Mémoire présenté devant

L'Institut de Statistique de l'université Pierre et Marie Curie

Pour l'obtention du

Diplôme de Statisticien Mention Actuariat

Par Virginie POUNA SIEWE

Sujet : **Modèles additifs généralisés : Intérêts de ces modèles en assurance automobile.**

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

Membre présent du jury de l'Institut des Actuaires

Arnaud Cohen

Jean-Marie Nessi

Entreprise :

Direction Assurance de Biens et responsabilités Allianz France

Membres présents du jury

Christian Hess

Directeur de mémoire en entreprise :

Laure LACAZE

Patrick LEVEILLARD

Directeur de mémoire ISUP

Abder OULIDI

Sécrétariat : 01 44 27 70 49

Signature du candidat : VP

Remerciements

Je tiens à remercier Mlle Laure Lacaze et M. Patrick Leveillard, pour leur implication et leur disponibilité tout au long de mon stage au sein du département Tarif et Rentabilité Automobile. Je remercie aussi M. Jean-François Sutter, Directeur Automobile d'Allianz France, pour m'avoir offert cette opportunité de travailler sur ce sujet passionnant. Je remercie tout particulièrement M. Abder Oulidi pour sa patience et son aide à la réalisation de ce mémoire.

Ensuite, je remercie l'ensemble des équipes de la Direction Assurance de Biens et Responsabilités, pour leur accueil chaleureux, leur disponibilité et tout leur soutien.

Pour finir, je dédie ce mémoire à mon oncle, M. André Ngalaho qui m'a permis dans tous les sens du terme, de compléter mes études d'actuariat. A ma famille et à tous mes amis.

Résumé

En 1958, l'assurance automobile devient obligatoire :

- Le parc automobile de 4 roues explose (le parc automobile 4 roues de particuliers passe de 2 millions à 30 millions de véhicules entre 1950 et 2005).

- L'inflation et une sinistralité élevée permettent de continuer à développer le chiffre d'affaire.

Cette situation a attiré des acteurs très différents[1] sur le marché automobile : les compagnies dites « traditionnelles » (avec intermédiaires), puis les Mutuelles sans intermédiaires, les Bancassureurs au début des années 1990, les Directs et enfin les Comparateurs (type Assurland ou HyperAssur). Mais la tendance est maintenant à une saturation du marché de l'assurance automobile avec des évolutions de parc comprises entre 1% et 2% ces dernières années. La compétition prix est inévitable et le marché est complexe du fait des différences entre les intervenants en termes d'atouts, d'exigence de rentabilité, de frais de fonctionnement ou de modes de distribution.

Ce marché ultra-concurrentiel impose évidemment une technicité de plus en plus accrue dans l'approche de tarification des contrats et d'appréhension des risques gérés.

C'est dans ce cadre qu'il m'a été confié cette mission au sein du département Tarif et Rentabilité. Mes recherches ont porté sur l'exploration d'une nouvelle classe de modèles pour la tarification : les modèles additifs généralisés (GAM) qui commencent à être utilisés pour le provisionnement stochastique.

Dans ce mémoire, nous présentons l'intérêt de ce modèle pour la tarification : le regroupement des modalités de variables continues à intégrer dans les modèles d'estimation de la prime pure qui utilisent les modèles linéaires généralisés (GLM). Dans le contexte que nous avons décrit, l**es doivent être faits tarifs de plus en plus faits à la mesure du client** et les sauts de tarif doivent être justifiés techniquement.

Mots clés

Estimation - Modèles additifs généralisés – Modèles linéaires généralisés – Lissage – linéarité – non linéarité- Paramétrique – Non paramétrique – Variables continues - Regroupement de modalités d'une variable

[1] Annexe 1 : Le contexte marché

Abstract

In 1958, auto insurance is mandatory:

- The fleet of 4 wheels explodes (fleet of "4 wheel individuals" increases from 2 million to 30 million vehicles between 1950 and 2005).
- Inflation and a high number of claims allow traditional companies to continue to develop sales.

This has attracted many different actors on the auto insurance market: after traditional companies, the insurance funds without intermediaries, bancassurance in the early 1990s, and finally Direct Comparators (type or Assurland HyperAssur). But the trend is now a saturated market of auto insurance with changes in park between 1% and 2% per year recently. The price competition is inevitable and the market is complex because of differences among stakeholders in terms of assets, requirement of profitability, operating costs or methods of distribution.

This ultra-competitive market requires a technical course more increased in the approach to pricing of contracts and understanding of risk management.

It is in this context that I was assigned this task within the department and Tariff Profitability. My research has focused on the exploration of a new class of models for pricing: generalized additive models (GAM) that get starting to be used for provisioning stochastic.

In this paper, we present the interest that we found this model for pricing: detect the nonlinear effects of continuous variables and grouping terms of continuous variables to incorporate into models to estimate the pure premium using the generalized linear models (GLM). In the context we have described, rates must more take into account the specificities of the client and the hopping rate must be technically justified.

Keys words

Generalized additive models - Generalized linear models – Smoothing – Linearity – Non linearity – parametric – non parametric – continuous variables – grouping terms of a variable

Table des matières

REMERCIEMENTS **3**

RESUME **6**

ABSTRACT **8**

TABLE DES MATIÈRES **10**

OBJECTIFS DE L'ETUDE **12**

AGF ET LE GROUPE ALLIANZ **14**

CHAPITRE I. CHAPITRE I : DESCRIPTION DU PORTEFEUILLE ET ANALYSE DES DONNEES 16

I.1. PRESENTATION DU PERIMETRE DE L'ETUDE 18
I.2. BASES DE DONNEES 18
I.2.1. Exposition aux risques *19*
I.2.2. Sinistralité *19*
I.2.3. Informations descriptives des risques *19*
I.2.4. Variables de mesure de sinistralité *21*
I.3. ANALYSE DES CORRESPONDANCES MULTIPLES 22
I.3.1. Choix du nombre d'axes factoriels *22*
I.3.2. Etude du plan 1-2 ..... *23*
I.4. MESURES D'ASSOCIATION 24
I.4.1. Rappels théoriques : *25*
I.4.2. Analyse des résultats obtenus sur les portefeuilles observés *27*
I.5. ANALYSE DE L'INFLUENCE DES VARIABLES CONTINUES SUR LA SINISTRALITE OBSERVEE 27
I.5.1. Variables concernant les assurés eux-mêmes *27*
I.5.2. Variables concernant les véhicules assurés *29*
I.5.3. Variables concernant les contrats *31*

CHAPITRE II. CHAPITRE II : DU MODELE LINEAIRE GENERALISE AU MODELE ADDITIF GENERALISE **34**

II.1. ELEMENTS DE THEORIE DES MODELES LINEAIRES GENERALISES 36
II.1.1. Introduction *36*
II.1.2. Composante aléatoire Y *37*
II.1.3. Fonction de lien *39*
II.1.4. Composante déterministe : Introduction au modèle additif généralisé *39*
II.2. DU MODELE LINEAIRE GENERALISE AU MODELE ADDITIF GENERALISE 41
II.2.1. Modèle additif *41*
II.2.1.1. Introduction 41
II.2.1.2. Maximum de vraisemblance pénalisée et splines de lissage cubiques 42
II.2.1.3. Régression locale pondérée : fonctions lœss 43
II.2.1.4. Nombre de degrés de libertés 43
II.2.1.5. Degré de lissage λ 44
II.2.2. Modèles additifs généralisés *45*
II.2.3. Utilisation conjointe GLM et GAM : PROC GAM sous SAS *46*
II.2.3.1. Entrées de la procédure GAM sous SAS 46
II.2.3.2. Algorithmes de résolution d'un modèle additif généralisé 47
II.2.3.3. Lecture des résultats 50

II.3. GAM COMME MODELE PRIVILEGIE D'ESTIMATION DE LA FREQUENCE ANNUELLE EN ASSURANCE AUTOMOBILE : NON 51
II.4. INTERETS DE LA GAM POUR LA TARIFICATION 52

CHAPITRE III. CHAPITRE III :INTEGRATION D'UNE VARIABLE CONTINUE DANS LE MODELE D'ESTIMATION DE FREQUENCE : EXEMPLE DE L'AGE DU CONDUCTEUR 54

III.1. LA GARANTIE RESPONSABILITE CIVILE CORPORELS 56
III.1.1. Non linéarité 56
III.1.2. Catégorisation de la variable 58
III.1.2.1. Graphe de l'influence partielle de l'âge sur la fréquence de sinistres 58
III.1.2.2. Choix des classes 59
III.1.2.3. Comparaison avec le découpage initial de l'âge 63
III.2. LA GARANTIE DOMMAGES 63
III.2.1. Non linéarité 63
III.2.2. Catégorisation de la variable 65
III.2.2.1. Graphe de l'influence partielle de l'âge sur la fréquence de sinistres 65
III.2.2.2. Choix des classes 66
III.2.2.1. Comparaison avec le découpage initial de l'âge 69
III.3. COMPARAISON DE MODELES ANCIEN DECOUPAGE- NOUVEAU DECOUPAGE DES VARIABLES AGE DU CONDUCTEUR, AGE DU VEHICULE ET ANCIENNETE DU CONTRAT 70
III.3.1. La garantie Responsabilité civile Corporels 70
III.3.2. La garantie Dommages 71

CONCLUSION 72

ANNEXES 74

BIBLIOGRAPHIE 84

Objectifs de l'étude

Les modèles linéaires généralisés sont largement utilisés dans le domaine de l'assurance et il existe plusieurs algorithmes de résolution robustes et mathématiquement bien définis. Ils permettent entre autre de construire une structure tarifaire technique très fine tout en utilisant un grand nombre de variables explicatives. Le modèle linéaire généralisé est par définition un modèle paramétrique. Cela permet de réduire le temps de calcul des algorithmes de résolution utilisés pour l'estimation des paramètres du modèle. Cependant, le modèle linéaire généralisé fait l'hypothèse d'une relation linéaire (ou autre forme paramétrique) entre la variable à expliquer et chacune des variables explicatives, cette hypothèse a retenu notre attention.

Le modèle additif généralisé, introduit en 1990 par Trevor Hastie et Robert Tishbarani dans un livre : « Generalized additive models » est une extension du modèle additif. Le modèle additif relâche l'hypothèse de linéarité de la relation entre la variable réponse et les variables explicatives, le modèle remplace le prédicteur linéaire par une somme de fonctions estimées par des techniques de lissage non paramétriques. Ces techniques de lissage nécessitent la continuité des variables car elles consistent à l'estimation de fonctions locales d'ajustement aux données.

En 2006, dans le mémoire intitulé « *Une méthode alternative de provisionnement stochastique en Assurance Non Vie: Les Modèles Additifs Généralisés »,* Elise Lheureux démontre l'intérêt de ces modèles comme alternative aux modèles linéaires généralisés pour le provisionnement stochastique. En effet, en provisionnement les deux principales variables explicatives du montant des réserves sont l'année de développement de paiements et l'année d'origine des sinistres. Ces deux variables peuvent être considérées comme continues et de ce fait le modèle additif généralisé a pu être testé et approuvé comme méthode alternative de provisionnement stochastique.

Ces modèles nous ont semblé intéressants à explorer pour la tarification technique car plusieurs variables qui expliquent la sinistralité automobile peuvent être considérées comme continues :

- L'âge du conducteur
- L'âge d'obtention du permis
- L'âge du véhicule
- L'ancienneté du contrat.
-

Notre objectif n'est pas de remplacer les modèles linéaires généralisés « classiques » par des modèles additifs généralisés. Essentiellement pour des raisons de complexité (commerciale, statistiques ou informatiques), les modèles additifs sont jugés comme étant peu adaptés aux cas de tarification en assurance automobile.

Ce qui nous intéresse davantage ici, est de déceler la forme précise (qui est supposée linéaire par les GLM) de l'influence sur la sinistralité des variables dites continues et ainsi **déterminer un regroupement optimal des modalités à effectuer en amont de la mise en œuvre d'un GLM**. Un des principaux leviers d'action de l'Actuaire pour améliorer l'estimation de la réelle charge de

sinistre à attendre sur une police est **de segmenter au mieux le portefeuille** regroupant, en amont de la modélisation des risques qui ont un niveau de sinistralité à peu près équivalent ; ceux-ci auront en effet un même niveau de sinistralité estimé, nous nous intéressons aux variables continues qui, le plus souvent sont regroupées de façon intuitive ou commerciale, parfois indépendamment des sinistres étudiés. Nous voulons une segmentation plus justifiée techniquement.

En construisant des modèles qui mêlent GLM et GAM, nous allons déterminer la forme réelle de la relation entre chaque variable continue et la réponse. Ainsi, en observant les courbes fournies par ces modèles, nous proposerons un meilleur découpage de ces variables. En effet, pour chaque variable continue, le GAM lissera la forme de la relation entre ces variables et la réponse. Il permet ainsi, en observant la fonction de lissage, d'identifier quelles sont les valeurs successives à regrouper pour améliorer la vraisemblance des modèles construits et diminuer l'erreur de prédiction faite par les modèles implémentés.

Dans le chapitre I, nous présenterons les données à notre disposition. Nous allons aussi observer l'influence des variables continues sur la sinistralité (fréquence de sinistres). Dans la suite, nous nous intéresserons principalement à la modélisation du nombre de sinistres comme support de notre étude. Dans le chapitre II, en présentant la théorie des modèles linéaires généralisés, nous motiverons l'utilisation des modèles additifs généralisés pour compléter ces modèles. Nous présenterons aussi des éléments de théorie des GAM ainsi que les algorithmes utilisés pour résoudre ces modèles. Dans le chapitre III, nous allons mettre en œuvre l'utilisation conjointe des modèles GLM et GAM pour :

- Déceler les influences non linéaires des variables continues
- Regrouper les modalités des variables continues de manière optimale

AGF et le groupe ALLIANZ

- **Historique**

« AGF devient ALLIANZ » : Le changement de nom de la compagnie AGF le 14 septembre 2009 a été un évènement particulièrement marquant pour l'environnement assurantiel français et européen. Ainsi, nous avons trouvé intéressant de revenir sur l'histoire du groupe Allianz, cet assureur allemand devenu leader en matière de services financiers internationaux et sur son histoire avec AGF.

Allianz voit le jour en 1890 à Munich en réponse à l'augmentation des risques liés aux activités de l'ère industrielle. Offrant, au départ, des assurances pour le transport et les accidents, Allianz développe rapidement une offre de couverture incendie. Au début du XXème siècle, les activités internationales prennent leur envol. En 1913, 20% des revenus proviennent de l'étranger. Les deux guerres mondiales restreignent toutefois le développement international, jusqu'à la reprise des années 50 : des bureaux ouvrent alors à Paris puis en Italie, au Royaume-Uni, au Pays-Bas, en Espagne, au Brésil et aux Etats-Unis.

Allianz est introduite le 3 novembre 2000 au NYSE (New-York Stock Exchange) et Allianz est aujourd'hui le numéro un européen de l'assurance par la capitalisation boursière (35,8 milliards d'euros), devant Axa (32,9 milliards).

- **D'AGF à ALLIANZ FRANCE**

1818 Martin d'André fonde la Société Anonyme des assurances générales (« La générale ») ; en France, c'est la première société qui regroupe des activités d'assurance incendie, d'assurance maritime et d'assurance vie.

1849 En pleine révolution industrielle, « La générale » signe un traité de réassurance avec le Phénix de Londres. Elle devient alors l'une des meilleures compagnies d'assurance au monde. Elle diversifie son activité notamment en investissant dans l'agriculture (assurance contre la grêle) et l'industrie (assurance contre le risque d'explosion des machines à vapeur).

1945 Charles de Gaulle nationalise les grands secteurs de l'économie, notamment le secteur de l'assurance. Les Assurances générales et le Phénix deviennent propriété de l'Etat.

1968 Dans le cadre de restructuration des assurances nationalisées, la Générale fusionne avec le Phénix pour devenir « Les AGF ».

1996 « Les AGF » retrouvent un statut privé quand l'état revend 51% de ses parts.

1997 « Les AGF » se rapprochent de l'allemand Allianz AG, avant de fusionner avec Athéna. Le groupe AGF devient la 3ème compagnie d'assurance de France.

1998 Allianz possède 57,9% des parts d'AGF.

2007 Allianz rachète les actions minoritaires. AGF devient ainsi une filiale d'Allianz à 100%.

2009 L'entreprise AGF change de nom pour prendre celui du groupe auquel elle appartient depuis 10 ans : ALLIANZ, leader européen de l'assurance et des services financiers.

Ce changement permet de prendre pleinement appui sur les forces de la marque Allianz : dimension internationale, solidité et expertises financières, modernité et innovation, pour répondre au mieux aux attentes des consommateurs français.

- **Allianz (ex-AGF) en France**

L'activité d'Allianz en France s'étend à de nombreux domaines : assurance de biens et de responsabilités (38% de l'activité), assurance de personnes (45%), assistance et assurance-crédit (15%), mais aussi gestion d'actifs, services bancaires et financiers.

Allianz en France c'est aujourd'hui :

- La 2ème filiale d'Allianz dans le monde
- Le 5ème assureur français
- 6,6% de part de marché
- 12275 collaborateurs
- 12 milliard d'euros de chiffre d'affaires en assurances
- 409 millions de résultat opérationnel en assurances
- Un réseau de distribution fondé sur la proximité et le conseil

Les réseaux et partenaires

CHAPITRE I. Chapitre I : DESCRIPTION DU PORTEFEUILLE ET ANALYSE DES DONNEES

Le but de cette partie est de :

- Décrire le portefeuille sur lequel nous allons effectuer notre étude,
- Analyser les données que nous avons à notre disposition,
- Faire une première analyse de l'influence des différentes variables continues sur la sinistralité.

I.1. Présentation du périmètre de l'étude

Notre étude porte sur le portefeuille automobile d'Allianz France. Nous nous intéressons uniquement aux particuliers et aux véhicules à quatre roues. Ce segment est communément appelé « 4 roues de particuliers » ou 4RP.

Au 31/12/2008, le parc Allianz 4RP compte environ 2 millions de polices. Pour les 4RP, Allianz propose une « palette » de garanties regroupées sous forme de formule à souscrire. Présentons les garanties sur lesquelles nous allons travailler.

Le tableau I-1 suivant décrit les différentes garanties concernées par l'étude :

Garantie	Description succincte
La responsabilité civile	C'est une assurance obligatoire pour l'assuré et couvre les dommages matériels ou corporels causés à autrui.
La garantie Vol	Elle couvre la disparition ou la détérioration, suite à un vol ou tentative de vol du véhicule (ou de ses éléments).
La garantie Incendie	Elle couvre tous les dommages matériels causés par un incendie.
La garantie Bris de Glace	Elle couvre les éléments du véhicule suivants : Pare brise, glaces latérales, vitre arrière, toit ouvrant, feux avant.
La garantie Dommages Tous accidents	Elle couvre tous les risques y compris le vandalisme.

Tableau I-1 : Description des garanties proposées

I.2. Bases de données

Nous désignerons par « risque », l'ensemble des caractéristiques **« client »**, **« véhicule »**, **« géographie »** et **« contrat »** de la police sur laquelle porte l'assurance à un instant donné. En effet, ce sont ces différentes caractéristiques qui permettent de mesurer le niveau de risque représenté par une police à un instant donné.

I.2.1. Exposition aux risques

Toutes les informations sur l'évolution du risque associé à un contrat pendant une année sont stockées dans un fichier dit « fichier par période ». En effet, au cours de l'année toute modification de la police susceptible de changer la nature du risque sous jacent au contrat est enregistrée dans ce fichier.

On peut enregistrer par exemple,

- Un changement de conducteur principal ou l'association d'une nouvelle personne au contrat,
- Un changement de véhicule,
- Un changement de domicile,
- La modification des garanties souscrites etc.

Chaque risque est repéré par un numéro de police, une date de début de période et une date de fin de période. La date de fin représente la date de modification de la police et la différence entre la date de fin de période et la date de début nous permet d'obtenir la durée d'exposition à un risque sous jacent particulier sur la police. La durée d'exposition moyenne de 5 mois pour chaque risque.

Les fichiers par période contiennent les informations liées au contrat que l'assuré fournit à l'assureur lors de la souscription du risque. Ces informations sont les bases de connaissance du « risque contenu » dans le portefeuille. Elles constituent donc des variables potentiellement explicatives de la sinistralité (fréquence et coût de sinistres) observée sur le portefeuille.

