Mémoire d'actuariat

Approche innovante de la tarification de la garantie RCD

Juan Calderon

Tuteur académique :
M. Alexandre You
Tuteur en entreprise :
Mme. Véronique Marpillat

11 septembre 2016

Remerciements

J'adresse tout d'abord mes remerciements à Laurence LEMERLE, responsable de l'actuariat IARD Entreprises, de m'avoir accueilli dans ses équipes.

Je tiens à remercier Véronique MARPILLAT, mon maître de stage et responsable du service Actuariat Construction, RI et RC, pour sa disponibilité, son suivi et pour les connaissances du métier qu'elle m'a apportées.

Un grand merci à Mahmoud LAKKIS, conseiller d'études actuarielles, pour ses conseils et son suivi tout au long de ce stage.

Je remercie également l'ensemble de la direction Actuariat d'AXA Entreprises, pour leur accueil chaleureux et leur soutien.

Enfin, mes remerciements s'adressent à Alexandre YOU, responsable du cours Tarification IARD à l'Université de Strasbourg et mon tuteur académique, pour son encadrement, sa relecture et ses conseils avisés qui ont grandement contribué à l'accomplissement de ce mémoire.

Confidentialité

Le présent rapport est confidentiel.

De plus, nous avons volontairement modifié les chiffres, sans impacter les conclusions qu'on peut en déduire.

Table des matières

Chapitre 1

Introduction

L'assurance construction vise à indemniser les personnes physiques ou morales qui font construire un bien immobilier pour leur compte. Son objectif est de protéger l'acquéreur de ce bien suffisamment longtemps après la fin des rapports contractuels avec le constructeur.

En France, cette assurance est obligatoire et est régie par la loi de 1978 dite loi Spinetta, qui a instauré un mécanisme à double détente : une assurance de choses, au bénéfice du maître d'ouvrage, préfinance les réparations et exerce ensuite un recours sur l'assurance de responsabilité des intervenants. La principale particularité de cette assurance est liée à sa durée de couverture de risque qui est de dix ans à partir de la fin des travaux, contrairement aux assurances dommages classiques qui couvrent des risques sur une période d'un an. Une prime unique est perçue à l'ouverture du chantier et sert à payer les sinistres se produisant dans les dix années suivant la réception du chantier. Pour les assureurs, il s'agit d'un mode de gestion complexe appelé gestion en capitalisation.

Cette complexité de gestion s'accompagne depuis plusieurs années d'un contexte contraignant du marché. En effet, la crise financière qui a débuté en 2008 a sévèrement impacté le marché français de la construction qui aujourd'hui, 8 ans après, ne présente toujours pas de conditions favorables : le nombre de mises en chantier, qui est le moteur de ce marché, devrait poursuivre sa baisse en 2016 et les entreprises du bâtiment restent fragiles, en témoigne le nombre de faillites d'entreprises, toujours très élevé. A cela s'ajoute un contexte financier contraignant avec des taux d'intérêt particulièrement bas, qui ne favorisent pas la rentabilité des branches longues comme la construction.

Dans ce contexte, une refonte complète de la gamme construction d'AXA Entreprises est menée par la Direction d'Actuariat, Coordination et Suivis Techniques (DACST) afin d'améliorer la segmentation des risques et de redresser les résultats de la branche. Une approche nouvelle est alors proposée pour la modélisation de la prime pure de la garantie décennale (RCD), se basant notamment sur des hypothèses concernant le choix de la période d'observation.

En effet, quand il s'agit d'étudier la garantie décennale, le choix d'une période d'observation suffisamment longue est nécessaire. Cependant en choisissant un historique trop long, nous pourrions être confrontés à des effets générationnels, le marché étant instable, les différentes générations ne reflètent pas forcément le comportement du marché d'aujourd'hui. Or ce sont ces générations qui ont la sinistralité la plus manifestée et qui vont servir de point de repère. Il faut donc trouver un compromis entre ces deux contraintes. Notre mémoire qui s'attache à la garantie décennale a pour objectif de juger de la pertinence la méthode utilisée par la DACST et de répondre aux questions suivantes :

- Quel impact a le choix du périmètre sur la modélisation de la prime pure en RCD ?
- Le modèle proposé est-il robuste et fiable, peut-on l'améliorer ? Existent-ils des modèles plus adéquats ?

Pour répondre à ces questions, nous aborderons la démarche de la modélisation de la prime pure sur des années différentes. Dans un premier temps nous déterminons les variables les plus explicatives en considérant deux visions de sinistralité différentes : une vision complète et une autre limitée à 5 années de survenance. Nous comparons ensuite les résultats.

Nous entamons ensuite la recherche du meilleur modèle, en considérant d'une part l'approche directe de la modélisation des charges totales, et d'autre part, l'approche fréquence-coût moyen. Pour chacune des approches, diverses lois sont testées : les lois dites Tweedie et les modèles Zero-inflated par exemple.

Enfin, nous retenons un modèle, qui sera appliqué sur deux périmètres puis sur deux visions de sinistralité différentes, puis nous comparons les résultats afin de juger de la sensibilité du modèle au périmètre retenu.

Première partie

Contexte de l'étude

L'étude faisant l'objet de ce mémoire a été réalisée au sein de la Direction de l'Actuariat, Coordination et Suivi Technique d'AXA Entreprises. Elle porte sur le produit d'assurance construction et plus précisément, sur la garantie décennale de ce produit.

L'objectif de cette partie du mémoire est tout d'abord, de décrire les assurances liées à l'acte de construire, de présenter les différents intervenants, de détailler les différentes garanties et de décrire le cadre réglementaire dans lequel évoluent les assurances construction.

Nous décrirons ensuite le marché français de l'assurance construction.

Chapitre 2

L'assurance construction

2.1 Principe générale

L'assurance construction a pour vocation d'indemniser les différents intervenants d'une opération de construction.

En France, cette assurance est réglementée selon la loi Spinetta de 1978, qui définit les risques couverts ainsi que les responsabilités de chaque partie.

Il est donc indispensable de distinguer la nature et les rôles de chaque intervenant.

L'opération de construction s'accompagne nécessairement de nombreux contrats qui lient entre eux les différents intervenants de l'acte de construire. Décrivons-les.

Les intervenants

Il existe trois principaux intervenants dans une opération de construction :

1. **Le maître d'ouvrage** :

   C'est la personne physique ou morale pour le compte de laquelle sont effectués les travaux. Il définit le programme de construction : besoins fonctionnels que l'ouvrage devra satisfaire, exigences en matière de qualité, de délai et de prix. Le maître d'ouvrage peut agir pour son compte (particulier qui fait construire), ou pour le compte d'autrui (industriel, promoteur vendeur) ; il peut être privé (particulier, sociétés, commerçants, industriels, promoteurs) ou public (État, collectivités, établissements publics).

2. **Le maître d'œuvre** :

   C'est la personne physique ou morale qui reçoit la mission du maître d'ouvrage de concevoir l'ouvrage, de diriger et contrôler l'exécution des travaux et de l'assister et

le conseiller lors de leur réception et leur règlement. Le maître d'œuvre, qui peut être un architecte, un technicien, un ingénieur-conseil ou un bureau d'études techniques, est tenu d'une obligation de résultat.

3. **Les entrepreneurs de la construction ou réalisateurs** :

   L'entrepreneur est la personne physique ou morale qui exécute les travaux de construction. Il est responsable envers le maître d'ouvrage, et lié à celui-ci par un contrat de louage.

   En plus de ces trois intervenants, il peut y avoir aussi : des sous-traitants, qui sont tenus d'obligation de résultat envers les réalisateurs, et des fournisseurs, qui sous certaines conditions engagent leur responsabilité avec les réalisateurs.

4. Les autres intervenants sont :

   - le négociant, qui fournit les matériaux et peut apposer sa marque ;
   - le fabricant ou le détenteur du procédé de construction ;
   - le contrôleur technique.

Chacun des intervenants est tenu par la loi de souscrire une assurance construction avec des garanties différentes, que nous décrivons.

2.2 Les garanties de l'assurance construction

En France, l'assurance construction est régie par la loi Spinetta de 1978, qui instaure un mécanisme à double détente. Ce système est composé de des assurances complémentaires et obligatoires. Nous les classons selon la durée de couverture.

- **Sinistres survenus après réception des travaux**
  - **L'assurance Dommages Ouvrages (DO)** : elle est destinée aux maîtres d'œuvre. En cas de sinistre, l'assureur DO couvre immédiatement la totalité du coût de réparation des faits. Il se retournera par la suite vers le responsable du sinistre pour le remboursement des frais. Cette assurance est appelée assurance de préfinancement.
  - **L'assurance RCD** : le réalisateur des travaux doit souscrire obligatoirement une assurance décennale. Cette responsabilité s'étend sur 10 ans après la réception des travaux, d'où le nom de cette garantie. Dès lors que l'assureur DO prouve que le réalisateur est responsable des faits, l'assureur RCD remboursera les charges des sinistres à l'assureur DO.
- **Sinistres survenus avant réception des travaux** :
  - **L'assurance Responsabilité Civile (RC)** :

Elle couvre les sinistres pendant la réalisation des travaux.

– **Autres assurances facultatives** :

  Le maître d'œuvre et le constructeur peuvent souscrire des assurances facultatives pour des sinistres survenus avant la réception des travaux.

2.3 L'offre assurance construction d'AXA

L'offre d'AXA regroupe les différentes garanties construction en deux familles de produits : contrats abonnement et contrats chantier.

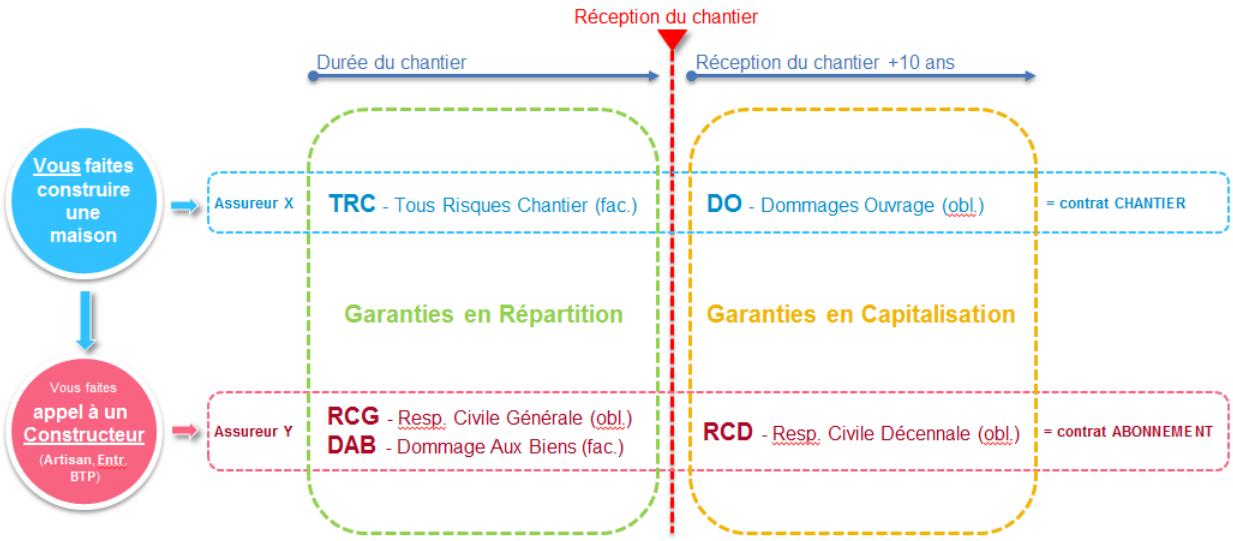


FIGURE 2.1 – L'offre d'AXA en assurance construction

L'offre abonnement est destinée au maître d'œuvre, qui est un professionnel du bâtiment. Ces contrats sont reconduits par tacite reconduction. L'offre chantier, quant à elle, est destinée au maître d'ouvrage et propose des contrats ponctuels limités à un chantier donné.

Ces deux offres proposent des garanties qui couvrent les sinistres survenus pendant les travaux (garanties en répartition), et ceux survenus après la réception du chantier (garanties en capitalisation).

Le sujet de ce mémoire concerne l'offre abonnement, et plus particulièrement la garantie RCD, que nous traitons en détail.

Chapitre 3

La garantie RCD

3.1 Cadre Réglementaire

Instauration de la RC Décennale

L'origine de l'assurance construction en France date du code civil de 1804 ou les grands principes de la Responsabilité Civile Décennale des architectes et des entrepreneurs sont établis.

Cette loi a pour objet de protéger l'acquéreur d'un bien immobilier contre les risques consécutifs d'un défaut de fabrication pendant une période de 10 ans suivant la date de livraison de l'ouvrage.

Bien que cette législation donne à l'acquéreur une protection très importante, ce système présente de nombreuses insuffisances :

- des délais de règlements très longs liés aux multiples expertises nécessaires pour déterminer le partage des responsabilités,
- un faible taux d'assurance des professionnels concernés par cette loi,
- un champ d'application trop restreint : seuls les vices de construction sont couverts,
- une sinistralité sans cesse croissante.

Pour pallier ces lacunes, le législateur a édité des nouvelles règles visant à apporter d'avantage de protection aux maîtres d'ouvrage vis-à-vis de leur bien.

La Loi Spinetta du 4 janvier 1978

Cette loi apporte les modifications suivantes au régime existant :

- Une extension du champ de responsabilité à tous les intervenants ayant contribué à la livraison du bien (par exemple, les promoteurs ou les fabricants de composants).
- une définition très précise des garanties couvertes. Celles-ci sont au nombre de trois :
  - la garantie de parfait achèvement qui pèse pendant un an sur les entrepreneurs qui doivent réparer tous les désordres signalés par le maître d'ouvrage, qu'ils soient ou non responsables.
  - la garantie décennale, qui voit son champ d'application étendue à tous les dommages :
    * compromettant la solidité de l'ouvrage,
    * le rendant impropre à sa destination,
    * affectant la solidité des équipements faisant corps avec le bâtiment.
  - La garantie de bon fonctionnement des équipements dissociables du bâtiment. Elle est donnée pour une durée de deux ans.
  - La souscription obligatoire d'une police d'assurance (dite de Dommage-Ouvrage) par le maître d'ouvrage lui permettant d'obtenir rapidement réparation auprès de son assureur en cas de sinistre. Cette garantie évite ainsi à l'assuré l'avance des fonds nécessaires à la réhabilitation de son bien. Charge ensuite à l'assureur de rechercher les responsabilités et d'effectuer les recours. Il n'existe pas d'autres cas dans le droit français où il est obligatoire de souscrire une assurance pour se couvrir contre les dommages que l'on est susceptible de supporter.
  - La souscription obligatoire d'une police d'assurance par tous les intervenants dont les travaux entrent dans le champ de la responsabilité décennale. Lorsque cette responsabilité est établie, son assureur rembourse le sinistre à l'assureur dommage, qui a lui-même déjà indemnisé son client.

Dans la suite de ce mémoire, le terme DROC sera utilisé pour désigner un exercice de souscription. Il signifie "Date Réglementaire d'Ouverture de Chantier", et correspond à la date à laquelle a été souscrite la police d'assurance.

A noter qu'il ne s'agit pas de la date de début de garantie, celle-ci ne prenant effet qu'à la fin du chantier, pour une durée de 10 ans.

3.2 la DROC et la période garantie

Bien que la garantie RCD ne commence à courir qu'après réception des travaux, l'assuré souscrit et paie sa prime au moment de la Déclaration Réglementaire d ?Overture de Chantier (DROC, Appelé aussi Date d'Overture de Chantier (DOC)).

Par ailleurs l'assureur, ne connaissant pas la date de fin des travaux, assure le client à partir de la Déclaration Réglementaire d'Overture de Chantier (DROC), et jusqu'à 10 ans après la réception. C'est pourquoi, l'on peut constater des sinistres RCD jusqu'à 13 ans après les débuts des travaux.

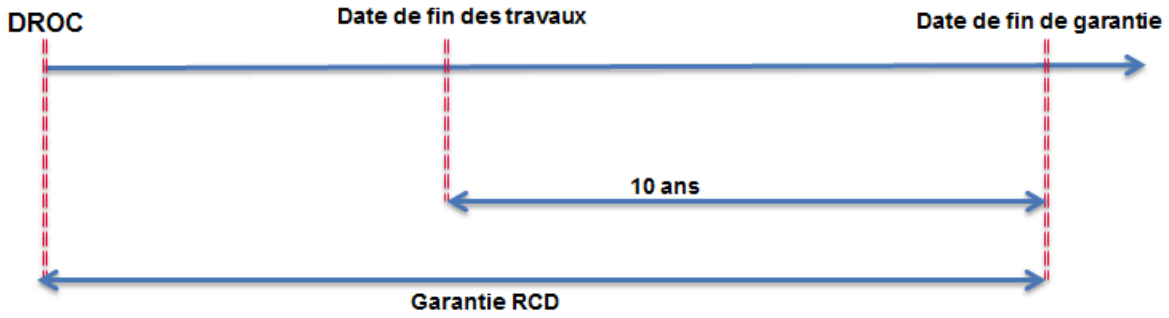


FIGURE 3.1 – DROC et période de garantie

Si un sinistre survient, le mécanisme de règlement est déterminé par le principe de double détente.

Le principe de double détente

Le paiement d'un sinistre se fait en trois étapes :

1. Lorsqu'un sinistre survient, le maître d'ouvrage notifie son assureur DO de son sinistre, et celui-ci doit rembourser les coûts de réparation ou de remplacement de sa propriété sans se soucier de la responsabilité du maître d'ouvrage.
2. Suite au remboursement des frais, l'assureur DO se retourne contre l'assureur RCD et deux cas sont possibles :
   (a) La responsabilité repose sur le constructeur : l'assureur RCD s'acquitte des frais de réparation ou de remplacement.
   (b) L'assureur RCD parvient à prouver que la faute est liée à une cause extérieure, et se libère de du remboursement des frais. Ainsi, et dans ce cas ce sera l'assureur DO qui assumera la totalité de la charge.

Remarque *Dans aucun cas le maître d'ouvrages devra s'acquitter des frais de réparation ou de remplacement.*

Sont obligatoires en France les assurances :

- RCD

- DO
- RCG

Et facultatives :

- Tous Risques Chantier
- Dommages Aux Biens

L'offre d'AXA en RCD Revenons vers l'offre de contrats **abonnements** d'AXA.

Supposons que le constructeur A débute les travaux d'un chantier en année t (DO). Il est donc tenu de souscrire une assurance RCD pour son chantier. Néanmoins il souscrira, non pas un contrat pour chaque chantier, mais un seul, qui couvrira **tous les chantiers dont la DROC est au cours de l'année** t.

Cela suppose deux difficultés pour l'assureur pour mesurer son risque. D'une part, il ne connaît pas exactement la fin de la période garantie, et d'autre part, il ne connaît pas le nombre de chantiers qu'il assure. C'est pourquoi, il se servira du chiffre d'affaires déclaré pour estimer la matière assuré.

Ensuite, l'assureur va estimer les charges futures liées à la DROC t, et fera payer une prime au constructeur en année t. Ceci donne lieu à un nouveau mode de gestion de primes et à un autre type de provisions à calculer.

3.3 Les modes de gestion des primes, et la PSNEM

Les modes de gestion des primes

Deux modes de gestion des primes existent en assurance Incendie, Accidents et Risques Divers (IARD) :

1. la gestion en répartition,
2. la gestion capitalisation.

L'assurance IARD classique correspond au mode de gestion en répartition : les primes encaissées en année n serviront à payer les sinistres survenus au cours de l'année n.

Par opposition, en capitalisation, les primes encaissées pendant l'année n serviront à payer tous les sinistres qui surviendront pendant les années futures. **Ce système est propre à l'assurance construction (DO et RCD).** [1]

1. Le principe de ces modes de gestion est équivalent à celui appliqué en assurance retraite

Chaque mode de gestion généré un décalage différent entre le paiement des primes et le règlement des sinistres. L'assureur sera donc tenu de constituer des provisions de différente nature.

Les provisions pour sinistres en RCD

Deux types de provisions existent en assurance IARD :

1. Les Provisions pour sinistres à payer (PSAP)
2. la Provision pour sinistres non-encore manifestés (PSNEM)

PSAP

Nous introduisons la PSAP, pour ensuite établir un parallèle avec la PSNEM.

En assurance IARD, les sinistres survenus au cours d'une période peuvent être déclarés plusieurs années après la survenance, et vont voir leur coût augmenter au fil des années. Les assureurs sont donc tenus de constituer des provisions pour remédier au décalage entre le paiement des primes et le règlement des sinistres.

Deux types de sinistres sont à l'origine de ce décalage :

- "Incurred But Not Yet Reported" (IBNYR) des sinistres survenus pendant la période n mais déclarés dans les années suivantes.
- "Incurred But Not Enough Reported" (IBNER) des sinistres survenus et déclarés pendant l'année n mais dont le coût évolue dans les années qui suivent.

Les PSAP se calculent donc comme :

$$PSAP = IBNR + Provision\ Dossier/Dossier \tag{3.1}$$

avec :

$$IBNR = IBNYR + INBER \tag{3.2}$$

Dans le cadre de l'assurance construction RCD, l'impact des IBNR est très restreint puisque la déclaration du sinistre se fait presque immédiatement après la survenance. Nous constatons néanmoins un impact très considérable des IBNER, puisque la charge d'un sinistre peut évoluer fortement après la survenance [2].

2. Sur notre périmètre d'étude, qu'on décrira plus tard, 93% des sinistres sont déclarés pendant l'année de survenance

Il reste maintenant à traiter les provisions pour les sinistres qui surviennent des années après la souscription.

La PSNEM

Cette provision s'applique uniquement aux contrats gérés en capitalisation, et découle du principe de double détente.

Puisque l'assureur encaisse en année n la prime RCD, et s'engage à couvrir des 10 années futures, il fait face à nouveau à un décalage de paiements. Il est tenu donc de constituer des primes pour lisser ses résultats, et pouvoir faire face à ses engagements. Cette provision est appelée la PSNEM.

Pour résumer, l'assureur RCD fait face à 3 types de provisions (IBNYR, IBNYR, PSEM), et l'on peut représenter ses engagements sus un espace à trois dimensions.

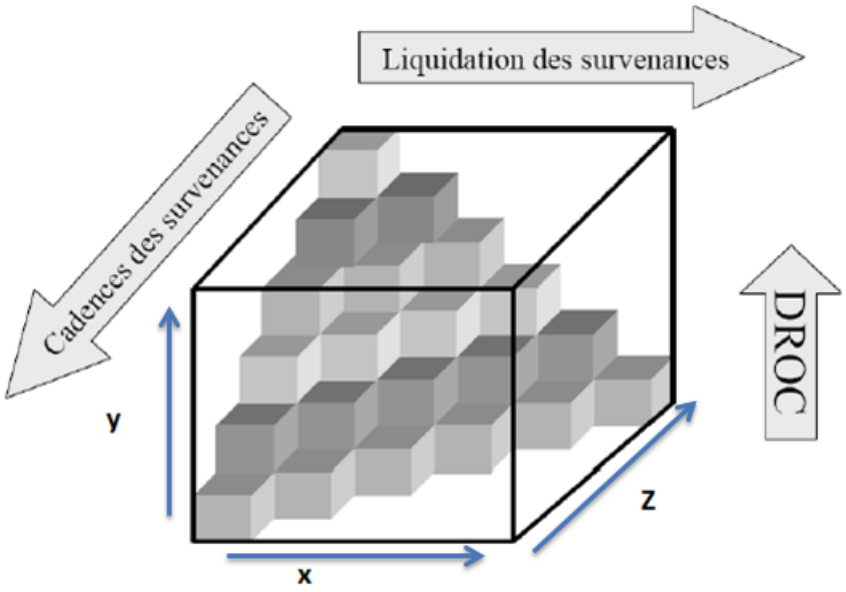


FIGURE 3.2 – schéma DROC, survenance, année de compte, source : (Luzi, 2006)

1. un premier axe x qui représente la liquidation des survenances, c'est-à-dire l'année de déclaration-paiement des sinistres,
2. un deuxième axe y, qui représente les années de DROC,
3. un troisième axe z, qui représente la cadence des survenances, c'est à dire l'année de survenance du sinistre.

En connaissance des trois axes, nous pouvons calculer pour une DROC donnée, le montant total de charges futures. Nous appellerons la cadence de manifestation (axe z) requiert une analyse plus approfondie, nous traiterons cet aspect dans la partie suivante.

Passons finalement au contexte économique sur lequel s'inscrit ce mémoire.

Chapitre 4

Contexte économique du secteur BTP

L'activité de la construction en France représente plus de 450 000 entreprises, composées majoritairement des Très Petites Entreprises (TPE). Elle emploie à elle seule, 1.15 millions de salariés, et représente près de 5% de l'économie française [1].

	2015		2014	
Nombre de salariés	Nombre entreprises	Travaux réalisés en 2015(HT)	*Nombre entreprises*	*Travaux réalisés en 2014(HT)*
0 à 10 salariés	380 300*	46 Mds €	*361 000***	*47 Mds €*
11 à 50 salariés	19 300	41 Mds €	*19 400*	*40 Mds €*
51 à 200 salariés	1 300	18 Mds €	*1 400*	*18 Mds €*
Sup 200 salariés	200	19 Mds €	*200*	*19 Mds €*
Ensemble	**401 100**	**124 Mds €**	***382 000***	***124 Mds €***

FIGURE 4.1 – Nombre d'entreprises par tranche des salariés source : FFB

Les chiffres du premier trimestre l'année 2016 montrent une légère amélioration du secteur (+0.5%), suite à un fort recul en 2015 (-2.2%). [2]. Mais le secteur est loin de retrouver son niveau d'avant la crise de 2008.

1. Chiffres de l'année 2014, Le Figaro
2. Note de conjoncture, juin 2016, INSEE

4.1 La crise financière de 2008 et ses impacts

La crise financière, qui a débuté en 2008, a eu des conséquences sur le marché de la construction. La baisse des mises en chantier a fragilisé davantage ce secteur, dans lequel la disparition des petits acteurs par dépôt de bilan s'est accélérée. Ceci a engendré un alourdissement des sinistres à charge des assureurs qui ont perdu, de facto, l'avantage du Service Après-Vente assuré par les promoteurs les deux premières années.

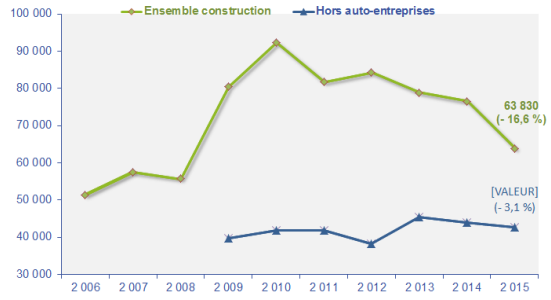


FIGURE 4.2 – Création d'entreprises, secteur de la construction et de l'immobilier source : INSEE

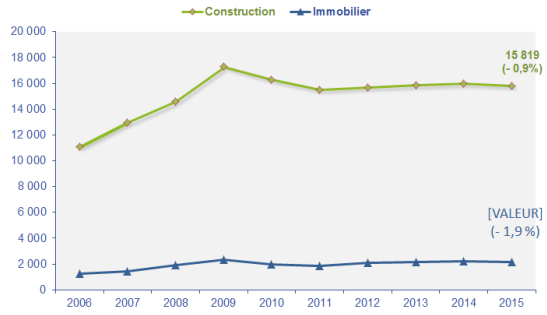


FIGURE 4.3 – Défaillances des entreprises, secteur de la construction source : Banque de France

La figure 3.2 montre que le nombre de défaillances, qui a fortement augmenté en 2008, reste élevé malgré une légère baisse en 2016.

L'impact de la crise est multiplié dans les garanties décennales, puisque l'assureur est engagé dès la souscription du contrat et ce jusqu'à la fin de la période décennale, sans possibilité de renégocier le contrat.

	Montant 2015 (M€) prov	Variations				
		2015/2014 (p)	2014/2013	2013/2012	2012/2011	2011/2010
DO	***560***	**+ 4,7 %**	- 12,3 %	- 3,2 %	-13,3 %	+ 1,7 %
RCD	***1 570***	**- 3,5 %**	- 2,8 %	- 2,8 %	+ 2,4 %	+ 5,7 %
Ensemble	**2 130**	**- 1,5 %**	- 2,9 %	- 2,9 %	- 2,4 %	+ 4,4 %
Ass. Biens et Resp.	***52 300***	***+ 2,2 %***	+ 2,0 %	+ 1,5 %	+ 3,1 %	+ 4,2 %

FIGURE 4.4 – Evolution de l'encaissement des primes en assurance construction

Cette situation de marché se traduit logiquement par une baisse des primes encaissées par

les assureurs et notamment en RCD, et par un fort recul des cotisations sur la période 2014/2015 (-3.5%).

Ainsi, les constructeurs fragilisés, et plus particulièrement les petites structures, n'assurent plus le même niveau de qualité (main-d'œuvre moins qualifiée, baisse de la qualité des matériaux de construction, etc.), entraînant une augmentation du nombre de sinistres.

Tout cela nous amène à un contexte particulièrement difficile pour les assureurs, dont AXA.

4.2 La part AXA du marché

Le marché de l'assurance construction reste relativement concentré avec la présence de sociétés spécialisées dans les risques liés à la construction. Ainsi, les 7 premières sociétés représentent 85% du marché en Assurance Construction.

Malgré un fort recul ces deux dernières années, AXA reste l'un des principaux assureurs de la branche construction et se positionne en troisième place, derrière SMABTP et COVEA.

Rang	Groupes	Cotisations 2015 (M€)	Evolution (%)	Part de marché (%)
1	SMABTP	394	- 6	25
2	COVEA	288	- 1	18
3	AXA	245	- 7	16
4	ALLIANZ	131	- 10	8
5	MUT. des ARCHITECTES	121	- 6	8
6	GROUPAMA GAN	110	- 5	7
7	AUXILIAIRE	48	+ 3	3
8	AVIVA	47	+ 7	3
9	GRP. ASS. MUT. BTP	43	- 4	3
10	GENERALI	34	- 13	2
	ENSEMBLE 10 groupes	1 461	- 5	93 %

FIGURE 4.5 – Principaux acteurs de l'assurance construction

D'ailleurs, la construction est l'une de principales activités d'AXA Entreprises IARD. Elle représente 22% de son CA soit 550 M (en 2015).

La branche construction, et plus précisément les contrats abonnement(RCD), sont aujourd'hui un enjeu d'envergure pour AXA Entreprises IARD. D'où la refonte tarifaire menée par la direction, qui vise à bien identifier les risques dans un contexte économique difficile.

Les spécificités techniques de la garantie RCD rendent ce travail plus complexe, car nous faisons face à des engagements sur 10 ans, et à un bouleversement du portefeuille, suite à la crise économique.

Ce mémoire s'inscrit dans cette logique. Il cherche les méthodologies les plus pertinentes qui permettent une meilleure modélisation de la prime pure de cette garantie, afin de corroborer la validité des hypothèses prises en compte dans le cadre de la refonte tarifaire, et de proposer des améliorations à mener dans le futur.

Deuxième partie

Présentation de l'étude

Dans cette partie nous décrivons la problématique de notre étude, ainsi que les analyses des données préliminaires qui nous ont servi à définir les périmètres d'étude.

Nous étudierons en détail :

- les bases de données disponibles,
- la cadence de manifestation en RCD,
- les variables de tarification disponibles.

Chapitre 5

Problématique de l'étude

5.1 Le choix de la durée d'observation

Rappelons qu'un contrat d'abonnement en assurance construction couvre l'assuré pour tous ces chantiers commencés l'année n (qui correspond donc à la DROC dans ce cas).

La garantie RCD intervient, à partir de la date de réception de l'ouvrage, pendant une durée de 10 ans pour tout dommage rendant l'ouvrage impropre à sa destination.

Par conséquent, une période d'observation dépassant les 10 ans, et pouvant même atteindre 15 ans, est nécessaire pour connaître la sinistralité totale liée à cette DROC.

Rappelons qu'en assurance construction RCD, l'assureur souscrit un contrat pour assurer tous les chantiers commencés en année n (qui correspond à la DROC d'un contrat). L'assuré sera par la suite couvert pour tous les sinistres survenus jusqu'à 10 ans après la fin des travaux, qui peut, dans certains cas, atteindre 15 ans après la DROC.

Cependant, en choisissant un historique trop long, nous pourrions être confrontés à des effets générationnels dus aux bouleversements du marché. En effet, le marché de construction ayant subi plusieurs crises, reste instable. Les différentes générations de DROC, et notamment les plus anciennes, ne reflètent pas forcément le comportement du marché d'aujourd'hui. Or, ce sont ces générations qui ont la sinistralité la plus manifestée. Il a donc fallu trouver un compromis entre ces deux contraintes.

La solution proposée par la direction est d'estimer les charges futures d'un contrat à partir de x premières années d'observation. Mais là encore, le choix de la méthode et de la maille de projection peut affecter les résultats.

Exemple : Nous pouvons considérer deux entreprises de type A et B observées pendant x ans toutes les deux. Sachant que les sinistres des grandes entreprises se manifestent

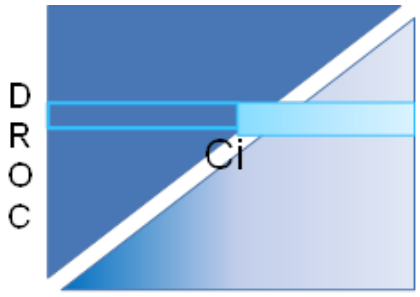


FIGURE 5.1 – Estimation de la CFP par DROC

plus tardivement, la modélisation de la prime pure diffère selon le nombre d'années observées.

