Université Claude Bernard – Lyon 1

INSTITUT DE SCIENCE FINANCIERE ET D'ASSURANCES

Mémoire présenté devant l'Institut de Science Financière et d'Assurances pour l'obtention du diplôme d'Actuaire de l'Université de Lyon

le 27 Janvier 2010

Par : Guillaume GONNET

Titre: ETUDE DE LA TARIFICATION ET DE LA SEGMENTATION EN ASSURANCE AUTOMOBILE.

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

Membre du jury de l'Institut des Actuaires

M. Pierre ARNAL

Membres du jury I.S.F.A.

M. Jean Claude AUGROS
M. Alexis BIENVENÜE
Mme Diana DOROBANTU
Mme Anne EYRAUD-LOISEL
M. Jean-Paul LAURENT
M. Nicolas LEBOISNE
M. Stéphane LOISEL
Mlle Esterina MASIELLO
Mme Véronique MAUME-DESCHAMPS
M. Frédéric PLANCHET
M. François QUITTARD-PINON
Mme Béatrice REY-FOURNIER
M. Didier RULLIERE

Secrétariat

Mme Marie-Claude MOUCHON

Bibliothèque :

Mme Michèle SONNIER

Entreprise :

Mutant Assurances

Directeur de mémoire en entreprise :

M. Stéphane LARTY

Invité :

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

Signature du responsable entreprise

Signature du candidat

50 Avenue Tony Garnier
69366 Lyon Cedex 07

RESUME

Mots clés : assurance automobile, tarification, segmentation, analyse factorielle, modèles linéaires généralisés, coût moyen, fréquence.

Ce mémoire a pour objet l'étude de la tarification et de la segmentation en assurance automobile. Nous allons ainsi nous intéresser à la problématique du marché de l'assurance auto en France et brièvement à ce que l'on appelle le « Pay as you drive » avant d'entrevoir les différentes étapes à suivre pour l'analyse et la modélisation du risque automobile. Les premières notions abordées ici sont l'extraction de données et l'analyse descriptive, préliminaires nécessaires qui seront illustrés par des exemples afin d'en comprendre l'importance. L'analyse factorielle et la modélisation linéaire généralisée du coût moyen et de la fréquence seront ensuite mis en avant. Nous nous attacherons plus particulièrement à ces deux derniers points. Nous détailleront ainsi les principes théoriques de différents types d'analyses factorielles telles que l'ACP, l'AFC ou l'ACM avant de nous intéresser à la théorie de la régression linéaire et à sa généralisation. La grande majorité des notions théoriques abordées ici seront accompagnées d'un exemple afin d'en faciliter la compréhension et d'en saisir l'utilité dans le cadre de cette étude.

ABSTRACT

Key words: car insurance, pricing, segmentation, factorial analysis, generalized linear models, average cost, frequency.

This dissertation is based on the study of pricing and segmentation in the case of car insurance. We are so going to be interested in the problematic of the market of the automobile insurance in France and shortly in the « Pay as you drive » systems before to see the different steps to be followed for the analysis and the modeling of the automobile risk. The first stages approached here are the extraction of data and the descriptive analysis, necessary preliminary which will be illustrated by an example in order to understand their importance. Then the factorial analysis and generalized linear modeling of average cost and frequency will be advanced. We'll see more particularly these last two points. We'll so detail the theoretical aspects of different type of factorial analysis such as PCA, COA or MCA before to see the theory of linear regression and its generalization. Most of the theoretical aspects studied here will be illustrated by an example in order to facilitate the understanding and to appreciate the utility of them in the case of the study.

REMERCIEMENTS

Je tiens tout d'abord à remercier Olivier Réant, directeur technique de Mutant Assurances, Stéphane Larty, responsable du service technique produit, et Alexandra Travincek, ma tutrice en entreprise, qui m'ont permis de réaliser ce mémoire autour d'une étude intéressante lors de mon stage de fin d'études, et qui m'ont fait bénéficier de leur expérience et de leurs conseils.

Je remercie également l'ensemble de la société Mutant Assurances, en particulier le plateau sinistres, pour leur accueil et leur disponibilité durant toute la durée de mon stage, et spécialement Serge Drevon, pour toutes ses connaissances qu'il n'hésite pas à partager.

Je remercie ensuite l'ISFA, notamment Alexis Bienvenue, mon tuteur pédagogique, et particulièrement Frédéric Planchet, pour ses précieux conseils et l'attention qu'il a pu porter à mon travail.

Je souhaite enfin remercier mon entourage, qui de près ou de loin, a su m'apporter sa confiance et son soutien.

SOMMAIRE

RESUME **2**

ABSTRACT **3**

REMERCIEMENTS **4**

INTRODUCTION **7**

PARTIE I : LE MARCHE DE L'ASSURANCE AUTOMOBILE EN FRANCE. **8**

Chapitre 1 : Problématique du marché de l'assurance auto en France. 8

Chapitre 2 : Principes de tarification et nécessité de segmenter. 10

Chapitre 3 : Une innovation constante. 14

PARTIE II : PRELIMINAIRES A LA MODELISATION DU RISQUE AUTOMOBILE. **19**

Chapitre 1 : Pré requis, limite et pertinence de la segmentation. 19

Chapitre 2 : Premières analyses et intuition de modélisation. 21

2.1 Analyse descriptive préalable. 21

2.2 Analyses factorielles et classification. 23

2.2.1 Principe. 23

2.2.2 Définitions et propriétés. 24

2.2.3 L'analyse en composantes principales (ACP). 27

2.2.4 L'analyse factorielle des correspondances (AFC) 38

2.2.5 L'analyse des correspondances multiples (ACM). 45

2.2.6 L'analyse dite mixte de données qualitatives et quantitatives 52

Chapitre 3 : Extraction, analyses préliminaires (Exemple numérique). 53

3.1 Extraction des données. 53

3.1.1 Première extraction et définition du périmètre. 53

3.1.2 Détection des erreurs, épurement des données. 55

3.2 Analyses descriptives. 58

3.3 Analyses factorielles. 63

3.3.1 Confrontation du zonier et de l'analyse des départements. 63

3.3.2 Analyse factorielle globale. 66

PARTIE III : MODELISATION DU RISQUE AUTOMOBILE ET ETUDE DES SEGMENTS DE TARIFICATION. **73**

Chapitre 1 : Les modèles linéaires généralisés. 73

1.1 Préliminaire : La régression linéaire. 73

1.1.1 Fondements théoriques. 73

1.1.2 Estimations et intervalles de confiances. 77

1.1.3 Choix du modèle et tests de significativité. 78
1.1.4 Détection d'erreurs et validation des hypothèses de modélisation. 81
1.2 Modèles linéaires généralisés. 84
1.2.1 La régression logistique. 84
1.2.2 La famille exponentielle. 86
1.2.3 Définition d'un modèle linéaire généralisé 89
1.2.4 Estimations des paramètres par maximum de vraisemblance. 91
1.2.5 Adéquation du modèle et tests de significativité. 96
1.2.6 Analyse des résidus et détection d'erreurs. 99
Chapitre 2 : Modélisation du risque en assurance automobile. 101
2.1 Fréquence et coût moyen. 101
2.2 Régression de poisson, sur dispersion et solutions. 102
2.2.1 Régression de Poisson. 102
2.2.2 Sur dispersion : définition, causes et détection. 103
2.2.3 Solutions d'améliorations. 104
Chapitre 3 : Un exemple de modélisation. 109
3.1 Présentation des données, méthodologie et enjeux. 109
3.2 Modélisation des montants de sinistres. 111
3.3 Modélisation des sinistres graves. 119
3.4 Modélisation de la fréquence de sinistre 120
3.5 Analyse des segments sur ou sous tarifés 123

CONCLUSION 127
BIBLIOGRAPHIE 129
SITES INTERNET DE REFERENCES 133
ANNEXES 134
ANNEXE 1 : ARTICLES DE LOI CONCERNANT LE COEFFICIENT BONUS MALUS 134
ANNEXE 2 : DELIBERATION DE LA CNIL (COMMISSION NATIONALE INFORMATIQUE ET LIBERTE) AU SUJET DE LA GEOLOCALISATION DE CONDUCTEURS. 136
ANNEXE 3 : DEMONSTRATION DE LA PROPRIETE 3 : THEOREME SPECTRAL EN DIMENSION FINIE 137
ANNEXE 4 : CARTES DES DEPARTEMENTS DE FRANCE. 139
ANNEXE 5 : REPRESENTATION D'UNE ACP SUR UNE CARTE DES REGIONS DE FRANCE. 141
ANNEXE 6 : EXEMPLE D'INTERFACE UTILISATEUR POUR L'ANALYSE DESCRIPTIVE. 143
ANNEXE 7: ENONCE DU THEOREME DE COCHRAN ET DEMONSTRATION. 147
INDEX DES ILLUSTRATIONS 149

INTRODUCTION

L'assurance automobile est une catégorie d'assurance omniprésente et diverse qui représente un marché très important (17,9 milliards d'euros de chiffre d'affaire en 2008, source : www.ffsa.fr [1]). L'obligation d'assurance est tout d'abord à signaler, de même que l'importance du parc automobile français, car ceci porte alors à un niveau conséquent le nombre d'assurés potentiels. Le chiffre d'affaire occasionné est également très élevé, ce qui amène logiquement tous les assureurs à proposer des contrats d'assurance automobile. Cette importance du marché automobile ne fait cependant pas de lui un marché où il est aisé de se développer et d'attirer la clientèle tout en conservant ses assurés d'origine, bien au contraire. L'assurance automobile est aujourd'hui un marché « hyperconcurrentiel ». C'est pourquoi on peut apercevoir une telle diversité de contrats au sein de cette branche de l'assurance. De nombreuses garanties annexes telles que la protection juridique ou l'assistance sont ainsi de plus en plus développées. De nombreuses offres existent également, bien souvent autour de l'évolution du coefficient bonus malus ou de la personnalisation du contrat d'assurance. La concurrence observable autour des garanties existe bien évidemment aussi autour des prix, en baisse depuis plusieurs années. Cette baisse, en partie due au désir de la sécurité routière de voir les évolutions de comportements sur la route récompensées, porte aujourd'hui les primes d'assurance automobile à un niveau proche du coût technique.

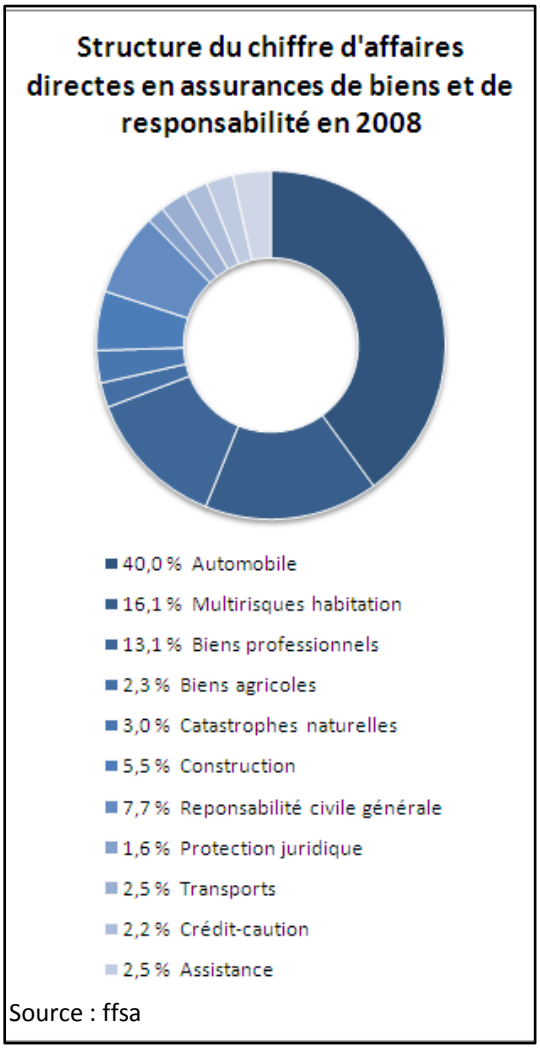
Dans un tel contexte de concurrence, on comprend la nécessité de posséder une tarification précise et adaptée à son portefeuille. Cette tarification doit de même s'appuyer sur une segmentation plus ou moins poussée mais pertinente. De plus, cette même segmentation ne saurait rester fixe dans le temps et il est alors nécessaire d'effectuer un suivi. Ce suivi doit alors avoir pour but de contrôler l'adéquation de la tarification à la population assurée mais également la mise en évidence de segments sur ou sous tarifés, représentations éventuelles de sous populations soit porteuses d'un risque supérieur à la moyenne, soit porteuses d'un risque plus faible. Une telle étude, qui doit être comparée avec les objectifs de l'entreprise afin d'être interprétée de façon sensée permet ainsi de faire évoluer la tarification en fonction du portefeuille et des opportunités d'actions qui se présentent à l'assureur.

Elle se base sur l'utilisation d'outils actuariels et se décompose en plusieurs étapes que nous allons décrire en développant le plan suivant. Nous commencerons par nous intéresser au contexte économique et à la problématique du marché de l'assurance automobile en France dans une première partie qui décrira alors le milieu de l'étude et exposera des grands principes tels que le coefficient bonus malus ou la segmentation, tout en ouvrant le débat sur l'évolution vers un tarif d'assurance individualisé avec ce que l'on appelle le « Pay as you drive » (PAYD). Puis nous nous tournerons dans une deuxième partie vers les étapes préliminaires de la modélisation du risque automobile. Nous évoquerons ainsi l'importance des données extraites et les premières intuitions procurées par des analyses descriptives et factorielles, dont nous décriront la théorie d'une part et que nous illustrerons avec un exemple concret d'autre part. Enfin nous nous attarderons sur le principe de la modélisation linéaire généralisée, outil capital de l'étude de la tarification, que nous illustrerons encore une fois avec un exemple numérique concret.

Partie I : Le marché de l'assurance automobile en France.

Chapitre 1 : Problématique du marché de l'assurance auto en France.

Le marché de l'assurance automobile en France est l'un des plus importants, comme en témoigne le rapport annuel 2008 de la ffsa (fédération française des sociétés d'assurances) [1]. En effet, il est le marché le plus développé de l'assurance de biens avec 17,9 milliards d'euros de cotisations en 2008, bien loin devant l'assurance multirisque habitation ou MRH avec 7,2 milliards d'euros de cotisations. De même, le chiffre d'affaire direct en assurances de biens et de responsabilité en 2008 est composé à 40 % par l'assurance automobile (voir graphiques infra). L'importance de cette branche d'assurance s'explique en grande partie par l'obligation d'assurance automobile minimale en responsabilité civile dite au tiers, mais également par la volonté des assurés prudents considérés comme de bons risques de se couvrir au mieux contre ce risque quotidien, tout en voyant leurs primes diminuer en récompense de leur bon comportement. En 2008, 82 % des automobilistes ont souscrit une garantie incendie-vol et 62 % une garantie pour les dommages subis par leur véhicule. Il est également à noter que près des deux tiers des automobilistes ont atteint un CRM (Coefficient de réduction majoration ou coefficient bonus malus) de 50 %. Il faut enfin noter que *«l'automobile, régulièrement renouvelée dans les foyers, représente un enjeu de taille, car elle peut faire changer d'assureur »* (Clotilde Briard, journaliste).



Source : ffsa

Illustration 1

Les assurances de biens et de responsabilité		
	Cotisations 2008 en milliards d'euros	Variation en %
Automobile	17,9	+ 1,0
Multirisques habitation	7,2	+ 6,4
Biens professionnels	5,9	+ 1,0
Biens agricoles	1,0	+ 8,5
Catastrophes naturelles	1,4	+ 0,6
Construction	2,5	+ 0,2
Reponsabilité civile générale	3,4	+ 3,5
Protection juridique	0,7	+ 8,0
Transports	1,1	+ 5,1
Crédit-caution	1,0	+ 0,3
Assistance	1,1	+ 7,7
Pertes pécuniaires	1,6	+ 1,1
Total des affaires directes	44,8	+ 2,5

Source : ffsa

Illustration 2

Ce marché d'assurances de biens est certes le plus important, mais il est également le plus concurrentiel. De nombreux acteurs se disputent inlassablement les parts de marchés. De plus, et contrairement aux autres assurances qui sont en augmentation en termes de cotisations, l'assurance automobile stagne depuis quelques années, de même que le parc automobile dont la croissance avoisine environ 1 % depuis 2002 (Source ffsa [1]). Le marché de l'assurance est donc aujourd'hui qualifié d'hyperconcurrentiel. Les principaux acteurs du marché de l'assurance automobile, représentés ci-dessous dans le classement des assureurs automobile en 2007 (Argus de l'assurance [5]), doivent ainsi rivaliser d'une part en termes de prix, et d'autre part en termes d'inventivité.

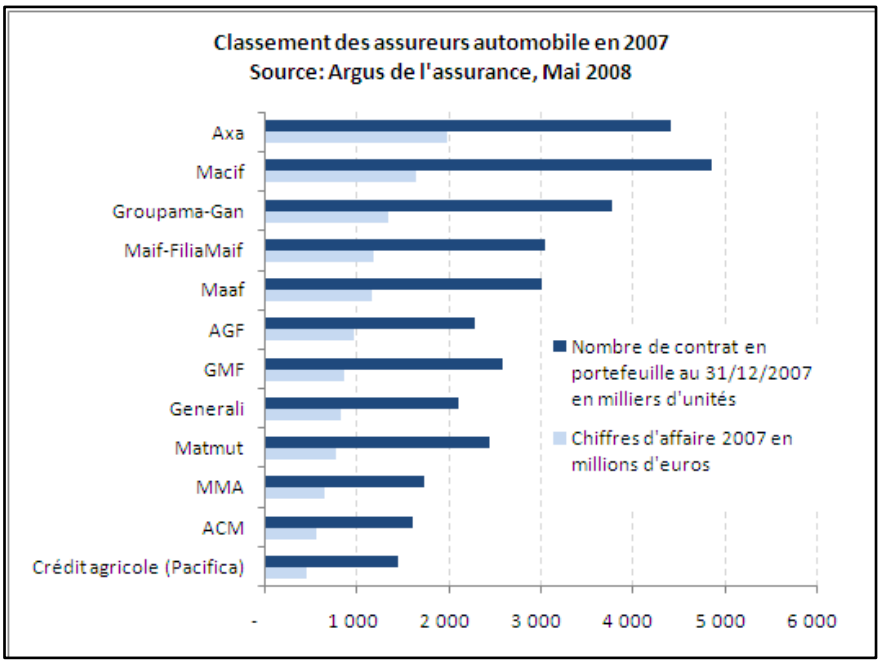


Illustration 3

Les acteurs habituels de ce marché que sont les sociétés d'assurances et les mutuelles doivent de plus faire face à l'arrivée sur le marché des bancassurances. Ces dernières peuvent utiliser un portefeuille de clients déjà important et source d'informations appropriées afin d'attirer de nombreux clients à partir d'offres intéressantes et de bas prix permis par un réseau d'agence important et l'absence d'intermédiaire, quand d'autres utilisent les services d'agents généraux et de courtiers.

Durant ces dernières années, on a assisté à une baisse non négligeable des primes moyennes demandées aux assurés (-0,9 % en 2008 et -2,0 % en 2007, source ffsa [1]). Cette baisse des prix encouragée par l'Etat, dans le but de récompenser l'amélioration des comportements des Français au volant, a tendance à s'essouffler aujourd'hui. Ce temps d'arrêt dans la baisse des tarifs a plusieurs explications. Tout d'abord, les tarifs ont presque rejoint les coûts techniques, diminuant ainsi très fortement la marge de manœuvre des assureurs. De plus, bien que la mortalité sur les routes française ait reculée de 1,9 % en 2007 et 8,2 % en 2008, le nombre de blessés à quant à lui augmenté de 2,4 % en 2007, avant de baisser à nouveau en 2008 (source sécurité routière [12]). A cela se rajoute l'inflation des coûts des dommages corporels, de 6,5 % par an sur les dix dernières années (source ffsa [1]), en raison entre autre d'hospitalisation viagère plus fréquente qu'auparavant. Il en est de même pour les coûts de réparations qui ont fortement augmentés. Malgré cela, on imagine mal qu'un assureur puisse subitement relever ses tarifs dans un tel contexte de concurrence.

Chapitre 2 : Principes de tarification et nécessité de segmenter.

L'assurance automobile s'inscrit dans un cadre particulier. En effet, elle joue un rôle important en lien avec la sécurité routière et prend une réelle part à l'incitation à la prudence ainsi qu'à la prévention. De nombreux acteurs de l'assurance mettent d'ailleurs en place des journées de prévention et des stages afin de sensibiliser leurs assurés (70 000 personnes par an suivent un stage de sensibilisation à la conduite par le biais de Groupama par exemple). Dans ce cadre, la législation française impose la prise en compte par les assureurs d'un coefficient de réduction majoration (CRM) ou coefficient bonus malus dont nous allons détailler les principales caractéristiques ;

- Les modalités de calcul et de prise en compte du CRM sont définies par l'article A 121-1 du code des assurances (Voir Annexe 1).
- Lors de la première année d'assurance, le CRM est fixé à 1 (aucune modification de la prime d'assurance de base fixée par l'assureur).
- Le CRM est transférable d'un assureur à un autre, il est propre à l'assuré en cas de changement d'assureur ou de changement de véhicule, mais il est assigné au véhicule durant la vie du contrat.

Exemple :

Monsieur X est assuré pour son véhicule avec un CRM de 0,6. Madame X, qui a un CRM de 0,9 au titre de son véhicule, emprunte la voiture de son mari et a un accident responsable. En théorie (sauf cas particulier en fonction des assureurs), le CRM qui sera majoré sera celui de monsieur X. Dans ce cas le CRM est bien assigné au véhicule.

Suite à l'accident, monsieur X décide de changer de véhicule, le CRM qui sera pris en compte lors de l'avenant à son contrat sera le sien, soit 0,6. Arrivé à son échéance, il décide de changer d'assureur, là encore et pour son nouveau contrat, le CRM pour l'année à venir sera le sien, ici de 0,75 (à cause du sinistre qui a eu lieu, voir plus loin).

- Chaque année sans sinistre procure une réduction de 5 % du coefficient de l'année précédente, pour l'année à venir. Cette réduction est conditionnelle à une durée d'assurance d'au moins 9 mois lors de l'année précédente. Ainsi, une suspension d'assurance d'une durée supérieure à 3 mois ne permettra pas d'obtenir une réduction du coefficient.
- Le bonus est cumulable dans une certaine limite. Un assuré sans sinistre verra en effet son CRM diminuer de 5 % chaque année jusqu'à un minimum de 0,5 (réduction de 50 % de la prime). Treize années sans sinistre sont nécessaires pour obtenir ce CRM minimum de 0,5 la quatorzième année.
- En cas de sinistre responsable et s'il y a plus de 2 mois entre la date d'échéance du contrat et la date de sinistre, une majoration du CRM sera appliquée pour l'année à venir. Le CRM sera multiplié par 1,25 pour chaque sinistre totalement responsable et par 1,125 pour chaque sinistre à responsabilité partielle.
- En cas de sinistre responsable et s'il y a moins de 2 mois entre la date d'échéance et la date du sinistre, il y a deux cas à distinguer. S'il n'y a pas de changement d'assureur, ce dernier prendra en compte le sinistre et donc la majoration du CRM non pas pour l'année à venir mais pour l'année après celle-ci. En cas de changement d'assureur, le nouvel assureur prendra en compte le sinistre et donc la majoration du CRM immédiatement pour l'année à venir.

Exemple :

Evolution du CRM selon le changement ou non d'assureur de l'année N à l'année N+1 dans l'hypothèse où l'assuré à un coefficient de 0,6 au titre de l'année N, un sinistre responsable l'année N moins de deux mois avant l'échéance, puis aucun sinistre l'année N+1.

	Pas de changement d'assureur	Changement d'assureur
Année N	**0,6**	**0,6**
Année N+1	0,6*0,95 = **0,57**	0,6*1,25 = **0,75**
Année N+2	0,57*1,25 = **0,71**	0,75*0,95 = **0,71**

Illustration 4

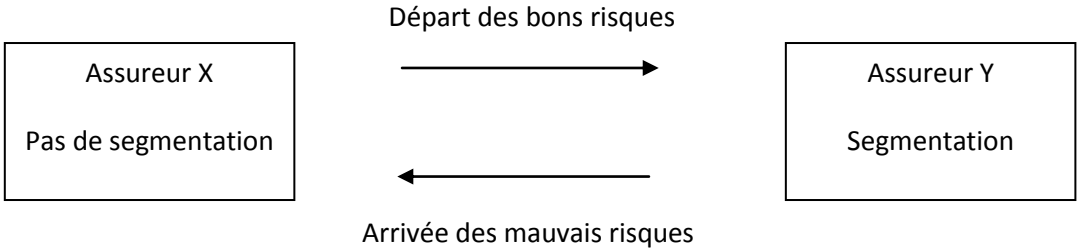
On peut voir sur cet exemple que dans cette hypothèse, il est préférable pour l'assuré d'attendre une année supplémentaire pour changer d'assureur. Son CRM sera le même pour l'année N+2, mais il aura économisé durant l'année N+1 grâce à un CRM de 0,57 au lieu de 0,75.

- Le CRM ne peut atteindre une valeur supérieure à 3,5.
- En cas d'absence de sinistre responsable durant deux années successives, le CRM est automatiquement ramené à 1. On parle alors de descente rapide.
- Si le bonus de 0,5 est acquis depuis au moins trois années d'assurance pleines, alors le premier sinistre responsable n'impliquera pas de majoration du coefficient.
- Ne sont pas pris en compte dans le calcul du CRM les sinistres non responsables, les sinistres causés par un auteur conduisant le véhicule à l'insu de l'assuré (sauf s'il vit habituellement au foyer d'un des conducteurs désignés), les sinistres survenus sur un véhicule en stationnement par les faits d'un tiers non identifié, les sinistres dont la cause est assimilable à la force majeure ou encore les sinistres pour lesquels la garantie concernée est le bris de glace, le vol ou l'incendie.
- Lorsque le véhicule est utilisé pour un usage « tournée » ou « tout déplacement », la réduction en cas d'absence de sinistre est alors de 0,93 (soit -7 %) au lieu de 0,95 et la majoration en cas de sinistre responsable est de 20 % par sinistre au lieu de 25 %.
- Les règles ci-dessus concernant le calcul du CRM sont obligatoires et applicables par toutes les sociétés d'assurances en France, le coefficient de réduction majoration devant impérativement apparaitre clairement sur le contrat d'assurance automobile. Cependant et en pratique, on observe un très grand nombre de dérogations à ces règles, car l'assureur, bien qu'obligé d'inscrire le CRM de l'assuré au contrat, peut parfaitement introduire des réductions ou des majorations supplémentaires lors de la tarification du contrat et de la détermination de la prime sur laquelle viendra enfin s'appliquer le CRM. On observe ainsi un très grand nombre de campagnes publicitaires des assureurs promettant un bonus à vie ou une réduction du CRM deux fois plus rapide que la normale. En pratique, l'assureur ne change pas la méthode de calcul du CRM, il ajoute simplement une réduction supplémentaire, valable uniquement dans sa compagnie, et donc non transférable en cas de changement d'assureur.

Au-delà de ce mécanisme de bonus malus, chaque assureur possède une tarification et une segmentation qui lui est propre, en adéquation avec le profil de son portefeuille d'assurés. Dans un contexte de marché très concurrentiel, la segmentation des risques est une nécessité. Celle-ci consiste à différencier les assurés et le risque qu'ils portent. On obtient ainsi différentes catégories de risques en fonction des caractéristiques de l'assuré et des garanties consenties par l'assureur. Chaque catégorie se verra ainsi attribuer un tarif qui lui sera propre, en adéquation avec le risque associé. Nous donnons ci-dessous une définition de la segmentation donnée en cours d'actuariat en master « ingénierie statistique et actuariat » (IMA).

La segmentation consiste à analyser et contrôler l'adaptation des primes aux sinistres suivant des classes de risques homogènes, de façon à en tirer des conséquences du point de vue technique. La segmentation permettra de prendre des mesures techniques à chacun des niveaux de segmentation, tant en tarification, qu'en souscription.

Prenons un exemple simple afin de comprendre la nécessité de segmenter, tout particulièrement en assurance automobile. L'assureur X décide de n'effectuer aucune segmentation. Il fait donc un profit avec les clients à faible probabilité de sinistre et une perte avec les assurés à forte probabilité de sinistre. L'assureur Y quant à lui, décide de faire payer plus cher les conducteurs à risques, déterminés à partir de variables telles que l'âge ou les antécédents de sinistres. En contrepartie, l'assureur Y diminuera la prime demandée aux assurés à risque faible. Suite à cette segmentation, les assurés prudents, ayant une faible probabilité d'avoir un sinistre et que l'on qualifiera de bons risques, décideront de s'assurer auprès de l'assureur Y, proposant des primes plus faibles. L'assureur X qui propose quant à lui un tarif unique, verra venir s'assurer chez lui la plupart des personnes ayant un comportement à risque, et donc une forte probabilité de sinistre. En effet, le tarif qui sera proposé à ces individus qualifiés de mauvais risques sera bien plus intéressant que celui de l'assureur Y. En conclusion, l'assureur X devra faire face à un risque de sinistre très élevé qu'il ne pourra pas absorber étant donné les primes demandées aux assurés. L'assureur X ne fera alors que des pertes, tandis que l'assureur Y ne fera que du profit.



Cet exemple simplifié exprime bien la nécessité pour l'assureur de segmenter. On peut également s'apercevoir à partir de cet exemple que la segmentation, si elle diminue la prime des conducteurs prudents, augmentera également sensiblement la prime des conducteurs à risque. Ce constat se révèle particulièrement problématique en ce qui concerne l'assurance automobile. En effet, les jeunes conducteurs, qui ne bénéficient pas d'ancienneté d'assurance, sont considérés comme une catégorie à risque et doivent souvent faire face soit au refus des assureurs, soit à des primes importantes. Conduire un véhicule peut donc devenir extrêmement coûteux pour un nouveau

titulaire du permis de conduire. Se pose alors une question de déontologie en ce qui concerne la tarification et la segmentation en assurance automobile. Cette classe de risque des 18-25 ans est en effet particulièrement délaissée par les acteurs du marché, bien qu'un traitement de faveur soit souvent consenti pour les enfants des assurés (faveur que l'on retrouve particulièrement chez les sociétaires des mutuelles). On pourra alors citer Etienne Couturier, Directeur Général Délégué de la MAAF : *« Personne n'a trouvé la bonne solution. Techniquement, c'est un mauvais risque. Et, commercialement, chacun s'occupe des enfants de ses clients en priorité ».* De nouveaux acteurs du marché cherchent cependant à attirer cette classe d'assurés, il s'agit des bancassurances, qui cherchent à renouveler leur portefeuille de clients en attirant de nouveaux assurés qui deviendront probablement de nouveaux clients pour la banque.

Les jeunes ne sont pas la seule catégorie que la segmentation amène à délaisser. On peut alors citer les assurés résiliés pour non paiements et tous les assurés ayant vu leur permis retiré pour une raison ou pour une autre. Il n'en reste pas moins que l'assurance automobile est obligatoire. Un conducteur porteur d'un risque important ne possède alors pas beaucoup de solutions. En premier lieu il peut se tourner vers des assureurs spécialisés dans ce que l'on appelle parfois les « mauvais risques », on retrouve alors les bancassurances, certains courtiers tels que « Mutant Assurances » et même une compagnie comme AXA. La garantie d'assurance sera cependant minimale, avec parfois des montants de primes très importants. En dernier recours, le conducteur « indésirable » a toujours la possibilité de se tourner vers le Bureau Central de Tarification (BCT), qui obligera un assureur à le couvrir (uniquement pour la garantie responsabilité civile qui est obligatoire), et qui fixera une prime en fonction des tarifs de référence de la société choisie par le conducteur.

On comprend à travers ces constats que la tarification, bien que nécessaire pour l'assureur, soulève un certain nombre de problématiques, et particulièrement en ce qui concerne la couverture des jeunes conducteurs. Il convient donc de rester prudent et de ne pas segmenter à l'extrême sans réflexion préalable, auquel cas l'assureur prendrait le risque de ne pas renouveler son vivier d'assurés, qui est bien souvent vieillissant dans les grandes compagnies d'assurances.

Chapitre 3 : Une innovation constante.

Au-delà de la baisse des prix, on assiste à une surenchère permanente en matière d'innovation. On voit ainsi arriver sur le marché de nouvelles réductions offertes aux meilleurs conducteurs, de nouvelles garanties telles que l'assistance ou la protection juridique, et une segmentation toujours plus poussée des assureurs. Les objectifs des assureurs sont alors multiples, il est tout d'abord nécessaire d'avoir une politique de prix en adéquation avec le marché. On assiste ainsi à une baisse quasi constante des prix, mais également à une démarche grandissante de comparaison effectuée par les assurés et facilitée par Internet. Certains assureurs ont même choisi de mettre en place des propositions d'assurances dites « low cost ». Ensuite, et afin de conserver une certaine marge malgré la baisse des prix, les assureurs se doivent d'une part d'attirer puis de garder les assurés considérés comme de bons risques, et d'autre part *« d'offrir des services dont la valeur est difficilement chiffrable par les consommateurs »* (Cyrille Chartier-Kastler, Solving). Dans ce but la plupart des assureurs sont aujourd'hui associés à un spécialiste de l'assistance, tel que « Mondial Assistance » pour AGF, « Europ Assistance » pour Generali ou « GIE Ima » pour les mutuelles Maif, Maaf, Macif et Matmut. On assiste également à la multiplication des partenariats avec un nombre limité de réparateurs chez lesquels l'assuré est prié de se rendre afin que l'assureur puisse négocier les prix et ainsi réduire ses coûts de sinistres. Un autre terrain d'innovation et de concurrence entre assureurs est celui d'Internet, qui malgré une mise en place et une acceptation relativement lente en France, ne doit surtout pas être négligé aujourd'hui. Enfin, la mise en place d'un tarif personnel et sur mesure devient réellement envisageable avec l'apparition de nouvelles technologies d'acquisition d'informations utilisées dans les contrats d'assurances automobile que l'on appelle PAYD (Pay as you drive ou littéralement payez comme vous conduisez), qui poussent la segmentation à l'extrême et offrent de nouveaux horizons à l'assurance automobile. Nous allons dans la suite de cette partie revenir sur chacun de ces points afin de les étudier plus en profondeur.

Dans un contexte de concurrence, les principaux acteurs mettent en avant par le biais de la publicité des offres toutes plus intéressantes les unes que les autres. Le secteur de l'assurance automobile ne déroge pas à cette règle, bien au contraire. On pourra ainsi évoquer le « Bonus à vie » accordé par la Maaf aux conducteurs justifiants de trois années pleines avec un CRM de 0,5, et ce même en cas de sinistre responsable. On pensera également au « Bonus double effet » de MMA qui permet d'acquérir le CRM minimal en 8 ans au lieu de 13 et de le conserver définitivement. Les assurés automobilistes sont très friands de ce genre d'approches, quelle aubaine pour les conducteurs prudents d'avoir la garantie de garder définitivement ce coefficient minimal si difficile à atteindre. Bien entendu, ce type de garantie ne reste valable qu'au sein de l'assurance ou de la mutuelle qui veux bien la consentir, et l'on comprend alors la grande capacité de fidélisation d'une telle démarche, d'autant plus qu'elle concerne ceux que l'on appelle les bons risques, qui représentent aujourd'hui plus de 60 % des conducteurs. Il se trouve de plus que cette mesure, qui soulève un risque limité, a un coût raisonnable. Les assureurs qui souhaitent enrayer le phénomène grandissant de résiliation que l'on retrouve en assurance automobile ont donc tout intérêt à mettre en place un système de réduction attractif similaire à ceux évoqués plus haut.

Nous avons évoqué à plusieurs reprises la guerre des prix livrée par les différents acteurs de l'assurance automobile. Celle-ci devient d'autant plus importante que les assurés n'hésitent plus à comparer les prix et à se tourner vers le plus offrant. Cette comparaison a bien souvent lieu sur Internet, par le biais de comparateurs tels que « Assurland » ou « HyperAssur ». Internet devient

également un lieu de souscription, ou tout du moins de premier moyen d'information, de première prise de contact. La quasi-totalité des compagnies proposent ainsi à d'éventuels futurs clients d'effectuer un devis gratuit en direct, et s'ils le souhaitent, ils pourront alors souscrire un contrat. Malgré des débuts difficiles, la souscription par Internet prend petit à petit de l'ampleur. Internet devient alors un nouveau terrain de concurrence et une source de profit qui ne peut être négligée par les assureurs. Le défi à relever est alors le suivant ; proposer des garanties équivalentes à celles proposées en agence tout en permettant à l'assuré de bénéficier de coûts plus faibles grâce à la diminution ou à l'absence de coûts d'acquisition et de gestion pour l'assureur. Il est cependant nécessaire de ne pas perdre de vue l'accompagnement de l'assuré en mettant à sa disposition la possibilité de joindre un conseiller par le biais d'une plateforme téléphonique de qualité.

Toutefois, les offres que l'on retrouve sur Internet sont parfois bien différentes de celles proposées par les agences. On trouve en effet sur Internet des offres d'assurances automobile à bas prix répondant au phénomène grandissant de « low cost ». Les principaux acteurs de ce marché des bas prix sont aujourd'hui « Ineas », « Direct Assurance », la « Macif » par le biais du site internet « Idmacif.fr » et « Groupama » avec le site « Amaguiz.com ». Tous parient sur la capacité de ces offres « low cost » à amorcer le décollage d'Internet comme *« canal de distribution autonome, pour des clients autonomes, sur des produits standards répondant à des besoins de masse »*, selon la définition de l'Observatoire de l'évolution des métiers de l'assurance [13]. Ce type d'offre ne se retrouve néanmoins pas uniquement sur Internet. Le « low cost » devient en effet le fer de lance de certaines compagnies souhaitant attirer les jeunes conducteurs, telles que « Mutant Assurances ». Ces bas prix ne sont cependant pas sans contrepartie. Ils sont en effet possible grâce à la suppression d'un certains nombres de garanties, le produit est déshabillé afin de ne retenir que l'essentiel. Sont également à surveiller la présence de franchises, de surprimes ou de plafonds de remboursements. L'enjeu des assureurs est alors de proposer des contrats à faibles coûts tout en prenant garde de ne pas ternir leur image. En effet, la mise en place de tels contrats est à double tranchant selon l'observatoire de l'évolution des métiers de l'assurance qui rappelle dans son baromètre 2008 [13] que *« gagner sur le montant des sinistres payés est (…) une stratégie extrêmement risquée en terme d'image du fait des insatisfactions que générerait une réparation trop limitée des dommages subis »*. Le marché de l'assurance automobile à bas prix est un marché très délicat qui reçoit de vives critiques comme celles d' « Assurland » envers « Mutant Assurances » : *« C'est typiquement du faux low-cost (…) On abaisse les prix au détriment de la qualité ».* Ce type d'assurance n'est pourtant pas dénué de sens, comme on le soutient chez « Mutant Assurances » ; *« Notre offre de produit est centrée sur l'essentiel, nous ne fournissons pas de garanties gadgets, mais celles les plus plébiscitées par les clients et les plus adaptées à la majeure partie des besoins. »* (Carole Bérard, directrice de la communication chez Mutant Assurances).

Si l'on assiste à un véritable déshabillage des contrats d'une part, on constate l'apparition de multitudes de garanties optionnelles souvent packagées d'autre part. Les assurés français s'avèrent en effet être de plus en plus exigeant en ce qui concerne les services d'assistance procurés par leur assureur. Comme nous l'avons souligné précédemment, la grande majorité des assureurs a profité de cette occasion pour personnaliser son offre à travers tout un panel de garanties d'assistance en partenariat avec des professionnels de ce métier. L'assistance zéro kilomètre est aujourd'hui une garantie qui ne surprend plus personne, de même que la protection juridique de l'assuré, le remplacement du véhicule en cas de sinistre ou bien encore la mise à disposition d'un chauffeur, autant de garanties annexes difficilement chiffrables par l'assuré, mais qui rassurent. Ce type de

garanties donne alors tout son sens au terme de « qualité » des produits et services offerts par l'assureur. Un assuré satisfait des services procurés par son assureur sera plus enclin à renouveler son contrat, ou tout du moins à ne pas le résilier. Il en de même en ce qui concerne la rapidité d'action, la possibilité de joindre un conseiller, d'être informé, le temps d'attente au téléphone ou en agence, autant de notions qui reflètent la qualité des services procurés par l'assureur. Au-delà du tarif proposé, l'assureur se doit donc d'innover, et au-delà même des garanties qu'il peut offrir, il se doit de les accompagner, en partenariat avec différentes structures, d'une qualité à laquelle il s'engage auprès de ses assurés.

Nous allons enfin nous intéresser aux contrats utilisant le système PAYD (Pay as you drive ou payez comme vous conduisez). Ce système consiste à demander à l'assuré d'installer dans son véhicule un instrument de mesure appelé odomètre, permettant de connaitre la distance parcourue. Cet élément peut alors être couplé avec un GPS (Global Positionning System ou Système de géo localisation par satellite). Les données enregistrées sont ensuite transmises via le réseau de téléphonie mobile (GSM : Global System for Mobile communication). La quantité et la diversité des données pouvant être exploitées sont alors très importantes, allant du simple relevé du nombre de kilomètres parcourus au décompte du nombre de freinages brusques. Toutes ces données sont alors agrégées et stockées par un intermédiaire puis redirigées après retraitement dans le système de gestion de la compagnie. L'assureur peut alors à partir de ces données établir un tarif personnalisé pour chaque assuré. Le mécanisme du système PAYD est repris par le schéma ci-dessous.

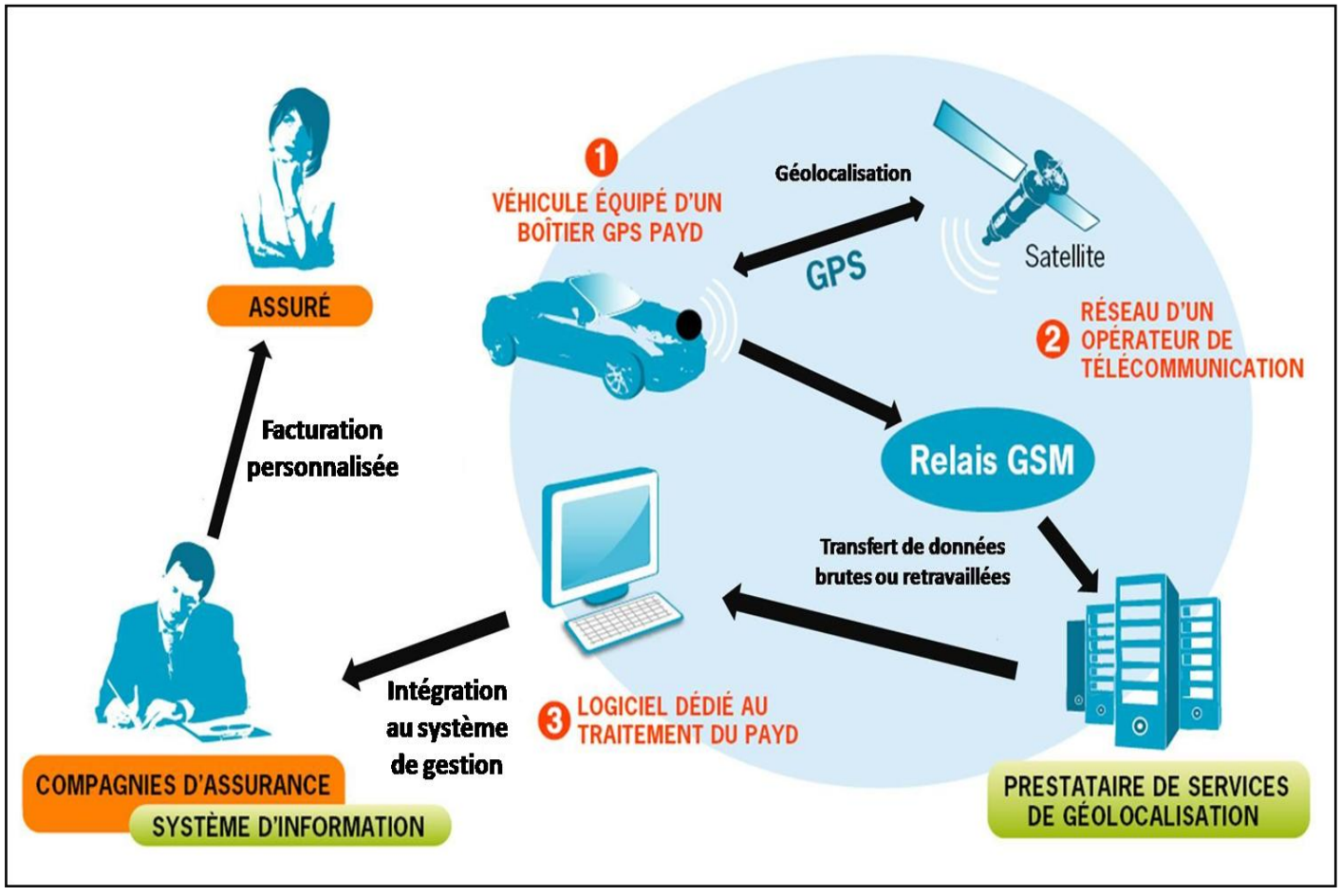


Illustration 5 : Schéma récapitulatif du système PAYD

Le PAYD a vu le jour en 1996 aux Etats-Unis par le biais d'une offre de « Progressive Insurance ». Ce système a été lancé entre 2004 et 2006 en Grande Bretagne et en Italie par un grand nombre d'acteurs. Le succès de ce type d'assurance n'a alors pas tardé à se faire sentir, suscitant la curiosité dans le monde de l'assurance. L'engouement autour de cette nouveauté est alors général, et tous les grands groupes présents sur le marché Français s'intéressent de près ou de loin à cette innovation prometteuse capable de bouleverser le monde de l'assurance automobile. En effet, les solutions apportées par le PAYD sont étonnantes.