I.2.2. Sinistralité

Les sinistres sont enregistrés dans des tables dédiées qui comprennent notamment des informations sur la date de survenance, la date d'ouverture, les règlements, les provisions, la responsabilité ou la nature de la garantie sinistrée. Pour qu'un sinistre soit associé à un risque particulier, il faut qu'il concerne la police qui porte ce risque et il faut que la date de survenance de sinistre soit entre la date de début de période et la date de fin de période de ce risque. Pour l'estimation de la fréquence de sinistres, les sinistres concernant les garanties que nous étudions sont regroupés en 5 « groupes » de sinistres :

1- Responsabilité civile matérielle (RC Mat. ou RCM)
2- Responsabilité civile corporelle (RC Corp. ou RCC)
3- Bris de glace (BDG)
4- Vol ou incendie (VI)
5- Autres dommages matériels (DOM).

En définissant ces 5 « groupes », notre objectif est de traiter ensemble des risques supposés homogènes en termes de fréquence de survenance de sinistre.

I.2.3. Informations descriptives des risques

Ce sont les informations fournies par l'assuré au contrat. Elles sont descriptives de la police et permettent à l'assureur de mesurer le risque pris en acceptant de souscrire à la police. Elles sont supposées être des variables potentiellement explicatives de la sinistralité observée sur un risque

donné. Le tableau décrit quelques informations disponibles dans les bases de données et leur impact potentiel sur la fréquence et le coût des sinistres. Ces variables sont :

Informations	**Description**
Niveau de risque du véhicule	Le risque Responsabilité civile est influencé par les éléments déterminant la fréquence et la gravité des dommages causés aux tiers ou aux passagers tels que la vitesse, puissance, poids, éléments de sécurité. Les garanties dommages au véhicule intègrent les facteurs liés au coût du véhicule lui-même et celui de sa réparation. L'ensemble est pris en compte par les classifications SRA[2]
Age véhicule	Selon la garantie concernée, la vétusté du véhicule peut avoir un effet positif ou négatif sur la sinistralité liée à un contrat. Par exemple, l'évolution technologique confère aux voitures récentes un système de protection des passagers en cas de choc qui sont susceptibles de diminuer les coûts en cas de dommages corporels par exemple. D'un autre côté, la valeur du véhicule est moindre au fil des années en cas de vol ou de destruction.
Age conducteur	Le risque automobile étant un risque de comportement, il est donc très lié au conducteur habituel (maturité etc.)
Age obtention permis	Ce critère est basé sur l'idée que les personnes qui ont commencé à conduire très tôt (moins de 18 ans) seront de meilleurs conducteurs que les autres. Aussi, une date d'obtention de permis très récente pour un conducteur pour un assuré âgé peut signifier que celui-ci a perdu son permis et l'a repassé. Il est donc potentiellement un assuré plus risqué qu'un conducteur qui a eu son permis à 18 ans et ne l'a jamais repassé.
Ancienneté du permis	Permet de juger de l'expérience de conduite de l'assuré ;
Catégorie socio-prof.	C'est un indicateur du niveau et du type de déplacements de l'assuré, cette information complète le critère usage du véhicule
Usage	Ce critère permet d'approcher l'intensité et les conditions d'utilisation du véhicule assuré. Ainsi, véhicule au repos sera moins à même de subir des sinistres qu'un véhicule à usage professionnel.
Expérience d'assurance	Les assurés sont distingués aussi par leur expérience d'assurance et leur situation de famille. Par exemple, une distinction est faite entre les novices femmes ou hommes, les novices enfant d'assuré ou pas etc.
Zones géographiques	La densité de circulation, la qualité des infrastructures routières, le niveau de délinquance amènent à définir des zones de tarification selon la garantie souscrite. Un zonier de risque est établi et pour un contrat, la zone est déterminée par la commune du lieu de garage habituel du véhicule
Formule de garantie	Le choix de la formule par un assuré peut donner des informations et prévenir l'aléa moral.
Franchises et niveaux de franchise	Le choix de l'assuré de prendre une franchise ainsi que le niveau de franchise sur le contrat influence le comportement de l'assuré. S'il sait qu'il assumera une partie du risque, il sera un peu plus responsable.

Tableau I-2 : Description des différentes informations fournies par le client et leur intérêt

[2] Sécurité et réparation automobile

Catégorisation des variables

Pour notre étude, nous nous concentrons sur les variables quantitatives continues :

- Age du conducteur
- Age d'obtention du permis
- Age du véhicule
- Ancienneté du contrat

Nous allons chercher à optimiser leur catégorisation en utilisant la GAM.

I.2.4. Variables de mesure de sinistralité

- Le nombre de sinistres

C'est le nombre de sinistres déclarés par l'assuré. Chaque sinistre déclaré est classé dans un groupe de sinistres (RCC, RCM, VI ou DOM). Cette information est connue avec précision et est donc extrêmement fiable pour des analyses statistiques. Ainsi, cette variable sera le support de notre étude dans la détermination des classes optimales pour les variables continues.

- Le coût de sinistres

C'est le coût total mis à la charge de la compagnie pour un sinistre donné. Lors de l'étude des coûts, il convient de tenir compte du nombre de sinistres engendrés. On analysera le plus souvent le coût moyen des sinistres. Cette valeur peut être connue très longtemps après la survenance d'un sinistre.

- La durée d'exposition au risque

C'est le nombre de jours où la police a été exposée à un risque sous – jacent particulier (nombre de jours entre le début et la fin d'une période[3] donnée). Cette variable permet de quantifier l'exposition du portefeuille à un risque donné sur une période donnée.

Pour se faire une première idée de la structure de no

[3] Voir paragraphe 1-2 sur les bases de données

tre portefeuille, nous faisons une analyse des correspondances multiples du portefeuille. Nous allons aussi calculer des mesures d'association entre les variables afin de déceler d'éventuelles corrélations entre elles. Aussi, nous faisons une analyse uni-variée du portefeuille en fréquence de sinistre. Pour des raisons de confidentialités, nous ne présenterons que les résultats obtenues pour les variables continues qui nous intéressent.

I.3. Analyse des correspondances multiples

Les méthodes d'analyse de données (ACP, AFC, ACM) visent à synthétiser les informations sur les variables sans trop les déformer. Pour cela, il s'agit de trouver des axes (qui sont alors des indices graphiques) qui respectent la forme de la relation entre les variables (forme du nuage multidimensionnel que constituent nos données). Ces axes résument l'information sur nos données (oppositions, liaisons, corrélations) et permettent une interprétation plus simple de la structure des variables (nuage unidimensionnel ou en deux dimensions).

L'ACM permet ainsi d'observer

- les relations entre les modalités des différentes variables ;

- les relations entre les variables, telles qu'elles apparaissent à partir des relations entre modalités.

L'ACM prend en entrée un tableau disjonctif complet qui résume l'information contenu dans les données. Nous avons construit ce tableau disjonctif complet à l'aide des différentes variables à notre disposition. Un résumé du principe de l'ACM est donné en annexe.

I.3.1. Choix du nombre d'axes factoriels

Le choix du nombre d'axes factoriels s'effectue à l'aide des pourcentages cumulés d'inertie, (le pourcentage d'inertie expliquée devant être suffisant, mais aussi par la décroissance des valeurs propres. La règle du coude consiste à retenir le nombre d'axes comprenant le dernier axe avant un « saut » et fournissant un pourcentage d'inertie expliquée suffisant. **Enfin les axes retenus doivent être interprétables à partir des variables qui contribuent à leur construction. La « cohérence » de l'axe doit être vérifiée, c'est à dire qu'une structure doit en ressortir.**

L'inertie des axes factoriels en ACM étant toujours faible, J.P. Benzécri propose de recalculer le taux d'inertie cumulée extrait par les axes retenus, pour pouvoir le comparer à celui d'une analyse factorielle classique :

- Les valeurs propres supérieures à 1/N sont retenues où N est le nombre de variables actives dans l'ACM.

- Pour ces valeurs propres l_k, on calcule $\boxed{d_k = (l_k - 1/N)^2}$

- Le nouveau taux d'inertie cumulée est $I=\frac{\sum_{k} d_k}{\sum_{k \geq 1/N} d_k}$

Sur notre base de données, on obtient les résultats suivants :

Décomposition de l'inertie et du Khi-2					
Valeur propres	Inertie principale	Khi-2	Pourcentage d'inertie	Pourcent. Cumulé	2 4 6 8 10 ----+----+----+----+----+---
0.40545	0.16439	7625722	9.29	9.29	**********
0.39670	0.15737	7300067	8.89	18.19	**********
0.31383	0.09849	4568713	5.57	23.75	**********
0.30873	0.09531	4421336	5.39	29.14	**********
0.30195	0.09117	4229335	5.15	34.29	**********
0.29424	0.08658	4016120	4.89	39.19	**********
0.29022	0.08423	3907089	4.76	43.95	**********
0.28036	0.07860	3646124	4.44	48.39	**********
0.27888	0.07777	3607649	4.40	52.79	**********
0.27711	0.07679	3562009	4.34	57.13	**********
0.27596	0.07616	3532644	4.30	61.43	**********
0.27520	0.07574	3513150	4.28	65.71	**********
0.27175	0.07385	3425600	4.17	69.89	**********
0.26869	0.07220	3349001	4.08	73.97	**********
0.26282	0.06907	3204113	3.90	77.87	**********
0.25711	0.06611	3066479	3.74	81.61	*********
0.25248	0.06375	2957113	3.60	85.21	*********
0.24715	0.06108	2833406	3.45	88.66	*********
0.23155	0.05361	2487038	3.03	91.69	********
0.22939	0.05262	2440965	2.97	94.67	*******
0.21266	0.04522	2097807	2.56	97.22	******
0.16050	0.02576	1194990	1.46	98.68	****
0.15283	0.02336	1083539	1.32	100.00	***
Total	1.76923	8.207E7	100.00		
Degrés de liberté = 1225					

Figure I-1: Valeurs propres associées aux axes factoriels

Le saut d'inertie entre l'axe 1 et l'axe 2 est de (9,29-8,89)/8,89=4,49%. Le saut entre l'axe 2 et l'axe 3 est de (8,89-5,57)/5,57=59,60%. La règle du coude nous incite donc à étudier les deux premiers axes.

Les 2 premiers axes expliquent 18,19% de l'inertie totale. En recalculant le taux d'inertie selon la formule de J.P. Benzécri, la part d'inertie expliquée par les deux premiers axes est de 70,1%. Nous retenons les deux premiers axes dans notre interprétation.

I.3.2. Etude du plan 1-2

Pour les variables ordinales, les modalités sont rangées dans un ordre naturel. On peut ainsi tracer un faisceau qui joint les modalités entre elles.

Les variables « âge », « âge d'obtention du permis » et « bonus-malus » ont des faisceaux de même allure. Ce phénomène appelé effet Gutmann, indique une corrélation entre ces variables.

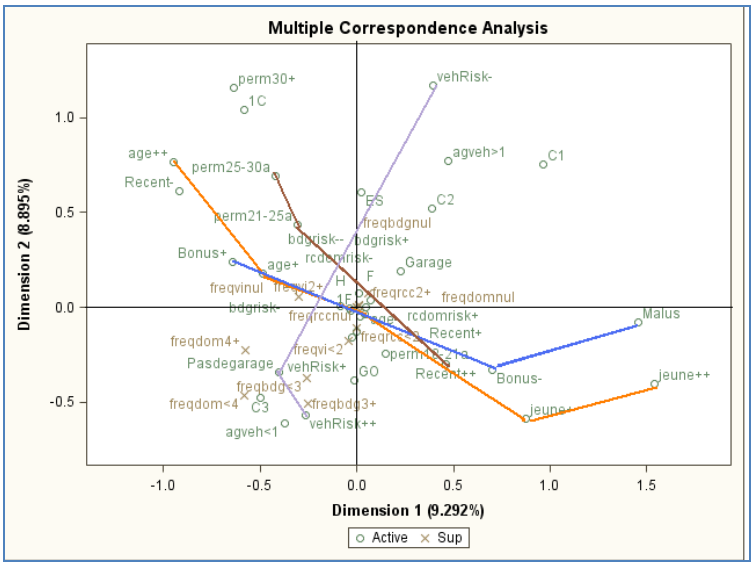


Figure I-2: Projection des données sur le plan 1-2

Les personnes très âgées (âge++) ont plus souvent obtenu leur permis à plus de 30 ans, leurs contrats sont anciens (récent -). Les jeunes (jeune ++ et jeune+) ont eu leur permis entre 18 et 21 ans, les contrats sont très récents (récent++). Il se dessine là une dépendance entre l'âge d'obtention du permis et l'âge du conducteur.

Certains résultats « logiques » sont mis en évidence : les personnes qui possèdent un véhicule de plus de 10 ans (ageveh>10) choisissent des formules dites RC, tandis que les personnes qui ont un véhicule plus récent choisissent naturellement des formules dites « tous risques ». Il se dessine là, une forte dépendance entre la variable formule et l'âge du véhicule.

En conclusion, l'ACM nous a permis de vérifier la fiabilité de nos données, nous ne retrouvons pas de conclusion aberrante concernant celles-ci. Elle nous a aussi permis de mieux caractériser les informations et les profils de risques contenus dans le portefeuille Automobile 4RP (l'âge du conducteur, sa profession, son ancienneté en portefeuille, etc. Pour des raisons de confidentialité, nous ne pouvons pas donner plus de détails).

I.4. Mesures d'association

Dans la suite, à travers les mesures d'association, nous voulons mesurer les dépendances entre variables, afin de détecter d'éventuels effets croisés. L'intérêt des mesures d'association est de fournir un indicateur numérique de liaison entre variables qualitatives prises deux à deux. Leur utilisation va ainsi nous permettre de quantifier puis d'ordonner les différentes associations entre les variables. Ainsi pour calculer ces mesures d'association, nous allons catégoriser les variables continues en classes équi-distribuées.

I.4.1. Rappels théoriques :

Définitions

On dit qu'il y a association (ou corrélation) entre 2 variables si la distribution des observations par modalités d'une variable diffère selon les modalités de la deuxième variable.

Une **mesure d'association** indique avec quelle force deux variables sont reliées entre elles sur la base de l'échantillon étudié.

Les mesures d'association se calculent à partir du tableau de contingence de deux variables qualitatives.

Soient deux variables X et Y,

x_1 à x_r les modalités de X

y_1 à y_k les modalités de Y.

N_{ij} le nombre d'occurrences communes aux deux modalités x_i et y_j.

Le tableau de contingence de X et Y est sous la forme :

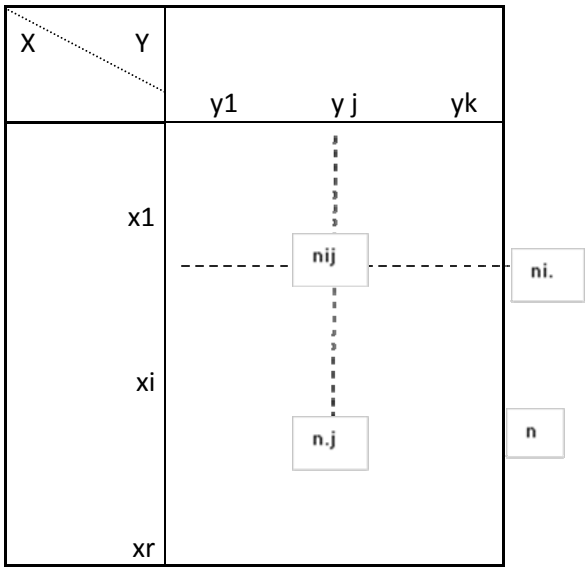


Figure I-3: Illustration d'un tableau de contingence

Nous allons nous intéresser aux mesures basées sur la statistique de χ^2.

L'intérêt de la statistique du χ^2 réside dans le fait qu'elle permet de tester l'indépendance entre deux variables qualitatives X et Y en mesurant, à partir du tableau de contingence, l'écart entre les valeurs observées et les valeurs attendues en cas d'indépendance.

Soit d² une statistique qui s'écrit :

$$d^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Sous Ho : X et Y sont indépendants, la statistique d^2 suit une loi du χ^2 à (r-1)*(k-1) degrés de libertés, où

- r et k sont respectivement le nombre de modalités de X et Y.
- n_{ij} le nombre d'occurrences communes aux deux modalités x_i et y_j.
- n le nombre d'observations

Le test de niveau α consiste à rejeter Ho lorsque $d^2 > \chi^2_{(r-1)*(k-1)}(1-\alpha)$ où $\chi^2_{(r-1)*(k-1)}(1-\alpha)$ est le quantile d'ordre 1- α d'une loi du χ^2 à (r-1)*(k-1) degrés de libertés.

L'inconvénient majeur du d^2 est que les liaisons entre les différentes variables ne peuvent pas être ordonnées par simple comparaison des d^2 car cette statistique dépend du nombre d'observations et du nombre de modalités des variables croisées. Ainsi un d^2 plus élevé ne traduit pas forcément une liaison plus importante, cela peut être le simple fait d'un nombre d'observations ou d'un nombre de modalités plus élevé.

Pour remédier à ce problème, diverses mesures d'association ont été proposées afin d'obtenir une valeur comprise entre 0 (indépendance) et 1(liaison fonctionnelle) : le coefficient de corrélation de Pearson, le coefficient de Spearman pour les variables ordinales et le V de Cramer.

Ce dernier est le plus utilisé, il a l'avantage d'annuler l'effet du nombre de modalités.

$$V\ de\ Cramer = \frac{d^2}{\sqrt{n(k-1)}}$$

Où k est le plus petit nombre de modalités des deux variables mises en relation. Nous nous baserons donc principalement sur cet indicateur.

I.4.2. Analyse des résultats obtenus sur les portefeuilles observés

Le test du χ^2 rejette très fortement l'indépendance des variables, et ce quel que soit le niveau de test qu'on s'impose puisque la p-value est toujours égale à 0.0001.

Bien que toutes les variables soient plus ou moins corrélées les unes aux autres du fait de leur nature, certaines le sont plus que d'autres. Le V de Cramer permet de mesurer le degré d'association des variables et d'établir un « classement » des liaisons. Une valeur du V de Cramer proche de 1 montre une liaison forte entre les deux variables observées. Plus la valeur est proche de 0, plus les variables sont peu corrélées l'une de l'autre.

Les résultats que nous interprétons sont les suivants :

Variable 1	Variable 2	V de cramer
Formule	Age du véhicule	0,63829223
Age du conducteur	Age du permis	0,278408

Tableau 3 : Mesure de V de Cramer interprétées

Concernant les variables continues, nous faisons deux remarques :

✓ Le V de Cramer pour la formule et l'âge du véhicule est supérieure à 0,5. Cela confirme l'observation faite de dépendance entre ces deux variables.

Cependant fondamentalement, on ne saurait associer l'influence de l'âge du véhicule observée sur la sinistralité à la seule influence de la formule choisie, au contraire.

✓ La dépendance forte entre l'âge d'obtention du permis et l'âge du conducteur n'est pas confirmée par les mesures d'association (V de Cramer faible).

I.5. Analyse de l'influence des variables continues sur la sinistralité observée

Une analyse uni-variée des fréquences de sinistres permet de se faire une première idée de l'influence des variables continues sur la sinistralité indépendamment des autres variables. Pour des raisons de confidentialité, nous ne dévoilerons pas la vraie distribution de notre portefeuille.

I.5.1. Variables concernant les assurés eux-mêmes

- **Age du conducteur**

La moyenne d'âge est comprise entre 40 et 50 ans (donnée confidentielle). Nous supprimons les personnes morales et les âges « erronés » pour la cohérence de nos études.