TABLE 5.1 – Exemple, typologie des contrats

	CA	Charges moyennes à x ans	Charges moyennes à 10 ans
A	1M	0	50k
B	100k	10k	10k

Dans ce cas, la prime pure du contrat de type A est plus élevée que celle du contrat de type B, qui est en réalité le contrat le plus risqué.

Pour pallier cela, il faudrait provisionner séparément chaque typologie de contrat, en appliquant un coefficient de passage plus élevé aux grandes entreprises par exemple.

Ainsi nous faisons face à un arbitrage sur la définition de l'âge pivot x :

- en prenant x "trop petit", nous pénalisons les typologies de contrats [1] dont les sinistres se manifestent très tôt (les petites entreprises par exemple),
- en choisissant un périmètre très ancien, nous négligions le changement de la structure de notre portefeuille, et sous-estimons le montant moyen des charges par contrat.

La méthodologie testée pour la modélisation de la prime pure a été la suivante :

1. à partir des contrats des années 2000-2005, observer les charges totales à **5 ans**,
2. estimer à partir de ces 5 années, les charges totales pour chaque contrat, en tenant compte de la typologie de contrat,
3. modéliser la prime pure sur les charges finales projetées.

1. les entreprises avec des caractéristiques similaires

5.2 Validation de la méthode à 5 ans

L'hypothèse du point pivot pour la modélisation étant choisie, l'on se soucie de l'impact de ce choix sur la modélisation.

Notre intérêt est particulièrement de savoir :

1. si les variables significatives sont les mêmes, indépendamment du choix du périmètre ;
2. quels sont les modèles le mieux adaptés, du point de vue de l'adéquation des données vis à vis du modèle ;
3. quelles différences aurait-on pu obtenir, en modélisant cette charge sur un autre périmètre ?

Chapitre 6

Données disponibles et périmètre d'étude

Voici le périmètre d'étude établi :

6.1 Périmètre d'étude

- Contrats souscrits entre 1999 et 2015 entre 1999 et 2015.
- Seules les entreprises avec un chiffre d'affaires inférieur à 10M sont conservées.
- Un écrêtement des sinistres à 50k est effectué. Les sinistres graves feront l'objet d'une autre étude.

6.2 Données disponibles

Dans le cadre de cette étude, nous possédons deux types de bases de données à exploiter ;

- base sinistres
- base contrats

Nous avons également à disposition des bases externes permettant de compléter les informations sur les entreprises assurées.

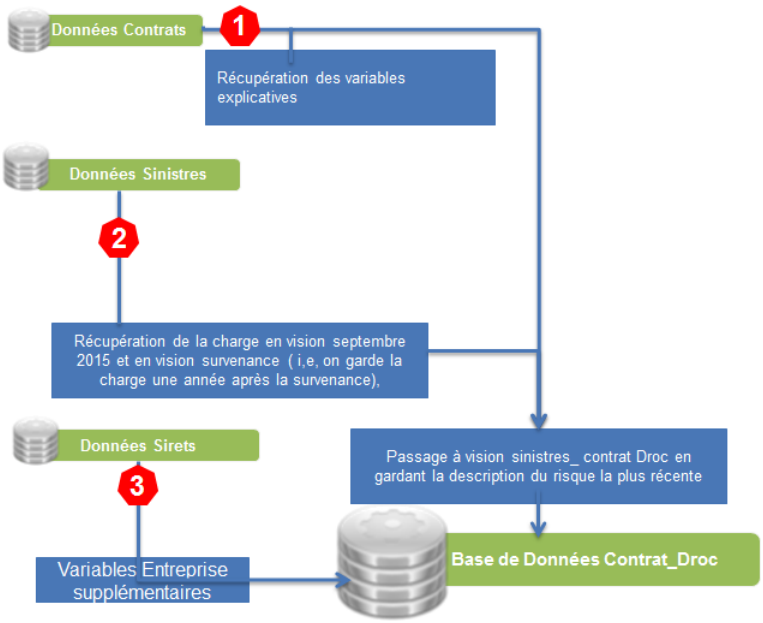


FIGURE 6.1 – Schéma création base de données

Nous expliquerons plus tard le concept de contrat-DROC.

Données Sinistres

La base sinistres d'assurance construction contient les variables suivantes :

- état du sinistre
  - en cours : la charge constatée dans la base n'est qu'une provision.
  - clos sans suite : ce sont des sinistres qui ne seront pas pris en charge par l'assureur, par contre ils peuvent représenter un coût lorsque l'assureur s'acquitte des frais d'expertise au moment de l'évaluation.
  - clos : le sinistre est entièrement réglé.
  - annulé : ne représente aucun coût pour l'assureur.
- date de survenance
- date de déclaration
- charge (différentes visions)
- garantie mise en jeu
- contrat

- DROC, année à l'origine du sinistre.

Les traitements suivants sont effectués :

Traitements

1. repérage des sinistres relatifs à la garantie RCD,
2. exclusion des sinistres annulés,
3. deux visions de la charge sont conservées :
    - vision survenance : nous regardons les charges un an après la déclaration du sinistre,
    - vision 12/2015 : vision vieillie des charges.

Afin de construire notre base finale, nous fusionnons cette base et la base contrats.

Données contrats

La base contrats comporte pour chaque client des variables descriptives du risque telles que :

- le chiffre d'affaires,
- la classe de risque,
- le nombre d'employés,
- la note financière.

Puisque ces variables peuvent évoluer au fil du temps, nous avons pour chaque contrat des visions différentes de son profil de risque.

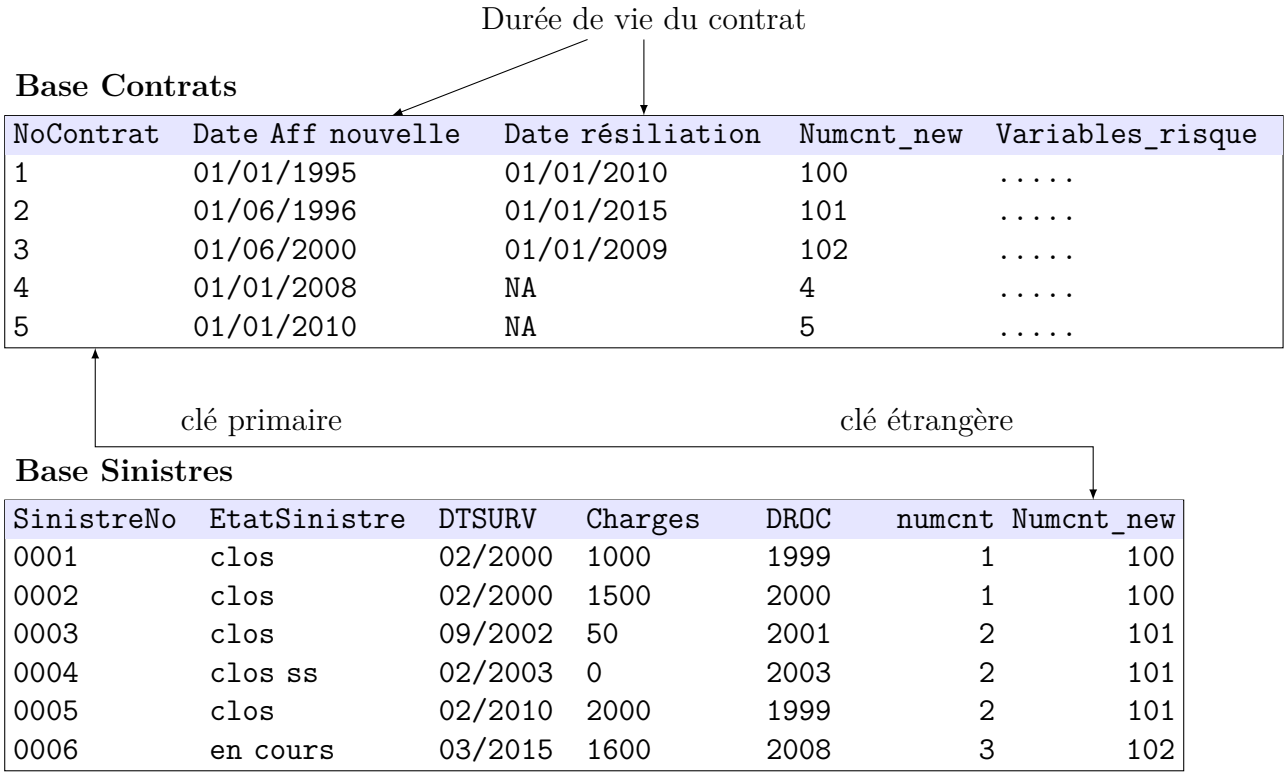
Nous possédons par ailleurs pour chaque contrat :

- la date d'affaire nouvelle,
- la date de résiliation.

Finalement, cette base comporte deux gammes de produit distinctes :

- nouvelle gamme (à partir de 2007),
- ancienne gamme (entre 1999 et 2007).

La structure tarifaire de l'ancienne gamme repose sur peu de variables tarifaires. Le passage à la nouvelle gamme nous apporte des informations supplémentaires sur le client. Nous évaluerons dans la section suivante la pertinence de ces nouvelles variables.



NoContrat	Date Aff nouvelle	Date résiliation	Numcnt_new	Variables_risque
1	01/01/1995	01/01/2010	100	.....
2	01/06/1996	01/01/2015	101	.....
3	01/06/2000	01/01/2009	102	.....
4	01/01/2008	NA	4	.....
5	01/01/2010	NA	5	.....

SinistreNo	EtatSinistre	DTSURV	Charges	DROC	numcnt	Numcnt_new
0001	clos	02/2000	1000	1999	1	100
0002	clos	02/2000	1500	2000	1	100
0003	clos	09/2002	50	2001	2	101
0004	clos ss	02/2003	0	2003	2	101
0005	clos	02/2010	2000	1999	2	101
0006	en cours	03/2015	1600	2008	3	102

FIGURE 6.2 – Schéma création base sinistres

Étant donné que les variables de l'ancienne gamme sont de faible qualité (peu de modalités, description introuvable, forte présence de valeurs manquantes), nous décidons de ne pas conserver ces variables. Nous effectuons donc un chainage des contrats i.e. : on conserve uniquement les contrats souscrits sous l'ancienne gamme et reconduits sous la nouvelle offre.

Ce traitement vient au détriment de la volumétrie des données. Nous perdons ce faisant, 46% des Contrats-DROCs.

Nous passons maintenant à la construction de notre base **Contrat-DROC**, cette notion sera évoquée tout au long de ce mémoire.

Passage à la vision contrat-DROC

Tout d'abord nous construisons une nouvelle base sinistres en associant à chaque sinistre les variables tarifaires du contrat renouvelé.

Nous avons ainsi, pour chaque sinistre, non pas la situation de l'entreprise au moment du sinistre, mais la dernière vision existante dans la base. La démarche contraire, qui consisterai à retrouver la vision du moment de la DROC, est très laborieuse et peut être inefficace, car la qualité des données est nettement meilleure pour les années récentes.

La base sinistres résultante comporte, pour chaque sinistre, la DROC de rattachement du sinistre, le montant des charges, et les variables contrat.

Construction de la base Contrat-DROC

Un sinistre possède une unique DROC. En revanche, un contrat possède autant de DROCs que d'années de souscription.

Nous construisons une base contrats dans laquelle, pour chaque contrat et année de DROC i, nous aurons une ligne avec le montant de charges associés. De plus le nombre de sinistres et leurs charges pour un contrat DROC seront séparés selon différentes visions.

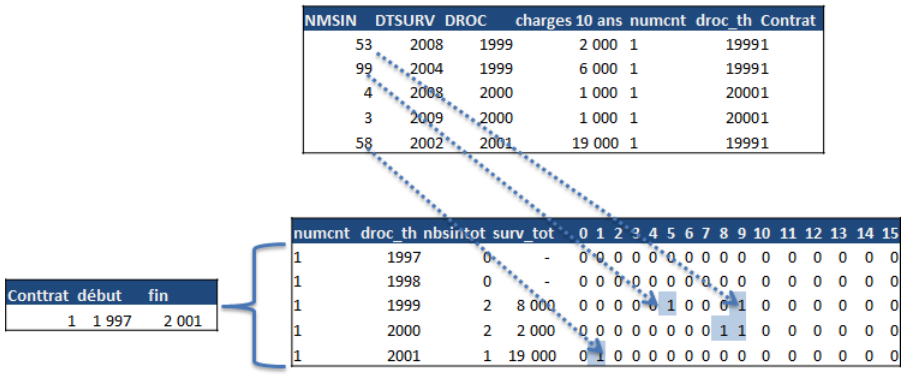


FIGURE 6.3 – Schéma création base de données

Ci-dessus le traitement effectué pour le contrat 1. Ce contrat est représenté par 5 lignes, qui correspondent aux années d'observations du contrat [1]. Pour chaque Contrat-DROC nous avons les nombre de sinistres et la charge totale associée, à chacune des années de développement. [2].

Cette représentation nous permet de calculer les charges et le nombre de sinistres d'un Contrat-DROC, vus à un âge de 5 ans.

Après avoir construit notre base finale de travail, nous passons à l'analyse de la qualité de nos données.

Qualité des données

Nous avons trouvé quelques incohérences dans notre base qui peuvent provoquer des erreurs dans la suite de notre modélisation.

1. Cette représentation de la base correspond à une analyse longitudinale des individus
2. $Annéedéveloppement = Annéesurvenance - AnnéeDROC$

DROC théorique Au cours du traitement de la base de données nous avons constaté des anomalies sur la variable DROC :

$$DROC\_TH = \begin{cases} DTDROC & \text{si DTDROC renseignée,} \\ Estimation & \text{si DTDROC = NA.} \end{cases}$$

Pour certains sinistres, la date de DROC était associée à une année en dehors de la durée de souscription du contrat [Date d'affaire nouvelle (DTFAN),Date de résiliation (DTFRS)], et ceci dû à deux causes :

- date de DROC mal renseignée ;
- Le régleur n'ayant renseigné aucune DROC, l'estimation de cette dernière était incorrecte.

Ces anomalies représentent environ 6% des sinistres dans notre périmètre. Nous décidons cependant de les conserver afin d'éviter une perte de volumétrie supplémentaire.

Cadence de manifestation et choix de l'âge x

Rappelons que chaque sinistre est lié à un contrat et à une DROC. Nous appelons "âge de la DROC", le nombre d'années écoulés depuis la DROC. Ainsi lorsqu'on se limite à l'âge 5 de la DROC, nous faisons allusion à tous les sinistres survenus entre DROC et DROC + 5, pour une DROC donnée.

Dans cette partie nous étudions la cadence de manifestation en RCD, c'est-à-dire à quelle vitesse se manifestent les sinistres après le début des travaux. Nous chercherons également à savoir dans quelle mesure la typologie du contrat influe sur cette cadence.

Commençons par analyser la cadence de manifestation par DROC, pour les DROCs 1999-2003 :

Nous observons que les sinistres se manifestent de façon plus importante la troisième et la quatrième année après la DROC. Ensuite la cadence de manifestation commence à ralentir à partir de la cinquième année, pour représenter moins d'un pourcent après la douzième année.

On peut aussi ventiler la distribution par année de DROC pour voir si la distribution est similaire pour deux DROC différentes.

Ci-dessous les quantiles de la distribution empirique des sinistres par année de développement en ventilant par année de DROC.

- Nous pouvons nous apercevoir que les sinistres peuvent se manifester jusqu'à 17 ans après l'année de DROC (voir quantile à l'ordre 1 pour la DROC 1999). En revanche la cadence en nombre se manifeste dans environ 90% des cas avant 10 ans.

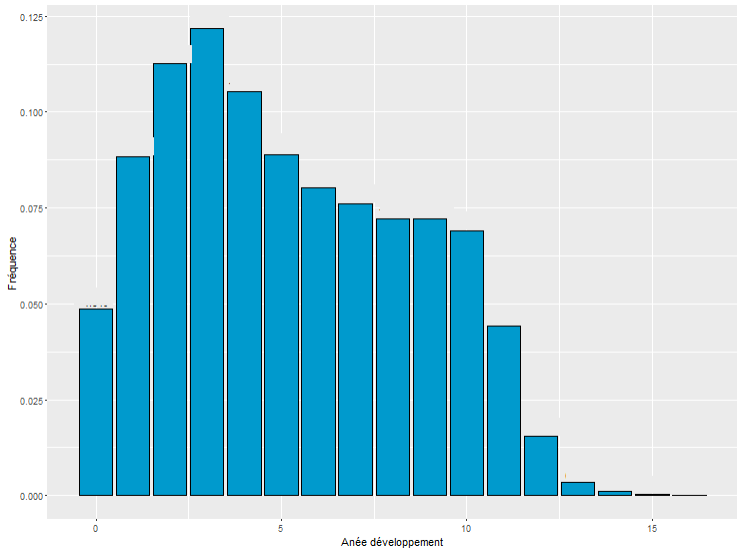


FIGURE 6.4 – Distribution des sinistres par année de développement

Quantiles empiriques en jours à l'ordre											
droc	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	1,00	$\widehat{F}(10ans)$
1 999	470	757	1 011	1 275	1 601	2 013	2 519	3 069	3 621	6 087	91%
2 000	429	739	1 029	1 309	1 654	2 054	2 504	3 050	3 603	5 648	91%
2 001	417	779	1 095	1 406	1 775	2 197	2 683	3 176	3 680	5 355	89%
2 002	444	833	1 159	1 484	1 862	2 310	2 776	3 243	3 719	5 034	89%
2 003	490	900	1 227	1 568	1 968	2 393	2 829	3 276	3 691	4 733	89%

Quantiles empriques en années à l'ordre											
droc	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	1,00	$\widehat{F}(10ans)$
1 999	1,3	2,1	2,8	3,5	4,4	5,5	6,9	8,4	9,9	16,7	91%
2 000	1,2	2,0	2,8	3,6	4,5	5,6	6,9	8,4	9,9	15,5	91%
2 001	1,1	2,1	3,0	3,9	4,9	6,0	7,4	8,7	10,1	14,7	89%
2 002	1,2	2,3	3,2	4,1	5,1	6,3	7,6	8,9	10,2	13,8	89%
2 003	1,3	2,5	3,4	4,3	5,4	6,6	7,8	9,0	10,1	13,0	89%

FIGURE 6.5 – Quantiles de la distribution des sinistres par année de développement

En considérant un âge de DROC à 5 ans, nous pouvons analyser les courbes de cadences pour chaque DROC.

Cette courbe est donnée par :

$$(x, \frac{\sum_{j=1}^{x} N_j}{\sum_{j=1}^{5} N_j})$$

avec $x = 1....5.$, et N_i le nombre de sinistres manifestés à l'age i.

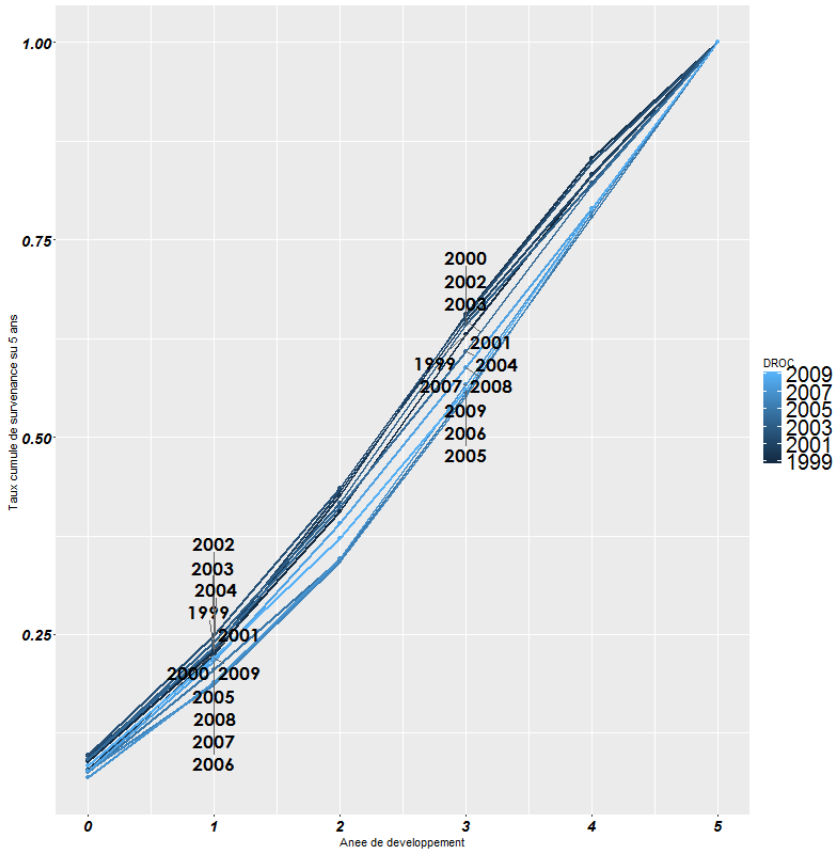


FIGURE 6.6 – Evolution de la cadence de manifestation sur 5 par année de DROC

Plus la courbe est concave (vers la gauche) plus les sinistres se manifestent rapidement ; nous parlons alors de cadence de manifestation rapide. A l'inverse plus la courbe est convexe (vers la droite) plus le développement est lent.

Ceci nous laisse penser qu'il existe en effet un effet temporel. La cadence de manifestation a tendance à décélérer (se manifester plus tardivement).

Une manière d'analyser la tendance est d'observer le pourcentage de sinistres manifestés à 3 ans selon l'année de DROC (sur une base de 5 ans).

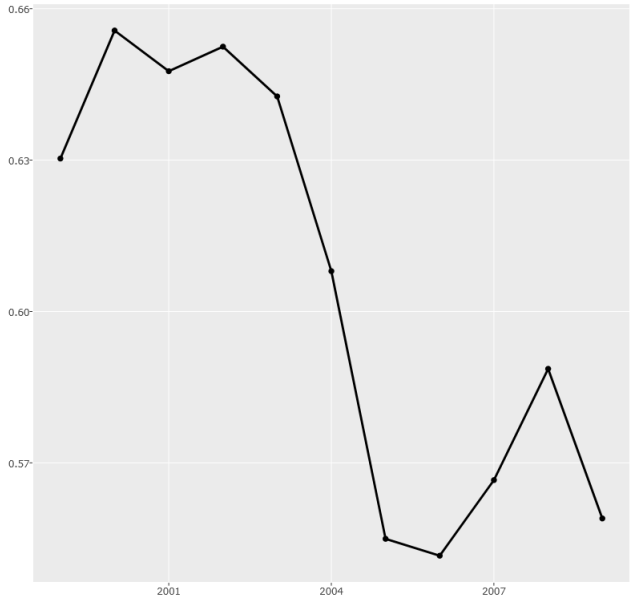


FIGURE 6.7 – Pourcentage de sinistres manifestés 3 ans après la DROC en fonction de la DROC

On observe en effet que la cadence de manifestation a effectivement ralenti entre les DROCS 2000 à 2006 mais que ce ralentissement a aujourd'hui tendance à se corriger. Nous observons cet effet sur tous les horizons de développement (5 à 10 ans).

Cette tendance peut être problématique pour notre modélisation car, dans notre **méthode à 5 ans**, les coefficients de passage sont supposés d'être indépendants des DROCs. [3]. Nous cherchons donc à comprendre l'origine de ce décalage.

Nous regardons donc, par modalité, la cadence de manifestation en fonction des variables tarifaires actuelles.

La cadence de manifestation a été ventilée par chacune des variables suivantes :

- tranche du chiffre d'affaires,
- ancienneté,
- nombre d'employés,
- classe de risque.

Nous constatons une discrimination uniquement pour la variable chiffre d'affaires : le décalage des cadences est plus ou moins marqué selon les tranches de chiffre d'affaires.

3. Ceci équivaut à l'hypothèse d'indépendance des coefficients de passage par rapport années de survenance dans la méthode de Chain Ladder

En effet, nous constatons que la cadence de manifestation ralentie avec la taille de l'entreprise : plus le chiffre d'affaires est élevé, moins cette cadence est rapide.

Si le portefeuille se tournait vers les grandes entreprises, ce décalage aurait pu donc être expliqué, or c'est justement le contraire qui est arrivé ; le portefeuille s'est tourné vers les petites entreprises ces dernières années.

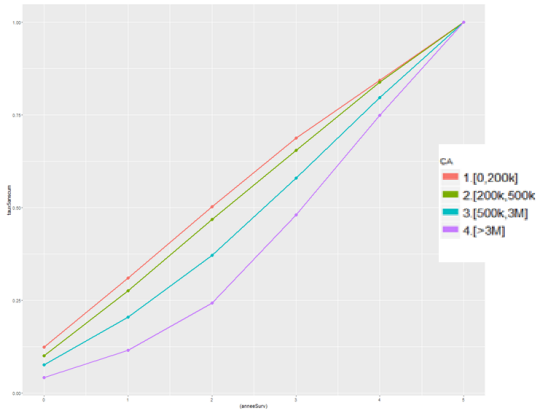


FIGURE 6.8 – Cadence de manifestation selon la tranche du CA

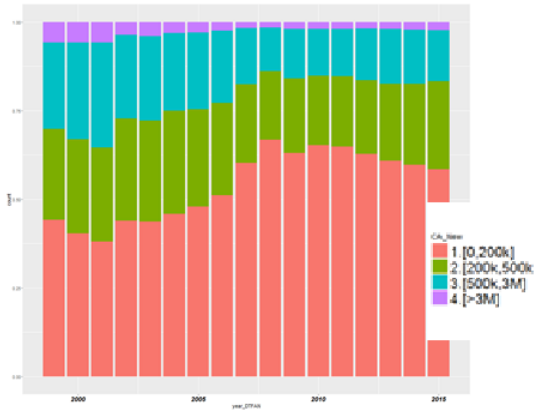


FIGURE 6.9 – Structure du portefeuille par CA

Nous concluons que le décalage de la cadence de manifestation est un phénomène global de marché, même s'il est plus marqué sur les petites entreprises.

De plus, ce décalage est moins visible lorsqu'on prend en compte une fenêtre de 10 ans de survenance.

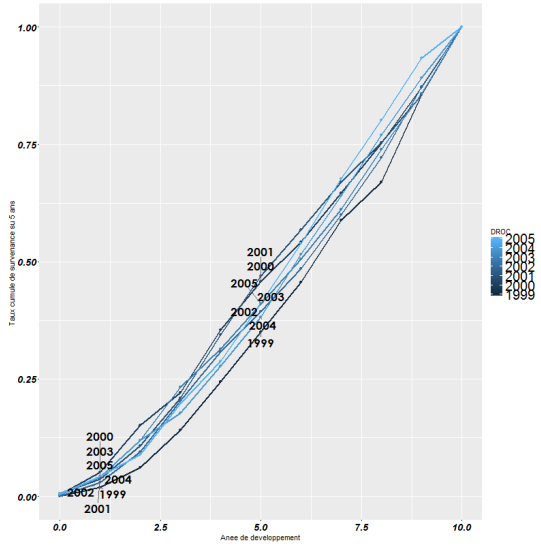


FIGURE 6.10 – Cadence de manifestation sur 10 ans, ventilation par DROC

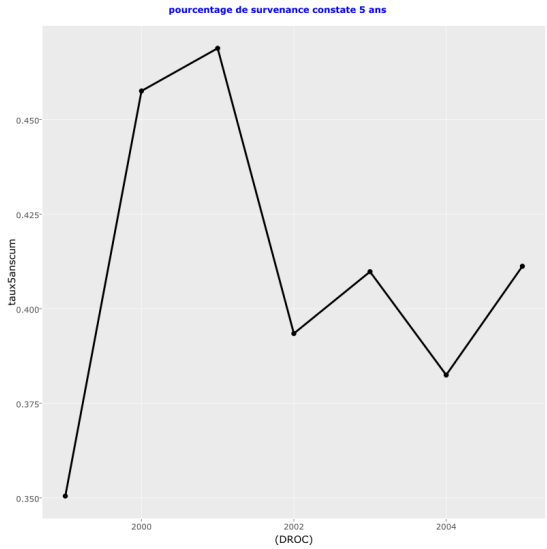


FIGURE 6.11 – pourcentage de sinistres manifestés 5 ans après la DROC, par DROC

En regardant la cadence de manifestation sur 10 ans de survenance, les DROCs 2000-2005 semblent se comporter de façon similaire.

Nous établissons deux périmètres d'étude pour la suite de nos travaux.

1. DROCS 2000-2005 à l'âge 10
2. DROCS 2005-2000 à l'âge 5

L'analyse des variables tarifaires sera effectuée, pour chacun de deux périmètres, avec comme objectif d'expliquer les charges totales par contrat.

Chapitre 7

Analyse des variables tarifaires

Rappelons que nous cherchons à tester **la méthode à 5 ans**, qui consiste à se limiter à 5 années d'observation des contrats récents pour modéliser la prime pure de la garantie RCD.

Néanmoins, seule l'observation de ces contrats à l'ultime pourra nous dire avec exactitude la pertinence de ce méthode.

Notre seule façon aujourd'hui d'analyser cette méthode, est de comparer ses résultats à ce qu'on aurait obtenu en modélisant sur une période plus longue en utilisant des contrats plus anciens.

En résumé nous allons effectuer toute la démarche de modélisation de la prime pure sur deux sous-périmètres distincts :

1. DROCS 2000-2005 à l'âge 10,
2. DROCS 2005-2000 à l'âge 5.

La première étape consiste à déterminer quelles sont les variables les plus explicatives de la fréquence et de la charge, et à étudier les corrélations entre les variables explicatives, de façon séparée pour chaque périmètre.

Nous commençons par introduire les outils mathématiques qui seront utilisés.

7.1 Méthodes utilisées pour l'analyse des données

L'analyse de la variance

(Besse, 2016)

Cette méthode sera utilisée pour étudier la corrélation entre une variable à expliquer Y, quantitative, et une variable explicative X qualitative.

Nous possédons d'observations y_i, x_i des variables Y, X, avec X prenant m modalités distinctes, notées $1, ...m$.

Notons $n_l = \sum_i 1_{x_i=l}$ le nombre d'individus prenant la modalité l pour la variable X et $\overline{y}_l = \frac{1}{n_l} \sum_{x_i=l} y_i$ la moyenne de la variable Y au sein de la modalité l de X.

Notons aussi $\sigma_l^2 = \frac{1}{n_l} \sum_i (Y_i - \overline{y}_l)^2$ la variance partielle de Y sur les individus prenant la modalité l, et $\overline{y}$ la moyenne de Y sur l'ensemble de la population : $\overline{y} = \frac{1}{n} \sum_l n_l * \overline{y}_l$.

Alors l'indice de liaison de Y avec X peut se déduire de la formule de la décomposition de la variance :

$$\sigma_Y^2 = \frac{1}{n} \sum_i (y_i - \overline{y})^2 = \frac{1}{n} \sum_{l=1}^{m} n_l (\overline{y}_l - \overline{y})^2 + \frac{1}{n} \sum_{l=1}^{m} n_l \sigma_l^2$$
$$\sigma_Y^2 = \sigma_E^2 + \sigma_R^2$$

Variance Totale= Variance Expliquée + Variance résiduelle

L'indice de corrélation que nous utiliserons sera donc donné par le R^2 d'un modèle d'ANOVA à un facteur :

$$S_{Y|X} = \sqrt{\frac{\sigma_E^2}{\sigma_R^2}}$$

La distance du χ^2 et le V de Cramer

Nous utiliserons cette méthode pour mesurer la liaison entre variables qualitatives.

Soit un tableau de contingence de deux variables qualitatives, X avec I modalités, et Y avec J modalités. Notons :

- $n_{i,j}$ l'effectif observé pour la modalité i de la variable X, et la modalité j de la variable Y,
- $n_{i,.}$ l'effectif total observé pour la modalité i de la variable X,
- $n_{.,j}$ l'effectif total observé pour la modalité j de la variable Y,
- n l'effectif total du tableau de contingence,
- $t_{i,j}$ l'effectif théorique pour la case (i, j).

Le V de Cramer est un indicateur de corrélation qui repose sur la distance du χ^2 du tableau.

Hypothèses pour le test du χ^2

- $\mathcal{H}_0$: les variables aléatoires X et Y sont indépendantes
- $\mathcal{H}_1$: les variables aléatoires X et Y ne sont pas indépendantes

En cas d'indépendance l'effectif théorique de la case (i, j) est égal au produit des fréquences empiriques fois n :

$$t_{i,j} = \frac{n_{.,j} * n_{i,.}}{n^2} * n = \frac{n_{.,j} * n_{i,.}}{n}$$

Nous calculons en suite la distance du tableau, qui compare les effectifs théoriques aux effectifs observés ;

$$D = \sum_{i}^{I} \sum_{j}^{J} \frac{(n_{i,j} - t_{i,j})^2}{t_{i,j}}$$

Sous l'hypothèse $\mathcal{H}_0$, D suit asymptotiquement une loi du χ^2 à $(I-1)*(J-1)$ degrés de liberté.

La $p - value$ du test est donnée par : $P(\chi^2 > D(obs))$ avec χ^2 une variable aléatoire suivant la loi du χ^2 à $(I-1)*(J-1)$ degrés de liberté.

Les variables X et Y seront d'autant plus corrélées que la distance du χ^2 sera élevée.

Conditions d'application du test Afin de pouvoir appliquer ce test, les effectifs totaux n doivent être supérieurs à 50, et les effectifs observés $n_{i,j}$ doivent être supérieurs à 5.

Ces conditions d'application sont tout à fait remplies car nous possédons, pour chacun des périmètres plus de 50 000 observations.

Inconvénients du test La valeur de D dépend du nombre des effectifs, et donc, ne nous permet pas d'avoir une idée de la corrélation entre ces deux variables.