Le principal intérêt de ce système réside dans les possibilités de segmentation qu'il apporte, et qui deviennent alors quasiment illimitées. Alors que les variables explicatives étaient limitées aux caractéristiques de l'assuré ou de son véhicule ainsi qu'aux évènements passés, il devient aujourd'hui possible de faire payer l'assuré en fonction de son comportement en temps réel. Il est ainsi possible de contrôler le kilométrage parcouru, de connaitre l'heure d'utilisation du véhicule ainsi que le type de routes empruntées (Ville, Nationale, Autoroute…). On peut également s'intéresser à la vitesse, aux chocs, au temps de voyage entre deux arrêts (on rappellera qu'il est conseillé de faire une pause toutes les deux heures, le temps de réaction étant deux fois plus long au-delà de deux heures de conduite, d'où l'intérêt du dernier point). Il devient même possible d'étudier le comportement de l'assuré en termes de coups de volant ou de freinages brusques, même si cette dernière analyse est très critiquée et pose la question de savoir quelle est la définition d'une conduite « brusque » et son impact sur la probabilité de sinistre.

Le développement d'un tel système en France est cependant très limité de nos jours, et l'évolution du PAYD y est bien plus faible que l'engouement constaté en Grande Bretagne ou en Italie ne l'aurait laissé présagé. Il existe un certain nombre de freins qui bloquent en effet l'évolution de ce système en France. Le premier problème se situe au niveau de la réglementation. Une première tentative de lancement d'un PAYD par la Maaf en 2005 fut en effet stoppée par une décision de la CNIL (Commission Nationale Informatique et Liberté). Cette interdiction était due au fait que le contrat prévoyait la géolocalisation permanente des conducteurs, et la détention illégale de données telles que les dépassements de vitesses maximales autorisées, que seule l'autorité publique peut avoir en possession (Voir Annexe 2). Un deuxième frein au développement réside dans les coûts. Les primes d'assurances demandées par les assureurs en France sont 2 à 3 fois plus faibles que celles demandées en Italie ou en Grande Bretagne (en moyenne 400 euros annuel pour la France). Ceci s'explique par une forte fréquence de vols et de fraudes en Italie et un médiocre réseau routier en Grande Bretagne. Les gains potentiels pour les assurés de ces deux pays sont alors élevés, ce qui attire de nombreuses personnes, mais il n'en est pas de même en France où les primes sont déjà très basses. Se pose alors la question de savoir si un assuré acceptera de payer pour l'installation d'un boitier au sein de son véhicule dans le but d'obtenir une moindre réduction de prime.

Le PAYD trouve cependant d'autres avantages qui permettraient, malgré les limitations et problématiques précités, d'acquérir un avantage concurrentiel important chez les assureurs le mettant en place. Le système offre en effet de nombreuses possibilités concernant d'éventuelles options supplémentaires. On peut en effet imaginer une garantie de traçage en cas de vol, de dépannage d'urgence, d'assistance immédiate en cas de sinistre voir même d'aide ou de preuve concernant le remplissage d'un constat. On voit bien que les possibilités sont multiples et que l'inventivité des assureurs sera sollicitée. Le PAYD possède également des vertus concernant la sécurité routière et l'écologie, qui sont deux notions très en vogue en ce moment, aussi bien pour les

assurés que pour les assureurs. Ce système possède en effet un aspect psychologique non négligeable capable d'influencer les comportements, amenant les assurés à conduire moins, et plus prudemment. Un tel système pourrait de plus être utilisé par les secours, augmentant ainsi considérablement la rapidité d'action et donc les chances de réussites d'une intervention. Enfin, un atout de plus est que ce type d'assurance est susceptible d'attirer les bons conducteurs. Ces derniers, une fois satisfait, encourageraient alors d'autres conducteurs, initialement plus retissant à l'utilisation de ce type de données, à venir à leur tour profiter d'un tel système. En conclusion et selon le même principe que l'exemple concernant la nécessité de segmenter, une entreprise sans formule PAYD supporterait de mauvais risques tandis qu'une compagnie ayant mis en place un tel système attirerait un grand nombre de clients porteurs d'un risque faible.

Un lancement sur ce marché n'est cependant pas à prendre à la légère. Les grands acteurs de l'assurance automobile sont mitigés quant aux éventuels résultats d'un tel système. La réaction des clients n'est pas à négliger, il n'existe pas de certitude quant au fait que les français soient prêts à accepter la collecte d'informations personnelles de ce type. La réglementation en vigueur sera bien évidemment à suivre de très près, afin de pouvoir définir clairement les droits d'accès aux données. La sécurisation des données, que ce soit dans le stockage ou durant le transfert, est alors essentielle car en aucun cas un tiers ne doit avoir accès aux données concernant un assuré. L'installation et l'encombrement de la « boite noire » recueillant toutes les informations est également à prendre en compte. MMA a par exemple privilégié la simplicité via la « MMA Box », une petite boite qu'il suffit de placer dans le véhicule. Il faut de plus prendre en compte la gestion des données, qui peuvent rapidement devenir très volumineuses. On pourra citer l'exemple de « Norwich Union » qui décida de s'allier à « Teradata » afin de concevoir un « entrepôt de données actives » capables aujourd'hui de traiter plus d'un milliard de lignes d'informations par jour.

Les principaux acteurs de ce marché en France sont aujourd'hui MMA, la Maaf, AXA, Aviva, Groupama via Amaguiz et Solly Azar avec leur offre « easy drive » que nous allons détailler ici à titre d'exemple. Cette offre concerne les jeunes de 18 à 25 ans. Un boitier GPS spécifique développé par « Cobra » relève chaque mois le kilométrage ainsi que le nombre d'heures de conduite de nuit (entre 1h00 et 6h00). L'assuré possède alors un forfait de 1 000 kilomètres mensuel, chaque kilomètre supplémentaire étant facturé à 30 centimes d'euros. L'assuré reçoit un SMS d'alerte pour l'informer de la distance parcourue dès lors qu'il atteint 800 kilomètres dans le mois. Un forfait incitatif à la prudence de 20 euros est appliqué pour chaque conduite de nuit, quelque soit le nombre de kilomètres parcourus ou le nombre d'heures de conduite. L'assuré bénéficiera alors d'un rabais allant jusqu'à 40 % de la prime initiale en fonction du respect de ces conditions. Ce produit représentait plus de 7 % des contrats automobile du courtier grossiste un an après son lancement en 2008. Un tel résultat laisse donc sous entendre le fort potentiel du système PAYD.

On pourra conclure que l'assurance automobile, qui parait pourtant être un risque simple à gérer et bien connu des assurances, est un marché très concurrentiel et en perpétuel mouvement. La facturation des garanties d'assurance automobile est de plus en plus personnalisée et tend à se rapprocher d'un modèle de facturation dépendant de la consommation. On pourrait bien voir apparaitre à moyen terme et sous la bannière de la sécurité routière, de l'écologie et de prix avantageux des forfaits d'assurances automobile sous la forme de ce que l'on rencontre actuellement dans la téléphonie mobile. Bien entendu la France est encore loin d'en arriver à un tel point, mais les importantes possibilités d'évolution sont bien réelles.

Partie II : Préliminaires à la modélisation du risque automobile.

Chapitre 1 : Pré requis, limite et pertinence de la segmentation.

La mise en place ou la révision d'une segmentation, aussi méthodique et précise soit elle, est soumise d'une part à d'éventuels manques d'informations, et d'autres part à des enjeux marketing qu'il n'est pas toujours possible de modéliser mathématiquement parlant. De ce fait, la modélisation d'un risque tel que le risque automobile ne saurait se faire sans pré requis et intuitions préalables.

Il est tout d'abord nécessaire de prendre connaissance et de comprendre une tarification et donc une segmentation préexistante. Cette première étape primordiale permettra par la suite de pouvoir interpréter rapidement et clairement des résultats qui viendront alors s'inscrire dans une certaine logique. On pourra alors parler de logique actuarielle, mais également de logique d'entreprise ou d'enjeux marketing. En effet, si une tarification s'appuie essentiellement sur le principe de neutralité actuarielle (idéalement la prime pure doit couvrir le risque exact garanti par l'assureur), ce fonctionnement est altéré par des positions stratégiques et concurrentielles qui modulent le monde de l'assurance. Chaque entreprise possède un cœur de cible particulier, une clientèle spécifique qu'elle cherche à attirer, et cette clientèle ne correspond pas toujours à ce que l'on pourrait appeler un « bon risque ». Les cibles visées par les assureurs sont bien souvent des populations importantes en besoin d'assurances et qui représentent par conséquent d'éventuelles parts de marché à conquérir. Cette notion importante ne doit pas être perdue de vue lors de la modélisation ou tout du moins lors de l'interprétation de la segmentation ou des évolutions à mettre en place.

La modélisation, de même que l'analyse de donnée qui lui est préalable, connait un certain nombre de limites. En effet, la qualité et la quantité des données utilisées doivent être contrôlées. A titre d'exemple, un portefeuille contenant une très faible quantité de véhicules d'une marque X auxquels sont associés une charge de sinistre très élevée ne signifiera pas que la marque X est à bannir du portefeuille. La cause de cette charge trop importante peut être liée à un unique véhicule, ou encore à la présence d'un sinistre grave ayant une probabilité d'occurrence très faible. On comprend donc que la modélisation d'un risque doit pouvoir s'asseoir sur une quantité suffisante de données afin d'être pertinente. La qualité des données est également primordiale. Une modélisation ne peut être correcte si elle s'appuie sur des données erronées. Toute modélisation doit donc faire l'objet d'une extraction de données rigoureuse ainsi que d'une épuration de la base de données utilisée.

Enfin, une autre limite de la modélisation reste l'absence de données, ou tout du moins l'existence de données non accessibles. On évoquera tout d'abord les données non connues, soit par manque d'information, soit par la non collecte de ces informations qui occasionnerait un coût jugé trop élevé, soit enfin par l'impossibilité matérielle d'acquérir de telles données. L'absence de données peut également être la conséquence d'une franchise ou d'un écrêtage, voir même de l'utilisation de recours en assurance. Pour ce qui est de la franchise, il s'agit d'un montant, pas toujours fixe, qui reste à la charge de l'assuré en cas de sinistre. Dans ce cadre et plus particulièrement dans l'assurance automobile, on comprend bien la volonté de l'assuré de ne pas déclarer d'éventuels sinistres dont le coût serait inférieur à la franchise. En effet, la déclaration d'un tel sinistre non seulement ne leur rapporterait pas un sou de la part de leur assureur, mais en plus

pourrait faire augmenter leur coefficient bonus malus. Il existe ainsi des sinistres non déclarés, ce qui peut alors se révéler problématique si l'assureur souhaite modéliser le risque sous-jacent à une franchise inférieure. On parle alors de troncature à gauche, ce qui signifie que la variable Y reprenant le montant d'un sinistre (et donc l'existence d'un sinistre) n'est pas observable lorsqu'elle est inférieure à un seuil C (la franchise). L'écrêtage correspond quant à lui à un processus différent. L'assureur qui a la possibilité de se réassurer, a la possibilité de limiter la charge à laquelle il devra faire face en cas de sinistre important. De même, un écrêtage peut être effectué lors de la modélisation du risque, permettant ainsi de ne pas prendre en compte des sinistres exceptionnels, et donc peu nombreux, qui viendraient troubler la modélisation et seront traités séparément. On parle alors de censure à droite, ce qui signifie que la variable Y représentant le montant d'un sinistre ici n'est pas observable au-delà d'un seuil C fixe ou variable, mais que C est alors observable à la place. En clair, on n'observe plus Y, mais le minimum entre Y et C.

On retrouve alors très souvent les notations suivantes. Soit $(X_1,\dots,X_n)$ un échantillon (exemple : le montant d'un sinistre) et C > 0 fixé (exemple : le montant correspondant à l'écrêtage), on dit qu'il y a censure à droite si au lieu d'observer $(X_1,\dots,X_n)$ on observe $(T_1,D_1),\dots, (T_n,D_n)$ définis comme suit ;

$$T_i = X_i \wedge C = \min(X_i, C) \ et \ D_i = \begin{cases} 0 \ si \ X_i \leq C \\ 1 \ si \ X_i > C \end{cases}$$

On définit de même la censure à gauche, qu'il faut bien différencier de la troncature à gauche (exemple de la franchise). Dans le premier cas on observe tout de même une donnée (fixée à C) tandis que pour une troncature, il n'y a pas d'observation en dessous d'un seuil C > 0. Si X est la variable qui représente le montant d'un sinistre, la variable tronquée observée est en réalité la variable X sachant X > C que l'on pourra éventuellement noter {X | X > C} par abus de notation. On rajoutera également que la censure peut être aléatoire (et on remplace alors C par un échantillon $(C_1,\dots,C_n)$).

Notons de plus que l'absence de données peut être liée à l'utilisation de recours en assurance contre un assureur, ce recours pouvant être l'objet d'une convention entre assureurs. Il est alors nécessaire de savoir ce que représentent les données sur lesquelles on désire travailler. Il peut s'agir du montant réel occasionné par un sinistre, ou bien du montant d'un sinistre conditionnellement à une franchise ou à un écrêtage, ou bien encore de la charge réelle pour l'entreprise, prenant compte des provisions mises en place ou d'éventuels recours. Ce dernier montant peut ainsi être très différent du montant réel occasionné par un sinistre et il faut alors définir très clairement quel sera l'objet de la modélisation, afin d'éviter toute confusion.

Un dernier point à souligner est l'historique de données. Une modélisation ayant pour but de créer une tarification devra s'appuyer sur un historique de données conséquent (tout en gardant à l'esprit que le risque ne doit pas avoir changer durant la période d'observation), tandis qu'une modélisation de « contrôle » de la tarification ou de suivi pourra éventuellement se concentrer sur une année d'exercice. Ce suivi de la tarification est nécessaire, d'autant plus lorsque l'historique utilisé lors de la création de la tarification en vigueur n'était pas assez important. L'apparition de nouvelles informations permet alors de combler les absences de la première étude, ou tout du moins d'apporter une information supplémentaire, ce qui permet ensuite une évolution des grilles tarifaires, voir même de la segmentation.

Chapitre 2 : Premières analyses et intuition de modélisation.

2.1 Analyse descriptive préalable.

Une fois les données clairement déterminées et l'outil de tarification actuel connu, il convient d'effectuer des travaux d'analyses préliminaires avant même de commencer la modélisation. Malgré une épuration préalable de la base de données, il est possible que certaines variables ne soit pas pertinentes ou ne correspondent pas à la segmentation issue de la tarification en vigueur. Une première analyse descriptive peut permettre de le repérer. On s'intéressera alors plus particulièrement aux variables classiques, soit afin de confirmer une intuition, soit dans le but d'en obtenir une. Les enjeux de l'analyse descriptive sont alors multiples. Elle permet tout d'abord de déterminer les caractéristiques d'un individu moyen afin de connaitre la population assurée et de vérifier son adéquation avec le cœur de cible de l'entreprise. Elle permet ensuite de vérifier la pertinence des variables tout en étudiant de façon plus ou moins succincte la corrélation entre les variables, notion primordiale lors de la modélisation et sur laquelle nous reviendront plus loin. Elle permet enfin d'obtenir un grand nombre d'intuitions concernant les populations à risques et leur importance mais également concernant les variables déterminantes dans la mesure du risque. Dans la suite de ce paragraphe, nous allons tout d'abord revenir sur les variables classiques que l'on retrouve dans l'assurance automobile, puis nous évoquerons la notion de coût moyen et de fréquence, et nous nous intéresserons enfin plus particulièrement aux méthodes de classification et d'analyses de données telles que les analyses en composantes principales (ACP) ou encore les analyses en composantes multi variées (ACM).

Commençons alors par décrire les variables classiques que l'on retrouve en assurance automobile. Nous avons d'ores et déjà évoqué le CRM ou coefficient de réduction majoration (Bonus/Malus). Cette variable est généralement déterminante dans une tarification car elle rend compte de la probabilité que l'assuré a d'avoir un sinistre. Nous parlerons de fréquence de sinistres qui sera alors plus faible pour les assurés ayant un faible CRM, et inversement. Le risque, qui dépend alors de l'assuré lui-même, sera en lien avec son âge et son sexe. La première variable est essentiellement utilisée dans le but de différencier ce que l'on appelle les jeunes conducteurs des automobilistes expérimentés. En effet et comme évoqué dans la première partie, les jeunes conducteurs sont porteurs d'un risque important, et cette variable sera alors d'une grande utilité dans la modélisation et la tarification. Le sexe de l'assuré est quant à lui moins usité, mais implique tout de même souvent une réduction aux jeunes conductrices, porteuses d'un risque plus faible que leurs homologues masculins.

Le risque assuré dépend également du véhicule. Les variables classiques sont alors la puissance du véhicule, qui augmente le risque en croissant, la valeur à neuf du véhicule, qui informe sur le coût éventuel d'un sinistre ou encore la date de première mise en circulation. Il existe cependant de nombreuses autres variables concernant le véhicule telles que la marque, le nombre de places assises, le type de boite de vitesse et bien d'autres encore. Deux variables que l'on retrouve également très souvent sont la zone d'utilisation et l'usage du véhicule. On comprend qu'un véhicule dont le lieu de garage est Lyon ou Paris est porteur d'un risque plus important qu'un véhicule situé en rase campagne. De même, le véhicule d'un VRP (Voyageur Représentant Placier) ou tout autre véhicule utilisé dans le cadre d'un emploi (véhicule de livraison ou autre), sera représentatif d'un risque plus important que le véhicule d'un conducteur occasionnel.

Enfin, Le risque dépend bien entendu des garanties octroyées par l'assureur. Un assuré « au tiers » ne payera pas la même prime d'assurance qu'un assuré « tous risques ». En effet, les sinistres couverts par l'assureur seront alors différents, le premier ne sera assuré qu'en responsabilité civile et donc couvert pour les dégâts qu'il pourra causer à un tiers tandis que le second sera couvert en garantie dommage, pour les dégâts qu'il pourra causer à un tiers ou à lui-même. Bien entendu les contrats proposés par les assureurs sont plus complexes et plus divers, mais l'on conçoit ainsi l'importance de cette variable dans la modélisation et dans la tarification.

Comme nous l'avons vu précédemment, chaque variable est porteuse d'une information, on parle alors de variables explicatives, qui vont en l'occurrence expliquer le risque, que ce soit par la probabilité de survenance d'un sinistre ou son montant. Certaines variables concernent en effet plus particulièrement la fréquence de sinistres tandis que d'autres informent sur les montants qui feront suite à un sinistre. La fréquence et le coût moyen sont alors des indicateurs qui permettent de mesurer, de quantifier le risque. Un troisième indicateur très souvent usité est le rapport S/P ou ratio sinistre sur prime. Ce dernier permet d'analyser si un segment de tarification ou une sous population a été tarifé correctement, il renseigne sur les résultats de l'assureur. Les trois indicateurs dépendent alors de variables telles que le nombre de sinistres, leurs montants mais aussi le nombre de contrats et leur prime associée. On note cependant que le nombre de contrats est un mauvais indicateur pour le calcul de la prime totale ou moyenne et donc du rapport S/P. En effet, le nombre de contrat ne rend pas compte de l'exposition réelle au risque. Un contrat dont la durée de garantie associée est de deux jours doit être comptabilisé comme tel, et l'intégralité de la prime ne doit donc pas être prise en compte. Nous parlerons alors du nombre de « contrats acquis ». Le calcul des indicateurs est repris à partir du tableau ci-dessous. Nous exposons ainsi le calcul de l'exposition réelle au risque, de la prime moyenne, du coût moyen ou encore de la fréquence. On notera que le calcul du nombre de contrats acquis comprend une division par 360 et non par 365 car l'on considère une année commerciale et non civile. On remarque également que la prime totale correspond à la somme des primes individuelles ramenées à leur exposition réelle au risque et multipliée par 12, si l'on considère des primes mensuelles.

Nombre de sinistres	Donnée
Charge totale	Donnée
Coût moyen	$Charge\ totale / Nombre\ sinistre$
Exposition ou Contrats acquis	$Nombre\ de\ jours\ objets\ d'une\ garantie / 360$
Prime totale acquise	$\sum (Prime\ individuelle) * Exposition * 12$
Prime moyenne	$Prime / Contrats\ acquis$
S/P	$Charge\ totale / Prime$
Fréquence	$Nombre\ sinistres / Exposition$

Illustration 6

Dans le cadre de la modélisation, deux études séparées seront opérées, l'une pour la fréquence de sinistre et l'autre pour le coût moyen d'un sinistre. L'indépendance entre ces deux indicateurs est bien souvent admise mais il est plus prudent de vérifier cette hypothèse afin de s'assurer que cette dernière reste cohérente. Un recoupement devra ensuite être opéré entre ces deux variables et la prime moyenne afin de constater la présence ou non de segments sur ou sous tarifés. On note que le rapport S/P se retrouve alors facilement avec l'égalité suivante ;

$$S/P = \frac{Coût\ moyen * Fréquence}{Prime\ moyenne}$$

2.2 Analyses factorielles et classification.

2.2.1 Principe.

Nous allons à présent nous intéresser à différentes théories de classification ou d'analyses de données, autres qu'une simple analyse descriptive et permettant d'acquérir une intuition plus poussée quant à la tarification et à la mise en place de la modélisation. Une analyse porte sur un certain nombre d'individus que l'on fixera à « n », chaque individu étant décrit par « p » variables. Une analyse descriptive classique permet d'étudier séparément chaque variable, voir de les étudier deux par deux. Ce type d'étude est une source d'information importante, mais elle ne permet pas de prendre en compte et de comprendre les liaisons ou les relations entre les variables, alors que l'on recherche bien souvent une vision d'ensemble. Les méthodes factorielles d'analyses de données ont cette vocation. Ces méthodes sont variées et toutes s'appliquent dans un cadre d'étude particulier. Nous citerons ainsi l'ACP (Analyse en Composantes Principales), pertinente lorsque les variables sont quantitatives et si possible dans la même unité de mesure et l'AFC (Analyse Factorielle des Correspondances) utilisée pour l'étude de tableaux de contingences que nous décriront plus loin. Ces deux analyses de base font l'objet de nombreuses généralisations telles que l'ACM (Analyse des Correspondances Multiples) que nous décriront plus loin et qui concerne l'étude de tableaux de variables qualitatives représentées dans ce que l'on appelle un tableau disjonctif complet.

Commençons alors par décrire le principe de l'analyse factorielle. L'approche classique consiste en l'étude du nuage de points représentant les individus dans l'espace géométrique des variables à p dimensions. On comprend bien que la visualisation d'un tel nuage de points s'avère complexe dès lors que l'on dépasse les trois dimensions. Nous allons donc chercher à projeter ce nuage de points dans un plan tout en préservant un maximum d'information. On s'efforce ainsi d'obtenir une représentation la plus fidèle possible du nuage de points dans un espace facilement visualisable et interprétable.

2.2.2 *Définitions et propriétés.*

Avant de rentrer plus en détails dans la description des aspects théoriques et pratiques de l'ACP, nous allons énoncer certaines définitions et propriétés mathématiques qui seront utilisées plus loin. On suppose connu les notions élémentaires concernant les matrices et les probabilités, ainsi que les notions de base orthonormale et de projection sur un espace. L'opérateur $.^T$ désignera la transposée d'une matrice, l'opérateur $\bar{.}$ le conjugué d'un nombre complexe.

Définition 1

Soit A une matrice à n lignes et p colonnes, V un vecteur non nul de $\mathbb{R}^p$ et λ un nombre réel. On dira que V est un vecteur propre et λ une valeur propre de A si et seulement si la relation suivante est vérifiée ;

$$AV = \lambda V \Leftrightarrow (A - \lambda \mathbb{I}_p)V = 0$$

Définition 2

Soit A une matrice à n lignes et n colonnes, A est dite régulière ou inversible si et seulement si il existe une matrice noté A^{-1} telle que $AA^{-1} = A^{-1}A = \mathbb{I}_n$.

Définition 3

Soit A une matrice à n lignes et n colonnes, A est dite diagonalisable si et seulement si il existe une matrice P inversible et une matrice D diagonale telles que $P^{-1}AP = D$. Les vecteurs colonnes de P sont alors les vecteurs propres de A et les composantes de D sont les valeurs propres de A.

Définition 4

Une matrice A est dite orthogonale si et seulement si $A^T A = I$, où I désigne la matrice identité.

Définition 5

Soit une fonction de $\mathbb{R}^p \times \mathbb{R}^p$ dans $\mathbb{R}$ qui à un couple de point (x,y) associe un nombre réel $< x|y >_\phi$. On appelle cette fonction un produit scalaire si elle vérifie les propriétés suivantes ;

- $< x|y >_\phi =< y|x >_\phi$ pour tout couple (x,y) de $\mathbb{R}^p \times \mathbb{R}^p$.
- $< x|y+z >_\phi = < x|y >_\phi + < x|z >_\phi$ pour tout triplet (x,y,z) de $\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$.
- $< x|ay >_\phi = a< x|y >_\phi$ pour tout couple (x,y) de $\mathbb{R}^p \times \mathbb{R}^p$ et pour tout a de $\mathbb{R}$.
- $< x|x >_\phi \geq 0$ pour tout x de $\mathbb{R}^p$.
- $< x|x >_\phi = 0 \Rightarrow x=0$.

Définition 6

Soit $<-|->_\phi$ un produit scalaire, on définit alors la norme associée $\|x\|_\phi =< x|x >_\phi$. On définit également la distance associée $d_\phi^2(x,y) = \|x-y\|_\phi$. Dans toute la suite, on notera <-|-> le produit scalaire standard de l'espace euclidien, associé à la norme et à la distance euclidienne.

Définition 7

Soit M une matrice symétrique d'ordre n. M est dite définie positive si elle vérifie au moins une des trois relations suivantes (En pratique, seule la dernière sera utilisée ici);

- $x^T Mx > 0$ pour toute matrice colonne non nulle à n éléments réels.
- Toutes les valeurs propres de M sont strictement positives.
- $< x|y >_M := x^T My$ est un produit scalaire sur $\mathbb{R}^n$.

Propriété 1 : Décomposition de Cholesky.

Soit X une matrice symétrique définie positive. Alors X peut s'écrire sous la forme $X = TT^T$ ou T est une matrice triangulaire inférieure à diagonale positive, et on en déduit que X est inversible.

Démonstration de la propriété 1

Cette propriété se démontre par récurrence inverse (en partant de la fin) en résolvant le système d'équation fourni par l'égalité, on ne le démontrera pas ici (voir [20] p88-89) mais on fournit tout de même l'algorithme de construction de la matrice T ;

$$Pour\ i \neq k, t_{ii} = \sqrt{x_{ii} - t_{i1}^2 - \cdots - t_{i(i-1)}^2}\ et\ t_{ik} = \frac{s_{ki} - t_{k1}t_{i1} - \cdots - t_{k(i-1)}t_{i(i-1)}}{t_{ii}}$$

Définition 8

Soit M une matrice symétrique définie positive (et donc inversible) d'ordre n. On parlera dans toute la suite de métrique M associée au produit scalaire $< x|y >_M := x^T My =< x|My >$. La métrique identité qui est la plus usitée correspond ainsi au produit scalaire standard et donc à la distance euclidienne. On peut cependant rencontrer des métriques réduites ou la métrique du Khi² dont nous reparlerons plus loin.

Propriété 2

Soit X une matrice que l'on souhaite étudier autour d'une métrique M. Alors il existe une matrice T telle que $< x|y >_M =< Tx|Ty >$ et il est alors équivalent de travailler sur la matrice TX en considérant la métrique identité.

Démonstration de la propriété 2

Soit X une matrice de n lignes et p colonnes. On définit alors l'espace $\mathbb{R}^p$ muni du produit scalaire $< -|->_M$ associé à la métrique M. On sait d'après la propriété 1 que toute matrice symétrique M définie positive peut s'écrire sous la forme $M = T^T T$ avec ici T une matrice triangulaire supérieure. Ainsi on obtient que ;

$$< x|y >_M = y^T Mx = y^T T^T Tx = (Ty)^T Tx =< Tx|Ty >$$

x et y représentant des vecteurs de coordonnées dans $\mathbb{R}^p$. On peut alors utiliser la métrique identité sur des données transformées (à savoir TX).

Définition 9

Deux vecteurs u et v sont dits M-orthogonaux si et seulement si $< x|y >_M = 0$.

De même on généralise la notion de vecteur x unitaire en M-unitaire si $\|x\|_{\mathrm{M}} = 1$.

Définition 10

Soit A une matrice et M une métrique, alors on dit que A est M-symétrique si et seulement si $MA = A^T M = (MA)^T$. On en déduit que la matrice MA est symétrique au sens classique du terme.

Propriété 3: Théorème spectral en dimension finie.

Soit A une matrice réelle symétrique (respectivement M-symétrique), alors il existe une matrice P orthogonale (respectivement M-orthogonale) et une matrice D diagonale dont tous les coefficients sont réels et telles que $P^{-1}AP = D$. Autrement dit, toute matrice symétrique (respectivement M-symétrique) réelle admet une base de vecteurs propres orthonormale (respectivement M-orthonormale) dont les valeurs propres associées sont réelles. (Voir démonstration en Annexe 3)

Propriété 4

Soit A une matrice carré de dimension k et X un vecteur de k colonnes. Alors on a l'égalité suivante ;

$$\frac{d(X^T AX)}{dX} = (A + A^T)X$$

Démonstration de la propriété 4

On pose $A = (a_{ij})_{\substack{1 \le i \le k \\ 1 \le j \le k}}$ et $X = (x_i)_{1 \le i \le k}$. On a alors ;

$$\frac{\partial(X^T AX)}{\partial x_1} = \frac{\partial}{\partial x_1}\left(a_{11}x_1^2 + \sum_{i=2}^{k} a_{i1}x_i x_1 + \sum_{j=2}^{k} a_{1j}x_1 x_j\right) = \left(2a_{11}x_1 + \sum_{i=2}^{k} a_{i1}x_i + \sum_{j=2}^{k} a_{1j}x_j\right)$$

$$\frac{\partial(X^T AX)}{\partial x_1} = \sum_{i=1}^{k} a_{i1}x_i + \sum_{j=1}^{k} a_{1j}x_j = \sum_{i=1}^{k} (a_{i1} + a_{1i})x_i$$

On généralise ce résultat aux dérivées partielles par rapport à x_i pour tout i et l'on en déduit le résultat.

Propriété 5

Un corollaire simple de la propriété précédente existe pour A une matrice symétrique, et on a alors ;

$$\frac{d(X^T AX)}{dX} = 2AX$$

2.2.3 L'analyse en composantes principales (ACP).

2.2.3.1 Cas Général, aspect théorique.

Afin de formaliser la problématique de l'analyse factorielle décrite plus haut d'un point de vue mathématique, commençons par définir les éléments de l'étude. Soit une matrice $X = [x_{ij}]_{\substack{1\leq i\leq n \\ 1\leq j\leq p}}$ qui est la représentation des individus en lignes et des variables en colonnes. Le nuage de points que l'on notera « N » est la représentation de chaque individu associé à un point M_i de coordonnées $\begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$ dans l'espace $\mathbb{R}^p$. On définit de plus les pondérations $(p_i)_{1\leq i\leq n}$ de chaque individu, bien souvent constantes et égales à 1/n. L'espace des individus est muni d'une distance (pas forcément euclidienne) associée à une métrique M d'ordre p. L'espace est donc muni du produit scalaire $< x|y >_M = y^T M x$ pour x et y dans $\mathbb{R}^p$.

On recherche alors un sous espace affine $\mathbb{E}_k$ de dimension k<p (souvent k=2) tel que l'inertie noté $I_{\mathbb{E}_k}$ du nuage N par rapport à $\mathbb{E}_k$ soit minimale (représentation la plus fidèle possible). L'expression de cette inertie est alors la suivante ;

$$I_{\mathbb{E}_k} = \sum_{i=1}^{n} p_i \times d_M^2(M_i, \mathbb{E}_k)$$

$d_M(M_i, \mathbb{E}_k)$ représente la distance entre le point M_i et sa projection sur l'espace $\mathbb{E}_k$ (que l'on notera m_i). Cette distance peut alors s'écrire $d_M(M_i, m_i) = \|M_i - m_i\|_M$ avec $\|-\|_M$ la norme associée à la métrique M. Nous gardons ici la notation M_i par abus de langage, en effet, la notation ne représente alors plus un point mais le vecteur associé $\overrightarrow{OM_i}$ avec O l'origine. On note que l'espace $\mathbb{E}_k$ passe nécessairement par le point moyen du nuage également appelé centre de gravité. Ceci amène logiquement à considérer des variables centrées.

Nous allons alors décomposer l'espace $\mathbb{E}_k$ en k sous espaces orthogonaux. Les k sous espaces seront alors des droites que l'on appellera axes principaux. On pourra ensuite représenter le nuage de points sur différents plans composés de deux axes. On définit alors le vecteur $u = \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix}$ comme un vecteur directeur d'un axe Δu égal au premier sous espace de $\mathbb{E}_k$, on impose de plus que le vecteur u soit unitaire au sens de la métrique M, soit $u^T M u$=1. L'objectif est alors la recherche d'un tel vecteur minimisant l'inertie projetée $I_{\Delta u}$, soit qui minimise la distance entre les points et cette même droite (la moyenne des carrés des distances des points à cette même droite dans le cadre euclidien). Puis nous rechercherons un vecteur v M-unitaire M-orthogonal au premier minimisant à nouveau l'inertie projetée et ainsi de suite. Dans toute la suite, m_i désignera les coordonnées du point M_i après projection sur l'axe Δu .

Revenons alors sur la notion d'inertie. Notons que l'inertie totale du nuage de points vaut ;

$$I_T = \sum_{i=1}^{n} p_i \times \|M_i\|_M^2$$

Or nous savons que $M_i = m_i + (M_i - m_i)$, et que l'on se place alors dans un triangle rectangle, ce qui nous permet d'utiliser le théorème de Pythagore $\|M_i\|_M^2 = \|m_i\|_M^2 + \|M_i - m_i\|_M^2$ (*Voir schéma ci-dessous*).

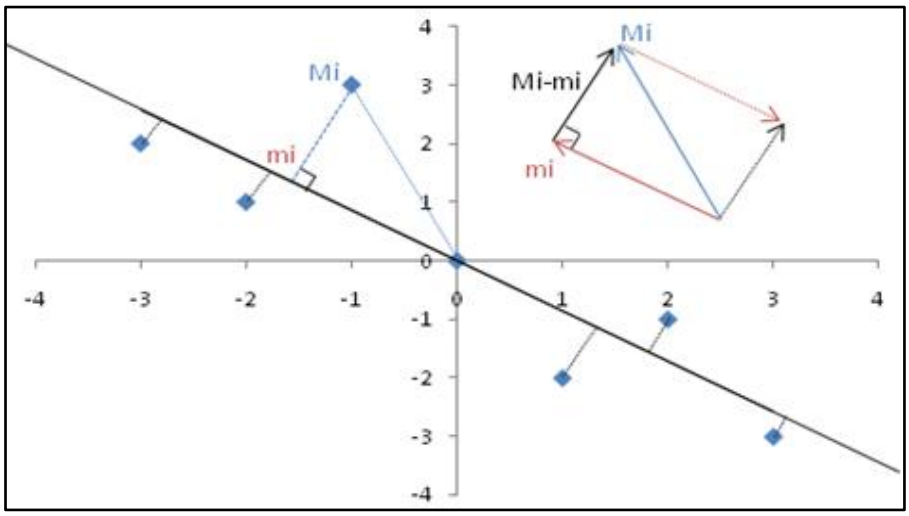


Illustration 7 : Régression affine d'un nuage de points et visualisation d'un triangle rectangle

L'utilisation du théorème de Pythagore nous permet ainsi de décomposer l'inertie totale du nuage de points en inertie projetée sur un axe appelée également inertie statique et en inertie autour de l'axe dénommée inertie mécanique. On formalise cette égalité comme suit ;

$$I_T = I_{Stat}(u) + I_{Meca}(u) = \sum_{i=1}^{n} p_i \times \|m_i\|_M^2 + \sum_{i=1}^{n} p_i \times \|M_i - m_i\|_M^2$$

Nous constatons alors que le terme d'inertie mécanique correspond à ce que nous avons noté plus haut $I_{\mathbb{E}_k}$ pour k=1 (détermination d'un axe u). I_T étant fixé par le nuage de points, minimiser l'inertie mécanique revient alors à maximiser l'inertie statique ou projetée.

Nous commencons alors par donner l'expression des m_i (on rappelle que les coordonnées de la projection s'obtiennent à partir du produit scalaire), avant d'exprimer $I_{Stat}(u)$.

$$m_i = < M_i | u >_M \frac{u}{\|u\|_M} = \left(M_i^T M u\right) \frac{u}{u^T M u} = \left(M_i^T M u\right) u$$

On obtient alors

$$I_{Stat}(u) = \sum_{i=1}^{n} p_i \times \|m_i\|_M^2 = \sum_{i=1}^{n} p_i \times \left(M_i^T M u\right)^2 \times u^T M u = \sum_{i=1}^{n} p_i \times \left(M_i^T M u\right)^2$$

$$I_{Stat}(u) = \sum_{i=1}^{n} p_i \times u^T M M_i M_i^T M u = u^T M \left(\sum_{i=1}^{n} M_i p_i M_i^T \right) M u$$

Nous pouvons alors écrire le terme de la somme en fonction de la matrice centrée de départ X et la matrice diagonale des pondérations. En effet ;

$$X^T DX = X^T \begin{bmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{bmatrix} \begin{bmatrix} M_1^T \\ \vdots \\ M_n^T \end{bmatrix} = [M_1 \quad \dots \quad M_n] \begin{bmatrix} p_1 M_1^T \\ \vdots \\ p_n M_n^T \end{bmatrix} = \left[\left(\sum_{i=1}^{n} M_i p_i M_i^T \right) \right]$$

Il reste a rajouter finalement la notation $V = X^T DX$, permettant ainsi d'écrire ;

$$\boxed{I_{Stat}(u) = u^T MVMu}$$

MVM est appelée la matrice d'inertie du nuage de points, elle se confond avec la matrice de variance covariance lorsque l'on considère la métrique identité. Rappelons que l'objectif est alors la recherche d'un vecteur u qui maximise $I_{Stat}(u)$ (u n'est alors pas forcément unitaire), ce qui équivaut à la recherche d'un vecteur u qui annule la dérivée, on démontre tout d'abord rapidement que VM est M-symétrique puis l'on calcule la dérivée ;

$$(MVM)^T = M^T V^T M^T = M(X^T DX)^T M = MVM$$

On peut ainsi calculer aisément la dérivée de $u^T MVMu$ et $u^T Mu$ grâce à la *propriété 5* ;

$$\frac{dI_{Stat}(u)}{du} = \frac{d(\frac{u^T MVMu}{u^T Mu})}{du} = \frac{(u^T Mu) \times 2MVMu - (u^T MVMu) \times 2Mu}{(u^T Mu)^2} = 0$$

$$MVMu = \frac{(u^T MVMu)}{(u^T Mu)} Mu$$

Or $(u^T Mu)$ et $u^T MVMu$ sont des scalaires et M est inversible donc $VMu = \lambda u$, d'où u est un vecteur propre de VM associé à la valeur propre λ. Si ceci est vérifié, on a alors $u^T MVMu = \lambda u^T Mu = \lambda$ pour u unitaire. On obtient donc l'inertie maximale en sélectionnant le vecteur propre u associé à la plus grande valeur propre.

Nous savons de plus que VM est M-symétrique (car MVM est symétrique), on en déduit donc d'après le *théorème spectral* que VM est diagonalisable dans une base M-orthonormale formée de vecteurs propres. On détermine ainsi aisément le premier axe en choisissant le vecteur propre associé à la plus grande valeur propre, puis l'on calcule le deuxième axe en sélectionnant un deuxième vecteur propre, M-orthogonale au premier et associé au deuxième plus grand vecteur propre et ainsi de suite.

Le problème initial est alors transformé en la diagonalisation de la matrice $VM = X^T DXM$. Nous pouvons pour ce faire écrire à partir de la *définition 3* et de la *définition 4* VM=$U\Lambda U^T$ avec U une matrice orthogonale dont les vecteurs colonnes représentent les axes principaux et Λ une matrice diagonale dont les composantes sont les valeurs propres de VM. Il est également possible d'ordonner les valeurs propres pour plus de lisibilité. Le premier axe est alors la première colonne de U et ainsi de suite.

Il reste alors à sélectionner le nombre d'axes, à visualiser les individus, à interpréter leur répartition et enfin à interpréter les axes en projetant les variables initiales.

Si u_k représente le vecteur propre de rang k (le k^{ième} plus important), u_k est appelé l'axe principal de rang k. On mesure alors la contribution de chaque axe (désigné par le vecteur propre u_k et la valeur propre λ_k) à l'inertie par la formule ;

$$\boxed{\frac{\lambda_k}{Trace(VM)} * 100}$$

Nous pouvons ensuite représenter la contribution de chaque vecteur propre à l'inertie totale sur un graphique. On sélectionne enfin un minimum d'axes pour un maximum d'informations (nous garderons en règle générale aux alentours de 80 % de l'inertie totale).