- Données atypiques

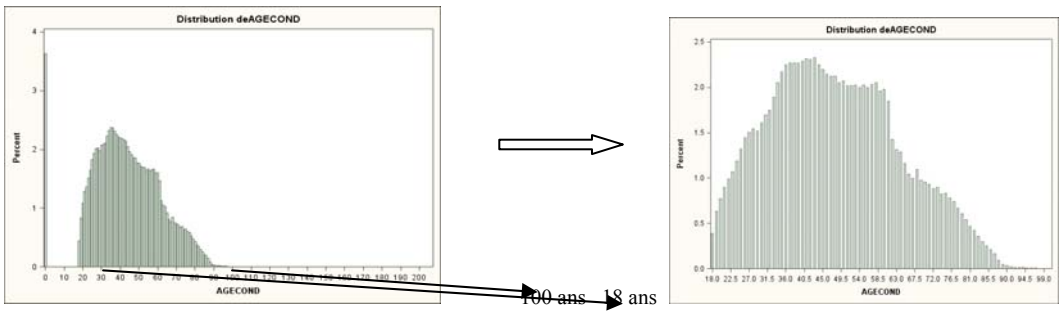
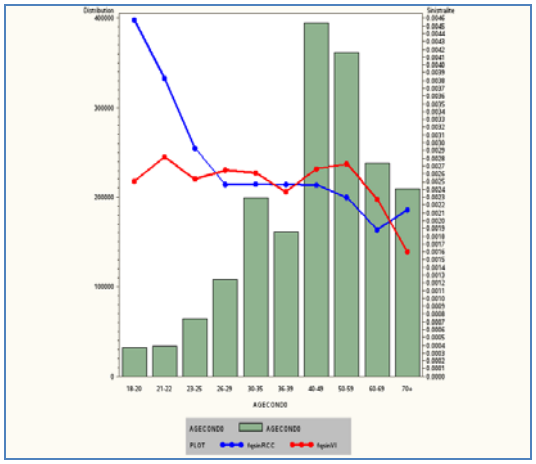


Figure I-4 : Distribution avant et après suppression des données atypiques : Age du conducteur

- Sinistralité observée

Sur les figure I-2 et I-3, nous avons regroupé les modalités d'âge du conducteur en 10 tranches et nous avons observé les nuages **(tranche d'âge du conducteur *(en années)*; fréquence de sinistre moyenne observée*(en ‰)*)** pour les différentes garanties : bris de glace (BDG), dommages (DOM), vol/incendie (VI) et responsabilité civile corporels (RCC). Nous séparons les garanties BDG, DOM d'une part et les garanties RCC et VI d'autre part à cause des échelles différentes.



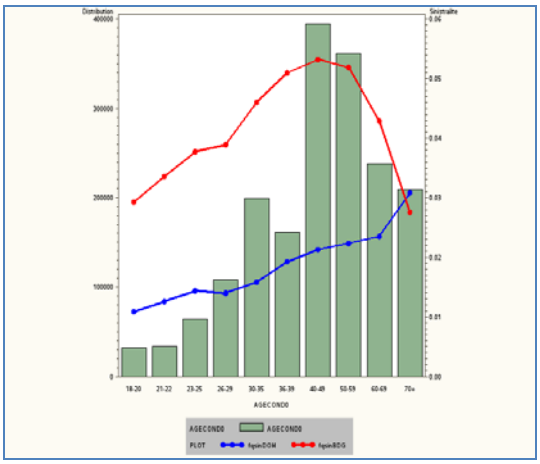


Figure I-5 : Fréquence de sinistres RCC et VI en fonction de l'âge du conducteur

Figure I-6 : Fréquence de sinistres DOM et BDG en fonction de l'âge du conducteur

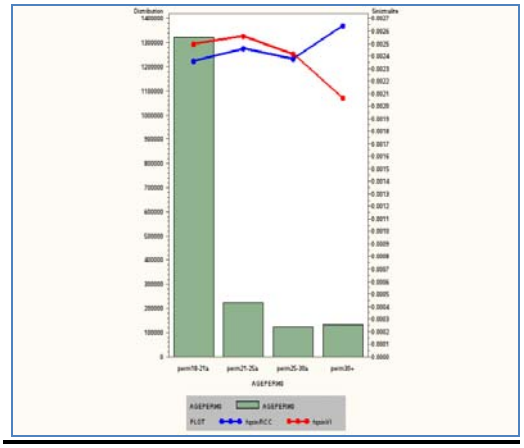
- ✓ Les jeunes conducteurs ont une sinistralité élevée pour la garantie RC corporels. Les jeunes conducteurs (18 - 20 ans) ont une fréquence de sinistres 2 fois supérieure à la moyenne pour cette garantie.
- ✓ L'influence de l'âge du conducteur sur les fréquences de sinistres diffère selon la garantie qu'on considère.
- ✓ On observe une influence forte de prime abord injustifiée de l'âge du conducteur sur la fréquence BDG, cette influence est contraire à celle sur la garantie RCC, les jeunes sont peu risqués.

✓ Le regroupement de ces variables est en général fait de la même manière pour toutes les garanties. Avec la GAM, nous allons construire garantie par garantie de nouveaux regroupements pour cette variable.

➢ **Age d'obtention du permis**

Nous observons l'influence de l'âge d'obtention du permis sur la fréquence de sinistres observée après avoir découpé les modalités en 4 classes d'âge d'obtention du permis :

• Sinistralité observée



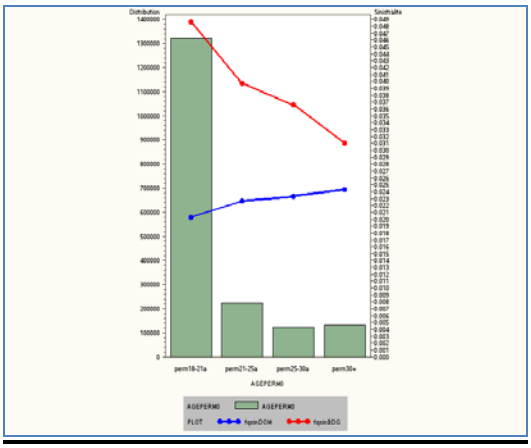


Figure I-7 : Fréquence de sinistres RCC et VI en fonction de l'âge d'obtention du permis

Figure I-8 : Fréquence de sinistres DOM et BDG en fonction de l'âge d'obtention du permis

✓ L'influence de l'âge d'obtention du permis sur les fréquences de sinistres diffère selon la garantie qu'on considère.

✓ Ici aussi, nous observons une décroissance forte de la fréquence de sinistres en fonction de l'âge d'obtention du permis.

I.5.2. Variables concernant les véhicules assurés

➢ **Age du véhicule**

La distribution des âges de véhicules est proche de celle du marché. Nous retirons là aussi des valeurs atypiques (véhicules de collection par exemple).

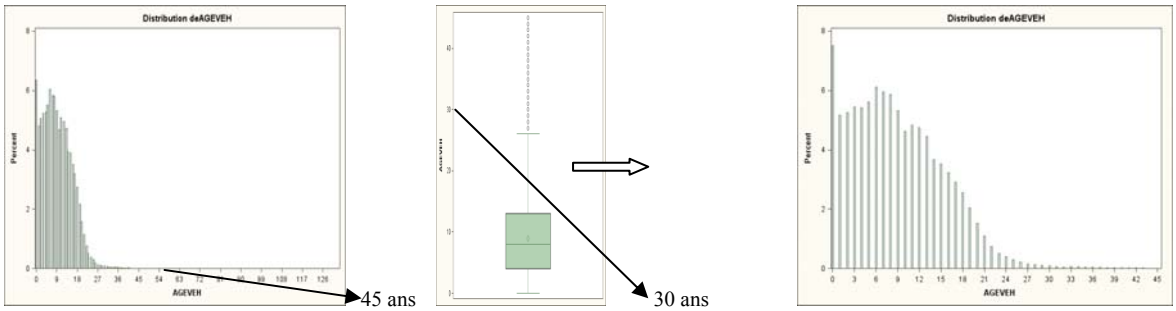


Figure I-9 : Distribution « âge du véhicule » avant suppression des données atypiques

Figure I-10 : Boîte à moustaches « âge du véhicule » avant suppression des données atypiques

Figure I-11 : Distribution « âge du véhicule » après suppression des données atypiques

Sur les graphes suivants, nous avons regroupé les modalités d'âge du véhicule en 7 tranches et nous avons observé les nuages **(tranche d'âge du véhicule** ***(en années)*****; fréquence de sinistre moyenne observée*****(en ‰)*****)** pour les différentes garanties : BDG, DOM, VI et RCC. Nous séparons les garanties; BDG, DOM d'une part et RCC et VI d'autre part à cause des échelles différentes.

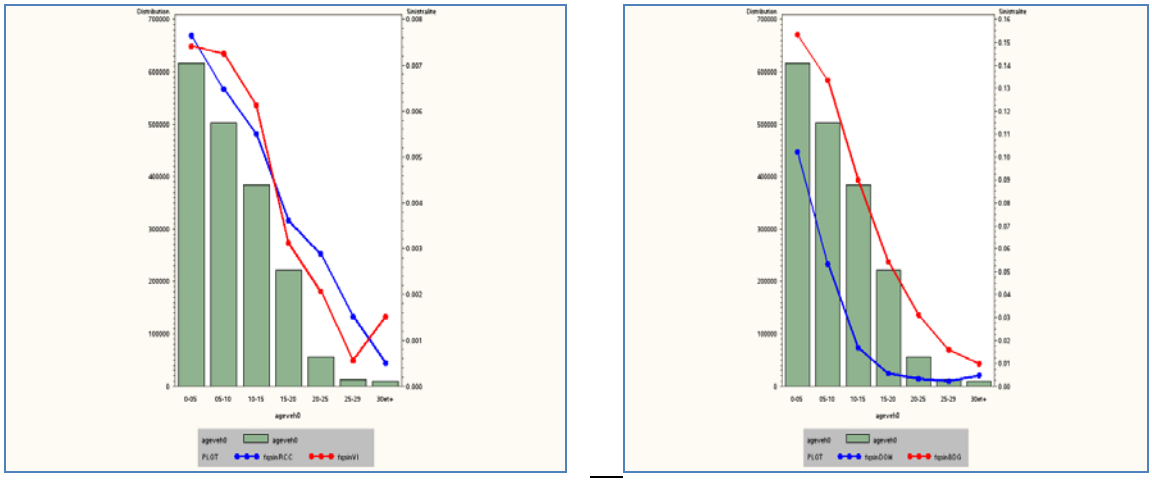


Figure I-12 : Fréquence de sinistres RCC et VI en fonction de l'âge du véhicule

Figure I-13 : Fréquence de sinistres DOM et BDG en fonction de l'âge du véhicule

- ✓ Le critère âge du véhicule est une variable qui s'avère très discriminante de la fréquence de sinistres et ceci pour toutes les garanties observées. La fréquence de sinistres décroit fortement avec l'âge du véhicule.

I.5.3. Variables concernant les contrats

➢ **L'ancienneté du contrat**

Les données relatives à l'ancienneté du contrat sont confidentielles. Les graphiques donnés ici le sont à titre illustratif mais ont été revus en conséquence. Nous avons retiré les données rares comme un âge de contrat supérieur à 50 ans pour les besoins de l'étude.

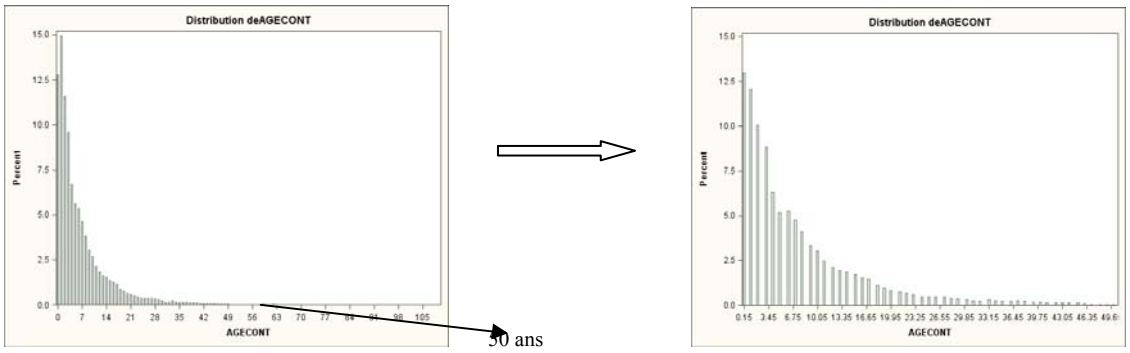


Figure I-14 : Distribution « âge du contrat » avant suppression des données atypiques

Figure I-15 : Distribution « âge du contrat » après suppression des données atypiques

Sur les graphes I-15 et I-16, nous avons découpé les modalités d'ancienneté du contrat en 6 tranches et nous avons observé les nuages **(tranche d'ancienneté** ***(en années)*****; fréquence de sinistre moyenne observée*****(en ‰)*****)** pour les différentes garanties : BDG, DOM, VI et RCC. Nous séparons les garanties; BDG, DOM d'une part et RCC et VI d'autre part à cause des échelles différentes.

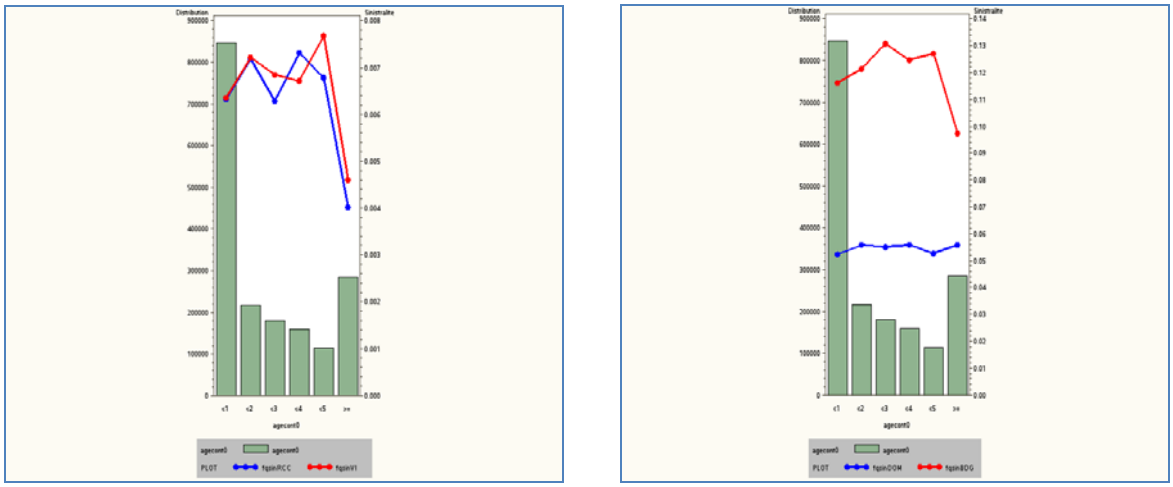


Figure I-16 : Fréquence de sinistres RCC et VI en fonction de l'ancienneté du contrat

Figure I-17 : Fréquence de sinistres DOM et BDG en fonction de l'ancienneté du contrat

La majeure partie des contrats sont très récents. Il n'apparaît pas de forme particulière de l'influence de l'âge du contrat sur la fréquence de sinistres. On observe néanmoins une forte décroissance de la fréquence de sinistres pour les contrats de plus de 5 ans et ceci sur les garanties RCC, VI et BDG.

En conclusion, l'analyse des données nous a permis de retirer des données qui pourraient nuire à la lecture de nos résultats. L'ACM nous a permis de nous faire une idée de la structure du portefeuille étudié en observant les liens entre les variables observées, les mesures d'associations ont permis de quantifier ces dépendances. Enfin, nous avons observé une influence forte des variables âge du conducteur, ancienneté du permis et âge du véhicule et ancienneté du contrat sur la fréquence de sinistres observée et ceci sur toutes les garanties

CHAPITRE II. DU MODELE LINEAIRE GENERALISE AU MODELE ADDITIF GENERALISE

Dans ce chapitre, en présentant la structure des modèles linéaires généralisés et ses différentes composantes, nous introduirons l'intérêt des modèles additifs généralisés pour le regroupement des modalités des variables continues dans un modèle d'estimation. Nous présenterons ensuite les éléments de théorie du modèle additif généralisé, en insistant sur les éléments de ce modèle qui diffèrent du modèle linéaire généralisé.

II.1. ELEMENTS DE THEORIE DES MODELES LINEAIRES GENERALISES

II.1.1. Introduction

Les modèles linéaires généralisés sont basés sur le principe de la régression linéaire. Dans le modèle de régression linéaire, la variable à expliquer $Y=(Y_1,...,Y_n)$ est exprimée comme une fonction affine des variables explicatives X

$$\boxed{\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}} \quad Avec\ Y_i \approx Nor(\eta_i, \sigma^2)$$

Ce modèle dit modèle linéaire gaussien s'est trouvé désuet à cause de la complexité des problèmes statistiques qui se posent à l'Actuaire. En effet l'hypothèse de normalité est peu conciliable avec la réalité de grand nombre de phénomènes aléatoires observés Les modèles linéaires généralisés, introduits par Nelder et Wedderburn en 1972, permettent de **s'affranchir de l'hypothèse de normalité de Y_i**. Celle- ci est remplacée par l'hypothèse que **la variable à expliquer** suit une loi de probabilité qui appartient à **la famille exponentielle**.

Ainsi, les modèles linéaires généralisés permettent l'analyse d'un grand nombre de phénomènes. De plus, ils sont relativement **faciles à estimer**, **à interpréter** et à **représenter** à cause de l'hypothèse de linéarité de la relation entre la variable réponse et les variables explicatives. De ce fait, avec l'inclusion, dans les logiciels de statistiques les plus usités, de procédures permettant d'appliquer cette technique (GENMOD sous Sas, en l'occurrence), on a observé une percée de l'utilisation de ces méthodes. Ces modèles sont à présent un outil de choix pour l'estimation de plusieurs phénomènes en assurance automobile.

- L'estimation du nombre de sinistres espéré sur une police
- L'estimation du coût de sinistres espéré sur une police
- Les probabilités d'entrées et de sorties du portefeuille etc.

Présentons ici les éléments de théorie des modèles linéaires généralisés en insistant sur l'hypothèse de linéarité faite par ces modèles.

Les modèles linéaires généralisés (GLM – Generalized Linear Models) peuvent être caractérisés par les 3 composantes suivantes.

▪ Une composante aléatoire : Il s'agit de variable de réponse $Y = (Y_1,...,Y_n)$, n-uplet de **variables aléatoires** indépendantes non identiquement distribuées et de loi de densité appartenant à la famille des lois exponentielles.

▪ Une composante déterministe : Nous disposons d'un p-uplet (X_1…. X_p) de variables explicatives auxquelles nous associons un p-uplet de paramètres réels (β_0, β_1,..., β_p). La combinaison linéaire des X_i qui en découle définit le prédicteur**.**

▪ La fonction de lien décrit la relation fonctionnelle entre la combinaison linéaire des variables explicatives et l'espérance mathématique de la variable de réponse Y_i. Elle est strictement monotone et différentiable.

Le modèle s'écrit pour un risque donné i :

$$\begin{aligned} &Y_i \approx L_{\exp}(\mu_i) \quad où\ \mu_i = \mathrm{E}(Y_i) \\ &\eta_i = g(\mu_i) \\ &\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \end{aligned}$$

▪ Yi, variable aléatoire

▪ $L_{\exp}$, une loi de la famille exponentielle

▪ μ_i, espérance de Yi,

▪ η_i, prédicteur,

▪ g, fonction de lien

▪ β_j paramètre à déterminer.

Les modèles linéaires généralisés sont utilisés ici pour la modélisation du nombre de sinistres. La construction de tout modèle passe par le choix de chacune de ces composantes.

II.1.2. <u>Composante aléatoire Y</u>

C'est la variable à expliquer : par exemple le nombre de sinistres. En observant la quantité à estimer, on fait une hypothèse sur la loi de distribution à associer cette quantité. Le modèle linéaire généralisé fait l'hypothèse que cette loi de probabilité appartient à la famille **exponentielle,** c'est-à-dire que la loi de densité est de la forme :

$$f_{Y_i}(y_i/\theta,\varphi)=\exp\left\{\frac{\theta y_i-b_i(\theta)}{a_i(\varphi)}+c(y_i,\varphi)\right\}$$

- θ, paramètre canonique, fonction de l'espérance mathématique de Y
- φ, paramètre de dispersion (la variance pour la loi normale par exemple)
- a_i , fonction définie sur les réels et non nulle,
- b_i , fonction définie sur les réels et non nulle,
- c_i , fonction définie sur R².

Puisque les $(Yi)_i$ sont tous indépendants, la densité du vecteur f_Y du vecteur Y est donnée par :

$$f_Y(y/\theta,\varphi)=f_{(Y_1\ldots,Y_n)}(y_1,\ldots,y_n/\theta,\varphi)=\exp\left\{\sum_{i=1}^{n}\frac{\theta y_i-b_i(\theta)}{a_i(\varphi)}+\sum_{i=1}^{n}c(y_i,\varphi)\right\}$$

Parmi les lois de distribution appartenant à la famille exponentielle on peut citer :

- La loi normale
- La loi exponentielle
- La loi Gamma
- La loi binomiale
- La loi géométrique
- La loi binomiale négative
- La loi de Poisson

Pour l'estimation du nombre de sinistres en assurance automobile, la loi qui est utilisée en général est la loi de Poisson :

Soit N le nombre de sinistres enregistrés sur un risque assuré. On fait l'hypothèse que N suit une loi de Poisson de paramètre m. La probabilité de l'évènement {N= k} est donné par par :

$$P(N=k)=\frac{m^k}{k!}e^{-k}$$

La justification de l'utilisation de cette loi pour modéliser un tel phénomène est donnée en annexe

La deuxième composante à choisir est la fonction de lien. Elle et qui est très liée au choix de la loi à associer à la variable à expliquer.