Par contre, le V de cramer nous donne une meilleure idée de la liaison entre les deux variables. Il n'est rien d'autre qu'une normalisation de la distance du χ^2.

$$V = \sqrt{\frac{D}{n * min(I-1, J-1)}} \tag{7.1}$$

Il n'existe pas une valeur du V de Cramer pour laquelle le test d'indépendance est rejeté (pour un niveau de confiance donné). Nous conserverons un seuil à 10% pour distinguer les variables discriminantes.

L'ACP

L'ACP nous permet d'analyser les corrélations linéaires entre nos variables quantitatives.

Notons X la matrice de données de taille I, J qu'on cherche à réduire, avec :

- I= nombre d'individus totaux,
- J= nombre de variables totales.

L'objectif de l'ACP est de trouver les axes $u_1.., u_p$ qui maximisent l'inertie projetée [1].

On peut montrer que pour un axe donné i, l'inertie s'écrit :

$$I_1 = \frac{1}{n}(Xu_i)^t(Xu_i)^1 \tag{7.2}$$

On peut également montrer que ce programme d'optimisation revient à chercher les vecteurs propres associés à la matrice $\frac{1}{n}X^tX$, et de les ranger pour trouver les axes les plus discriminantes.

Sans entrer plus dans le détail, il convient de préciser que l'ACP détecte uniquement les interactions linéaires des variables.

Les arbres de régression

Les arbres de régression sont des méthodes très populaires en "data mining", qui permettent de capter des liaisons non linéaires entre les variables explicatives et à expliquer (Stéphane, 2015).

Le principe est de prédire une variable réponse Y à partir des covariables $X_1, X_2...X_p$ en découpant l'espace des covariables en 2 groupes successivement ; lorsque la variable Y est nominale nous parlons d'arbre de décision, et lorsqu'elle est continue nous parlons d'arbre de régression (Shalizi, 2015).

A chaque noeud de l'arbre, nous cherchons la meilleure segmentation de la population selon un test appliqué à une des variables explicatives, qui sera de la forme $(X < s)$ ou $(X_i = x_{i,j})$. Les deux groupes issus de cette séparation sont appelés feuilles, ou "leaves"

1. la distance moyenne des individus au centre de gravité

en anglais. Cet algorithme est répété de façon récursive jusqu'à garder seulement une observation dans chaque feuille.

Puisque la meilleure segmentation tient compte de la variable réponse, les arbres de régression sont des algorithmes de classification supervisés, à différence des méthodes du type "k-means" ou Classification Ascendante Hiérarchique (CAH).

La segmentation optimale dépendra du critère à optimiser, le critère par défaut pour les arbres de régression est l'erreur en moyenne quadratique donnée par :

$$MSE(T) = \frac{1}{n} \sum_{c \in leaves(T)} \sum_{i \in c} (y_i - m_c)^2 \tag{7.3}$$

$$m_c = \frac{1}{n_c} \sum_{i \in c} y_i \tag{7.4}$$

Finalement, l'arrêt de l'algorithme dépendra d'un critère à définir, comme par exemple :

- la variation lorsque la variation du risque (MST est inférieur à un seuil cp,
- le nombre d'observations minimal dans une feuille, qui es définie par $minbucket$ sous R,
- le nombre d'observations minimal dans un noeud pour être séparé $minsplit$ sous R.

On peut également définir une profondeur optimale au préalable avec $maxdepth$

La complexité des arbres repose donc sur le moment d'arrêt.

Validation croisée et élagage des arbres ("pruning")

(Breiman et al., 1984) donne l'un des résultats le plus importants pour les arbres de régression :

soit :

- $|T|$ le nombre de noeuds de l'arbre,
- $R(T)$ une fonction de risque appliqué à l'arbre,(pour les arbres de régression l'erreur moyen quadratique),

et le coût-complexité d'un arbre pour un paramètre de complexité cp, donné par :

$$R_{cp} = R(T) + cp * |T| \tag{7.5}$$

Alors il existe un arbre T_{cp} qui minimise R_{cp}, et plus encore une suite d'arbres emboités :

$$T_m \subset ..... \subset T_2 \subset T_1 \subset T_{max} \tag{7.6}$$
$$CP_m > ...CP_2 > CP_1 > CP_0 = 0 \tag{7.7}$$

Ainsi la méthode la plus utilisée pour sélectionner le meilleur arbre consiste à trouver l'arbre optimal pour chaque valeur de cp, et de calculer l'erreur par validation croisée de chaque arbre. Dans le cadre de ce mémoire nous utilisons le package *rpart* de R.

Les méthodes mentionnées ci-dessus nous serviront d'outil pour l'analyse des variables tarifaires et des corrélations.

7.2 Application à Étude des variables tarifaires

Traitements des variables explicatives

Rappelons tout d'abord nos deux sous-périmètres d'étude :

1. DROCs 2000-2005 à l'age 10 [2],
2. DROCs 2005-2009 à l'age 5.

Nous commençons par détailler les analyses qui ont été faits indépendamment du périmètre.

Exclusion des variables Avant traitement nous avons plus de 200 variables tarifaires provenant de sources internes et externes. Pour la modélisation, nous ne retenons que les plus cohérentes d'entre elles. Nos critères d'exclusion ont été les suivants :

- valeurs manquantes pour plus de 50% de l'exposition,
- concentration de l'exposition dans une des modalités,
- incohérence des modalités, exemple : indicatrices avec valeur égal à 2,
- incohérence avec les autres variables tarifaires,
- avis de l'entreprise.

En outre, nous créons les variables supplémentaires suivants :

Création des nouvelles variables

1. région d'implantation : notre base comporte déjà la variable départements, nous regroupons cette variable en fonction des 13 nouvelles régions de la France ;
2. âge de l'entreprise-DROC : nous calculons une nouvelle variable âge qui varie en fonction de la DROC, elle est égale au nombre d'années écoulées entre la création de l'entreprise et la DROC ;
3. Indicatrice sinistre : prend la valeur 1 lorsque le contrat-DROC comporte au moins un sinistre.

Pour les variables restantes, nous regroupons les modalités des variables comptant le moins d'effectifs.

2. On conserve 10 ans d'observations après la DROC

Fusion des modalités

- pour les variables catégorielles : fusion des modalités semblables avec très peu d'effectifs. Par exemple :

Nous créons également des tranches pour les variables explicatives numériques :

TABLE 7.1 – Variables tarifaires quantitatives

Variable	**Description**
ca	chiffre d'affaires
age_entreprise_droc	années écoulées entre la date de la création et la DROC
age_contrat_droc	années écoulées entre la première souscription et la DROC
effectif	nombre d'employés
nbetact	nombre d'établissement actifs de l'entreprise
nb_fam30	nombre d'activités de l'entreprise
ancienneté	ancienneté de l'entreprise

Tout d'abord nous construisons des classes d'effectif égal. La fonction *Classintervals* de R nous permet d'effectuer ces tranches à partir de plusieurs critères, nous utilisons les quantiles empiriques.

Néanmoins, ces méthodes ne tiennent pas compte de la variable à expliquer. Nous utilisons donc des arbres de régression pour en tenir compte, et cela sur nos **deux périmètres préétablis**.

Création des tranches en utilisant des arbres de régression

Les arbres de régression sont des algorithmes supervisés, ils considèrent une variable réponse.

Les variables réponses sont les suivantes :

TABLE 7.2 – Variables réponse

Notation mathématique	**Variables**	**type**
S	les charges totales	continue
N	la fréquence des sinistres	discrète
Y	le coût moyen des sinistres des contrats sinistrés	continue
B	l'occurrence d'un sinistre (fréquence >0)	dichotomique

Nous utilisons la variable S charges totales **à 10 ans** par contrat DROC pour la construction des tranches des variables suivantes :

- Chiffre d'affaires,
- ancienneté de l'entreprise (3 types),
- note financière de l'entreprise,
- nombre de familles d'activité.

Nous abordons la démarche utilisée pour la variable : **Nombre de familles d'activité**, qui a été la même pour les autres.

On commence par estimer l'arbre maximal, qu'on obtient en définissant $cp = 0$ dans le package "rpart" de R.

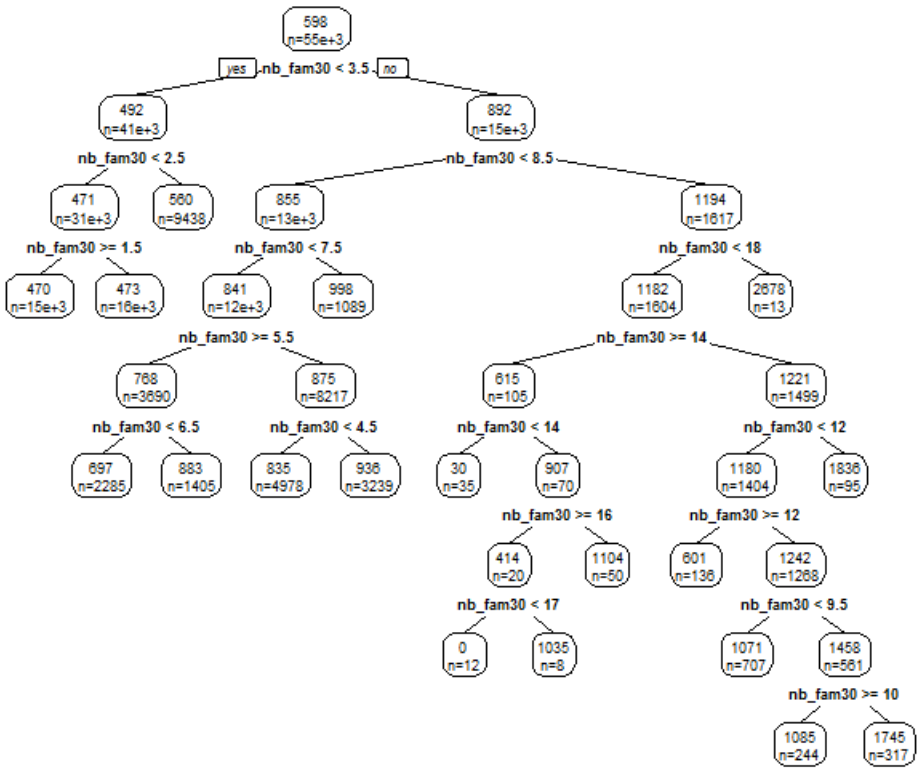


FIGURE 7.1 – Arbre de régression pour la discrétisation des variables, nombre de famille d'activités

Nous cherchons ensuite le meilleur sous-arbre par validation croisée ; la commande *printcp* divise la population en 10 groupes, et calcule l'erreur la moyenne des erreurs quadratique (normalisée) de chaque arbre T_{cp} sur les 10 groupes.

La colonne *xerror* nous donne la moyenne des erreurs sur les 10 groupes, dans notre cas l'arbre $T_{2.3089e-04,nsplit=1}$ est le meilleur arbre de régression, qui a deux feuilles.

Nous coupons la variable nbfam30 de la façon suivante :

- 1.[0-3]
- 2.[>3]

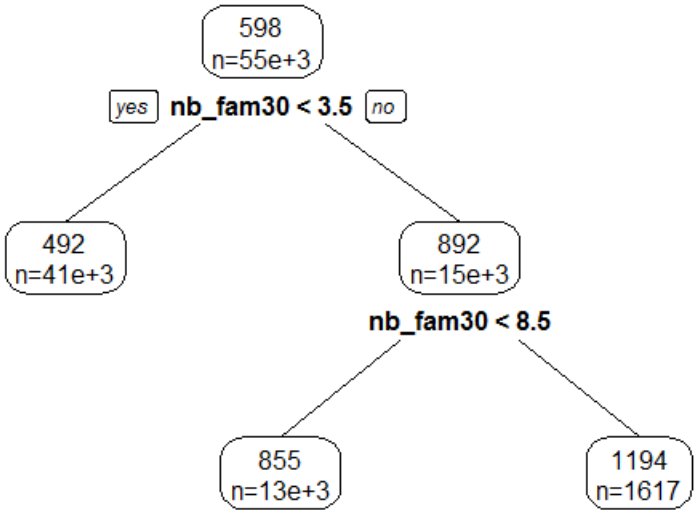


FIGURE 7.2 – Arbre de régression optimale d'après la validation croisée pour la discrétisation des variables, nombre de famille d'activités

Après avoir traité toutes nos variables explicatives, nous étudions les corrélations entre elles, et entre la variable à expliquer et les variables explicatives.

7.3 Corrélation entre variables

Nous abordons dans cette partie les corrélations entre variables. D'une part, nous cherchons à connaître les variables qui expliquent le mieux les charges totales par contrat, et d'autre part nous cherchons à exclure les variables très fortement corrélées, car cela empêche les algorithmes d'optimisation des GLM de converger.

Ces analyses sont à nouveau, à mener en distinguant nos **deux périmètres**

Nous avons deux types d'analyse à mener :

1. entre les variables quantitatives ;
2. entre les variables qualitatives.

Entre les variables tarifaires

Variables quantitatives On commence par effectuer une ACP sur les variables tarifaires quantitatives, les cercles de corrélation de l'ACP sont les suivantes :

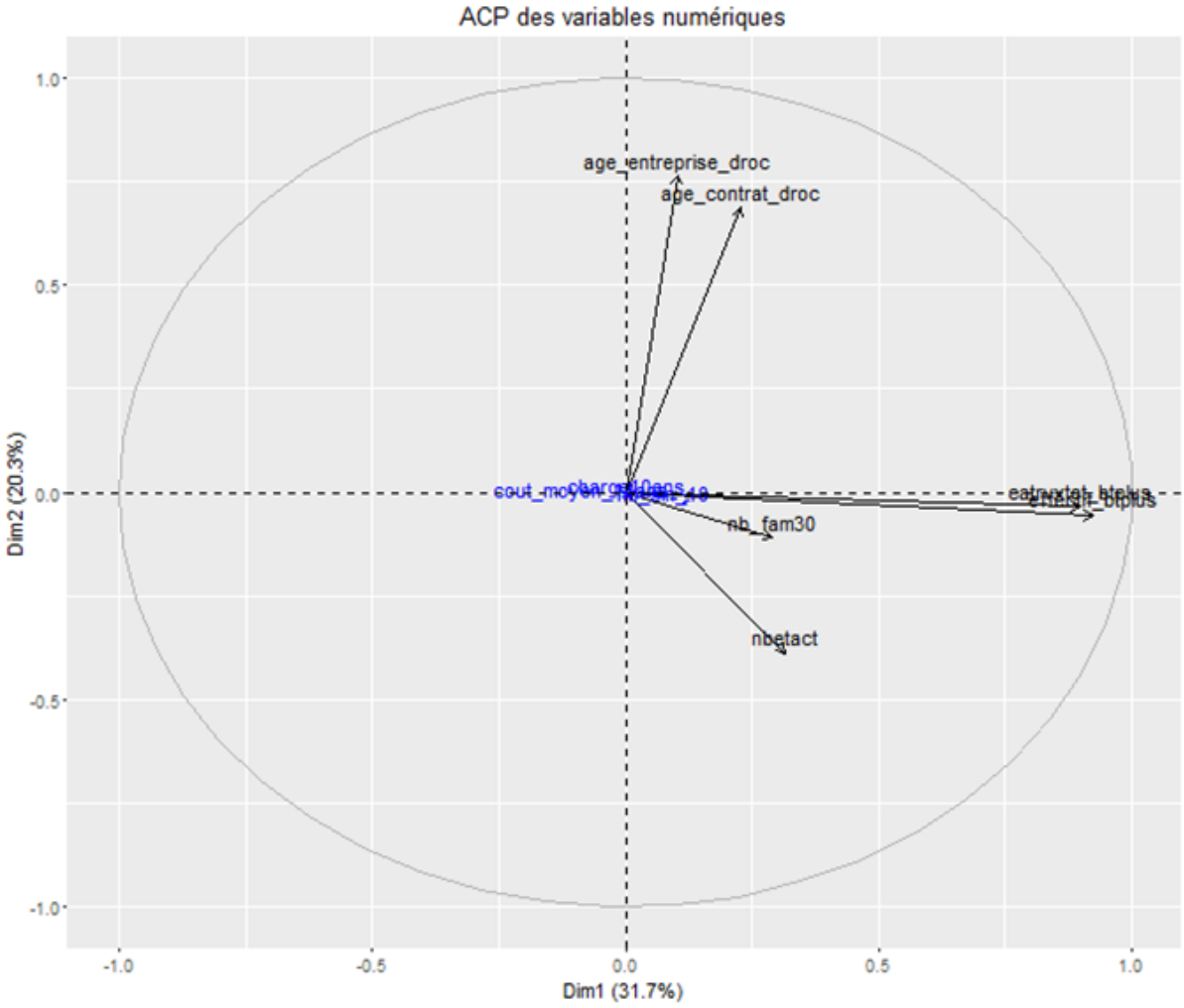


FIGURE 7.3 – ACP, variables quantitatives, périmètre 1

Dans les deux cas, les deux premiers axes expliquent plus de 50% de la variance. Le premier axe correspond, pour les deux cas, à la taille de l'entreprise, représenté par les variables

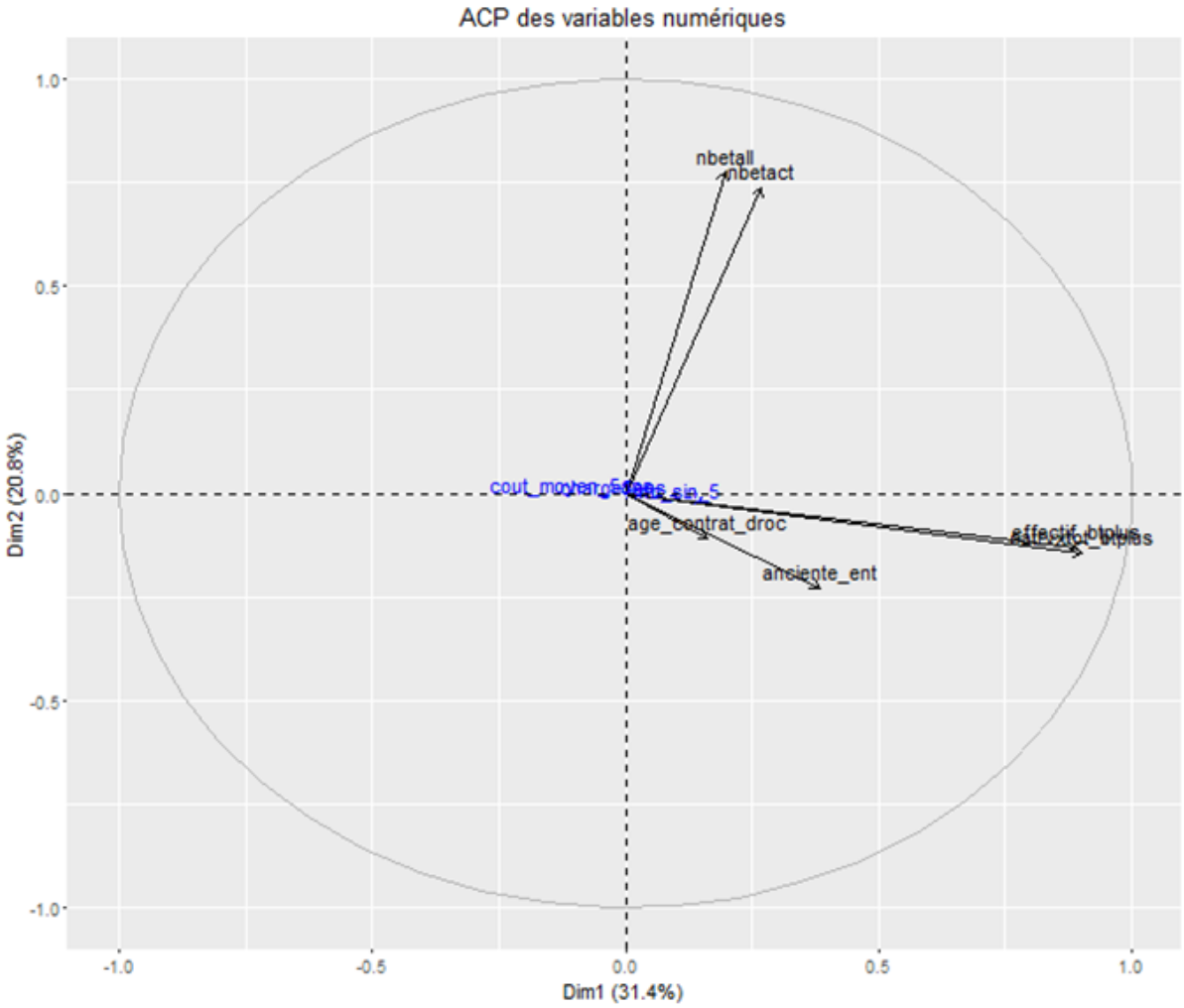


FIGURE 7.4 – ACP, variables quantitatives, périmètre 2

effectif et chiffre d'affaires. Nous observons aussi que les variables nombre d'établissements et nombre d'activités contribuent partiellement à cet axe.

Le deuxième axe correspond dans un cas à l'ancienneté des contrats, sans d'autre apport particulier des variables.

Nous considérons uniquement deux axes car le pourcentage de variance expliquée décroit "fortement" pour le 3ième axe. Nous nous apercevons avec l'aide du critère du coude.

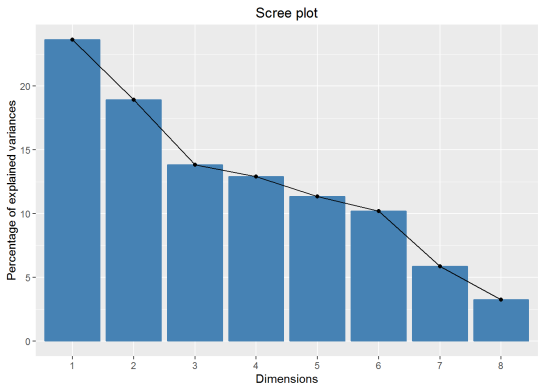


FIGURE 7.5 – Critère du coude

Les résultats de l'ACP nous révèlent une corrélation très forte de la variable chiffre d'affaires et effectif, ce qui nous amène à considérer uniquement la variable chiffre d'affaires pour la modélisation.

Nous présentons également les corrélations de Pearson de ces variables :

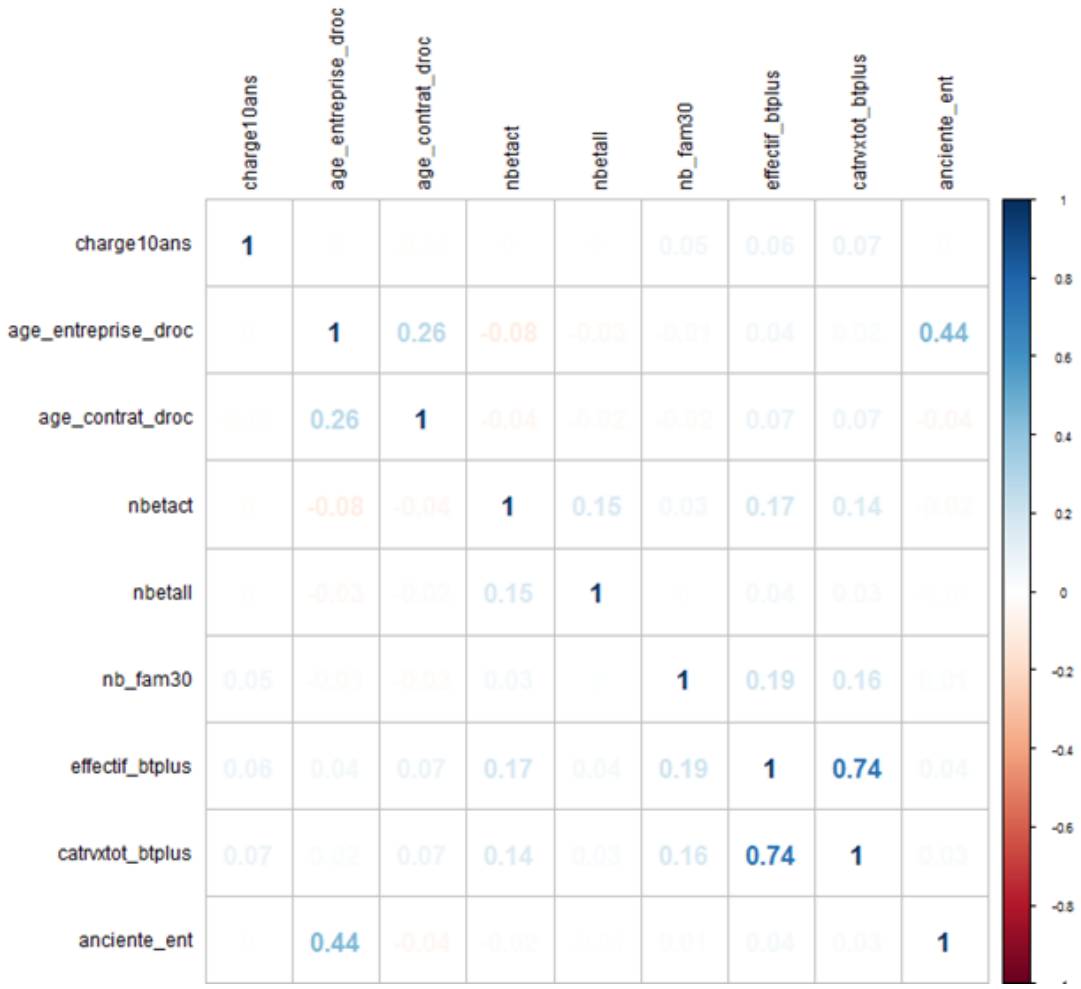


FIGURE 7.6 – Matrice de corrélation de Pearson, périmètre 1

Nous observons des corrélations remarquables entre les variables du type âge, nous décidons de conserver uniquement la variable ancienneté qui correspond au nombre d'années écoulées entre la création de l'entreprise et la souscription.

Par ailleurs, rien ne semble expliquer linéairement les charges totales (variables illustratives en bleu dans l'ACP).

L'étude des variables qualitatives ne peut pas être fait à partir d'une ACP, nous avons envisagé une Analyse de Composantes Multiples (ACM), mais qui n'a pas produit de résultat satisfaisant ; les deux premiers axes représentent uniquement **10% de la variance**.

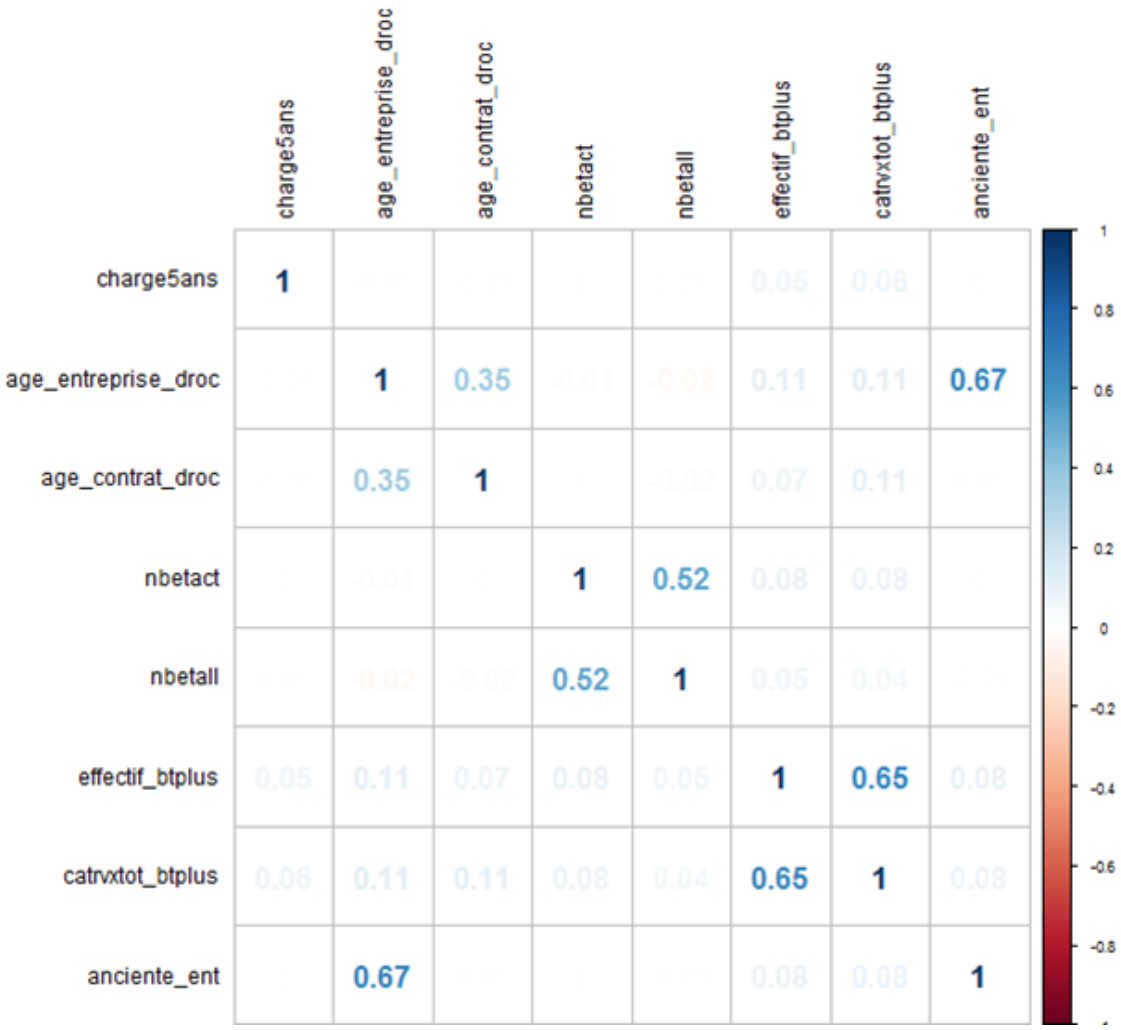


FIGURE 7.7 – Matrice de corrélation de Pearson, périmètre 2

Étude des variables qualitatives

Nous étudions tout d'abord les corrélations entre les variables tarifaires et choisissons le V de Cramer comme mesure de corrélation.

Les variables qualitatives étudiées sont :

- activité majoritaire,
- région d'implantation,
- famille d'activités majoritaire,
- nombre d'établissements (tranche),
- origine,
- antécédents,
- qualification,
- tranche d'effectif,
- note financière (tranche),
- chiffre d'affaires, (tranche),
- ancienneté (tranche).

Nous calculons les corrélations à partir du V de Cramer 2 à 2, l'application $Cramer\_V$ d'AXA nous permet de visualiser les corrélations sous forme d'un cercle ;

Les variables famille d'activités, classe de risque et activités majoritaires sont fortement corrélées. Ces trois variables représentent des groupements d'activités du secteur de la construction. Nous conservons uniquement la variable **classe de risque** qui est le groupement le plus robuste.

On conclut cette partie par la recherche des variables qui expliquent le mieux les variables d'intérêt, à savoir la sinistralité. Plus précisément nous effectuons une "Analysis of Variance" (ANOVA) à 1 facteur pour expliquer les charges totales, la fréquence et le coût moyen d'un contrat-DROC en fonction d'une variable explicative.

Le graphique suivant classe les variables en fonction du R^2 de l'ANOVA. Nous représentons uniquement les variables avec un $R^2 > 10\%$.