Nous obtenons ensuite ce que l'on appelle les coordonnées sur l'axe principal de rang k en projetant X sur u_k , soit en considérant le vecteur $C^k = \begin{bmatrix} {C^k}_1 \\ \vdots \\ {C^k}_n \end{bmatrix} = XMu_k = \begin{bmatrix} < M_1 | u_k >_M \\ \vdots \\ < M_n | u_k >_M \end{bmatrix}$. Ce vecteur représente alors les coordonnées du nuage de points sur l'axe. En pratique, nous représenterons les n individus sur différents couples d'axes u_i et u_j, formant alors un plan, à l'aide de leurs coordonnées inscrites dans les vecteurs C^i et C^j. Ceci permettra de distinguer différents groupes d'individus. On rajoute ensuite la notion de contribution de l'individu i à l'axe u_j. Nous savons d'ores et déjà que l'inertie associée à un axe vaut ;

$$\sum_{i=1}^{n} p_i \times ({C^j}_i)^2 = \sum_{i=1}^{n} p_i \times ({M_i}^T M\, u_j)^2 = {u_j}^T M \sum_{i=1}^{n} M_i p_i {M_i}^T M\, u_j = {u_j}^T M X^T D X M u_j = \lambda_j$$

On en déduit donc que la contribution relative d'un individu i à l'axe u_j vaut ;

$$\boxed{\frac{p_i \times ({C^j}_i)^2}{\lambda_j}}$$

Il est préférable que l'ensemble des individus possède environ la même valeur, on retirera de l'étude un individu trop atypique. On doit cependant faire enfin remarquer que deux points proches dans le plan ne sont pas forcément proches dans l'espace. Un bon indicateur de qualité est alors ce que l'on appelle la contribution relative de l'axe k à la représentation du point i qui s'écrit ;

$$\boxed{\frac{({C^k}_i)^2}{\sum_{k=1}^{p} ({C^k}_i)^2}}$$

Plus cette valeur sera proche de 1 et plus l'axe sera représentatif du point i et donc le point proche de l'axe. Cette valeur correspond au cosinus carré de l'angle entre le point et l'axe. On peut alors conclure que deux points proches dans le plan et proches d'un axe sont proches dans l'espace. Une façon claire et simple de visualiser les individus dans un plan est d'adapter la taille de leur représentation en fonction de cette valeur (ou somme de la valeur de chaque axe pour une représentation dans le plan). Ceci peut facilement s'écrire sous un logiciel tel que R et évite de voir des similarités là ou il n'y en a pas.

Il reste enfin à interpréter les axes principaux. Pour ce faire, nous allons projeter les variables initiales. Nous savons que les vecteurs propres de la base orthonormale sont logiquement des combinaisons linéaires des variables initiales, qui sont alors plus ou moins importantes. On regarde alors la corrélation entre les variables initiales et les vecteurs $\left(C^k\right)$. Les vecteurs C^k définiront les axes de projection, à normalisation près (soit en divisant par $\sqrt{\lambda_k}$), ce qui permettra de visualiser les corrélations entre les variables initiales et les nouvelles variables fictives dans un cercle de rayon 1. On parle classiquement de composantes principales. Nous calculons alors la covariance qui correspond à $\frac{1}{\sqrt{\lambda_k}} \times < y_i \mid C^k >$ avec y_i la i^{ième} variable, soit $y_i = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{bmatrix}$. Cette covariance se note alors de la façon suivante et vaut ;

$$r\left(C^k, y_i\right) = \sqrt{\lambda_k} u_{k_i} \text{ car } r\left(C^k, X\right) = \frac{1}{\sqrt{\lambda_k}} X^T DXMu_k = \sqrt{\lambda_k} u_k$$

On peut ainsi représenter les variables sur ce que l'on appelle le cercle des corrélations. La position des variables vis-à-vis des axes donne alors une information sur l'interprétation de ces derniers. Il reste enfin à préciser que seuls les variables situées près de la circonférence du cercle sont à prendre en compte (forte corrélation positive ou négative).

2.2.3.2 ACP en dimension 2 et métrique euclidienne.

A titre d'exemple et afin de clarifier ce point, plaçons nous dans le cas où p=2 et M est la métrique identité. On se situe ainsi dans $\mathbb{R}^2$ muni de la distance euclidienne. On suppose de plus que la pondération est équiprobable ($\forall i \in [\![1:n]\!],\ p_i = 1/n$). On considère que la matrice X est centrée et s'écrit $X = \begin{bmatrix} x_1 - m(x) & y_1 - m(y) \\ \vdots & \vdots \\ x_n - m(x) & y_n - m(y) \end{bmatrix}$ où m() désigne la fonction moyenne, x et y les deux variables. Le centre de gravité du nuage de points est alors à l'origine. On prend un vecteur u unitaire de la forme $\begin{bmatrix} a \\ b \end{bmatrix}$ avec $a^2 + b^2 = 1$. On obtient ainsi $Xu = \begin{bmatrix} < M_1 | u > \\ \vdots \\ < M_n | u > \end{bmatrix}$. L'expression de l'inertie statique est alors la suivante ;

$$I_{Stat}(u) = u^T \left(\frac{1}{n} X^T X\right) u = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} v(x) & c(x,y) \\ c(x,y) & v(y) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

$$v(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - m(x))^2$$

$$v(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - m(y))^2$$

$$c(x,y) = \frac{1}{n} \sum_{i=1}^{n} \big(x_i - m(x)\big)\big(y_i - m(y)\big)$$

Nous reconnaissons ainsi la matrice de variance covariance que l'on notera C. Cette matrice est symétrique, elle admet donc une base de vecteurs propres orthonormale et est ainsi diagonalisable. On peut donc écrire C sous la forme ;

$$C = U \Lambda U^T = \begin{bmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} & \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} (u_{11} & u_{12}) \\ (u_{21} & u_{22}) \end{bmatrix}.$$

On en déduit alors ;

$$I_{Stat}(u) = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} & \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} (u_{11} & u_{12}) \\ (u_{21} & u_{22}) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \alpha & \beta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

$$I_{Stat}(u) = \lambda_1 \alpha^2 + \lambda_2 \beta^2 \leq \lambda_1 \alpha^2 + \lambda_1 \beta^2 \leq \lambda_1$$

$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ représente ainsi les coordonnées du vecteur u dans la nouvelle base orthonormée des vecteurs propres. L'inertie maximale vaut λ_1 et est atteinte pour $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, et donc pour le premier vecteur propre. On en conclut que l'axe principal d'un nuage de points bivarié est le premier vecteur propre de la matrice de variance covariance des deux variables, l'axe secondaire étant le second.

2.2.3.3 Exemple pratique.

Nous allons à présent exposer un exemple concret d'utilisation de l'ACP dans le cadre de l'assurance automobile. La grande majorité des tarifications inclut une segmentation qui concerne la zone géographique de l'automobile, et plus particulièrement la zone de garage du véhicule. Il existe donc un zonier propre à chaque compagnie qui amène à se poser la question suivante ;

Existe-t-il une géographie du risque automobile ?

Avant toute chose, on renvoie le lecteur intéressé en Annexe 4 pour visualiser des cartes des départements de France en cas de question ou de doute sur leurs noms, numéros ou sur leurs positions sur la carte. Nous considérons ensuite un tableau noté TabACP. Ce tableau est composé de 94 individus représentant les départements de France métropolitaine. Chaque ligne est caractérisée par son exposition réelle au risque ou nombre de contrats acquis, la prime moyenne, la charge moyenne, le CRM moyen, le coefficient tarifaire moyen (un coefficient supplémentaire appliqué au tarif) et la fréquence, pour l'année d'exercice 2008 de la compagnie « Mutant Assurances ». On utilise également une table nommée DepReg qui contient les numéros et noms de départements ainsi que les noms de régions de France métropolitaine. Nous allons alors effectuer une ACP sur les 5 dernières variables, la première sera utilisée pour le poids des lignes. Nous utilisons ici le logiciel R et plus particulièrement le package « ade4 » qu'il faut donc charger, on rentre ainsi le code suivant ;

```
library("ade4")
acp0=dudi.pca(TabACP[,2:6],row.w=TabACP[,1]/sum(TabACP[,1]),scannf=FALSE,nf=5)
```

L'objet « *acp0* » contient un grand nombre d'informations. On peut tout d'abord extraire les valeurs propres que l'on retrouve dans « *acp0$eig* », mais également les composantes principales désignées par « *acp0$li* » ou encore la contribution du premier plan factoriel à la représentation d'un point avec l'objet « *inertia.dudi(acp0,row.inertia=TRUE)$row.cum[,2]* ». Nous allons alors représenter différents graphiques. On commence par exposer les valeurs propres en pourcentage d'inertie projetée ;

```
barplot(acp0$eig/sum(acp0$eig))
```

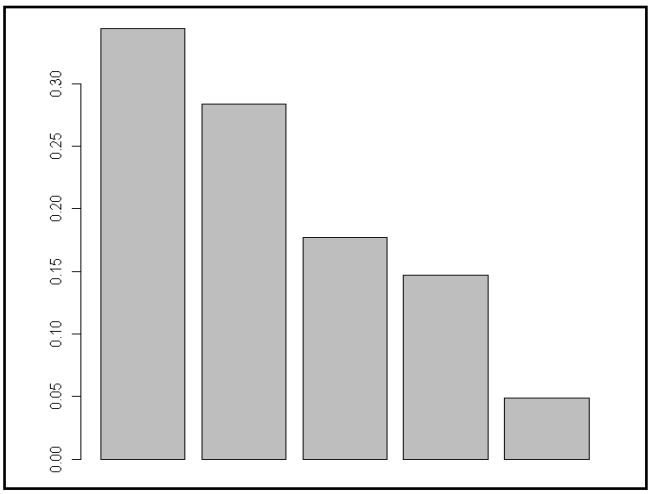


Illustration 8 : Représentation des valeurs propres

Nous constatons ainsi que plus de 60 % de l'information peut être représentée sur les deux premiers axes principaux. Nous allons donc représenter le nuage de points des individus dans ce premier plan de manière classique pour commencer.

s.label(acp0$li,xax=1,yax=2)

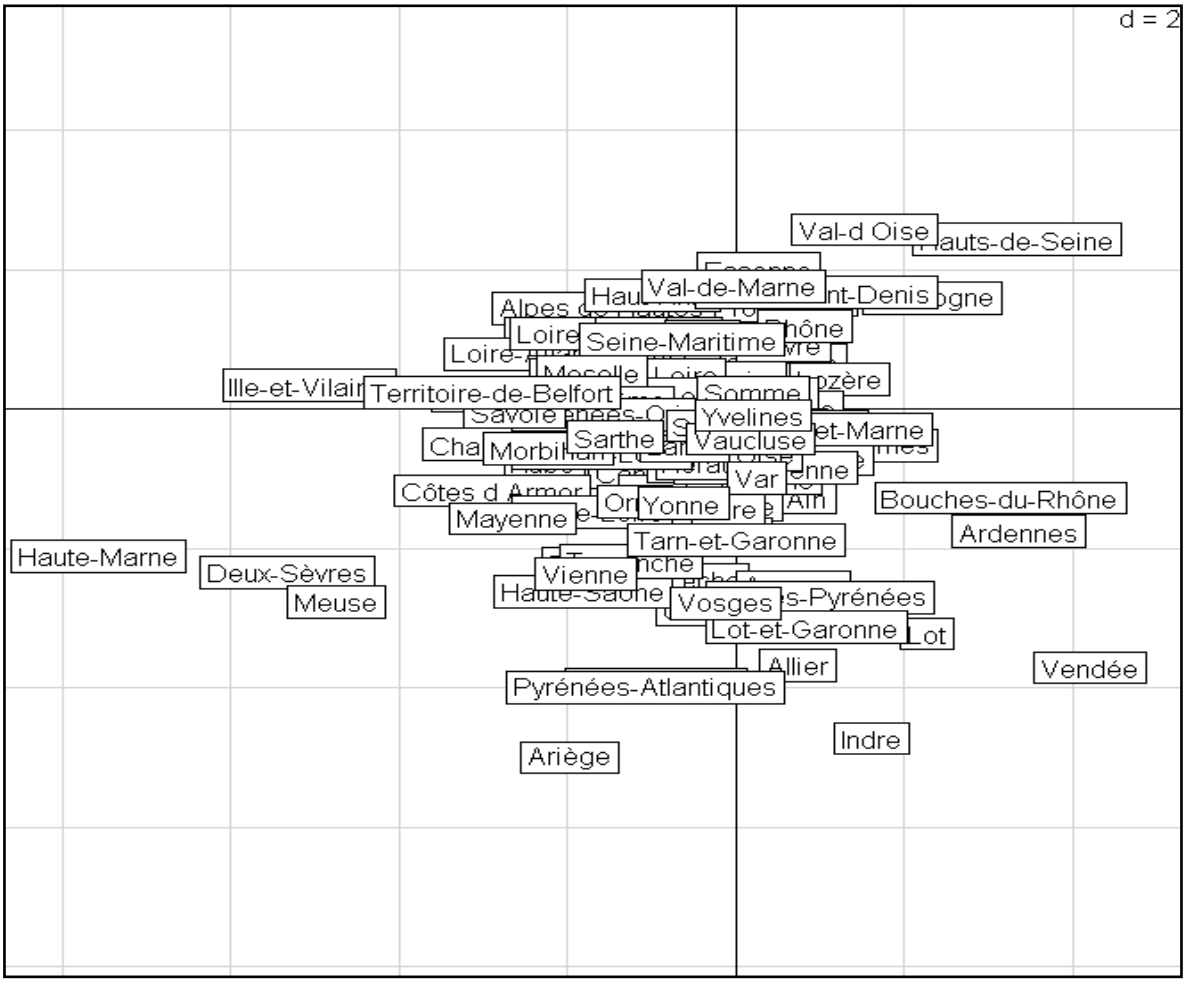


Illustration 9 : Représentation du nuage de points sur les deux premiers axes principaux d'un exemple d'ACP

Nous remarquons d'ores et déjà que certains départements se détachent du centre du nuage. Cependant, nous ne savons toujours pas à quoi correspondent les axes d'une part, et nous rappelons que deux points proches dans le plan ne sont pas forcément proches dans l'espace d'autre part. Nous allons donc représenter le même graphique, à ceci près que la taille des individus dépendra de leur qualité de représentation sur ce plan. Concrètement, nous représenterons les numéros de département, les individus les mieux représentés seront inscrit en rouge et les autres en bleu. De plus et pour éviter que le graphique devienne illisible, nous ne représenterons que par un point les départements les moins significatifs dans le plan. Nous allons de plus représenter le cercle des corrélations afin de comprendre ce que les axes signifient. Pour des raisons de lisibilité et d'interprétation, nous placerons tout ceci sur un seul objet graphique, on note qu'il est également possible de superposer graphe des valeurs propres, nuages des individus et variables initiales avec la fonction « *scatter* ».

```
par(mfrow=c(2,2))
barplot(acp0$eig/sum(acp0$eig))
s.label(acp0$li,xax=1,yax=2)
s.label(acp0$li,xax=1,yax=2,label=NULL)
Taille=inertia.dudi(acp0,row.inertia=TRUE)$row.cum[,2]/10000
for(i in 1:(length(Taille))){
if(Taille[i]<0.8) Taille[i]=0.01}
text(acp0$li[,1],acp0$li[,2],labels=DepReg$num,cex=Taille,col="red")
Taille=inertia.dudi(acp0,row.inertia=TRUE)$row.cum[,2]/10000
for(i in 1:(length(Taille))){
if(Taille[i]<0.6) Taille[i]=0.01
if(Taille[i]>=0.8) Taille[i]=0.01}
text(acp0$li[,1],acp0$li[,2],labels=DepReg$num,cex=Taille,col="blue")
s.corcircle(acp0$co)
```

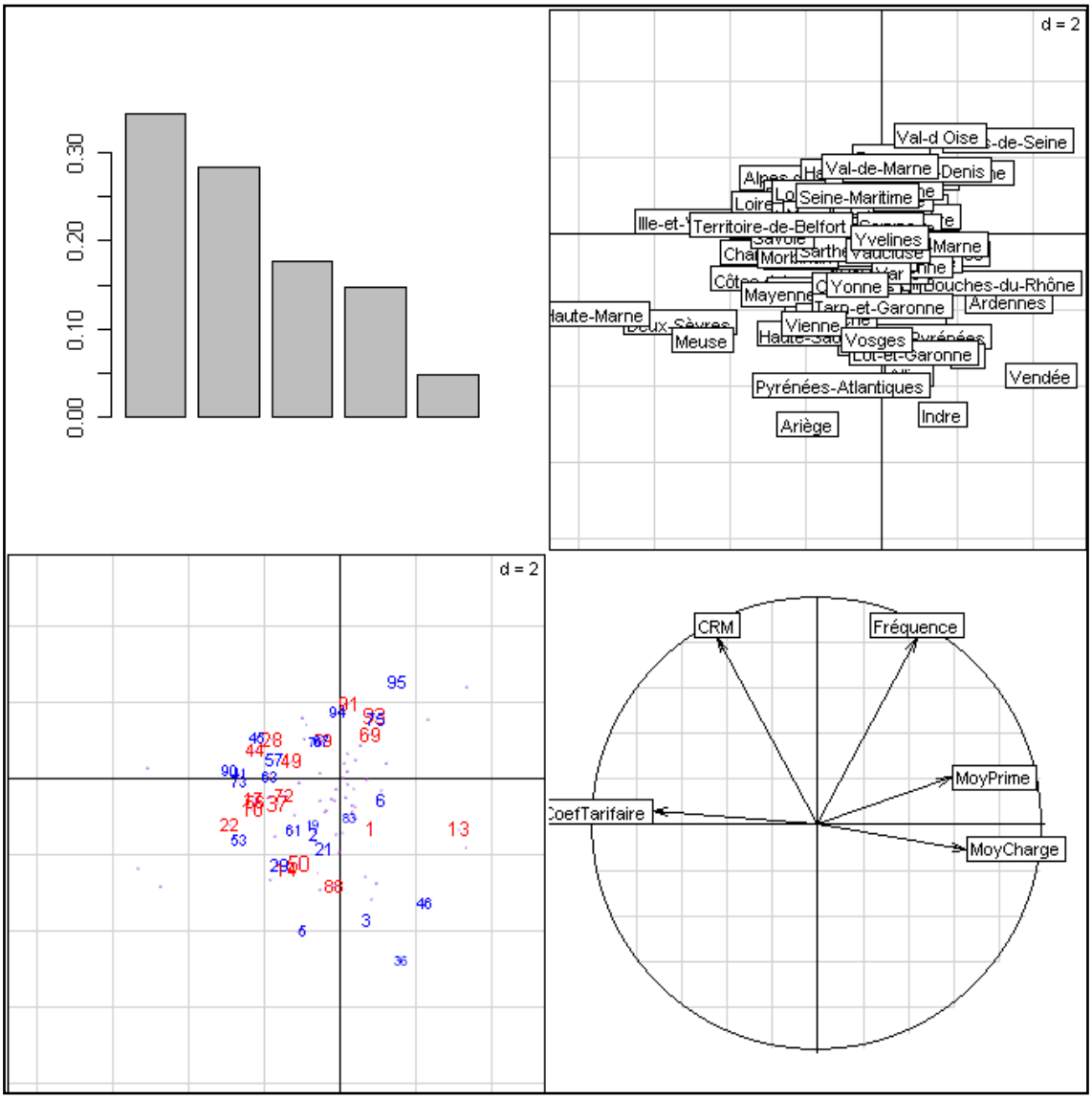


Illustration 10 : Valeurs propres, nuages de points et cercle de corrélations d'un exemple d'ACP

Nous pouvons alors constater que le premier axe est représenté par la prime moyenne et la charge moyenne, qui semblent corrélées. On note cependant que ces deux variables ne sont pas très proches de la circonférence, il faudra donc être prudent dans l'interprétation. On note que le deuxième axe semble composé de la fréquence et du CRM, qui parait alors complétement décorrélé de la charge moyenne. Ceci confirme alors l'éventuelle intuition que le coefficient bonus malus est par construction représentatif de la fréquence de sinistre et non du montant des sinistres. Il semblerait enfin que les départements dont le coefficient tarifaire moyen est élevé soient associés à une prime moyenne et une charge moyenne faible. Ce dernier constat est discutable et provient en réalité du fait que les départements à forts coefficients ont un poids faible et peu de sinistres, cet aspect ne sera donc pas pris en compte. Nous rappellons ici qu'un retour aux données est parfois nécessaire, comme ici, pour comprendre et interpréter une analyse. Nous pouvons enfin observer que conformément à une certaine intuition, le CRM donne une réelle information sur la fréquence de sinistre tandis que la prime moyenne donne une information sur la charge moyenne, en effet, plus la prime est élevée, plus le sinistre potentiel sera coûteux (coûts de réparations plus élevés, garanties plus larges etc…).

Revenons alors sur les départements. On constate d'ores et déjà que des départements tels que l'Ariège, la Meuse ou l'Indre qui étaient représentés sur le premier nuage ne le sont pas ou peu dans le second, il ne sont donc pas proches du plan et donc non interprétables ici. On pourra dire que les départements situés dans le carré supérieur droit tels que le Val-d'Oise, Seine Saint-Denis, l'Essonne ou encore le Rhône possèdent à la fois une charge moyenne élevée et une fréquence importante. A l'inverse, les départements du carré inférieur gauche tels que les Côtes d'Armor, le Morbihan ou l'Aube ont une faible fréquence et une charge moyenne faible. Sans aller plus dans le détail ici, nous pouvons déjà constater que les départements d'île de France et le Rhône possèdent une forte fréquence de sinistre, tandis que les Bouches du Rhônes possèdent une forte charge moyenne. Nous constatons ainsi un risque plus important dans les départements de grandes agglomérations. En contrepartie, on observe un risque plus faible dans la région de la Bretagne ainsi que dans d'autres départements du nord-ouest. Il est alors possible de représenter les coordonnées des points sur une carte de France. Ceci nous limite à une seule dimension, et nous décidons ici de représenter les deux premiers axes séparément. On utilisera pour ce faire des données situées dans le package « ade4 » dans la table « elec88 », qui contient les informations nécessaires au traçage d'une carte de France.

```
data(elec88)
par(mfrow=c(2,2))
area.plot(elec88$area,val=acp0$li[,1])
s.corcircle(acp0$co[,1:2])
area.plot(elec88$area,clab=1)
area.plot(elec88$area,val=acp0$li[,2])
```

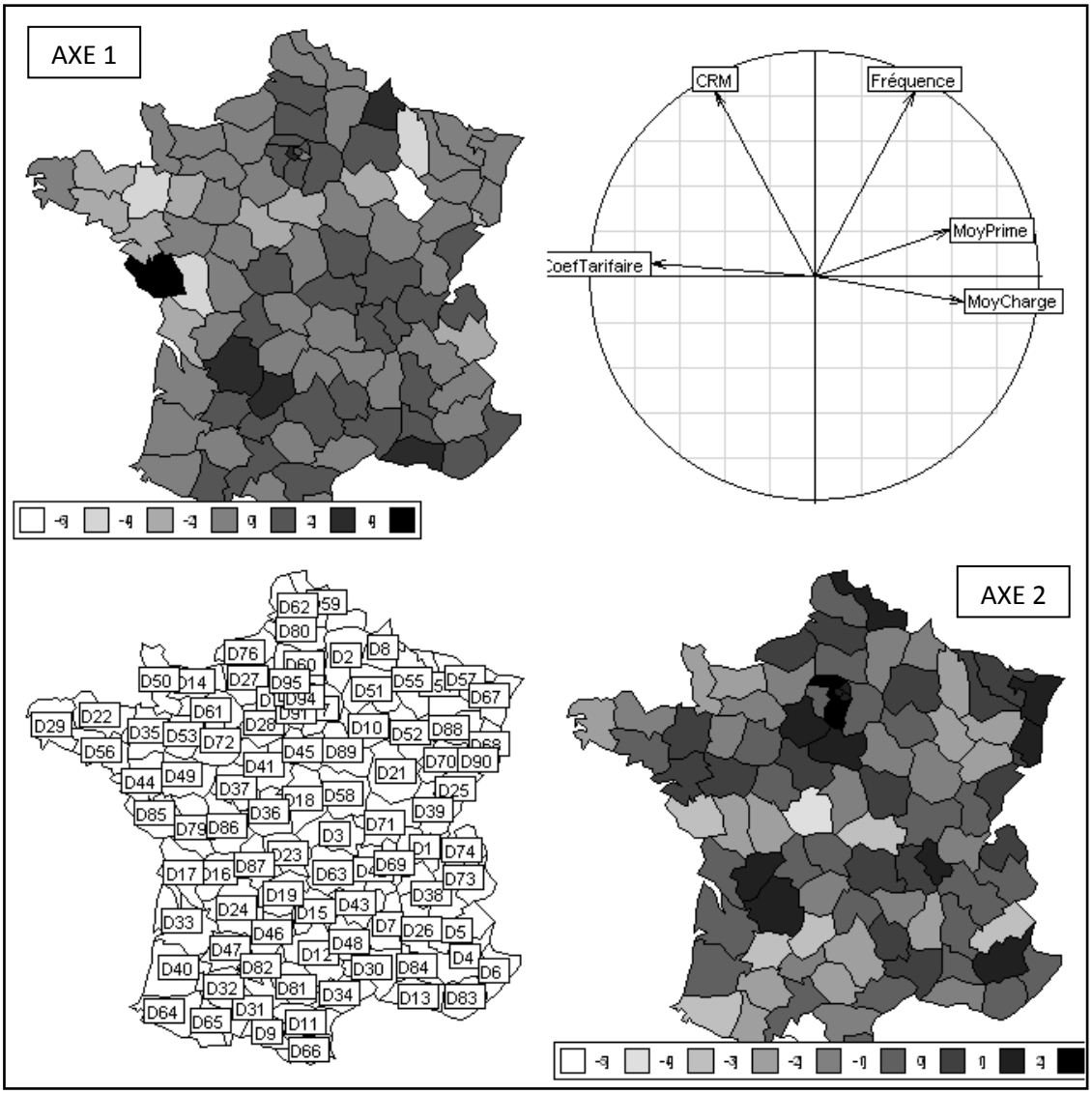


Illustration 11 : Représentation d'un exemple d'ACP sur une carte de France

Une telle représentation peut s'avérer intéressante mais doit être interprétée avec prudence. En effet, on se trouve ici en dimension 1 et une interprétation croisée est plus difficile. De plus, la notion de qualité de représentation n'est pas prise en compte ici. On note qu'il est également possible d'effectuer une représentation par région, si les données s'y prêtent, à condition de modifier « *elec88$area* » grâce à une fonction transcrite en Annexe 5. Nous remarquerons ici que les départements représentés en sombre dans les deux cartes de France sont des départements à risque et inversement les départements représentés en clair sont des zones géographiques à faible risque. Un tel tracé, pour être plus facilement interprétable, peut prendre en compte la qualité de représentation des données sur chaque axe comme précédemment avec l'objet « Taille ». Il faut alors utiliser la table obtenue par « *inertia.dudi(acp0,inertia=TRUE)$row.rel* », ce qui clarifie considérablement le graphique obtenu. Nous conclurons cependant qu'une telle analyse est insuffisante pour établir un zonier car non seulement elle doit être plus localisée (étude par département par exemple), mais elle doit en plus tenir compte d'autres facteurs tels que la présence d'agglomérations ou de zones de circulation particulière.

2.2.4 L'analyse factorielle des correspondances (AFC).

2.2.4.1 Aspect théorique.

Nous avons vu précédemment le principe théorique de l'ACP, qui est la base de l'analyse factorielle. Cette méthode est très utilisée pour des variables numériques et est d'autant plus intéressante lorsque le nombre d'individus est limité (étude par département par exemple). Cependant cette méthode n'est pas réellement applicable en ce qui concerne l'assurance automobile qui utilise un grand nombre de variables qualitatives, c'est pourquoi nous allons nous intéresser à l'AFC tout d'abord puis à l'ACM qui en est une généralisation.

L'AFC porte sur ce que l'on appelle un tableau de contingence de deux variables qualitatives. Nous considérons un tableau T de nombres positifs ou nuls, comportant I lignes et J colonnes, occurrences respectives de deux variables X et Y. On note n_{ij} son terme générique qui coïncide bien souvent à l'effectif correspondant à l'occurrence i de la variable X et à l'occurrence j de la variable Y. On note $n_{i.}$ et $n_{.j}$ les sommes marginales et n la somme de tous les éléments du tableau. Nous calculons alors les fréquences conjointes et les fréquences marginales ;

$$p_{ij} = \frac{n_{ij}}{n} \quad p_{i.} = \frac{n_{i.}}{n} \quad p_{.j} = \frac{n_{.j}}{n}$$

Nous définissons alors ce que l'on appelle le profil ligne qui est associé à la matrice $P_1 = D_1 T$ avec D_1 la matrice diagonale des poids des lignes (soit $1/n_{i.}$). Les sommes marginales des lignes sont alors égales à 1. Nous allons réaliser une ACP sur ce tableau du profil des lignes. On identifie tout d'abord les poids de chaque lignes de la matrice P_1, qui correspondent logiquement aux $\frac{n_{i.}}{n} = \frac{1}{n} D_1^{-1}$.

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2$$

La matrice associée à la métrique est alors en fait nD_2 qui correspond à la diagonale des $\frac{n_{.j}}{n}$. La principale justification de l'utilisation de cette métrique vient du fait que si deux colonnes j et j' ont même profil, alors il est logique de les regrouper en une seule d'effectifs $n_{ij} + n_{ij}$, sans que ceci implique une modification des distances. Nous vérifions facilement ceci pour $\frac{n_{ij}}{n_{.j}} = \frac{n_{ij'}}{n_{.j'}}$;

$$\alpha = (n_{.j} + n_{.j'}) \left[\frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2 + \frac{n}{n_{.j'}} \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)^2\right]$$

$$= n \left[\left(1 + \frac{n_{.j'}}{n_{.j}}\right) \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2 + \left(1 + \frac{n_{.j}}{n_{.j'}}\right) \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)^2\right]$$

$$= n \left[\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2 + \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)^2 + \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right) \left(\frac{n_{.j'}}{n_{.j}} \frac{n_{ij}}{n_{i.}} - \frac{n_{.j'}}{n_{.j}} \frac{n_{i'j}}{n_{i'.}}\right)\right.$$
$$\left. + \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right) \left(\frac{n_{.j}}{n_{.j'}} \frac{n_{ij'}}{n_{i.}} - \frac{n_{.j}}{n_{.j'}} \frac{n_{i'j'}}{n_{i'.}}\right)\right]$$

$$\alpha = n\left[\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2 + \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)^2 + \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)\left(\frac{n_{ij'}}{\cancel{n_{ij}}}\frac{\cancel{n_{ij}}}{n_{i.}} - \frac{n_{i'j'}}{\cancel{n_{ij}}}\frac{\cancel{n_{ij}}}{n_{i'.}}\right)\right.$$
$$\left. + \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)\left(\frac{n_{ij}}{\cancel{n_{ij'}}}\frac{\cancel{n_{ij'}}}{n_{i.}} - \frac{n_{i'j}}{\cancel{n_{ij}}}\frac{\cancel{n_{ij}}}{n_{i'.}}\right)\right]$$

$$= n\left[\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2 + \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)^2 + 2\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)\left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)\right]$$

$$= n\left[\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} + \frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right]^2$$

$$\alpha = n\left(\frac{n_{ij} + n_{ij'}}{n_{i.}} - \frac{n_{i'j} + n_{i'j'}}{n_{i'.}}\right)^2$$

Nous obtenons bien la relation suivante ;

$$Si\ \frac{n_{ij}}{n_{.j}} = \frac{n_{ij'}}{n_{.j'}}\ Alors\ \left[\frac{n}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2 + \frac{n}{n_{.j'}}\left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}}\right)^2\right] = \frac{n}{\left(n_{.j} + n_{.j'}\right)}\left(\frac{n_{ij} + n_{ij'}}{n_{i.}} - \frac{n_{i'j} + n_{i'j'}}{n_{i'.}}\right)^2$$

Fait remarquable, le terme de métrique du χ^2 vient du fait que l'inertie du nuage de points des profils lignes correspond à la quantité mesurant l'écart à l'indépendance que l'on retrouve dans le test d'indépendance du χ^2. En effet, l'inertie du nuage s'écrit, en notant g le centre de gravité de coordonnées $\frac{n_{.j}}{n}$;

$$\sum_{i=1}^{I}\frac{n_{i.}}{n}d^2_{\chi^2}(i,g) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{i.}}{n}\frac{n}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{i.}}{n_{.j}}\left(\frac{n_{ij} - \frac{n_{.i}n_{.j}}{n}}{n_{i.}}\right)^2 = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - \frac{n_{.i}n_{.j}}{n}\right)^2}{\frac{n_{.i}n_{.j}}{n}}$$

Le paragraphe précédent traitant du profil ligne peut être reproduit par symétrie pour le profil colonne. La réalisation de l'ACP du profil ligne revient alors à la diagonalisation de la matrice ${P_1}^T\frac{1}{n}D_1^{-1}P_1\ nD_2 = T^TD_1TD_2$. La réalisation de l'ACP du profil colonne revient alors à la diagonalisation de la matrice ${P_2}^T\frac{1}{n}D_2^{-1}P_2\ nD_1 = TD_2T^TD_1$.

Nous allons alors établir des relations entre les deux approches. Nous constatons d'ores et déjà que les deux approches conduisent aux mêmes valeurs propres. En effet, soit λ une valeur propre du profil ligne associé au vecteur propre u. On a alors la relation $(T^TD_1)(TD_2)u = \lambda u$ d'où $(TD_2)u = v$ non nul. On en déduit $(TD_2)(T^TD_1)(TD_2)u = (TD_2)(T^TD_1)v = \lambda(TD_2)u = \lambda v$. On peut alors conclure que λ est également valeur propre du profil colonne mais pour un vecteur propre associé différent $v = (TD_2)u$. Nous obligeons alors une normalisation par $\sqrt{\lambda}$, qui revient à rechercher un vecteur v qui après normalisation est $\frac{1}{n}D_2^{-1}$-normé, soit ;

$$v = k \times (TD_2)u\ et\ v^T\frac{1}{n}D_2^{-1}v = \lambda \Rightarrow\ k^2 \times u^TD_2T^T\frac{1}{n}D_2^{-1}(TD_2)u = \lambda$$

$$\frac{k^2}{n} \times u^T T^T D_2 D_2^{-1} T D_2 u = \lambda \Rightarrow \frac{k^2}{n} \times u^T \times D_1^{-1} \times \lambda u = \lambda \Rightarrow k^2 \left(u^T \frac{D_1^{-1}}{n} u \right) = 1 \Rightarrow k = \frac{1}{\sqrt{\lambda}}$$

$$\text{Donc } v = \frac{1}{\sqrt{\lambda}} T D_2 u \; et \; inversement \; u = \frac{1}{\sqrt{\lambda}} T^T D_1 v$$

Nous pouvons alors en déduire une relation simple entre les composantes principales d'une approche et les axes de l'autre approche. En effet, il est possible d'écrire les composantes principales du profil colonne comme suit ;

$$D_2 T^T n D_1 v = \frac{1}{\sqrt{\lambda}} D_2 T^T n D_1 T D_2 u = \frac{n}{\sqrt{\lambda}} D_2 \lambda u = n\sqrt{\lambda} D_2 u \; et \; inversement \; D_1 T n D_2 u = n\sqrt{\lambda} D_1 v$$

Nous pouvons donc conclure que les valeurs propres sont conservées et que les axes d'une analyse sont les composantes de l'autre et réciproquement. Il y a donc équivalence entre les deux approches et en pratique, on diagonalise dans la plus petite des deux dimensions. Il est également à noter que les occurrences de X et de Y seront représentées simultanément sur un seul et même graphique.

2.2.4.2 Exemple pratique.

Nous allons à présent mettre en avant un exemple concret. Pour ce faire nous allons étudier le tableau de contingence sur la classe et le groupe du véhicule (deux variables qualitatives) de l'ensemble des assurés ayant eu une garantie en 2008 chez « Mutant Assurances ». La valeur de comptage ou de remplissage du tableau correspond quant à elle à la somme de l'exposition réelle des assurés ou nombre de contrats acquis rentrant dans chaque catégorie.

Nous entendons par classe du véhicule la valeur à neuf de ce dernier. Cette variable a pour occurrences les lettres allant de A à M auxquelles s'ajoute la lettre X. Un véhicule de la classe A est un véhicule dont la valeur à neuf est faible, inversement, un véhicule de la classe M a une valeur à neuf élevée. La classe X représente soit des contrats suspendus (pour non paiements ou autre), soit des véhicules spécifiques partiellement assurables (garantie minimale), il s'agit essentiellement de véhicules à toit ouvrant.

Nous entendons par groupe du véhicule la puissance de ce dernier. Cette variable peut prendre les valeurs allant de 5 à 18 auxquelles se rajoute la valeur 99. Plus le véhicule est puissant, plus son groupe est élevé. Le groupe 99 est révélateur d'une suspension de contrat.

Nous notons TabAFC le tableau de contingence de ces deux variables (en plaçant la classe en lignes et le groupe en colonnes). Le tableau est représenté ci-dessous ;

	5	7	8	9	10	11	12	13	14	15	16	17	18	99
A	554	5623	2954	9327	9798	4051	718	23	1	0	0	0	0	0
B	0	37	53	2203	7351	13371	3928	714	89	0	0	0	0	0
C	0	0	12	35	645	8880	8784	4956	534	16	2	0	0	0
D	0	0	0	0	0	214	1760	3135	1339	272	1	0	0	0
E	0	0	0	0	0	107	583	1870	1703	301	76	5	0	0
F	0	0	0	0	0	0	80	405	809	354	114	1	1	0
G	0	0	0	0	0	0	6	132	434	259	81	5	0	0
H	0	0	0	0	0	0	1	3	28	100	58	3	2	0
J	0	0	0	0	0	0	0	12	26	54	60	3	5	0
K	0	0	0	0	0	0	0	1	7	18	43	5	3	0
L	0	0	0	0	0	0	0	0	1	5	10	8	5	0
M	0	0	0	0	0	0	0	0	0	0	15	5	11	0
X	67	3	1	13	12	29	41	37	120	29	16	2	16	1232

Illustration 12 : Tableau de contingence sur la classe et le groupe du véhicule (exemple d'AFC)

Nous allons alors réaliser une AFC de ce tableau, puis projeter les nuages de points des lignes et des colonnes dans un plan retenant un maximum d'informations. Nous utiliserons également les contributions des axes à la représentation des points lignes et des points colonnes pour interpréter la pertinence de la représentation de ces derniers. Enfin, nous effectuerons une AFC sur la transposée de ce tableau (inversion des lignes et des colonnes), pour nous convaincre qu'il y a équivalence entre la réalisation de l'AFC sur les lignes et la réalisation de l'AFC sur les colonnes. Là encore, nous utilisons le package « ade4 » du logiciel R.

```
afc0=dudi.coa(TabAFC,scannf=FALSE,nf=2)
par(mfrow=c(2,2))
barplot(afc0$eig/sum(afc0$eig))
s.label(afc0$co)
s.label(afc0$li)
scatter(afc0)
```

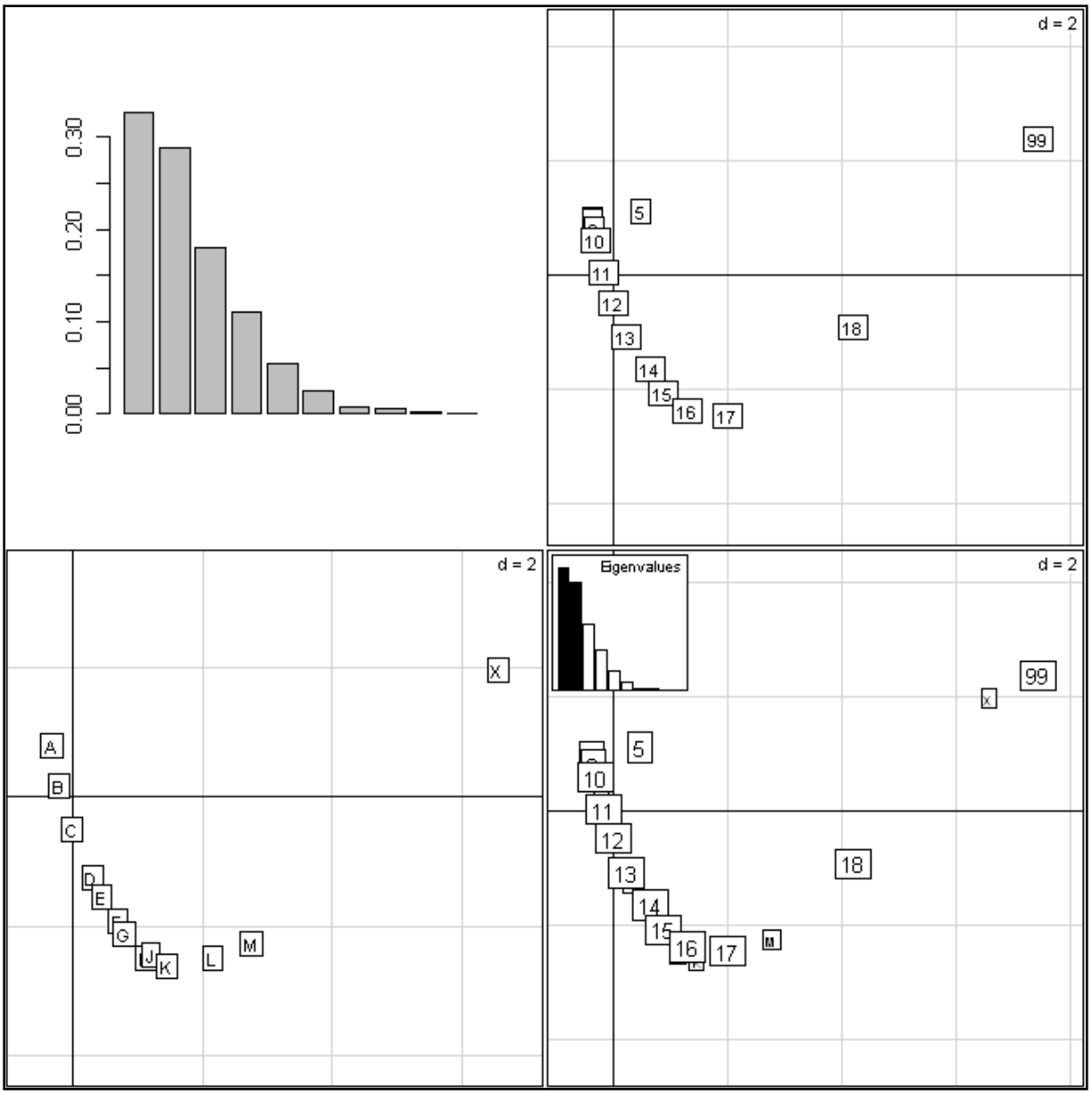


Illustration 13 : Représentation des valeurs propres et des nuages de points d'un exemple d'AFC

Nous constatons que les deux premiers vecteurs propres regroupent plus de 60 % de l'information du tableau, ce qui parait convenable. Nous remarquons ensuite une forte corrélation entre les deux variables qualitatives. Il semblerait que plus la classe est élevée, plus le groupe est élevé. Ceci signifie simplement que les voitures ayant une valeur à neuf plus élevée sont bien souvent des voitures plus puissantes. Nous noterons également le lien entre la classe X et le groupe 99 qui semblent relativement écartés des autres catégories. Ceci vient du fait qu'un nombre

important de véhicules dont la garantie à été suspendue par le passé sont automatiquement classifiés en groupe 99 et classe X. Or la valeur étudiée est l'exposition au risque et un contrat suspendu possède une exposition nulle. Après retour sur les données, nous constatons que la levée d'une suspension n'est pas suivie d'une mise à jour, ce qui constitue une erreur informatique. Enfin du fait de la corrélation entre les deux variables, le dernier graphique obtenu par la fonction « *scatter* » devient très peu lisible (superposition des points). De plus, les graphiques précédents ne nous donnent aucune information concernant la qualité de représentation des points. Nous allons donc représenter les deux nuages de points en tenant compte de la qualité de la représentation ;

```
TailleLig=inertia.dudi(afc0,row.inertia=TRUE,col.inertia=TRUE)$row.cum[,2]/10000
TailleCol=inertia.dudi(afc0,row.inertia=TRUE,col.inertia=TRUE)$col.cum[,2]/10000
plot(afc0$co[,1],afc0$co[,2],col="white",axes=TRUE,xlab='',ylab='')
grid()
text(afc0$co[,1],afc0$co[,2],colnames(TabAFC),cex=3*TailleCol,col="red")
text(afc0$li[,1],afc0$li[,2],rownames(TabAFC),cex=3*TailleLig,col="blue")
```

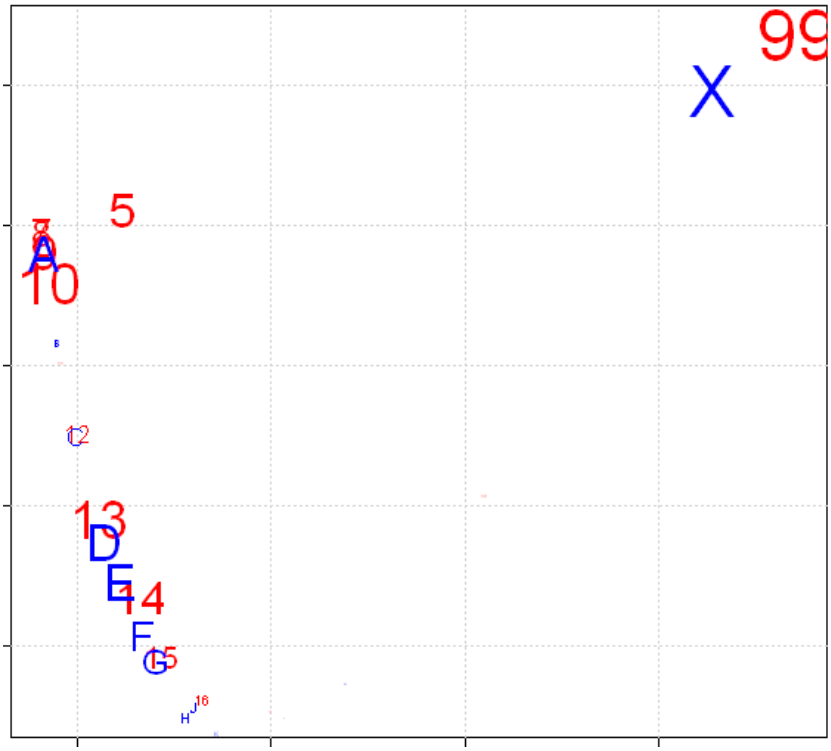


Illustration 14 : Représentation des deux nuages de points d'un exemple d'AFC en tenant compte de la qualité de représentation

Nous remarquons bien ici l'importance des occurrences « 99X ». Il faudrait alors retirer ces deux modalités de l'étude. Nous ne le ferons cependant pas ici dans le but de ne pas exposer une trop grande quantité de graphiques. Nous remarquons de plus que la classe A couvre une grande quantité de groupes (7 à 10), ce qui n'est pas le cas des classes suivantes. Le groupe 5 ne parait pas quant à lui être dans la continuité des autres. Les classes B et C étant mal représentées, nous ne pouvons interpréter leur proximité relative dans le plan. Nous pourrons tout de même admettre une certaine continuité d'association entre groupes et classes (B11, C12, D13 etc…), bien que cette association s'essouffle en fin de parcours (F et G avec 15, H et J avec 16).