II.1.3. Fonction de lien

La fonction de lien notée g détermine la relation entre l'espérance mathématique d'Y, μ et le prédicteur linéaire η.

$$\eta = g(\mu) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \quad où \quad \mu = \mathrm{E}(Y)$$

Dans de nombreux cas les modèles linéaires généralisés sont construits en utilisant une fonction de lien canonique, associée à une loi de probabilité donnée.

La fonction de lien canonique est définie par :

Soit $\theta = (b')^{-1}(E(Yi))$, il existe une fonction g, telle que $\theta = g(\mu_i)$; g est la fonction inverse de b' ; c'est la fonction de lien associée à la loi de Y_i.

Le tableau 8 donne les fonctions de lien pour différentes lois de probabilité de la famille exponentielle.

Loi de probabilité	Fonction de lien canonique
Normale	g= μ
Poisson	g=lnμ
Gamma	g=1/μ
Binomiale	g=lnμ-ln(1-μ)

Tableau II-1 : Liens canoniques associés aux lois de probabilité usuelles

La troisième composante d'un GLM est la composante déterministe ou prédicteur linéaire. Résoudre un GLM revient à déterminer entièrement le prédicteur linéaire qui explique une quantité donnée Y.

II.1.4. Composante déterministe : Introduction au modèle additif généralisé

Une fois que la loi à associer aux données et la fonction de lien à utiliser est choisie, l'analyse de la variance permet de sélectionner les variables expliquent le mieux la sinistralité d'une garantie donnée. L'analyse de la variance recouvre un ensemble de tests et d'estimation destinées à apprécier l'effet de variables qualitatives sur une variable numérique. Par exemple, pour chacune des garanties, nous retenons grâce aux tests des p-uplets $(X_1 \ldots X_p)$ qui expliquent la fréquence de sinistres. Pour des raisons de confidentialité, nous ne dévoilerons pas leur constitution exacte.

Une fois les p variables sélectionnées, il reste à estimer p-uplet de paramètres réels $\hat{\beta}_1, \cdots, \beta_p$ et la constante $\hat{\beta}_0$ à leur associer pour déterminer une estimation du prédicteur $\beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. Il est dit linéaire. C'est l'hypothèse très forte du modèle GLM ; chaque variable explicative a une relation linéaire sur la variable à expliquer. Un modèle de ce type est dit **linéaire et paramétrique.**

Il arrive que l'on soupçonne une forme non linéaire dans l'influence d'une (ou plusieurs) variable quantitative sur la réponse. Dans ce cas, deux types de démarches peuvent être envisagés.

✓ **Un modèle non linéaire et paramétrique**

Il s'agit de substituer à la variable elle-même X_i différentes transformations de la variable. Le prédicteur est une combinaison linéaire de fonctions (de variables) qui ont **une expression algébrique i.e. dépendant d'un ensemble de paramètres** (forme polynômiale ou trigonométrique etc.). La forme de la transformation est choisie en observant les courbes individuelles de variables. Cela permet de capturer une influence non linéaire.

Par exemple, on peut avoir un modèle de la forme :

$$\eta_i = \alpha + \beta_1 X_i^3 + \beta_2 \cos(X_2)$$

Plusieurs études ont concerné le choix des transformations en fonction de la forme des courbes individuelles observées. (cf. Premiers pas en régression sous Sas, Josiane Confais, Monique Leguen). Celles-ci restent cependant, très intuitives.

✓ **Un modèle non linéaire et non paramétrique**

Le prédicteur est constitué de fonctions (de variables) partiellement définies sans paramètres accessibles au calcul analytique ; construites sur la base d'un ou de plusieurs algorithmes. Le modèle peut être de la forme :

$$\eta_i = \alpha + f(X_{i1}, \cdots, X_{ip}) \text{ ou}$$

$$\eta_i = \alpha + \sum_{j=1}^{k} f_j(X_{ij})$$

Le modèle additif généralisé rentre dans ce deuxième cas (additif), le caractère additif de ce modèle diminue la complexité des calculs f fonctions sont à estimer et non une fonction en dimension p. Ainsi, il permet la modélisation des mêmes phénomènes que le modèle linéaire généralisé sans faire d'hypothèse sur la forme de la relation entre les variables et la réponse. Il s'agit ici d'estimer le vecteur $\underline{f} = (f_1, \cdots, f_k)$ à l'aide de techniques de lissage

Les fonctions f_j sont estimées par des fonctions de lissages. Les techniques de lissage nécessitent beaucoup de calculs. Elles deviennent très intéressantes et accessibles aujourd'hui grâce à de la performance croissante des ordinateurs. C'est une des techniques modernes les plus attractives puisqu'elle ne nécessite pas de préciser la forme d'un modèle, elle laisse « **parler les données** ». Cependant, cette méthode nécessite la continuité des variables.

Présentons quelques éléments de théorie du modèle additif généralisé, nous discuterons ensuite de l'intérêt de ces modèles pour la tarification en particulier.

II.2. Du modèle linéaire généralisé au modèle additif généralisé

Le modèle additif généralisé est une extension du modèle additif et du GLM. Comme le GLM, il permet l'ajustement de la variable à expliquer aux lois de la famille exponentielle. Comme le modèle additif, il ne suppose pas que la relation entre les variables explicatives et la réponse est linéaire. Le prédicteur d'un GAM est dit additif ; il est composé d'une somme de fonctions qui ne sont pas nécessairement paramétriques.

II.2.1. Modèle additif

II.2.1.1. Introduction

Le modèle additif est né de l'inconnu autour de la relation réelle entre les variables explicatives et la variable à expliquer, ainsi, un modèle additif s'écrit :

$$\eta_i = \alpha + \sum_{j=1}^{p} f_j(X_{ij})$$

Avec :

- α constante,
- $Y_i \approx \text{Nor}(\eta_i, \sigma^2)$
- les $f_j(.)$ à estimer sont supposées régulières et traduisent l'influence de chaque variable X_j sur la variable réponse Y_i.

Pour chaque variable X_j, la fonction f_j est à estimer à partir des données, soit l'observation d'un certain nombre de couples entrée-sortie $\{(x_{ij}, y_i): i=1,\ldots, n\}$ où n est la taille du portefeuille.

En considérant une seule variable explicative X, regardons comment il est possible d'estimer f à l'aide de techniques de lissage.

Nous nous intéressons particulièrement aux lisseurs dits linéaires en ce sens que :

Si on définit le vecteur $f = (f_0(x_0),..., f_n(x_n))^t$ $\hat{f}$ se réécrit $\hat{f} = S_\lambda y$ où S_λ est appelée matrice de lissage associée à l'estimation de $\hat{f}$.

En d'autres termes, $\hat{f}(x_0)$ peut s'exprimer comme combinaison linéaire des valeurs y_1, y_2, …,y_n

$\hat{f}(x_0) = \sum_{j=1}^{n} s_{0j} y_j$ Les poids s_{0j} dépendent de l'endroit x_i où la réponse $f(x_i)$ doit être estimée.

Soit un nombre d'observations (x_i, y_i) i=1,…,n où

- la variable à expliquer est y_i
- x_i et y_i sont continus.

L'influence de x_i sous y_i est modélisée à l'aide d'une fonction f(.) : $y_i = f(x_i) + \varepsilon_i$

- Où f est supposée régulière
- les ε_i sont supposés indépendantes avec E (ε_i) = 0 et var (ε_i) = σi^2 $\forall i \in [1,n]$.

Le lissage « fabrique », pour une valeur de x_i donnée, x_0 par exemple, des valeurs $\hat{f}(x_0)$ proches des vraies valeurs.

Les fonctions de lissage « linéaires » et non paramétrique les plus utilisées sont les splines de lissage cubiques et les fonctions lœss. Des méthodes de lissage sont associées à celles-ci. Nous les présentons dans les paragraphes suivants.

II.2.1.2. Maximum de vraisemblance pénalisée et splines de lissage cubiques

Cette estimation utilise le principe des moindres carrés pénalisés qui consiste à minimiser la fonction PML définie par :

$$PML(f) = \sum (y_i - f(x_i))^2 + \lambda \int_a^b [f''(t)]^2 dt$$

Hypothèses :

- f , deux fois continûment différentiable,
- f'' , de carré intégrable,
- λ , un paramètre,
- a et b réels tels que : $a \leq X_1 \leq X_2 \leq \ldots \leq X_n \leq b$.

Le terme $\sum (y_i - f(x_i))^2$ assure que f(.) ajustera au mieux les données, on peut le voir comme une log-vraisemblance tandis que le terme $\lambda \int_a^b [f''(t)]^2 dt$ pénalise l'irrégularité de l'estimateur avant de le maximiser. λ est appelé paramètre de lissage. Cette méthode est appelée approche PML (Penalized maximum log-likehood) et remonte aux travaux de l'actuaire E. Whittaker qui l'utilisa pour lisser des tables de mortalité. Pour plus de détails, on pourra se référer à l'ouvrage de Hastie &Tibshirani [2].

On parle de splines de lissage car la solution de cette minimisation est un spline (régression polynômiale par morceaux) cubique dont les nœuds sont $x_1,...,x_n$. Ce qui signifie que $\hat{f}$ coïncide avec un polynôme du 3éme degré sur chaque intervalle (x_i, x_{i+1}) et possède **des dérivées première et seconde continues** en chacun des xi.

Pour des grandes valeurs de λ, l'estimateur résultant de la minimisation de PML(f) présentera une courbure très faible tandis que lorsque λ tend vers 0, la pénalisation disparaît et on obtient une interpolation parfaite.

II.2.1.3. Régression locale pondérée : fonctions læss

Cette méthode consiste à approximer **localement** f(.) par **une droite**. L'idée est d'utilise les λ [4] plus proches voisins de x pour estimer f(x). Le voisinage s'apprécie par rapport aux variables explicatives : on utilise les λ observations dont les variables explicatives sont les plus proches de x pour estimer la réponse f(x). λ est le paramètre de lissage

Les fonctions de régression locale pondérée (ou fonctions lœss) remplacent un point (x_0, y_0) par une régression linéaire sur les points (x_i, y_i) du voisinage de (x_0, y_0), affectés d'une pondération dépendant de la distance $|x_i - x_0|$. Les fonctions lœss réalisent ainsi un lissage non paramétrique déterminé par l'étendue du voisinage de points participant aux régressions locales. On pourra se référer au livre « Mathématiques de l'assurance non-vie, Tome II » d'Arthur Charpentier et Michel Denuit [1] pour plus de détails.

Du fait de ces fonctions de lissage à estimer dans un modèle GAM, il découle de nouveaux paramètres à prendre en compte dans la résolution d'un modèle GAM (par rapport à un modèle GLM).

Ce sont entre autres :

- Le nombre de degrés de libertés
- Le paramètre de lissage

II.2.1.4. Nombre de degrés de libertés

Pour un GLM, le nombre de degrés de liberté (ddl ou df – degree of freedom) est clairement la différence entre le nombre d'observations et le nombre de paramètres à estimer.

Pour un modèle additif, le calcul est plus complexe. Il faut mesurer la complexité de l'estimation des fonctions de lissage. Pour une fonction de lissage linéaire, soit S_λ la matrice de lissage,

[4] Communément le paramètre de lissage est désigné par λ, il n'a pas ici la même signification que dans le paragraphe précédent

$$S_\lambda = \begin{bmatrix} a_{11} & & a_{1n} \\ & a_{22} & \\ & \cdots & \\ a_{n1} & & a_{nn} \end{bmatrix}$$

La trace de S_λ est donnée par $Tr(S_\lambda) = \sum_{i=1}^{n} a_{ii}$

Pour un spline de lissage, le nombre de degrés de libertés et donné par :

- $DF(spline\ \ de\ lissage) = Tr(S_\lambda)$

Pour une fonction lœss, il est défini par

- $DF(loess) = Tr(S_\lambda S_\lambda^{'})$

Ce paramètre permet de comparer le niveau de complexité de deux modèles d'estimation.

II.2.1.5. <u>Degré de lissage λ</u>

Le degré de pénalité du modèle est mesuré par λ

- Quand λ tend vers 0, il n'y a pas de pénalité de lissage et le modèle fournit un ajustement parfait : les valeurs ajustées sont les données elles-mêmes.

- Quand λ, tend vers l'infini, l'ajustement est un lissage « parfait » : les valeurs ajustées tombent le long d'une ligne droite forçant ainsi la liaison à être linéaire. Il en résulte une faible variance mais un biais important ; une erreur de lissage globale faible mais des erreurs d'estimation locales fortes. Par exemple, pour la méthode loess, cela correspond au fait que des valeurs relativement éloignées de x_0 ont été utilisées pour estimer $f(x_0)$ du fait d'une fenêtre trop grande λ.

- Le paramètre λ pour chaque lisseur doit être compris entre ces deux cas extrêmes pour produire le niveau désiré de lissage qui permet d'avoir un bon ajustement du modèle avec un équilibre entre une différence peu importante entre les valeurs estimées et les valeurs observées.

Le choix de ces paramètres de lissage peut se faire à l'aide de la méthode par validation croisée. Une valeur particulière pour λ peut également être choisie en fixant soi-même le nombre de degrés de liberté.

La validation croisée CV (λ) pour le choix de λ consiste à minimiser le critère CV (λ) défini par :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{\lambda}^{-i}(x_i))^2$$

Où $\hat{f}_{\lambda}^{-i}$ est l'estimation de $f(x_i)$ obtenue à l'aide de l'échantillon {(y_i, xi) j différent de i} de taille n-1.

On a aussi parfois recours au critère de validation croisée généralisé GCV (λ) :

Ici, la fonction GCV à minimiser est alors :

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(y_{ij} - \hat{\eta}_{ij})^2}{(n - tr(S_{\lambda}))^2}$$

II.2.2. Modèles additifs généralisés

Comme pour les modèles linéaires généralisés, les modèles additifs généralisés permettent de **s'affranchir de l'hypothèse de normalité de Y_i**. Ils offrent la possibilité aussi de mettre en œuvre des modèles de régression de Poisson, binomiale et Gamma. Le modèle additif généralisé (GAM – Generalized additive models) s'écrit:

$$\begin{aligned} &Y_i \approx L_{\exp} \quad et \quad \mu_i = \mathrm{E}(Y_i) \\ &\eta_i = g(\mu_i) \\ &\eta_i = \alpha + \sum_{j=1}^{p} f_j(X_{ij}) \end{aligned}$$

- α une constante.
- Y_i, une variable aléatoire
- $L_{exp,}$ une loi de la famille exponentielle,
- μ_i, espérance de Y_i,
- η_i, prédicteur,
- g, fonction de lien, strictement monotone et différentiable

La principale différence entre un GAM et un GLM réside dans le lien entre les variables explicatives et la variable réponse. Utiliser un GLM revient à estimer un vecteur β réel et utiliser un GAM revient à estimer un vecteur F de fonctions.

II.2.3. Utilisation conjointe GLM et GAM : PROC GAM sous SAS

II.2.3.1. Entrées de la procédure GAM sous SAS

PROC GAM <options> ;
CLASS variables<(options)> </options> ;
MODEL dependent = <PARAM(effects)> <smoothing effects> </model options> ;
SCORE data=SAS-data-set out=SAS-data-set ;
OUTPUT <out=SAS-data-set> keyword=prefix < keyword=prefix> ;
BY variables ;
FREQ variable ;
RUN ;

MODEL

Dependent: Variable à expliquer

PARAM (effects): Ce sont les paramètres à estimer pour déterminer la relation entre ces variables et la réponse.

Smoothing effects: Variables continues à lisser à l'aide de lisseurs non-paramétriques spécifiées. Pour chaque variable à lisser, il faut spécifier le lisseur utilisé.

Exemple de spécification du modèle :

Type de modèle	Syntaxe	Forme mathématique
Paramétrique	**y=param(x1 x2)**	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
Non paramétrique	**y=spline(x)**	$E(Y) = \beta_0 + f(x)$
Non paramétrique	**y=loess(x)**	$E(Y) = \beta_0 + f(x)$
Semi-paramétrique	**y=spline(x1) param(x2)**	$E(Y) = \beta_0 + f(x_1) + \beta_2 x_2$
Additive	**y=spline(x1) spline(x2)**	$E(Y) = \beta_0 + f_1(x_1) + f_2(x_2)$

Eléments à préciser dans le modèle

- **La distribution associée à la variable aléatoire-réponse:**

DIST=	Distribution	Type de données -réponse
GAUSSIAN | GAUS | NORM	gaussienne	Variables continues
BINOMIAL | LOGI | BIN	Binômiale	Variables binaires

POISSON | POIS | LOGL	Poisson	Variables non négatives discrètes
GAMMA | GAMM	Gamma	Variables positives discrètes
IGAUSSIAN | IGAU | INVG	Inverse gaussienne	Variables continues positives

▪ **La fonction de lien choisie:**

Pour les distributions précédentes, la fonction de lien canonique est utilisée.

▪ **La méthode de sélection du paramètre de lissage**

OPTION METHOD=GCV spécifie que la méthode de validation croisée générale va être utilisée. Notons que si le nombre de degrés de libertés DF est spécifié par ailleurs, l'option METHOD est ignorée.

▪ **Le seuil de convergence de l'algorithme:**

OPTION EPSILON=NUMBER pour le « backfitting algorithm »

OPTION EPSSCORE=NUMBER pour le « local scoring algorithm »

La procédure SAS intègre deux algorithmes d'estimation des fonctions de lissage que nous avons trouvé intéressant de présenter.

II.2.3.2. Algorithmes de résolution d'un modèle additif généralisé

1. Algorithme d'ajustement arrière ou backfitting algorithm

Considérons l'estimation des fonctions de lissage $f_1(\,),\cdots, f_p(\,)$ dans le modèle suivant :

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

On suppose que ε est indépendant des X_j et tels que $E(\varepsilon) = 0$ et que $var(\varepsilon) = \sigma^2$.

L'algorithme qui permet de faire face à toutes les situations de régression dans le cas d'un modèle additif (simple) est l'algorithme d'ajustement arrière, plus connu sous ne nom de « backfitting algorithm ».

On veut estimer les $f_1(\,),\cdots, f_p(\,)$**.** Les étapes de l'algorithme sont les suivantes :

De l'écriture précédente de Y, on déduit

$$Y - \alpha - \sum_{k=1}^{p} f_{k\neq p}(X_k) = f_j(X_j) + \varepsilon\,,$$

D'où,

$$E\left(Y - \alpha - \sum_{k=1}^{p} f_{k\neq j}(X_k) \middle| X_j \right) = f_j(X_j),$$

Ceci signifie que la valeur de $f_j(X_j)$ ou, ce qui revient au même, les valeurs $\{f_j(X_{1j}), f_j(X_{2j}), \cdots, f_j(X_{nj})\}$ sont exprimables à partir des fonctions des autres variables.

Le principe de l'algorithme est qu'à chaque étape, la valeur prise par la fonction d'une variable donnée est calculée à partir des valeurs prises par les fonctions des autres variables à l'étape précédente.

L'algorithme « d'ajustement arrière » (ou « rétrograde ») repose sur les étapes suivantes :

1°) On choisit des valeurs initiales pour les $f_j(X_j) : \alpha = \sum_{i=1}^{n} \frac{Y_i}{n}$,

En d'autres termes, on prend la moyenne des Y_j pour valeur initiale de α ; pour les autres, on choisit p valeurs de départ: f_1^0, $f_2^0, \cdots, f_p^0$.

Comme on n'a pas d'a priori sur la forme des fonctions, on peut choisir des valeurs déduites d'une régression de Y_i sur les X_j.

2°) On « fabrique » f_1^1 tel que $f_1^1 = S_j(Y - \alpha - \sum_{k=1}^{p} f_{k\neq j}^0(X_j)$,

Ceci veut dire que l'on calcule f_1 au stade 1 à partir des f_k au stade 0 à l'aide d'une fonction de lissage prédéterminée.

3°) On fait la même chose avec $f_2^1, f_3^1, ..., f_p^1$

4°) On recommence pour $f_2^2, f_3^2, ..., f_p^2$ etc.

5°) On s'arrête quand les fonctions ne diffèrent plus trop d'une étape à l'autre. Il est donc nécessaire de définir un seuil.

