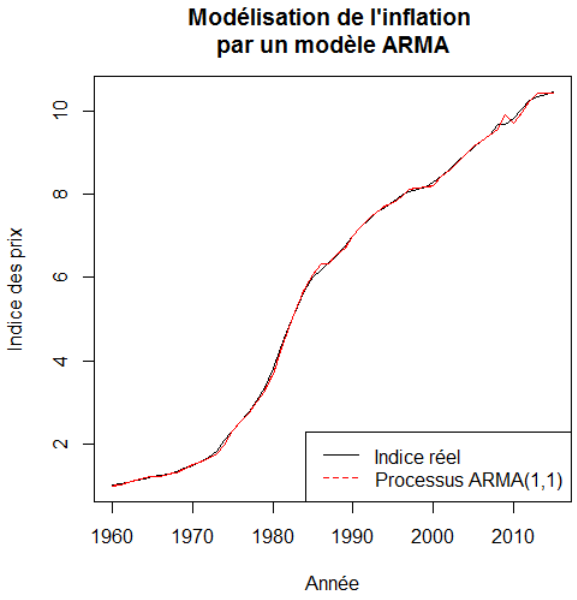


FIGURE 8.3 – Modélisation de l'indice des prix par un modèle AR(1)

Nous conservons alors ce modèle pour projeter l'indice des prix futurs. Nous pouvons maintenant ré-inflater les paiements futurs.

8.3 L'influence de l'inflation sur les montants de réserves obtenus

Nous disposons désormais d'un triangle en *as if* vu en 2011, et des taux d'inflation futurs. Nous pouvons utiliser nos modèles sur les règlements nets d'inflation afin de prédire les règlements futurs nets d'inflation, puis nous réinjectons l'inflation prédite

dans nos nouveaux règlements simulés.

Finalement, en calculant les réserves avec nos différents modèles sur des règlements nets d'inflation et en réinjectant l'inflation dans nos cash flow prévisionnels, l'estimation des réserves est relativement stable. La modélisation de l'inflation montre que cette dernière entraine un sur-provisionnement d'environ 1,4%. La correction de l'inflation n'influence donc pas significativement le montant des réserves. L'impact aurait probablement été plus important sur un historique de sinistres plus long. Un léger changement intervient au niveau des cadences de règlement, avec des facteurs de développement inférieur dans le cas où l'inflation est prise en compte.

La prise en compte de l'inflation permet toutefois de diminuer légèrement la volatilité des réserves, de l'ordre de 7%. Cela nous permet donc de nous rapprocher théoriquement de la valeur de la provision recherchée.

Conclusion générale

Dans ce mémoire, nous nous sommes intéressés au calcul des provisions techniques. Les méthodes classiques utilisent des données agrégées par année de survenance et année de développement dans des triangles de liquidation. Ces méthodes ne sont toutefois pas infaillibles ; comme nous l'avons montré, sur notre jeu de données, les méthodes standards ne sont rigoureusement pas utilisables. Nous avons alors voulu présenter un autre type de modèle de provisionnement, qui utilise des données détaillées et revient à la maille du sinistre individuel. Un tel modèle présente différents avantages : prise en compte de la réassurance proportionnelle et non proportionnelle, mesure de l'incertitude des provisions,. . .

Pour développer un tel modèle, nous avons dû étudier les différentes étapes de la vie d'un sinistre. Nous avons donc modélisé différents phénomènes : la durée de vie des sinistres, les dates des flux de règlement des sinistres et les montants de règlement associés à ces dates.
Pour l'étude de la durée de vie des sinistres, les modèles de survie nous ont semblé répondre le mieux possible au problème. Ils permettent en effet de prendre en compte le phénomène de censure des données, c'est-à-dire les sinistres non clos au moment de l'étude.
Pour la modélisation des flux de règlements, notre première idée fut d'utiliser des variables exogènes pour expliquer les différents phénomènes. Les résultats n'étant pas concluants, nous avons alors ajusté des lois de probabilité paramétriques à nos variables étudiées.

Nous avons ensuite comparé les résultats obtenus avec notre modèle individuel à ceux obtenus par les méthodes standard de provisionnement. De manière générale, les méthodes traditionnelles sont inadaptées pour ce type de données. Il s'agit en effet de sinistres longs, avec des montants réglés volatils. Des résultats peu précis ont ainsi été obtenus avec ces méthodes. Au contraire, le modèle individuel a permis d'obtenir des réserves moins élevées et surtout moins volatiles que celles obtenues avec les méthodes agrégées, bien que le modèle individuel pourrait sous-estimer légèrement ces réserves.