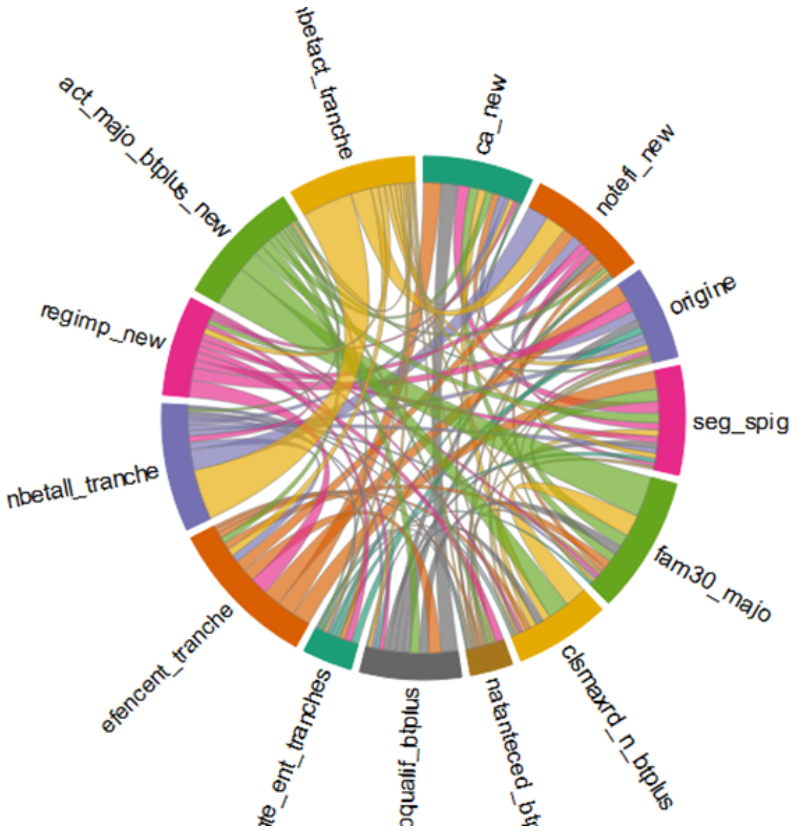
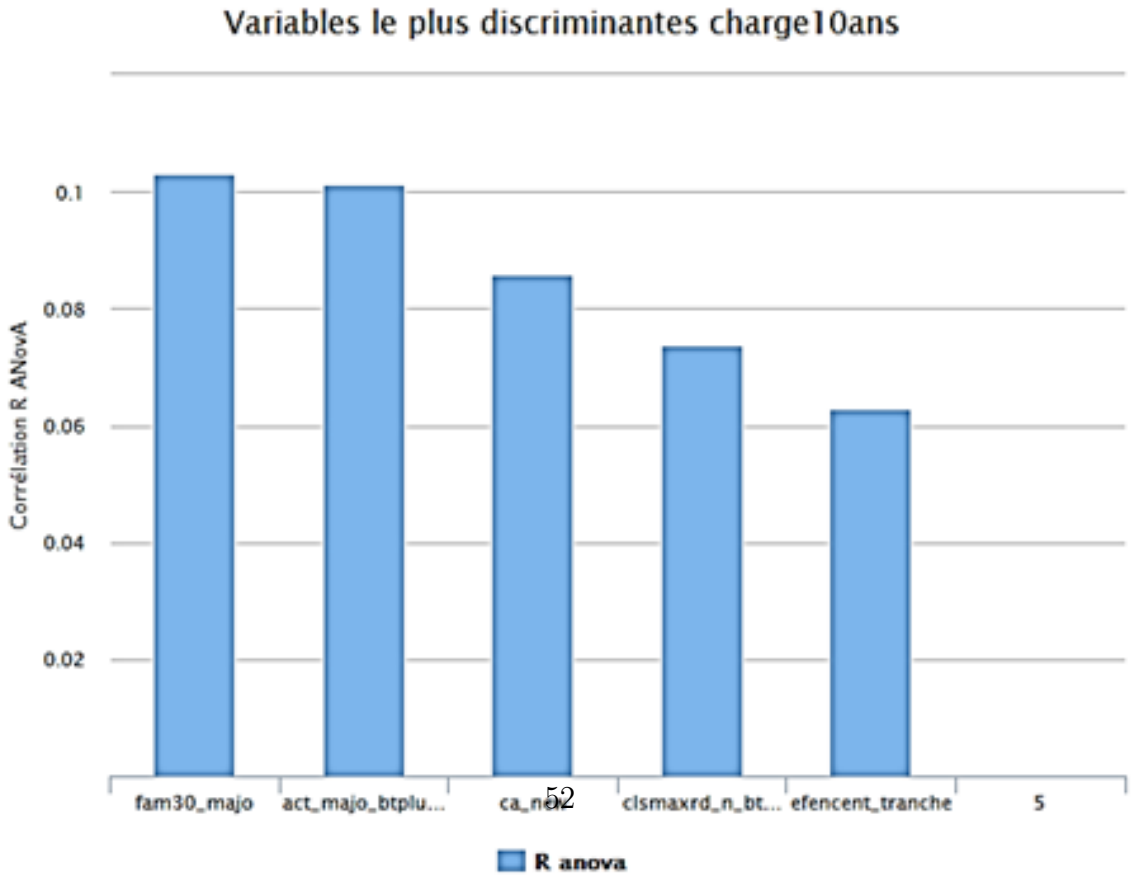


FIGURE 7.8 – Critère du coude



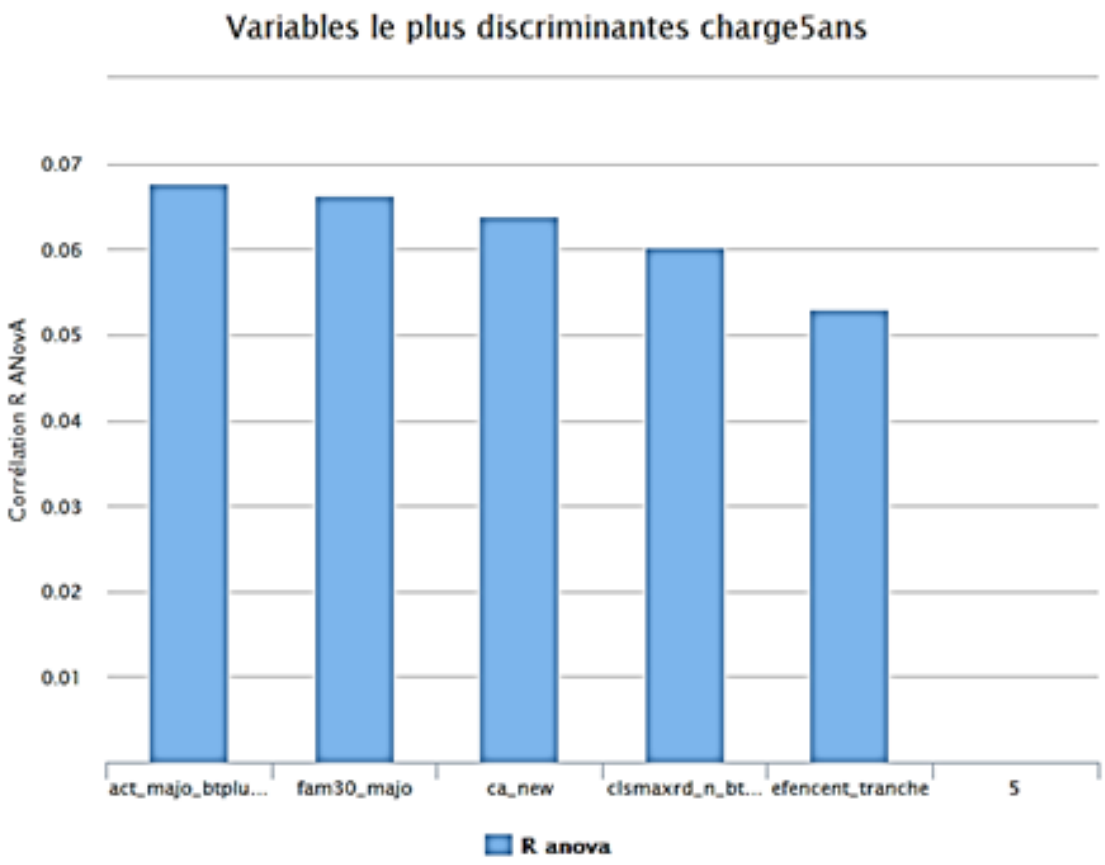


FIGURE 7.10 – Variables explicatives charge, périmètre 2

Cette analyse nous permet de capter des corrélations non linéaires non présentes lors de l'ACP.

Les variables le plus significatives sont le chiffre d'affaires (en tranches), l'effectif, la classe de risque et la note financière.

Nous effectuons la même analyse pour les variables fréquence et coût moyen.

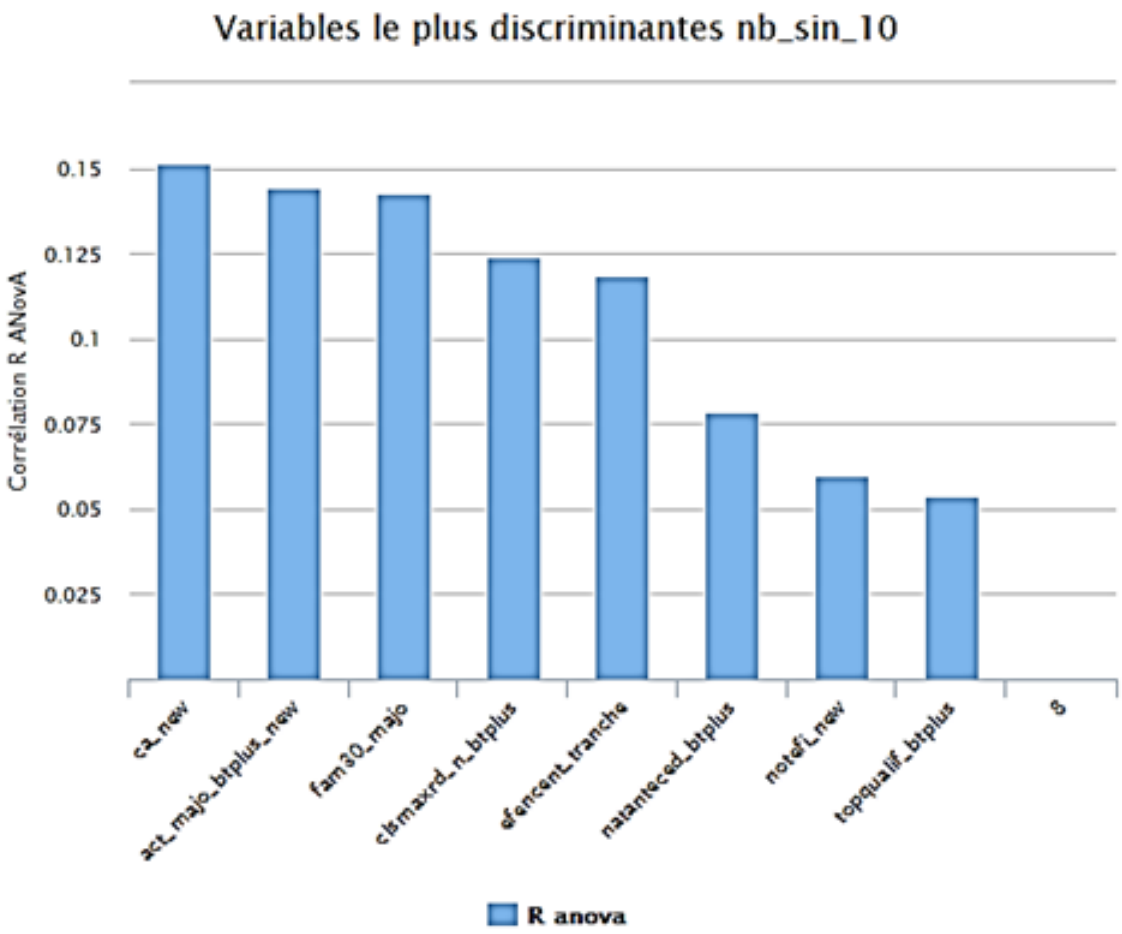


FIGURE 7.11 – Variables explicatives fréquence, périmètre 1

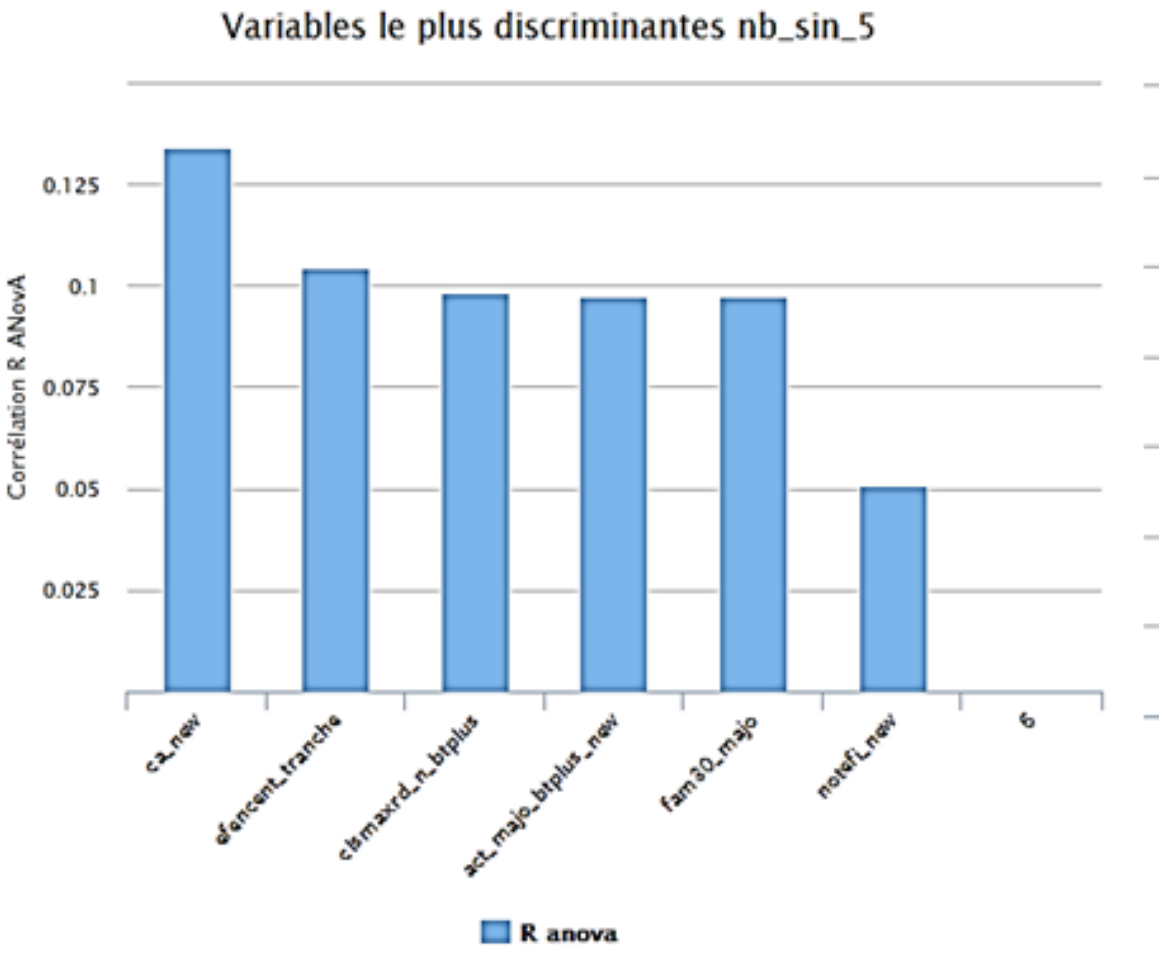


FIGURE 7.12 – Variables explicatives fréquence, périmètre 2

Pour la variable fréquence les mêmes variables ressortent comme significatives, à savoir le chiffre d'affaires, la classe de risque et la note financière.

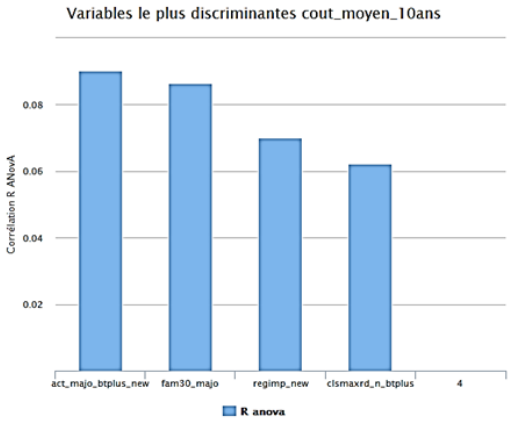


FIGURE 7.13 – Variables explicatives coût moyen, périmètre 1

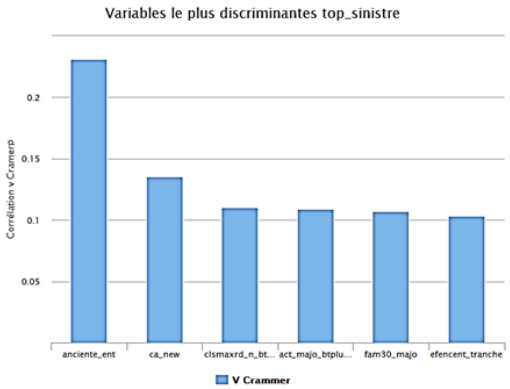


FIGURE 7.14 – Variables explicatives coût moyen, périmètre 2

Finalement, rien ne semble expliquer le coût moyen du sinistre, dans le graphique ci-dessous nous relâchons le critère du 10%. Nous verrons plus tard, dans la modélisation de la prime pure, que nous ne possédons pas de variables explicatives du coût moyen, le pouvoir de segmentation du modèle de coût moyen sera donc très faible.

Conclusions de cette partie Nous avons effectué une analyse exhaustive de notre base de données. Cela nous a permis d'identifier les variables potentiellement explicatives de la sinistralité.

Par ailleurs nous avons détecté des corrélations qui pourraient nuire à notre modélisation ; notamment entre le chiffre d'affaires et l'effectif, et entre les variables d'activité. Nous avons décidé de n'en conserver qu'une pour chaque type.

Finalement, nous avons construit des tranches des variables quantitatives, qui nous permettront dans la section suivante de mieux prédire les charges d'un contrat à partir des GLM.

Troisième partie

La tarification de la garantie RCD

Dans cette partie nous abordons les différentes approches pour modéliser la prime pure de la garantie RCD en assurance construction.

En Assurance IARD, nous procédons le plus souvent selon les approches suivantes :

1. modélisation des charges totales
2. modélisation de la fréquence et du coût moyen séparément

Nous abordons dans un premier temps l'approche des charges totales, qui est traditionnellement modélisée par une loi de Poisson Composée. Nous expliquons la démarche utilisée et les résultats et défauts du modèle. Nous mettrons un accent particulier sur la sélection des variables et des modalités discriminantes.

Cette approche sera élargie, pour modéliser non pas un montant de prime, mais un taux applicable au chiffre d'affaires. Pour ce faire, nous étudions brièvement la prise en compte de l'exposition dans les GLM et tout particulièrement dans les modèles de Tweedie.

Pour conclure avec cette approche, nous testerons des modèles non paramétriques de "machine learning". Ces algorithmes s'avéreront très efficaces et robustes, mais ne nous permettent pas d'interpréter les résultats d'un point de vue économique.

Dans un deuxième temps nous modéliserons séparément la fréquence et le coût moyen en utilisant les GLM. Pour commencer, nous utilisons les lois classiques (Gamma, Poisson), pour ensuite palier au problème de surdispersion dans nos données, à travers des modèles dits "Zéro inflated".

Cette étude nous permettra, d'une part, d'établir les modèles les mieux adaptés à nos données, et de proposer des améliorations aux modèles utilisés actuellement, ainsi que d'évaluer les différences entre les approches, charges totales et fréquence-coût moyen.

Chapitre 8

Éléments théoriques

Nous commençons par expliciter les outils mathématiques dont nous nous servirons pour la modélisation de la prime pure en RCD.

Avant tout, nous rappelons l'origine des deux approches de modélisation : charges totales et fréquence - coût moyen.

8.1 Bases théoriques de la modélisation de la prime pure

Le cadre général de la tarification *a priori* (Charpentier, 2010-2011) suppose que la charge totale d'un contrat est égale à :

$$S = \sum_{i=1}^{N} Y_i \tag{8.1}$$

Avec $Y_1, ..., Y_n$ des variables aléatoires de même loi simulant le coût des sinistres, et N le nombre de sinistres pendant la période observée.

Sous l'hypothèse

Hypothèse 1 *N et $Y_{i=1....n}$ sont des variables aléatoires indépendantes*

la prime pure d'une police d'assurance est égale d'après l'identité de Walder à :

$$E[S] = E[N] * E[Y]. \tag{8.2}$$

C'est l'approche fréquence-coût moyen.

On peut élargir ce cadre très simpliste en supposant que N et $Y_i, i = 1.....n$ sont conditionnellement indépendants par rapport à un ensemble d'information noté Ω, qui englobe l'information apportée par les variables tarifaires. Nous obtiendrons ainsi comme estimation de la prime pure :

$$E[S|\Omega] = E[N|\Omega] * E[Y|\Omega] \tag{8.3}$$

Cette identité est démontrée dans Foata (2004).

Ainsi, nous avons deux approches possibles pour modéliser la prime pure $E[S|\Omega]$:

- Estimer $E[N|\Omega]$ et $E[S|\Omega]$ séparément,
- Estimer $E[S|\Omega]$ directement.

Il est plus courant parmi les praticiens d'effectuer une modélisation du type fréquence -coût moyen car ceci permet d'évaluer le risque de fréquence et de sévérité séparément, en retenant un jeu de variables explicatives distinctes pour chaque variable d'intérêt. Néanmoins ceci est plus couteux en temps de calcul.

8.2 Les Modèles linéaires généralisés

Les modèles linéaires généralisés introduits par (Nelder and Weddernurn, 1972) sont très populaires en tarification. Ils permettent d'englober un vaste ensemble de modèles paramétriques en utilisant un seul algorithme d'optimisation (Kaas, 2005).

Les GLM possèdent trois composantes principales :

1. une composante aléatoire : des observations des variables aléatoires $Y_1....Y_n$ de densité appartenant à la famille exponentielle caractérisées par leur moyennes $\mu_i, i = 1...n$ et paramètre de dispersion ψ_i.
2. un prédicteur linéaire $\eta_i = \sum_j \beta_j * x_{i,j}$, avec $X = (x_{i,j})$ la matrice de covariables (variables explicatives) et $(\beta_1, ..\beta_p)$ le vecteur de coefficients à estimer.
3. une fonction de lien $g(.)$ telle que $\eta_i = g(\mu_i)$ supposée une fois différentiable et inversible.

La famille des lois exponentielles

On dit que la distribution de Y appartient à la famille des lois exponentielles si sa densité s'écrit comme :

$$f_Y(y_i) = c(y_i, \psi_i) * exp\{\frac{\theta_i y_i - b(\theta_i)}{\psi_i}\} \tag{8.4}$$

- $\psi = \frac{\phi}{w}$. avec ϕ le paramètre de dispersion identique pour toutes les observations, et w_i le poids *a priori*, exposition ou poids de crédibilité (Kaas, 2005),
- $c(.,.)$ est une fonction de normalisation, qui ne dépend pas de θ_i,
- θ_i est le paramètre naturel, qui est à valeurs dans un ouvert,
- b est une fonction deux fois différentiable, et sa dérivée est inversible.

L'espérance et la variance d'une loi appartenant à la famille exponentielle sont données par :

$$E[Y_i] = b'(\theta_i)$$
$$Var[Y_i] = \phi * b''(\theta_i) = V(\mu_i) * \frac{\phi}{w_i}$$

On explicite également la fonction de variance, qui détermine entièrement la loi d'une loi appartenant à la famille exponentielle.

$$V(\mu) = (b'' \circ (b')^{-1})(\mu)$$

Pour une telle distribution la log-vraisemblance s'écrit :

$$l(y, \theta, \phi, w) = \sum_{i=1}^{n} w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + \sum_{i=1}^{n} c(y_i, \phi_i)$$

Estimation par maximum de vraisemblance

La méthode la plus courante pour l'estimation des paramètres est le maximum de vraisemblance. Étant donnée la log-vraisemblance explicitée précédemment, les conditions de première ordre pour un maximum local donnent comme équations à résoudre :

$$\frac{\delta l}{\delta \beta_k} = \sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{V(\mu_i) * g'(\mu_i)} * x_{i,k} = 0$$

La Déviance

Tous les tests d'adéquation et de significativité des variables reposent sur la notion de déviance.

Remarquons que la solution naturelle équations précédents consiste à estimer μ_i par y_i, cette solution correspond au modèle dit "saturé". La déviance est donc une sorte de distance entre le modèle saturé et le modèle issu de la régression.

$$D(y, \mu) = 2(l(y_1, ...y_n) - l(\mu_1, ...\mu_n)) \tag{8.5}$$

Remarques

- Entre deux modèles nous choisissons celui avec la déviance la plus petite.
- La déviance D suit asymptotiquement une loi du $\chi^2(n-p)$ à n-p degrés de liberté, avec n le nombre d'observations et p le nombre de paramètres. Ce résultat nous permettra d'effectuer un test d'adéquation globale.
- Nous pouvons nous faire une idée de l'adéquation globale d'un modèle en regardant le rapport $D/(n-p)$ qui doit être inférieur à 1.

L'exposition au risque

Soit N_i la fréquence annualisée de sinistre pour l'assuré i ; avec $N_i \sim \mathcal{P}(\lambda_i)$, si l'assuré a été observé pendant une période E_i, le nombre de sinistres observés $\sim \mathcal{P}(\lambda_i * E_i)$. Ceci découle de la propriété d'accroissements indépendants du processus de Poisson (Charpentier, 2010-2011).

Ainsi l'espérance de la fréquence de sinistralité pour un contrat observé pendant e_i années est : $\lambda_i * e_i$.

Variable offset

Considérons le GLM de Poisson, réduit à une variable explicative.

$$Y_i|_{X_i,E_i} \sim \mathcal{P}(\lambda_i * E_i) \tag{8.6}$$
$$\lambda_i = exp(\beta_0 + \beta_1 * X_1) \tag{8.7}$$

Ce modèle est équivalent à :

$$Y_i|_{X_i,E_i} \sim \mathcal{P}(\tilde{\lambda}_i) \tag{8.8}$$
$$\tilde{\lambda}_i = E_i * exp(\beta_0 + \beta_1 * X_1) = exp(\beta_0 + \beta_1 * X_1 + 1 * log(E_i)) \tag{8.9}$$

On conclut que dans la fonction *glm* de R, ajouter la variable offset comme le logarithme de l'exposition revient à supposer que :

$$E[Y_i|_{X_{1,i},E_i}] = \mu_i * E_i \tag{8.10}$$
$$\mu_i = E[Y_i|_{X_{1,i},E_i=1}] \tag{8.11}$$

Ce résultat reste valide lorsqu'on définit la fonction de lien comme le logarithme népérien.

Néanmoins, l'introduction de la variable offset peut compromettre la reproductibilité du modèle.

Definition 8.2.1: *Modèle reproductible On dit que le modèle est reproductible lorsqu'en combinant des observations avec le même profil de risque nous obtenons un modèle équivalent.*

$$\frac{Y_1 * w_1 + Y_2 * w_2}{w_1 + w_2} = Y_{1+2} \tag{8.12}$$

avec

- Y_i *la variable réponse ;*
- w_i *le poids a priori.*

Exemples de lois utilisées pour la famille exponentielle

Les lois de Tweedie

On dit que Y appartient à la famille de lois de Tweedie si sa fonction de variance peut s'exprimer comme :

$$V(\mu) = \mu^p \tag{8.13}$$

(Jose, 2014)

Connaissant la forme de la fonction de variance, la déviance d'un tel modèle s'écrit, pour $p \in (1, 2)$ comme :

$$D = \sum_i^n \frac{1}{\phi} * w_i \{ \frac{y_i}{1-p} (\mu_i^{1-p} - y_i^{1-p}) - \frac{1}{2-p} (y_i^{2-p} - \mu_i^{2-p}) \}$$

Introduction de la variable Offset pour le modèle de Tweedie

De plus, lorsqu'on utilise la fonction de lien logarithmique $g(x) = ln(x)$, les conditions du premier ordre du maximum de vraisemblance s'écrivent :

$$\frac{\delta l}{\delta \beta_k} = \sum_{i=1}^n w_i \frac{y_i - \mu_i}{\mu_i^{p-1}} * x_{i,k} = 0$$
$$\forall k = 1....p$$

En introduisant une composante offset $z_i = ln(e_i)$ dans le prédicteur linéaire (e_i étant l'exposition) :

$$\eta_i^* = z_i + \sum_{j=1}^p \beta_j * x_{i,j} = z_i + \eta_i$$

la moyenne μ_i du nouveau modèle devient

$$\mu_i^* = exp(\eta_i^*) = \mu_i * e_i \tag{8.14}$$

Et la vraisemblance du nouveau modèle s'écrit :

$$\frac{\delta l}{\delta \beta_k} = \sum_{i=1}^{n} w_i * e_i^{2-p} \frac{\frac{y_i}{e_i} - \mu_i}{\mu_i^{p-1}} * x_{i,k} = 0 \tag{8.15}$$

$$\forall k = 1....p \tag{8.16}$$

Nous en déduisons qu'il existe deux modèles équivalents dans le cadre de la modélisation par une loi de Tweedie ($p \in (1,2)$) :

1. Modèle 1
   (a) variable réponse y_i charges totales
   (b) offset = années polices e_i
   (c) poids *a priori* w_i
2. Modèle 2
   (a) variable réponse $\frac{y_i}{e_i}$ charges totales
   (b) pas de variable offset
   (c) poids *a priori* $w_i' = w_i * e_i^{2-p}$

La loi Gamma

On dit que Y suit la loi Gamma $\Gamma(\alpha, \beta)$ si sa densité s'écrit comme :

$$f_Y(y) = \frac{1}{\Gamma(\alpha)} e^{-\beta y} \beta^{\alpha} y^{\alpha-1} * \mathbf{1}_{y \geq 0} \tag{8.17}$$

Le paramètre α est un paramètre de forme ou "shape" en anglais.

Le paramètre β est appelé le paramètre d'intensité ou "rate" en anglais. Dans le monde anglo-saxon il est plus commun de définir la densité par le paramètre d'échelle ou "shape" en anglais.

$$shape = \frac{1}{rate} \tag{8.18}$$

On peut facilement la réécrire de la façon suivante :

$$f_Y(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)} * \alpha^{\frac{1}{\alpha}} exp\Big\{ \frac{(-\frac{\beta}{\alpha})y - \Big(- ln(-(\frac{-\beta}{\alpha})) \Big)}{\frac{1}{\alpha}} \Big\}$$

On peut reconnaitre donc :

$$\theta = \frac{-\beta}{\alpha} \qquad \phi = \frac{1}{\alpha}$$
$$b(x) = -ln(-x) \qquad b'(x) = -\frac{1}{x} \qquad b''(x) = \frac{1}{x^2}$$

La fonction de lien canonique pour une loi de Gamma est donc $1/x$ et non pas $log(x)$.

Son espérance est donnée par

$$E[Y] = b'\Big(\frac{-\beta}{\alpha}\Big) = \frac{\alpha}{\beta}$$

La variance est donnée par :

$$Var[Y] = \phi * b''(\theta) = \frac{\alpha}{\beta^2}$$

Et finalement la fonction de variance est égale à :

$$V(\mu) = (b'' \circ b')(\mu) = \mu^2$$

On reconnait donc que la loi Gamma appartient à la famille des lois Tweedie avec $p = 0$.

La loi de Poisson Composée $GPG(\lambda, \alpha, \beta)$

Cette loi introduite par Jorgensen (1987) correspond à la loi de :

$$S = \sum_{i}^{N} Y_i$$

Avec $Y_0 = 0, Y_i \sim gamma(\alpha, \beta_i); \forall i > 0$ et $N \sim \mathcal{P}(\lambda)$, $Y_{i=1...n} \perp\!\!\!\perp N$.

Sa distribution est donnée par :

$$f_{S_N}(ds) = \delta_0(ds) * e^{-\lambda} + \sum_{n \geq 1} \frac{s^{n\alpha-1}\beta^{n\alpha}e^{-\beta s}}{\Gamma(n\alpha)} \frac{e^{-\lambda}\lambda^k}{k!} ds$$

La loi de Poisson Composée est donc une loi de mélange. Elle n'est pas absolument continue par rapport à la mesure de Lebesgue et sa masse en zéro est donnée par :

$$P(S_N = 0) = e^{-\lambda}$$

Son espérance et sa variance sont données par :

$$E[S_N] = E[S] * E[N] = \frac{\lambda * \alpha}{\beta} = \mu$$
$$Var[S_N] == \frac{\alpha * \lambda}{\beta^2} * (1 + \alpha)$$

On peut ainsi démontrer que la loi de Poisson Composée correspond à un cas particulier des familles de loi Tweedie avec $p \in (1, 2)$.

Et on a par ailleurs le résultat suivant :

Proposition (Jorgensen, 1987) : Soit un modèle de Poisson Composé $CPG(\lambda, \alpha, \beta)$ la puissance p est donnée par $\frac{\alpha+2}{\alpha+1}$.

La loi de Poisson

La fonction de densité d'une loi de Poisson est donnée par :

$$f(k, \lambda) = \frac{exp(-\lambda)\lambda^k)}{k!} \tag{8.19}$$

Qui est absolument continue par rapport à la mesure de comptage. Cette loi est communément utilisée pour modéliser la fréquence d'un évènement et a comme particularité la propriété d'équidispersion : $E[N] = Var[N] = \lambda$ (Charpentier, 2010-2011).

Elle correspond également à un cas particulier des lois de Tweedie, avec paramètre de puissance p=1. Son paramètre de dispersion ϕ est connu et vaut 1.

La loi Binomiale Négative dans le modèle de Poisson mélangé

En présence de surdispersion nous pouvons considérer un modèle Poisson mélangé (Charpentier, 2010-2011). Il consiste à supposer qu'il existe une variable aléatoire positive Θ, $E[\Theta] = 1$ telle que :

$$P(N = k|\Theta) = \frac{[\lambda * \Theta]^k * exp(-\lambda * \Theta)}{k!} \tag{8.20}$$

Lorsque $\Theta \sim \gamma(\theta, \theta)$, cette loi correspond à une loi binomiale négative de densité :

$$f(k, \lambda, \theta) = \frac{\Gamma(k+\theta)}{\Gamma(\theta) * k!} \frac{\lambda^k * \theta^\theta}{(\lambda+\theta)^{k+\theta}} \tag{8.21}$$

Et on a :

$$E[N] = E[E[N|\Theta]] = E[\Theta * \lambda] = \frac{\theta}{\theta} * \lambda \tag{8.22}$$

$$Var[N] = Var[E[N|\Theta]] + E[Var[N|\Theta]] = \tag{8.23}$$
$$var[\Theta * \lambda] + E[\Theta * \lambda] = \tag{8.24}$$
$$\frac{\lambda^2}{\theta} + \lambda > \lambda \tag{8.25}$$

Nous avons donc une méthode qui nous permet de prendre en compte la surdipersion à travers du paramètre θ, sans impacter l'estimation de l'espérance $E[N_i] = \lambda_i$.

Lorsque le paramètre θ est connu, cette loi appartient à la famille exponentielle. Dans le cas contraire, la fonction glm de R réalise une estimation itérative de β pour une valeur donnée de θ et vice-versa jusqu'à convergence.

Modèles "Zero-inflated"

En cas de surdispersion, et particulièrement, en cas d'une forte présence de zéros, on peut envisager de modifier la loi de probabilité pour accorder un poids supplémentaire aux 0.