La procédure GAM de SAS utilise comme critère de convergence pour le « backfitting algorithm » la mesure suivante :

$$\frac{\sum_{j=1}^{n}\sum_{i=1}^{k}\left[f_i^{(m-1)}(X_j) - f_i^{(m)}(X_j)\right]^2}{1 + \sum_{j=1}^{m}\sum_{i=1}^{k}(f_i^{(m-1)}(X_j))^2} < \varepsilon$$

L'algorithme converge la plupart du temps dans le cas des splines de lissage mais dans le cas des fonctions lœss, ceci est moins sûr.

Le local scoring algorithm répond à la problématique de l'ajustement du GAM. Cet algorithme reprend, en le généralisant, celui de la méthode des scores de Fischer. On retrouve l'expression de Taylor mais avec la notion de lissage (toutes les fonctions de lissage peuvent convenir).

2. *Local scoring algorithm*

Rappelons la définition du GAM :

$$\begin{aligned} &Y_i \approx L_{\exp} \quad et \quad \mu_i = \mathrm{E}(Y_i) \\ &\eta_i = g(\mu_i) \\ &\eta_i = \alpha + \sum_{j=1}^{p} f_j(X_{ij}) \end{aligned}$$

Il faut estimer α et les f_j. Les étapes de l'algorithme sont les suivantes.

1°) On démarre l'algorithme en donnant une valeur arbitraire à α et aux fj :

$\alpha^0 = g(\sum_{i=1}^{n} \frac{Y_i}{n})$ *et* $f_1^0 = f_2^0 = ... = f_p^0 = 0$, g fonction de lien du modèle.

Ainsi, $\eta_i^0 = \alpha^0 + \sum_{j=1}^{p} f_j^0(X_{ij})$ et $\mu_i^0 = g^{-1}(\eta_i^0)$ avec, on rappelle, les $f_j^0 = 0$ et g connue.

2°) On construit une nouvelle variable dépendante que l'on pondère

Soit z_i, telle que :

$$z_i = \eta_i^0 + (Y_i - \mu_i^0)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0$$

Soit w_i une pondération de z_i; l'expression de cette pondération dépend de la dérivée de g^{-1} et de l'inverse de la variance de Y_i en μ_i^0.

$$w_i = \left(V_i^0\right)^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0^2$$

Où V_i est la variance de Y_i en μ_i^0.

3°) On ajuste un modèle additif pondéré sur les valeurs de la nouvelle variable dépendante z_i.

L'ajustement se fait par le « backfitting algorithm ». On obtient ainsi une estimation des fonctions f_j^1. Ainsi que la valeur des η_i^1 et des μ_i^1.

On calcule le critère de convergence basé sur la somme relative des valeurs absolues des différences entre les f_j^0 et f_j^1.

4°) On répète 2) et 3) en remplaçant les valeurs du stade 0 par les valeurs du stade 1

5°) L'algorithme opère jusqu'à ce que le critère de convergence soit suffisamment proche de 0. Ceci suppose, là encore, le choix d'un seuil de convergence.

La procédure GAM de SAS utilise comme critère de convergence pour le « local scoring algorithm » la mesure suivante :

$$\frac{\sum_{j=1}^{n}\sum_{i=1}^{k} w(x_i)\left[f_i^{(m-1)}(X_j) - f_i^{(m)}(X_j)\right]^2}{\sum_{j=1}^{m}\sum_{i=1}^{k} w(x_i) + \sum_{j=1}^{m}\sum_{i=1}^{k}\left(f_i^{(m-1)}(X_j)\right)^2} < \varepsilon$$

Avec $\varepsilon=10^{-8}$ par défaut.

II.2.3.3. Lecture des résultats

Les notions statistiques qui permettent de lire les résultats sont expliquées en annexe.

Il s'agit notamment de :

- La notion de quasi-vraisemblance
- La notion de déviance
- Le test de significativité des coefficients de lissage
- Les notions de résidus de Pearson et de résidus de déviance.

- En ce qui concerne les sorties graphiques, il convient d'expliciter leur contenu. La procédure GAM sépare l'effet linéaire d'une variable de l'effet non linéaire de celle-ci, ceci pendant l'estimation mais aussi dans la présentation des résultats finaux. Il est donc facile de déterminer si la signification d'une variable de lissage est associée à une tendance linéaire simple ou un modèle plus complexe.

Par exemple, supposons que l'on veuille ajuster un modèle semi-paramétrique de la forme :

$$\log(y) = \alpha + \alpha_1 z + f_1(x_1) + f_2(x_2)$$

L'estimation du GAM pour ce modèle est : $\log(y) = \hat{\alpha} + \hat{\alpha}_1 z + [\hat{\beta}_1 x_1 + \hat{f}_1(x_1)] + [\hat{\beta}_2 x_2 + \hat{f}_2(x_2)]$

où $\hat{f}_1$ et $\hat{f}_2$ sont des fonctions non paramétriques estimateurs de f_1 et f_2.

- Les p-valeurs pour $\hat{\alpha}_1$, $\hat{\beta}_1$, $\hat{\beta}_2$ sont rapportés dans le tableau estimations des paramètres.

- $\hat{\beta}_1$ *et* $\hat{\beta}_2$ sont les estimations de β_1 *et* β_2. Ils seront marqués linear (x_1) et linear (x_2) dans le tableau.

- La p-valeurs pour $\hat{f}_1$ et $\hat{f}_2$ sont présentés dans le tableau d'analyse de la déviance.

Seules les valeurs de $\hat{f}_1(x_1)$, $\hat{f}_2(x_2)$ et $\hat{\log}(y)$ sont données, elles correspondent à P_x_1, P_x_2 et P_y dans les tables de données estimées. La prédiction de l'influence complète de x_1 sur y, il faut calculer: $\hat{\beta}_1 x_1 + P\_x_1$.

Les courbes d'estimation fournies par les graphes en sortie donnent $\hat{\beta}_1(x_1 \_ \bar{x}_2) + P\_x_1$. SAS fait une transformation linéaire qui ne modifie pas la forme de la fonction initiale. Elle sert à centrer les données autour de la moyenne de x_2.

II.3. GAM comme modèle privilégié d'estimation de la fréquence annuelle en assurance automobile : Non

La procédure GAM permet de construire un modèle semi-paramétrique, on peut donc traiter à la fois les variables dont le découpage est prédéfini (Usage, catégorie socioprofessionnelle etc.) et les variables continues. Cependant cette procédure est plus adaptée pour des problèmes où l'on a peu de variables explicatives car le temps de calcul et d'estimation des fonctions non paramétriques est très long. Lorsqu'on teste un modèle « optimal » au sens du nombre de variables incluses et du découpage des autres variables, la procédure GAM prend un espace mémoire supérieure à la capacité du système. Une erreur « Out of Memory » apparaît dans le journal.

Le document en ligne SAS : http://www.sas.com/offices/europe/france/services/support/download/allo_support_12.pdf met en garde contre ce type d'erreur.

« Memsize » représente la mémoire allouée au processus SAS au moment du démarrage de la session. Par défaut ce paramètre est positionné à 64Mo et peut prendre jusqu'à la valeur de 2Go (limite système). Il est possible aussi de spécifier une valeur nulle et dans ce cas SAS va allouer au fur et à mesure des traitements la mémoire qui lui est nécessaire. Ce dernier cas est recommandé uniquement pour les besoins d'un test, pour un utilisateur donné. L'emploi d'une telle option en production et pour l'ensemble des utilisateurs peut en effet provoquer une consommation trop importante de mémoire. Donc, parfois un message «out of memory» est préférable à une défaillance générale du serveur.

Cette erreur est donc liée à la lourdeur des calculs à effectuer pour cette procédure.

Dans la rubrique : Ressources computationnelles » de l'aide sur la procédure GAM, le besoin en mémoire et en temps de la procédure SAS est calculée en fonction des paramètres d'entrée. Voir Annexe. Notons tout de même qu'au sortir de cette analyse, nous choisissons d'utiliser les fonctions splines pour notre étude.

En tarification, une dizaine de variables sont retenues pour l'estimation de la fréquence ou du coût de sinistres. Chaque variable est divisée en 10 modalités en moyenne. Ce grand nombre de variables à inclure dans le modèle GAM écarte l'hypothèse de remplacer les modèles GLM par ceux-ci. De plus, lorsque l'on associe ces variables à une ou plusieurs variables continues telles quelles, les

segments homogènes deviennent trop petits (du fait du grand nombre de modalités), les hypothèses statistiques faites ne seront plus acceptables. Le modèle sera donc moins fiable et les algorithmes sont susceptibles de diverger. Enfin, pour le tarif, on peut difficilement discriminer les individus âge par âge cela induirait une segmentation infinie.

II.4. Intérêts de la GAM pour la tarification

Idéalement, l'objectif est de se ramener à un modèle paramétrique. En effet, même si les hypothèses prises sont assez fortes dans le cas d'un modèle paramétrique (GLM), celui a plusieurs avantages :

- Facilement interprétable
- Simple
- Représentable

Ainsi, des algorithmes mathématiques très performants permettent d'intégrer un grand nombre de variables tarifaires dans les modèles GLM à condition d'avoir au préalable redéfinies les modalités de chaque variable. Dans le cas du GLM, c'est cette étape préalable qui permet de « personnaliser » au maximum le tarif.

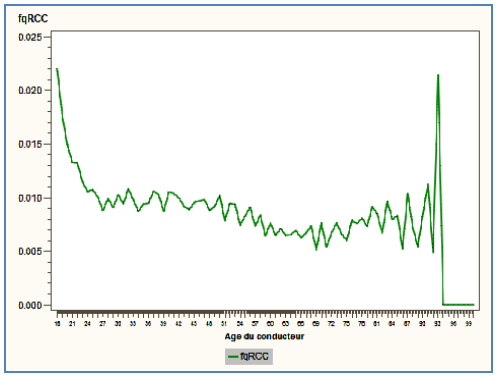
En effet, le GLM calcule une relativité pour chaque modalité de chaque variable. C'est cette quantité qui donne une idée de la part dans la sinistralité du « risque » provenant de la caractéristique décrite par la modalité concernée.

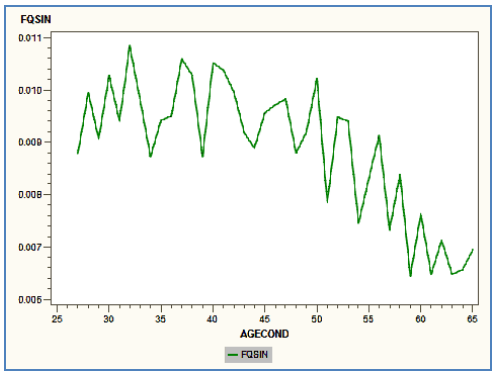
Concrètement, pour un risque j, pour un critère X_j (groupe de risque SRA par exemple) et pour une modalité X_{ij} de ce critère (groupe 25), le GLM fournit la quantité $\exp(\beta_{ij})$. C'est la part du nombre de sinistres N_i attendu sur i qui provient du fait que le véhicule concerné par le risque appartienne au groupe 25. Si le véhicule lié cette police appartient au groupe 25, a 5 ans, que son propriétaire a 22 ans etc. Le produit des quantités $\exp(\beta_{ij})$ correspondant à chacune des informations sur l'assuré constitue l'estimation du nombre de sinistres attendu sur cette police par la GENMOD. Ainsi si les modalités 25 à 30 par exemple ont été regroupées ensemble comme une seule modalité de la variable « groupe SRA » telle qu'elle sera incluse dans le modèle, l'influence de cette variable sur ces polices sera estimée être la même. Le GLM ne remet pas en cause le découpage effectué.

Ainsi pour éviter des sauts de tarif injustifiés techniquement, il convient en amont de la GLM de constituer des classes d'âge de niveaux de risques homogènes. C'est là l'intérêt des modèles additifs généralisés :

Les données brutes sont très erratiques, elles ne permettent pas une interprétation aisée des niveaux de risques des différentes modalités.

Exemple:





II-1 : Données brutes de la fréquence de sinistres RCC en fonction de l'âge du conducteur
II-2: Données brutes de la fréquence de sinistres RCC en fonction de l'âge du conducteur entre 27 et 65 ans

La GAM fournit une estimation basée sur les données observées de la part marginale de chaque variable continue sur la fréquence de sinistres et un intervalle de confiance à 95% de cette estimation.

Les fonctions $\hat{f}_j$ à estimer sur la base des données **lisseront** le lien entre les variables continues X_{i1}, …, X_{il} et chaque variable continue. **La forme additive** (on rappelle : au lieu d'avoir à estimer une fonction en dimension p ($f(X_1, X_2, ..., X_p)$), on estime p fonctions f,*)* **du GAM permet ainsi, d'extraire du modèle l'influence marginale non linéaire de chaque variable continue incluse.**

Nous proposons une méthode essentiellement basée sur l'observation des courbes de lissage fournies par la GAM de classification des variables continues à intégrer dans le tarif. En effet, les courbes de lissage fournies par l'estimation GAM sont plus faciles à interpréter et fournissent un intervalle de confiance à 95% de l'estimation fournie. Cela permettra d'optimiser le découpage des variables continues préalable à la mise en œuvre de modèles de type GLM.

CHAPITRE III. INTEGRATION D'UNE VARIABLE CONTINUE DANS LE MODELE D'ESTIMATION DE FREQUENCE : Exemple de l'âge du conducteur

Dans ce chapitre, nous proposons une méthode de regroupement des variables continues qui utilise le modèle additif généralisé. Nous présentons l'exemple de l'âge pour les garanties RC corporels et Dommages. Avant de regrouper les modalités de variables, le modèle additif permet de tester l'hypothèse de non linéarité de la relation avec la fréquence de sinistres. Enfin, nous allons comparer des modèles GLM afin de mettre en évidence l'importance de la catégorisation des variables pseudo-continues par le modèle additif généralisé.

III.1. La garantie Responsabilité civile corporels

III.1.1. Non linéarité

Pour tester l'hypothèse de non linéarité de la relation avec la fréquence de sinistres, nous voulons estimer la part due à l'influence de l'âge sur la réponse. Ainsi, construisons le modèle suivant de type GAM. $\log(\text{IE}(\text{Nombre de sinistres}_i)) = \alpha_0 + f(age_i))$

Où le nombre de sinistres est supposé suivre une loi de Poisson.

La structure additive du modèle GAM permet cette démarche d'estimation individuelle. L'inclusion d'autres variables explicatives dans le modèle ne changerait pas la forme de la relation estimée ici.

Synthèse de la table d'entrée	
Number of Observations	83
Number of Missing Observations	0
Distribution	Poisson
Link Function	Log

Tableau III-1 : Synthèse de la table en entrée pour l'estimation du nombre de sinistres pour la garantie RCC

Le modèle converge, le tableau 2 fournit le résumé de la résolution du modèle. Dans le chapitre précédent, nous avons donné le détail de la théorie des algorithmes «Local scoring» et « backfitting » qui se complètent pour la résolution d'un modèle additif généralisé.

Synthèse des itérations et statistiques d'ajustement	
Number of local scoring iterations	13
Local scoring convergence criterion	1.2889788E-9
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	5.9258988E-9
The Deviance of the Final Estimate	80.312792137

Tableau III-2 : Synthèses des itérations et statistiques d'ajustement

La procédure GAM sépare l'influence linéaire de l'âge de son influence non linéaire. Le tableau 3 donne l'ajustement de la partie linéaire du modèle. La p-valeur est inférieure à 0.0001.

En d'autres termes, pour la variable, on accepte l'hypothèse que le coefficient β_j estimé par le modèle est significativement non-nulle. Sur le tableau 4, on observe l'ajustement non linéaire de l'âge.

Ainsi, on constate que l'influence de l'âge sur la fréquence annuelle de sinistres comporte une partie linéaire et une partie non non-linéaire (p-valeur <0.0001),.

Analyse du modèle de régression Valeurs estimées des paramètres				
Paramètre	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	6.10095	0.02868	212.73	<.0001
Linear(AGECOND)	-0.02370	0.00061811	-38.34	<.0001

Tableau III-3 : Estimations des paramètres intervenants dans la partie non linéaire du modèle Poisson GAM.

Analyse du modèle de lissage Récapitulatif d'ajustement pour composantes du lissage						
Composante	Paramètre de lissage	DDL	GCV	Nombre d'observations	Khi-2	Pr > Khi-2
Spline(AGECOND)	0.999706	12.743325	1.543763	83	1717.9051	<.0001

Tableau III-4: Récapitulatif d'ajustement pour composantes de lissage

Ainsi, nous obtenons un modèle

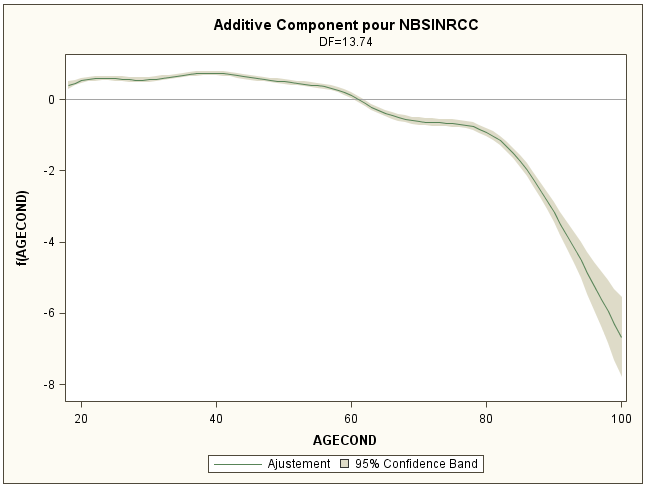
$$\log(\mathrm{IE}(\mathrm{Nombre\,de\,sinistres}_i)) = \hat{\alpha}_0 + \hat{\beta}_{age} * \hat{age}_i + \hat{f}_{age}(age_i))$$

Avec $\hat{\beta}_{age}$ =-0.02370 et f est un spline de lissage cubique.

Notons que pour ce modèle, le modèle de type GLM diverge. Voir résultats en annexe.

Le modèle GAM fournit la courbe suivante qui représente :

$\hat{y}_{RCC} = \hat{\alpha}_0 + \hat{\beta}_{age} * (\hat{age}_i - \overline{\hat{age}}) + \hat{f}_{age}(age_i))$, elle apparaît bien non-linéaire.



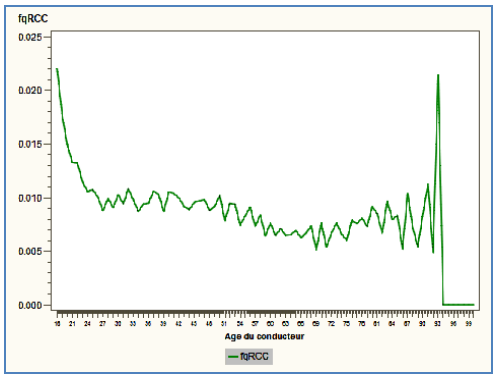
III-1: $\hat{y}_{RCC}$ en fonction de l'âge du conducteur

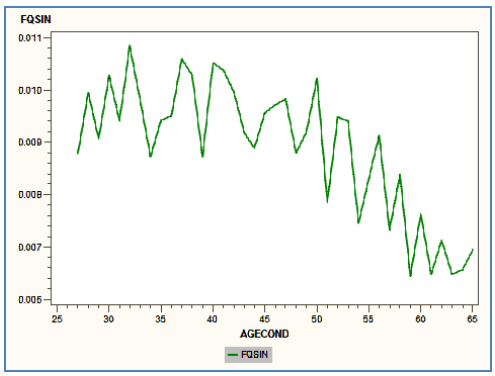
En conclusion, sur cet exemple, nous avons testé et accepté l'hypothèse de non linéarité de l'influence de la variable âge du conducteur sur la fréquence de sinistres RCC. Les tests de

significativités des coefficients de lissage (présentés en annexe) ne rejettent pas cette hypothèse avec un seuil de 0.0001.

III.1.2. Catégorisation de la variable

L'âge du conducteur est un critère de tarification très discriminant, il traduit l'expérience de conduite et l'expérience d'assurance de l'assuré. Ici, nous observons son influence sur la sinistralité RCC Les figures ci-dessous illustrent le caractère volatil de l'influence de l'âge sur la fréquence de sinistres RCC.





III-2 : Données brutes de la fréquence de sinistres RCC en fonction de l'âge du conducteur
III-3: Données brutes de la fréquence de sinistres RCC en fonction de l'âge du conducteur entre 27 et 65 ans

Nous allons construire les courbes lissées de la fréquence de sinistres RCC en fonction de l'âge pour construire de nouvelles classes de modalité pour cette variable.

III.1.2.1. Graphe de l'influence partielle de l'âge sur la fréquence de sinistres

Le modèle que nous avons présenté plus haut estime le nombre de sinistres en fonction de l'âge mais ce qui nous intéresse d'observer ici, c'est le l'influence de l'âge sur la fréquence annuelle de sinistres RCC observée.