Nous allons à présent effectuer une AFC sur la transposée du tableau initial et confronter les différentes représentations, dans le but de nous convaincre de l'équivalence entre l'étude des lignes et l'étude des colonnes ;

```
afc2=dudi.coa(t(TabAFC),scannf=FALSE,nf=2)
par(mfrow=c(2,2))
s.label(afc0$co)
s.label(afc0$li)
s.label(afc2$co)
s.label(afc2$li)
```

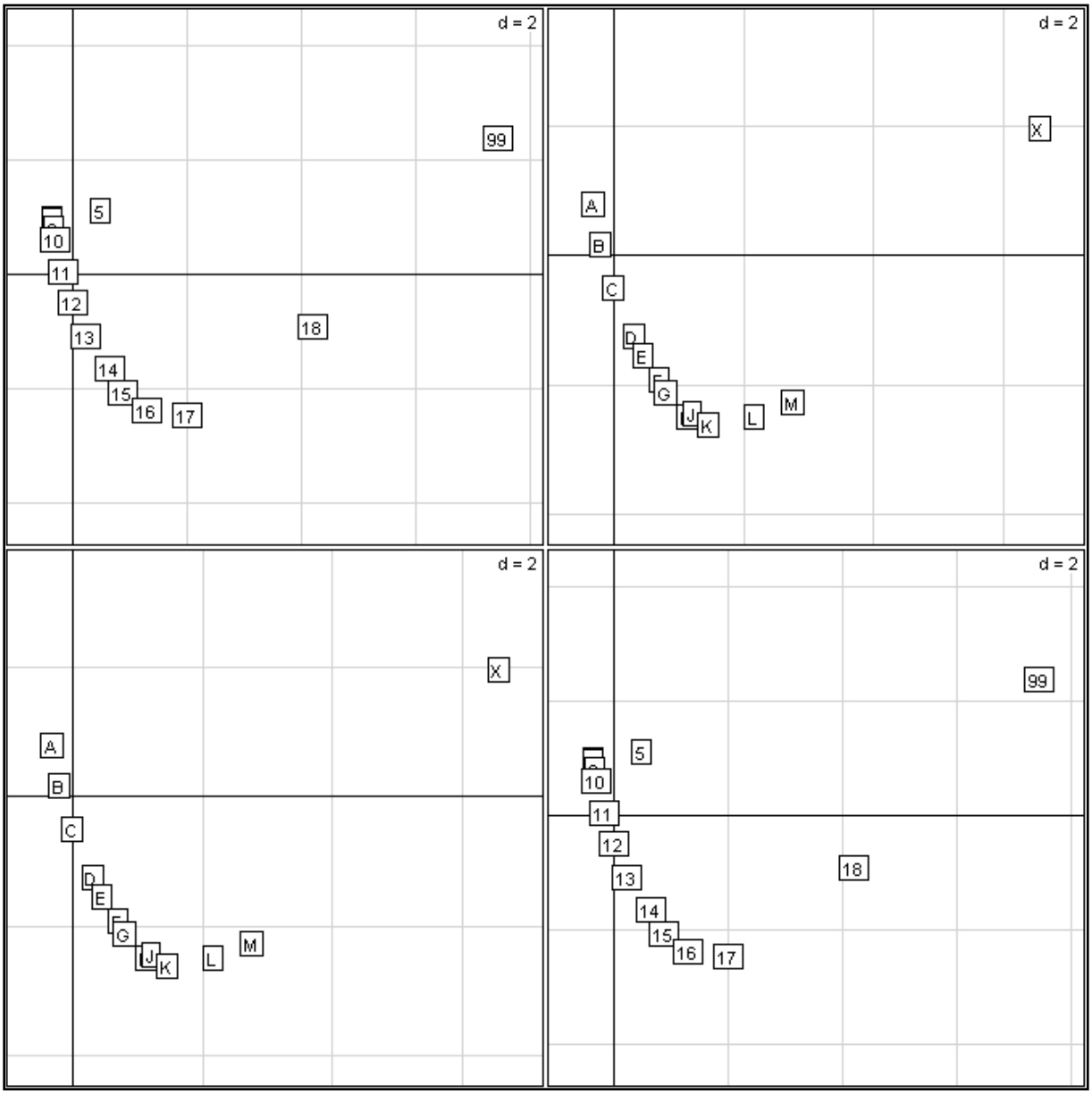


Illustration 15 : Confrontation du profil lignes et du profil colonnes d'un exemple d'AFC

Nous constatons en effet l'équivalence parfaite entre le graphique supérieur gauche et le graphique inférieur droit (et réciproquement entre le graphique inférieur gauche et le graphique supérieur droit). Il est donc équivalent de travailler sur les lignes et de travailler sur les colonnes.

2.2.5 L'analyse des correspondances multiples (ACM).

2.2.5.1 Aspect théorique.

Nous avons vu précédemment le fonctionnement de l'AFC, utile pour le traitement de deux variables qualitatives. Il est cependant souhaitable de pouvoir effectuer une analyse portant sur un nombre plus important de variables qualitatives. Nous allons pour ce faire décrire le fonctionnement de l'analyse des correspondances multiples ou ACM appelée également analyse factorielle des correspondances multiples ou AFCM. Cette analyse porte sur n individus caractérisés par p variables que l'on notera $(X^i)_{1\leq i\leq p}$. On note $(m_1, \quad \dots \quad , m_p)$ les nombres de modalités de chaque variable. Le nombre de modalités total est $= m_1 + \cdots + m_p$. On appelle alors tableau disjonctif complet le tableau contenant n lignes (les individus) et m colonnes (toutes les modalités). Chaque ligne contient alors des 1 ou des 0 en fonction de la modalité de chaque variable qui lui correspond. On remarque alors que la somme des éléments sur une ligne est toujours égale à p. On définit de plus le porteur de la j[ième] modalité de la variable X^k qui correspond au nombre d'individus l'utilisant et que l'on note ;

$$n_j^k = \sum_{i=1}^{n} x_{il} \ \ avec\ l = m_1 + \cdots + m_{k-1} + j \text{ et } \sum_{j=1}^{m_k} n_j^k = n$$

On définit alors ce que l'on appelle le tableau de Burt noté $\mathcal{B} = X^T X$. Il est alors possible d'écrire $\mathcal{B} = \left(\mathcal{B}_{ij}\right)_{\substack{1\leq i\leq p \\ 1\leq j\leq p}}$ avec $\mathcal{B}_{ij}$ la table de contingence de X^i et X^j si $i \neq j$ ou $\mathcal{B}_{ij}$ la diagonale des porteurs des modalités de j $diag(n_1^j, \dots, n_{m_k}^j)$ sinon.

Nous obtenons alors facilement certaines propriétés. Tout d'abord $\mathcal{B}$ est symétrique, ensuite la somme sur une ligne (ou sur une colonne) vaut p fois le porteur de la modalité associée à la ligne (ou à la colonne) et enfin la somme de tous les éléments vaut p²n. Nous résumons alors ces résultats à travers les deux tableaux ci-dessous ;

	X^1 : $1 \ \dots \ m_1$	$\dots \ \dots \ \dots$	X^p : $1 \ \dots \ m_p$	
1 ⋮ ⋮ ⋮ n	$0\ ou\ 1$	$0\ ou\ 1$	$0\ ou\ 1$	p ⋮ ⋮ ⋮ p
	$n_1^1 \ \dots \ n_{m_1}^1$	$\dots \ \dots \ \dots$	$n_1^p \ \dots \ n_{m_p}^p$	np

⇓

		X^1 : $1 \ \dots \ m_1$	$\dots \ \dots \ \dots$	X^p : $1 \ \dots \ m_p$	
X^1	1 ⋮ m_1	$n_1^1\ 0\ 0$ / $0\ \ddots\ 0$ / $0\ 0\ n_{m_1}^1$	$\rightarrow$ *Tableaux de contingence*	x_{il} pour $l \in [\![(p-m_p), p]\!]$ $1 \leq i \leq m_1$	$p \times n_1^1$ ⋮ $p \times n_{m_1}^1$
⋮		↓ *Tableaux de contingence*	↘ *Diagonales* ↖	*Tableaux de contingence* ↑	⋮
X^p	1 ⋮ m_p	x_{lj} pour $l \in [\![(p-m_p), p]\!]$ $1 \leq j \leq m_1$	*Tableaux de contingence* ←	$n_1^p\ 0\ 0$ / $0\ \ddots\ 0$ / $0\ 0\ n_{m_p}^p$	$p \times n_1^p$ ⋮ $p \times n_{m_p}^p$
		$p \times n_1^1 \ \dots \ p \times n_{m_1}^1$	$\dots \ \dots \ \dots$	$p \times n_1^p \ \dots \ p \times n_{m_p}^p$	$p^2 \times n$

Illustration 16 : Tableaux de contingences et de Burt

Nous allons alors effectuer une AFC de ce tableau de Burt. Il est tout d'abord à noter que par symétrie, il est équivalent de travailler sur le profil ligne ou sur le profil colonne. Nous considérons ici le profil ligne correspondant alors à la matrice $\frac{1}{p}D\mathcal{B}$ avec $= diag(\frac{1}{n_1^1}, \dots, \frac{1}{n_{m_1}^1}, \dots, \frac{1}{n_1^p}, \dots, \frac{1}{n_{m_p}^p})$. Le poids des lignes correspond alors à $\frac{1}{np}D^{-1}$ et la métrique utilisée est npD. Les vecteurs propres recherchés sont ceux de la matrice $\frac{1}{p}\mathcal{B}D\frac{1}{np}D^{-1}\frac{1}{p}D\mathcal{B}npD = (\frac{1}{p}\mathcal{B}D)^2$. Les vecteurs de coordonnées principales correspondent alors à $C^k = \frac{1}{p}D\mathcal{B}npDu_k = nD\mathcal{B}Du_k$.

Revenons alors sur le tableau initial noté X. Nous allons également effectuer une AFC de ce tableau. Nous décidons de travailler sur le profil ligne et donc sur la matrice $\frac{1}{p}X$, avec la matrice des poids $\frac{1}{n}$ et la métrique nD. La matrice à diagonaliser est donc $\frac{1}{p}X^T\frac{1}{n}XnD = \frac{1}{p}\mathcal{B}D$. Les vecteurs de coordonnées principales sont alors $C^k = \frac{1}{p}XnDu_k = \frac{n}{p}XDu_k$ et représentent les coordonnées des individus sur l'axe déterminé par le vecteur u_k. Nous remarquons alors le lien existant entre les individus et les variables. Si z représente les coordonnées de la modalité d'une variable et u les coordonnées d'un individu, alors ;

$$z = \frac{1}{\sqrt{\lambda}}XDu \; et \; u = \frac{1}{\sqrt{\lambda}}\frac{1}{p}Xz$$

Cela signifie d'une part que les coordonnées d'une catégorie correspondent à la moyenne arithmétique des coordonnées des individus de cette même catégorie, et d'autre part que les coordonnées d'un individu correspondent à la moyenne arithmétique des catégories auxquelles il appartient, tout ceci à un facteur $\frac{1}{\sqrt{\lambda}}$ près. La représentation d'une catégorie est donc le barycentre des individus qui la compose, et pour s'assurer cette visualisation, on se portera sur l'axe $\sqrt{\lambda}u$ et non u.

Nous notons alors que le vecteur (1,…,1) est un vecteur propre trivial associé à la valeur propre 1 (car les marges ou sommes en lignes de $\frac{1}{p}\mathcal{B}D$ valent 1). Nous savons alors que les autres vecteurs propres lui sont orthogonaux, et sont donc de moyenne nulle. Le centre de gravité de la représentation des individus et des catégories est donc au centre du plan. Nous ajouterons sans plus nous attarder que les effectifs et le nombre de modalités de chaque variable influent sur l'inertie qui leur est associée et que en pratique, on cherchera à éviter les trop grandes disparités entre variables et les effectifs trop faibles.

Nous remarquerons enfin un lien entre les deux approches (AFC du tableau de Burt ou AFC du tableau initial) par le fait que les vecteurs propres obtenus sont les mêmes, mais pas les valeurs propres. En effet, si λ, u sont une valeur propre et un vecteur propre résultant de l'AFC du tableau disjonctif complet, alors u est un vecteur propre de l'AFC du tableau de Burt associé à la valeur propre λ^2. Ceci se démontre facilement comme suit ;

$$\frac{1}{p}BDu = \lambda u \; \Rightarrow \left(\frac{1}{p}BD\right)^2 u = \left(\frac{1}{p}BD\right)\lambda u = \lambda^2 u$$

2.2.5.2 Exemple pratique.

Nous allons à présent exposer un exemple concret afin de mettre en avant l'intérêt de la mise en œuvre d'une ACM et un exemple d'interprétation tout en illustrant graphiquement la relation entre la réalisation d'une AFC sur le tableau disjonctif complet (principe de l'ACM) et l'exécution d'une AFC sur le tableau de Burt défini plus haut.

Nous allons considérer un tableau composé de clients de « Mutant Assurances » sinistrés en 2008, en France métropolitaine. Les variables caractérisant ces individus sont la nature du sinistre (Corporel ou Matériel), une variable dénommée « Option » qui représente une segmentation particulière effectuée par « Mutant Assurances » en fonction de l'ancienneté d'assurance du client, l'usage du véhicule (de modalités U2, U3 ou U4), la zone du véhicule déterminée grâce à un zonier (cartographie du risque par le biais de différentes modalités ; Z2, Z4, Z5, Z6 ou Z9), la classe et le groupe du véhicule évoqués dans l'exemple concret de l'AFC (les groupes sont ici recodés de manière à éviter toute confusion, 5 devient G5, 6 devient G6, etc…), et enfin le département pour une éventuelle comparaison avec la zone. Les premières lignes du tableau nommé TabACM sont représentées pour faciliter la visualisation ;

```
head(TabACM)
```

	NatureSinistre	Options	Usage	Zone	Classe	Groupe	Département
1	Matériel	SEL	U2	Z5	C	G12	Rhône
2	Matériel	C07	U2	Z5	B	G12	Rhône
3	Matériel	ESS	U2	Z4	C	G11	Gard
4	Matériel	ESS	U2	Z2	A	G11	Côtes d Armor
5	Matériel	BON	U2	Z5	E	G13	Rhône
6	Matériel	BON	U2	Z5	E	G13	Rhône

Nous réalisons ensuite l'ACM directement sur ce tableau, nous représentons alors les valeurs propres puis nous demandons à visualiser les modalités de chaque variable avec un nuage de points des individus ;

```
acm0=dudi.acm(TabACM)
3
x11()
scatter(acm0)
```

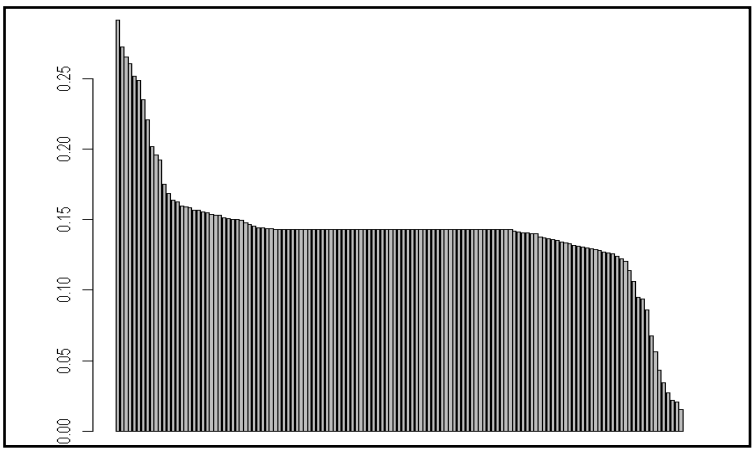


Illustration 17 : Représentation des valeurs propres

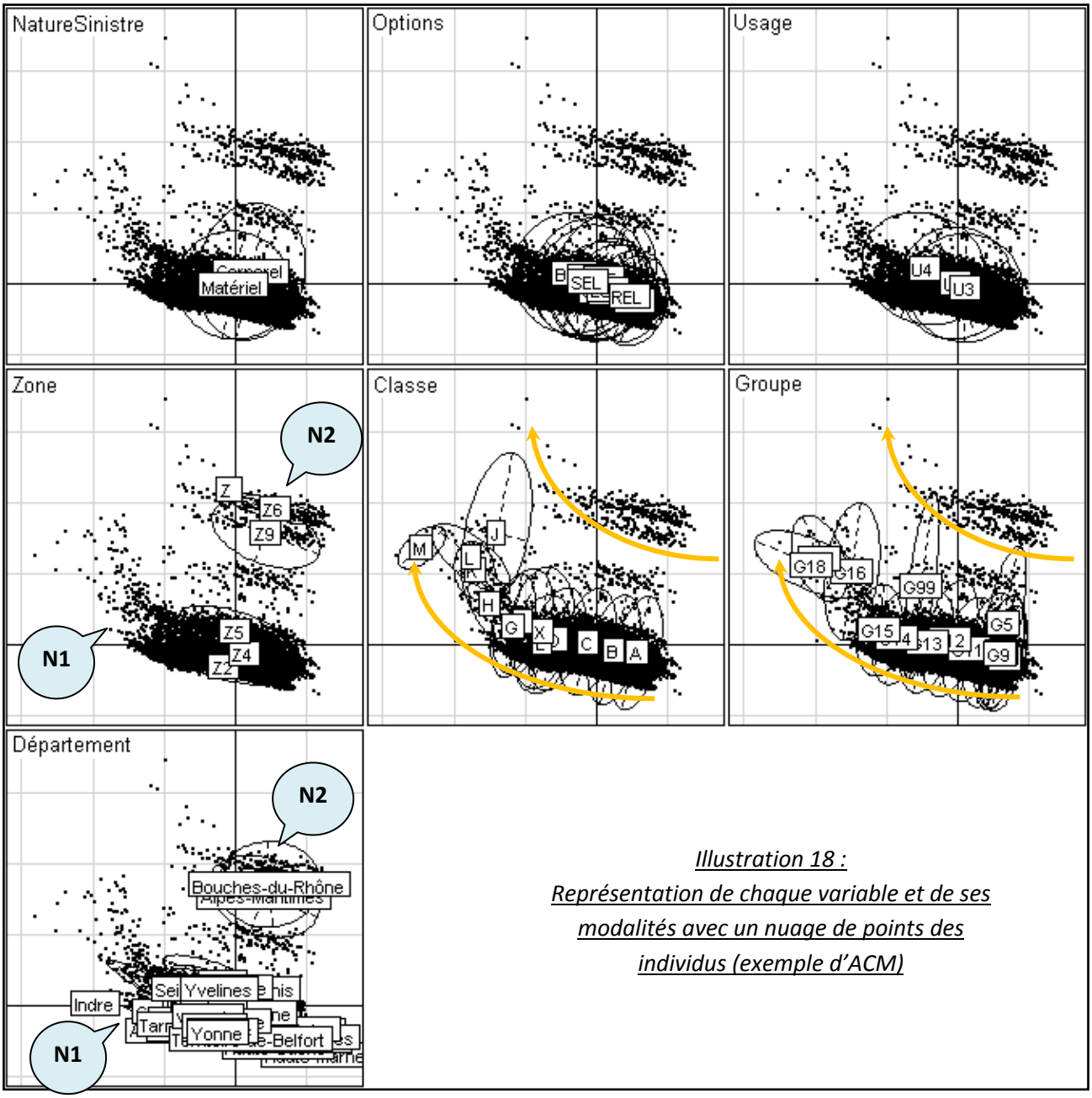


Illustration 18 :
Représentation de chaque variable et de ses modalités avec un nuage de points des individus (exemple d'ACM)

Les graphiques ci-dessus nous donnent un grand nombre d'informations. Tout d'abord, nous retrouvons la forte corrélation entre la classe et le groupe du véhicule, que nous avions aperçu lors de l'exemple concret concernant l'ACP. Nous remarquons ensuite que l'ensemble des individus sinistrés sont associés à une valeur à neuf et à une puissance du véhicule faible ou moyenne tandis que les véhicules de classes plus élevées sont possédés par une poignée d'individus se différenciant des autres et représentés plus en haut à gauche sur les graphiques. Ce même schéma se retrouve sur deux groupes distincts d'individus (voir les flèches sur les graphiques). En effet, deux groupes d'individus symbolisés par les nuages de points nommés N1 et N2 se distinguent. Cette distinction est essentiellement due à la zone et au département. Nous constatons ainsi un lien fort entre deux départements (les Bouches du Rhône et les Alpes Maritimes) et les deux zones Z6 et Z9. Ceci vient du fait que ces zones géographiques sont des zones à risques et donc fortement associées aux zones Z6 et Z9, qui caractérisent les zones les plus risquées. Enfin, l'interprétation des variables restantes est plus complexe car leurs modalités sont très regroupées.

En pratique, certaines variables sont retirées petit à petit pour parfaire l'analyse. L'ensemble des graphiques réalisés ne sera cependant pas représenté ici. Nous commencerions ici par retirer les départements et le groupe, étant donnée leur corrélation respective avec la zone et la classe. Puis la variable classe, qui est alors prédominante et aura tendance à étaler les graphiques tout en laissant les variables restantes dans un périmètre très condensé, sera extraite. Enfin, serait retirées les lignes pour lesquelles la variable zone prend la valeur « Z », car ces dernières possèdent un comportement atypique qui de plus ne peut être analysé ici car elles représentent des valeurs manquantes. Il sera alors possible de regarder plus en détails les variables restantes, qui ne sont donc pas des variables prédominantes dans la caractérisation des individus sinistrés. Une première explication plausible viendrait du fait que les différences entre usages sont faibles (la preuve en est qu'un retour aux sources des données et à la tarification en vigueur nous informe qu'il n'existe pas de différence de tarif entre l'usage 2 et l'usage 3). Ceci n'est en revanche pas vérifié pour l'usage 4 caractérisant l'utilisation professionnelle du véhicule qui se distingue quelque peu sur le graphique initial comme étant attaché à des véhicules de puissance et de valeur à neuf plus élevés que la moyenne. Une deuxième explication concerne les options, qui sont en réalité une segmentation plus fine mise en place par « Mutant Assurances ». Cette segmentation étant plus fine, elle n'est donc pas un facteur prédominant dans la différenciation des individus, et il est donc logique que cette différence ne soit pas évidente. Le retrait de certaines variables nous permet, comme signalé plus haut, d'affiner le raisonnement (Voir schéma ci-dessous), et nous pouvons alors constater que les options commençant par un C (CLE, C12, C20, C07) sont en majorité associées à des sinistres corporels, de même que les zones Z6 et Z9. Or les sinistres corporels sont révélateurs d'un risque plus important car ils sont souvent plus coûteux. Ceci nous indique donc que les options en question sont associées à des individus à risque, de même que les zones Z6 et Z9. Un tel renseignement peut facilement se vérifier en consultant la documentation appropriée au sein de l'entreprise.

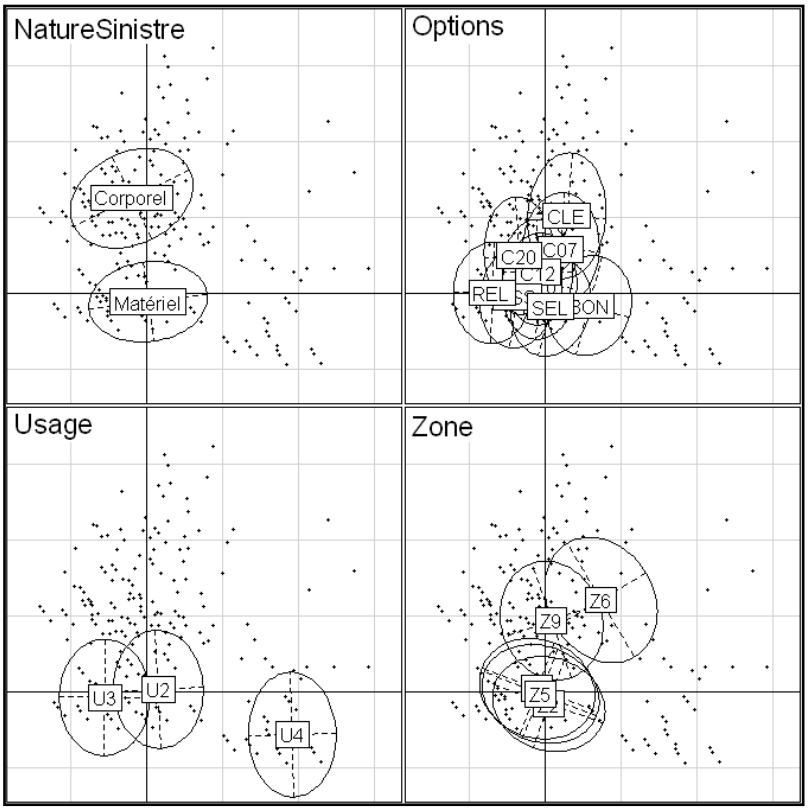


Illustration 19 : Exemple d'ACM pour une quantité de variables restreinte

Nous allons à présent, afin de comprendre et de visualiser les fondements théoriques de l'ACM, exécuter l'AFC du tableau disjonctif complet (composé des indicatrices des modalités et donc de 1 ou de 0), ainsi que l'AFC du tableau de Burt, composé de matrices diagonales et des tableaux disjonctifs des variables deux à deux. Les graphiques des valeurs propres et des nuages de points de ces deux approches sont alors confrontés ;

```
par(mfrow=c(2,2))
TDisj=acm.disjonctif(TabACM)
acpDisj=dudi.coa(TDisj)
3
scatter(acpDisj)
TBurt=acm.burt(TabACM,TabACM)
acpBurt=dudi.coa(TBurt)
3
scatter(acpBurt)
```

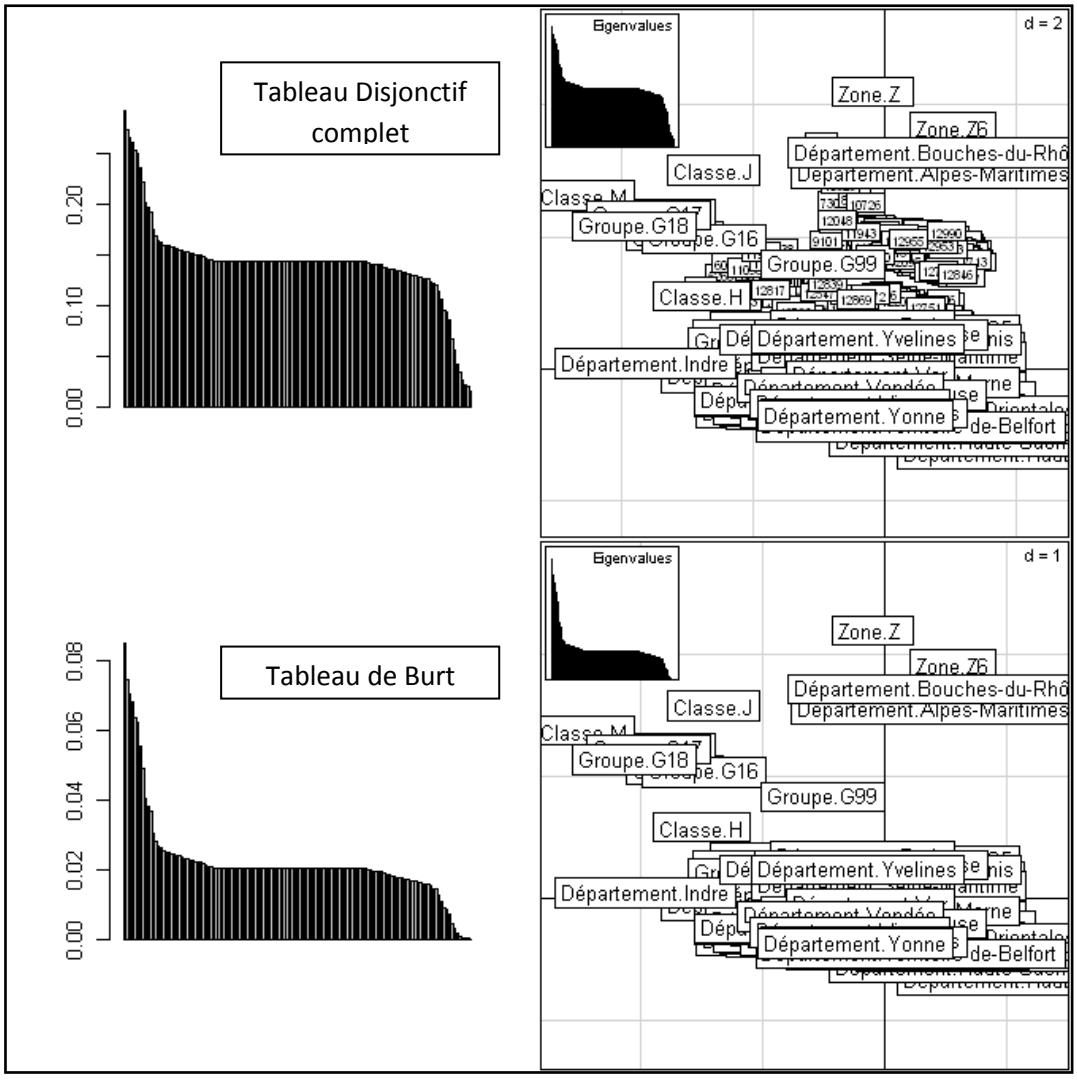


Illustration 20 : Représentation des valeurs propres et nuages de points pour le tableau disjonctif complet et le tableau de Burt d'un exemple d'ACM

Nous constatons tout d'abord que l'ACP du tableau disjonctif complet correspond exactement à la réalisation de l'ACM (et c'est ainsi que nous l'avions défini). La seule différence est alors que la représentation issu de la fonction « *scatter* » contient le nuage de points des individus (sous forme de numéros encadrés) ainsi que toutes les modalités des variables, ce qui rend le graphique illisible en comparaison avec les résultats obtenus directement avec la fonction « *scatter* » lors de l'ACM.

En ce qui concerne l'AFC du tableau de Burt, le graphique des valeurs propres possède une forme tout à fait semblable, mais plus écrasé. Ceci vient du fait que comme nous l'avons démontré dans l'approche théorique, les vecteurs propres de l'AFC du tableau disjonctif complet sont les mêmes que ceux obtenus lors de l'AFC du tableau de Burt, à ceci près que les valeurs propres associées à la deuxième méthode sont le carré des valeurs propres obtenues avec la première méthode. Nous vérifions rapidement et simplement ceci ;

```
> head(acpBurt$eig)
[1] 0.08495277 0.07432470 0.07049449 0.06800188 0.06340144 0.06201577
> head(acpDisj$eig*acpDisj$eig)
[1] 0.08495277 0.07432470 0.07049449 0.06800188 0.06340144 0.06201577
```

Nous remarquons ensuite que la fonction « *scatter* » nous renvoie uniquement les modalités des variables, mais ceci est logique car dans le tableau de Burt, les variables sont reportées aussi bien en lignes qu'en colonnes. La position des variables sur le graphique en revanche est exactement la même que lors de l'AFC du tableau disjonctif complet, à ceci près que l'échelle de grandeur est différente (symbolisée par d=2 et d=1 ici). Cet écart vient du fait que lors de la représentation en AFC des individus du profil colonne et des individus du profil ligne, un facteur correspondant à la racine de la valeur propre associée à l'axe de projection intervient. Or la racine de 0,08495 vaut environ 0,29146 et la racine de 0,29146 vaut environ 0,53987, ce qui explique la différence d'échelle qui s'effectue avec un arrondi (0,53987 vaut grossièrement le double de 0,29146).

2.2.6 *L'analyse dite mixte de données qualitatives et quantitatives.*

Nous avons pu voir précédemment comment analyser un tableau de n individus et p variables toutes quantitatives ou toutes qualitatives. Cependant, il arrive bien souvent que l'ensemble des variables ne soit pas conforme à un seul type de données. On parle alors de données mixtes (données quantitatives + qualitatives). La solution la plus simple et la plus fréquente est alors de regrouper les variables quantitatives en classes et donc de les transformer en variables qualitatives. Le choix des classes sera alors déterminant et tout découpage devra être issu d'une réflexion avancée et justifiable. Il existe cependant une généralisation des principes de l'ACP et de l'ACM qui utilisées de manière combinée permettent l'étude de données mixte.

Nous décrirons ici le principe de cette analyse que l'on retrouve parfois sous le nom d'AFDM (Analyse factorielle de données mixte). Une approche de l'ACP normée consiste à rechercher une variable synthétique (la première composante principale) qui maximise la somme des rapports de corrélations linéaires entre cette variable fictive et les variables initiales. De même il est possible d'interpréter l'ACM comme la recherche d'une variable fictive qui maximise la somme des rapports de corrélations multiples entre la variable synthétique et les indicatrices initiales. L'AFDM a pour principe de mettre dans un même tableau les variables quantitatives centrées et normées ainsi que le tableau disjonctif complet des variables qualitatives. Nous allons alors diviser les éléments des variables qualitatives par la racine du porteur de leur modalité (équivalent à considérer des poids en colonne). La maximisation du rapport de corrélation multiple obtenu avec une ACM s'obtient alors à présent avec une ACP et l'on va chercher à maximiser la somme des rapports de corrélations linéaires et multiples en effectuant une ACP sur le tableau ainsi transformé. Une logique inverse et symétrique existe et consiste à travailler sur les variables quantitatives afin que ces dernières soient traitées par une ACM. Nous donnons à titre indicatif la modification à opérer ; il s'agit en fait de diviser chaque variable quantitative en deux prenant les valeurs $\frac{1-x}{2}$ et $\frac{1+x}{2}$.

Nous conclurons alors avant d'exposer un exemple global reprenant les étapes d'extraction et d'analyses que les méthodes d'analyses de données sont diverses et variées et qu'il existe un grand nombre de dérivées ou de généralisations des méthodes énoncées plus haut. Le lecteur intéressé par la théorie de l'analyse factorielle est invité à se reporter vers des ouvrages associés à des auteurs tels que Y.Escoufier [34], B. Escoffier [32], [33], J. Pagès [51], ou encore G. Saporta [56].

Chapitre 3 : Extraction, analyses préliminaires (Exemple numérique).

3.1 Extraction des données.

3.1.1 Première extraction et définition du périmètre.

Nous allons à présent mettre en avant une application numérique et concrète reprenant les principales étapes décrites plus haut de façon théorique. Nous exposerons par la suite une analyse descriptive et une analyse factorielle ainsi que les intuitions de modélisation qu'elles apportent. Mais auparavant, nous allons définir clairement un périmètre d'étude, tout en donnant un certain nombre de contraintes et de problématiques rencontrées lors d'une extraction de données.

Le périmètre de l'exemple étudié porte sur l'année d'exercice 2008. Nous allons ainsi nous restreindre à l'ensemble des contrats justifiant d'au moins un jour de garantie durant l'année 2008, et à l'ensemble des sinistres déclarés en 2008. Les types de contrats dits « flotte » (assurance collective d'un ensemble de véhicules) et les assurés salariés de « Mutant Assurances » (société de référence) seront retirés de l'étude étant donné le fait qu'ils représentent un risque et une gestion particulière.

La première étape consiste donc à extraire l'ensemble des données pouvant correspondre à ce périmètre. Auparavant, il est nécessaire de connaitre certains fonctionnements des bases de données au niveau informatique ainsi que les variables pertinentes pour l'extraction. Dans le cadre de notre exemple, plusieurs points doivent être respectés. Il est tout d'abord essentiel de ne considérer que les contrats automobiles, ce qui ne pose en règle générale et dans notre exemple aucun problème, le type de garantie étant bien souvent renseigné dans une variable. Mais il est également nécessaire de sélectionner les contrats actifs en 2008. Il est donc capital d'avoir accès à un identifiant de contrat et à un renseignement concernant la date de début d'effet du contrat et la date de fin d'effet. Ce point est ici problématique, et nous allons le détailler plus bas.

Afin de replacer les choses dans leur contexte, nous allons expliquer brièvement la logique informatique qui accompagne cet exemple. Chaque client est représenté par un numéro de client, chaque contrat est repéré par un numéro de contrat et chaque avenant se voit attribué un identifiant de contrat. A un client peut donc correspondre plusieurs numéros de contrat et à chaque numéro de contrat peut correspondre plusieurs identifiants de contrat. Chaque avenant est également repéré par une variable nommée « identprecedent ». Cette variable prend la valeur 0 si le contrat est une affaire nouvelle, et la valeur de l'identifiant de l'ancien contrat en cas de création d'un avenant. Nous allons reprendre ce principe à travers un tableau afin d'en faciliter la compréhension ;

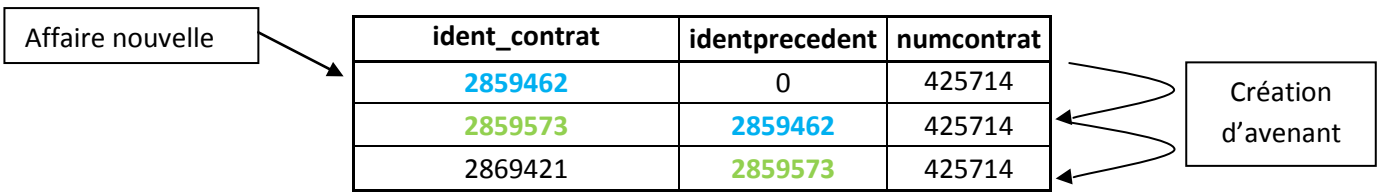


Illustration 21

En ce qui concerne les tables mises à notre disposition, il s'agit d'une part de l'ensemble des contrats automobiles ayant fait l'objet d'une quittance en 2008 et d'autre part de l'ensemble des informations concernant les contrats et les sinistres, accessibles par extraction (une requête d'extraction doit donc être rédigée). Nous allons ainsi, à partir des variables précitées et des tables mises à notre disposition, extraire l'ensemble des avenants correspondant au périmètre défini plus haut.

A partir de la table des quittances 2008, il est ici possible d'extraire les numéros de contrats automobiles correspondant au périmètre, mais pas les identifiants de contrats. L'information obtenue est donc incomplète. Afin d'obtenir la totalité de l'information (c'est-à-dire tous les avenants), il est nécessaire d'effectuer une requête renvoyant l'ensemble des avenants correspondants aux numéros de contrats obtenus plus haut. Malheureusement, les avenants associés aux numéros de contrats extraits ne correspondent logiquement pas tous à l'année d'exercice 2008 (un numéro de contrat reste le même au fil des années, mais l'identifiant de contrat change au fil des avenants). Il faut alors effectuer un tri dans les identifiants de contrats obtenus afin de ne sélectionner que ceux correspondant au périmètre.

Pour effectuer ce tri, la date d'effet et la date de fin du contrat doivent être utilisées. Si la date d'effet est présente dans les bases de données informatiques, il n'en est pas de même pour la date de fin, que nous allons donc créer. Il existe alors plusieurs possibilités à prendre en compte. Soit l'avenant a fait l'objet d'une résiliation, et la date de fin correspond alors à la date de résiliation. Soit l'avenant est le dernier en date, et la date de fin correspond à une date fictive que nous fixeront au 31/12/9999. Soit enfin l'avenant n'est pas le dernier en date, et la date de fin correspond alors à la date d'effet de l'avenant suivant, qu'il est possible de repérer grâce à la variable « identprecedent » décrite plus haut. Une fois la variable représentant la date de fin crée, il ne reste plus qu'à sélectionner les avenants pour lesquelles la date d'effet précède le 31/12/2008 et la date de fin se situe après le 01/01/2008. Nous avons ainsi obtenu l'ensemble des avenants correspondants au périmètre prédéfini, à ceci près qu'il reste à retrancher les contrats flottes et les salariés de « Mutant Assurances ». Nous ne détaillerons pas cette dernière opération qui fait simplement appel à deux tables supplémentaires, l'une contenant l'ensemble des numéros de contrats flottes et l'autre contenant l'ensemble des numéros de clients salariés de « Mutant Assurances » car il suffit alors de retirer tous les avenants associés soit à un numéro de contrat flotte soit à un numéro de client salarié de « Mutant Assurances ».

3.1.2 Détection des erreurs, épurement des données.

Comme nous l'avons fait remarquer plus haut, l'extraction doit faire l'objet d'un travail d'épurement des données, afin d'obtenir la certitude que les tables utilisées sont pertinentes et fiables. Le paragraphe précédent nous a décrit l'obtention de l'ensemble des identifiants de contrats conformes au périmètre. Ces derniers nous permettent alors d'obtenir dans le cas présent trois tables distinctes. La première contient les informations de base concernant les contrats ainsi que certaines caractéristiques principales de la tarification. La seconde contient des informations sur les sinistres et la troisième des caractéristiques détaillées sur les conducteurs désignés au contrat. L'enjeu de cette partie est de créer certaines variables utiles et de certifier que les données sont fiables car une étude technique, aussi poussée et sérieuse soit elle, ne peut donner de résultats corrects si elle se base sur des données erronées. Nous verrons également l'impact que peut avoir la détection d'une erreur dans l'évolution des systèmes informatiques.