Pour la loi de Poisson "Zero-Inflated", la densité devient :

$$P(N = k) = \begin{cases} w + (1-w) * exp(-\lambda) & k = 0 \\ (1-w)exp(-\lambda)\frac{\lambda^l}{k!} & k > 0 \end{cases}$$

Cette loi correspond à la loi de $Y = X * B$ avec $X \sim \mathcal{P}(\lambda)$ et $B \sim Ber(1-w)$, $X \perp\!\!\!\perp B$.

Où w est le poids supplémentaire accordé aux zéros. La variance et l'espérance de cette loi sont :

$$E[Y] = \mu = (1-w)\lambda < \lambda$$
$$Var[Y] = (\lambda - \lambda^2) * (1-w) - \lambda^2 * (1-w)^2 =$$
$$\mu + \frac{w}{1-w} * \mu^2$$

On remarque que le paramètre w modifie cette fois-ci l'espérance, contrairement au cas de la loi Binomiale Négative.

Cette loi ne correspond pas à un cas particulier de la famille exponentielle, mais ses paramètres peuvent s'estimer en utilisant les mêmes algorithmes. Le package $Zeroinfl$ de R permet de construire des modèles de régression à inflation de zéros.

Choix entre un modèle simple et Inflated

Le test de Vuong permet de tester l'adéquation d'un modèle Zero-Inflated, face à un modèle simple.

La statistique du test est (Vuong, 1989) :

$$LR_n(\hat{\theta}, \hat{\gamma})/\hat{w_n}\sqrt{n}$$

Avec $LR_n(\hat{\theta}, \hat{\gamma})$ le logarithme du rapport de vraisemblance des deux modèles, w la variance de $LR_n(\hat{\theta}, \hat{\gamma})$, et n la taille de l'échantillon.

L'hypothèse nulle du test est :

- $LR_n(\hat{\theta}, \hat{\gamma}) = 0$

Sous cette hypothèse $LR_n(\hat{\theta}, \hat{\gamma})/\hat{w_n}\sqrt{n}$ suit une loi $\mathcal{N}(0, 1)$.

Sous l'hypothèse nulle les deux modèles sont aussi adéquats vis à vis des données (ou aussi loin de la vérité). Dès lors que le test est rejeté, nous choisirons le modèle avec la log-vraisemblance la plus élevée.

Critères de sélection des variables

Il existe plusieurs méthodes de sélection des variables explicatives. Dans le cadre de cette étude nous avons testé les méthodes pas à pas basées sur l'AIC et le BIC.

"Akaike's Information Criterium" AIC

Soit un modèle d'ajustement paramétrique, l'AIC est donné par :

$$AIC = -2 * l(\beta) + 2 * p \tag{8.26}$$

Avec

- p le nombre total de coefficients estimés,
- β les paramètres estimés du modèle,
- $l(.)$ la log-vraisemblance sous les hypothèses de distribution implicites dans le modèle.

Ce modèle pénalise l'ajustement du modèle par sa complexité.

"Bayesien Information Criterion" BIC

Le critère de "Bayesian Information Criterion" (BIC) part du même principe de pénalisation de la complexité de l'AIC, mais en utilisant comme poids le logarithme du nombre d'observations :

$$BIC = -2 * l(\beta) + ln(n) * p \tag{8.27}$$

Avec

- p le nombre total de coefficients estimés, (somme des modalités des variables),
- β les paramètres estimés du modèle
- $l(.)$ la log-vraisemblance sous les hypothèses de distribution implicites dans le modèle,
- n le nombre d'observations.

Sélection pas à pas des variables

- **La méthode ascendante** : Cette méthode consiste à partir du modèle réduit à la constante et d'ajouter à chaque étape la variable qui augmente le plus l'AIC (ou le BIC), jusqu'à ce que l'ajout d'une variable supplémentaire réduise le critère choisi.
- **La sélection descendante** : Ici, nous partons du modèle qui contient toutes les variables explicatives, et nous enlevons la variable qui fait augmenter le plus l'AIC (ou BIC) à chaque étape, jusqu'à ce que la suppression d'une variable supplémentaire fasse incrémenter l'AIC (ou BIC).

Tests d'adéquation

Test de Wald type 1

On cherche à tester une hypothèse de la forme :

$$\mathcal{H}_0 : h(\beta) = 0$$

Où $h = (h_1, ..., h_r)$ est une fonction régulière de β.

Ce test de type 1 est basé sur la normalité asymptotique du estimateur maximum de vraisemblance. Sous certaines hypothèses de régularité et dans la limite d'un grand nombre de données on a la propriété (Bérard, 2016).

$$Loi(\hat{\beta} - \beta) \approx \mathcal{N}(0, \mathcal{I}(\beta)^{-1}) \approx \mathcal{N}(0, \mathcal{I}(\hat{\beta})^{-1})$$

et

$$Loi(h(\hat{\beta})^t Q(\hat{\beta}^{-1} b(\hat{\beta})) = \chi^2(r) \tag{8.28}$$

Avec

$$B(\beta) = [\frac{\partial h_l}{\partial \beta_j}]_{l=1...r, j=000d} \tag{8.29}$$

$$Q(\beta) = B(\beta)(\mathcal{I}(\beta)^{-1})^t B(\beta) \tag{8.30}$$

Avec I la matrice d'information de Fisher qui est donnée par :

$$\mathcal{I}_{j,k}(\beta) = \mathbb{E}\Big(\Big[\frac{\partial logL}{\partial \beta_j}(\beta, Y, X)\Big]\Big[\frac{\partial logL}{\beta_k}(\partial\beta, Y, X)\Big] | X = x\Big) \tag{8.31}$$

Pour une loi appartenant à la famille exponentielle, cette matrice est donnée par :

$$\mathcal{I}_{j,k} = \sum_{i=1}^{n} w_i \frac{x_{i,j} x_{i,k}}{\phi V(\mu_i)} \frac{1}{g'(\mu_i)^2} \tag{8.32}$$

Ce test est effectué automatiquement par la fonction glm de R et affiche le résultat du test $h(\beta) = \beta_k = 0$ pour chaque coefficient du modèle.

Intervalles de confiance

La loi asymptotique de l'estimateur de vraisemblance nous permet d'estimer les intervalles de confiance à un seuil α (Rouvière, 2015).

Notons $\hat{\sigma}_j^2$ le $j - ième$ terme de la diagonal de $\mathcal{I}_{j,k}(\hat{\beta})$,(l'estimateur de σ_j^2).

Un estimateur de $?_j^2$ est donné par le j ?ième terme de la diagonal de $\mathcal{I}_{j,k}(\hat{\beta})$.

Nous avons asymptotiquement que :

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_j} \sim \mathcal{N}(0,1)$$

Remarquons que cela nous permet également de construire le lest : $\mathcal{H}_0 : \beta_j = 0$, qui a comme $p - value$ $P(N > \frac{\hat{\beta_j^{obs}}}{\sigma_j})$ avec N une v.a $\mathcal{N}(0,1)$.

$$IC(\beta_j) = [\hat{\beta}_j - u_{1-\alpha/2} * \hat{\sigma}_j; \hat{\beta}_j + u_{1-\alpha/2} * \hat{\sigma}_j]$$

Avec u_α le quantile d'ordre α de la loi Normale centrée réduite.

Ces intervalles seront calculés lors des graphiques des coefficients multiplicatifs.

Test du rapport de vraisemblance, type III

On considère deux modèles emboités et une partition de la matrice de covariances X.

$$X = [X_1 X_2]$$
$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Avec X_1 une matrice $n * q$ et X_2 $n * (p - q)$

- **Modèle Complet** $\eta = X_1\beta_1 + X_2\beta_2$
- **Modèle réduit** $\eta = X_1\beta_1$

L'objectif du test est de déterminer si le modèle réduit est adéquat vis à vis des données :

- $\mathcal{H}_0 : \beta_2 = 0$
- $\mathcal{H}_1 : \beta_2 \neq 0$

Les déviances du modèle réduit D_R, et du modèle complet D_F "Full Model" sont données par :

$$D_R = 2(l(\tilde{\theta}) - \sup_{\beta_2=0,\beta_1} l(\theta))$$
$$D_F = 2(l(\tilde{\theta}) - \sup_{\beta_2,\beta_1} l(\theta))$$

avec $l(\tilde{\theta})$ la log-vraisemblance du modèle saturé.

Sous l'hypothèse $\mathcal{H}_0$ on a :

$$\frac{D_R - D_F}{\phi} \to \chi^2_{p-q} \tag{8.33}$$

Ce test est conduit de deux façons différentes :

- Lorsque le paramètre de dispersion ϕ est connu, nous comparons la valeur de $P(\chi^2_{p-q} > \frac{D_R - D_F}{\phi}(obs))$ à un seuil de risque (5% généralement) ;
- lorsque le paramètre de dispersion ϕ n'est pas connu, on utilise la statistique $\frac{D_R - D_F}{\hat{\phi}}$ qu'on compare à une loi de Fisher $F_{(p-q),(n-p)}$.

Test d'adéquation globale

Ce test consiste à tester (Rouvière, 2015) :

- $\mathcal{H}_0$ Le modèle $\mathcal{M}_\beta$ de dimension p est adéquat ;
- $\mathcal{H}_1$ Le modèle saturé est adéquat.

Au moyen de la déviance on sait que le modèle est d'autant mieux ajusté que la déviance est petite. Sous $\mathcal{H}_0$ la déviance du modèle $D_{\mathcal{M}_\beta}$ converge en loi vers une χ^2_{n-p} avec n le nombre d'observations et p le nombre de paramètres ; ceci nous donne la p-value du test.

Il est également courant de comparer le rapport $\frac{D_{\mathcal{M}_\beta}}{n-p}$ à 1, $n-p$ étant l'espérance de la loi asymptotique sous $\mathcal{H}_0$.

Critères pour la comparaison des modèles

La courbe de Lorenz

La courbe de Lorenz permet de juger le pouvoir de segmentation du modèle. Elle est construite de la façon suivante.

Notons :

- $\mu_i = \hat{y}_i$ la prédiction de la variable réponse pour l'individu i,
- y_i les valeurs observés de la variable réponse pour l'individu i.

Nous rangeons ensuite les observations d'après les prédictions en ordre descendant tel que : $\hat{y}_{(1)} \geq \hat{y}_{(2)} .... \geq \hat{y}_{(n)}$. Sous cet ordre nous construisons la courbe à partir des observations (Charpentier, 2010-2011)

$$\{\frac{i}{n}, \frac{\sum_k^i y_{(k)}}{\sum_k^n y_{(k)}}\} \tag{8.34}$$

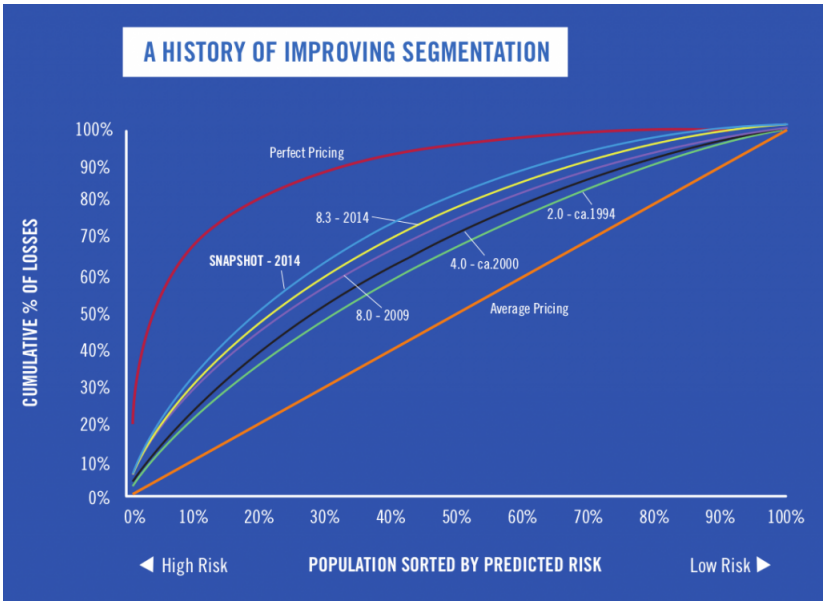


FIGURE 8.1 – Courbe de Lorenz, source : Arthur Charpentier `http://freakonometrics.hypotheses.org/20144`

Il est courant de calculer l'Area under the curve (AUC), qui correspond tout simplement à la surface en dessous de la courbe.

La segmentation parfaite correspond à un AUC égal à 1, à l'inverse une segmentation aléatoire correspond à un AUC égale à $\frac{1}{2}$.

Cette courbe nous permet de savoir si les individus les plus risqués, d'après le modèle, sont ceux qui ont encouru le plus de sinistres. Le modèle parfait correspond à la courbe en rouge.

Néanmoins elle ne permet pas de voir si, au sein d'un profil de risque, la prédiction se rapproche de la moyenne des charges observées.

La "Lift Chart"

Nous reprenons ce critère d'ajustement de (Jose, 2014). Cette courbe permet simultanément de juger le pouvoir de segmentation et de prédiction du modèle.

Sa construction est très similaire à celle de la courbe de Lorenz ; à partir des prédictions rangées en ordre décroissant : $\hat{y}_{(1)} \geq \hat{y}_{(2)} .... \geq \hat{y}_{(n)}$, nous divisons la population en k groupes avec exposition uniforme (généralement 10 groupes).

Pour chaque groupe k, nous calculons la moyenne des charges prédites $\frac{\sum_i^{n_k} \hat{y}_{(i)}}{n_k}$, la moyenne des charges observées $\frac{\sum_i^{n_k} y_{(i)}}{n_k}$ et l'exposition totale du groupe.

Voici un exemple, calculé sur le premier modèle qu'on traitera plus tard.

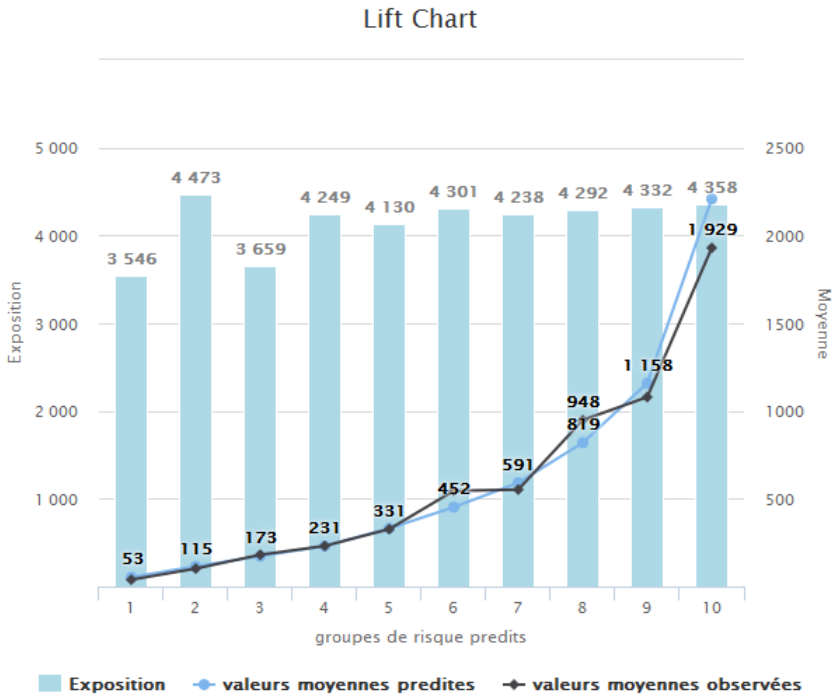


FIGURE 8.2 – Exemple liftchart

8.3 Le GBM

Nous abordons finalement le GBM, qui sera utilisé pour la modélisation de la prime pure.

Le GBM est une méthode d'agrégation d'arbres de prédiction très populaire en Data Science. Elle découle de l'algorithme Adaboost proposé par Friedeman et Schapireen en 1997. Nous abordons cette méthode sans entrer en détail dans les algorithmes d'optimisation.

Considérons une matrice de covariables X de taille (n, p), avec n individus et p variables explicatives et $y = (y_1, ..., y_n)$ le vecteur des variables réponses.

Le but de l'algorithme GBM est de minimiser l'espérance d'une fonction de perte L :

$$\tilde{F}() = arg \min_{F() \in \mathcal{F}} E_{y,x}[L(y, F(x)] \tag{8.35}$$

Cette fonction, en présence d'un grand nombre d'individus, peut s'approcher par la fonction de risque empirique :

$$\tilde{F}() = arg \min_{F() \in \mathcal{F}} \frac{1}{n} \sum_{i}^{n} [L(y_i, F(x_i)] \tag{8.36}$$

Dans l'algorithme GBM chaque fonction $F() \in \mathcal{F}$ correspond à une agrégation d'arbres de la forme suivante :

$$F(x) = F^{[0]} + \sum_{m=1}^{M} \beta^{[m]} h(x, \xi^{[m]}) \tag{8.37}$$

Pour estimer $F(x)$ le GBM procède de façon récurrente. Décrivons brièvement l'algorithme.

1. Estimer $\hat{F}^{[0]}$ par :

$$\hat{F}^{[0]}(x) = arg \min_{\xi in \mathbf{R}} \sum_{i=1}^{n} L(y_i, \xi)) \tag{8.38}$$

2. Calculer le vecteur de descente du gradient

$$r_{i,m} = -[\frac{\partial L(y_i, F(x_i)}{\partial F(x_i)}]_{F=\hat{F}_{[m-1]}} \tag{8.39}$$

3. Ajuster $r_{i,m}$ sur X, le résultat de cette étape est l'arbre $h(x, \xi^{[m]})$
4. Calculer

$$\hat{F}^{[m]} = \hat{F}^{[m-1]} + \beta^0 h(x, \xi^{[m]}) \tag{8.40}$$

   Pour une valeur de β^0 donnée, jusqu'à convergence.

5. Ainsi F est estimé par $F^{[0]} + \sum_{m=1}^{M} \beta h(x, \xi^{[m]})$, et il nous reste à trouver les vrais coefficients $\beta^{[m]}$:

$$\overrightarrow{\beta} = arg \min_{\beta}(y_i - F^{[M]}(x))^2 \tag{8.41}$$

Dit de façon grossière, le GBM consiste à appliquer des arbres de régression (ou de décision), de façon successive, sur le vecteur de descente du gradient de l'arbre précédent.

Chapitre 9

Application : modélisation de la prime pure par l'approche charges totales

Dans ce chapitre nous abordons la modélisation des charges totales par un modèle de Tweedie. Nous présentons la démarche, les difficultés rencontrées et des améliorations possibles.

9.1 Données et hypothèses

Rappel sur les données disponibles

Pour chaque contrat ayant souscrit la garantie RCD, nous possédons autant de lignes que d'années de souscription. Notre idée initiale a été de mener une étude **longitudinal**, c'est-à-dire de d'expliquer les charges du contrat i correspondantes à la DROC t y_i^t par des variables de risque x_i^t.

Les données disponibles ne nous permettent pas de continuer avec cette approche. Chaque observation y_i^t est expliquée par la vision du risque la plus récente.

Exemple Pour un contrat actif entre 2000 et 2010 nous avons 11 observations y_i^t, $t = 1...10$, mais un unique vecteur de covariables x_i^{2010}.

Périmètre d'étude de référence

- DROCs entre 2000 et 2005,

- 55 000 observations (Contrat-DROC),
- vision à 10 ans (on observe les sinistres pendant les 10 ans suivant l'ouverture des travaux),
- sinistres inférieurs à 50 k euros,
- nous excluons les très grandes entreprises.

Prise en compte de l'inflation

Puisque notre base de données comporte les montants de charges en euros de l'année de règlement, il convient d'ajuster tous les valeurs des charges en euros de 2015. Ceci nous permettra de comparer plus facilement les résultats obtenus dans les étapes suivantes.

L'indice d'inflation utilisé a été le BT01, qui est propre au secteur de la construction.

Voici l'ajustement réalisé :

$$S_{k,i} = \sum_{j=0}^{10} Y_{i,j}^{(k)} * \lambda_{i+j} \tag{9.1}$$

avec :

- $S_{k,i}$ les charges totales du Contrat k, DROC i,
- $Y_{i,j}^{(k)}$ les charges du contrat k concernant la DROC i et l'année de développement j,
- λ_i coefficient pour passer d'euros de l'année i en euros de 2015,
- λ_t un coefficient d'actualisation utilisé pour le passage d'euros de t en euros de 2015.

Nous simplifions la notation $S_{k,i}$ en S_k, car nous considérons que chaque année d'observation i correspond à un individu différent.

Exemple Prenons par exemple trois contrats, le contrat **1** présent dans le portefeuille pendant 7 ans, et les contrats **2** et **3** seulement présents pendant une année chacun.

Pour le contrat 1, l'assureur est engagé à payer tous les sinistres engendrés par les travaux ouverts entre 1999 et 2005. Cependant les garanties ne sont pas liées, et il est tout à fait possible de considérer séparément chaque année de DROC du contrat 1.

TABLE 9.1 – Exemple traitements charges contrat DROC

No Contrat	DROC	Charges	Exposition
1	1999	1000	1
1	2000	0	0.5
1	2001	0	1
1	2002	0	0.5
1	2003	0	1
1	2004	1000	1
1	2005	0	1
2	2000	0	1
3	2001	1	100

Pour le portefeuille ci-dessus, nous avons donc 9 lignes distinctes.

Remarque

- Lorsque la vision du risque ne change pas en fonction de la DROC, cette approche est équivalente à effectuer la régression sur le portefeuille suivant.

TABLE 9.2 – Exemple approche équivalente

No Contrat	Charges	Exposition
1	2000	6
2	0	1
3	100	1

Hypothèses

- Nous faisons l'hypothèse que les charges à l'ultime correspondent aux 10 premières années de développement.
- **L'exposition au risque**

  Quelle est l'exposition au risque en assurance construction ?

  Dans un contrat d'assurance ordinaire l'exposition au risque correspond à la fraction d'année pendant laquelle que le contrat a été en cours (années polices). Il est logique que la sinistralité d'un contrat sera d'autant plus élevée que le contrat aura une exposition plus longue.

Le même principe de proportionnalité entre exposition et sinistralité s'applique par exemple en assurance auto entreprises, où l'on considère l'exposition comme le nombre de voitures de la flotte assurée, pendant la période de couverture.

En assurance construction, l'on souscrit un contrat pour assurer tous les travaux effectués pendant l'année de paiement de la prime. Cependant l'assureur ne connaît pas le nombre exact de travaux. L'assureur devra donc se servir du chiffre d'affaires déclaré et des années polices pour estimer le nombre de travaux réalisés.

Nous avons ainsi deux choix possibles pour mesurer l'exposition au risque :

1. **exposition = années polices :** dans ce cas le chiffre d'affaires n'est qu'une variable explicative, et nous chercherons à estimer son influence.
2. **exposition=années polices * chiffre d'affaires**, dans ce cas on fait l'hypothèse que la sinistralité est proportionnelle au chiffre d'affaires, et nous modélisons un taux de prime applicable au chiffre d'affaires.

   En utilisant la variable offset = log(anp*chiffre d'affaires), nous modélisons donc directement un taux de chiffre d'affaires, tout en gardant les hypothèses du modèle Tweedie sur les charges. Et plus encore, nous supposons que le chiffre d'affaires évolue proportionnellement avec la sinistralité.

$$E[S_i|_{X_i,E_i}] = E[S_i|_{X_i,E_i=1}] * E_i \tag{9.2}$$

   Comme nous l'avons exprimé auparavant, les deux méthodes suivantes sont équivalentes :

   (a) Modèle 1

      i. variable réponse y_i charges totales

      ii. offset = années polices e_i

      iii. pas de poids *a priori* w_i

   (b) Modèle 2

      i. variable réponse $\frac{y_i}{e_i}$ charges totales

      ii. pas de variable offset

      iii. poids *a priori* $w_i' = w_i * e_i^{2-p}$

   Il serait donc théoriquement incorrect d'utiliser la méthode suivante, qui été choisie pour la modélisation :

   (c) Modèle 3 (modèle pas reproductible)

      i. variable réponse y_i/e_i charges totales

ii. pas d'offset

iii. pas poids *a priori*

Segmentation de la base

Afin de valider nos modèles nous divisons la base en deux :

- une base d'apprentissage ou "training data" sur laquelle nous ajusterons le modèle, pour cela nous prenons un échantillon de 80% des observations,
- une base de test qui nous servira pour voir la qualité de prédiction du modèle, pour cela nous prenons un échantillon de 20% des données.

Analyse préliminaire des charges

Nous possédons 90% des charges égales à 0. Les charges égales à 0 écartent toute de suite la possibilité de modéliser avec une loi continue.

Nous commençons par tracer l'histogramme des observations des charges totales, et aussi des charges strictement positives.

Notons s_i les observations de charges totales de la variable aléatoire S, nous présentons les quantiles empiriques $\hat{q}(\alpha) = inf\{x|\hat{F}_S(x) >= \alpha\}$, pour $\alpha \in \{0.05 * t\}_{t=1..20}$.

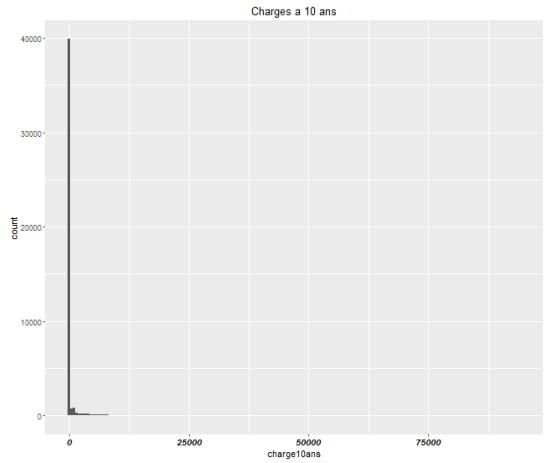


FIGURE 9.1 – Histogramme de la distribution des charges, 10 ans de survenance

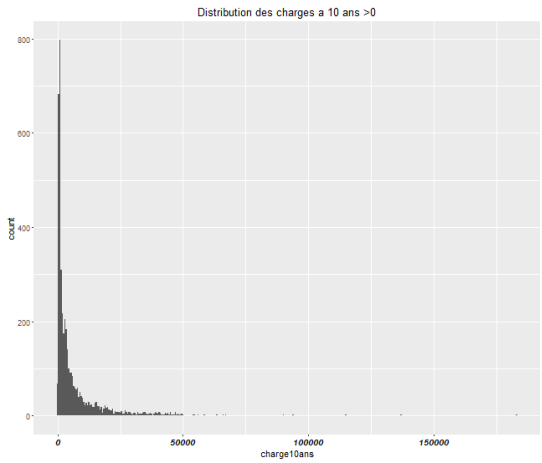


FIGURE 9.2 – Histogramme de la distribution des charges supérieures à 0, 10 ans de survenance

TABLE 9.3 – Quantiles de la distribution des charges totales

α	$\hat{q}(\alpha)$
0.00	0.00
0.05	0.00
0.10	0.00
0.15	0.00
0.20	0.00
0.25	0.00
0.30	0.00
0.35	0.00
0.40	0.00
0.45	0.00
0.50	0.00
0.55	0.00
0.60	0.00
0.65	0.00
0.70	0.00
0.75	0.00
0.80	0.00
0.85	0.00
0.90	0.00
0.95	2 170.55
1.00	182 847.27

Nous remarquons que cette distribution est très concentrée en 0, mais possède une composante continue par rapport à la mesure de Lebesgue.

D'après les résultats ci-dessus, il est raisonnable de penser que les réalisations des charges proviennent d'une loi de Poisson Composée.

9.2 Application : modélisation des charges par une loi de Poisson Composée CPG

La première étape de notre modélisation consiste à estimer la puissance p telle que $V(\mu) = \mu^p$.

Afin de comparer cette distribution à notre échantillon nous simulons 4 échantillons de loi $CPG(\alpha, \beta = 0.01, \lambda = 0.5)$ avec α prenant 4 valeurs différentes.

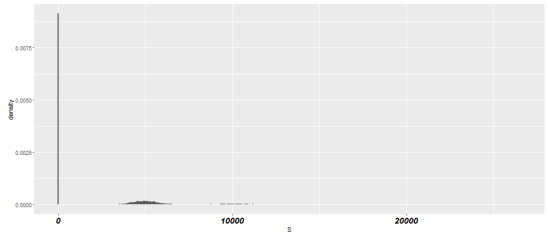


FIGURE 9.3 – histogramme simulations CPG α=50

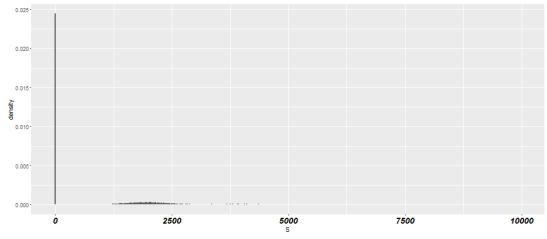
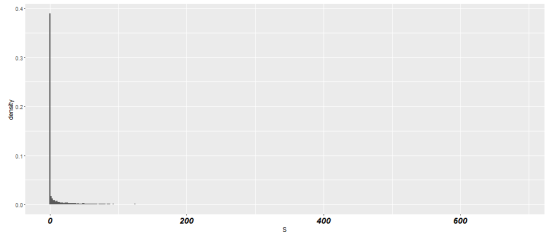


FIGURE 9.4 – histogramme simulations CPG α=20



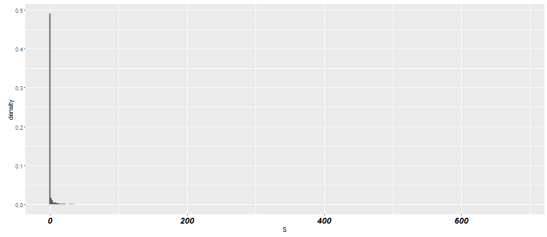


FIGURE 9.5 – histogramme simulations CPG α=0.1

Remarques

- Soit un modèle de Poisson Composé $CPG(\lambda, \alpha, \beta)$ la puissance p est donnée par $\frac{\alpha+2}{\alpha+1}$.
- La loi de Poisson et de Gamma correspondent aussi à des cas particuliers de la loi Tweedie avec p valant 1 et 2 respectivement.
- Puisque le paramètre de dispersion de la loi de Gamma est $\phi = \frac{1}{\alpha}$, le paramètre p de la loi de Tweedie est aussi un paramètre de dispersion indépendant des variables explicatives.

Estimation du paramètre de puissance p La fonction $tweedie.profile$ de r permet d'estimer aisément la puissance p.

Nous utilisons cette fonction, en incluant les années polices comme poids des observations.

L'optimisation de l'algorithme donne comme $p = 1.6$.

Nous essayons également d'effectuer une régression pour différentes valeurs de p et de calculer la déviance du modèle.

Ci-dessous, les différentes valeurs de la déviance en fonction du paramètre de puissance p.

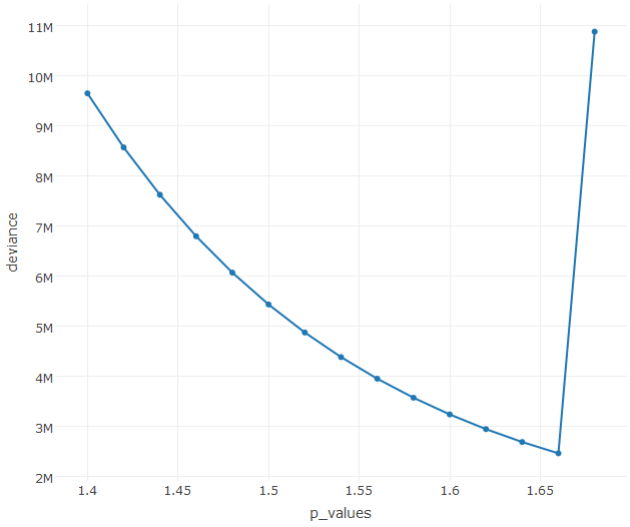


FIGURE 9.6 – Déviance du modèle en fonction de la puissance p

Cette fois ci la valeur optimale de p est 1.66. Nous conservons cette valeur pour la suite

Sélection des variables discriminantes

Les variables discriminantes issues de la sélection des variables sont les suivantes :

TABLE 9.4 – Variables discriminantes pour la modélisation

Variable	Description	Type de variable
ca_new	tranche de chiffre d'affaires	Classe tarifaire
notefi_nex	note financière	classe tarifaire
clasmaxrd	classe de risque	qualitative
natanteced	antécédents de sinistralité	qualitative
topqualif	qualification	indicatrice
anciente_ent_tranches	ancienneté de l'entreprise	classe tarifaire
efencent_tranche	effectif de l'entreprise, source INSEE	classe tarifaire
nbfam_30	nombre de familles de métiers	quantitative (nombre)
regimp_new	région d'implantation	qualitative
nbetact_tranche	nombre d'établissements actifs	classe tarifaire
exp_pro	expérience professionnelle	indicatrice

Nous avons effectué une sélection ascendante et une sélection descendante basées sur le critère AIC, le critère BIC s'avérant très pénalisant.

La fonction **glm** ne calcule pas directement l'AIC du modèle, nous recodons l'algorithme de sélection ascendante et descendante en utilisant la fonction **AICtweedie** de **R**.

Les deux sélections nous donnent les mêmes variables significatives.

TABLE 9.5 – Variables sélectionnées

	variables_AIC_forward
1	chiffres d'affaires
2	antécédents
3	classe de risque
4	note financière
5	effectif INSEE
6	expérience professionnelle
7	nombre de familles de métier
8	nombre d'établissements actifs

Modélisation des charges totales par le modèle de Poisson Composé

Le modèle final, après sélection des variables discriminantes est le suivant :

TABLE 9.6 – Modèle Poisson Composé

Composante du modèle	**Description**
Variable réponse	S_i, Charges totales pour chaque Contrat-DROC
fonction de lien	log
exposition	e_i= années polices pour chaque Contrat-DROC
offset	$log(e_i)$
poids *a priori*	1
distribution	$Tweedie(p \in (1,2))$ $p = 1.66$

Nous effectuons le test de rapport de vraisemblance (de type 3), en prenant comme test le test de Fisher, car le paramètre de dispersion ϕ est inconnu pour la loi de Poisson Composée.