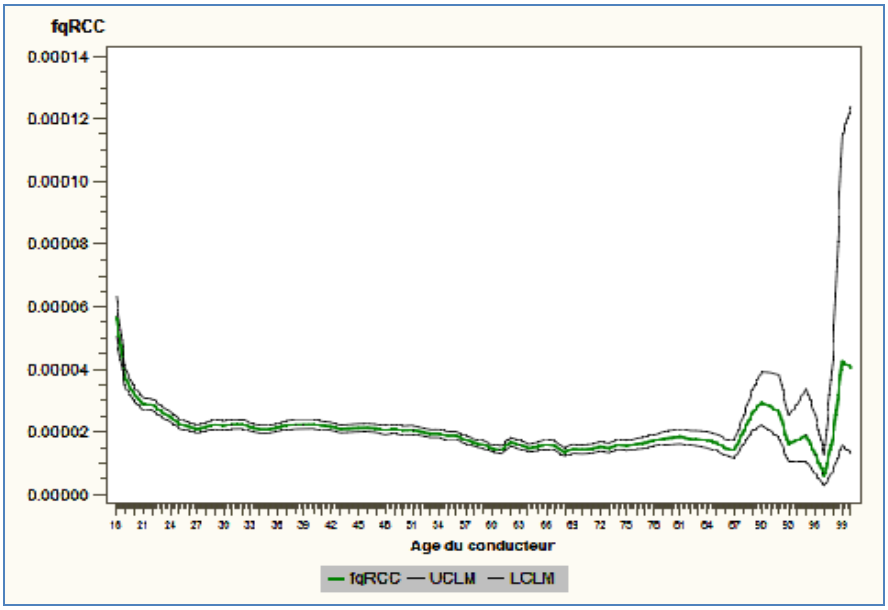
Ainsi, nous allons extraire la partie marginale correspondant à l'influence de l'âge : $\boxed{\hat{\beta}_{age} * \hat{a}ge_i + \hat{f}_{age}(age_i)}$

Nous avons ainsi la part de la fréquence de sinistres due uniquement due à l'âge du conducteur.par la formule :

$$\boxed{FqSINRCC_{i,\hat{a}ge} = \exp(-0.01060 * \hat{a}ge_i + \hat{f}_{age}(age_i)) / durée\,d'exposition_i}$$

Le modèle donne aussi des intervalles de confiances à 95% du nombre de sinistres estimés. De la même manière, nous retrouvons des IC à 95% de FqSINRCC.

Ces quantités sont représentées sur le graphe suivant :



III-4: Fréquence de sinistres en fonction de l'âge du conducteur et IC supérieur et inférieur à 95%

La borne supérieure est très large pour les valeurs supérieures à 96 ans. Cela est due à la faible distribution sur se segment, nous allons enlever cette borne du graphique pour avoir un schéma plus lisible.

III.1.2.2. Choix des classes

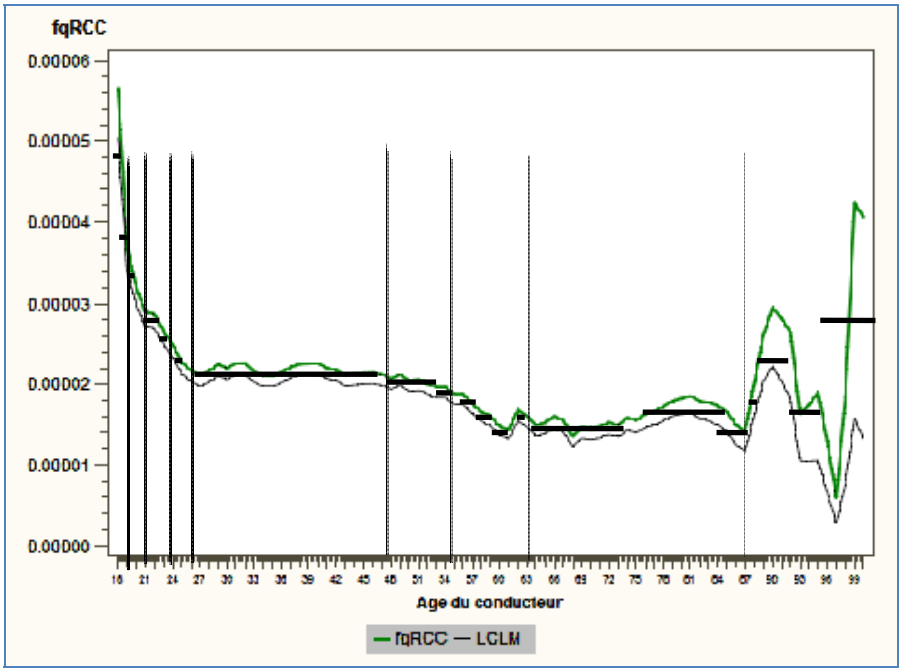
- Ainsi, à partir des intervalles de confiance de l'estimation, nous pouvons définir des niveaux de risques. Nous prenons une valeur moyenne de fréquence (représenté par le trait noir foncé) dans chaque intervalle de confiance qui correspond à un niveau de risque moyen de la classe.

- Les classes doivent être disjointes

- Nous tenons compte aussi de la distribution du portefeuille (pour éviter dans la mesure du possible des segments trop petits).

- Ceci permet de catégoriser la variable âge en mettant ensemble des modalités de niveaux de risques proches. Bien entendu, ceci comporte une part d'arbitraire.

Sur la figure suivante, nous schématisons cette démarche.



III-5:Fréquence de sinistres en fonction de l'âge du conducteur et IC inférieur à 95%

1. *Age compris entre 18 et 26 ans*

Entre 18 et 26 ans, on observe une décroissance forte de la fréquence de sinistres RCC. Je décide de diviser ce segment en 3 parties.

Age du conducteur	
Classe 1	18 – 19 ans
Classe 2	20 - 21 ans
Classe 3	22 – 26 ans

2. *Age supérieur à 65 ans*

Aussi, la distribution sur la tranche de plus de 65 ans est faible. A partir du graphe ci-dessous, nous divisons cette partie en deux classes.

Age du conducteur	
Classe 1	65 – 85 ans
Classe 2	85 ans et plus

Sur ce segment, on peut noter un saut de segmentation (et donc de tarif) qui était injustifié dans l'approche intuitive de segmentation. En effet, ce segment était divisé en trois classes :

Age du conducteur

Classe 1	60– 64 ans
Classe 2	65 - 70 ans
Classe 3	70 ans et plus

Cependant sur la courbe ci-dessous, on observe une constance de la fréquence de sinistres entre 65 et 85 ans.

3. *Age compris entre 27 et 65 ans*

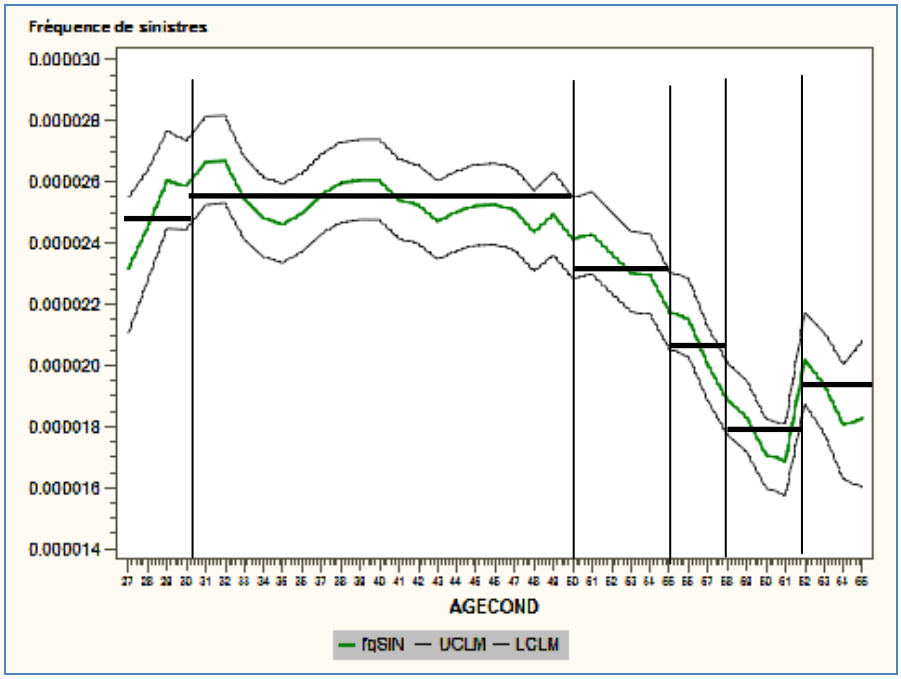
Une bonne partie du portefeuille est concentrée entre 27 et 65 ans, nous allons refaire une estimation pour cette partie du portefeuille afin d'induire une segmentation plus fine de cette partie du portefeuille. Nous procédons de la même manière. Le modèle de type GAM converge.

$\hat{\beta}_{age}$ = -0.01668. Nous observons donc sur les graphes suivants

:

$$FqSIN_i = \exp(-0.01668 * \hat{a}ge_i + \hat{f}_{age}(age_i)) / durée\ d'exposition_i$$

Le modèle donne aussi des intervalles de confiances à 95% du nombre de sinistres estimés. De la même manière, nous retrouvons des IC à 95% de FqSIN.



III-6:Fréquence de sinistres RCC pour les âges compris entre 27 et 65 ans

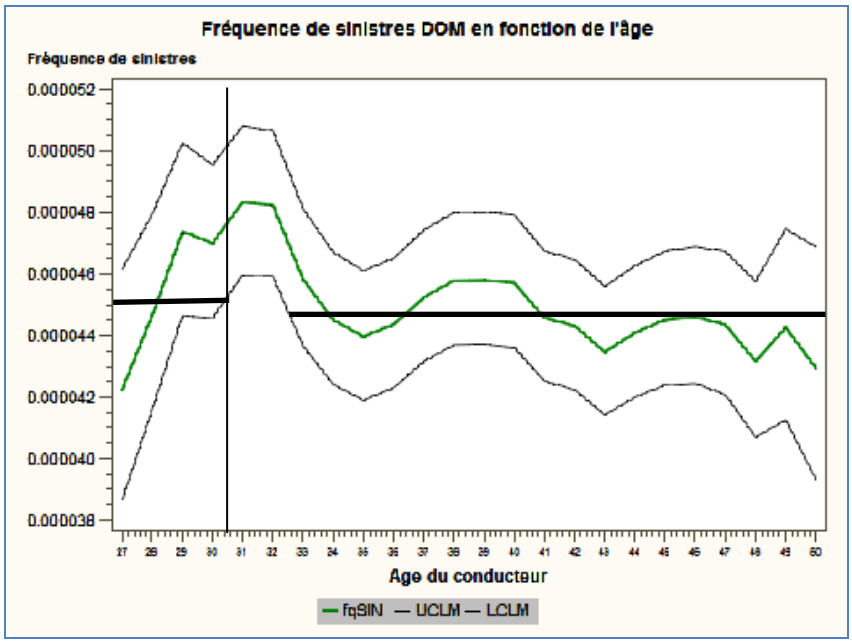
Avec un arbitrage, distribution/ niveaux de risque, nous allons faire le découpage suivant :

Classe 1	27 – 30 ans
Classe 2	31 – 50 ans
Classe 3	51 – 55 ans
Classe 4	56 – 59 ans
Classe 7	60- 62 ans
Classe 8	63 – 65 ans

Pour les 27 – 50 ans, nous refaisons une estimation. Le modèle converge, $\hat{\beta}_{age}$ = -0.00101.

Nous observons donc sur les graphes suivants :

$$FqSIN_i = \exp(-0.00101 * \hat{a}ge_i + \hat{f}_{age}(age_i)) / durée\ d'exposition_i$$



III-7:Fréquence de sinistres RCC pour les âges compris entre 27 et 50 ans

Nous allons conserver le découpage effectué précédemment. La sinistralité en fréquence est constante entre 31 et 50 ans. En conclusion, nous obtenons le découpage suivant :

III.1.2.3. Comparaison avec le découpage initial de l'âge

Ancien découpage	
Classe 1	18 – 20 ans
Classe 2	21 – 22 ans
Classe 3	23 – 24 ans
Classe 4	25 - 29 ans
Classe 5	30 – 34 ans
Classe 6	35 – 39 ans
Classe 7	40 – 44 ans
Classe 8	45 – 49 ans
Classe 9	50 – 54 ans
Classe 10	55 – 59 ans
Classe 11	60 – 64 ans
Classe 12	65 – 69 ans
Classe 13	70 – 99 ans

Nouveau découpage	
Classe 1	18 – 19 ans
Classe 2	20 - 22 ans
Classe 3	23 – 26 ans
Classe 4	27 – 30 ans
Classe 5	31 – 50 ans
Classe 6	51 – 55 ans
Classe 7	56 – 59 ans
Classe 8	60- 62 ans
Classe 9	63 – 65 ans
Classe 10	66 – 85 ans
Classe 11	85 ans et plus

Nous venons de mettre en évidence :

- ✓ Il est important de faire un saut d'âge à 19 ans,
- ✓ Il n'est pas nécessaire de diviser les 30-50 ans en 4 classes car la fréquence de sinistres est constante en moyenne sur ce segment.
- ✓ La sinistralité est constante entre 65 et 85 ans. Il n'y a pas de saut notable à 70 ans comme pourrait le laisser penser le découpage initial, il est plutôt à 65 ans.

III.2. La garantie Dommages

III.2.1. Non linéarité

Nous voulons estimer la part due à l'influence de l'âge sur le modèle. Ainsi, construisons le modèle suivant de type GAM. $\log(\text{IE}(\text{Nombre de sinistres}_i)) = \alpha_0 + f(age_i))$

Où le nombre de sinistres est supposé suivre une loi de Poisson.

Synthèse de la table d'entrée	
Number of Observations	83
Number of Missing Observations	0
Distribution	Poisson
Link Function	Log

Tableau III-5 : Synthèse de la table en entrée pour l'estimation du nombre de sinistres pour la garantie DOM

Le modèle converge, le tableau 6 fournit le résumé de la résolution du modèle

Synthèse des itérations et statistiques d'ajustement	
Number of local scoring iterations	4
Local scoring convergence criterion	5.6241991E-9
Final Number of Backfitting Iterations	2
Final Backfitting Criterion	6.3661502E-9
The Deviance of the Final Estimate	82.254211512

Tableau III-6 : Synthèses des itérations et statistiques d'ajustement

Les tableaux 7 et 8 indiquent que l'influence de l'âge sur la fréquence annuelle de sinistres comporte une partie linéaire et une partie non non-linéaire (p-valeur <0.0001), l'influence de l'âge sur la fréquence de sinistres comporte bien une partie linéaire et une partie non-linéaire.

Analyse du modèle de régression Valeurs estimées des paramètres				
Paramètre	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	7.34566	0.01281	573.64	<.0001
Linear(AGECOND)	-0.01060	0.00024524	-43.23	<.0001

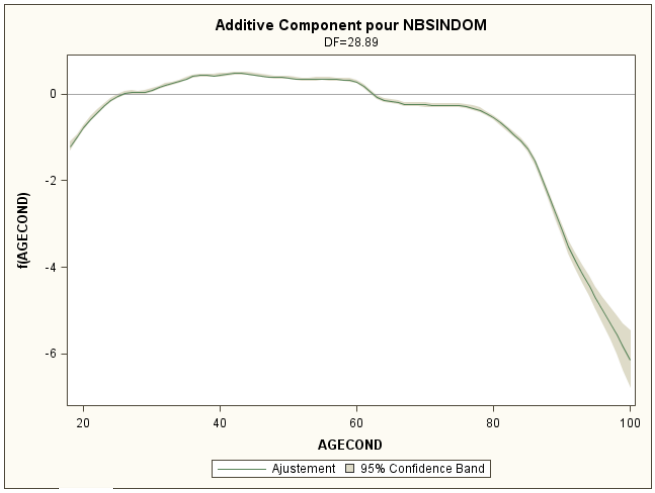
Tableau III-7 : Estimations des paramètres intervenants dans la partie non linéaire du modèle Poisson GAM.

Analyse du modèle de lissage Récapitulatif d'ajustement pour composantes du lissage						
Composante	Paramètre de lissage	DDL	GCV	Nombre d'observations	Khi-2	Pr > Khi-2
Spline(AGECOND)	0.989803	27.889190	0.212924	83	19534.6551	<.0001

Tableau III-8: Récapitulatif d'ajustement pour composantes de lissage

Ici $\hat{\beta}_{age}$ =-0.01060.

Le modèle fournit la courbe suivante qui représente : $\hat{y}_{DOM} = \hat{\alpha}_0 + \hat{\beta}_{age} * (\hat{a}ge_i - \overline{\hat{a}ge}) + \hat{f}_{age}(age_i))$



III-8: $\hat{y}_{DOM}$ en fonction de l'âge du conducteur

La courbe est bien non linéaire.

En conclusion, sur cet exemple, nous avons testé et accepté l'hypothèse de non linéarité de l'influence de la variable âge du conducteur sur la fréquence de sinistres DOM. Les tests de significativités des coefficients de lissage ne rejettent pas cette hypothèse avec un seuil de 0.0001.

III.2.2. Catégorisation de la variable

III.2.2.1. Graphe de l'influence partielle de l'âge sur la fréquence de sinistres

Nous avons estimé le nombre de sinistres en fonction de l'âge mais ce qui nous intéresse à observer ici, c'est le l'influence de l'âge sur la fréquence annuelle de sinistres DOM observée.

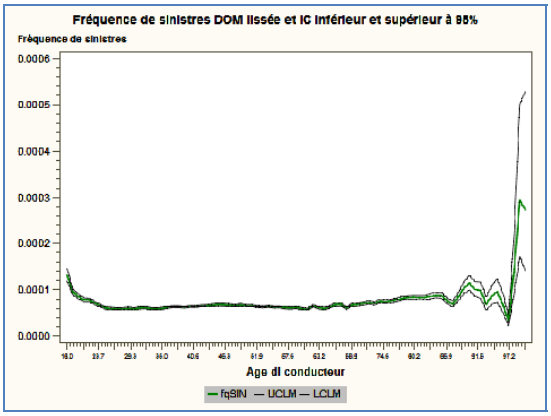
Ainsi, nous allons extraire la partie marginale correspondant à l'influence de l'âge, ici

$$\hat{\beta}_{age} * \hat{age}_i + \hat{f}_{age}(age_i)$$

Nous avons ainsi la part de la fréquence de sinistres due à l'âge du conducteur seul.

$$FqSINDOM_{i,age} = \exp(-0.01060 * \hat{age}_i + \hat{f}_{age}(age_i)) / durée\ d'exposition_i$$

Le modèle donne aussi des intervalles de confiances à 95% du nombre de sinistres estimés. De la même manière, nous retrouvons des IC à 95% de FqSINDOM.
Ces quantités sont représentées sur le graphe suivant :



III-9: Fréquence de sinistres DOM en fonction de l'âge du conducteur et IC supérieur et inférieur à 95%

III.2.2.2. Choix des classes

D'ores et déjà, les clients de plus de 90 ans seront mis ensemble. La distribution est trop faible pour ce segment et l'intervalle de confiance de l'estimation trop large. Notons aussi que globalement une telle courbe confirme l'influence faible de l'âge du conducteur sur la garantie Dommages que nous avions observé dans le chapitre 1 (analyse des données).

Le graphique n'est pas très lisible, nous divisons le reste du portefeuille en trois et nous refaisons une estimation GAM séparée.

Nous observons séparément : les personnes de moins de 27 ans, les personnes entre 2è ans et 50 ans et les personnes entre 50 et 90 ans. Les modèles restent convergents.