En ce qui concerne la table des contrats, nous allons chercher à créer une variable contenant l'exposition réelle au risque en nombre de jours. En effet, un avenant pour lequel la durée de garantie associée est de quelques jours seulement ne doit pas être comptabilisé comme une entité entière. Il est alors nécessaire de créer deux variables dates fictives. La première notée « DateEffetFictif » correspond soit au 01/01/2008 soit à la date d'effet car dans le but de respecter le périmètre, nous devons nous limiter à l'exposition réelle en 2008. La variable « DateFinFictif » est définie de la même manière et l'exposition réelle au risque est obtenue en calculant la différence entre ces deux dates. Il est cependant nécessaire de fixer à 0 l'exposition concernant d'éventuels contrats suspendus restés en portefeuille avant résiliation et nous prenons également soin de supprimer les incohérences informatiques telles que la présence d'exposition négatives.

En ce qui concerne la table des sinistres, une simple vérification de l'adéquation de la date d'accident avec les dates d'effet et de fin calculées ci-dessus nous permet de mettre en évidence une incohérence pour un pourcentage de sinistres non négligeable. Une recherche approfondie permet alors de constater que l'identifiant de contrat associé à un sinistre n'est pas toujours le bon. En effet, un sinistre est dans plus de 7 % des cas associé à un identifiant de contrat antérieur à celui réellement impacté. Cette incohérence est représentée ci-dessous à travers deux tableaux pour en faciliter la compréhension. Le premier tableau représente un cas où l'avenant associé au sinistre est l'avenant directement antérieur à celui qui aurait dû l'être, le deuxième tableau est une généralisation pour un avenant encore plus antérieur.

Information initiale

ident_contrat	ident_precedent	DateEffet	DateFin	DATE ACC
~~3110720~~	~~2940576~~	~~09/05/2008~~	~~09/05/2008~~	~~27/11/2008~~
3324245	3110720	09/05/2008	09/12/2008	**27/11/2008**

Réalité

Information initiale

ident_contrat	ident_precedent	numcontrat	DateEffet	DateFin	DATE ACC
~~2962098~~	~~2801834~~	~~769571~~	~~31/12/2007~~	~~29/05/2008~~	~~01/08/2008~~
~~3129466~~	~~2962098~~	~~769571~~	~~29/05/2008~~	~~29/05/2008~~	~~01/08/2008~~
~~3134886~~	~~3129466~~	~~769571~~	~~29/05/2008~~	~~29/05/2008~~	~~01/08/2008~~
~~3264042~~	~~3134886~~	~~769571~~	~~29/05/2008~~	~~29/05/2008~~	~~01/08/2008~~
3264203	3264042	769571	29/05/2008	31/12/2008	**01/08/2008**

Réalité

Illustration 22

Afin de reconstituer l'information réelle, il faut donc comparer l'adéquation de la date d'accident non pas avec la date d'effet et la date de fin de l'identifiant de contrat associé, mais avec l'ensemble des dates d'effet et des dates de fin pour le numéro de contrat associé. Cette étape permet alors de récupérer l'identifiant de contrat correct qui servira à effectuer un lien avec les autres tables dans le but de croiser des informations.

Pour ce qui est de la table contenant des informations sur les conducteurs, de nombreux contrôles doivent être opérés. Mais commençons tout d'abord par décrire rapidement les données que contient cette table. Comme nous l'avons dit plus haut, cette base de données contient des informations sur les conducteurs désignés au contrat, à savoir leur sexe, leur âge, la date de passage de leur permis, leur lien avec le conducteur principal, leur position dans la désignation des conducteurs (1^{er} , $2^{ème}$ ou $3^{ème}$ conducteur), mais également de nombreuses informations concernant les cas spécifiques tels que les retraits de permis pour alcoolisme ou autre.

Un premier contrôle doit donc s'effectuer autour de la numérotation de la désignation des conducteurs car dans notre exemple, seul trois conducteurs peuvent être inscrits au contrat. La cohérence entre les dates de naissance et de passage du permis doit également être vérifiée. Nous allons donc nous appuyer sur différentes variables, à savoir un identifiant de famille qui est en réalité unique pour chaque conducteur, la date de naissance du conducteur, le numéro du conducteur (1, 2 ou 3 selon sa position dans la désignation des conducteurs) et enfin la date de passage du permis.

Ce contrôle conduit dans le cas présent à la détection d'un nombre important d'incohérences. En effet, un premier constat est alors que la date de naissance et la date de passage du permis peuvent varier pour un même conducteur désigné sur différents avenants. Une seconde incohérence provient de la numérotation dans la désignation du conducteur, qui révèle la présence de conducteurs en réalité non désignés au contrat d'une part, et de conducteurs désignés plusieurs fois sur un même contrat d'autre part. Contrairement aux incohérences détectées dans la table des sinistres, et suite à une discussion avec les responsables informatiques et un contrôle sur des contrats papiers et informatisés, il s'avère que ces erreurs ne sont pas rectifiables, et qu'aucune logique permettant de corriger ces incohérences n'apparait. De plus, bien que cette erreur ne soit pas significative en termes de pourcentage relatif à la totalité de la base (3% de taux d'erreur), elle s'avère assez importante en termes de nombre de lignes, et donc en termes de nombre d'avenants concernés. De plus, la détection d'erreur concernant les dates de naissance et de passage de permis est effectuée sur la base des avenants possédant une garantie en 2008. Ainsi, une erreur dont la source serait antérieure à 2008 ne saurait être détectée. L'exemple donné ci-après illustre ce point ;

<u>Exemple</u> :

Madame X souscrit un contrat en 2006 chez « Mutant Assurances », elle a passé son permis à 39 ans. Le périmètre étant limité à l'année d'exercice 2008, la ligne correspondant à cette conductrice et cet avenant n'apparait pas dans nos tables. Un avenant au contrat est crée en 2006, et sans réelle explication, l'âge de passage du permis est changé et passe à 82 ans. Le contrat est reconduit annuellement et reste ensuite inchangé durant les années suivantes. La base de données contient alors un renseignement sur une personne ayant passé son permis à l'âge de 82 ans, sans possibilité simple et directe de détecter le fait que cette information est erronée.

Le nombre d'erreurs sous jacent devient donc potentiellement important, et difficilement contrôlable, car il faudrait alors effectuer une extraction volumineuse et pas forcément suffisante pour détecter la totalité des incohérences. En admettant que le taux de changements inexpliqués de la date de naissance ou de passage du permis soit constant depuis ces 5 dernières années, le taux d'erreur pourrait se porter à 15 %, ce qui est important, d'autant plus que la totalité des incohérences ne saurait être détectée ici pour des raisons matérielles (extraction volumineuse). Ainsi et malgré les informations importantes contenues dans cette base de données, telles que l'âge du conducteur ou le recensement des cas spécifiques d'alcoolisme ou autres, cette table ne pourra être convenablement utilisée lors de l'étude, et il sera préférable de ne pas en tenir compte, afin de ne pas émettre des conclusions qui s'avèreraient non pertinentes.

Nous avons ainsi, à travers le contrôle de ces trois tables distinctes, envisagé différentes possibilités quant à la gestion des erreurs et à la qualité des données. La table des contrats ne révèle pas d'erreurs assez significatives pour faire l'objet d'une vive recommandation pour d'éventuelles évolutions informatiques. Elle peut en revanche faire l'objet d'une réclamation pour la création de variables représentant la date de fin ou l'exposition réelle au risque. Le contrôle de la deuxième table concernant les sinistres révèle la présence d'une incohérence qui peut être corrigée. La correction est bien évidemment faite pour les besoins de l'étude, mais il sera également nécessaire de signaler l'erreur et la méthode de résolution, afin de la rectifier pour les années à venir et ainsi augmenter la qualité des bases de données. Enfin, la table des conducteurs nous révèle la présence d'une erreur difficilement quantifiable mais potentiellement importante, et donc non négligeable. Un arbitrage en discussion avec des collaborateurs doit alors être effectué, et nous amène ici à retirer de l'étude cette table, malgré les informations intéressantes qu'elle contient. Dans ce dernier cas, une recommandation ou un avertissement doit être communiqué, soit aux supérieurs hiérarchiques, soit aux personnes en charges de la gestion informatique, afin qu'il puisse être décidé des démarches à suivre pour la suppression ou la réduction de l'erreur pour les années à venir.

Ce point essentiel de contrôle des données et de communication des incohérences détectées est à souligner car l'actuaire se doit, au-delà de l'expertise technique, d'apporter des recommandations d'évolution et de participer à l'amélioration, tant en qualité qu'en contenu, des bases de données de la compagnie, première source d'information des métiers de l'assurance, et donc probablement la plus importante.

3.2 Analyses descriptives.

Nous allons à présent exposer un exemple d'analyse descriptive. Nous commencerons alors par décrire la possibilité d'une automatisation de cette étape, puis nous mettrons en avant une partie des résultats et conclusions qui découlent d'une analyse concrète fondée sur les données obtenues par l'extraction décrite plus haut.

Lors de l'extraction, la possibilité d'une automatisation n'a pas été évoquée, et pour cause, la détection d'éventuelles incohérences ne peut se faire dans ce cas que par une action humaine. Une automatisation devient cependant possible pour une analyse descriptive qui vise à exposer des résultats généraux et à caractériser les individus et sinistres composant le portefeuille. Cette automatisation peut alors s'effectuer en utilisant conjointement les logiciels ACCESS, EXCEL et R. En effet, bien que nous ne l'ayons pas signalé auparavant, le logiciel ACCESS se prête parfaitement à la mise en place de l'extraction. Le code utilisé est alors un code « SQL », et le pilotage d'ACCESS dans le but d'une analyse descriptive (téléchargement de tables de données ou autre), peut s'effectuer sous R grâce au package « RODBC ». EXCEL et le langage « vba » sont quant à eux propices à la mise en place d'une automatisation. De plus, la création de tableaux et plus particulièrement de tableaux croisés dynamiques et de graphiques est facilité sur EXCEL et très utile pour une analyse descriptive. Le logiciel R peut être quant à lui utilisé pour importer et exploiter les tables présentes sous ACCESS, les retravailler rapidement et les renvoyer sous la forme désirée sous EXCEL pour ensuite les mettre en forme. De plus, l'établissement du lien entre le logiciel R et EXCEL est facilité par un package nommé « RExcel ». Il sera alors possible d'effectuer une programmation codée en « vba » sous EXCEL qui pourra faire appel au logiciel R, qui pourra lui-même faire appel au code « SQL » et à ACCESS. En raison du volume de cette programmation et par souci de clarté, nous n'exposerons pas le code de programmation écrit à cet effet ni la totalité des interfaces d'interaction avec l'utilisateur. Nous expliciterons cependant les possibilités qu'offre une telle programmation ainsi que les principaux résultats qu'il est possible d'obtenir.

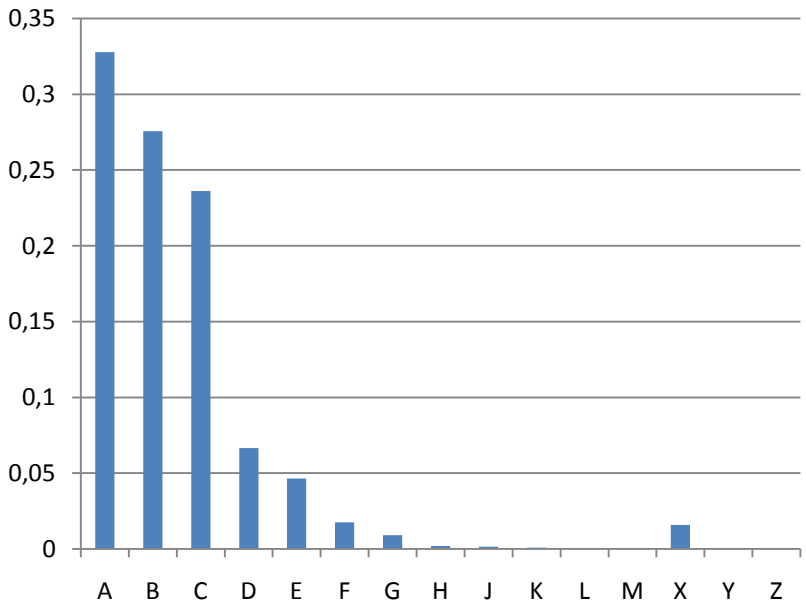
Nous allons commencer par décrire les enjeux et principes de cette programmation. L'outil informatique qui a été construit est composé d'un grand nombre d'interfaces avec l'utilisateur, qui pourra alors télécharger les tables après extraction, obtenir des informations générales sur les données, demander des analyses sur les variables explicatives soit pour les contrats en portefeuille, soit pour les individus sinistrés, et enfin effectuer des analyses croisées de deux variables de son choix. Bien entendu, un tel programme ne se suffit pas à lui-même et l'utilisateur se doit de demander les analyses pertinentes pour ensuite interpréter les résultats obtenus et les mettre en forme dans un rapport complet. Cette automatisation permettra cependant de réaliser la présente étude sur de nouvelles tables (pour l'année d'exercice 2009 par exemple) et représentera alors un gain de temps considérable. Nous ne détaillerons pas l'étape de téléchargement des tables mais nous allons revenir sur les autres points, tout en exposant des résultats et en explicitant leur intérêt pour l'analyse et la future modélisation.

Les premiers résultats obtenus concernent ce que l'on pourra appeler des chiffres clés, qui permettront de se placer dans le contexte de l'entreprise étudiée et de préciser le cadre de l'étude. Nous cherchons ainsi à obtenir le nombre de clients concernés, le nombre d'avenants, l'exposition au risque et sa répartition au sein du portefeuille, le nombre de sinistres, leur répartition entre sinistres matériels et sinistres corporels ou encore la répartition des montants de sinistres.

La deuxième étape consiste à obtenir des informations purement descriptives sur le portefeuille de contrats respectant le périmètre. Nous cherchons ainsi à obtenir des tableaux croisés dynamiques représentant à titre d'exemple les caractéristiques du véhicule telles que la valeur à neuf, la puissance, la marque ou l'année de première mise en circulation. Ceci nous permet alors de connaitre la répartition de l'exposition réelle au risque entre les différentes occurrences des caractéristiques. Le tableau et le graphique représentés ci-dessous donnent un exemple concret du résultat qui peut être obtenu autour de la variable « Classe », qui représente la valeur à neuf du véhicule et qui est ici codifiée de A à Z, A représentant un véhicule associé à une valeur à neuf faible.

Classe	Exposition
A	32,78%
B	27,56%
C	23,62%
D	6,66%
E	4,64%
F	1,75%
G	0,91%
H	0,19%
J	0,16%
K	0,08%
L	0,03%
M	0,03%
X	1,59%
Y	0,00%
Z	0,00%
Total général	**100,00%**

Illustration 23 : Exposition en fonction de Classe



Le tableau et le graphique représentés ci-dessus nous permettent ainsi de conclure que la population des assurés de « Mutant Assurances » possède de manière générale un véhicule de puissance faible (A, B ou C). Ce graphique nous indique également que lors de la modélisation, un regroupement des classes supérieures à D ou à F devrait probablement être mis en place. Cette étude effectuée au niveau de chaque variable (et que nous ne représenterons pas pour chaque variable), nous permet alors de décrire un ou plusieurs individus types présents chez « Mutant Assurances ». Ceci nous permet également d'effectuer une comparaison avec le cœur de cible de la compagnie et de voir s'il est respecté. Ceci permettra enfin par la suite de tenir compte de ces notions lors de la modélisation du risque. Dans notre exemple, l'étude révèle que l'assuré moyen possède un contrat en responsabilité civile pour une prime allant de 20 à 40 euros et bien souvent sans franchise. Une distinction s'effectue alors entre les conducteurs confirmés possédant un faible coefficient bonus malus et les jeunes conducteurs ayant un coefficient variant entre 0,8 et 1. L'analyse nous révèle également que l'assuré est généralement sans antécédent de sinistre et âgé de plus de 21 ans, bien que la population des moins de 21 ans reste non négligeable. L'étude démontre enfin que l'assuré moyen possède un véhicule de puissance moyenne et de faible valeur à neuf, généralement de marque Renault ou Peugeot et dont la date de première mise en circulation remonte entre 8 et 14 ans. En revanche, la zone et l'usage du véhicule sont variables et ne se distinguent pas particulièrement.

Nous reviendront plus loin dans l'étude sur ces résultats qui représentent une première information à ne pas perdre de vue lors de l'analyse factorielle et de la modélisation.

Une troisième étape consiste ensuite à obtenir des informations sur les sinistres, et plus particulièrement sur la population sinistrée et ses caractéristiques, afin de commencer à entrevoir d'éventuels segments sur ou sous tarifés. Le principe de programmation est alors le même car les variables explicatives restent inchangées. En revanche, alors que l'on étudiait la répartition de l'exposition entre les différentes occurrences des variables explicatives, il convient ici de s'intéresser non seulement à l'exposition réelle au risque en nombre de jours divisés par 360 (appelé aussi nombre de contrats acquis), mais également au nombre de sinistres, à la fréquence de sinistre, à la charge totale et moyenne, à la prime totale et moyenne et enfin à un rapport sinistres sur primes noté S/P. Il est cependant à noter que dans le cadre de notre étude, la prime pure n'est pas accessible directement, et il s'agit donc d'une prime chargée. Le rapport S/P n'est donc pas le rapport conventionnel et ne pourra être utilisé comme tel, il ne prendra ici que le rôle d'un outil de comparaison entre les occurrences des différentes variables. Mais revenons-en à l'automatisation, qui aura ici pour but d'obtenir des tableaux, qu'ils soient croisés dynamiques ou non, représentant la répartition des différents indicateurs cités plus haut entre les occurrences de chaque variable.

Le tableau ci-dessous est un exemple concret de résultat qui peut être obtenu suite à cette étape. Pour chaque colonne, des mises en forme conditionnelles sur le caractère ou la couleur pour rendre plus visible les valeurs extrêmes peuvent éventuellement permettre d'en faciliter la lecture. Le tableau suivant nous permet de constater en toute logique que les individus ayant un coefficient bonus malus bas possèdent une fréquence plus faible que la moyenne mais pas forcément une charge moyenne plus faible. La prime moyenne est logiquement décroissante avec le CRM. Le détail de l'interprétation de ce tableau ne sera pas inscrit ici, car ce dernier à principalement pour but la visualisation des types de résultats qui peuvent être obtenus.

CRM	Contrats acquis	Nombre Sinistres	Freq	Charge totale	Total Prime	S/P	Charge moyenne	Prime moyenne
<0,5	**0,05**	-	-	-	**18,15**	-	-	**384,36**
[0,5-0,6[	**20 356,86**	2 620	**12,87%**	2 618 502,14	7 455 224,36	0,35	**999,43**	**366,23**
[0,6-0,7[	13 210,90	1 875	**14,19%**	2 782 775,02	6 060 894,52	**0,46**	1 484,15	458,78
[0,7-0,8[	13 248,77	1 939	14,64%	2 388 421,92	6 561 104,15	0,36	**1 231,78**	495,22
[0,8-0,9[	17 672,45	2 838	16,06%	4 199 316,62	9 247 019,46	0,45	1 479,67	523,25
[0,9-1[	**22 264,41**	**4 057**	18,22%	**5 710 961,82**	**13 828 748,28**	0,41	1 407,68	621,11
[1-1,1[	13 807,85	**2 900**	21,00%	**4 278 235,32**	**11 299 192,50**	0,38	1 475,25	818,32
[1,1-1,2[	3 871,23	846	**21,85%**	1 114 529,31	3 386 731,45	0,33	1 317,41	874,85
[1,2-1,3[	1 534,33	339	**22,09%**	1 743 625,91	1 549 153,82	**1,13**	**5 143,44**	1 009,66
[1,3-1,4[	**634,94**	**134**	21,10%	**245 990,05**	**655 523,84**	0,38	1 835,75	1 032,42
[1,4-1,5[	835,46	170	20,35%	**253 784,15**	912 044,97	**0,28**	1 492,85	**1 091,67**
>=1,5	719,94	**152**	21,11%	291 215,41	907 179,61	**0,32**	**1 915,89**	**1 260,08**
Total général	**108 157,17**	**17 870**	**16,52%**	**25 627 357,67**	**61 862 835,09**	**0,41**	**1 434,10**	**571,97**

**On rappelle que le S/P est ici un simple outil de comparaison car il porte sur la Prime IARD chargée*

Illustration 24

Les principales conclusions de l'analyse dans notre cas nous amènent à penser que le zonier en vigueur mériterait d'être reconsidéré, que la variable consacrée à l'usage du véhicule, bien qu'intéressante, est peu usitée, que les véhicules de « classe » B et C semblent être sous tarifés tandis que la classe A est sur tarifé, que la fréquence de sinistre est significativement plus élevée pour les véhicules récents que pour les vieux véhicules et enfin que les contrats tous risques souffrent d'une fréquence très élevée impliquant un mauvais rapport S/P. Tous ces résultats donnent des indications qui représentent une information en elle-même et des intuitions de modélisation qui seront réutilisées par la suite (Retrait de certaines variables, regroupement de certaines occurrences et autres).

Une dernière fonctionnalité, qui sera un préliminaire de l'analyse factorielle, est d'effectuer des analyses croisées de deux variables, dans le but d'étudier une éventuelle corrélation. Nous exposerons ci-dessous le cas particulier de l'analyse croisée de la classe et du groupe du véhicule, analysant la répartition de l'exposition au risque entre les différentes occurrences. Il est à noter que ces variables ont fait l'objet d'un exemple d'analyse factorielle des correspondances en « 2.2.4.2 Exemple pratique ». La programmation renvoie en réalité ici un simple tableau avec une mise en forme automatique (voir ci-dessous). La courbe colorée représente bien le lien entre les deux variables. En effet, la classe A semble fortement associée aux groupes 9 et 10, puis la classe B au groupe 11, la classe C aux groupes 11 et 12, et ainsi de suite. Ce tableau nous permet également de détecter la présence importante de classes X et groupes 99. Hors, un retour aux données nous permet de constater que cette particularité représente les contrats suspendus, qui possèdent pourtant une exposition nulle et ne devrait donc pas apparaitre. Une recherche plus poussée nous permet alors de détecter une incohérence informatique. En effet, lorsqu'un contrat est suspendu, il est automatiquement assimilé au groupe 99 et à la classe X, mais si ce contrat est réhabilité, les variables classes et groupes ne sont pas mises à jour. Il sera donc nécessaire de supprimer les avenants concernés ou de les faire apparaitre en valeurs manquantes, puis une recommandation devra être établie afin que cette erreur ne persiste pas dans les bases de données.

Groupe / Classe	A	B	C	D	E	F	G	H	J	K	L	M	X	Y	Z	Total général
5	579	0	0	0	0	0	0	0	0	0	0	0	73	0	1	653
7	6 132	41	0	0	0	0	0	0	0	0	0	0	3	0	0	6 176
8	3 162	59	12	0	0	0	0	0	0	0	0	0	1	1	0	3 235
9	10 008	2 337	37	0	0	0	0	0	0	0	0	0	13	0	0	12 395
10	10 395	7 826	680	0	0	0	0	0	0	0	0	0	12	0	1	18 914
11	4 378	14 397	9 455	227	117	0	0	0	0	0	0	0	31	0	0	28 605
12	775	4 264	9 450	1 886	645	85	7	1	0	0	0	0	43	0	0	17 156
13	23	784	5 326	3 363	2 025	424	138	3	12	1	0	0	42	0	0	12 141
14	1	98	567	1 437	1 822	877	458	29	29	7	1	0	128	0	0	5 455
15	0	0	19	291	325	386	281	109	58	19	5	1	34	0	0	1 527
16	0	0	2	2	80	116	92	63	61	49	11	17	15	0	0	508
17	0	0	0	0	5	2	6	3	5	6	10	5	4	0	0	47
18	0	0	0	0	0	1	0	3	5	3	5	12	21	0	0	48
99	0	0	0	0	0	0	0	0	0	0	0	0	1 297	0	0	1 297
Total général	**35 452**	**29 806**	**25 549**	**7 207**	**5 019**	**1 890**	**983**	**210**	**169**	**85**	**32**	**34**	**1 718**	**1**	**2**	**108 157**

Illustration 25

Nous noterons enfin que dans le but de laisser une certaine liberté à l'utilisateur du programme, il peut être bon de ne pas mettre uniquement en place des analyses prédéfinies. Dans le cas présent, l'utilisateur du programme possède à tout moment la possibilité de demander une analyse croisée de deux variables de son choix (comme l'exemple ci-dessus avec la classe et le groupe) ou de demander un tableau récapitulatif d'une variable de son choix (comme l'exemple précédent sur le CRM). Le lecteur intéressé pourra trouver en Annexe 6 des exemples d'interfaces avec l'utilisateur qui ont été programmées dans le cadre de l'étude.

Nous conclurons alors que l'analyse descriptive, bien que d'apparence simpliste, est porteuse d'une grande quantité d'information. Dans le cadre de notre exemple, elle nous permet entre autre d'envisager et de justifier des regroupements d'occurrences de certaines variables explicatives (par exemple regrouper tous les groupes de véhicules supérieurs à D), de mieux comprendre et caractériser le portefeuille qui va faire l'objet de l'étude, de détecter éventuellement de nouvelles incohérences qui seront amenées à être corrigées et également de commencer à entrevoir des segments de populations plus ou moins à risques et éventuellement sur ou sous tarifées. Nous avons entrevu une infime partie d'un exemple d'automatisation de l'analyse descriptive. Bien entendu, les idées et principes mis en avant ici ne sont pas figés, et les solutions d'automatisation sont extrêmement vastes. Cependant, la logique des étapes décrites plus haut et les résultats obtenus resteront sensiblement les mêmes d'une automatisation à l'autre.

3.3 Analyses factorielles.

Plus en avant dans ce document, nous nous sommes intéressés en détail à l'analyse factorielle. Nous avons alors, dans le but d'illustrer la théorie, exposé un exemple au sujet de l'analyse en composante principale ou ACP, reprenant l'étude des départements en fonction de différentes variables. Puis nous avons analysé la corrélation entre le groupe et la classe du véhicule dans le but d'illustrer le fonctionnement de l'analyse factorielle des correspondances ou AFC. Enfin nous avons mis en avant une étude de certaines variables qualitatives pour illustrer l'analyse en composantes multi variées ou ACM. Chacun des exemples choisis pour illustrer les principes théoriques de l'analyse factorielle le fut pour de bonnes raisons. En effet, le premier d'entre eux abordait la notion de zonier. Hors nous avons vu dans le paragraphe précédent que la notion de zonier mériterait d'être abordée plus en détails étant donné une analyse descriptive décevante à ce sujet. Le second exemple nous a permis de préciser la relation existant entre la classe et le groupe du véhicule, qui représentent deux variables importantes de l'étude. Et enfin le dernier exemple représente en réalité les premières conclusions de l'analyse factorielle des données mixtes qui constitue en réalité l'analyse globale.

Nous allons représenter ici une partie de cette étude, la totalité de l'analyse se révélant beaucoup plus volumineuse. Mais nous commencerons par revenir sur l'étude des départements, que nous allons cette fois-ci confronter au zonier en vigueur chez « Mutant Assurances ».

3.3.1 Confrontation du zonier et de l'analyse des départements.

Nous avons pu voir précédemment que certains départements se distinguaient particulièrement en termes de risque automobile. Pour rappel, le département des Bouches du Rhône est apparu comme possédant une charge moyenne importante. En parallèle, la région d'île de France et le département du Rhône se distinguaient comme des régions à forte fréquence de sinistre. Nous avons également évoqué le fait qu'il serait bon de mettre en place un écrêtage du montant des sinistres, afin de ne pas conclure trop vite à la dangerosité d'un département. Cependant, nous n'avons à aucun moment confronté les résultats à la réalité du zonier, et il se trouve que l'exemple portant sur l'analyse en composante multi variées a mis en évidence une distinction significative des départements des Bouches du Rhône et des Alpes Maritimes. Ce dernier constat en légère contradiction avec l'étude des départements nous amène logiquement à penser qu'une comparaison entre le zonier réel et l'étude des départements s'impose. Nous savons que la variable « zone » est renseignée pour chaque avenant, et que cette dernière prend des valeurs allant de 2 à 9 (hors valeurs manquantes). Ces valeurs correspondent à une codification informatique de leur signification réelle. Après une légère modification et le retrait de la zone 7 spécifique aux DOM-TOM, il est possible d'ordonner les valeurs prises par la zone comme suit ; 2 / 4 / 5 / 5,5 / 6. Bien entendu, cette variable est qualitative, mais nous admettrons, afin de faire ressortir des évidences, que cette variable peut être assimilée à une variable quantitative. Il est tout de même à noter qu'admettre ceci n'est pas dénué de sens étant donné le fait que la variable est ordonnée (2 représente une zone de moindre risque et 6 une zone à risque élevée). Il nous sera alors possible de calculer une zone moyenne fictive pour chaque département. Nous représentons ainsi une carte de France où les départements sont coloriés en niveaux de gris en fonction de la zone moyenne fictive grâce à la fonction « *area.plot* », puis nous rajoutons des carrés (avec la fonction « s.value ») dont la taille et la couleur dépendent d'une valeur représentant le risque automobile calculée grâce à l'ACP des données écrêtées (on ne précisera pas ici la valeur choisie pour l'écrêtage).

La valeur utilisée pour la représentation des carrés est calculée de manière arbitraire. Afin de bien comprendre le choix du calcul de cette valeur, effectuons un bref rappel. Nous avons vu plus haut que le risque automobile pouvait être représenté par les deux premiers axes obtenus par ACP. Nous avons également entrevu la possibilité de placer les départements sur ce premier plan factoriel, tout en tenant compte de la qualité de représentation. Cette représentation, combinée avec l'interprétation des deux premiers axes factoriels, nous avait alors conduit à la conclusion et aux tracés suivants ;

Rappel :

« On pourra dire que les départements situés dans le carré supérieur droit tels que le Val-d'Oise, Seine Saint-Denis, l'Essonne ou encore le Rhône possèdent à la fois une charge moyenne élevée et une fréquence importante. A l'inverse, les départements du carré inférieur gauche tels que les Côtes d'Armor, le Morbihan ou l'Aube ont une faible fréquence et une charge moyenne faible. »

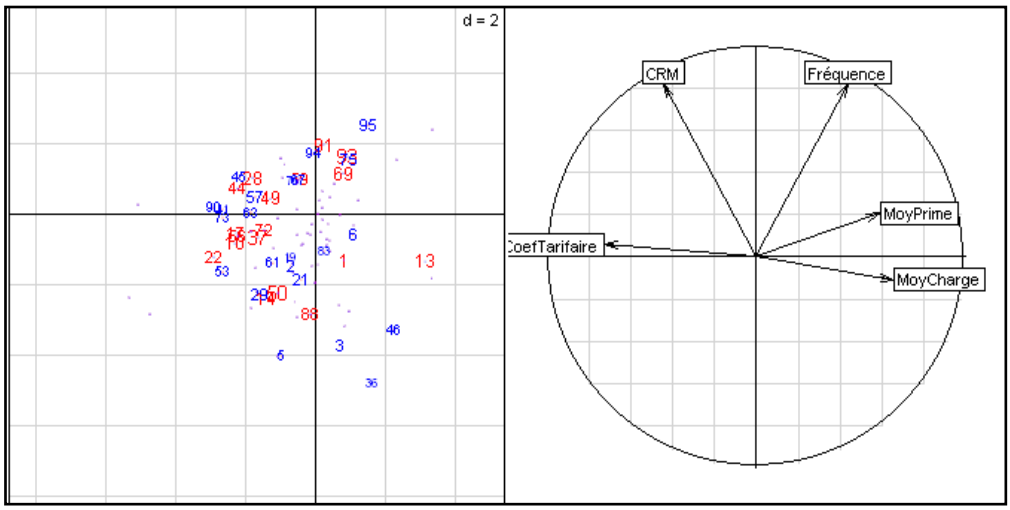


Illustration 26 : Représentation des départements sur le premier plan factoriel et interprétation des axes (exemple d'ACP)

Suite à cette conclusion, nous décidons ainsi arbitrairement de prendre comme valeur de référence la somme des coordonnées sur le premier et le second axe, multipliées respectivement par la qualité de représentation des départements sur le premier et le second axe. Un département associé à une valeur élevée sera alors bien représenté dans le carré supérieur gauche, et sera donc un département à risque, inversement, un département possédant une valeur faible sera une zone à faible risque. Le seul inconvénient de cette méthode est que tout département dont la qualité de représentation sur les axes est faible sera automatiquement rapproché d'une valeur médiane nulle (le centre du plan). Nous serons donc amener à interpréter les départements extrêmes et uniquement ceux là. Nous inscrivons ci-dessous le code R (sans revenir sur l'ACP notée « *acp0* ») ainsi que la carte de France obtenu finalement et les conclusions que nous en tirons.

```
Taille1=inertia.dudi(acp0,row.inertia=TRUE)$row.rel[,1]/1000
Taille2=inertia.dudi(acp0,row.inertia=TRUE)$row.rel[,2]/1000
valeur=acp0$li[,1]*abs(Taille1)+acp0$li[,2]*abs(Taille2)
```

```
x11()
area.plot(elec88$area,val=MoyZone,clegend=0.75)
s.value(elec88$xy,valeur,xax=1,yax=2,area=elec88$area,add.plot=TRUE,meth="squaresize",clegend=0)
```

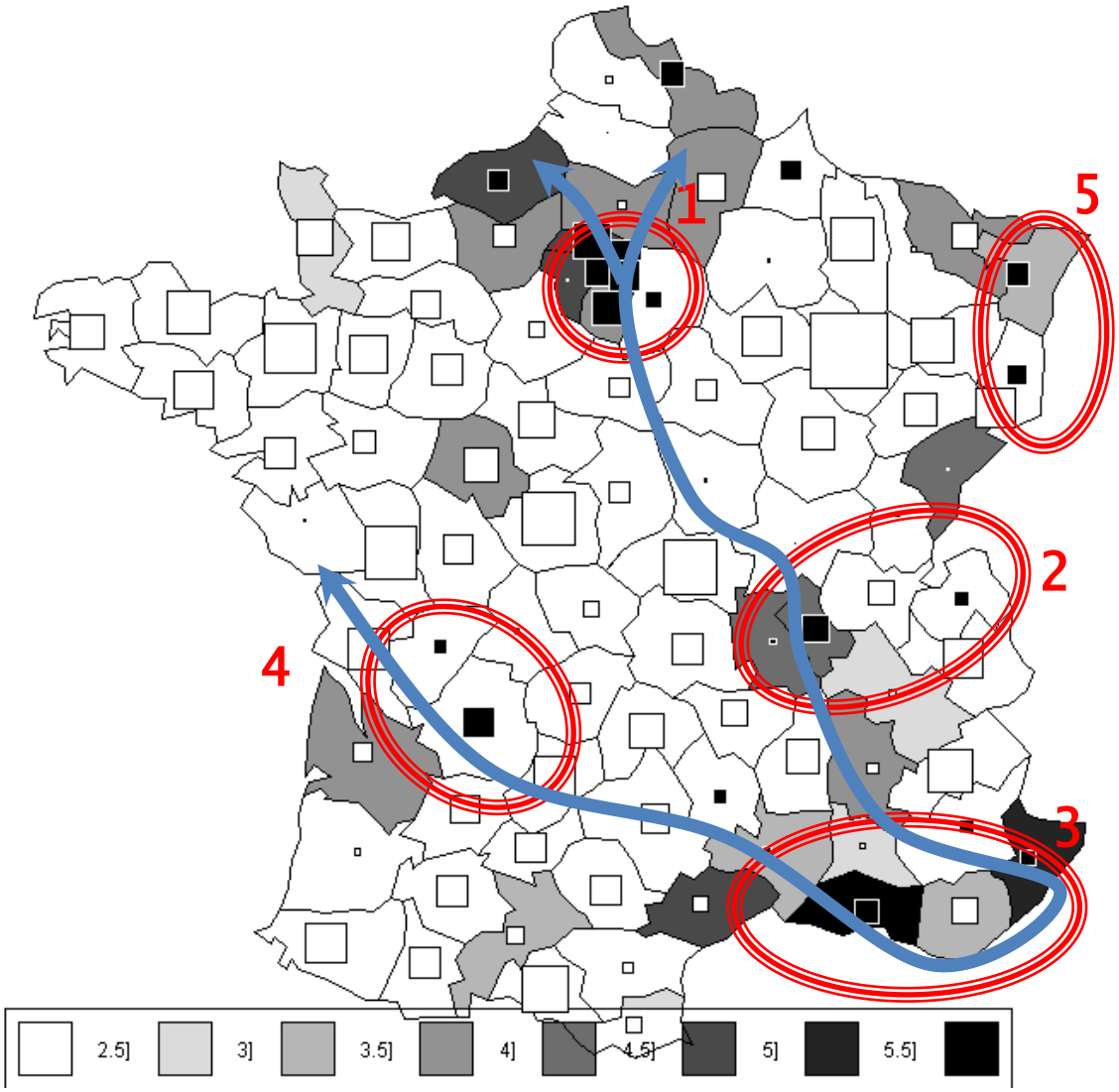


Illustration 27 : Carte de France récapitulative

- Le fond de carte symbolise le zonier, les carrés représentent le risque automobile.
- On distingue cinq zones où le risque automobile est accru, la région parisienne ou nord parisienne, la région Rhône Alpes, les côtes méditerranéennes, les départements Charente et Dordogne et enfin les départements du Bas Rhin et du Haut Rhin.
- Le zonier actuel prend bien en compte la troisième zone ainsi que les deux premières dans une moindre mesure, mais pas du tout la quatrième et peu la cinquième.
- Il est possible d'imaginer une courbe passant par l'ensemble des zones à risques (excepté par le Bas Rhin et le Haut Rhin), partant du Nord de Paris, redescendant jusque dans le Rhône, passant par la vallée du Rhône jusqu'aux côtes méditerranéennes puis remontant en direction de la Vendée.
- La Bretagne, les départements aux alentours de la Haute-Marne et le Sud-Ouest sont des zones où le risque est plus faible.
- Concrètement, suite à cette étude, les zones des départements du Rhône et de la Loire (très importantes en termes de nombre d'assurés) ainsi que de la région parisienne se sont révélées contenir un certain nombre d'incohérences et ont été revues à la hausse. La justification du zonier a de plus été discutée, et une échelle plus détaillée du risque fait l'objet d'un certain intérêt.

3.3.2 *Analyse factorielle globale.*

Nous allons à présent mettre en avant une analyse factorielle de données mixtes sur l'ensemble des variables. L'étude complète étant très volumineuse, nous ne reviendront pas sur l'ensemble des résultats mais uniquement sur les plus significatifs et les plus intéressants. De plus, nous ne reviendront pas en détails sur des résultats similaires à ceux obtenus dans l'exemple consacré à l'analyse en composantes multi variées.

Nous commençons alors par réaliser une AFDM sur un tableau contenant l'ensemble des individus sinistrés caractérisés par l'ensemble des variables explicatives. Nous notons le tableau en question TabAFDM et nous représentons les noms des variables utilisées ci-dessous.

```
> names(TabAFDM)
 [1] "AnneeDepart"     "intCoefTarifaire" "fltFranchiseMin"  "fltFranchiseMax"
 [5] "fltFranchiseVol"  "fltFranchiseRC"   "fltBonusAuto"     "chaChoixTarif"
 [9] "Options"         "Usage"           "Zone"            "Classe"
[13] "Groupe"          "Département"
```

Les sept premières variables sont des variables quantitatives, les sept suivantes sont des variables qualitatives. L'année de départ représente l'année de première mise en circulation du véhicule, le coefficient tarifaire est un bonus ou malus supplémentaire, la franchise minimale et la franchise maximale correspondent à des bornes que la franchise variable ne peut pas dépasser en cas de dommage (la franchise est un pourcentage du montant du sinistre dans notre exemple). La variable *"chaChoixTarif"* représente le type de contrat souscrit (1 pour des garanties minimales, 2 pour des garanties tous risques, R pour des garanties intermédiaires). L'option est une segmentation supplémentaire mise en place chez « Mutant Assurances » qui dépend de l'ancienneté d'assurance et de l'âge du conducteur. Nous allons représenter ci-dessous le code R nécessaire pour effectuer l'AFDM de ce tableau, puis nous représenterons chaque variable et le nuage de points afin d'en tirer certaines conclusions.

```
afdm0=dudi.mix(TabAFDM)
10
par(mfrow=c(4,2))
for(i in 8:14){
s.class(afdm0$li[,1:2],fac=TabAFDM[,i],clabel=0,cstar=0,cpoint=0.5,cellipse=0)
s.class(afdm0$li[,1:2],fac=TabAFDM[,i],cstar=0,cpoint=0,
col=rep("red",times=length(levels(TabAFDM[,i]))),add.plot=TRUE)}
s.corcircle(afdm0$co[1:7,])
```

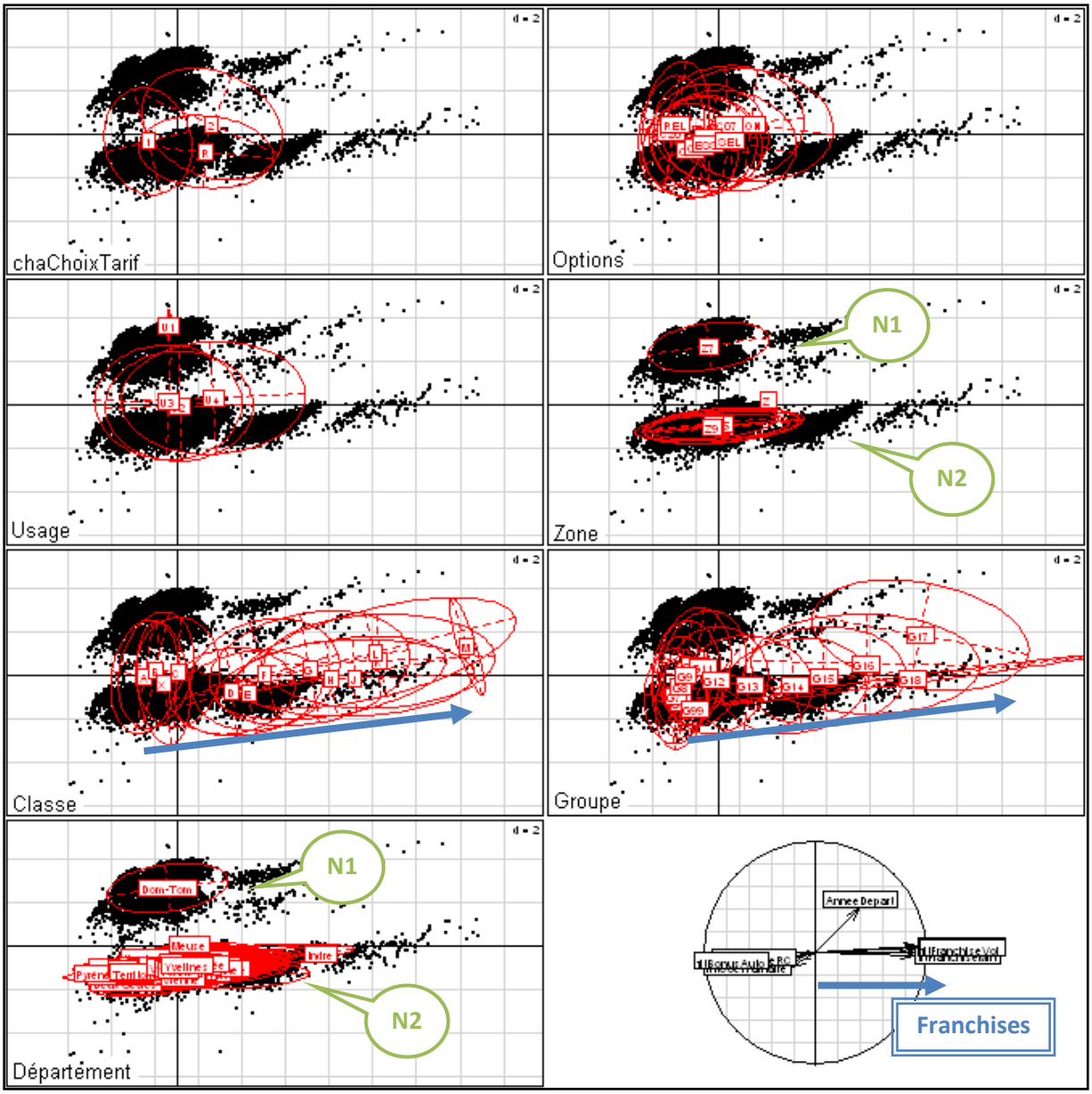


Illustration 28 : Représentation des variables d'un exemple d'AFDM

Cette première analyse factorielle nous permet d'arriver à certaines conclusions et intuitions de modélisation. En effet, nous constatons que deux nuages de points se distinguent fortement. Le premier noté N1 représente les DOM-TOM, très fortement associés à la zone Z7, tandis que le deuxième noté N2 représente les autres départements. Ceci nous invite à effectuer une modélisation du risque spécifique pour les DOM-TOM. Nous retrouvons sans grande surprise la forte corrélation entre le groupe et la classe du véhicule, mais nous constatons de plus que les franchises minimales, maximales et la franchise vol sont également liées de façon significative à ces variables. Afin de parfaire l'analyse factorielle, il sera donc nécessaire de ne sauvegarder qu'une seule de ces variables. Dans le cas présent et afin de respecter une certaine logique sous jacente à la tarification en vigueur que l'on ne détaillera pas ici, nous sauvegarderons la variable représentant la classe du véhicule. Cette dernière information devra également être prise en compte lors de la modélisation.