La fonction *dropterm* de R conduit ce test en enlevant une à une les variables explicatives.

TABLE 9.7 – Test du rapport de vraisemblance

	Df	Deviance	F value	Pr(>F)
<none>		2481919.41		
Chiffre d'affaires	3	2531365.64	292.58	0.0000
antécédents_btplus	1	2494571.22	224.59	0.0000
note financière_new	4	2488101.75	27.44	0.0000
classe de risque_n_btplus	9	2498818.77	33.33	0.0000
effectif (INSEE)_tranche	4	2493021.56	49.27	0.0000
expérience professionnelle	2	2486278.18	38.69	0.0000
nombre d'établissements actifs	3	2487296.80	31.82	0.0000

D'après le test de du rapport de vraisemblance (en comparant à la loi de Fisher), toutes les variables sélectionnées sont significatives. Cependant le modèle n'est pas optimal car plusieurs modalités des variables ne sont pas significatives.

Nous effectuons le test de Wald pour fusionner plusieurs modalités non significatives. Nous utilisons la fonction *linearhypothesis* du "package" car.

- **Chiffre d'affaires** Toutes les variables sont significatives.
- **Note financière** *A priori*, on s'attend que les entreprises avec la meilleure santé financière soient les moins risquées, néanmoins les coefficients issus de la régression ne sont pas strictement décroissants.

  On verra plus tard que dans la modélisation d'un taux applicable au chiffre d'affaires, le "taux de sinistralité" décroit avec la note financière.

  Par ailleurs, seule une des modalités de la note financière est significative, nous testons le modèle emboité limité à deux modalités plus la modalité de référence avec

le test de Wald.

– $\mathcal{H}_0$ Les coefficients β_j des tranches "(4,8]=(8,12]" sont tous égaux.

	Res.Df	Df	Chisq	Pr(>Chisq)
1	44067			
2	44066	1	1.31	0.2518

Le test n'est pas significatif, nous décidons de conserver l'hypothèse nulle. Ainsi nous gardons deux 3 tranches de la note financière plus la modalité de référence.

Nous testons également le remplacement de la variable Note financière avec le découpage obtenu par l'arbre de régression dans la section 6.2.

Voici les résultants en incluant cette variable.

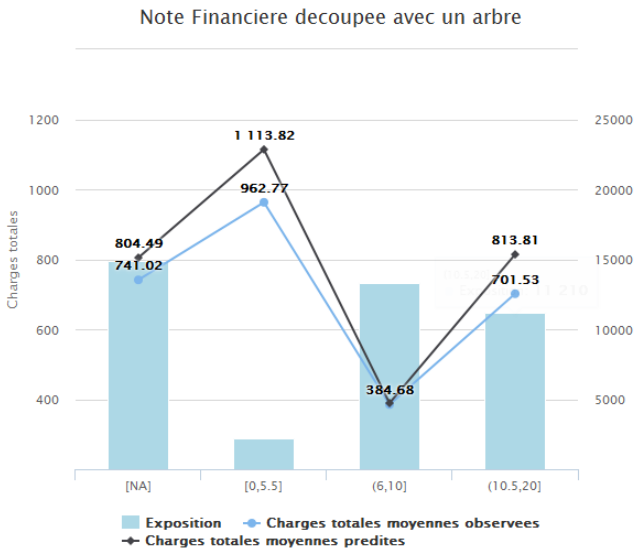


FIGURE 9.7 – Charges moyennes prédites et observées, note arbre de régression

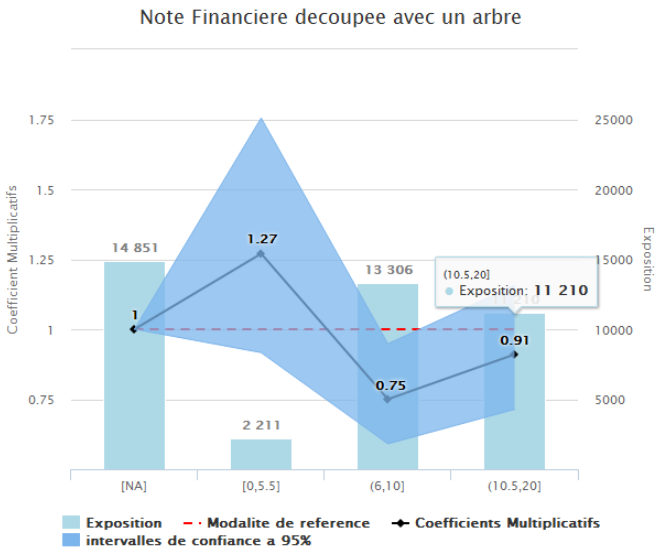


FIGURE 9.8 – Coefficients multiplicatifs de la variable note financière, découpée avec un arbre de régression

Cependant nous observons que le problème de cohérence de la variable **Note financière** persiste, nous essayons finalement d'ajouter une interaction entre le chiffre d'affaires et la note financière, mais seule une des modalités est significative.

Nous décidons ainsi de conserver seulement deux modalités pour cette variable.

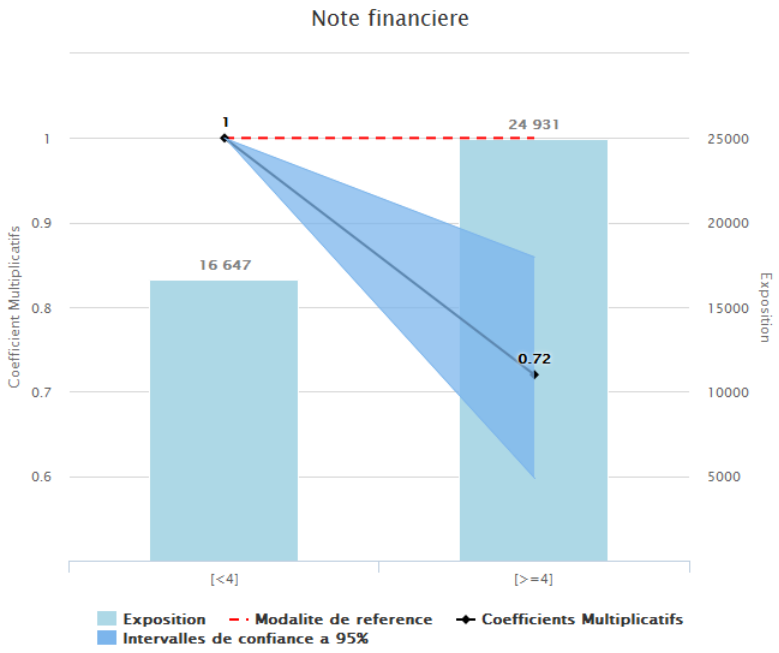


FIGURE 9.9 – Coefficients multiplicatifs finaux de la variable note financière

- **Classe de risque**

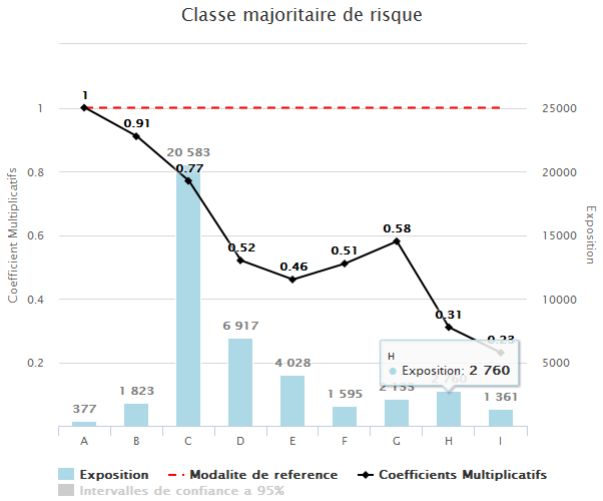


FIGURE 9.10 – coefficients multiplicatifs pour la classe de risque

Les modalités "B" "C" et "G" ne sont pas significativement différentes de la modalité de référence. Nous regroupons la modalité A et B et par soucis de cohérence du modèle nous regroupons les modalités E, F, G.

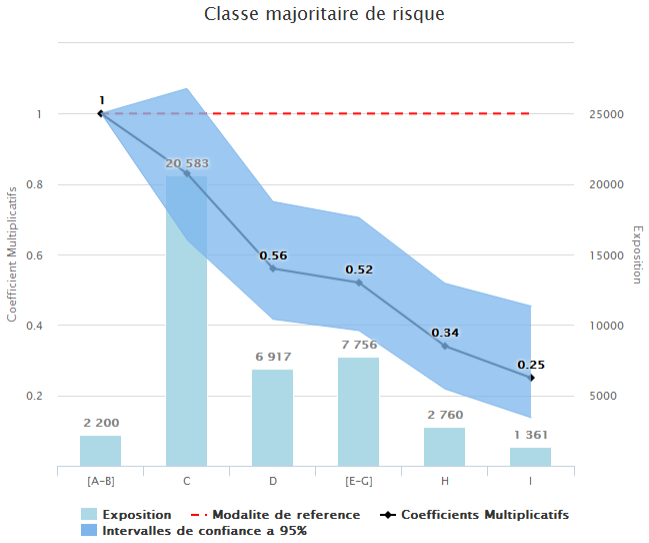


FIGURE 9.11 – Nouveaux coefficients multiplicatifs pour la classe de risque

- **Nature d'antécédents** Cette variable comporte une seule modalité qui s'avère significative :

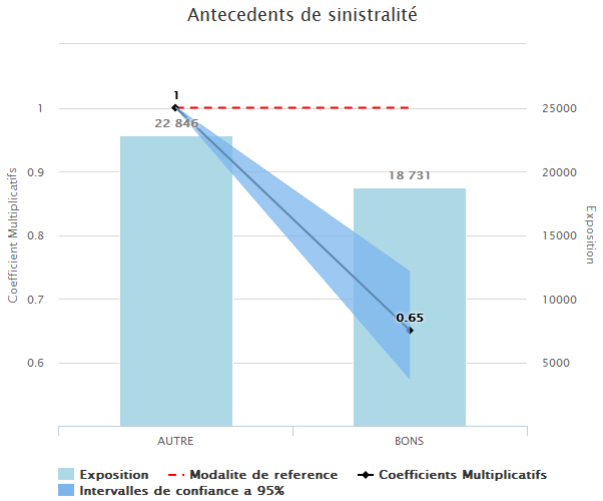


FIGURE 9.12 – Nouveaux coefficients multiplicatifs pour la classe de risque

- **Effectif INSEE** Nous regroupons les modalités 0-1, et nous mettons ensemble les valeurs manquantes avec la modalité "[>=10]", ce faisant, nous appliquons un coefficient de majoration pour les entreprises dont nous ne possédons pas cette information.

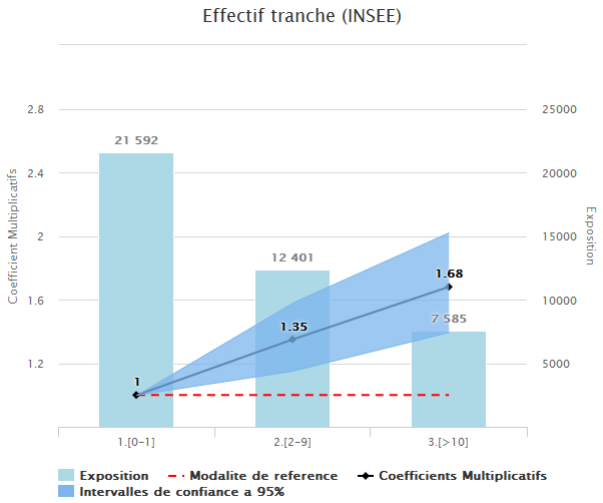


FIGURE 9.13 – Nouveaux coefficients multiplicatifs pour la classe de risque

- **Nombre d'établissements actifs** Les modalités 1 et 2 sont significativement différentes de la modalité de référence (0 établissements actifs), cependant par un souci de cohérence nous fusionnons les 3 premières modalités.

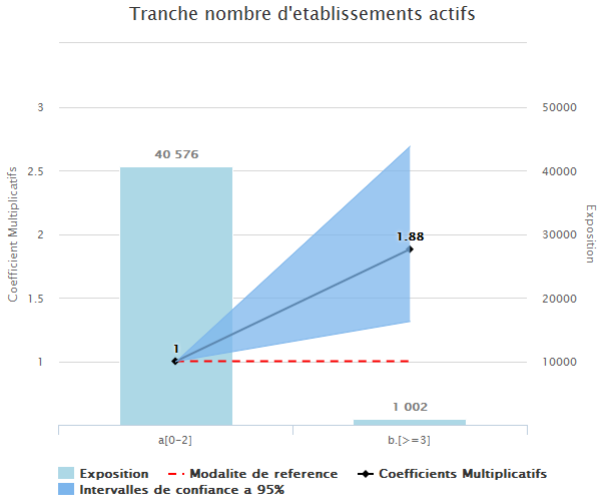


FIGURE 9.14 – Nouveaux coefficients Nombre d'établissements actifs

Validation du modèle

Une fois les modalités significatives choisies, nous passons à la validation du modèle.

Test d'adéquation globale

La déviance du modèle est de 2822081 avec 44102 degrés de liberté. La p-value du test d'adéquation globale est de 0, cela nous amène à rejeter l'adéquation du modèle.

Ce mauvais ajustement provient du fait que notre base de données contient plusieurs individus avec des variables réponse y_i distinctes, mais avec des covariances x_i égales.

Pour pallier cela, nous pourrions envisager de regrouper les observations d'un même contrat puisqu'ils ont la même espérance de sinistres, car les variables explicatives sont les mêmes.

La déviance du nouveau modèle est de 831000 avec 10357 dégrées de liberté, qui correspond encore à un ajustement de mauvaise qualité.

Remarquons encore, que ce premier modèle n'est pas reproductible, par conséquent les coefficients du modèle varient.

Illustration de la reproductibilité pour ce modèle

D'après les propriétés de la distribution de Tweedie exposées dans le chapitre précédent, le modèle de type 1 est équivalent à :

TABLE 9.8 – Modèle équivalent du Modèle de Poisson composé 1

Composante du modèle	**Description**
Variable réponse	S_i/e_i, Charges totales pour chaque Contrat-DROC
fonction de lien	log
exposition	e_i= années polices pour chaque contrat DROC
offset	$NULL$
poids *a priori*	e_i^{2-p}
distribution	$Tweedie(p \in (1,2))$ $p = 1.66$

Ainsi, lorsqu'on agrège deux observations possédant le même profil de risque nous avons :

$$\frac{Y_1 * w_1 + Y_2 * w_2}{w_1 + w_2} = \frac{\frac{S_1}{e_1} * e_1^{2-p} + \frac{S_2}{e_2} * e_2^{2-p}}{e_1^{2-p} + e_2^{2-p}} \neq \frac{\frac{S_1}{e_1} * e_1 + \frac{S_2}{e_2}}{e_1 + e_2} = \frac{S_1 + S_2}{e_1 + e_2} = Y_{1+2} \qquad (9.3)$$

- avec Y_i la variable réponse ;
- w_i le poids *a priori*.

En résultat, aucun des deux modèles (le modèle original et le reproductible) ne nous permettent pas de conclure à une bonne adéquation des données.

Néanmoins on peut toujours juger le **pouvoir explicatif** du modèle à travers des critères suivants.

1. La courbe de Lorenz
2. La "Lift Chart"
3. Les résidus de déviance

Courbes de Lorenz

En revanche ce modèle représente une bonne segmentation de la population (AUC=0.75). De plus il a l'avantage de ne pas conduire à un surapprentissage sur les données ; la capacité de prédiction sur la base de données de test est plus que satisfaisante.

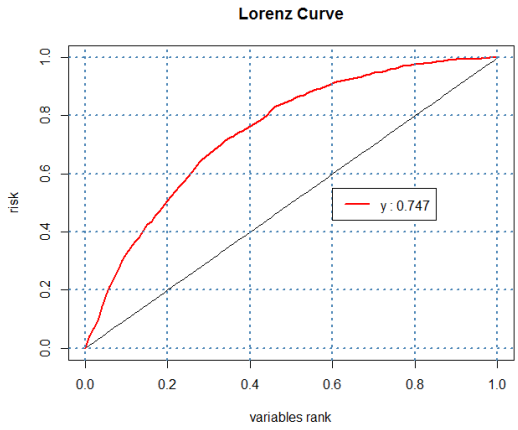


FIGURE 9.15 – Courbe de Lorenz sur la base d'apprentissage

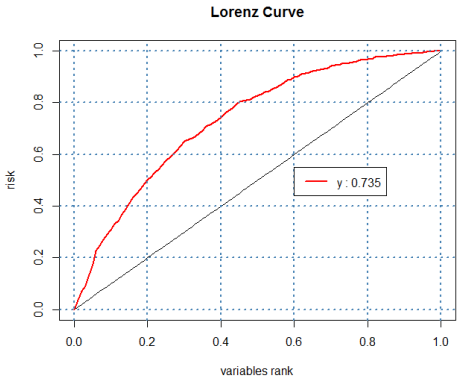


FIGURE 9.16 – Courbe de Lorenz sur la base de test

Les Liftcharts Les graphiques ci-dessous corroborent la qualité de segmentation du modèle, et valident également la capacité de prédiction du modèle.

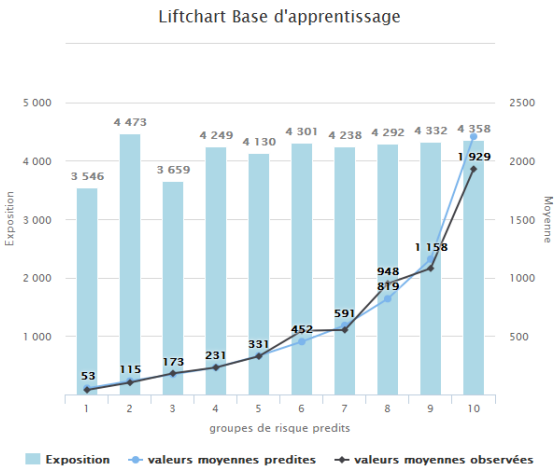


FIGURE 9.17 – Courbe de Lorenz sur la base d'apprentissage

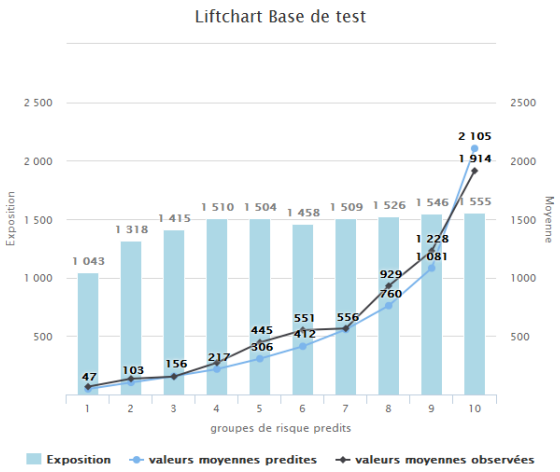


FIGURE 9.18 – Courbe de Lorenz sur la base de test

On conclut l'analyse de ce modèle par l'analyse des résidus.

Tout d'abord nous remarquons tout d'abord une forte densimétrie des résidus de déviance dans le graphique des qq-plot [1]. Cela est dû à la forte présence de contrats-DROCs non sinistrés dans notre base.

On observe aussi une surdispersion de nos données dans les box-plots croisés de la variable chiffre d'affaires et classe d'activité, mais qui n'indiquent cependant pas de tendance particulière.

1. Ce graphique compare les quantiles empiriques de la distribution avec ceux d'une loi normale

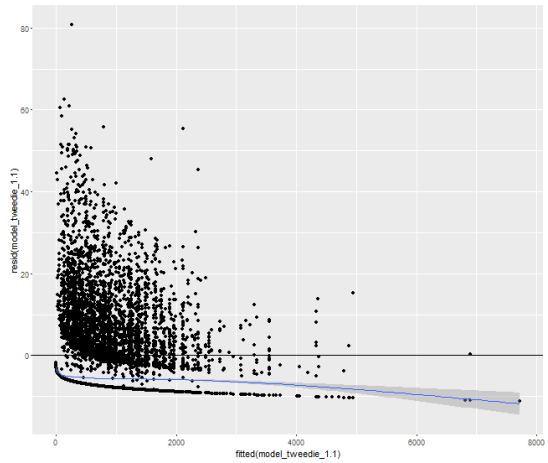


FIGURE 9.19 – Graphique des valeurs prédites versus résidus

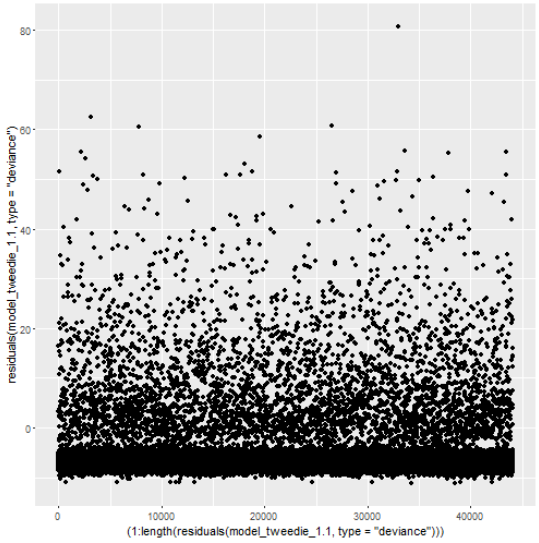


FIGURE 9.20 – Résidus de déviance, modèle de Tweedie

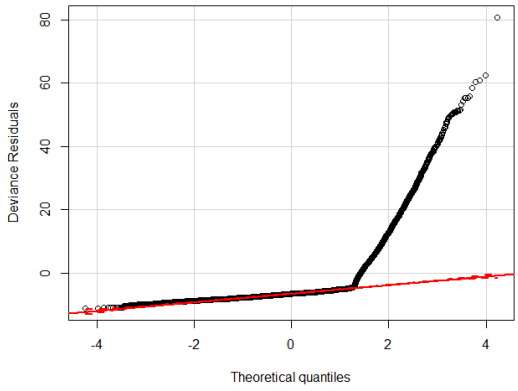


FIGURE 9.21 – Normal QQ-Plot, résidus de déviance Tweedie

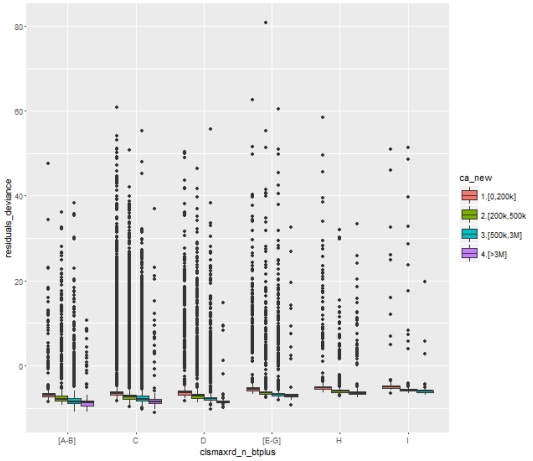


FIGURE 9.22 – Boxplot des résidus croisé entre les variables chiffre d'affaires et classe de risque, modèle de Poisson 1

Ainsi nous proposons les améliorations suivantes pour ce modèle.

9.3 Le Tweedie Boosted model

Rappelons la structure tarifaire derrière un GLM avec fonction de lien égale au logarithme :

$$log(E[Y_i|X_i = x_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j * x_{i,j} \tag{9.4}$$

Autrement dit le logarithme de la prime pure est une fonction linéaire des variables tarifaires. Cette approche consistant à faire l'hypothèse de relations linéaires peut s'avérer restrictive car l'on sait qu'en théorie on ne possède pas des fonctions linéaires.

Pour pallier cela, nous avons construit des tranches pour toutes les variables quantitatives ; certaines avec des arbres de régression et d'autres prédéterminées par les experts métier. Ainsi, certaines de ces tranches ont été déterminées *a priori*, et peuvent donc affecter le résultat de notre segmentation.

Une stratégie proposée par (Yi Yang, 2016) est de ne pas spécifier une structure log-linéaire, mais d'ajuster un GBM à nos données. Nous reprenons les notations introduites dans la section précédente.

$$log(E[Y_i|X_i = x_i]) = F(x_i) \tag{9.5}$$

Rappelons que le but de l'algorithme du Gradient Boosting est de minimiser une fonction de risque empirique, à partir d'une agrégation d'arbres de prédiction.

$$F(x) = arg \min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(y_i, f(x)) \tag{9.6}$$

La méthode du "Tweedie Boosted" consiste à spécifier la fonction de perte L comme moins la log-vraisemblance du modèle Tweedie. Et en remplaçant μ_i par $exp(F_{x_i})$, ce qui nous amène à conserver la structure multiplicative du modèle.

Pour effectuer cet algorithme nous utilisons le package TDboost de Yi Yang (2016). Nous utilisons 1000 arbres de régression.

Nous commençons par analyser les variables les plus significatives et leur importance. Cette importance est liée à la réduction obtenue sur la fonction de perte apportée par la variable en question.

TABLE 9.9 – Variables significatives Modèle Tweedie Boosted

var	Variable	Importance, base 100
1	Commune d'établissement	90%
2	Famille d'activités	5%
3	activités (700)	3%
4	Chiffre d'affaires	1%
5	Classe de risque	1%

Nous observons une la variable commune d'établissement est la variable la plus explicative pour le modèle de Boosted Tweedie.

Regardons les valeurs prédites pour chaque commune :

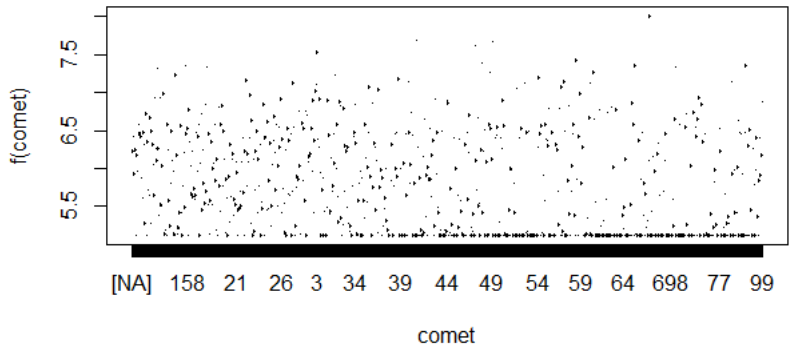


FIGURE 9.23 – Valeurs moyennes prédites pour chaque commune, Modèle Tweedie Boosted

On observe clairement un effet de sur-apprentissage, puisqu'il y a une estimation différente pour chaque commune. En l'état actuel, ce modèle est inutilisable. Nous excluons donc cette variable.

En excluant la commune d'établissement, les variables le plus significatives sont les suivantes ;

TABLE 9.10 – Nouvelles variables significatives, modèle Tweedie Boosted

	Variable	Importance
1	famille d'activités (30)	41%
2	classe de risque	29%
3	Chiffre d'affaires	24%

Nous donnons les valeurs prédites moyennes pour chacune des variables :

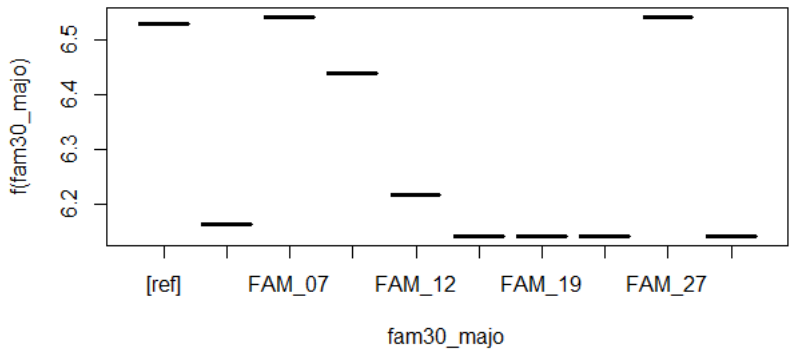


FIGURE 9.24 – Valeurs moyennes prédites pour chaque commune, Modèle Tweedie Boosted

Bien que nous avions écarté cette variable dans un premier temps car elle correspond à un regroupement plus fin des activités que la variable classe de risque, l'algorithme GBM est sensé de pouvoir tenir compte de toutes ces interactions. Les résultats du modèle nous donnent un aperçu des familles d'activités le plus risqués.

Cependant ce regroupement n'est pas à intégrer dans une tarification future, car l'on sait qu'il n'est pas assez robuste.

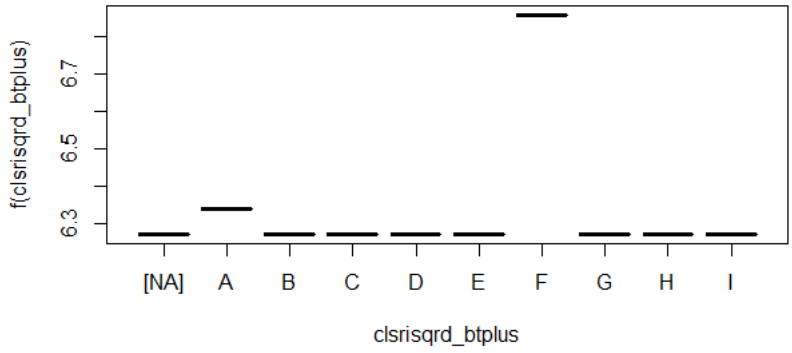


FIGURE 9.25 – Valeurs moyennes prédites pour chaque commune, Modèle Tweedie Boosted

La variable classe de risque présente des tendances anormales, avec une classe F plus risquée que les autres.

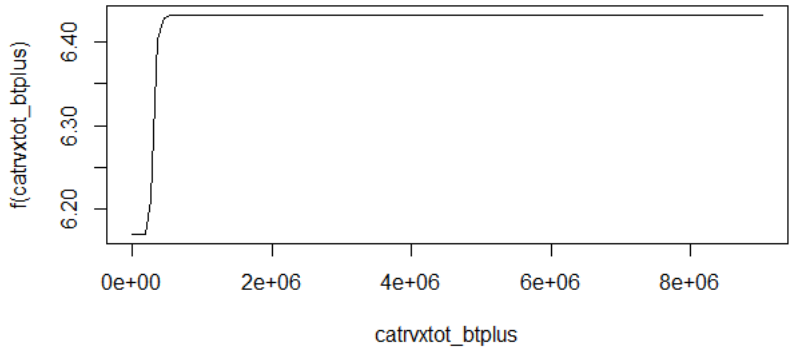


FIGURE 9.26 – Valeurs moyennes prédites pour chaque commune, Modèle Tweedie Boosted

Finalement la variable chiffre d'affaires, sur laquelle nous possédons un intérêt particulier, affiche une allure exponentielle, qui peut effectivement se traduire par une log-linéarité des charges par rapport au chiffre d'affaires.

Nous ne nous attarderons pas plus sur ce modèle car d'un point de vue tarifaire, il n'est pas optimal.

De plus nous sommes conscients de la technicité nécessaire pour atteindre un modèle plus adéquat, et cela n'est pas l'objectif principal de ce mémoire. Néanmoins, les résultats de ce modèle nous laissent penser que le chiffre d'affaires peut être bien une mesure de l'exposition au risque.

Chapitre 10

Application : Modélisation d'un taux applicable au chiffre d'affaires

En absence d'information sur le nombre des travaux effectués par chaque compagnie, il convient de modéliser, non pas un montant de prime pure, mais un taux applicable au chiffre d'affaires.

Le modèle utilisé a été le suivant :

TABLE 10.1 – Modèle 1.2 Modèle Poisson Composé, taux applicable au CA

Composante du modèle	**Description**
Variable réponse	S_i/e_i, Charges totales pour chaque contrat DROC
fonction de lien	*log*
exposition	e_i=anp*Chiffre DROC
offset	Null
poids *a priori*	1
distribution	$Tweedie(p \in (1,2))$ $p = 1.66$

Tout d'abord nous estimons la puissance de la loi Tweedie, avec la même méthode utilisée précédemment. Nous obtenons une puissance $p = 1.69$.

Quant à la sélection des variables, nous testons plusieurs combinaisons de variables, mais seules les variables suivantes permettent l'algorithme d'optimisation de converger.

TABLE 10.2 – Variables discriminantes pour la modélisation d'un taux applicable au CA

Variable	Description	Type de variable
ca_new	Tranche de chiffre d'affaires	Classe tarifaire
clasmaxrd	Classe de risque	qualitative
notefi_new	Note financière	classe tarifaire
natanteced	antécédents de sinistralité	qualitative
efencent_tranche	effectif de l'entreprise, source INSEE	classe tarifaire
regimp_new	region d'implantation	qualitative
nb_fam_30_tranches	nb de familles d'activités	catégorielle
anciennete.entreprise	ancienneté de l'entreprise	catégorielle
nbetact_tranche	nombre d'etablissements actifs	classe tarifaire

Validation du modèle

Nous testons la significativité des variables avec le test du rapport de vraisemblance.

Toutes les variables sont significatives avec une p-value inférieure à 0.001. Nous passons à l'analyse de cohérence des variables.

Nous commençons par exclure le nombre d'établissements actifs et l'effectif de l'entreprise (source INSEE), ces deux variables sont non seulement corrélées linéairement avec le chiffre d'affaires (d'après l'ACP), mais sont également incohérentes d'un point de vue tarifaire (les entreprises avec plus d'établissements paient moins cher).