1. *Pour le segment 18 à 27 ans :*

Ici $\hat{\beta}_{age}$ =0,14735, la courbe suivante représente

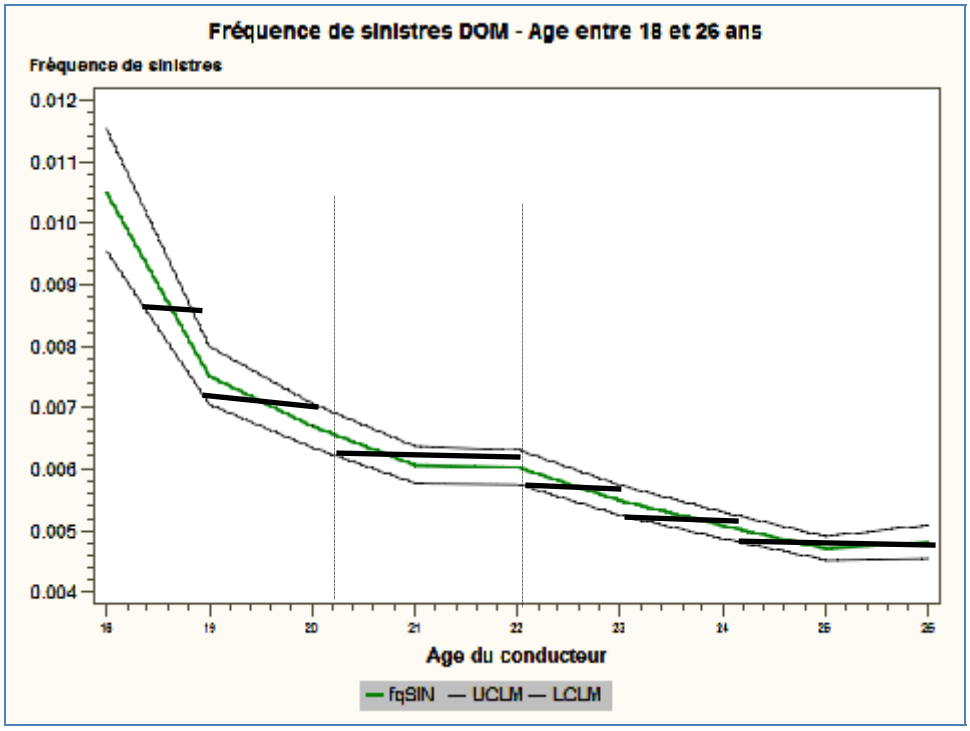
$$FqSINDOM_{i,age} = \exp(0.14735 * \hat{a}ge_i + \hat{f}_{age}(age_i)) / durée\ d'exposition_i$$

, et un intervalle de confiance à 95%. La courbe suivante (figure 10 de la page suivante[5]) représente $\hat{FqSINDOM}$ l'estimation GAM du nombre de sinistres. A sa lecture, nous décidons de diviser ce segment en trois.

Age du conducteur	
Classe 1	18 – 19 ans
Classe 2	20 - 22 ans
Classe 3	23 – 26 ans

- ✓ Le saut à 19 ans est assez marqué, il est inapproprié de mettre ensemble les personnes entre 19 et 20 ans

[5] Nous avons fait un schéma assez grand pour être facilement lisible



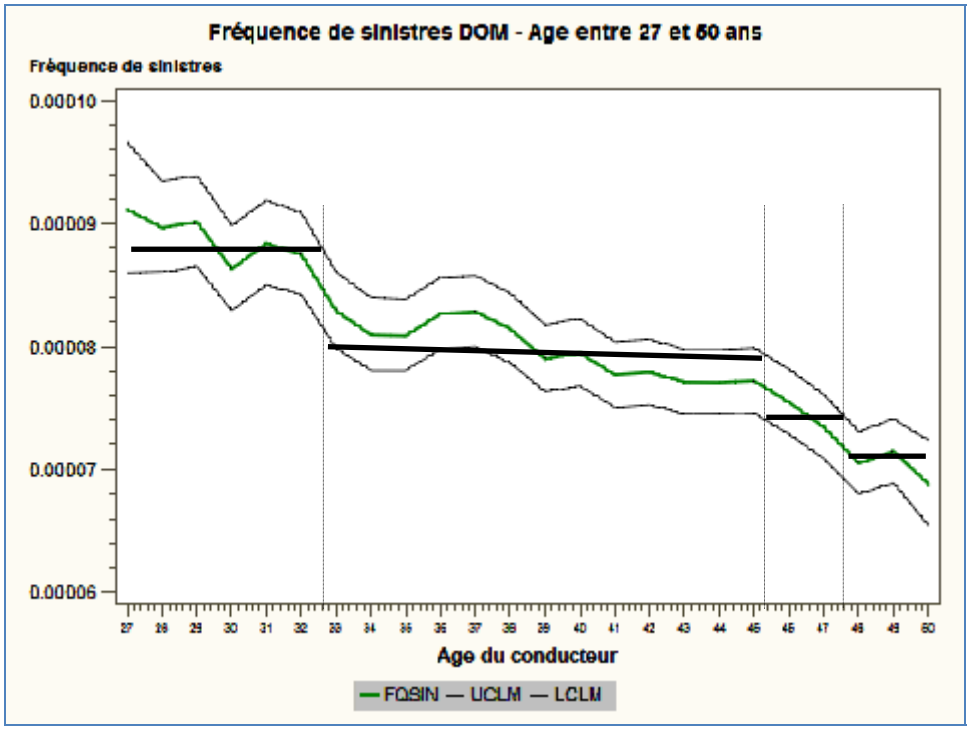
III-10:Fréquence de sinistres entre 18 et 26 ans et IC inférieur à 95%

2. *Age entre 27 à 50 ans :*

$\hat{\beta}_{age}$ =-0.01591, la courbe suivante représente

$\boxed{FqSINDOM\hat{}_{i,age} = \exp(-0.01591 * \hat{a}ge_i + \hat{f}_{age}(age_i)) / durée\ d'exposition_i}$, et un intervalle de confiance à 95% de courbe suivante représente $FqSINDOM\hat{}$ l'estimation GAM du nombre de sinistres.

Age du conducteur	
Classe 1	27 – 32 ans
Classe 2	33 - 40 ans
Classe 3	41 – 45 ans
Classe 4	46 - 50 ans
Classe 5	70 - 80 ans
Classe 6	80 ans et plus



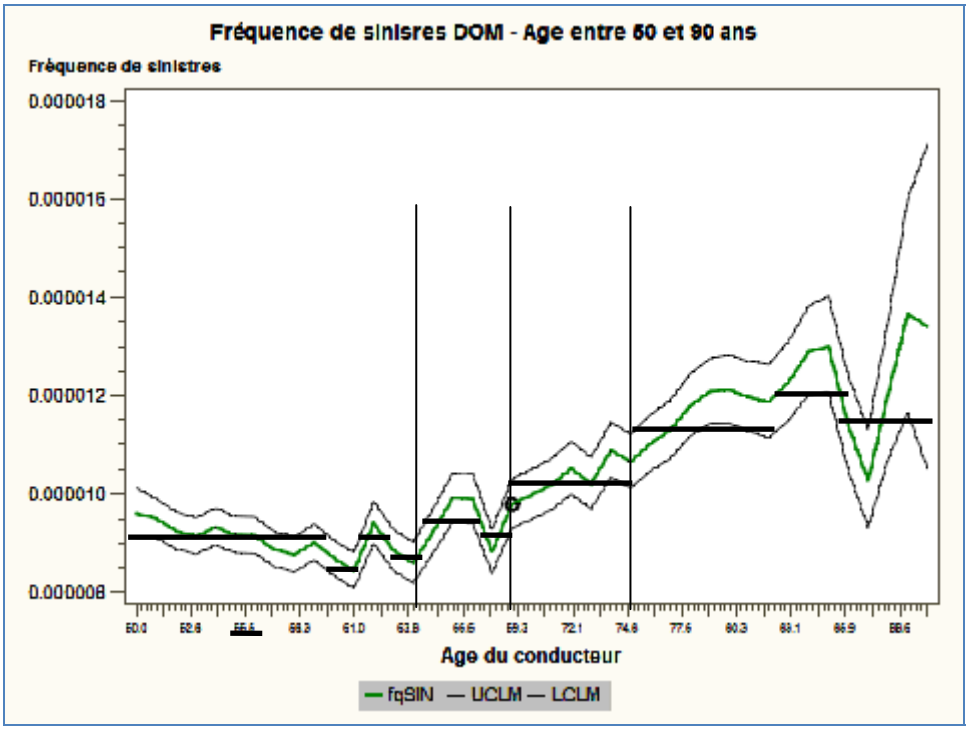
III-11:Fréquence de sinistres entre 27 et 50 ans et IC inférieur à 95%

3. *Age entre 50 à 90 ans :*

Ici $\hat{\beta}_{age}$ =-0.04051, la courbe suivante représente

$$\boxed{FqSINDOM\hat{}_{i,age} = \exp(-0.01591 * \hat{a}ge_i + \hat{f}_{age}(age_i)) / durée\ d'exposition_i}$$, et intervalle de confiance à 95% de courbe suivante représente $FqSINDOM\hat{}$ fourni par l'estimation GAM du nombre de sinistres.

Classe 4	46 - 50 ans
Classe 5	70 - 80 ans
Classe 6	80 ans et plus



III-12 :Fréquence de sinistres entre 27 et 50 ans et IC inférieur à 95%

Les nouvelles classes d'âge que nous constituons sont :

III.2.2.1. Comparaison avec le découpage initial de l'âge

Ancien découpage	
Classe 1	18 – 20 ans
Classe 2	21 – 22 ans
Classe 3	23 – 24 ans
Classe 4	25 - 29 ans
Classe 5	30 – 34 ans
Classe 6	35 – 39 ans
Classe 7	40 – 44 ans
Classe 8	45 – 49 ans
Classe 9	50 – 54 ans
Classe 10	55 – 59 ans
Classe 11	60 – 64 ans
Classe 12	65 – 69 ans
Classe 13	70 – 99 ans

Nouveau découpage	
Classe 1	18 – 19 ans
Classe 2	20 – 22 ans
Classe 3	23 – 26 ans
Classe 4	27 – 32 ans
Classe 5	33 - 40 ans
Classe 6	41 – 45 ans
Classe 7	46 - 50 ans
Classe 8	51 - 69 ans
Classe 9	69 – 80 ans
Classe 10	80 et plus

- ✓ Le découpage n'est pas le même que pour la garantie RCC,
- ✓ Le saut à 19 est assez marqué, il est inapproprié de mettre ensemble les personnes entre 19 et 20 ans
- ✓ Entre 46 et 70 ans, on observe une constance de sinistralité.

Nous ne présenterons pas les résultats obtenus sur les autres variables et les autres garanties. La démarche utilisée est similaire.

En visualisant les courbes de lissage fournies par la GAM, nous avons effectué une catégorisation des variables âge du conducteur, âge du véhicule et ancienneté du contrat. Dans le paragraphe suivant, en comparant les différents critères de mesure d'ajustement fournis par Sas :

- La log-vraisemblance du modèle qui doit être maximale
- La déviance du modèle qui doit être minimale
- Les critères d'Akaike à minimiser,

Nous allons quantifier l'apport de la catégorisation des variables continues à l'aide de la GAM en termes d'amélioration des modèles paramétriques d'estimation de la fréquence de sinistres construits. Présentons les résultats obtenus sur les garanties RCC et Dommages.

III.3. Comparaison de modèles ancien découpage- nouveau découpage des variables âge du conducteur, âge du véhicule et ancienneté du contrat

III.3.1. La garantie Responsabilité civile Corporels

En combinant toutes modifications effectuées sur l'âge du conducteur, l'âge du véhicule et l'âge du contrat, nous construisons de nouveaux modèles d'estimation de la fréquence de sinistres annuelle. Pour des raisons de confidentialité, nous ne dévoilons pas la structure du modèle.

Les deux modèles sont satisfaisants au sens de l'analyse de type3, comparons les résultats d'adéquation ; le tableau 9 donne les différentes mesures d'adéquation aux données des modèles.

Critère d'évaluation de l'adéquation		
Critère	**Modèle initial**	**Nouveau modèle**
Deviance	69806.3523	63683.4681
Scaled Deviance	69806.3523	63683.4681
Pearson Chi-Square	1843615.0134	1285059.7138
Scaled Pearson X2	1843615.0134	1285059.7138
Log Likelihood	-42930.1601	-40068.7115
Full Log Likelihood	-43941.7004	-40763.5545
AIC (smaller is better)	88133.4008	81719.1090
AICC (smaller is better)	88133.4295	81719.1309
BIC (smaller is better)	89621.6655	82837.9062

Tableau III-9: RCC : Comparaison des critères d'adéquation aux données

Du modèle initial au nouveau modèle, la vraisemblance du modèle augmente, la déviance diminue, la somme des carrés des résidus est plus petite. Les critères AIC et BIC sont aussi moins élevé.

Au regard de tous les critères d'évaluation à notre disposition, le nouveau modèle estime au plus proche des données le nombre de sinistres RCC à attendre sur un risque donné.

III.3.2. La garantie Dommages

Les deux modèles sont satisfaisant, comparons les résultats d'adéquation ; les figures suivantes résument tous les critères d'évaluation d'adéquation disponibles pour chacun des modèles.

Critère d'évaluation de l'adéquation		
Critère	**Modèle initial**	**Nouveau modéle**
Deviance	359553.3575	304911.4118
Scaled Deviance	359553.3575	304911.4118
Pearson Chi-Square	2062472.6605	1534756.4921
Scaled Pearson X2	2062472.6605	1534756.4921
Log Likelihood	-236092.7257	-204198.7164
Full Log Likelihood	-237793.3778	-208424.2792
AIC (smaller is better)	475876.7556	417080.5584
AICC (smaller is better)	475876.7859	417080.5850
BIC (smaller is better)	477638.5258	418453.5644

Tableau III-10: Dommages - Comparaison de l'adéquation aux données du modèle initial

Du modèle initial au nouveau modèle, la vraisemblance du modèle augmente, la déviance diminue, la somme des carrés des résidus est plus petite. Les critères AIC et BIC sont aussi moins élevé.

CONCLUSION

Les modèles économétriques privilégient en général les modèles paramétriques (linéaires en l'occurrence) du fait de leur facile interprétation, ceci des fois en rejetant complètement les modèles non paramétriques. Nous avons démontré ici l'intérêt des modèles additifs généralisés pour l'estimation de la fréquence de sinistres. Même si le modèle additif généralisé ne remplace pas ici le modèle linéaire généralisé pour des raisons techniques, nous avons proposé une méthode de segmentation du portefeuille avant l'estimation, basée sur ce modèle.

Nous avons démontré que l'utilisation conjointe des modèles généralisés et des modèles additifs généralisés permet d'améliorer la vraisemblance et l'adéquation aux données des modèles d'estimation de la fréquence de sinistres.

De plus, des algorithmes puissants peuvent être implémentés sur des ordinateurs de plus en plus performants, ces outils sont des outils modernes très performants auxquelles il convient de s'y intéresser davantage. Le risque de modèle (risque lié au modèle choisi et l'utilisation du modèle) peut s'avérer sournois et dans un contexte où l'appréhension des risques pris par les compagnies devient de plus en plus primordiale, c'est un risque à ne pas négliger.

Nous avons aussi mis en évidence que la segmentation du portefeuille avant l'estimation doit se faire garantie par garantie. Une segmentation intuitive peut engendrer des sauts de tarifs injustifiés techniquement et dans le contexte de forte concurrence dans le domaine de l'assurance automobile, aucun détail n'est à négliger. Le modèle additif offre ici une méthode de segmentation plus « technique » des variables continues. Cette étude de l'intérêt de la GAM en tarification peut être étendue à d'autres problématiques de tarification, notamment l'estimation des coûts de sinistres, l'estimation des probabilités d'entrées et de sorties de portefeuille.

ANNEXES

[1] Evolution du marché de l'assurance automobile

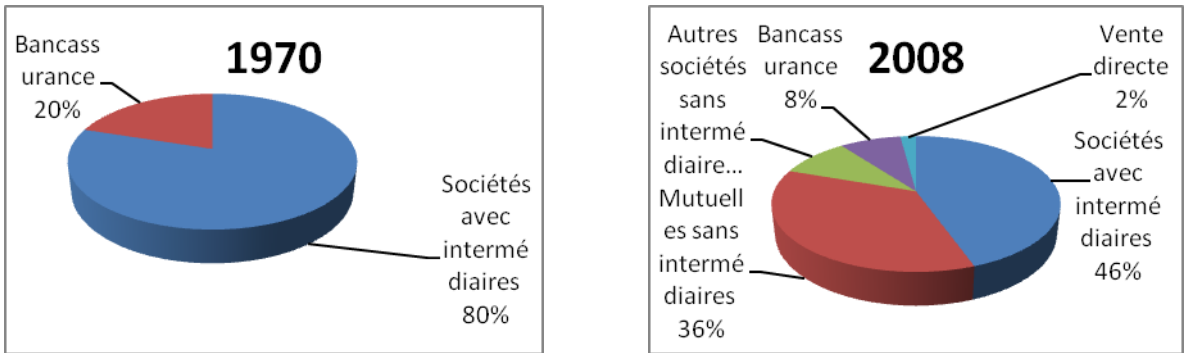


Figure 0-1 : Marché de l'assurance automobile en France (Source : Le marché de l'automobile – FFSA, juillet 2009).

[2] Interprétation de l'analyse en composantes principales

Principe

L'analyse en composantes multiples étudie le lien entre plusieurs variables qualitatives. Les données peuvent avoir deux formats :

• le tableau disjonctif complet.

• le tableau de BURT

Considérons un ensemble V=$\{v_1,...,v_k\}$ de n variables et un ensemble I de n individus

Pour chaque variable v_i de V, on note M_i l'ensemble des modalités de v_i et l'on désigne par M l'union disjointe des M_i

$$M = \bigcup_{i=1}^{n} Mi \qquad et \quad Card(M) = p$$

Le **tableau disjonctif complet** K_{IJ} est défini par :

$$\forall (i,j) \in I \times M, k(i,j) = \begin{cases} 1 \; si \; la \; \text{mod}alité \; présente \; la \; \text{mod}alité \; j \\ 0 \; \text{sin}on \end{cases}$$

On a donc : $k(i,j) \in \{0,1\}$ *et* $\sum_{j \in M_i} k(i,j) = 1$

Le **tableau de Burt** est défini comme le tableau des cooccurrences des modalités des différentes variables :

$$\boxed{\forall (j,j') \in M^2,\ B(j,j') = \sum_{i \in M} k(i,j) * k(i,j')}$$

Soit sous forme matricielle : $\boxed{B = {}^tK_{IJ} * K_{IJ}}$

L'analyse en composantes multiples revient à rechercher les axes principaux d'inertie des nuages suivants :

- nuage des profils des lignes i du tableau B

- nuage des profils des colonnes j du tableau B

Les facteurs sont identiques dans l'analyse des K_{IJ} et dans celle de B, les valeurs propres étant dans l'analyse de B les carrés de celles issues de K_{IJ}. Lorsqu'on ne travaille que sur l'étude des variables, on utilise le tableau de Burt, car il est beaucoup plus petit et permet un gain important de temps de calcul (K_{IJ} est un tableau de taille n.p alors que B est un tableau de taille p^2).

L'analyse en composantes multiples revient à rechercher les axes principaux d'inertie des nuages suivants :

- nuage des profils des lignes i du tableau B

- nuage des profils des colonnes j du tableau B

Les facteurs sont identiques dans l'analyse des K_{IJ} et dans celle de B, les valeurs propres étant dans l'analyse de B les carrés de celles issues de K_{IJ}. Lorsqu'on ne travaille que sur l'étude des variables, on utilise le tableau de Burt, car il est beaucoup plus petit et permet un gain important de temps de calcul (K_{IJ} est un tableau de taille n.p alors que B est un tableau de taille p^2).

Contribution relative

La contribution relative d'une modalité j à l'inertie de l'axe k est :

$$\boxed{C(k,j) = \frac{1}{c_k} \times \frac{n_j}{Nn} \times c_{j,k}^2}$$

Avec

§ n_j : l'effectif de la modalité j

§ N : le nombre de variables actives dans l'ACM

§ n : le nombre total d'individus

§ $c_{j,k}$, la coordonnée de la modalité j sur l'axe k

§ l_k : la valeur propre associée à l'axe k

Cette quantité permet de repérer les variables qui contribuent le plus à l'inertie des axes. Son étude permet de « décrire » les axes.

Qualité de représentation:

La qualité de représentation d'une modalité sur un axe s'apprécie en calculant le cosinus carré de l'angle entre l'axe et le vecteur joignant le centre de gravité du nuage et le point représentant la modalité étudiée.

Plus ce cosinus est élevé, plus la modalité est voisine de l'axe et plus la position de la modalité en projection est proche de sa position réelle dans l'espace. On apprécie ainsi la qualité de représentation d'une modalité dans un plan factoriel en faisant la somme des cosinus carrés sur les axes étudiés.

Le cosinus carré pour un axe k et une modalité j est la quantité :

$$\cos_k^2 = \frac{c_{j,k}^2}{d_j^2}$$

Avec :

§ $c_{j,k}$, la coordonnée de la modalité j sur l'axe k

§ d_j , la distance de la modalité j au centre de gravité du nuage.