Par souci de clarté et d'interprétation des variables, nous décidons de façon arbitraire de regrouper par classes les variables quantitatives de manière à les transformer en variables qualitatives. En effet, dans le cas particulier du coefficient bonus malus, il sera plus simple et plus parlant de s'intéresser à des tranches de CRM que l'on pourra alors associer avec la classe du véhicule ou avec l'option du contrat par exemple pour des analyses croisées facilitées. Nous ajoutons de plus à l'analyse les variables représentant le montant du sinistre, la nature du sinistre, le nombre de sinistres de l'assuré et la prime associée à l'assuré. Sans reproduire le code (ce qui constituerait une répétition inutile), nous représentons ci-dessous 4 groupes de graphiques parmi les plus représentatifs de l'étude globale puis nous fermerons ce paragraphe en reprenant l'ensemble des conclusions et intuitions obtenues par l'analyse et à prendre en compte pour la modélisation à venir.

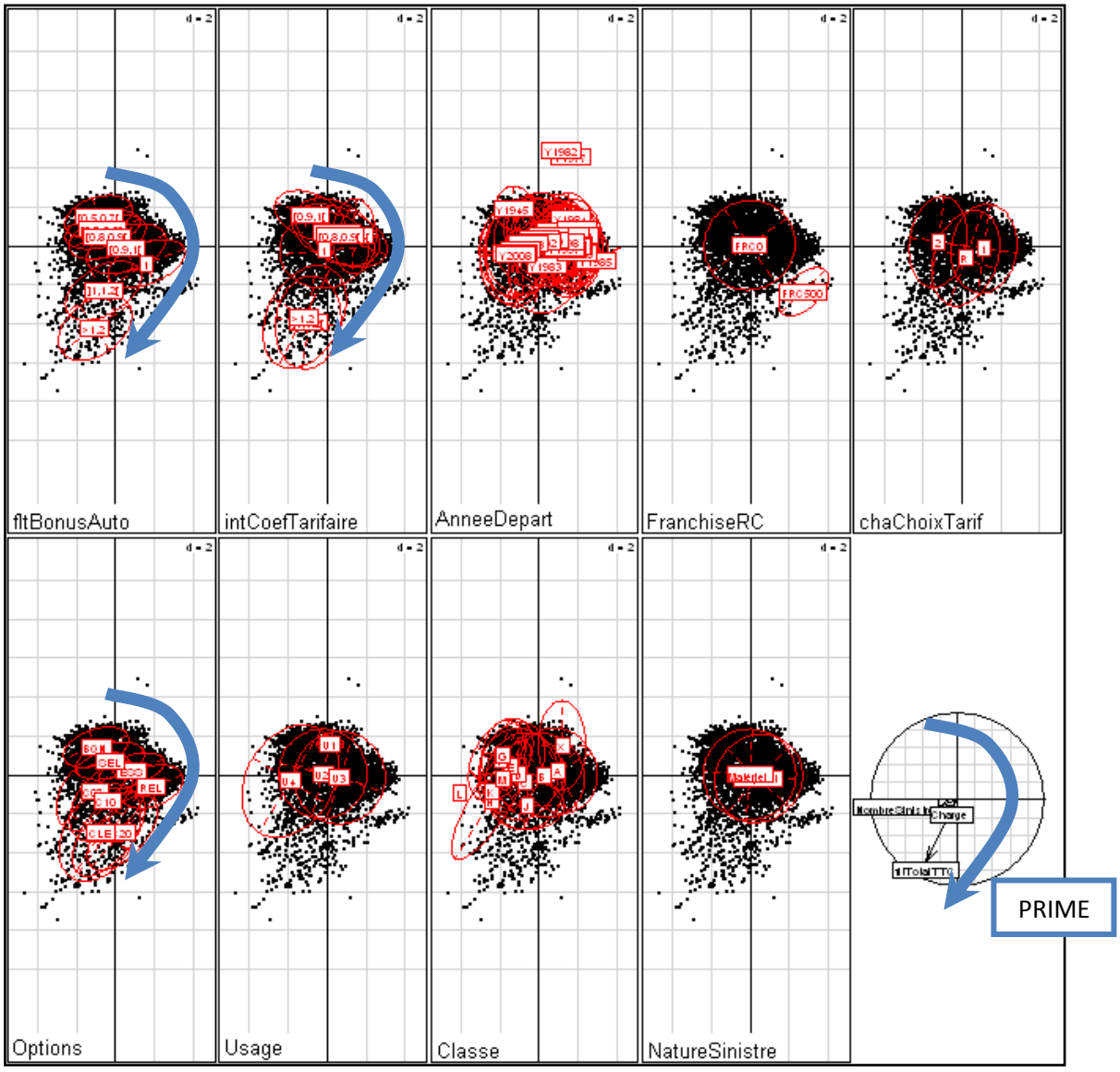


Illustration 29 : Première représentation graphique (exemple d'AFDM)

Cette première représentation indique un lien fort entre le CRM, le coefficient tarifaire, l'option et la prime (croissante avec les précédents). La significativité de ces trois variables sera donc à surveiller afin de savoir lesquelles d'entre elles sont les plus pertinentes.

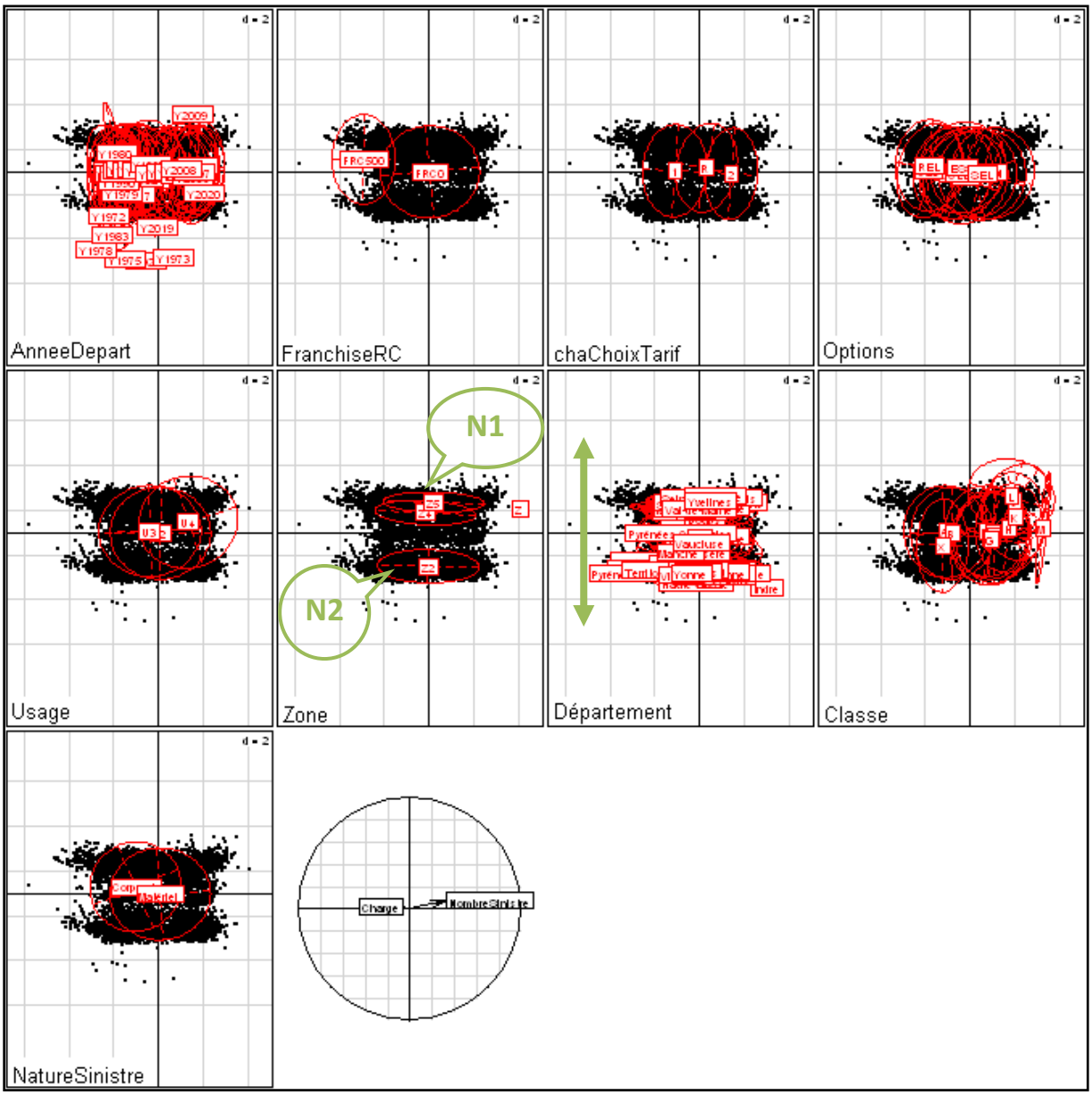


Illustration 30 : Deuxième représentation graphique (exemple d'AFDM)

La seconde représentation apporte une information sur la zone, venant compléter l'étude sur les départements. Dans cet exemple, la zone 6 et la zone 9, fortement attachées aux départements des Bouches du Rhône, des Alpes Maritimes et de la Corse ont été retirées préalablement. Nous remarquons alors une segmentation apparente entre la zone 2 et les zones 4 et 5, qui sont très proches. Une question quant à la pertinence de la zone pour l'étude du risque automobile se pose alors. Il conviendra par la suite de tenter de réaliser différentes modélisations soit avec la zone soit avec les départements (bien que cette dernière variable ne paraisse pertinente). La significativité respective de ces variables devra donc être surveillée.

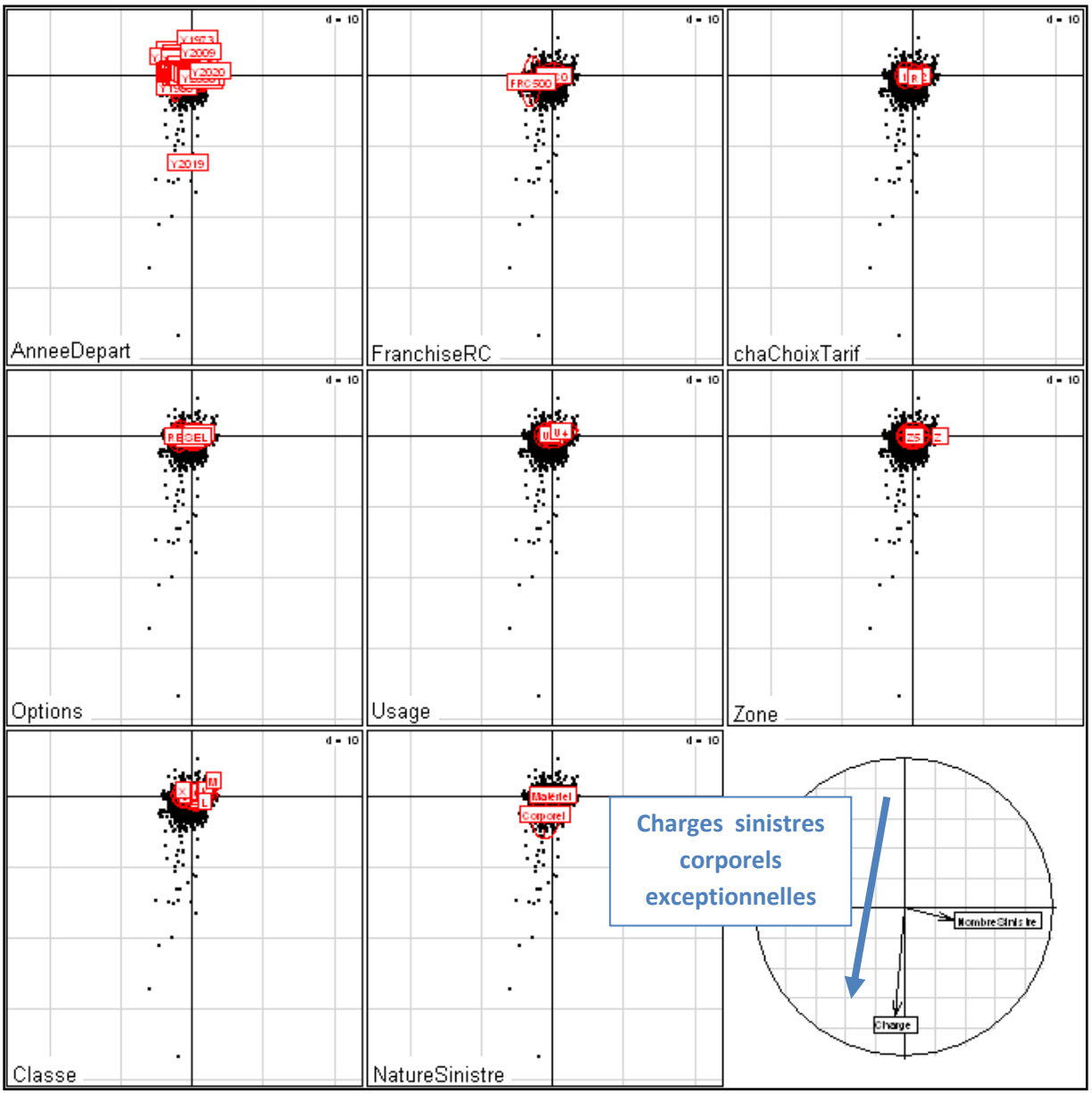


Illustration 31 : Troisième représentation graphique (exemple d'AFDM)

Cette dernière représentation, qui parait pourtant être mauvaise, nous apporte une autre indication, conforme à l'intuition. En effet, il pourra être judicieux d'opérer une distinction entre les sinistres matériels et les sinistres corporels (pour lesquels le coût est bien plus élevé). Cette distinction devra probablement être opérée en ce qui concerne la modélisation du montant des sinistres. En revanche, cette distinction peut paraitre discutable pour la modélisation de la fréquence de sinistre. Il parait de plus également nécessaire de distinguer les sinistres dits « graves », de montants exceptionnellement élevé et de fréquence très faible.

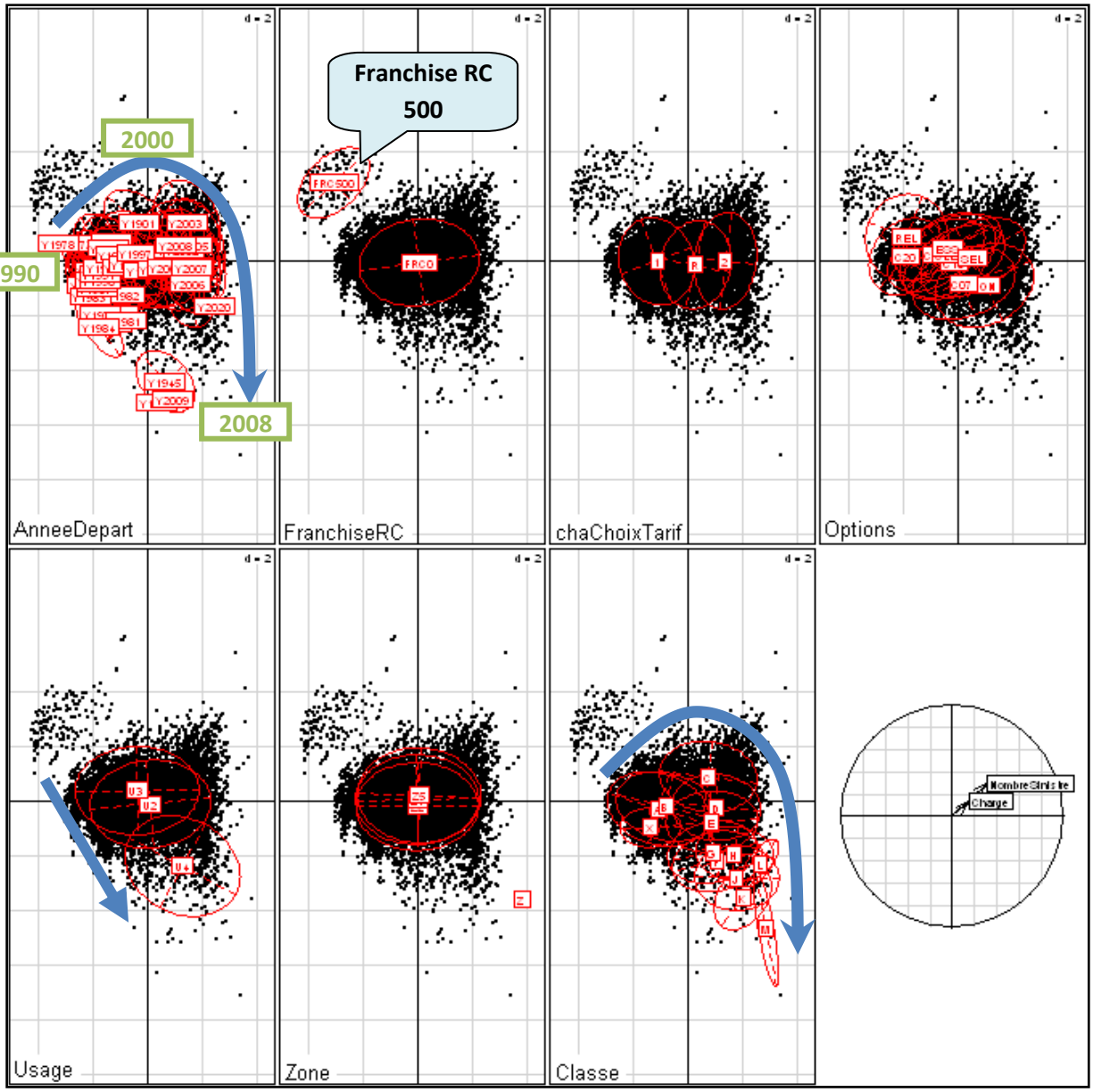


Illustration 32 : Quatrième représentation graphique (Exemple d'AFDM)

La représentation ci-dessus contient différentes informations. Nous constatons premièrement que les individus sinistrés possédant une franchise en responsabilité civile de 500 € se distinguent des autres, cette notion devra être prise en compte. Ensuite, nous remarquons que l'usage 2 et l'usage 3 restent très proches, nous avons en effet vu lors de l'exemple consacré à l'ACM que la tarification actuel n'effectuait pas de différence entre ces deux usages. L'usage 4 quant à lui, qui représente une utilisation professionnelle du véhicule, semble être associé à des véhicules plus récents, et de valeur à neuf plus élevée que la moyenne. Nous pouvons enfin remarquer le lien existant entre l'année de mise en circulation du véhicule et la classe du véhicule, ce qui semble respecter une certaine logique.

Nous allons enfin reprendre toutes les conclusions que l'analyse factorielle globale a pu nous apporter, et qui devront être prises en compte lors de la modélisation du risque automobile.

- Les variables « Franchise Min », « Franchise Max », « Franchise Vol », « Groupe » et « Classe » sont fortement corrélées. Il parait vraisemblable que la modélisation du risque amène à ne considérer comme significative qu'une ou deux de ces variables. Il serait souhaitable que la variable sauvegardée soit la classe du véhicule. Ces variables sont liées avec la prime du contrat et donc avec la charge sinistre, elles interviendront donc probablement plus clairement dans la modélisation du montant du sinistre.
- Les DOM-TOM doivent faire l'objet d'une étude spécifique. Ils sont en effet associés à une zone particulière et ne doivent pas être négligés car ils représentent environ 1/3 de la population sinistrée.
- La zone devient ainsi une variable significative importante, et ce même pour l'étude hors DOM-TOM. La Corse, les Bouches du Rhône, les Alpes Maritimes et dans une moindre mesure le Rhône se distinguent des autres départements car ils possèdent des zones 5/6 et 6. Une séparation existe également entre les zones 4, 5 et la zone 2. Cette variable devra alors être prise en compte et contrôlée, tout en effectuant un parallèle avec l'étude particulière concernant les départements. Il pourrait s'avérer intéressant d'effectuer une première modélisation prenant en compte la zone et une deuxième prenant en compte les départements, voir un maillage plus fin. Ces variables se rapportent aussi bien à la notion de fréquence qu'à la notion de montant.
- Un autre groupe de variables significatives est composé du CRM, du coefficient tarifaire et de l'option de tarification. Là encore il existe une certaine corrélation entre ces variables et la significativité de chacune d'entre elle lors de la modélisation sera à surveiller. On portera une attention particulière à l'option Relais bien souvent associée à un coefficient bonus malus de 1. Bien que cela puisse paraitre insuffisant, l'absence de la notion d'âge qui n'a pu être retenue pour l'étude pourra être palliée par l'étude des options. Un regroupement des options commençant par un C (CLE, C20, C07, C10, C12) pourra éventuellement être opéré. Ces trois variables, par construction, devraient prendre tout leur sens lors de la modélisation de la fréquence de sinistre.
- Les études précédentes semblent indiquer la nécessité de modéliser séparément les sinistres corporels et les sinistres matériels. La notion d'écrêtage, notamment pour les sinistres corporels, devra également être prise en compte. Il parait de plus justifié de distinguer les sinistres standards des sinistres graves. Seule la confrontation de plusieurs modélisations permettra d'éclaircir ce point et de déterminer la démarche à suivre.
- Malgré l'indépendance qui sera supposée entre la fréquence et le montant de sinistre, un lien subsiste entre les différentes variables explicatives. Ce lien s'opère par le type de contrat souscrit. On peut en effet conclure de façon grossière que les contrats de garantie minimale concernent des vieux véhicules, de classes faibles et d'assurés à CRM variés peu enclins à avoir des sinistres corporels. Les contrats « tout risque » concernent quand à eux des véhicules plus récents, pour des assurés à faible CRM en moyenne, mais plus enclin à avoir un grand nombre de sinistres matériels pour une charge de sinistre plus élevée.
- On conclut enfin que l'usage 4 est en règle générale associé à des véhicules de classe élevée et que la présence d'une franchise RC fixée à 500 € est une particularité apparente.

Partie III : Modélisation du risque automobile et étude des segments de tarification.

Nous avons pu entrevoir précédemment des généralités concernant l'assurance automobile ainsi que des techniques d'analyses descriptives et factorielles. Ces aspects constituent une première approche et un préliminaire à la réalisation de la modélisation du risque automobile que nous allons à présent décrire. Nous nous concentrerons tout d'abord sur l'aspect théorique des modèles linéaires généralisés, puis nous nous intéresserons plus particulièrement au cadre particulier de l'assurance automobile. Enfin nous exposerons un exemple concret de modélisation du risque automobile accompagné des conclusions que cette dernière peut apporter.

Chapitre 1 : Les modèles linéaires généralisés.

1.1 Préliminaire : La régression linéaire.

1.1.1 Fondements théoriques.

Les modèles linéaires étant une généralisation de la régression linéaire, nous allons commencer par évoquer cette dernière. L'enjeu de la modélisation est de prédire ou expliquer une variable Y, à partir de « p » variables explicatives appelées également prédicteurs et regroupées dans une variable notée $X=[X_1 , X_2 , \dots , X_p]$. Nous supposerons que l'on dispose de « n » observations des variables X et Y et nous noterons y_i et x_{ij} les $i^{ème}$ observations respectives de Y et X_j. Dans toute la suite, nous considérerons que « p » est le nombre de variables explicatives, incluant un éventuel intercept (vecteur unité composé de 1).

Dans l'hypothèse d'un lissage par régression linéaire simple, nous cherchons à modéliser Y par une droite linéaire de X (un seul prédicteur ici et un intercept, p=2) d'ordonnée à l'origine b et de coefficient directeur a, soit Y=aX+b. L'estimation des coefficients a et b s'effectue alors par la méthode des moindres carrés. Nous cherchons donc ;

$$\min_{a,b} \Delta(a,b) = \min_{a,b} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

La recherche des points a^* et b^* qui minimise $\Delta(a,b)$ passe par l'annulation des dérivés partielles. Nous notons directement les résultats obtenus que le lecteur intéressé pourra aisément retrouver à partir des notations suivantes ;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \; \sigma_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \; ; \; \sigma_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$\begin{cases} b^* = \bar{y} - a^*\bar{x} \left(La\ droite\ optimale\ passe\ par\ le\ centre\ de\ gravité\ C_g\right) \\ a^* = \dfrac{\sigma_{xy}}{\sigma_x^2} \end{cases}$$

Nous allons ensuite généraliser ce résultat en rajoutant un terme d'erreur gaussien. Nous nous plaçons alors dans le cadre de l'estimation pour $y_i=ax_i+b+\varepsilon_i$, en supposant que les ε_i sont indépendants et identiquement distribués (noté « iid » ou « vaiid » dans la suite) selon une loi normale centrée et de variance σ^2. Nous noterons $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$, ce qui implique alors $y_i \sim \mathcal{N}(ax_i+b,\sigma^2)$. L'estimation des facteurs que nous noterons $\hat{a}$, $\hat{b}$ et $\widehat{\sigma^2}$ peut alors s'effectuer à partir de l'estimation par maximum de vraisemblance notée « EMV ». Nous cherchons ainsi à maximiser $ln(\mathcal{L}_y)$ avec ;

$$\ell_{y_i} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y_i-(ax_i+b)]^2}{2\sigma^2}} \; densité\ de\ la\ loi\ \mathcal{N}(ax_i+b,\sigma^2)$$

$$\mathcal{L}_y = \prod_{i=1}^{n} \ell_{y_i} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^{n}[y_i-(ax_i+b)]^2}{2\sigma^2}} \; vraissemblance\ de\ Y$$

$$\ln(\mathcal{L}_y) = -nln(\sqrt{2\pi}) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(ax_i+b)]^2$$

La maximisation de $ln(\mathcal{L}_y)$ passe donc par la minimisation de $\sum_{i=1}^{n}[y_i-(ax_i+b)]^2$ qui correspond à ce que nous avions noté $\Delta(a,b)$. Ceci nous permet ainsi d'obtenir que ;

$$\boxed{\hat{a} = a^* \; et \; \hat{b} = b^*}$$

Il reste alors à estimer σ^2 en annulant la dérivée partielle correspondante. Nous obtenons donc ;

$$\boxed{\widehat{\sigma^2} = \frac{\Delta(a,b)}{n}}$$

Nous formalisons ensuite la présentation de la régression linéaire en présence de plusieurs prédicteurs ;

$$\boldsymbol{Y = X\theta + \varepsilon}$$

$Y\ variable$ à $expliquer$

$X = [X_1 \quad \dots \quad X_p]\ variables\ explicatives\ ou\ prédicteurs$

$\theta\ paramètres\ inconnus$ à $estimer$

$\varepsilon\ vecteur\ aléatoire\ appelé\ terme\ de\ bruit\ ou\ d'erreur$

La régression linéaire simple vue plus haut correspond alors au cas où ;

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \theta = \begin{pmatrix} b \\ a \end{pmatrix} \text{ et } (\varepsilon_i)_{1\le i\le n} \; iid\ de\ loi\ \mathcal{N}(0,\sigma^2)$$

Ce cas simple répond aux hypothèses que l'on retrouve classiquement en régression linéaire, à savoir ;

$$E(\varepsilon) = 0$$
$$Var(\varepsilon_i) = \sigma^2 (identique\ pour\ tout\ i, on\ parle\ d'homoscédasticité)$$
$$(\varepsilon_i)_{1 \le i \le n}\ indépendants\ et\ de\ loi\ normale\ \mathcal{N}(0, \sigma^2)$$

Sous ces hypothèses, nous pouvons énoncer un certain nombre de définitions et propriétés fondamentales ;

Si $\hat{\theta}$ est un estimateur de θ, alors $\hat{Y} = X\hat{\theta}$ est une estimation de la moyenne et $\hat{\varepsilon} = Y - \hat{Y}$ est appelé le vecteur des résidus.

Nous posons $SCR(\theta) = \|Y - X\theta\|^2 = (Y - X\theta)^T (Y - X\theta) = \|\varepsilon\|^2$
Alors l'estimation de θ par moindres carrés revient à minimiser $SCR(\theta)$
$Min\big(SCR(\theta)\big) = SCR\big(\hat{\theta}\big) = \|\hat{\varepsilon}\|^2$

$\hat{\theta} = (X^T X)^{-1}(X^T Y)$ et $\hat{\theta}$ est un vecteur gaussien

$E\big(\hat{\theta}\,\big) = \theta$ et *$Var\big(\hat{\theta}\,\big) = \sigma^2 (X^T X)^{-1}$*

$SCR\big(\hat{\theta}\big) \sim \sigma^2 \chi^2(n-p)$, de plus $SCR\big(\hat{\theta}\,\big)$ et $\hat{\theta}$ sont indépendants

$\widehat{\sigma^2} = \dfrac{SCR\big(\hat{\theta}\big)}{n-p}$ est un estimateur sans biais $\left[E\big(\widehat{\sigma^2}\big) = \sigma^2\right]$

$X^T X$ inversible $\Leftrightarrow rang(X) = p$

Si les termes d'erreurs sont centrés, homoscédastiques et indépendants,
Alors l'estimateur $\hat{\theta}$ des moindres carrés est l'estimateur linéaire sans biais de variance minimale

La démonstration du premier point se fait naturellement en partant de l'idée que $\hat{Y} = X\hat{\theta}$ est la projection orthogonale de Y sur l'espace des variables explicatives, ce qui implique que $\langle Y - X\hat{\theta} \big| X \rangle = 0$, ce qui démontre le résultat (Une estimation par maximum de vraisemblance ou EMV fonctionne également). Du fait que Y soit gaussien, on en déduit que $\hat{\theta}$ est un vecteur gaussien également. L'espérance et la variance en découlent directement en notant que pour A une matrice et Y une variable aléatoire, nous avons $E(A \times Y) = A \times E(Y)$ et $Var(A \times Y) = A \times Var(Y) \times A^T$. Bien entendu, il est nécessaire que la matrice $X^T X$ soit inversible, ce qui est le cas si et seulement si cette dernière est de plein rang (p), or le rang de $X^T X$ est le même que celui de X. En pratique lorsque ceci n'est pas le cas, cela signifie que certaines variables explicatives sont liées, et qu'il faut donc en retirer pour continuer l'étude. La démonstration du troisième point se base sur le théorème de Cochran, le lecteur intéressé pourra en trouver l'énoncé et la démonstration en Annexe 7. L'estimation de σ^2 découle alors directement du point précédent. Le dernier point se démontre enfin en considérant un estimateur linéaire sans biais $\ddot{\theta}$, il existe alors une matrice A telle que $\ddot{\theta} = AY$ ce qui implique $E(\ddot{\theta}) = AX\theta$ et donc $AX = I$. Il reste alors à écrire que $Var\big(\ddot{\theta}\big) = Var\big(\ddot{\theta} - \hat{\theta} + \hat{\theta}\big) = Var\big(\ddot{\theta} - \hat{\theta}\big) + Var\big(\hat{\theta}\big) - 2Cov(\ddot{\theta} - \hat{\theta}, \hat{\theta})$ et à montrer sans difficulté que le dernier terme est nul, ce qui implique que $Var\big(\ddot{\theta}\big) \geq Var\big(\hat{\theta}\big)$ et qui conclut la preuve.

Nous faisons alors remarquer que dans certains cas, il est possible d'introduire des poids aux observations. Ceci est par exemple le cas lorsque l'on possède plusieurs valeurs de Y pour des observations identiques de X, ou lorsque l'on sait que certaines observations de Y sont plus précises que d'autres. Sous cette hypothèse, la variance des résidus n'est pas constante, elle dépend des observations de Y et peut s'écrire σ^2/w_i pour la i^{ème} observation avec w_i le poids associé (ceci est le cas lorsque Y_i est la moyenne de w_i observations). Nous partons ainsi du principe que plus le poids d'une observation est grand, plus l'observation est pertinente et précise, et plus la variance de l'erreur est faible. Nous introduisons alors une matrice diagonale des poids noté $W = diag(1/w_i)$, à n lignes et n colonnes. La variance des résidus vaut alors $\sigma^2 W$ et l'ajout de cette matrice de poids implique, après estimation par la méthode du maximum de vraisemblance, la relation $\hat{\theta} = (X^T W X)^{-1}(X^T W Y)$ (non détaillé ici mais facilement réalisable en écrivant la vraisemblance et en faisant le lien avec la minimisation de l'expression $\left\| W^{\frac{1}{2}}(Y - X\theta) \right\|^2$).

Enfin, nous ferons remarquer que l'utilisation de variables explicatives qualitatives est tout à fait possible, et conduit à utiliser un codage avec des indicatrices. Ainsi, si l'on désire mettre en place un modèle avec Y la variable à expliquer, X_1 la première variable explicative (quantitative) et X_2 la deuxième variable explicative (qualitative) d'occurrences possibles X_{2a} , X_{2b} , X_{2c} , nous construirons un modèle à partir de trois prédicteurs en réalité. En effet, le premier d'entre eux sera X_1 , le deuxième sera l'indicatrice de X_{2a} et le dernier l'indicatrice de X_{2b} . L'occurrence X_{2c} sera ainsi considérée comme l'occurrence de référence, et contenu dans l'intercept s'il est présent.

1.1.2 *Estimations et intervalles de confiances.*

Les formules fondamentales précédentes permettent d'obtenir une estimation des paramètres de la régression $(\theta_i)_{1\leq i\leq p}$, de la variance de l'erreur σ^2 et des observations de Y en fonction des prédicteurs. De plus, il nous est possible de calculer des intervalles de confiances autour de ces estimations. Nous résumons ci-dessous les différents intervalles de confiances ;

Pour $(\theta_i)_{1\leq i\leq p}$;

Il est à noter que $E(\widehat{\theta_\iota}) = \theta_i$ et $\sigma^2\lambda_{ii}^2 := Var(\widehat{\theta_\iota}) = [\sigma^2(X^TX)^{-1}]_{ii}$

On obtient alors que $A1 := \frac{\widehat{\theta_\iota}-\theta_i}{\sigma\lambda_{ii}} \sim \mathcal{N}(0,1)$ et $A2 := (n-p)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-p)$

Et donc par construction, $\frac{\widehat{\theta_\iota}-\theta_i}{\hat{\sigma}\lambda_{ii}} = \frac{A1}{\sqrt{\frac{A2}{(n-p)}}} \sim Student(n-p)$

En notant $t_{n-p}^{1-\frac{\alpha}{2}}$ le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi de Student à (n-p) degrés de liberté, et par symétrie de la loi de Student, nous obtenons finalement l'intervalle de confiance suivant ;

$$IC_\alpha(\theta_i) = \left[\widehat{\theta_\iota} - \hat{\sigma}\lambda_{ii}\, t_{n-p}^{1-\frac{\alpha}{2}} \,;\, \widehat{\theta_\iota} + \hat{\sigma}\lambda_{ii}\, t_{n-p}^{1-\frac{\alpha}{2}}\right]$$

Pour σ^2;

Il suffit alors de réutiliser le fait que $(n-p)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-p)$, ce qui permet d'obtenir l'intervalle de confiance en notant $\chi^2_{n-p}(\alpha)$ le quantile d'ordre α de la loi du Khi² à (n-p) degrés de liberté ;

$$IC_\alpha(\sigma^2) = \left[\frac{(n-p)\widehat{\sigma^2}}{\chi^2_{n-p}(1-\frac{\alpha}{2})} \;;\; \frac{(n-p)\widehat{\sigma^2}}{\chi^2_{n-p}(\frac{\alpha}{2})}\right]$$

Pour $(Y_i)_{1\leq i\leq n}$;

Nous savons que $E(Y_i) = {X_i}^T\theta$, $\widehat{Y_\iota} = \widehat{E(Y_\iota)} = {X_i}^T\hat{\theta}$ et $Var(\widehat{Y_\iota}) = \sigma^2{X_i}^T(X^TX)^{-1}X_i$ et on peut donc en déduire que $\frac{\widehat{Y_\iota}-E(Y_i)}{\sqrt{Var(\widehat{Y_\iota})}} \sim \mathcal{N}(0,1)$ et nous savons de plus que $(n-p)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-p)$. Nous pouvons donc une fois de plus nous ramener à une loi de Student à (n-p) degrés de liberté et écrire l'intervalle de confiance de l'espérance de l'observation ;

$$IC_\alpha = \left[\widehat{Y_\iota} - \sqrt{\widehat{\sigma^2}{X_i}^T(X^TX)^{-1}X_i} \times t_{n-p}^{1-\frac{\alpha}{2}} \,;\, \widehat{Y_\iota} + \sqrt{\widehat{\sigma^2}{X_i}^T(X^TX)^{-1}X_i} \times t_{n-p}^{1-\frac{\alpha}{2}}\right]$$

En remarquant que $\frac{\widehat{Y_\iota}-Y_i}{\sqrt{Var(\widehat{Y_\iota}-\varepsilon_\iota)}} = \frac{(\widehat{Y_\iota}-\varepsilon_i)-E(Y_i)}{\sqrt{Var(\widehat{Y_\iota}-\varepsilon_i)}} \sim \mathcal{N}(0,1)$ avec $Var(\widehat{Y_\iota}-\varepsilon_i) = \sigma^2(1+{X_i}^T(X^TX)^{-1}X_i)$, nous pouvons construire un intervalle de confiance de Y_i selon le même principe. On parlera alors d'intervalle de prédiction ;

$$IC'_\alpha = \left[\widehat{Y_\iota} - \sqrt{\widehat{\sigma^2}(1+{X_i}^T(X^TX)^{-1}X_i)} \times t_{n-p}^{1-\frac{\alpha}{2}} \,;\, \widehat{Y_\iota} + \sqrt{\widehat{\sigma^2}(1+{X_i}^T(X^TX)^{-1}X_i)} \times t_{n-p}^{1-\frac{\alpha}{2}}\right]$$

1.1.3 *Choix du modèle et tests de significativité.*

Nous savons à présent comment effectuer une régression linéaire puis estimer les paramètres ainsi que la variable à expliquer ou son espérance. Il nous reste cependant à définir et à mettre en place des tests de validité de la modélisation mise en œuvre et des tests de significativité des variables utilisées.

Etant donné la projection orthogonale sur l'espace des variables explicatives, nous sommes en mesure de décomposer la variabilité de la variable expliquée Y en deux termes ;

$$\|Y\|^2 = \left\|X\hat{\theta}\right\|^2 + \left\|Y - X\hat{\theta}\right\|^2$$

Sous l'hypothèse de la présence de ce que l'on appelle un intercept (la première colonne de X est composée de 1), cette égalité peut se réécrire et se représenter graphiquement comme suit ;

$$\|Y - \bar{Y}\|^2 = \left\|\hat{Y} - \bar{Y}\right\|^2 + \left\|Y - \hat{Y}\right\|^2 \text{où } \bar{} \text{ désigne la moyenne}$$

$$\boldsymbol{SCT}\ (Somme\ des\ Carrés\ Totales) = \boldsymbol{SCM}(Somme\ des\ Carrés\ du\ Modèle) + \boldsymbol{SCR}\ (Somme\ des\ Carrés\ Résiduels)$$

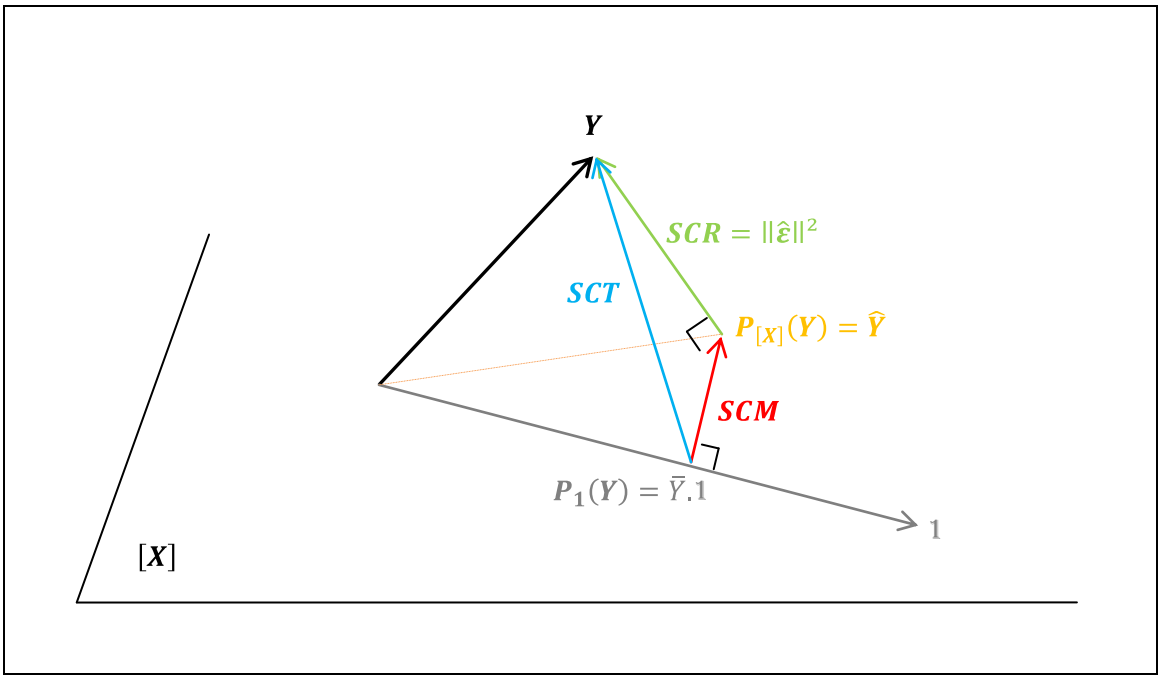


Illustration 33 : Représentation graphique des sommes de carrés dans le cas avec intercept

Cette égalité se démontre aisément en se remémorant que $\hat{\varepsilon}$ est orthogonal à l'espace $[\boldsymbol{X}]$ engendré par les variables explicatives et en utilisant le théorème de Pythagore. Nous pourront également noter que la présence d'un intercept implique la nullité de la somme des composantes de $\hat{\varepsilon}$. En effet, l'intercept appartient à l'espace engendré par les colonnes de X, et nous pouvons donc écrire;

$$\sum_{i=1}^{n} \widehat{\varepsilon_i} = \langle \hat{\varepsilon} | (1, \dots, 1) \rangle = 0$$

Un bon indicateur de la significativité du modèle peut ainsi être calculé en effectuant le rapport de la variance expliquée par la régression (SCM) sur la variance totale (SCT). Ce coefficient noté R^2 est parfois appelé coefficient de détermination.