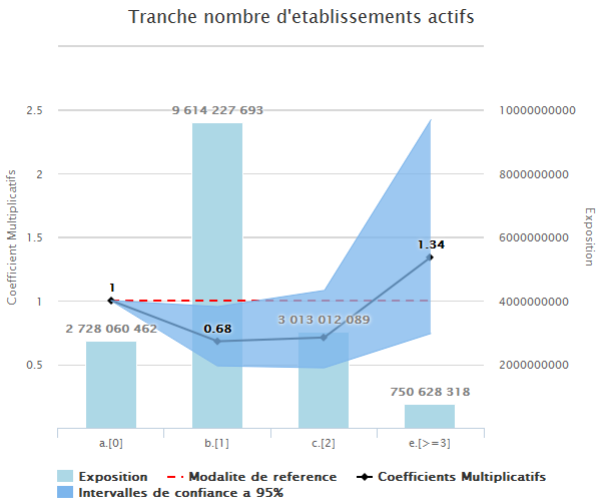


FIGURE 10.1 – Coefficients nombre d'établissements, modélisation du taux

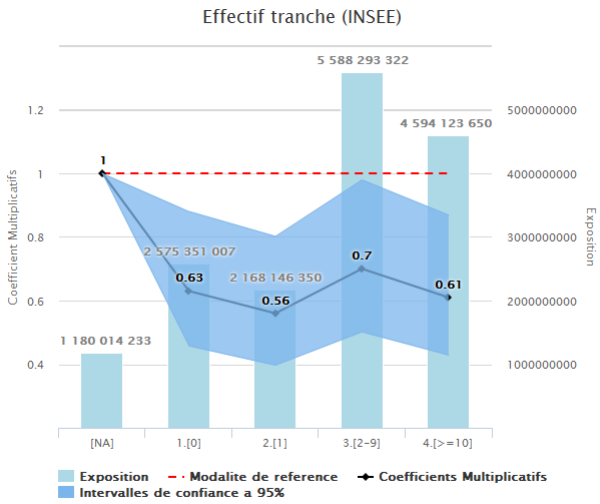


FIGURE 10.2 – Coefficients effectif, modélisation du taux

Pour les autres variables nous regroupons les modalités non significatives en utilisant le test de Wald.

Nous comparerons ces résultats par périmètre dans la partie suivante.

Validation du modèle

Nous donnons ci-dessus les courbes de Lorenz et les "Lift-charts" pour l'échantillon test et l'échantillon d'apprentissage. Nous conservons un bon pouvoir de segmentation dans le nouveau modèle, néanmoins l'adéquation vis à vis des données ne semble pas optimale puisqu'on a une déviance trop importante. Le test d'adéquation globale est rejeté avec une p-value égale à 0.

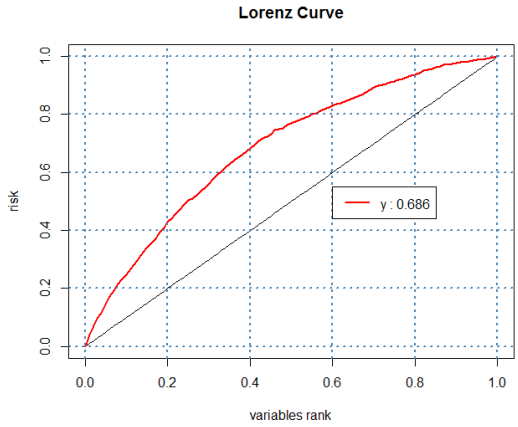


FIGURE 10.3 – Courbe de Lorenz sur la base d'apprentissage

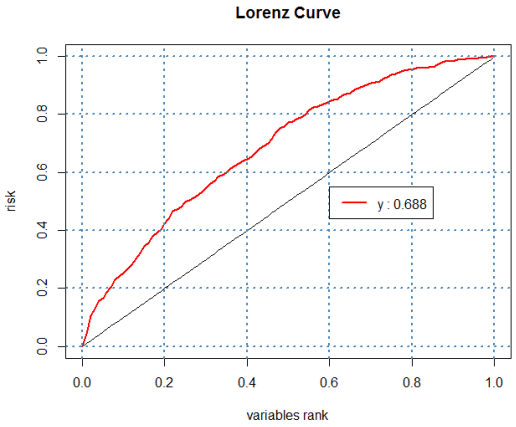


FIGURE 10.4 – Courbe de Lorenz sur la base de test

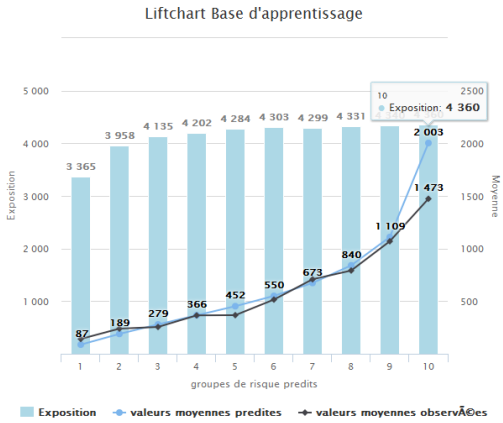


FIGURE 10.5 – "Liftchart" sur la base d'apprentissage

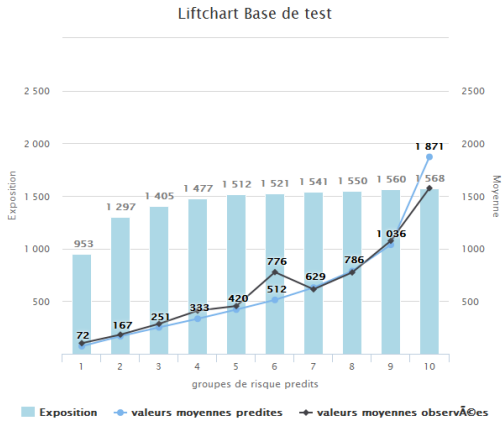


FIGURE 10.6 – "Liftchart" sur la base de test

A la fois les courbes de Lorenz et les Liftcharts nous indiquent que le modèle permet une bonne segmentation de la population entre individus risqués et non risqués.

Critiques du modèle

Le modèle présenté ci-dessus attribue le même poids aux grandes entreprises qu'aux petites entreprises, il modélise en quelque sorte une moyenne des taux et non pas un taux moyen.

Pour rendre ce modèle reproductible, il faudrait ajouter des poids $w_i = e_i$ dans la régression.

Cependant ce modèle pénalise trop fortement les grandes entreprises au profit des petites entreprises, qui dans certains déclarent un chiffre d'affaires très faible [1].

Cette tendance est davantage prononcée pour les classes d'activité les moins risquées d'après le modèle de charges de Tweedie et nous observons un taux moyen qui ne décroît pas avec les classes d'activité.

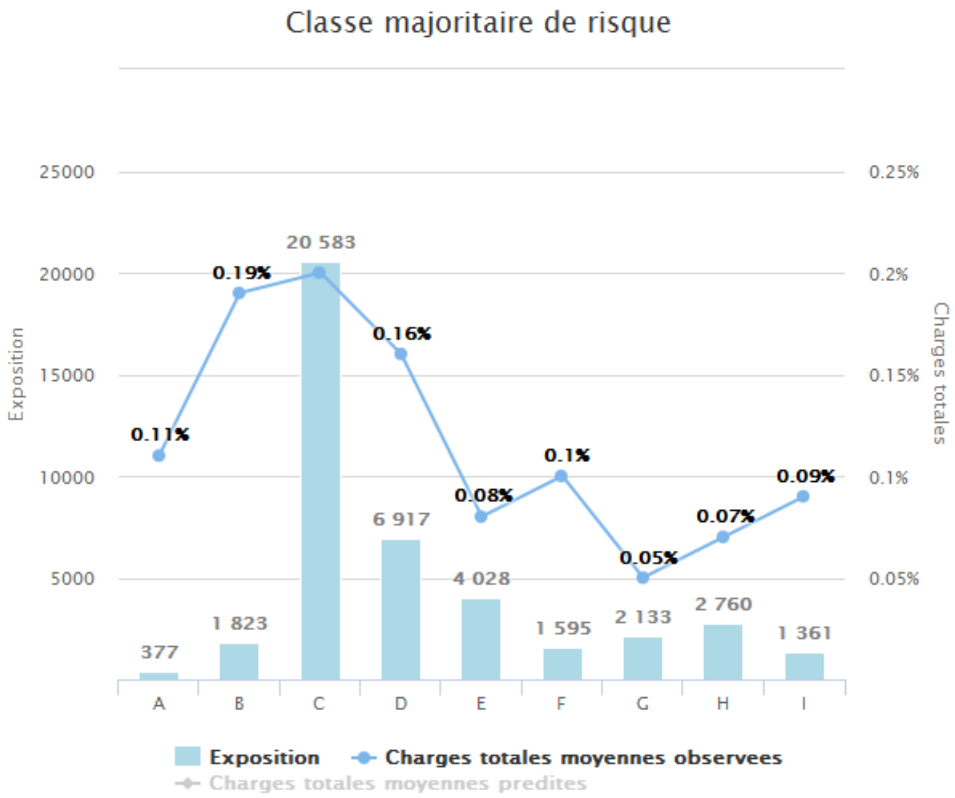


FIGURE 10.7 – Taux moyen de sinistralité par classe d'activité

On décide donc de conserver notre modèle initial qui sera l'?objet d'?une étude plus approfondie dans la section suivante.

1. Nous observons des entreprises déclarant moins de 1 000 comme chiffre d'affaires

Chapitre 11

Application : Modélisation de la fréquence

11.1 Analyse préliminaire de la distribution

Sur le périmètre étudié (DROCS entre 2000 et 2005, survenance à 10 ans), nous observons une forte proportion de zéros sur les contrats (90% des contrats non sinistrés).

On observe par ailleurs une surdispersion sur notre échantillon ($\overline{x} = 0.13$) $<$ ($\hat{\sigma}^2 = 0.3$), néanmoins nous considérerons toutefois une modélisation de poisson dans un premier temps.

L'estimateur de vraisemblance pour le paramètre λ coïncide avec la moyenne empirique. Nous donnons également la fréquence empirique et la fréquence théorique ($Nombredecontrats*$ $P(N = k|\lambda = 0.13)$ ci-dessous.

11.2 Modélisation de la fréquence par un GLM, avec distribution de Poisson

TABLE 11.1 – Modèle 2.1 Modèle de fréquence, distribution de Poisson

Composante du modèle	**Description**
Variable réponse	N_i Nombre de sinistres par contrat DROC
fonction de lien	log
exposition	e_i= années polices
offset	$log(e_i)$
poids *a priori*	1
distribution	$Poisson$

	Nombre	observés	estimés
1	0.00	44841.00	43195.19
2	1.00	3577.00	5986.18
3	2.00	762.00	414.80
4	3.00	264.00	19.16
5	4.00	84.00	0.66
6	5.00	37.00	0.02
7	6.00	21.00	0.00
8	7.00	4.00	0.00
9	8.00	10.00	0.00
10	9.00	2.00	0.00
11	10.00	3.00	0.00
12	11.00	3.00	0.00
13	12.00	2.00	0.00
14	13.00	1.00	0.00
15	14.00	0.00	0.00
16	15.00	0.00	0.00
17	16.00	0.00	0.00
18	17.00	0.00	0.00
19	18.00	1.00	0.00
20	19.00	2.00	0.00
21	20.00	0.00	0.00
22	21.00	0.00	0.00
23	22.00	0.00	0.00
24	23.00	0.00	0.00
25	24.00	0.00	0.00
26	25.00	0.00	0.00
27	26.00	0.00	0.00
28	27.00	2.00	0.00

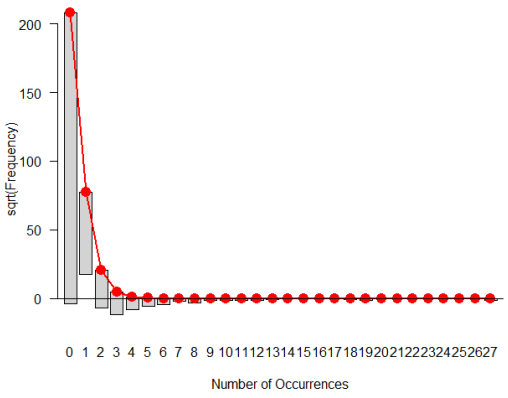


FIGURE 11.1 – Fréquence empirique versus fréquence théorique selon la loi de Poisson

Sélection des variables

Nous procédons de la même manière que lors de la modélisation des charges totales. La sélection automatique ascendante et descendante à partir du critère AIC coïncident, ce qui nous amène à conserver toutes les variables explicatives.

On effectue le test de rapport de vraisemblance pour la significativité des variables ; toutes les variables sont significatives d'après le test. Cependant ce modèle dans cet état n'est pas optimal car il comporte plusieurs modalités qui ne sont pas significativement différentes de la modalité de référence.

Table 11.2 – Test de rapport de vraisemblance pour les variables discriminantes, modèle de Poisson pour la fréquence

	Df	Deviance	AIC	LRT	Pr(>Chi)
<none>		29561.39	40226.43		
Chiffre d'affaires	3	30256.62	40915.66	695.23	0.0000
Antécédents	1	29945.47	40608.50	384.08	0.0000
Effectif INSEE	4	29813.12	40470.16	251.73	0.0000
Note financière	4	29635.98	40293.01	74.58	0.0000
Classe d'activité	8	29759.71	40408.74	198.32	0.0000
Région d'implantation	13	29738.07	40377.10	176.68	0.0000
Expérience professionnelle	2	29630.69	40291.72	69.30	0.0000
Nombre d'établissements actifs	3	29592.45	40251.49	31.06	0.0000
Ancienneté de l'entreprise	4	29574.72	40231.75	13.32	0.0098
nb_fam30_tranches	1	29565.97	40229.01	4.58	0.0323

Remarque Les variables nombre d'établissements actifs et ancienneté ont été exclues du modèle car elles ne pressentaient pas de tendance particulière.

Fusion des modalités non significatives

Nous effectuons systématiquement le test de Wald avec la commande *linearhypothesis* de R pour tester l'égalité multiple des coefficients. Après traitement, le modèle finale retenu pour la modélisation de la fréquence est le suivant :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0848	0.0814	-13.33	0.0000
Chiffre d'affaires [200k,500k	0.6125	0.0315	19.47	0.0000
Chiffre d'affaires[>500k]	1.0209	0.0337	30.25	0.0000
Chiffre d'affaires[>=4]	-0.3411	0.0253	-13.50	0.0000
Classe B	-0.5346	0.0811	-6.59	0.0000
Classe C	-0.6406	0.0699	-9.16	0.0000
Classe D	-0.9623	0.0747	-12.87	0.0000
Classe E-G	-1.1379	0.0789	-14.42	0.0000
Classe H	-1.6257	0.1032	-15.75	0.0000
Classe I	-2.1694	0.1664	-13.04	0.0000
Antécédents bons	-0.5335	0.0261	-20.44	0.0000
Effectif [>=10]	0.5346	0.0351	15.25	0.0000
Région [11 ;32]	-0.2041	0.0422	-4.83	0.0000
Région [52 ;53 ;75 ;84]	0.2013	0.0279	7.21	0.0000
Région	0.3559	0.0387	9.19	0.0000
Pas d'expérience	-0.2942	0.0374	-7.87	0.0000
Expérience pro	-0.5436	0.0651	-8.35	0.0000

Ajustement du modèle

La déviance du modèle est égale à 29982 avec 49595 degrés de liberté. Notre base comporte, sur ce périmètre, environ 50 000 observations, ce qui nous permet de mener le test d'adéquation globale du χ^2.

La p-value du test est égale à 1 ; l'hypothèse $\mathcal{H}_0$ ne peut être rejetée et on peut donc considérer que le modèle est adéquat.

Analyse des résidus

Nous traçons pour ce modèle 4 graphiques d'analyse des résidus :

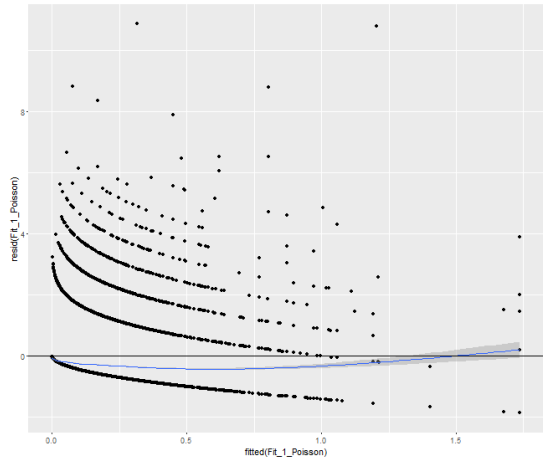


FIGURE 11.2 – Graphique des valeurs prédites versus résidus

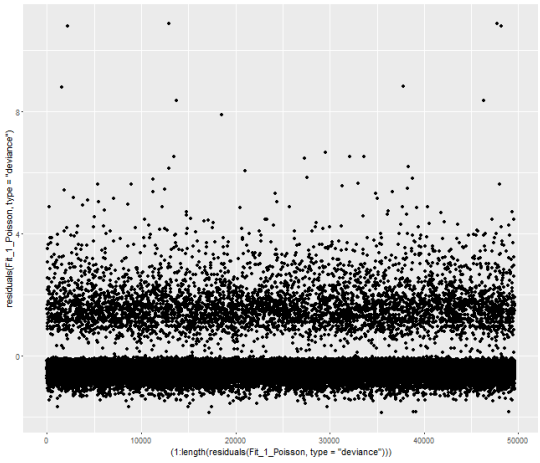


FIGURE 11.3 – Résidus de déviance, modèle de Poisson

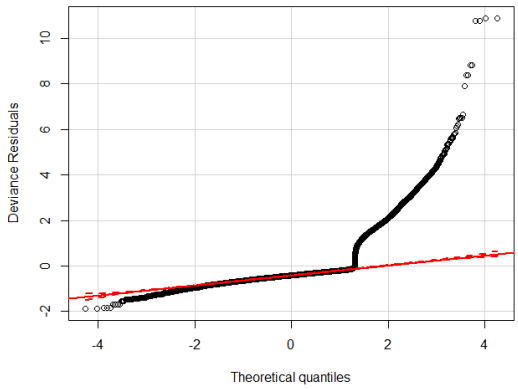


FIGURE 11.4 – Normal QQ-Plot, résidus de déviance de Poisson

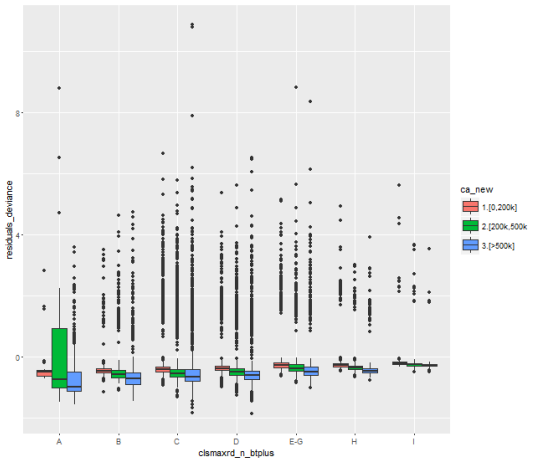


FIGURE 11.5 – Boxplot des résidus croisé entre les variables chiffre d'affaires et classe de risque, modèle de Poisson 1

Nous sommes en présence d'une forte dissymétrie, provenant de la forte proportion de zéros dans notre échantillon. Les résidus de déviance ne sont vraisemblablement pas normaux, puisqu'ils sont fortement asymétriques.

Néanmoins nous constatons qu'il n'y a pas une tendance des résidus sur l'espérance ni sur la variance. La courbe de "Loess" sur le premier graphique nous indique également qu'il

n'y a pas de lien entre les résidus et les valeurs prédites ; de ce point de vue le modèle est satisfaisant.

Analyse du pouvoir de segmentation

Nous présentons à nouveau les courbes de Lorenz et la "Liftchart" pour ce modèle.

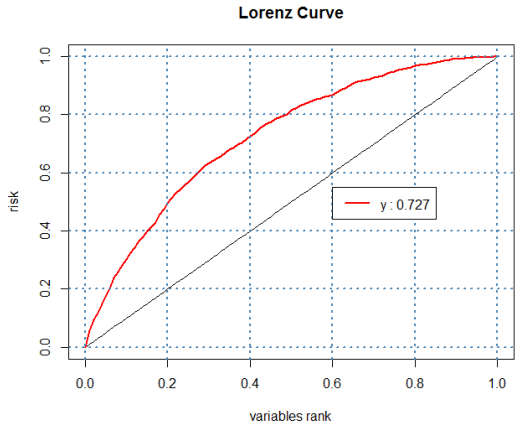


FIGURE 11.6 – Courbe de Lorenz, base d'apprentissage, modèle de Poisson

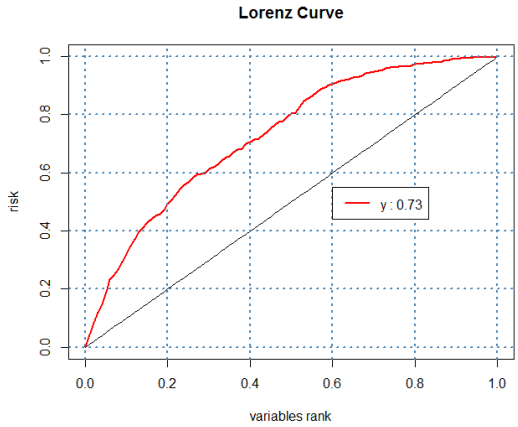


FIGURE 11.7 – Courbe de Lorenz, base de test, modèle de Poisson

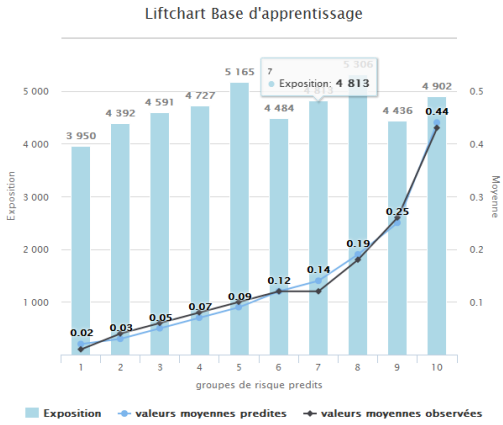


FIGURE 11.8 – Liftchart, base d'apprentissage, modèle de Poisson

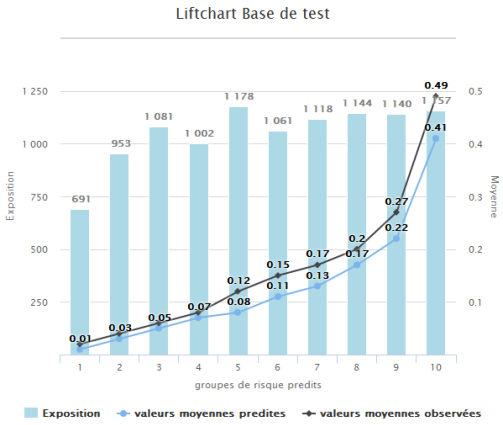


FIGURE 11.9 – Liftchart, base de test, modèle de Poisson

Ce modèle est, comme le modèle de charges totales, très satisfaisant du point de vue de la segmentation tarifaire. Il possède de même un très bon pouvoir de prédiction qu'on peut constater sur la courbe de Lorenz appliqué sur l'échantillon test.

Néanmoins l'asymétrie des résidus et la surdispersion des observations nous amènent à considérer d'autres modèles de fréquence.

11.3 Modélisation de la fréquence par la loi Binomiale négative

Tenant compte des défauts que peut avoir le modèle précédent, nous introduisons un modèle avec distribution binomiale négative avec paramètre de dispersion θ inconnu.

Nous utilisons la fonction *glm.nb* de R, qui estime automatiquement le paramètre de surdispersion θ.

TABLE 11.3 – Modèle de fréquence, loi binomiale négative

Composante du modèle	**Description**
Variable réponse	N_i, Nombre de sinistres par contrat-DROC
fonction de lien	log
exposition	e_i=anp
offset	$log(e_i)$
poids *a priori*	1
distribution	Binomiale négative, paramètre θ inconnu

Nous utilisons les mêmes variables explicatives que dans le modèle précédent afin de rendre facile une comparaison entre les modèles.

Cette méthode nous permet de réduire considérablement la déviance du modèle à 19473 avec 49615 degrés de liberté. Le test d'adéquation globale, n'est pas rejeté (p-value=1) nous pouvons conserver l'hypothèse que le modèle représente un ajustement adéquat.

Le paramètre de dispersion θ es estimé à 0.28, et la vraisemblance est égale à 18562.18. Nous constatons une amélioration nette de la qualité de l'ajustement face au modèle de Poisson.

Finalement le nombre de zéros prédit par le modèle s'approche plus des observations.

	Nombre de zéros prédits
obs	44841
NB	44855
ML_poisson	43533

Finalement nous présentons les résidus du modèle :

Résidus du modèle

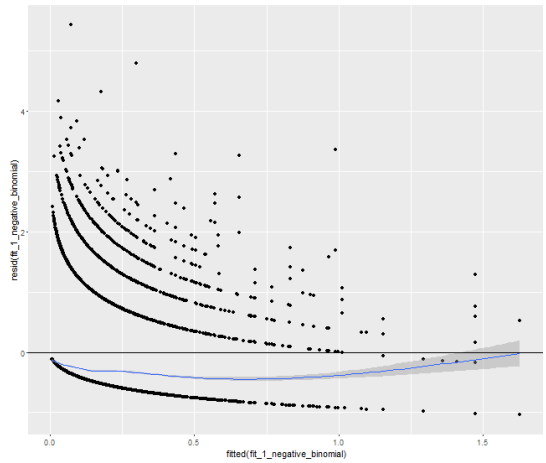


FIGURE 11.10 – Graphique des valeurs prédites versus résidus de déviance, modèle de fréquence NB

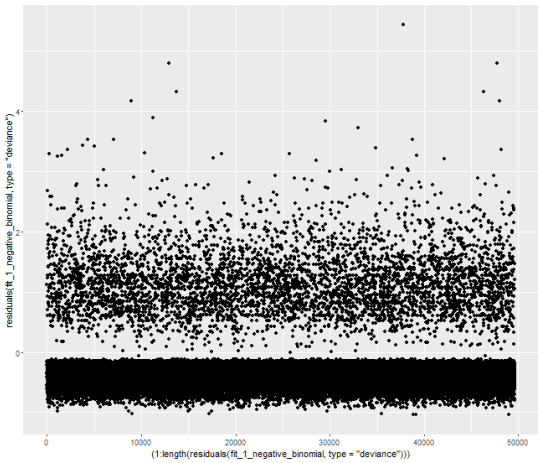


FIGURE 11.11 – Résidus de déviance, modèle de fréquence, BN

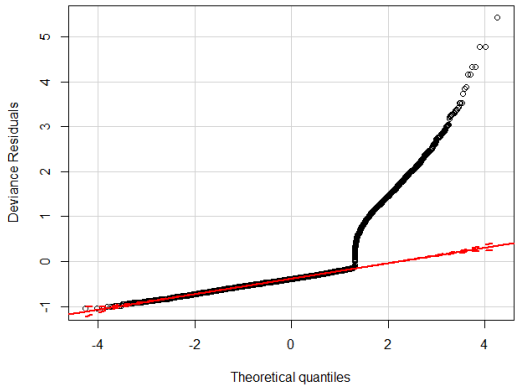


FIGURE 11.12 – qqplot, modèle de fréquence, BN

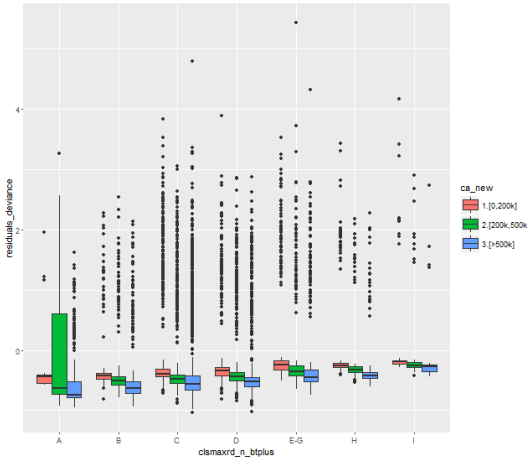


FIGURE 11.13 – boxplot, modèle de fréquence, BN

On constate toujours une surdispersion des résidus, avec une légère amélioration de la symétrie par rapport au modèle précédent.

Néanmoins, il est envisageable de proposer une amélioration du modèle en utilisant les modèles dits "Zero-inflated".

11.4 Modélisation de la fréquence avec des modèles "Zero-Inflated

Dans cette partie nous chercherons à modéliser l'excès de zéros dans la fréquence des sinistres. Cet excès peut provenir du fait que les sinistres ne sont pas déclarés quand ils n'excèdent pas un montant. En effet le coût des démarches administratives, ou même les franchises, peuvent inciter à ne pas déclarer.

La variable à modéliser est donc $N_i = N_i^* B_i$, avec $N_i^* \sim \lambda_\rangle$, $N_i \sim Ber(1 - w_i)$. Ni N_i^* ni B_i sont observables, donc nous ne pouvons pas effectuer une régression séparée pour les deux variables réponses.

Rappelons la loi de probabilité d'une telle variable :

$$P(N_i = k_i) = \begin{cases} w_i + (1 - w_i) * exp(-\lambda_i) & k_i = 0 \\ (1 - w_i) exp(-\lambda_i) \frac{\lambda_i^l}{k_i!} & k_i > 0 \end{cases}$$

Nous nous plaçons dans le modèle dit $ZIP(\tau)$ qui consiste à effectuer la régression logistique sur un sous-ensemble des variables utilisées pour la modélisation de Poisson.

Avec $logit(w_i) = \tau X\beta$, $log(\lambda) = X\beta$, et τ un vecteur des zéros et de 1.

Le modèle est le suivant :

TABLE 11.4 – Modèle de Fréquence, Poisson "Zero-Inflated"

Composante du modèle	**Description**
Variable réponse	N_i,
fonction de lien	*log* pour Poisson *logit* pour w_i
exposition	e_i=anp
offset	$log(e_i)$
poids *a priori*	1
distribution	ZIP

La log-vraisemblance du modèle est égale à -18747.97 avec 42 degrés de liberté ; en se basant sur ce critère il représente déjà un meilleur ajustement que le modèle de Poisson.

Néanmoins la difficulté du modèle repose sur la sélection des variables discriminantes pour chacune des régressions. Une idée a été de modéliser l'apparition d'un sinistre $B = 1_{N_i>0}$ par une régression logistique sur les variables explicatives.

TABLE 11.5 – Modèle, probabilité d'avoir un sinistre loi de Bernoulli

Composante du modèle	**Description**
Variable réponse	B_i,
fonction de lien	$logit$
exposition	e_i=anp
offset	$log(e_i)$
poids *a priori*	1
distribution	Binomiale(n=1)

Néanmoins toutes les variables sont significatives, ce qui nous empêche de réduire le modèle. Ci-dessous le test du rapport de vraisemblance du χ^2 car le paramètre de dispersion es connu pour la loi binomiale.

TABLE 11.6 – Test du rapport de vraisemblance, significativité des variables dans le modèle logistique

	Df	Deviance	AIC	LRT	Pr(Chi)
<none>		36719.78	36761.78		
chiffre d'affaires	2	37422.52	37460.52	702.73	0.00
note financière	1	36820.17	36860.17	100.39	0.00
classe de risque	6	37217.28	37247.28	497.49	0.00
antécédents	1	37077.09	37117.09	357.31	0.00
effectif INSEE	1	36782.96	36822.96	63.17	0.00
région d'implantation	3	36796.82	36832.82	77.03	0.00
expérience professionnelle	2	36769.93	36807.93	50.14	0.00

Cependant, nous pouvons donner un certain ordre d'importance aux variables en regardant la variation du critère AIC ; la suppression des variables chiffre d'affaires et classe de risque conduit à la plus forte augmentation de l ?AIC. Nous testerons donc le modèle emboité qui limite la régression logistique à ces deux variables.

- $\mathcal{M}_0 = log(\mu) = X\beta, logit(w) = X * \beta$
- $\mathcal{M}_1 = log(\mu) = X\beta, logit(w) = X\gamma$ avec $gamma$ prenant uniquement les paramètres des variables chiffre d'affaires et classe de risque.

Le test de Wald est significatif, les différences du modèle emboité et du modèle de base sont significatives. Il en est de même lorsqu'on ne considère pas toutes les variables explicatives pour la régression logistique.

Nous conservons donc ce modèle, qu'on comparera aux autres modèles de fréquence dans la section suivante.

TABLE 11.7 – Test de Wald modèles emboités

	Res.Df	Df	Chisq	Pr(>Chisq)
1	49586			
2	49574	12	124.87	0.0000

Modélisation de la fréquence par une loi Binomiale négative "Zero Inflated"

Le cadre probabiliste est très similaire au cas de Poisson Composé, nous considérons à nouveau : $N_i = N_i^* B_i$, avec $N_i^* \sim \mathcal{NB}(\lambda, \theta$ avec $\theta inconnu$, $N_i \sim Ber(1 - w_i)$.

La log-vraisemblance du modèle sur toutes les variables discriminantes est de -18426 avec 43 degrés de liberté. Le modèle semble bien s'ajuster aux données, et nous vérifierons ceci.

11.5 Comparaison des modèles de fréquence

Nous commençons par établir quel modèle estime le mieux la fréquence des zéros.

TABLE 11.8 – Estimation de la proportion de zéros dans le portefeuille

	Zeros_prediction
obs	44 841
NB	44 855
ML_poisson	43 533
ZI-Poisson	44 842
ZI-NB	44 874

Les modèles Binomial négative et ZIP estiment le mieux le nombre de zéros dans le portefeuille. Le modèle ZINB semble inapproprié car il sur estime la proportion des zéros.