[3] Justification de l'utilisation de la loi de poisson pour modéliser la fréquence de sinistres en assurance automobile

Une justification de l'emploi de la loi de Poisson pour modéliser le nombre de sinistres d'un risque sur une période fixée est fournie par les arguments suivants. Considérons une valeur t positive représentant un instant donné. Quel que soit l'entier 1 ≤k, l'intervalle]0, t] peut être partagé en k sous-intervalles ou périodes de longueurs égales, soit h = t/k, comme. Pour tout i allant de 1 `a k, on note Ni le nombre de sinistres sur le i-émet sous-intervalle de longueur h.

On introduit alors les trois hypothèses ci-après.

(a) A condition que h soit suffisamment petit, on considère que sur un sous-intervalle de longueur h, la probabilité de survenance de plus d'un sinistre est négligeable. Deux cas sont donc possibles : un sinistre se produit sur cet intervalle ou il ne se produit aucun sinistre. Les probabilités respectives de ces évènements sont p et 1 − p.

(b) Sur chaque sous-intervalle de longueur h, la valeur de la probabilité p est de la forme p = ch., où c est une constante positive,

(c) Il y a indépendance entre les nombres de sinistres des périodes successives.

Ces trois hypothèses entraînent que les variables aléatoires N_1, N_2,…,N_{OK} sont indépendantes et de même loi de Bernoulli B(1, p). Par conséquent, la variable aléatoire $N = N_1 + N_2 + \cdot + N_{OK}$, qui représente le nombre de sinistres sur l'intervalle de temps]0, t], suit la loi B(k, p), avec p = ct/k.

Cette dernière loi converge vers la loi de Poisson P(c t) lorsque k tend vers l'infini. Par conséquent, si k est suffisamment grand, la loi B(k, p), pourra être approchée par la loi de Poisson P(k p). On dit parfois pour cette raison que la loi de Poisson est la "loi des évènements rares".

Si on réalise n (n>q) mesures successives, la probabilité pour qu'au cours des (n – q) dernières mesures l'événement considéré n'ait pas lieu est donc égale à $(1 - p)^{n-q}$.

$p.q.(1 - p)^{n-q}$ représente la probabilité d'avoir q événements au cours des q premières mesures sur les n effectuées. Il existe C_n^q manières de choisir l'ordre d'apparition des q événements parmi les n mesures et la probabilité de trouver q événements au cours d'une série de n mesures est donc :

$P_{nb}(q) = C_n^q . p^q . (1 - p)^{n-q}$ (loi binomiale)

<u>Convergence vers une distribution de Poisson :</u>

Si **n est assez grand** et **p assez petit**, l'expression de la loi binomiale devient :

En remplaçant q par x, on obtient la loi de Poisson :

$$P_n(q) = \frac{1}{q!}(nq)^q(1-\frac{1}{n})...(1-\frac{q-1}{n})(1-p)^{n(1-\frac{q}{n})} \approx \frac{(nq)^p}{q!}(1-p)^n$$

Or, n grand donc $(1-p)^n \approx e^{-np}$

Si p<< n la loi binomiale s'écrit sous la forme approchée :

$$P_n(q) = \frac{(nq)^q}{q!} e^{-np} = \frac{m^q}{q!} e^{-np}$$

En remplaçant m par x, on obtient la loi de poisson

$$\boxed{P_n(x) = \frac{m^x}{x!} e^{-m}}$$

[4] Temps de calcul et espace mémoire utilisé par un modèle GAM sous SAS

Ces deux quantités permettent d'optimiser l'interprétation qui est faite de chacun des axes 1 et 2 retenus et du plan 1-2.

Soient

- N = nombre d'observations
- Pr =nombre de variables à intégrer de façon paramétrique
- Ps = nombre de variables à lisser par des splines de lissage
- Pl = nombre de variables à lisser par des fonctions de type lœss
- Pb= nombre de variables à lisser par des fonctions de type bivariante tin-plate
- P=Pr + Ps + Pl + Pb
- Pnb=Ps + Pl + Pb
- M=maximum d'itérations « backfitting »

En plus de l'espace pour stocker les données (en octets), l'espace de travail minimum (en octets) nécessaires à l'ajustement d'un modèle utilisant PROC GAM est

$$\boxed{(16+8\mathrm{Pr})(n+2pr)+(160+48p+16ps+8pl)n+8p+32pb+32ps+8m+8n+(14n+4)ps+4}$$

On doit rajouter $\boxed{80+120n+8n^2+8pb}$ octets de mémoire pour le montage pour les variables « bivariate thin-plate » et $\boxed{48n+16pt}$ octets de mémoire pour les variables « loess ». Si les limites de confiance sont demandées, la mémoire supplémentaire est requise.

Il est difficile de fournir des estimations précises du temps nécessaire pour ajuster un modèle GAM. Les deux algorithmes « backfitting » et l'algorithme « local scoring » sont des techniques itératives dont les taux de convergence dépendent des données particulières qu'on analyse. En outre, le temps requis dépend des types de lisseurs que l'on spécifie. Ils sont de l'ordre de :

- N^3 si au moins un à deux variables « bivariate thin-plate » est utilisé
- $N^{3/2}$ lorsque seulement lisseurs « lœss » sont utilisés

▪ N lorsque seulement les « splines de lissage » sont utilisés

[5] Quasi Vraisemblance Etendue (ou Extended Quasi-Likelihood)

Ce critère repose en partie sur la déviance : la déviance est une mesure de bon ajustement au sein du modèle et permet de comparer des modèles avec des fonctions de lien g différentes. Cette notion de déviance est détaillée ci-dessous.

Cependant, cette mesure de comparaison des modèles n'est plus valable lorsque la fonction de variance V(.), soit la distribution des Y_{ij}, est changée : si l'on considère V', telle que V'(s)≥V(s) □ s, alors la déviance diminue de façon mécanique sans rendre compte de la qualité des modèles. L'Extended Quasi-Likelihood q+ développée par Nelder et Pregibon puis Nelder et Lee un autre indicateur qui corrige cet effet, et qui permet de comparer les modèles avec des fonctions de variance et des fonctions de liens différentes. Il s'agit de:

$$-2q^{+}(\underline{y},\underline{\hat{u}}) = \frac{1}{\phi}\sum_{i+j\le n} d_{ij} + \sum_{i+j\le n}\log(2\pi\varphi V(y_{ij}))$$

- φ , est estimé par $\hat{\varphi} = \frac{\chi^2}{m} = \frac{1}{m}\sum_{i=j\le n}\frac{(y_{ij}-\hat{u}_{ij})^2}{V(\hat{u}_{ij})}$

- m étant le nombre de degrés de liberté des résidus du modèle (nombre de données-nombre de degrés de liberté du modèle), et

- χ_2 *étant la statistique de Pearson;*

- $\sum_{i=j\le n} d_{ij}$ est la quasi déviance avec $d_{ij} = 2\int_{\hat{u}_{ij}}^{y_{ij}} \frac{y_{ij}-u}{V(u)}du$

Maximiser la statistique q+ permet de comparer les modèles qui ont une structure (V, φ , g) différente. Cela revient à minimiser la statistique -2 q+.

[6] Notion de déviance

La déviance joue le rôle de la **somme au carré des résidus** pour les modèles généralisés et peut être utilisée pour évaluer la qualité de l'ajustement et pour comparer les modèles entre eux :

$$D = 2\sum_{i=j\le n} d_{ij}$$

Où $d_{ij} = 2\left(y_{ij}(\tilde{\theta}_{ij}-\hat{\theta}_{ij}) - \left[b(\tilde{\theta}_{ij}) - b(\hat{\theta}_{ij})\right]\right)$

$\tilde{\theta}_{ij} = b^{-1}(y_{ij})$ et $\hat{\theta}_{ij} = b^{-1}(\hat{u}_{ij})$

[7] Test de significativité des coefficients de lissage

Pour Ho : $f_\lambda = 0$, la statistique de test est la suivante :

$X_\lambda = p_\lambda' \, V_\lambda^{k-} \, p_\lambda$

Où p_λ est le vecteur de paramètres du terme de lissage sλ ,

V_λ est la matrice de covariance,

V_λ^{k-} est la pseudo-inverse de rang k de V_λ , avec k = rang estimé de V_λ

et elle vérifie :

$X_\lambda = p_\lambda' \, V_\lambda^{k-} \, p_\lambda \approx \chi^2(k)$

Cette statistique asymptotique est utilisable pour étudier la significativité de f_λ . Le paramètre φ intervient dans l'expression de X_λ. Lors des tests statistiques, nous disposons de $\hat{\phi}$ et nonde φ (ce dernier étant inconnu). Il faut donc prendre en compte la distribution asymptotique de φ dans notre statistique. Le calcul de φ est effectué à partir de la somme des résidus qui convergent vers une loi du chi deux :

$m\frac{\hat{\phi}}{\phi} \approx \chi^2(m)$ *où m est le nombre de degré de liberté des résidus du modèle (nombre de données-nombre de degrés de liberté du modèle).*

En combinant ces deux statistiques qui sont supposées être indépendantes, il est alors possible d'étudier la significativité des fonctions de lissage f_λ^j .

$F_\lambda \frac{\hat{\phi}}{k\phi} \approx F(k,m)$

Ainsi l'hypothèse contrainte sera rejetée au seuil α si $F\lambda \geq F1-\alpha\,(k,m)$.

Il est possible d'associer la région critique à l'hypothèse Ho du modèle contraint :

W= $F_\lambda \geq F_{1-\alpha}\,(k,m)$

La frontière de cette région critique fournit l'intervalle de confiance associée au terme de lissage estimé.

[8] Résidus

Les résidus de Pearson (avec paramètre d'échelle Φ) sont définis par :

$$r_{ij} = \frac{y_{ij} - \hat{u}_{ij}}{\sqrt{\phi V(\hat{u}_{ij})}}$$

Les résidus de déviance (avec paramètre d'échelle Φ) sont définis par :

$$r_{ij} = sign(y_{ij} - \hat{u}_{ij})\sqrt{\frac{d_{ij}}{\varphi}}$$

où *dij* a été défini précédemment.

[9]Détails de la non convergence du modèle GLM du nombre de sinistres en fonction de l'âge du véhicule.

Informations sur le modèle	
Data Set	BASES.ECHANTRCC
Distribution	Poisson
Link Function	Log
Dependent Variable	NBSINRCC

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		**DDL**	**Valeur estimée**	**Erreur type**	**Intervalle de confiance de Wald à 95 %**		**Khi-2 de Wald**	**Pr > Khi-2**
Intercept		1	-18.6931	11466.42	-22492.5	22455.07	0.00	0.9987
AGECOND	18	1	23.6490	11466.42	-22450.1	22497.41	0.00	0.9984
AGECOND	19	1	23.9242	11466.42	-22449.8	22497.69	0.00	0.9983
AGECOND	20	1	23.9610	11466.42	-22449.8	22497.72	0.00	0.9983
AGECOND	21	1	23.9864	11466.42	-22449.8	22497.75	0.00	0.9983
AGECOND	22	1	24.0113	11466.42	-22449.8	22497.77	0.00	0.9983
AGECOND	23	1	23.9558	11466.42	-22449.8	22497.72	0.00	0.9983
AGECOND	24	1	23.9402	11466.42	-22449.8	22497.70	0.00	0.9983
AGECOND	25	1	24.0402	11466.42	-22449.7	22497.80	0.00	0.9983
AGECOND	26	1	24.0014	11466.42	-22449.8	22497.76	0.00	0.9983
AGECOND	27	1	23.8805	11466.42	-22449.9	22497.64	0.00	0.9983
AGECOND	28	1	23.9713	11466.42	-22449.8	22497.73	0.00	0.9983
AGECOND	29	1	23.8406	11466.42	-22449.9	22497.60	0.00	0.9983
AGECOND	30	1	23.9964	11466.42	-22449.8	22497.76	0.00	0.9983
AGECOND	31	1	23.9026	11466.42	-22449.9	22497.66	0.00	0.9983
AGECOND	32	1	24.0684	11466.42	-22449.7	22497.83	0.00	0.9983
AGECOND	33	1	24.0402	11466.42	-22449.7	22497.80	0.00	0.9983
AGECOND	34	1	23.9713	11466.42	-22449.8	22497.73	0.00	0.9983

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		DDL	Valeur estimée	Erreur type	Intervalle de confiance de Wald à 95 %		Khi-2 de Wald	Pr > Khi-2
AGECOND	35	1	24.0776	11466.42	-22449.7	22497.84	0.00	0.9983
AGECOND	36	1	24.0913	11466.42	-22449.7	22497.85	0.00	0.9983
AGECOND	37	1	24.1903	11466.42	-22449.6	22497.95	0.00	0.9983
AGECOND	38	1	24.1527	11466.42	-22449.6	22497.91	0.00	0.9983
AGECOND	39	1	23.9864	11466.42	-22449.8	22497.75	0.00	0.9983
AGECOND	40	1	24.1696	11466.42	-22449.6	22497.93	0.00	0.9983
AGECOND	41	1	24.1696	11466.42	-22449.6	22497.93	0.00	0.9983
AGECOND	42	1	24.1181	11466.42	-22449.6	22497.88	0.00	0.9983
AGECOND	43	1	24.0355	11466.42	-22449.7	22497.80	0.00	0.9983
AGECOND	44	1	23.9661	11466.42	-22449.8	22497.73	0.00	0.9983
AGECOND	45	1	24.0063	11466.42	-22449.8	22497.77	0.00	0.9983
AGECOND	46	1	23.9964	11466.42	-22449.8	22497.76	0.00	0.9983
AGECOND	47	1	23.9915	11466.42	-22449.8	22497.75	0.00	0.9983
AGECOND	48	1	23.8861	11466.42	-22449.9	22497.65	0.00	0.9983
AGECOND	49	1	23.8861	11466.42	-22449.9	22497.65	0.00	0.9983
AGECOND	50	1	24.0014	11466.42	-22449.8	22497.76	0.00	0.9983
AGECOND	51	1	23.7104	11466.42	-22450.1	22497.47	0.00	0.9984
AGECOND	52	1	23.9026	11466.42	-22449.9	22497.66	0.00	0.9983
AGECOND	53	1	23.8971	11466.42	-22449.9	22497.66	0.00	0.9983
AGECOND	54	1	23.6419	11466.42	-22450.1	22497.40	0.00	0.9984
AGECOND	55	1	23.7745	11466.42	-22450.0	22497.54	0.00	0.9983
AGECOND	56	1	23.8406	11466.42	-22449.9	22497.60	0.00	0.9983
AGECOND	57	1	23.6490	11466.42	-22450.1	22497.41	0.00	0.9984
AGECOND	58	1	23.7869	11466.42	-22450.0	22497.55	0.00	0.9983
AGECOND	59	1	23.4889	11466.42	-22450.3	22497.25	0.00	0.9984
AGECOND	60	1	23.6490	11466.42	-22450.1	22497.41	0.00	0.9984
AGECOND	61	1	23.4116	11466.42	-22450.4	22497.17	0.00	0.9984
AGECOND	62	1	23.2364	11466.42	-22450.5	22497.00	0.00	0.9984
AGECOND	63	1	23.0876	11466.42	-22450.7	22496.85	0.00	0.9984
AGECOND	64	1	23.0752	11466.42	-22450.7	22496.84	0.00	0.9984
AGECOND	65	1	23.0239	11466.42	-22450.7	22496.79	0.00	0.9984
AGECOND	66	1	22.8040	11466.42	-22451.0	22496.57	0.00	0.9984
AGECOND	67	1	22.8520	11466.42	-22450.9	22496.61	0.00	0.9984
AGECOND	68	1	23.0369	11466.42	-22450.7	22496.80	0.00	0.9984
AGECOND	69	1	22.5643	11466.42	-22451.2	22496.33	0.00	0.9984
AGECOND	70	1	22.9416	11466.42	-22450.8	22496.70	0.00	0.9984
AGECOND	71	1	22.5643	11466.42	-22451.2	22496.33	0.00	0.9984
AGECOND	72	1	22.7362	11466.42	-22451.0	22496.50	0.00	0.9984
AGECOND	73	1	22.8828	11466.42	-22450.9	22496.64	0.00	0.9984
AGECOND	74	1	22.6634	11466.42	-22451.1	22496.43	0.00	0.9984
AGECOND	75	1	22.5850	11466.42	-22451.2	22496.35	0.00	0.9984
AGECOND	76	1	22.7875	11466.42	-22451.0	22496.55	0.00	0.9984
AGECOND	77	1	22.7005	11466.42	-22451.1	22496.46	0.00	0.9984
AGECOND	78	1	22.6821	11466.42	-22451.1	22496.44	0.00	0.9984
AGECOND	79	1	22.4773	11466.42	-22451.3	22496.24	0.00	0.9984
AGECOND	80	1	22.6052	11466.42	-22451.2	22496.37	0.00	0.9984
AGECOND	81	1	22.4067	11466.42	-22451.4	22496.17	0.00	0.9984

Paramètres estimés par l'analyse du maximum de vraisemblance								
Paramètre		**DDL**	**Valeur estimée**	**Erreur type**	**Intervalle de confiance de Wald à 95 %**		**Khi-2 de Wald**	**Pr > Khi-2**
AGECOND	82	1	22.0604	11466.42	-22451.7	22495.82	0.00	0.9985
AGECOND	83	1	22.2767	11466.42	-22451.5	22496.04	0.00	0.9984
AGECOND	84	1	21.9120	11466.42	-22451.9	22495.67	0.00	0.9985
AGECOND	85	1	21.7842	11466.42	-22452.0	22495.55	0.00	0.9985
AGECOND	86	1	21.1780	11466.42	-22452.6	22494.94	0.00	0.9985
AGECOND	87	1	21.6376	11466.42	-22452.1	22495.40	0.00	0.9985
AGECOND	88	1	20.6390	11466.42	-22453.1	22494.40	0.00	0.9986
AGECOND	89	1	19.7917	11466.42	-22454.0	22493.55	0.00	0.9986
AGECOND	90	1	19.7917	11466.42	-22454.0	22493.55	0.00	0.9986
AGECOND	91	1	19.7917	11466.42	-22454.0	22493.55	0.00	0.9986
AGECOND	92	1	18.6931	11466.42	-22455.1	22492.46	0.00	0.9987
AGECOND	93	1	20.3026	11466.42	-22453.5	22494.06	0.00	0.9986
AGECOND	94	1	-0.0000	16211.07	-31773.1	31773.10	0.00	1.0000
AGECOND	95	1	-0.0000	16211.07	-31773.1	31773.10	0.00	1.0000
AGECOND	96	1	-0.0000	16211.07	-31773.1	31773.10	0.00	1.0000
AGECOND	97	1	-0.0000	16211.07	-31773.1	31773.10	0.00	1.0000
AGECOND	98	1	-0.0000	16211.07	-31773.1	31773.10	0.00	1.0000
AGECOND	99	1	-0.0000	16211.07	-31773.1	31773.10	0.00	1.0000
AGECOND	100	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

BIBLIOGRAPHIE

Ouvrages

[1] Mathématiques de l'assurance non-vie Tome II : Tarification et provisionnement, Michel Denuit & Arthur Charpentier, Economica, 1995.

[2] Generalized Additive Models, Monographs on statistics and applied Probability 43, Hastie & Tibshirani, Chapman & Hall/CRC,1990.

[3] An Introduction to Generalized Linear Models, A. Dobson, London: Chapman &Hall, 1990

[4] Spline models for observational data, G. Wahba, CBMS-NSF Reg. Conf. Ser.Appl. Math. :59, 1990.

[5] Modelling and smoothing parameter estimation with multiple quadratic penalties, Simon Wood. , J. R. Statist. Soc. B 62, 413-428, 2000..

Mémoires et cours

[6] Cours de statistique non paramétrique, Lucien Birgé et Arnak Dalalyan ISUP,2008.

[7] Cours de mathématiques de l'assurance, Christian Hess, ISUP, 2008.

[8] Cours d'assurances dommages, Jean-Marie Nessi, ISUP, 2008.

[9] Construction d'un tarif automobile, techniques de regroupement des modalités d'une variable qualitative, Laurence Serant, ISUP, 1993

[10] Une méthode alternative de provisionnement en assurance non-vie : les modèles additifs généralisés, Elise Lheureux, 2006.

Documents en ligne

[11] Supports SAS

http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/statug_gam_sect021.htm

http://v8doc.sas.com/sashtml/stat/index.htm

http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/whatsnew_toc.htm

[12] Le marché de l'assurance automobile en 2008, FFSA, direction des études, des statistiques et des systèmes d'information, Juillet 2009.

[13] Comparaison de modèles linéaires généralisés http://www.cict.fr/~stpierre/glim-genmod/node24.html