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}$$

Nous pouvons remarquer que ce coefficient correspond alors au carré du cosinus de l'angle du vecteur $Y - \bar{Y}.1$ et le vecteur $\hat{Y} - \bar{Y}.1$, ce qui nous permet de noter que ;

$$R^2 = \frac{\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{SCT}\sqrt{SCM}} = \rho^2_{Y,\hat{Y}} \ \ avec\ \rho^2_{Y,\hat{Y}}\ le\ coefficient\ de\ corrélation\ de\ Y\ et\ \hat{Y}$$

Nous considèrerons alors que le modèle est significatif lorsque la valeur de R^2 est proche de 1. Cependant, R^2 est un indicateur qui augmente avec le nombre de variables explicatives. En effet, si le nombre de prédicteurs augmente, l'espace engendré $[X]$ devient plus important, ce qui engendre une baisse de la valeur du SCR alors que le SCT reste inchangé, ce qui implique automatiquement une augmentation du coefficient de détermination. Ainsi, si l'on doit choisir entre un modèle à p variables et le même modèle avec une variable en moins, nous seront toujours amené à choisir le modèle contenant le plus de variables. C'est pourquoi il est préférable de se baser sur un coefficient de détermination ajusté tenant compte du nombre de variables explicatives du modèle et défini comme suit ;

$$1 - R^2_{adj} = \frac{n - intercept}{n - p}(1 - R^2)$$
$$O\grave{u}\ intercept\ vaut\ 1\ en\ présence\ d'intercept, 0\ sinon$$

Il existe d'autres critères de sélection de modèle tenant compte du nombre de prédicteurs, nous représentons ci-dessous l'un d'entre eux, souvent utilisé, l'AIC (Akaïke Information Criterion) ;

$$AIC = 2p + n \times ln(\frac{SCR}{n})$$

Nous possédons donc à présent des outils de contrôle de significativité du modèle. Il nous reste alors à déterminer un test de significativité des variables afin de pouvoir construire un modèle significatif composé de variables pertinentes. Nous supposons qu'un modèle avec p variables explicatives $Y = X\theta + \varepsilon$ a été mis en place et que nous souhaiterions savoir si un sous modèle de celui-ci, comportant p_0<p prédicteurs et représenté par l'équation $Y = X_0\theta_0 + \varepsilon$ serait convenable. Nous parlerons ici de modèles emboités dans le sens où le sous modèle peut être obtenu en fixant certains coefficients de θ à 0. L'hypothèse à tester est donc H_0 : « Le sous-modèle convient », et la question qui se pose est de savoir si l'augmentation du pouvoir de prédiction apporté par les variables supplémentaires du premier modèle est suffisante pour justifier la préférence de ce modèle au second. Sous l'hypothèse H_0 , et en notant SCR et SCR_0 les sommes des carrés résiduels du modèle et du sous modèle respectivement, nous avons le résultat suivant ;

$$\hat{F} := \frac{\frac{\mathrm{SCR}_0 - \mathrm{SCR}}{p - p_0}}{\frac{SCR}{n-p}} \sim FS(p - p_0, n - p)$$

Ce résultat que nous ne démontrerons pas ici découle du théorème de Cochran qui nous permet de montrer que les termes $\mathrm{SCR}_0 - \mathrm{SCR}$ et SCR suivent des lois du Khi² (respectivement d'ordre $p - p_0$ et d'ordre $n - p$) et sont indépendantes (par appartenance à des espaces orthogonaux et étant gaussien). Ainsi, en notant $F_{p-p_0,n-p}(1-\alpha)$ le quantile d'ordre $(1-\alpha)$ de la loi de Fisher Snedecor à $p - p_0$ et $n - p$ degrés de liberté et $F \sim FS(p - p_0, n - p)$, nous rejetterons l'hypothèse H_0 si $|\hat{F}| > F_{p-p_0,n-p}(1-\alpha)$ soit si $p - value := P(F > |\hat{F}|) > \alpha$ avec $\alpha = 5\%$ en règle générale.

Il est alors possible de mettre en place deux types de tests de significativité d'une variable dans un modèle. Les termes de TYPE I et TYPE II ou TYPE III sont souvent employés. Considérons un modèle $Y = X\theta + \varepsilon$ composé de n observations et p prédicteurs notés $(X_1 \quad \cdots \quad X_{p-1})$ en plus de l'intercept ici. Nous nous proposons de tester la significativité de la k[ème] variable pour k compris entre 1 et (p-1) selon les deux types de test. Le premier type de test consiste à contrôler la pertinence du modèle $Y = X_0\theta_0 + \varepsilon$ avec $X_0 = (1 \quad X_1 \quad \cdots \quad X_{k-1})$, considéré comme emboité dans le modèle $Y = X_1\theta_1 + \varepsilon$ avec $X_1 = (1 \quad X_1 \quad \cdots \quad X_k)$. Le résultat de ce test dépendra donc de l'ordre d'introduction des prédicteurs dans le modèle. C'est pourquoi l'on retrouve dans la plupart des logiciels le second type de test qui consiste à contrôler la pertinence du modèle $Y = X_0\theta_0 + \varepsilon$ avec $X_0 = (1 \quad X_1 \quad \cdots \quad X_{k-1} \quad X_{k+1} \quad \cdots \quad X_{p-1})$, soit le modèle complet sans le prédicteur X_k, contre le modèle complet. Le résultat obtenu ne dépend plus de l'ordre d'introduction des variables explicatives, et permet de mettre en place différentes procédures de décisions que nous allons décrire.

En effet, il est possible à partir de ce test de mettre en place différentes méthodes de construction d'un modèle. Nous pouvons par exemple partir du modèle avec le seul intercept et rajouter des variables jusqu'à obtention du modèle optimal, on parle alors de méthode ascendante. Nous pouvons également partir du modèle comportant le maximum de variables explicatives, puis retirer une par une les variables les moins significatives jusqu'à obtention du modèle le plus pertinent, on parle alors de méthode descendante. Nous pouvons enfin utiliser une combinaison de ces deux méthodes, on parle alors de méthode hiérarchique.

Nous remarquerons enfin une particularité qui concerne la significativité des variables qualitatives. En effet, il faudra alors distinguer la significativité de la variable de la significativité des indicatrices des occurrences prises séparément (qui représentent chacune un prédicteur fictif). En effet, les deux tests sont différents, le second confronte le modèle complet avec le modèle sans une indicatrice précise tandis que le premier confronte le modèle complet avec le modèle sans la variable, soit sans aucune des indicatrices associées aux occurrences de la variable. En pratique, on parlera de significativité de la variable, et il faudra alors faire attention, car si une variable qualitative est non significative, cela ne signifie pas que toutes ces occurrences le sont. En cas de divergence entre les deux tests, une nouvelle construction de la variable qualitative peut éventuellement être envisagée (regrouper ou dissocier certaines classes ou occurrences de la variable).

1.1.4 *Détection d'erreurs et validation des hypothèses de modélisation.*

Nous savons à présent construire une régression linéaire et tester sa significativité, mais il ne faut cependant pas oublier que les aspects théoriques de la régression linéaire entrevus ci-dessus reposent sur des hypothèses fortes. Il est donc nécessaire de vérifier la véracité de ces hypothèses, bien que des écarts modérés ne se révèlent pas problématiques. Chaque hypothèse peut faire l'objet de tests spécifiques, qui vont cependant cibler une hypothèse particulière et ne pourront envisager le grand nombre de cas possibles suite à une modélisation. Nous préférerons ainsi mettre en avant des analyses graphiques, qui jouent un grand rôle dans la validation des hypothèses d'une régression linéaire, en permettant d'une part de détecter les hypothèses éventuellement non vérifiés, et d'autre part en orientant l'utilisateur vers d'éventuelles corrections ou vers l'utilisation de modèles plus généraux. L'ensemble des graphiques concerne généralement l'analyse des résidus de la régression linéaire, vu comme des estimateurs du terme d'erreur de la modélisation.

Avant toute chose, nous définissons ci-dessous une matrice qui nous sera utile par la suite et que l'on appelle la matrice chapeau (ou « hat matrix ») ;

$$H = \left(h_{ij}\right)_{1\leq i,j\leq n} := X(X^T X)^{-1}X^T$$

Il est à noter que la matrice H est symétrique et est un projecteur (H^2=H et H est la matrice qui fait passer de Y à $\hat{Y}$). Or il se trouve que la trace d'un projecteur est égale à son rang soit à la dimension du sous espace de projection (résultat admis ici). L'espace de projection étant ici l'espace engendré par les vecteurs colonnes de X, nous pouvons en déduire que la trace de H vaut p, le nombre de variables explicatives.

Les résidus analysés pour la validation des hypothèses sont en général les $(\widehat{\varepsilon_i})_{1\leq i\leq n}$, il faut cependant faire remarquer que même sous l'hypothèse d'homoscédasticité, ces résidus estimés ne possèdent pas la même variance. En effet, $Var(\hat{Y}) = Var(HY) = \sigma^2 H^2$ ce qui implique logiquement que $Var(\widehat{\varepsilon_i}) = Var(Y_i - \hat{Y}_i) = \sigma^2(1 + h_{ii})$. Il est donc préférable de travailler sur ce que l'on appelle les résidus normalisés, notés r_i (voir tableau ci-après). Or, σ est inconnu, nous travaillons donc avec une estimation de ce dernier et donc sur ce que l'on appelle les résidus standardisés notés t_i (voir tableau ci-dessous). Cependant ces résidus ne sont pas indépendants par construction, car $\widehat{\sigma^2}$ a été estimé avec toutes les données, il est alors possible de corriger ce problème en travaillant sur les résidus studentisés, issus d'une estimation de σ^2 pour chaque résidu mais sans l'observation qui lui est associée. Ces résidus sont noté t_i^* (voir tableau ci-dessous), avec $\hat{\sigma}_{(i)}$ l'estimateur de σ dans le modèle linéaire privé de l'observation i. L'hypothèse d'indépendance des résidus est généralement vérifiée à partir du test de Durbin-Watson (voir [25]p 46).

Résidus estimés	$\widehat{\varepsilon_i}$
Résidus normalisés	$r_i = \dfrac{\widehat{\varepsilon_i}}{\sigma\sqrt{(1+h_{ii})}}$
Résidus standardisés	$t_i = \dfrac{\widehat{\varepsilon_i}}{\hat{\sigma}\sqrt{(1+h_{ii})}}$
Résidus studentisés	$t_i^* = \dfrac{\widehat{\varepsilon_i}}{\hat{\sigma}_{(i)}\sqrt{(1+h_{ii})}}$

Illustration 34 : Tableau des différents types de résidus

Suite à ces définitions, nous allons dans un premier temps décrire les outils permettant de vérifier la nullité de l'espérance des résidus ainsi que l'homoscédasticité (constance de la variabilité de l'erreur). Cette vérification s'opère en effectuant le tracé des résidus ou de la racine de leurs valeurs absolues en fonction des prédictions du modèle (on trouve souvent le terme anglais « residuals versus fitted »). L'absence de tendance et la constance de la variabilité de l'erreur viennent alors confirmer ces deux hypothèses. Si une tendance est visible, la transformation d'une variable explicative par une fonction (logarithme ou racine carré par exemple) peut suffire à corriger le problème. Si une certaine hétéroscédasticité se dégage, une transformation de la variable à expliquer peut parfois stabiliser la variance. L'utilisation de la méthode de Box-Cox, non décrite ici, est un outil intéressant d'aide à la décision quant à ces transformations (voir [25] p 103 ou [55]).

Il est ensuite nécessaire de vérifier l'adéquation des résidus à une loi Normale. La représentation graphique de ce que l'on appelle le QQ-Plot (« Quantile to Quantile Plot ») permet de vérifier cette hypothèse. Le tracé de cette représentation graphique a lieu en deux étapes. Il faut alors commencer par ordonner les résidus, puis l'on représente les résidus en fonction des quantiles de la loi normale en associant le i^{ème} résidu ordonné avec le quantile d'ordre $\frac{i}{n+1}$. Si l'hypothèse de normalité des résidus est vérifiée, alors les points de la représentation graphique seront approximativement alignés autour de la première bissectrice. Une représentation en forme de U signifiera que la distribution des résidus est asymétrique tandis qu'une représentation en forme de S signifiera que la distribution des résidus présente une plus grande influence des valeurs extrêmes (ce qui peut éventuellement suggérer l'utilisation d'un modèle linéaire généralisé autour de la loi Gamma). Cette analyse graphique peut éventuellement être complétée par un test de Shapiro-Wilks, de Kolmogorov-Smirnov ou encore un test du Khi², que nous ne détailleront pas ici (voir [25] p 57-92, [50] ou [59])

Il est également nécessaire de mesurer l'influence d'une observation sur la modélisation. Pour ce faire, nous pouvons nous intéresser à l'influence de la i^{ème} observation sur la valeur prédite. Nous savons que $\hat{Y} = HY$ et donc $\frac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}$, mesure de l'influence de la i^{ème} observation sur la valeur prédite. On parle alors de levier de Y_i. Etant donné la propriété sur la trace de H citée plus haut, nous savons que $\sum_{i=1}^{n} h_{ii} = p$ et nous pouvons donc nous attendre à ce que les leviers avoisinent la valeur $\frac{p}{n}$ (influence idéalement répartie entre chaque levier). Nous considérerons ainsi comme important voir anormal un levier supérieur à $\frac{2p}{n}$ (selon Hoaglin et Welsch [42]). Il peut paraitre cependant plus pertinent de s'intéresser à l'influence de la i^{ème} observation non pas sur la valeur prédite mais sur l'ensemble des valeurs prédites. Nous cherchons donc à mesurer la distance des moindres carrés entre les prédictions avec la i^{ème} observation et les prédictions sans cette observation, à normalisation près. L'un des critères les plus utilisé est alors la distance de Cook (voir [23]) de la i^{ème} observation, noté D_i et définie ci-dessous, que l'on peut exprimer d'une part en fonction des leviers décrits plus haut, et d'autre part en fonction des résidus standardisés (avec p le nombre de variables explicatives).

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{Y_j^{(i)}} - \widehat{Y_j} \right)^2}{p\widehat{\sigma^2}} = \frac{1}{p} \frac{h_{ii}}{(1 - h_{ii})} t_i^2$$

Le seuil de tolérance associé à cette distance de Cook le plus souvent employé est la valeur 1. En règle générale, toute observation pour laquelle la distance de Cook est élevée sera retirée de l'étude, ou son influence sur les coefficients surveillée de près. Nous pourront remarquer que la distance de Cook peut être vue comme la contribution de deux termes. Le terme t_i^2 mesure le degré d'adéquation de l'observation au modèle tandis que le terme $\frac{h_{ii}}{(1-h_{ii})}$ mesure l'effet de levier vu plus haut. La distance de Cook mesure donc à la fois le caractère aberrant d'une observation et la notion de point levier. D'autres indicateurs classiques tels que DFBETAS, DFFITS ou COVRATIO sont souvent employés, mais ne seront pas abordés ici (voir [36] p 119-176 ou [40] p 321-323).

Il nous reste enfin à aborder une dernière problématique, celle d'une éventuelle colinéarité des variables explicatives (Nous parlons ici de colinéarité approximative ou statistique et non pas de colinéarité « réelle »). En effet, une telle corrélation entre les variables peut poser problème dans l'inversion de la matrice $X^T X$, qui existe tout de même (la corrélation n'étant pas parfaite), mais contient des éléments très grands, ce qui implique de très grandes valeurs pour la variance de $\widehat{\theta}$, et agrandit les intervalles de confiances. Un tel phénomène peut être détecté en observant des différences importantes entre les p-valeurs de tests de significativités de variables d'un modèle à l'autre ou bien en utilisant ce que l'on appelle les facteurs d'inflation de la variance (ou VIF en Anglais). Le principe lié à cet indicateur est d'effectuer une régression de la i^{ème} variable explicative sur les autres et de calculer alors le coefficient de détermination de cette régression noté R_i^2. L'indicateur utilisé est alors noté ;

$$VIF_i = \frac{1}{1 - R_i^2}$$

Nous savons que le coefficient de détermination est compris entre 0 et 1. Sous l'hypothèse de non colinéarité entre les variables explicatives, tous les facteurs d'inflation de la variance seront nuls, sinon, cette valeur tendra vers l'infini. En pratique, on considère que l'on est en présence de multi colinéarité lorsqu'un indicateur est supérieur à 5 ou supérieur à 10. Une solution est alors d'augmenter le nombre d'observations, de diminuer le nombre de variables ou d'imposer des contraintes. Si cela n'est pas suffisant, ou si l'on désire réellement utiliser les variables colinéaires, il est alors possible d'utiliser ce que l'on appelle la « méthode de régression de Ridge » qui visent à utiliser des estimateurs biaisés, mais de variance plus faible que les estimateurs non biaisés. Cette méthode ne sera pas décrite ici, mais le lecteur intéressé est invité à consulter les articles de Hoerl , Kennard et Baldwin à ce sujet (voir [43], [44], [45]).

1.2 Modèles linéaires généralisés.

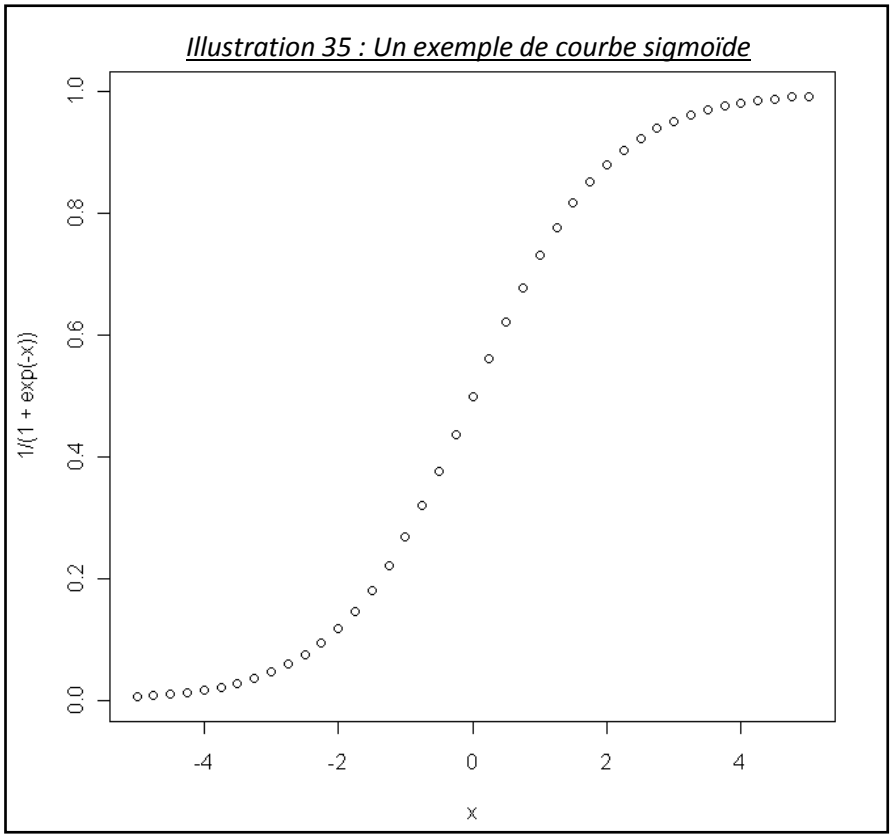
Nous avons entrevu dans le paragraphe précédent comment réaliser une régression linéaire. Nous allons à présent chercher à généraliser le modèle linéaire classique dans un cadre ou la variable à expliquer Y peut être qualitative, de plus, elle n'est pas forcément gaussienne, et enfin, l'hétéroscédasticité peut être retenue. Afin de permettre cette généralisation, nous considérerons que X le vecteur des prédicteurs est une variable aléatoire, nous allons ainsi travailler sur la distribution de Y conditionnellement aux observations de X. Nous avons d'ores et déjà entrevu le cas particulier de la régression linéaire, nous allons à présent considérer un second cas particulier des modèles linéaires généralisés que l'on appelle la régression logistique, puis nous nous intéresserons plus en détail à la généralisation à proprement parler.

1.2.1 La régression logistique.

Nous allons chercher à modéliser Y la notion d'absence ou de présence (d'une maladie par exemple), en fonction d'une variable explicative X (l'âge par exemple). Bien souvent, une représentation de la moyenne des observations de Y (notion de fréquence) par classes d'observations de X ressemble à une courbe sigmoïde (en forme de S). Nous somme donc en mesure de proposer un modèle théorique tel que l'espérance conditionnelle de Y sachant $X=x_i$ suit une courbe sigmoïde. De nombreuses fonctions de ce type existent, mais nous en choisissons une très simple, continue et dérivable, de la forme ;

$$f(x) = h(x\theta) = \frac{\exp(x\theta)}{1+\exp(x\theta)} = \frac{1}{1+\exp(-x\theta)}$$

Illustration 35 : Un exemple de courbe sigmoïde



Nous notons alors $\eta_i = x_i\theta$, appelé prédicteur linéaire. En effet, on remarque alors que η_i suit un modèle de régression linéaire sans terme d'erreur. Nous avons donc la relation suivante ;

$$h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{1}{1 + \exp(-\eta_i)}$$

h() étant une fonction inversible, nous pouvons poser g()=h()$^{-1}$ et réécrire notre modèle théorique sous les deux formes suivantes ;

$$\mu_i = E(Y|X = x_i) = h(\eta_i) \; ou \; g\big(E(Y|X = x_i)\big) = g(\mu_i) = logit(\mu_i) = \ln\left(\frac{\mu_i}{1 - \mu_i}\right) = \eta_i$$

Nous constatons alors que la deuxième écriture correspond exactement à la régression linéaire, à ceci près que la fonction g() n'est pas ici la fonction identité mais une fonction continue dérivable et inversible. Nous avons donc généralisé le modèle de régression simple en rajoutant une transformation g() de la moyenne. Cette fonction est appelée fonction de lien.

Ceci est cependant insuffisant, et le modèle incomplet. En effet, lors de la régression linéaire, la présence d'un terme d'erreur suivant une loi normale nous permettait d'obtenir des renseignements sur l'espérance mais également sur la variance de la variable à expliquer. Afin d'obtenir des résultats équivalents, nous n'allons pas explicitement rajouter un terme d'erreur dans le but de définir la loi de Y, mais nous allons définir directement et entièrement la loi de Y.

Nous cherchons ici à modéliser une proportion théorique. La loi la plus naturelle pour une telle modélisation est alors la loi binomiale fréquence $\mathfrak{B}(n_i, \mu_i)/n_i$. Remarquons que les effectifs, notés n_i , sont supposés connus et fixés (bien que variables d'une observation à l'autre). La fonction g() étant également fixée, les seuls paramètres inconnus sont alors les coefficients de θ. Enfin, étant donné que Y sachant X=x_i suit une loi binomiale $\mathfrak{B}(n_i, \mu_i)/n_i$ d'espérance $\mu_i = g^{-1}(x_i\theta)$, nous savons à présent que la variance de Y sachant X=x_i vaut $\mu_i(1 - \mu_i)/n_i$.

En conclusion, nous avons mis en place une modélisation linéaire généralisée d'une variable qualitative Y, ou plus justement de la transformation de l'espérance conditionnelle d'une variable qualitative $g\big(E(Y|X = x_i)\big)$. Cette modélisation, après transformation, est linéaire en x_i. Les termes d'erreurs ne suivent pas une loi normale mais une loi de Bernoulli (La loi binomiale est obtenu en sommant les termes d'erreurs autant de fois que l'effectif de la valeur x_i). Et enfin, l'hypothèse d'homoscédasticité n'est pas vérifiée, car la variance dépend à présent des variables explicatives.

Nous allons donc entrevoir les modèles linéaires généralisés à partir des principes exposés ci-dessus, mais de manière à englober les cas particuliers de la régression linéaire et de la régression logistique, entre autres. Les principales caractéristiques de cette modélisation sont alors ;

- Une transformation de l'espérance conditionnelle par une fonction de lien monotone et dérivable.
- La modélisation de cette transformation par une combinaison linéaire des variables explicatives.
- Des termes d'erreur non gaussien et non explicites.
- La définition de la loi conditionnelle de Y appartenant à la famille exponentielle, que nous allons définir et décrire ci-dessous avant de revenir sur la modélisation linéaire généralisée.

1.2.2 *La famille exponentielle.*

Dans un premier temps, nous allons définir de manière générale la famille exponentielle, puis nous nous placerons dans un cadre simplifié pour la modélisation linéaire généralisée et enfin nous nous intéresserons à certaines propriétés issues de cette famille de loi. Attention, il est à noter que le coefficient θ utilisé dans ce paragraphe concernant la famille exponentielle n'est pas le même que celui utilisé dans les cas particuliers de régressions.

Définition générale

- La loi de probabilité P appartient à une famille de loi de type exponentielle générale $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^q}$ s'il existe une mesure dominante ν telle que les lois P_θ ont pour densité par rapport à ν ;

$$f_\theta(y) = c(\theta)h(y)exp\left\{\sum_{j=1}^{p} \alpha_j(\theta)T_j(y)\right\}, y \in \mathcal{Y}$$

Avec $T_1(.), \dots, T_p(.), \alpha_1(.), \dots, \alpha_p(.)$ fonctions mesurables et $\mathcal{Y}$ l'ensemble de définition de $f_\theta(.)$.

- Les quantités $\{\alpha_j(\theta)\}_{1 \le j \le p}$ sont appelées paramètres naturels de la loi exponentielle générale et sont des éléments de l'espace naturel des paramètres ;

$$\mathcal{A} = \left\{\alpha \in \mathbb{R}^p \ , \ \int_{\mathcal{Y}} h(y)exp\left[\sum_{j=1}^{p} \alpha_j(\theta)T_j(y)\right] < +\infty\right\}$$

- Une famille exponentielle est dite naturelle si $f_\theta(y) = c(\theta)h(y)exp\left\{\sum_{j=1}^{q} \theta_j y_j\right\}$.
- La fonction qui à $\theta \in \Theta \subset \mathbb{R}^q$ associe le vecteur $\left(\alpha_1(\theta), \dots, \alpha_p(\theta)\right)$ n'est pas forcément bijective, et q n'est pas forcément égal à p.

Définition spécifique

Dans le cadre particulier des modèles linéaires généralisés, nous allons nous intéresser à une forme particulière de la famille de loi exponentielle. Nous supposerons en effet que la variable à expliquer Y possède une densité par rapport à une mesure dominante ν qui s'écrit ;

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

a(.), b(.) et c(.) sont des fonctions connues et dérivables, b(.) est trois fois dérivable et sa dérivée première est inversible, et le couple de paramètre (θ, ϕ) appartient à $\Omega \subset \mathbb{R}^2$. θ est parfois appelé le paramètre naturel et ϕ est appelé le paramètre de dispersion.

Si ϕ est connu, alors la densité de Y est un élément de la famille exponentielle, ce qui n'est pas toujours le cas lorsque ϕ est inconnu. En pratique, nous estimerons ϕ séparément puis nous le supposerons connu et fixé.

Nous allons à présent énoncer un certain nombre de propriétés et définitions avant d'exposer des exemples de lois classiquement utilisées dans la modélisation linéaire généralisée.

Soit Y une variable de densité $f_{\theta,\phi}$ comme défini plus haut. Nous savons alors que l'intégrale de cette densité vaut 1, et grâce au lemme de Fatou et à la linéarité de l'intégrale, il nous est possible d'écrire que ;

$$\int_y \frac{\partial f_{\theta,\phi}}{\partial \theta} d\nu = \frac{\partial \int_y f_{\theta,\phi}\, d\nu}{\partial \theta} = 0$$

$$\Rightarrow \int_y \frac{y - b'(\theta)}{a(\phi)} f_{\theta,\phi} d\nu = \frac{1}{a(\phi)} \left[\int_y y\, f_{\theta,\phi} d\nu - b'(\theta) \int_y f_{\theta,\phi}\, d\nu \right] = 0 \; avec \; a(\phi) \neq 0$$

$$\boxed{\Rightarrow E(Y) = b'(\theta)}$$

Nous pouvons alors, en utilisant la même logique, déterminer la variance de Y comme suit ;

$$\int_y \frac{\partial^2 f_{\theta,\phi}}{\partial^2 \theta} d\nu = \frac{\partial^2 \int_y f_{\theta,\phi}\, d\nu}{\partial^2 \theta} = 0$$

$$\Rightarrow \int_y \left[-\frac{b''(\theta)}{a(\phi)} + \frac{\left(y - b'(\theta)\right)^2}{a(\phi)^2} \right] f_{\theta,\phi} d\nu = \frac{-b''(\theta)}{a(\phi)} + \frac{1}{a(\phi)^2} \int_y (y - E(Y))^2 f_{\theta,\phi} d\nu = 0$$

$$\boxed{\Rightarrow Var(Y) = b''(\theta) a(\phi)}$$

Ceci nous permet alors de constater que le coefficient θ exerce un contrôle sur l'espérance et la variance de la variable à expliquer au moyen des fonctions $b'(.)$ et $b''(.)$. La moyenne et la variance sont donc liées entre elles, comme nous l'avons vu plus haut. Nous noterons dans toute la suite μ la moyenne de Y. La variance de Y est alors fonction de μ, et nous noterons V(μ) cette fonction, appelée fonction de variance. Nous notons de plus que par définition, la fonction $b'(.)$ est inversible, et nous avons ainsi la relation $\theta = (b')^{-1}(\mu)$. La fonction de lien g() évoquée plus haut sera alors appelée lien canonique lorsque la relation $g(\mu) = \theta = (b')^{-1}(\mu)$ est vérifiée. Le paramètre de dispersion ϕ n'influence quant à lui que la variance de Y à partir de la fonction a(.). En règle générale, cette fonction est de la forme $a(\phi) = \phi d$ où d représente un poids sur une observation. Nous fixerons par la suite d=1, sans perte de généralité, et ainsi $a(\phi) = \phi$.

Nous allons à présent représenter sous forme de tableaux les principales lois de la famille exponentielle utilisées en modélisation linéaire généralisée et leurs densités. Nous reprendrons alors les expressions de ϕ, $b(\theta)$, $c(y, \phi)$ ainsi que celle de la moyenne μ, de la fonction de variance V(μ), de la variance et enfin de la fonction de lien canonique. La vérification des résultats est laissée au soin du lecteur.

Illustration 36 : Lois de la famille exponentielle classiquement utilisée en modélisation linéaire généralisée

Lois	$\mathcal{Y}$; Ω	Densité/Loi	ϕ	$b(\theta)$	$c(y,\phi)$	$\mu = b'(\theta)$	$V(\mu)=b''(\theta)$	$Var(Y)=\phi V(\mu)$	Lien canonique
Loi Normale $\mathcal{N}(\mu,\sigma^2)$	$\mathbb{R}$; $\mathbb{R}\times\mathbb{R}^{+*}$	$\frac{1}{\sqrt{2\pi}\sigma} exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right]$	σ^2	$\frac{\theta^2}{2}$	$\frac{-\left(\frac{y^2}{\phi}+log\,(2\pi\phi)\right)}{2}$	θ	*1*	σ^2	*Identité*
Loi Gamma $\mathcal{G}(\mu,\tau)$	$\mathbb{R}^{+*}$; $\mathbb{R}^{+*}\times\mathbb{R}^{+*}$	$\frac{(\frac{\tau}{\mu})^\tau y^{\tau-1} exp\,(-\frac{\tau}{\mu}y)}{\Gamma(\tau)}$	$\frac{1}{\tau}$	$-ln(-\theta)$	$\tau ln(\tau y) - ln(y)$ $- ln\,(\Gamma(\tau))$	$-\frac{1}{\theta}$	$\frac{1}{\theta^2}=\mu^2$	$\frac{\mu^2}{\tau}$	$-\frac{1}{\mu}$ $(Inverse)$
Loi Poisson $\mathcal{P}(\mu)$	$\mathbb{N}$; $\mathbb{R}^{+*}$	$P(Y=y)=exp\,(-\mu)\frac{\mu^y}{y!}$	*1*	$exp\,(\theta)$	$-ln(y!)$	$exp\,(\theta)$	$exp(\theta)=\mu$	μ	$ln(\mu)$ *(Log)*
Loi Binomiale $\frac{\mathcal{B}(n,\mu)}{n}$	$\frac{[\![0;n]\!]}{n}$; $\mathbb{N}\times[0;1]$	$P(Y=y)=\binom{n}{ny}\mu^{ny}(1-\mu)^{n-ny}$	$\frac{1}{n}$	$ln\,(1+e^\theta)$	$ln\,(\binom{n}{ny})$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$	$\frac{\mu(1-\mu)}{n}$	$ln\left(\frac{\mu}{1-\mu}\right)$ *(Logit)*

- $\mathcal{Y}$ désigne l'ensemble de définition de la densité ou de la loi, Ω désigne l'ensemble de définition des coefficients associés à la loi.
- Classiquement, la loi Gamma est notée $\mathcal{G}(\alpha,n)$ avec pour densité $\frac{\alpha^n y^{n-1} exp\,(-\alpha y)}{\Gamma(n)}$ et $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$ pour tout $x\in\mathbb{R}^{+*}$. L'écriture du tableau a été préférée à cette dernière afin de faire en sorte que la moyenne apparaisse clairement dans un coefficient de la loi. Le parallèle entre les deux écritures se réalise sans complexité avec les relations $\frac{\tau}{\mu}=\alpha\ et\ \tau=n$. Nous retrouvons donc bien l'espérance $\mu=\frac{n}{\alpha}$.
- Classiquement, la loi binomiale est notée $\mathcal{B}(n,p)$ avec pour loi $P(Y=y)=\binom{n}{y}p^y(1-p)^{n-y}$ pour $y\in[\![0;n]\!]$. La loi représentée dans le tableau correspond à ce que l'on appelle la loi Binomiale-fréquence utilisée pour la régression logistique dans le but de modéliser une proportion.
- L'utilisation de la loi Normale correspond à la régression linéaire classique, avec des erreurs gaussiennes et sous l'hypothèse d'homoscédasticité. La loi Gamma est utile pour la modélisation de valeurs positives strictement avec une augmentation de la variance avec la moyenne, elle est souvent utilisée pour la modélisation de montants de sinistres. La loi de Poisson est utilisée pour modéliser des fréquences ou des taux d'un processus de comptage, la variance augmentant avec la moyenne. Enfin la loi Binomiale est utilisée pour la régression logistique entre autre et pour la modélisation de données binaires (ex : absence/présence), elle vise à estimer une proportion.

1.2.3 *Définition d'un modèle linéaire généralisé.*

Nous allons à présent définir proprement ce que l'on entend par modèles linéaires généralisés à partir des principes évoqués ci-dessus, avant d'aborder plus en détails l'estimation des paramètres et les outils de contrôle ou d'aide à la décision.

Un modèle linéaire généralisé a pour but de relier des variables explicatives $X = (X_1 \quad \cdots \quad X_p)$ à une variable à expliquer Y. La logique sous-jacente à un tel modèle peut alors être résumée à travers le schéma suivant ;

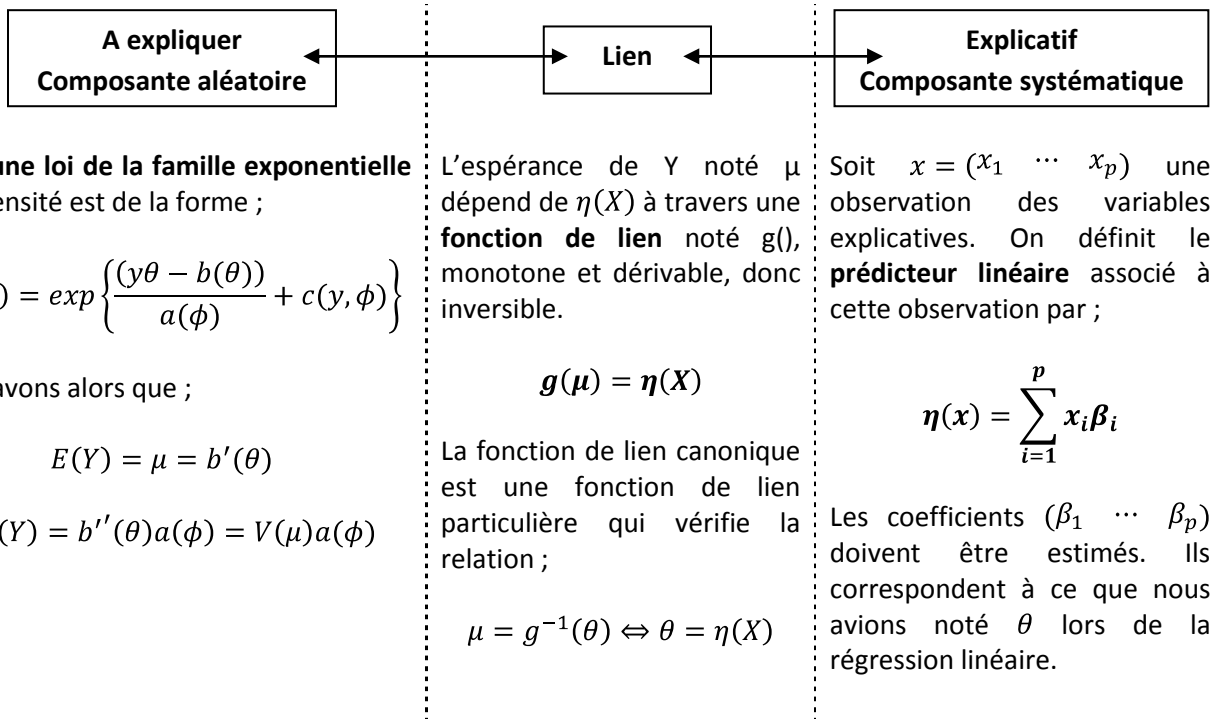


Illustration 37 : Schéma de construction d'un modèle linéaire généralisé

Pour construire un modèle linéaire généralisé, il faut donc commencer par choisir la loi de Y dans la famille exponentielle, ce qui fixe les fonctions *a(), b() et c().* Ce choix peut être orienté à partir du tableau de la page précédente. Il est également nécessaire de choisir une fonction de lien (attention, une fonction quelconque ne peut pas être fonction de lien). Un choix particulier qui simplifie les calculs est le choix de la fonction de lien canonique, soit $g = (b')^{-1}$.

Ensuite, pour utiliser le modèle linéaire généralisé, on commencera en règle générale par estimer séparément le paramètre de dispersion, afin de le considérer comme fixé par la suite. Puis il faudra estimer les paramètres $(\beta_1 \quad \cdots \quad \beta_p)$, ce qui fixera ainsi $\eta(X)$, ce qui détermine alors $\mu = g^{-1}(\eta(X))$, qui est la moyenne et donc la prévision par le modèle. Ceci fixera enfin θ, que l'on pourra déterminer à partir de la relation $\theta = (b')^{-1}(\mu)$, et qui permet également de calculer la fonction variance $V(\mu)$ et la variance de Y.

Nous avons donc vu ci-dessus qu'il est nécessaire d'effectuer deux choix pour construire un modèle linéaire généralisé. Le premier concerne la loi de la variable à expliquer. Comme nous l'avons signalé précédemment, ce choix peut être orienté par le type de la variable et des connaissances préalables. Le deuxième choix porte sur la fonction de lien. Nous reprenons ci-dessous à travers un tableau les fonctions de liens classiquement utilisées. Le choix de la densité peut alors dépendre de la loi. En effet, si Y est binaire, on préférera utiliser les liens *logit*, *probit* ou *clogclog*, si Y est un comptage, on utilisera classiquement le lien *log*, et enfin si Y est continue, on pourra utiliser les liens canoniques des lois normale et gamma. Le choix du lien peut également être déterminé par l'existence d'études passées, par une connaissance préalable du problème. Le choix de la fonction de lien canonique peut également être retenu. Enfin, il est possible d'utiliser des fonctions de lien possédant un paramètre, afin de garder une certaine souplesse. Nous noterons cependant que la fonction de lien *puissance* est souvent implémentée à paramètre constant dans les logiciels, et une estimation de ce paramètre est alors nécessaire. Nous remarquons également que le lien (ou la famille) *d'Aranda-Ordaz* tend vers le lien *cloglog* quand son paramètre tend vers 0 et que l'on retrouve le lien *logit* lorsque le paramètre vaut 1.

Nom du lien	**Fonction de lien**
Lien identité	$g(\mu) = \mu$
Lien log	$g(\mu) = \ln(\mu)$
Lien cloglog	$g(\mu) = \ln(-\ln(1-\mu))$
Lien logit	$g(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$
Lien probit	$g(\mu) = \Phi(\mu)$ Φ fonction inverse de la fonction de répartition d'une loi $\mathcal{N}(0,1)$
Lien réciproque	$g(\mu) = -1/\mu$
Lien puissance	$g(\mu) = \mu^{\gamma}\ avec\ \gamma \neq 0$ $g(\mu) = \ln(\mu)\ avec\ \gamma = 0$
Aranda Ordaz (asymétrique)	$g(\mu) = ln\left(\frac{(1-\mu)^{-\lambda} - 1}{\lambda}\right)$

Illustration 38 : Fonctions de lien usuelles

Nous connaissons à présent tous les éléments nécessaires à la construction d'un modèle linéaire généralisé. Cependant il nous faut encore pouvoir utiliser ce modèle, et nous allons donc nous intéresser plus en détail à l'estimation des coefficients par maximum de vraisemblance, ainsi qu'aux intervalles de confiances qui en découlent et à la notion de prédiction du modèle.

1.2.4 Estimations des paramètres par maximum de vraisemblance.

1.2.4.1 Calcul de la log-vraisemblance.

Nous considérons une variable à expliquer Y, pour laquelle nous possédons des observations pour n individus notées $(Y_1 \quad \cdots \quad Y_n)$. Nous cherchons alors à expliquer cette variable à partir de p variables explicatives notées $(X_1 \quad \cdots \quad X_p)$. Les observations des variables explicatives sont regroupées dans une variable X à n lignes et p colonnes, d'éléments $\left(X_{ij}\right)_{\substack{1\le i\le n\\ 1\le j\le p}}$. X_{ij} représente ainsi la i[ème] observation de la j[ème] variable X_j. La notation X_j représentera indifféremment l'ensemble des observations de la variable X_j ou la variable X_j elle-même. La notation $X_{i.}$ représentera le vecteur ligne de la i[ème] observation de l'ensemble des variables explicatives. Nous supposons qu'un modèle GLM a été construit, et que nous connaissons ainsi la densité de Y et la fonction de lien *g()*, toutes deux supposées connues et fixées.

La fonction de densité $f_{\theta,\phi}$ étant connue, nous pouvons écrire la log-vraisemblance (ou « profile-likelihood ») pour la i[ème] observation, en supposant que toutes les observations ont le même poids;

$$l_i = l(Y_i, \beta, \phi) = \ln\left(f_{\theta,\phi}\right) = \frac{(Y_i\theta - b(\theta))}{a(\phi)} + c(Y_i, \phi) = \frac{(Y_i\theta - b(\theta))}{\phi} + c(Y_i, \phi)$$

Rappelons que l'objectif de l'estimation par maximum de vraisemblance est de trouver les paramètres $\hat{\theta}\ et\ \hat{\phi}$ qui maximisent la log-vraisemblance. Nous notons alors que le paramètre de dispersion sert à ajuster la variance du modèle, il n'influe en aucun cas sur la partie moyenne de l'expression et donc sur la prévision ou l'explication du modèle. Il est donc d'usage de considérer ce paramètre comme fixé et de l'estimer séparément. Nous cherchons alors à estimer les paramètres $(\beta_1 \quad \cdots \quad \beta_p)$, qui n'apparaissent pas explicitement dans l'expression de la log-vraisemblance. Il nous faut ainsi calculer les différentes dérivées partielles, et nous allons utiliser la composition de fonction afin de pouvoir le faire. Nous définissons alors μ_i l'espérance de Y conditionnellement à la i[ème] observation, ainsi que θ_i la valeur du paramètre θ au point $X = X_{i.}$ (car si la moyenne dépend des observations, alors θ aussi car $\theta = (b')^{-1}(\mu)$), et nous définissons de même η_i le prédicteur linéaire de la i[ème] observation. Nous allons alors chercher à calculer les dérivés partielles de l_i à partir de la forme inscrite ci-dessous ;

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j}\ pour\ tout\ j \in [\![1, p]\!]$$

Nous allons alors calculer séparément chaque dérivée partielle ;

- La première expression s'obtient en dérivant simplement : $\frac{\partial l_i}{\partial \theta_i} = \frac{(Y_i - b\prime(\theta_i))}{\phi}$
- Nous savons que $b'(\theta_i) = \mu_i$ ce qui implique : $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = V(\mu_i)$
- Nous savons que $\mu_i = g^{-1}(\eta_i)$ ce qui implique en notant h=g^{-1} : $\frac{\partial \mu_i}{\partial \eta_i} = h'(\eta_i)$
- Enfin on écrit simplement : $\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \sum_{k=1}^{p} X_{ik}\beta_k}{\partial \beta_j} = X_{ij}$

A partir des calculs effectués ci-dessus, on obtient alors la relation suivante ;

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(Y_i - b'(\theta_i))}{\phi} \times \frac{1}{V(\mu_i)} \times h'(\eta_i) \times X_{ij} \; pour \; tout \; j \in [\![1, p]\!]$$

En posant $W_i = \frac{1}{V(\mu_i)} \times \left(h'(\eta_i)\right)^2$, on obtient finalement la relation suivante ;

$$\boxed{\frac{\partial l_i}{\partial \beta} = \frac{Y_i - \mu_i}{\phi} \times \frac{W_i}{h'(\eta_i)} \times {X_{i.}}^T}$$

Il est à noter que si la fonction de lien canonique est utilisée, alors nous savons que $\eta_i = \theta_i$, ce qui simplifie alors sensiblement les choses en limitant le calcul aux dérivés partielles $\frac{\partial l_i}{\partial \theta_i}$ et $\frac{\partial \eta_i}{\partial \beta_j}$. Ceci nous permet alors d'obtenir finalement la relation suivante ;

$$\boxed{\frac{\partial l_i}{\partial \beta} = \frac{Y_i - \mu_i}{\phi} \times {X_{i.}}^T}$$

Il subsiste alors un problème, car nous ne connaissons pas μ_i , qui dépend de β, et il n'est alors pas possible de parvenir à une expression exploitable de β en annulant cette dérivée. Nous allons donc avoir recours à une procédure itérative afin de maximiser la vraisemblance.