De notre étude, nous avons conclu que l'utilisation d'un modèle individuel peut être une bonne alternative à l'utilisation des modèles standards pour ce type de données. En effet, ce type de modèle permet de modéliser des phénomènes individuels complexes, qui sont mis de côté par l'agrégation des données dans les modèles standards.
En pratique, le développement d'un modèle individuel n'est pas chose aisée, du fait du manque de données disponibles. il n'est en effet pas toujours possible de disposer d'une

base suffisamment détaillée.
Enfin, au-delà de l'aspect technique du modèle, son utilisateur doit disposer d'une connaissance de la branche des sinistres concernée. Il lui faut en effet pouvoir anticiper l'évolution jurisprudentielle et les changements de procédure de gestion des sinistres par la compagnie sur les années à venir.

Dans le modèle proposé, nous avons effectué plusieurs restrictions : nous n'avons pas étudié les sinistres survenus mais non déclarés, et nous n'avons pas considéré le développement des sinistres au-delà de 5 ans. Il pourrait être intéressant donc de considérer des facteurs de queue à défaut d'avoir un historique plus important, de même qu'un utilisateur d'un modèle individuel pourrait vouloir considérer les sinistres tardifs dans un modèle plus global afin de calculer l'ensemble des réserves, IBNR compris.

Bibliographie

Articles

[1] XIE M., TANG Y., GOH T.N., *A modified Weibull expression with bathtub-shaped failure rate function*, Reliability Engineering and System Safety, 76,279-285, 2011

Livres

[2] DENUIT M., CHARPENTIER A., *Mathématiques de l'assurance non-vie, tome II : tarification et provisionnement*, 2005

[3] CHARPENTIER A., *Computational Actuarial Science with R*, 2014

[4] HARDIN J.W., HILBE J.M., *Generalized Linear Models and Extensions, Third Edition*, 2012

Mémoires d'actuariat

[5] BENETEAU G., *Modèle de provisionnement sur données détaillées en assurance non-vie*, ENSAE, 2004

[6] ROSE N., *Provisionnement en assurance non-vie : utilisation de modèles paramétriques censurés*, ISUP, 2009

[7] NGOC A.D., CHAU G., *Mesures de provisionnement cohérentes et méthodes ligne à ligne pour des risques non-vie*, ENSAE, 2013

Sites internet

[8] http ://www.legifrance.gouv.fr

[9] http ://france-inflation.com/inflation-depuis-1901.php

[10] http ://freakonometrics.hypotheses.org/

Liste des tableaux

Table des figures

Annexe

Démonstration des propriétés du modèle de Mack

Propriété . *Sous les hypothèses :*

(H1) les années de survenance sont indépendantes entre elles

(H2) l'espérance conditionnelle de $C_{i,j+1}$ sachant le passé $C_{i,1}$ à $C_{i,j}$ est proportionnel à $C_{i,j}$:

$$\mathbb{E}\left(C_{i,j+1}|C_{i,1}\dots C_{i,j}\right) = \lambda_j * C_{i,j}$$

les estimateurs standards de Chain Ladder $\hat{\lambda}_j = \dfrac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}$ *sont sans biais et non corrélés.*

Démonstration

D'après l'hypothèse (H2), $\mathbb{E}\left(C_{i,j+1}|C_{i,1}\dots C_{i,j}\right) = \lambda_j * C_{i,j}$ d'où

$$\begin{aligned}
\mathbb{E}\left(\hat{\lambda}_j\right) &= \mathbb{E}\left(\mathbb{E}\left(\hat{\lambda}_j|C_{i,1}\dots C_{i,j}\right)\right) \\
&= \frac{\sum_{i=1}^{n-j} \mathbb{E}\left(C_{i,j+1}|C_{i,1}\dots C_{i,j}\right)}{\sum_{i=1}^{n-j} C_{i,j}} \\
&= \lambda_j
\end{aligned}$$

De plus, pour tout j<k, on a :

$$\begin{aligned}
\mathbb{E}\left(\hat{\lambda}_j * \hat{\lambda}_k\right) &= \mathbb{E}\left(\mathbb{E}\left(\hat{\lambda}_j * \hat{\lambda}_k|C_{i,1}\dots C_{i,j}\right)\right) \\
&= \mathbb{E}\left(\hat{\lambda}_j * \mathbb{E}\left(\hat{\lambda}_k|C_{i,1}\dots C_{i,j}\right)\right) \\
&= \mathbb{E}\left(\hat{\lambda}_j\right) * \mathbb{E}\left(\hat{\lambda}_k\right)
\end{aligned}$$