Nous effectuons le test de Vuong sur les différents modèles pour établir lequel s'ajuste le mieux à nos données, tous les tests sont significatifs avec une p-value très proche de 0 (<0.00001). Nous rejetons l'hypothèse nulle $\mathcal{H}_0 : LR_n(\hat{\theta}, \hat{\gamma}) = 0$, qui veut dire que les deux modèles sont statistiquement différents. Nous préférerons le modèle avec la meilleure log-vraisemblance.

Nous récapitulons les résultats des tests dans le tableau suivant :

Table 11.9 – Test de Vuong 2 à 2

	ZI	**NB**	**ZI-Poisson**	**ZI-NB**
ZI	$=$	$\prec$	$\prec$	$\prec$
NB		$=$	$\succ$	$\prec$
ZI-Poisson			$=$	$\prec$
ZI-NB				$=$

Ce tableau résume les résultats du test de Vuong, effectués 2 à 2. Le modèle de Poisson est moins adéquat que tous les autres, comme l'on pouvait s'attendre. En revanche le modèle Binomial négatif représente un meilleur ajustement que le modèle de Poisson "Zéro Inflated".

Analyse des résidus

La fonction "Zeroinfl" ne calcule pas les résidus de déviance, nous donnons les "QQ-plot" des résidus de Pearson des deux modèles inflatés ainsi que leur graphique en fonction des individus.

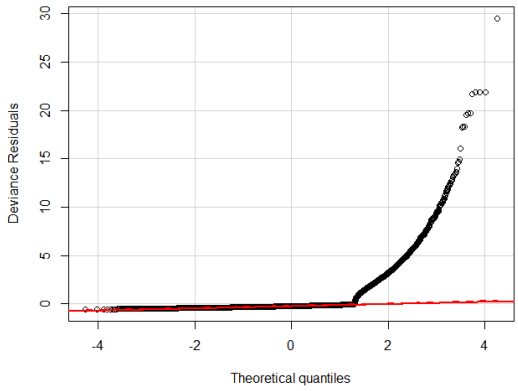


FIGURE 11.14 – QQplot, modèle de fréquence, ZIP

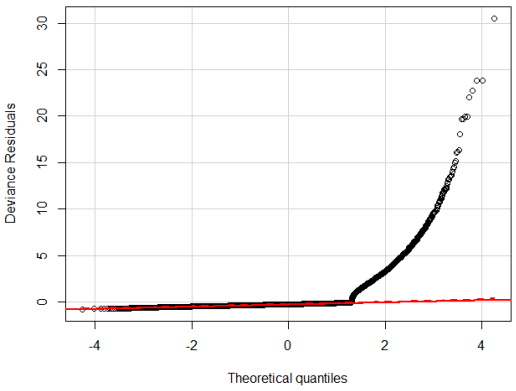


FIGURE 11.15 – QQplot, modèle de fréquence, ZINB

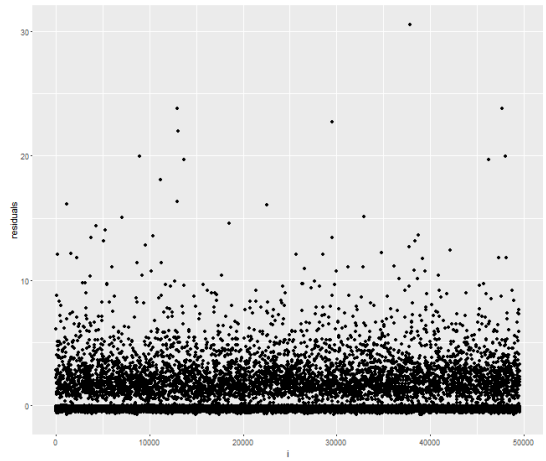


FIGURE 11.16 – Résidus de Pearson, modèle de fréquence, ZIP

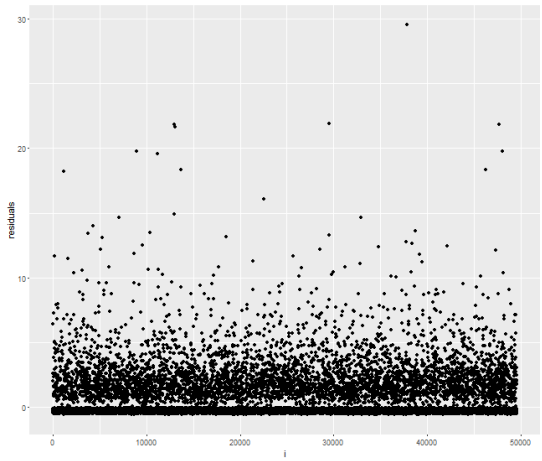


FIGURE 11.17 – Résidus de Pearson, modèle de fréquence, ZINB

L'asymétrie présente dans le modèle de Poisson persiste, il semble difficile d'améliorer l'ajustement des données par une loi paramétrique puisque l'estimation pour un contrat non sinistré est toujours > 0. Néanmoins les qqplot présentent une amélioration par rapport aux modèles précédents.

Conclusions

Bien que notre base de données comporte une surdispersion, et une fréquence trop forte des zéros, l'usage d'un modèle de type "Zero-Inflated" ne se justifie pas. Le gain apporté en ajustement ne compense pas la complexité supplémentaire.

TABLE 11.10 – Qualité d'ajustement des modèles de fréquence

	loglikelihood	AIC
Poisson	-20262.00	40565.00
ZIP	-18748.00	37580.00
NB	-18562.00	37168.00
ZI-NB	-18426.00	36938.00

En effet c'est la modélisation par loi Binomiale négative qui semble optimale, puisqu'elle prend en compte la surdispersion dans nos données, et reste utilisable d'un point de vue tarifaire.

Chapitre 12

Application : Modélisation du coût moyen des sinistres

Dans ce chapitre nous abordons la modélisation du coût moyen des sinistres.

Données disponibles et hypothèses

- Nous ne possédons pas le montant de charges de chaque sinistres, nous utiliserons comme variable réponse le coût moyen des sinistres par contrat $\overline{Y_i} = S_i/N_i = \sum_i^{N_i} Y_i$.
- Rappelons que notre périmètre d'étude comporte uniquement les sinistres inférieurs à 50 k, ceci a un impact sur le support de la loi de la variable réponse. Néanmoins nous négligerons cet impact dans notre étude par soucis de simplification.
- Nous prenons uniquement en compte les sinistres avec un montant de charge supérieur à 0.

On commence par une analyse des données de charges.

Les quantiles empiriques des charges moyennes sont :

Table 12.1 – Quantiles empiriques des charges

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.06	786	1 811	4 471	4 966	49 970

La distribution empirique a la forme suivante :

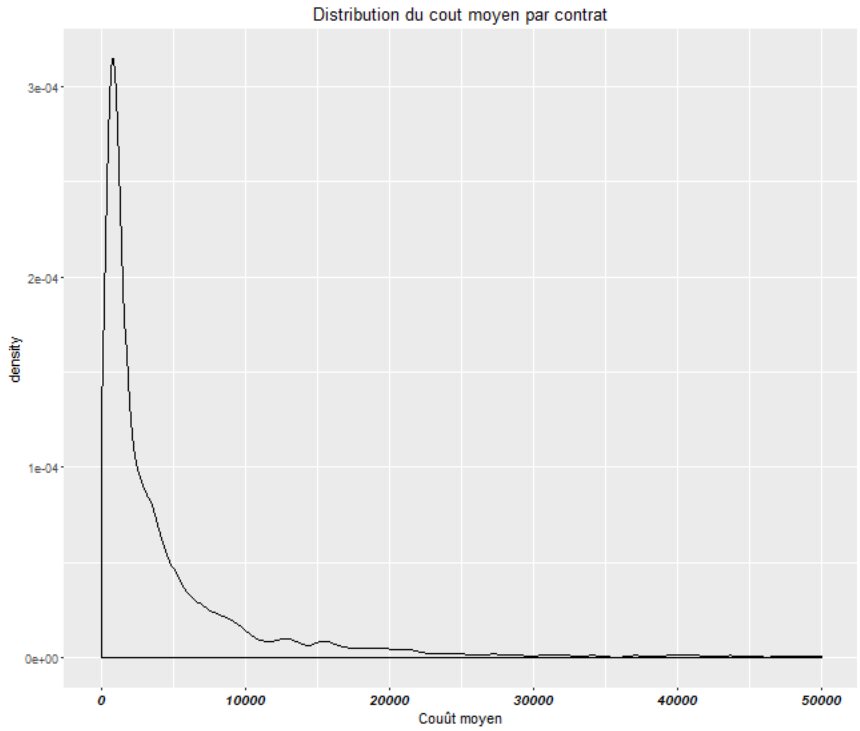


FIGURE 12.1 – Distribution empirique des charges moyennes par contrat

La distribution des charges moyennes a l'allure d'une distribution Gamma, qui a été initialement choisie pour la modélisation.

12.1 Modélisation du coût moyen des sinistres par une loi de Gamma

Le modèle utilisé est le suivant :

TABLE 12.2 – Modèle 3.1 Coût moyen, loi de Gamma

Composante du modèle	**Description**
Variable réponse	$\frac{S_i}{N_i}$, coût moyen des sinistres par contrat-DROC
fonction de lien	*log*
exposition	
offset	
poids *a priori*	n_i Nombre de sinistres par Contrat-DROC
distribution	Gamma

Nous effectuons une sélection des variables automatique ascendante avec le critère AIC, et la fonction *step* de R. Les variables les plus discriminantes sont les suivantes :

TABLE 12.3 – Test du rapport de vraisemblance, variables significatives

	Df	Deviance	AIC	F value	Pr(>F)
<none>		9700.50	128383.23		
nombre de familles métier	1	9715.60	128386.42	7.38	0.0066
effectif entreprise	4	9762.30	128396.48	7.56	0.0000
chiffre d'affaires	3	9736.90	128389.75	5.93	0.0005
antécédents	1	9710.77	128384.76	5.02	0.0251
classe de risque	8	9758.43	128387.15	3.54	0.0004
région d'implantation	13	9775.89	128383.16	2.84	0.0005

Toutes ces variables sont significatives d'après le test du rapport de vraisemblance. Néanmoins d'un point de vue de cohérence tarifaire, seules deux variables sont utilisables et le modèle sélectionné est le suivant :

TABLE 12.4 – Coefficients du modèle final de coût moyen

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.2143	0.0444	185.05	0.0000
chiffre d'affaires [>3M]	0.0819	0.0491	1.67	0.0955
nombre de familles métier (4,Inf]	0.1671	0.0491	3.40	0.0007
région d'implantation [27,28]	0.1563	0.0690	2.26	0.0236
région d'implantation 11	0.2468	0.0940	2.63	0.0086

La variable région d'implantation apporte un pouvoir discriminant, notamment car l'on observe une sévérité plus importante pour les constructions en Ile de France (Région 11).

Nous passons à l'analyse de la qualité de l'ajustement pour ce modèle.

Qualité d'ajustement globale

Nous effectuons le test d'ajustement globale du χ^2, néanmoins nous ne possédons pas assez de données pour dire que la loi de la déviance est proche d'une χ^2. Il restera à analyser les résidus pour conclure.

La déviance est égale à 9975 avec 4774 degrés de liberté, la p-value est égale à 0. La log-vraisemblance de ces modèles est égale à -64237.79 avec 6 degrés de liberté.

Nous remarquons par ailleurs que le pouvoir de segmentation du modèle est très restreint. L'AUC de la courbe de Lorenz est de :

Analyse des résidus

Nous effectuons les mêmes analyses que pour le modèle de coût moyen :

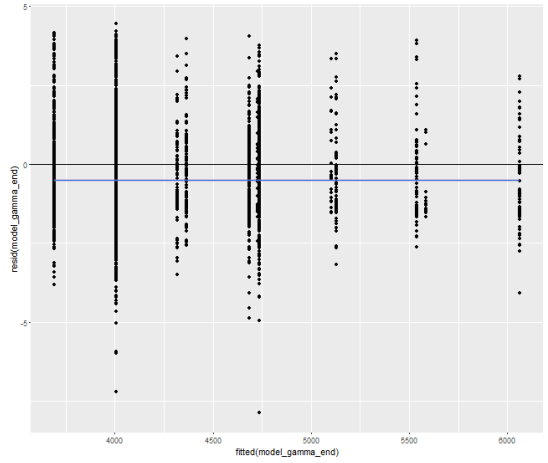


FIGURE 12.2 – Graphique des valeurs prédites versus résidus, modèle Gamma

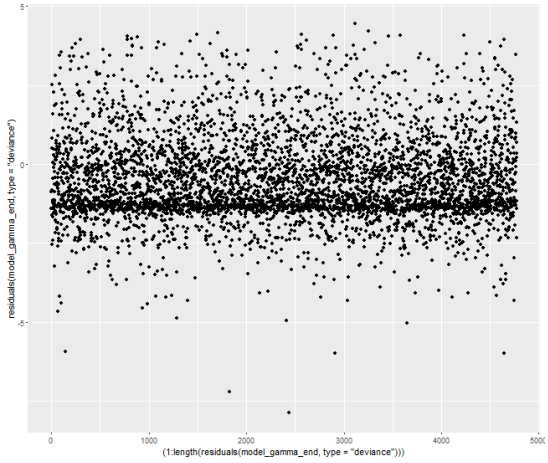


FIGURE 12.3 – Résidus de déviance, modèle Gamma

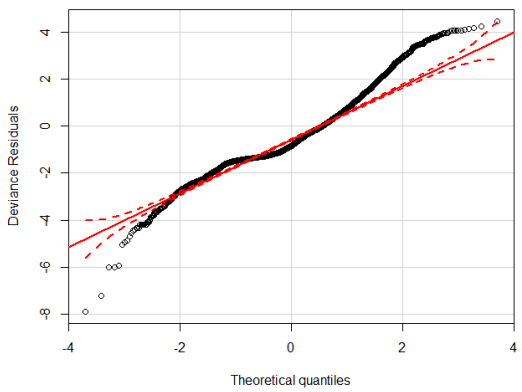


FIGURE 12.4 – Normal QQ-Plot, résidus de déviance de Gamma

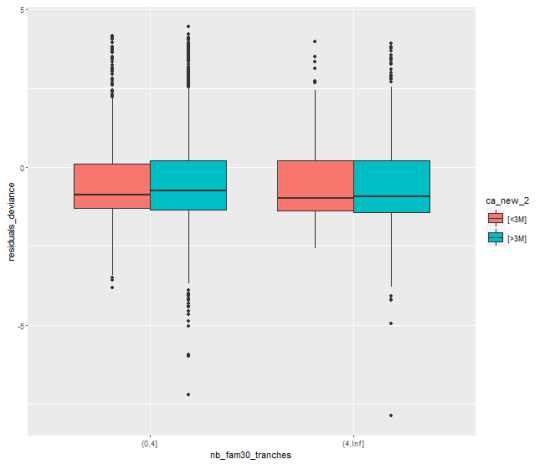


FIGURE 12.5 – Boxplot des résidus croisé entre les variables chiffre d'affaires et classe de risque, modèle de Gamma

En se basant sur les graphiques ci-dessus, nous pouvons dire que les résidus de déviance sont vraisemblablement normaux. Par ailleurs, nous n'observons pas de lien entre les variables chiffre d'affaires et nombre d'établissements actifs et les résidus. Finalement le box-plot nous donne un indice de symétrie des résidus.

Le modèle est jugé satisfaisant.

Pouvoir de segmentation du modèle

Nous analysons le pouvoir de segmentation de ce modèle à partir des courbes de Lorenz :

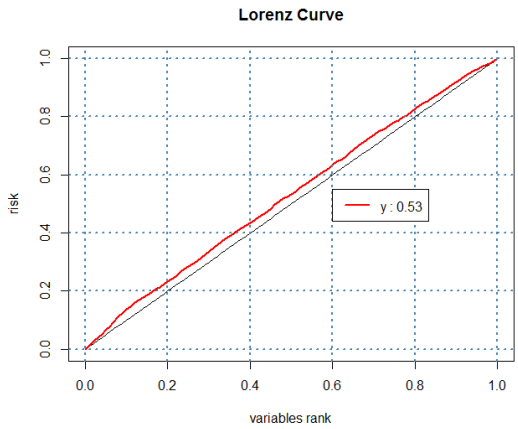


FIGURE 12.6 – Courbe de Lorenz sur la base d'apprentissage

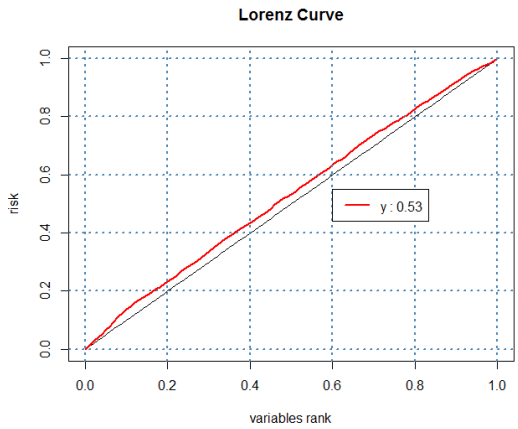


FIGURE 12.7 – Modèle Gamma, Lorenz Courbe, base de test

Notre prédiction pour le coût moyen est très proche de l'aléatoire, ceci est probablement dû au fait de modéliser uniquement les sinistres inférieurs à 50k.

12.2 Modélisation du coût moyen par une loi log-normale

La loi log-normale n'est pas un cas particulier des "GLM". Le modèle qu'on teste est le suivant :

TABLE 12.5 – Modèle 3.2 Coût moyen, loi-normale

Composante du modèle	**Description**
Variable réponse	$log(Y_i)$, coût moyen des sinistres par contrat-DROC
fonction de lien	$Identité$
exposition	
offset	
poids *a priori*	n_i nombre de sinistres par Contrat-DROC
distribution	normale

Rappelons que $Y \sim LN(\mu, \sigma^2)$ ssi, $log(Y) \sim N(\mu, \sigma^2)$. Nous estimons ce modèle donc en effectuant une régression linéaire du logarithme du coût moyen sur les variables explicatives. L'estimation de Y_i sera donnée par $exp(\mu_i + \sigma^2/2) \neq exp(\mu_i)$

Qualité d'ajustement du modèle

En se basant sur la Log-vraisemblance des deux modèles, nous retenons le modèle issu de la loi Gamma, car il possède la meilleure log-vraisemblance.

TABLE 12.6 – Log-vraisemblance modèles de coût moyen

	Log-vraisemblance
Gamma	-6244
Lognormale	-8170

Conclusions

Le modèle Gamma représente le meilleur ajustement du coût moyen de notre portefeuille. Néanmoins son pouvoir de segmentation, représente par la courbe de Lorenz et un AUC proche de 0.5. Cela nous amène à ne pas conserver ce modèle.

Rappelons que le but premier de l'approche fréquence-coût moyen est d'expliquer la fréquence et la sévérité séparément. Nos modèles de fréquence sont satisfaisants du point de vue de l'adéquation vis à vis des données. En revanche nous ne parvenons pas à expliquer la sévérité.

Ainsi, l'approche fréquence-coût moyen rajoute une couche de complexité supplémentaire, sans vraiment nous apporter plus de pouvoir explicatif. Nous avons donc décidé de conserver uniquement les modèles de Tweedie.

Quatrième partie

Comparaison des résultats selon les différents périmètres

Dans cette partie nous cherchons à répondre à la question suivante : quel est l'impact du choix de la vision de la sinistralité dans la modélisation de la prime pure ? est ce que la typologie des contrats reste la même indépendamment de la période observée ?

Pour ce faire nous considérons à présent 3 sous-périmètres d'observations distincts :

- DROCS 2000-2005
  - survenance à 10 ans,
  - survenance à 5 ans,
- DROCS 2005-2009
  - survenance à 5 ans

Sur chacun des périmètres, nous estimons les coefficients des modèles retenus dans la partie précédente, et nous comparons d'une part, les variables significatives pour chaque périmètre, et d'autre part les coefficients estimés.

Finalement sur le premier périmètre, nous estimons les coefficients en fonction de la période observé et analysons les différences.

Le modèle retenu pour cette partie est le modèle de Tweedie, car il est le seul qui nous permet de modéliser un taux applicable au chiffre d'affaires, qui est l'approche souhaitée par la direction.

Chapitre 13

Comparaison des résultats du modèle de Tweedie pour le taux sur chiffre d'affaires

Les variables significatives sont les mêmes indépendamment du périmètre :

1. chiffre d'affaires
2. classe de risque
3. note financière
4. nature des antécédents
5. nombre de familles métier
6. région d'implantation
7. ancienneté

Dans le but de comparer les coefficients des modalités utilisés pour la tarification actuelle, nous gardons toues les modalités, mêmes si elles ne sont pas significativement différentes des modalités de référence.

Nous présentons les coefficients issus du modèle selon l'âge pivot x.

Nous commençons avec les variables classe de risque et tranche de chiffre d'affaires.

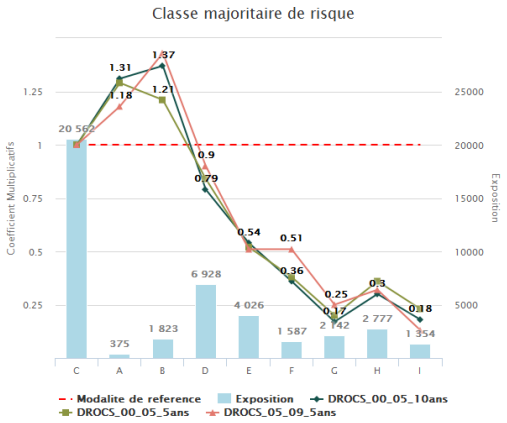


FIGURE 13.1 – Comparaison des coefficients, classe de risque

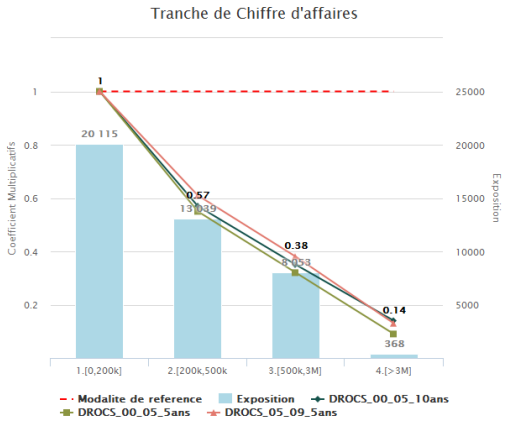


FIGURE 13.2 – Comparaison des coefficients, tranche de chiffre d'affaires

Nous constatons que les coefficients ne varient pas en fonction du périmètre, cette stabilité est particulièrement remarquable pour la variable chiffre d'affaires.

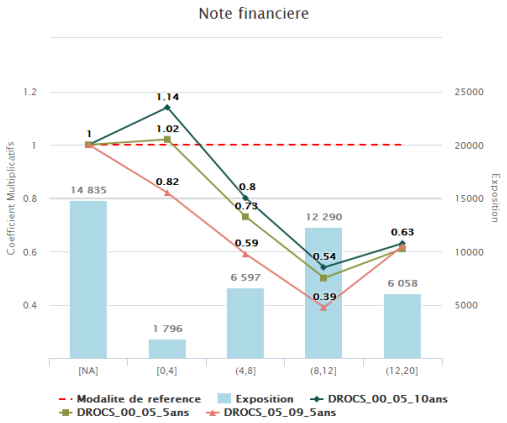


FIGURE 13.3 – Comparaison des coefficients, Note financière

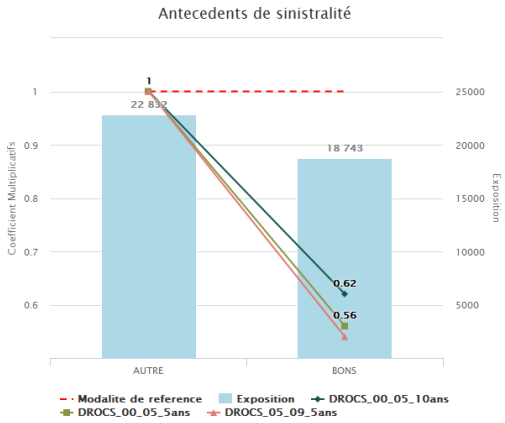


FIGURE 13.4 – Comparaison des coefficients, antécédents de sinistralité

Les coefficients des variables note financière et antécédents de sinistralité sont plus variables, mais nous conservons les mêmes tendances indépendamment du périmètre.

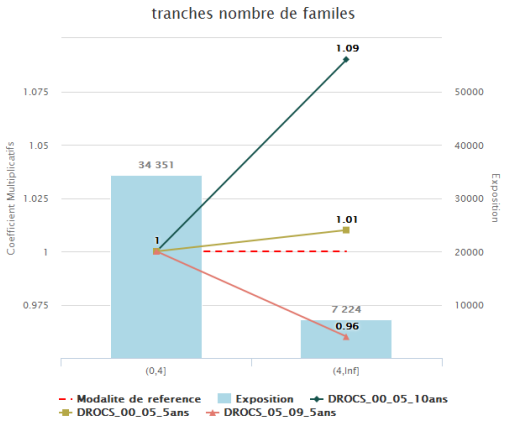


FIGURE 13.5 – Comparaison des coefficients, nombre de familles métier

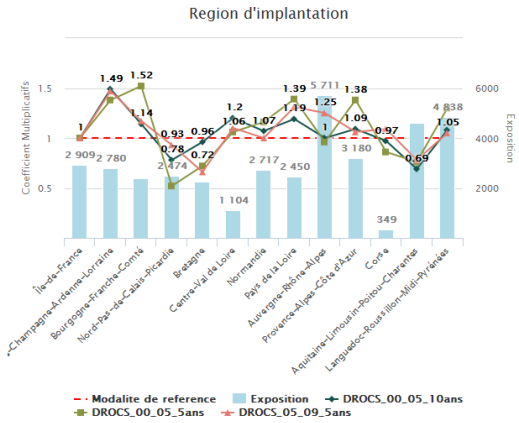


FIGURE 13.6 – Résidus de déviance, région d'implantation

En revanche la variable nombre de familles d'activité nous fournit une information tarifaire différente selon chacun des périmètres. Ce résultat nous amène à écarter cette variable de la tarification de la prime pure.

La variable région d'implantation nous donne des indices sur les régions qui seraient les plus risquées. Par exemple la région Alsace-Champagne-Ardennes-Lorraine ressort comme une région plus risquée que l'Ile-de-France. Néanmoins les coefficients sont très variables pour le reste des régions, ce qui nous amène à écarter également cette variable.

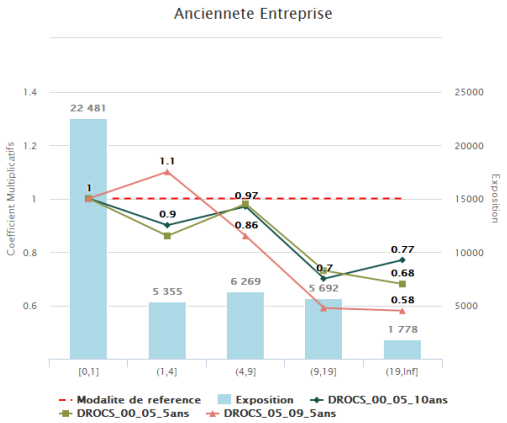


FIGURE 13.7 – Comparaison des coefficients, ancienneté

Finalement la variable ancienneté nous montre une différence claire entre les DROCS 2005-2009 et les DROCS 2000-2005. Cet effet n'est pas étonnant car sur les mêmes entreprises

nous observons un âge différent, et une sinistralité qui suit les mêmes ordres de grandeur. Nous parlerons plus tard des améliorations possibles pour cette variable.

Conclusions

Nous avons comparé les modélisations issues de 3 sous-périmètres afin de valider notre hypothèse **d'âge pivot à 5 ans**. Les résultats de cette étude s'avèrent satisfaisants.

Globalement, nous observons des tendances constantes indépendamment du périmètre, ce qui nous amène à accepter l'idée que l'âge pivot à 5 ans nous permet de discriminer les entreprises par rapport à leur sinistralité.

Cinquième partie

Conclusions générales

Ce mémoire qui s'attachait à étudier la garantie décennale d'un portefeuille d'assurance construction avait pour objectif de trouver le modèle le plus adéquat permettant de modéliser la prime pure de cette garantie, puis de tester l'impact de choix du périmètre d'observation sur ce modèle.

En effet, s'agissant d'une garantie à survenance lente, la sinistralité la plus mature (i.e. en termes de manifestation des sinistres) est celle des générations de chantier les plus anciennes, qui correspondent à des entreprises dont le comportement n'est plus forcément le même de nos jours. D'autant plus que le marché français de la construction a subi plusieurs crises ces dernières années. Une alternative consistait à modéliser la RCD en se basant sur des générations de chantier récentes avec un niveau de manifestation satisfaisant, sans pour autant qu'il soit complet, puis d'estimer la prime pure finale.

Ce mémoire a donc testé cette hypothèse en comparant deux périmètres d'observation : le premier avec une manifestation de sinistralité complète à 10 ans, puis un autre avec une période de manifestation incomplète, limitée à 5 ans de manifestation uniquement.

Dans un premier temps, nous avons testés plusieurs modèles, d'abord de Fréquence * Coût Moyen puis d'autres basés sur une modélisation directe de la charge.

Nous constatons que le modèle de Tweedie modélisant directement la charge était le plus fiable et le plus robuste. Il est en plus le plus simple, car il permet une modélisation directe de la prime pure. La structure de la prime pure donnée par ce modèle a ensuite été validée en utilisant les outils statistiques classiques (ACP, V de Cramer).

Ensuite, un test de sensibilité du modèle est effectué sur les deux périmètres : celui à 10 ans puis celui à 5 ans. Nous observons que les variables significatives restent les mêmes, et que les coefficients présentent une certaine stabilité. Ainsi, nous pouvons valider l'approche utilisée par la DACST pour la tarification de la garantie RCD sur plusieurs aspects :

- l'adéquation du modèle retenu
- la structure de la prime pure (variables discriminantes)

- la validité de l'approche 5 ans

Plusieurs points restent cependant en suspens, notamment concernant le ralentissement des cadences de manifestation. Ce décalage peut avoir de graves impacts sur l'approche à 5 ans s'il perdure dans le temps et doit être surveillé.

D'autre part, un enrichissement de données doit être envisagé, car le long déroulement des garanties décennales nécessite un suivi des entreprises dans le temps, pour un suivi plus précis du risque.

Bibliographie

Philippe Besse. Statistique descriptive bidimensionnelle. *Université de Toulouse*, 2016. URL http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-des-bi.pdf.

Jean Bérard. *Modèles de régression.* Université de Strasbourg, 2016. URL http://www-irma.u-strasbg.fr/~jberard/regression-transp.pdf.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees.* CRC press, 1984.

Arthur Charpentier. Statistique de l'assurance ii. Master's thesis, Université de Rennes 1, Université de Montréal, 2010-2011. URL http://www.dphu.org/uploads/attachements/books/books_330_0.pdf.

Dominique Foata. *Processus stochastiques.* Dunod, 2004.

Bent Jorgensen. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 127–162, 1987.

Quijano Oscar & Garrido Jose. *Generalised linear models for aggregate claims ; Tweedie or not.* 2. Concordia University, 2014. URL totweedieornot.sourceforge.net.

Rob Kaas. Compound poisson distributions and glm ?s tweedie ?s distribution. lecture, royal flemish academy of belgium for science and the arts. 2005. URL http://www.afmathconf.ugent.be/FormerEditions/Proceedings2005.pdf#page=11.

Michel Luzi. *Assurance IARD Interprétation des chiffres.* Economica, 2006.

J.A. Nelder and R.W.M. Weddernurn. Generalised linear models. *Journal of the Royal Statistical Society*, A :135 :370–384, 1972.

Laurent Rouvière. *Régression logistique avec R.* Université Rennes 2, 2015. URL http://perso.univ-rennes2.fr/system/files/users/rouviere_l/poly_logistique_web.pdf.

Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View.* Cambridge University Press, 2015. URL http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

Tufféry Stéphane. *Modélisation prédictive et apprentissage statistique avec R.* Editions TECHNIP, 2015.

Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica : Journal of the Econometric Society*, pages 307–333, 1989.

Hui Zou Yi Yang, Wei Qian. Insurance premium predictin via gradient tree-boosted tweedie compound poisson models. *McGil University*, pages 307–333, 2016.

Liste des acronymes

ACM Analyse de Composantes Multiples . . . 49
ANOVA "Analysis of Variance" . . . 51
AUC Area under the curve . . . 73
BTP Batiments et Travaux Publics . . . 2
BIC "Bayesian Information Criterion" . . . 69
CAH Classification Ascendante Hiérarchique . . . 40
DO Dommages Ouvrages . . . 8
DOC Date d'Overture de Chantier . . . 11
DROC Déclaration Réglementaire d'Overture de Chantier . . . 12
DTFAN Date d'affaire nouvelle . . . 30
DTFRS Date de résiliation . . . 30
GLM "Generalised Linear Models" . . . 3
GBM "Gradient Boosting Machine" . . . 76
IARD Incendie, Accidents et Risques Divers . . . 13
IBNER "Incurred But Not Enough Reported" . . . 14
IBNYR "Incurred But Not Yet Reported" . . . 14
PSAP Provisions pour sinistres à payer . . . 14
PSNEM Provision pour sinistres non-encore manifestés . . . 14
RC Responsabilité Civile . . . 8
RCD Responsabilité civile décennale . . . 2
TPE Très Petites Entreprises . . . 17

Liste des tableaux

*

Table des figures

*