1.2.4.2 Méthode itérative de maximisation.

Pour maximiser la log-vraisemblance notée $\mathcal{L} = \sum_{i=1}^{n} l_i$, nous allons minimiser $-\mathcal{L}$ par une procédure itérative. Nous fixons un point de départ noté β^0, puis on construit β^1 tel que $-\mathcal{L}(\beta^1) < -\mathcal{L}(\beta^0)$ et ainsi de suite jusqu'à ce que le gain obtenu par une itération soit peu significatif. Pour ce faire, nous utilisons la direction de la plus forte descente qui est le gradient $\nabla\mathcal{L}(\beta) = \left(\frac{\partial \mathcal{L}(\beta)}{\partial \beta}\right)^T$ à p lignes et 1 colonne. Avec β^k fixé, on construit alors β^{k+1} comme suit ;

$$\boxed{\beta^{k+1} = \beta^k + A_k \nabla\mathcal{L}_k \; avec \; \nabla\mathcal{L}_k = \nabla\mathcal{L}\left(\beta^k\right) \; et \; A_k \; la \; matrice \; de \; "pas"}$$

Il existe alors différentes possibilités pour le choix de la matrice de « pas » qui conduisent à des algorithmes de simplicité et d'efficacité différentes. Nous annotons ci-dessous différentes possibilités pour le choix de A_k.

- A_k est une constante, ce qui est le choix le plus simple mais le moins efficace.
- A_k est fonction de , $\nabla\mathcal{L}_k \; et \; \beta^k$, comme pour la méthode Polak Ribière.
- $A_k = -(\nabla^2\mathcal{L}_k)^{-1}$, ce qui correspond à la méthode de Newton avec $H(\beta) = \left(\frac{\partial \mathcal{L}(\beta)}{\partial \beta_i \partial \beta_j}\right)_{1 \le i,j \le p}$ appelée la matrice Hessienne et $\nabla^2\mathcal{L}_k = H(\beta^k)$.
- $A_k = -(E(\nabla^2\mathcal{L}_k))^{-1}$ et l'on parle alors de la méthode de Fisher scoring qui limite les problèmes éventuels d'inversibilité locale du hessien en prenant l'espérance. Cette méthode est la plus utilisée dans le cadre des modèles linéaires généralisés et nous allons donc nous restreindre à celle-ci.

Nous allons ainsi effectuer le calcul de la matrice hessienne, et étant donné que $\mathcal{L}(\beta) = \sum_{i=1}^{n} l_i$, il nous suffit d'effectuer le calcul pour une seule observation, d'en prendre l'espérance et de sommer. Nous connaissons déjà la dérivée partielle de premier ordre de la log-vraisemblance pour la i[ème] observation, ce qui nous permet d'en déduire l'expression ;

$$\frac{\partial l_i}{\partial \beta_j} = \frac{Y_i - \mu_i}{\phi} \times \frac{W_i}{h'(\eta_i)} \times X_{ij}$$

$$d'où \quad \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k}\left[\frac{Y_i - \mu_i}{\phi} \times \frac{W_i}{h'(\eta_i)} \times X_{ij}\right]$$
$$= \frac{\partial}{\partial \beta_k}\left[\frac{Y_i - \mu_i}{\phi}\right] \times \frac{W_i}{h'(\eta_i)} \times X_{ij} + \frac{Y_i - \mu_i}{\phi} \times \frac{\partial}{\partial \beta_k}\left[\frac{W_i}{h'(\eta_i)}\right] \times X_{ij}$$

Le dernier terme du membre de droite, que nous noterons K, est d'espérance nulle, nous ne cherchons donc pas plus à le développer, contrairement au premier terme, pour lequel nous allons une fois de plus utiliser la composition de fonction et le fait que : $\frac{\partial \mu_i}{\partial \eta_i} = h'(\eta_i)$. Nous pouvons ainsi réécrire le résultat comme suit ;

$$\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = -\frac{1}{\phi}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k} W_i \frac{\partial \eta_i}{\partial \mu_i} X_{ij} + K$$

$$\boxed{\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} = -\frac{1}{\phi} X_{ik} W_i X_{ij} + K}$$

Le passage à l'espérance permet alors d'annuler le second terme en laissant inchangé le premier, ce qui nous permet d'écrire le résultat final ;

$$-E(\nabla^2 \mathcal{L}) = \left(\sum_{i=1}^{n} \frac{1}{\phi} X_{ik} W_i X_{ij}\right)_{1 \le k,j \le p} = \frac{1}{\phi} X^T W X \;\; avec\, W = diag(W_1 \quad \cdots \quad W_n)$$

$$-\left(E(\nabla^2 \mathcal{L})\right)^{-1} = \phi (X^T W X)^{-1}$$

Nous pouvons alors écrire les équations d'itération de la méthode, en prenant la notation W^k pour insister sur le fait que W dépend de l'espérance, donc de β^k et donc de l'étape de l'itération.

$$\beta^{k+1} = \beta^k + \phi\left(X^T W^k X\right)^{-1} \nabla \mathcal{L}_k$$

Or nous pouvons écrire d'après les calculs de la dérivée partielle du premier ordre qui ont déjà été effectués que ;

$$\nabla \mathcal{L}_k = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\phi} \times \frac{W_i}{h'(\eta_i)} \times X_{i.}^{\;T} = \frac{1}{\phi} X^T W^k diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)(Y - \mu)$$

Nous réintégrons alors cette expression de $\nabla \mathcal{L}_k$ dans l'étape d'itération et nous composons à gauche par $\left(X^T W^k X\right)$ pour simplifier l'expression représentée ci-après.

$$\beta^{k+1} = \beta^k + \left(X^T W^k X\right)^{-1} X^T W^k diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)(Y-\mu)$$

$$X^T W^k X\beta^{k+1} = X^T W^k X\beta^k + X^T W^k diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)(Y-\mu) = X^T W^k \left[X\beta^k + diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)(Y-\mu)\right]$$

$$\boxed{\beta^{k+1} = \left(X^T W^k X\right)^{-1} X^T W^k \left[X\beta^k + diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)(Y-\mu)\right] = \left(X^T W^k X\right)^{-1} X^T W^k Z^k}$$

L'introduction de la notation $Z^k = \left[X\beta^k + diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)(Y-\mu)\right]$ nous permet de remarquer que l'algorithme de Fisher scoring correspond en fait à la réalisation d'une régression linéaire pondérée où Z^k est la variable à expliquer, X la variable explicative et W^k la matrice de poids (voir les formules des ***pages 75 et 76*** pour s'en convaincre). β^{k+1} est alors l'estimateur des moindres carrés des coefficients de cette régression. Nous laissons le soin au lecteur de vérifier que les dimensions des matrices obtenues sont cohérentes (étape sans grande complexité) et nous faisons remarquer que les éléments η_i, μ_i et μ de Z^k dépendent de l'étape d'itération et sont entièrement définis par le paramètre β^k de l'itération précédente. Enfin nous mettons en avant le fait qu'en pratique, le point de départ est souvent déterminé par un η^0 plutôt que par β^0 car il est plus facile de trouver un bon point de départ pour η. Ce point de départ est alors suffisant car ceci détermine $X\beta^0$ ainsi que les autres éléments de Z^0 étant donné que $\mu^0 = g^{-1}(\eta^0)$.

Nous sommes à présent en mesure d'effectuer une estimation de β. Nous pouvons alors remarquer que le paramètre de dispersion n'intervient pas dans l'estimation de β (le supposer connu et fixé n'a donc pas vraiment de conséquence). Nous faisons également remarquer que l'utilisation de la fonction de lien canonique permet de simplifier les calculs (comme signalé auparavant). Enfin nous remarquons que si le paramètre de dispersion est supposé connu, alors la matrice $\phi(X^T WX)^{-1}$ correspond à l'inverse de la matrice d'information de Fisher, d'où le nom de cet algorithme, que l'on appelle également souvent IRLS pour « Iterative reweighted least squares » (moindres carrés itérativement repondérés en traduction littérale).

1.2.4.1 Loi asymptotiques et intervalles de confiance.

L'utilisation de l'estimation par maximum de vraisemblance nous permet alors d'utiliser certaines bonnes propriétés asymptotiques dans le but de calculer des intervalles de confiance. Nous savons en effet que sous certaines hypothèses classiques de régularité, l'estimateur du maximum de vraisemblance est asymptotiquement normal, de moyenne nulle et de variance l'inverse de la matrice d'information de Fisher. Le lecteur intéressé pourra se référer aux ouvrages de Lejeune [48] (chapitre 9) ou de Schervish [57] pour de plus amples informations. En supposant que le paramètre de dispersion est connu et fixé, il nous est ainsi possible d'écrire que ;

$$\boxed{\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \phi(X^T WX)^{-1})}$$

De plus avec f de $\mathbb{R}^p$ dans $\mathbb{R}$ une fonction différentiable en β et de dérivées partielles continues, on obtient le résultat suivant à partir d'une propriété appelée la méthode Delta ;

$$\boxed{\begin{array}{c}\sqrt{n}\left(f\left(\hat{\beta}\right) - f(\beta)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla f^T . \phi(X^T WX)^{-1} . \nabla f) \\ avec\ \nabla f\ le\ vecteur\ colonne\ des\ dérivés\ partielles\ par\ rapport\ aux\ \beta_i\end{array}}$$

Il est alors cependant nécessaire d'évaluer la matrice des poids W en un point β inconnu. Deux solutions mises en avant par Schervish [57] (p 423) consistent soit à estimer W à partir de $\hat{\beta}$, soit à estimer la matrice $\phi(X^T W X)^{-1}$ par $n\phi(X^T W X)^{-1}$. Dans toute la suite nous noterons $W^* = W(\hat{\beta})$. Nous pouvons alors calculer des intervalles de confiance pour les coefficients $\widehat{\beta_\iota}$ en utilisant la méthode Delta avec une fonction f qui à β associe β_i , ce qui implique que le gradient est un vecteur colonne composé de 0 excepté pour la i[ème] coordonnée qui vaut 1. Nous obtenons ainsi finalement que ;

$$\boxed{\sqrt{n}(\widehat{\beta_\iota} - \beta_i) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2_{\widehat{\beta_\iota}}\right) \; avec \; \sigma^2_{\widehat{\beta_\iota}} = [\phi(X^T W^* X)^{-1}]_{ii} \; pour \; tout \; i \in [\![1; p]\!]}$$

A partir de cette expression, et en notant $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $\left(1 - \frac{\alpha}{2}\right)$ d'une loi normale centrée réduite, nous pouvons exprimer un intervalle de confiance à $(1-\alpha)$ pourcents pour la i[ème] composante de β ;

$$\boxed{IC_\alpha(\beta_i) = \left[\widehat{\beta_\iota} - \frac{\sigma_{\widehat{\beta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \; ; \; \widehat{\beta_\iota} + \frac{\sigma_{\widehat{\beta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right]}$$

Il est également possible d'obtenir différemment un intervalle de confiance pour les composantes de β, à partir de la vraisemblance, et en utilisant des résultats théoriques issus du test de rapport de vraisemblance (voir [57] p 459), que l'on ne démontrera pas ici. En effet, en posant $\mathcal{L}_i = \mathcal{L}(Y, \widehat{\beta_1}, \ldots, \widehat{\beta_{\iota-1}}, \beta_i, \widehat{\beta_{\iota+1}}, \ldots, \widehat{\beta_p})$ et $\mathcal{L}_{max} = \mathcal{L}(Y, \hat{\beta})$, nous obtenons que $2(\mathcal{L}_{max} - \mathcal{L}_i) \sim \chi^2(1)$, ce qui nous permet d'obtenir un intervalle de confiance pour β_i en posant $\chi^2_1(\alpha)$ le quantile d'ordre α d'une loi de Khi² à 1 degré de liberté ;

$$\boxed{IC_\alpha(\beta_i) = \left[\beta_i \; ; \; \mathcal{L}_i > \mathcal{L}_{max} - \frac{1}{2}\chi^2_1(1 - \frac{\alpha}{2})\right]}$$

Nous pouvons également obtenir des intervalles de confiances pour η_i d'une part et pour μ_i d'autre part en se rappelant que $\eta_i = X_{i.}{}^T \beta$ et $\mu_i = g^{-1}(\eta_i) = h(\eta_i)$ et en utilisant à deux reprises la méthode Delta. Nous exposons alors les intervalles de confiances obtenus ainsi, en laissant au soin du lecteur de vérifier l'obtention de ces derniers (aucune complexité) ;

$$\boxed{\begin{array}{c} IC_\alpha(\eta_i) = \left[\widehat{\eta_\iota} - \frac{\sigma_{\widehat{\eta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \; ; \; \widehat{\eta_\iota} + \frac{\sigma_{\widehat{\eta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \; avec \; \sigma^2_{\widehat{\eta_\iota}} = \phi X_{i.}(X^T W^* X)^{-1} X_{i.}{}^T \\ \\ IC_\alpha(\mu_i) = \left[\widehat{\mu_\iota} - h'(\widehat{\eta_\iota}) \frac{\sigma_{\widehat{\eta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \; ; \; \widehat{\mu_\iota} + h'(\widehat{\eta_\iota}) \frac{\sigma_{\widehat{\eta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right] \; avec \; \sigma^2_{\widehat{\eta_\iota}} = \phi X_{i.}(X^T W^* X)^{-1} X_{i.}{}^T \\ \\ IC_\alpha(\mu_i) = \left[h\left(\widehat{\eta_\iota} - \frac{\sigma_{\widehat{\eta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right) \; ; \; h\left(\widehat{\eta_\iota} + \frac{\sigma_{\widehat{\eta_\iota}}}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right)\right] \; dans \; certains \; logiciels \; (plus \; simple, moins \; rigoureux) \end{array}}$$

Nous faisons remarquer que ces intervalles de confiance sont approximatifs, car ils reposent sur des estimations de W. De plus, les intervalles pour μ_i sont des intervalles pour des prévisions, et non pas pour de nouvelles valeurs. Ceci nous renseigne uniquement sur la précision des prévisions du modèle.

1.2.5 Adéquation du modèle et tests de significativité.

Lorsque nous nous sommes intéressés à la régression linéaire, nous avons effectué des tests d'adéquation du modèle ou de significativité des variables à partir de la somme des carrés résiduels que nous avions noté SCR. En ce qui concerne les modèles linéaires généralisés, nous allons nous baser sur les résultats théoriques issus des tests de rapport de vraisemblances et de Pearson. Pour ce faire, nous allons définir ce que l'on appelle la déviance d'un modèle, ainsi que la statistique de Pearson.

Nous avons précédemment estimé le paramètre β par $\hat{\beta}$. Ceci nous permet d'exprimer la log-vraisemblance maximisée pour chaque observation, soit en fonction de $\hat{\theta}$, soit en fonction de $\widehat{\mu_i}$;

$$\phi \times l\left(Y_i, \hat{\beta}, \phi\right) = Y_i\hat{\theta} - b(\hat{\theta}) + cte$$

$$\phi \times l\left(Y_i, \hat{\beta}, \phi\right) = Y_i(b')^{-1}(\widehat{\mu_i}) - b((b')^{-1}(\widehat{\mu_i})) + cte$$

Or nous savons que si le modèle était parfait, la prévision $\widehat{\mu_i}$ par le modèle correspondrait à Y_i (ou à la moyenne des Y_i dans l'hypothèse où l'on possède plusieurs observations telles que $X = X_i$). On parle alors de modèle saturé (autant de paramètres que d'observations distinctes), et nous pouvons en calculer la log-vraisemblance maximisée comme suit (dans le cas où l'on possède une seule observation $X = X_i$);

$$\phi \times l_{saturé}(Y_i) = Y_i(b')^{-1}(Y_i) - b((b')^{-1}(Y_i)) + cte$$

Nous définissons alors ce que l'on appelle la déviance, qui mesure l'écart entre la vraisemblance du modèle et celle du modèle saturé correspondant (ayant la même loi);

$$\boxed{D = 2\phi \sum_{i=1}^{n} \left(l_{saturé}(Y_i) - l\left(Y_i, \hat{\beta}, \phi\right)\right) \geq 0}$$

Nous pouvons remarquer que la déviance est deux fois un écart de vraisemblance multiplié par le paramètre de dispersion (ce qui permet d'éliminer le paramètre du terme de droite). Lorsque ce facteur multiplicatif n'est pas présent, soit $D^* = D/\phi$, on parle de déviance standardisée (qui tire son nom du fait que dans le cadre de la loi normale, $\phi = \sigma^2$). Il est à noter que dans certains cas, le paramètre de dispersion vaut 1, les deux types de déviances sont alors égaux, et l'on parle ainsi de test de déviance, quand le terme de test de déviance standardisée serait plus correct.

Lorsqu'un modèle est adéquat, il parait logique que ce dernier soit proche du modèle saturé, et donc que la déviance soit proche de 0. Inversement, plus la déviance d'un modèle sera importante, moins le modèle en question sera pertinent. Nous allons exploiter cette notion afin de tester la significativité d'un modèle. Nous posons ainsi comme hypothèse nulle de test H_0 : « Le modèle à p variables explicatives est significatif ». Nous pouvons alors nous apercevoir que la déviance peut s'écrire sous la forme d'un rapport de vraisemblance (ou d'une différence de log-vraisemblance), ce qui nous permet d'obtenir, sous les hypothèses techniques classiques et en exploitant les résultats théoriques issus des tests de rapport de vraisemblance (voir [57] p 459), que sous H_0 la déviance standardisée D^* suit asymptotiquement une loi du Khi² à n-p degrés de liberté (avec n le nombre d'observations distinctes et p le nombre de variables du modèle).

Nous pouvons ainsi mettre en place un test de significativité du modèle de niveau α en comparant la déviance standardisée observée avec le quantile d'ordre 1- α d'une loi du Khi² à n-p degrés de liberté. Nous considérerons ainsi que le modèle est significatif lorsque la déviance standardisée observée est inférieure à ce quantile.

Il faut cependant faire remarquer que la validité de ce test n'est qu'asymptotique, il faut donc être prudent quant aux conclusions à en tirer. De plus, lorsque les données sont binaires, la déviance standardisée ne suit pas une loi du Khi², et le test de déviance n'est alors plus valide. Dans cette hypothèse, nous pouvons utiliser le test d'Hosmer-Lemershow. Ce dernier consiste à ordonner par ordre croissant les $\widehat{\mu_i}$, puis à les scinder en « g » groupes (en général g=10). Il reste alors à mener un test en tout point semblable à celui de la déviance, mais portant sur une statistique suivant approximativement une loi du Khi² à g degrés de liberté (validée par simulation, voir [21]) ;

$$C^2 = \sum_{k=1}^{g} \frac{\left(\sum_{i=1}^{c_k} y_i - m_k^* \overline{\mu_k}\right)^2}{m_k^* \overline{\mu_k}(1-\overline{\mu_k})}$$

$avec\ m_k^*\ (c_k)\ le\ nombre\ d'observations\ (distinctes)\ dans\ le\ k^{ième}\ groupe,$

$\overline{\mu_k} = \sum_{i=1}^{c_k} \frac{m_i}{m_k^*} \widehat{\mu_i}\ avec\ m_i\ le\ nombre\ d'observations\ telles\ que\ X = X_{i.}\ dans\ le\ k^{ième}\ groupe.$

Nous définissons de plus la statistique de Pearson, parfois nommée le Khi² de Pearson généralisé, qui est utilisée, au même titre que la déviance, comme mesure de qualité globale d'ajustement d'un modèle linéaire généralisé ;

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu_i})^2}{\widehat{Var(Y_i)}}$$

Lorsque la loi choisie est la loi normale avec le lien identité, cette statistique correspond à la somme des carrés des résidus, noté SCR dans la régression linéaire. Cette statistique, tout comme la déviance (excepté dans le cas de la loi normale), suit asymptotiquement une loi du Khi² à (n-p) degrés de liberté, ce qui fournit un second moyen d'obtenir une indication sur la qualité du modèle.

De plus cette statistique fournit un estimateur du paramètre de dispersion, plus simple que celui que l'on pourrait éventuellement déterminer avec la méthode du maximum de vraisemblance (et conseillé par McCullagh et Nelder [49]). Nous pouvons ainsi estimer simplement le paramètre de dispersion, que nous avions laissé de coté et supposé connu. En pratique, toutes les estimations et la construction du modèle seront établies en supposant le paramètre de dispersion connu, puis une fois le modèle construit, une estimation de ce paramètre sera établie à partir de la statistique de Pearson, afin de réintégrer la valeur de cette estimation (supposé une fois de plus fixée et connue) dans le calcul des intervalles de confiance, afin de les affiner.

Une fois le modèle reconnu comme significatif, il est nécessaire de pouvoir comparer deux modèles afin de sélectionner le meilleur d'entre eux. De la même façon que pour la régression linéaire avec la statistique $\widehat{F}$, nous allons mettre en place un test de comparaison de modèle qui nous servira également de test de significativité de variables, mais à partir de la déviance ici.

En effet, pour comparer deux modèles emboités, il suffit de comparer leurs déviances. En notant D_1 et D_2 les déviances respectives d'un modèle (1) et d'un sous modèle (2), nous pouvons nous apercevoir que $\Delta D = D_2 - D_1$ correspond en réalité à un rapport de vraisemblance (grâce aux propriétés de la fonction logarithme), ce qui nous permet d'utiliser la fait que sous certaines hypothèses classiques, la différence de déviance suit asymptotiquement une loi du Khi² à $p_1 - p_2$ degrés de liberté avec p_2 le nombre de paramètres du sous modèle et p_1 le nombre de paramètres du modèle (Tout se passe comme si le modèle D_2 était le modèle saturé et le modèle D_1 le modèle à tester). Si la différence observée est inférieure au quantile du Khi² associé, alors on retiendra le modèle initial. En revanche si la différence observée est supérieure au quantile, alors on préférera le sous modèle au modèle complet. De la même façon qu'en régression linéaire, il nous est alors possible de construire différents tests de significativité des variables en confrontant un modèle complet à un sous modèle (comportant une seule variable en moins). Nous renvoyons le lecteur à la *page 80* pour les notions de tests de Type I, II ou III, de méthodes de construction de modèle en découlant et de significativité autour des variables qualitatives.

Comme nous l'avons fait remarquer en régression linéaire, il existe d'autres indicateurs de qualité des modèles. Plus particulièrement, lorsque nous cherchons à comparer des modèles qui ne sont pas emboités (loi différente ou fonction de lien différente par exemple), il est possible d'utiliser les critères dit d'AIC pour « Akaïke Informative Criterion » ou de BIC pour « Bayesian Informative Criterion ». La philosophie de ces critères repose sur le fait que plus la vraisemblance du modèle est grande, plus la log-vraisemblance noté $\mathcal{L}$ est grande, et par conséquent meilleur est le modèle. De plus, et afin de faire en sorte que le rajout de variables explicatives ne signifie pas nécessairement une augmentation du critère, et pour obtenir un modèle de taille raisonnable, ces critères contiennent un facteur pénalisant fonction du nombre de paramètres p. Nous représentons ci-dessous l'expression de ces deux indicateurs, en apportant l'attention sur le fait que certains logiciels utilisent –AIC ou –BIC (le sens d'optimisation pouvant facilement être repéré en regardant le sens de variation d'un modèle « très mauvais » sans variable explicative à un autre avec une ou plusieurs variables explicatives).

$$AIC = -2\mathcal{L} + 2p$$

$$BIC = -2\mathcal{L} + plog(n) \; avec \; n \; le \; nombre \; d'observations$$

Un dernier principe souvent abordé en modèle linéaire généralisé dans un but de comparaison de modèle est le principe d'apprentissage et de validation. Ceci consiste simplement à séparer les données initiales en deux groupes de données. Le premier groupe de données servira alors à la modélisation à partir des différents modèles que l'on souhaite comparer. Ensuite, nous effectuons des prévisions à partir des variables explicatives du second groupe de données en fonction des coefficients de chaque modèle puis nous mesurons l'écart entre les observations réelles du second groupes et les prévisions des différents modèles. Le modèle pour lequel cette mesure est la plus faible sera alors considéré comme étant le plus pertinent. La mesure la plus connue est le PRESS dont l'expression est la suivante (avec Y^b les observations de la variable à expliquer pour le second groupe);

$$PRESS = \left\| \widehat{Y^b} - Y^b \right\|$$

Bien entendu, la réalisation d'un tel procédé sous entend une grande quantité de données. Pour ce qui est du découpage de la base de donnée, on considère généralement que ¾ des données va à l'apprentissage et que ¼ va à la validation. Dans l'hypothèse où les données sont en nombre insuffisant pour réaliser une telle comparaison entre modèles, il est possible d'utiliser ce que nous appelons la validation croisée. Cette dernière validation consiste à découper le jeu de données en « d » parties, puis à appliquer le principe d'apprentissage et de validation pour chaque partie, en prenant la partie en question comme jeu de validation et les autres comme jeu d'apprentissage. Ceci nous permet alors de calculer un PRESS pour chaque type de modélisation et pour chaque partie (on calcule « d » PRESS par modèle à comparer). Le critère à minimiser est alors la somme des PRESS de chaque partie.

$$PRESS_{CV} = \sum_{i=1}^{d} PRESS_i$$

*1.2.6 **Analyse des résidus et détection d'erreurs.***

Comme précédemment, nous allons devoir analyser les résidus afin de vérifier la cohérence du modèle. Pour ce faire, nous pouvons analyser les résidus estimés définis de la même façon qu'en régression linéaire comme suit ;

$$\widehat{\varepsilon_i} = Y_i - \widehat{\mu_i}$$

Ces résidus parfois appelés résidus bruts sont difficiles à analyser, les résidus de Pearson que nous allons définir ci-dessous ou les résidus de Pearson standardisés (ayant approximativement la même variance) leurs sont souvent préférés ;

$Résidus\ de\ Pearson :$ $$r_i = \frac{\widehat{\varepsilon_i}}{\sqrt{\widehat{Var(Y_i)}}}$$

$$r_i^* = \frac{\widehat{\varepsilon_i}}{\sqrt{(1-h_{ii})\widehat{Var(Y_i)}}}$$

$avec\ h_{ii}\ élément\ diagonal\ de\ la\ hat\ matrix\ H = X(X^T W^* X)^{-1} X^T W^*$

(voir définition *page 81* appliquée à la régression pondérée)

Nous définissons de plus les résidus de déviance (standardisés ou non) qui sont avec les résidus de Pearson les résidus les plus souvent conseillés.

$Résidus\ de\ Déviance :$ $$d_i = signe(\widehat{\varepsilon_i}) \times \sqrt{d_i}$$

$$d_i^* = signe(\widehat{\varepsilon_i}) \times \sqrt{\frac{d_i}{(1-h_{ii})}}$$

$$avec\ D^* = \sum_{i=1}^{n} 2\left(l_{saturé}(Y_i) - l\left(Y_i, \hat{\beta}, \phi\right)\right) = \sum_{i=1}^{n} d_i$$

Nous pouvons alors mener les mêmes analyses graphiques des résidus que celles entrevues au sujet de la régression linéaire. Nous définissons cependant ci-dessous ce que l'on appelle des résidus partiels, définis pour chaque prédicteur, soit pour chaque colonne de X. Un tracé d'une des variables explicatives en abscisse et du résidu partiel en ordonnée est alors une source d'information. En effet, si le tracé est linéaire, ceci signifie que tout est normal, mais si une tendance non linéaire se dégage, alors il faut remplacer la variable par une fonction de celle-ci donnant la même tendance que celle observée.

$$\widehat{\varepsilon_j^P} = \widehat{\varepsilon_\iota} g'(\mu_i) + \widehat{\beta_J} X_j$$

De la même façon que pour la régression linéaire, nous pouvons mesurer les points leviers à partir de la matrice chapeau issue de la dernière régression opérée dans l'algorithme de maximisation. Nous pouvons également mesurer l'influence des observations à partir de la distance de Cook, qui s'écrit alors ;

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{Y_J^{(i)}} - \widehat{Y_J} \right)^2}{p\widehat{\sigma^2}} = \frac{1}{p} \frac{h_{ii}}{(1 - h_{ii})} (r_i^*)^2$$

Chapitre 2 : Modélisation du risque en assurance automobile.

Nous avons entrevu précédemment la théorie et la méthodologie à suivre dans le cadre des modèles linéaires généralisés. Nous allons à présent nous intéresser plus en détails à la modélisation du risque automobile. Nous évoquerons ainsi dans un premier temps les notions de coût moyen et de fréquence, qu'il est d'usage de modéliser séparément sous l'hypothèse à contrôler d'indépendance de ces deux facteurs. Puis nous nous intéresserons plus particulièrement à la régression de poisson, et aux phénomènes de sur dispersion. Enfin, nous reviendrons sur l'objectif de la modélisation, à savoir l'analyse des segments sur ou sous tarifés et une éventuelle évolution des tarifs.

2.1 Fréquence et coût moyen.

Comme nous l'avons fait remarquer plus haut, il est d'usage en assurance automobile de modéliser séparément le coût moyen de sinistre et la fréquence de sinistre. La prime pure est ensuite calculée en multipliant le coût moyen par la fréquence. L'hypothèse sous jacente à cette méthodologie est l'indépendance entre ces deux notions. Cette indépendance est en règle générale admise, mais il est tout de même préférable de la contrôler. Pour ce faire, nous pourront utiliser un test d'indépendance basé sur le coefficient de corrélation de Pearson, sur le Tau de Kendall ou encore sur le Rhô de Spearman, que nous ne détailleront pas ici (voir [60] p 495-501).

En règle générale, les montants de sinistres seront modélisés à partir d'une loi Gamma. En effet, ces derniers correspondent bien à une distribution continue, définie sur les réels positifs, et ayant une variabilité qui augmente avec la moyenne. En pratique, nous observerons souvent une distinction entre les sinistres matériels et les sinistres corporels, l'échelle de valeurs associée à ces deux types de sinistres étant trop différente d'une catégorie à l'autre. De même, la modélisation des sinistres dits « graves », c'est-à-dire des sinistres d'un montant exceptionnel, fait souvent l'objet d'une attention particulière et utilise la théorie des valeurs extrêmes, qui ne sera pas abordée ici (voir [17], [18] ou [35]). Dans le cadre de cette étude, les sinistres dits graves feront l'objet d'une modélisation spécifique, d'une part de leur montant moyen (avec une loi gamma), et d'autre part de leur probabilité d'occurrence (avec une régression logistique).

En ce qui concerne la fréquence de sinistres, elle fait en règle générale l'objet d'une modélisation semblable à une régression de Poisson. En effet, nous observons bien un processus de comptage, pour lequel nous désirons modéliser une proportion (la fréquence). Cependant, la modélisation de la fréquence est généralement relativement complexe, d'une part parce que le nombre d'observations sans sinistres est très important, et d'autre part car l'hypothèse (sous jacente à l'utilisation d'une loi de poisson) selon laquelle la variance est égale à la moyenne est rarement vérifiée. Dans ce dernier cas, on parle alors soit de sous dispersion, soit de sur dispersion (voir paragraphe suivant), et l'on est souvent amené à effectuer des modélisations plus complexes afin de corriger ce phénomène.

2.2 Régression de poisson, sur dispersion et solutions.

2.2.1 *Régression de Poisson.*

En notant Y la variable à expliquer et X les variables explicatives, nous cherchons à maximiser la log-vraisemblance que l'on peut écrire facilement. La loi conditionnelle de Y sachant l'observation i étant une loi de poisson (de paramètre μ_i), nous pouvons écrire ;

$$\mathcal{L}(\beta) = ln\left(\prod_{i=1}^{n} exp(-\mu_i)\frac{{\mu_i}^{Y_i}}{Y_i!}\right) = \sum_{i=1}^{n} ln[-\mu_i - log(Y_i!) + Y_i ln(\mu_i)]$$

Or nous savons que dans le cadre des modèles linéaires généralisés, nous avons la relation $g(\mu) = \eta(X)$ avec g() la fonction de lien. En choisissant la fonction de lien canonique pour la régression de poisson, nous savons ainsi que $log(\mu_i) = \eta(X_i)$. Par concavité de la fonction de vraisemblance en β, il suffit alors de regarder les dérivés du premier ordre, qui s'écrivent ;

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij}(Y_i - \mu_i)$$

Ceci nous permet alors de remarquer que si l'on considère une catégorie de risque définie par l'occurrence d'une variable qualitative, alors le nombre de sinistres observés associés à ce niveau de risque est égal à son homologue théorique. En effet, si l'on ne sélectionne qu'une catégorie de risque particulière (les hommes par exemples), alors nous avons la relation ;

$$\boxed{\sum_{Niveau\ de\ risque} Y_i = \sum_{Niveau\ de\ risque} \widehat{\mu_i}}$$

Ceci nous indique donc que les « primes fréquences » attribuées aux différentes catégories de risque compensent exactement le nombre de sinistres (en supposant les montants de sinistres égaux à 1). De plus, le modèle reconstitue sans erreur le nombre total de sinistres observés (pour autant qu'il existe un intercept).

Nous ferons alors remarquer qu'en règle générale, nous possédons une information concernant la durée d'exposition au risque liée à chaque observation. Bien entendu, cette information est importante et joue un rôle capital dans la modélisation de la fréquence. Nous utiliserons cette information, mais sans estimer de coefficient associé (que nous fixons alors à 1). En notant t_i la durée d'observation associée à la i^{ème} observation, nous utiliserons alors la relation suivante ;

$$\boxed{ln(\mu_i) = \eta(X_i) + \ln(t_i)}$$

Ainsi, lorsque nous désirerons obtenir une fréquence annuelle associée à une observation, il nous suffira de calculer non pas $\widehat{\mu_i} = t_i \exp(\widehat{\eta_i})$ mais $\widehat{\mu_i} = 365 \times \exp(\widehat{\eta_i})$ en supposant ici que la durée d'exposition est exprimée en nombre de jours.

2.2.2 *Sur dispersion : définition, causes et détection.*

Comme nous l'avons vu précédemment, l'utilisation de la régression de poisson repose sur l'hypothèse forte d'égalité entre la variance et l'espérance de la variable à expliquer (on parle alors d'équidispersion). En pratique, cette équidispersion est rarement vérifiée, ce qui peut remettre en doute l'utilisation de la régression de poisson. Si la variance est supérieure à la moyenne, nous parlerons alors de sur dispersion, inversement si la variance est inférieure à la moyenne nous parlerons de sous dispersion.

Ce phénomène est généralement dû à l'omission de variables explicatives, pas toujours connues ou accessibles. Une interprétation simple de cette relation de cause à effet peut être mise en avant. Considérons deux classes de risque C_1 et C_2 de poids p_1 et p_2 sans effet de surdispersion ($\widehat{\sigma_1}^2 = \widehat{m_1}$ et $\widehat{\sigma_2}^2 = \widehat{m_2}$), mais que nous aurions omis de séparer. L'espérance de la classe $C_1 \cup C_2$ correspondrait à la somme pondérée de $\widehat{m_1}$ et $\widehat{m_2}$, tandis que la variance vaudrait ;

$$\widehat{\sigma^2} = p_1\widehat{\sigma_1}^2 + p_2\widehat{\sigma_2}^2 + p_1(\widehat{m_1} - \hat{m})^2 + p_2(\widehat{m_2} - \hat{m})^2 \geq \hat{m}$$

Nous constatons donc bien une surdispersion, l'égalité n'étant possible que dans l'hypothèse ou les classes de risque C_1 et C_2 ne sont pas différentiables et ainsi $\widehat{m_1} = \widehat{m_2} = \hat{m}$.

Il est ainsi possible de contrôler la présence d'une sur ou sous dispersion en représentant pour chaque classe de risque la variance empirique en fonction de la moyenne empirique. Si les points sont autour de la première bissectrice, nous pourront alors valider l'hypothèse d'équidispersion. Dans le cas contraire, le phénomène de sur dispersion sera celui le plus observé en pratique, et mis en évidence par une variance plus élevée que la moyenne, particulièrement pour les moyennes élevées en règle générale.

La présence de sur ou sous dispersion dans les données, relativement à un modèle de poisson, peut également être mise en évidence en estimant un paramètre de sur dispersion. Une estimation de ce paramètre est donnée en effectuant le rapport du Khi² généralisé de Pearson sur n-p, avec n le nombre d'observations et p le nombre de variables explicatives. Cette estimation peut également être effectuée en faisant le rapport de la déviance sur n-p. Dans l'hypothèse ou ce rapport est proche de 1, l'hypothèse d'équidispersion peut être retenue. A l'inverse, si ce rapport est supérieur à 1 (respectivement inférieur à 1), nous sommes en présence de sur dispersion (respectivement de sous dispersion).

Il est enfin possible de détecter une sur ou sous dispersion en considérant que les estimations des espérances sont des estimations des moyennes et les résidus au carré des estimations de la variance. Ainsi la représentation du rapport des résidus au carré sur les estimations des espérances, en fonction des estimations des espérances, ne doit laisser paraitre aucune tendance sous l'hypothèse d'équidispersion et une droite de régression de ce ratio devrait correspondre à une droite horizontale d'ordonnée 1. En pratique, cette analyse graphique permet d'obtenir une indication quant à la modélisation à mettre en œuvre en cas de sur ou sous dispersion. Le lecteur intéressé est invité à se rapporter aux ouvrages d'Allain et Brenac [16] ou de Hauer et al. [39] pour de plus amples informations.

2.2.3 ***Solutions d'améliorations.***

2.2.3.1 Approche préalable, la quasivraisemblance.

Avant de mettre en avant les possibilités d'améliorations en présence de sur dispersion, nous allons définir et mettre en avant l'utilité de ce que l'on appelle la quasivraisemblance. Nous nous plaçons dans le cadre classique des modèles linéaires généralisés, avec Y une variable à expliquer, X une matrice des variables explicatives, β un vecteur de coefficients à estimer, ϕ un paramètre de dispersion et g() une fonction de lien. Nous définissons alors la quasivraisemblance comme suit, en considérant y un vecteur d'observations iid de moyenne μ et de fonction de variance V(μ) ;

$$Q(y,\mu) = \int_y^{\mu} \frac{y-t}{\phi V(t)} dt$$

Nous pouvons alors vérifier que cette fonction possède trois propriétés communes avec la log-vraisemblance d'une loi de la famille exponentielle utilisée en GLM, à savoir ;

$$E\left(\frac{\partial Q}{\partial \mu}\right) = 0$$
$$E\left(\frac{\partial^2 Q}{\partial \mu^2}\right) = -\frac{1}{\phi V(\mu)}$$
$$Var\left(\frac{\partial Q}{\partial \mu}\right) = \frac{1}{\phi V(\mu)} = -E\left(\frac{\partial^2 Q}{\partial \mu^2}\right)$$

Ces propriétés dont la démonstration sera laissée aux soins de l'assuré (aucune complexité majeure), correspondent aux propriétés de la log-vraisemblance utilisées pour l'estimation des paramètres β, ainsi que pour les notions de convergence et de normalité asymptotique. Nous pouvons en effet chercher à maximiser la quasivraisemblance par rapport à β en calculant en une observation y_i la dérivée partielle par rapport à β;

$$\frac{\partial Q(\mu_i, y_i)}{\partial \beta} = \frac{\partial Q}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta} = \frac{Y_i - \mu_i}{\phi} \times {g^{-1}}'(\mu_i) \times {X_{i.}}^T$$

Nous retrouvons alors exactement le même résultat que pour la maximisation de log-vraisemblance *page 92*. De plus la variance de $\frac{\partial Q(\mu_i, y_i)}{\partial \beta}$ qui est égale à l'espérance de la dérivée seconde se calcule de la même façon que pour la log-vraisemblance et donne le même résultat. Il est alors possible de mettre en place un algorithme de maximisation de Fisher scoring qui conduit exactement à ce que nous avons appelé IRLS plus en avant. La normalité asymptotique de l'estimation par « quasi-maximum de vraisemblance » peut alors être montrée et l'on obtient ainsi exactement les mêmes résultats qu'avec la maximisation de la log-vraisemblance. L'utilisation de la quasivraisemblance nous permet ainsi de rester dans le cadre des modèles linéaires généralisés avec des résultats identiques, tout en fixant uniquement des hypothèses sur l'indépendance des observations ainsi que sur leurs deux premiers moments, alors qu'il était nécessaire de fixer une hypothèse concernant la loi toute entière de la variable à expliquer auparavant.

Notons enfin que par analogie, il est possible de définir la quasidéviance pour une observation y d'espérance μ comme $d(y,\mu) = -2\phi Q(y,\mu)$, la quasidéviance du modèle saturé étant nulle (pour plus d'informations sur la quasivraisemblance, voir [41], [49] ou [63]).

2.2.3.2 Le modèle quasi-poisson et (quasi) binomial négatif.

Suite à ce préliminaire concernant la quasivraisemblance, nous pouvons comprendre qu'une première solution permettant de palier à la sur dispersion réside dans la mise en place d'un modèle linéaire généralisé basé sur une loi proche de la loi de poisson, mais possédant une variance supérieure à la moyenne. Ainsi, nous allons considérer ce que l'on appelle un modèle de quasi-poisson en introduisant un paramètre ϕ (de sur dispersion) tel que la variance des observations de la variable à expliquer correspond à l'espérance multipliée par ce paramètre, que l'on estimera à partir du Khi² généralisé de Pearson.

Nous allons alors écrire la quasivraisemblance d'un tel modèle, afin de mettre en évidence le fait que l'estimation des coefficients β sera parfaitement identique à celle effectuée pour la loi de poisson.

$$Q(y_i, \mu_i) = \int_y^{\mu} \frac{y - t}{\phi t} dt = \frac{1}{\phi}(y_i log(\mu_i) - \mu_i)$$

Or, la vraisemblance pour une observation dans le modèle de poisson correspond à $y_i log(\mu_i) - \mu_i - \log(y_i!)$. La maximisation de la vraisemblance dans le modèle de poisson et la maximisation de la quasivraisemblance dans le modèle de quasi-poisson conduisent donc exactement aux mêmes estimations des coefficients β.

Il en est cependant différemment pour les écarts types, qui se verront attribuer un facteur $\phi^{1/2}$ au passage du modèle de poisson au modèle de quasi-poisson, car ces derniers dépendent de la variance du modèle et seront donc différents dans le modèle de poisson. Les intervalles de confiance et la significativité des variables pourront ainsi varier entre le modèle de poisson et le modèle de quasi-poisson. La mise en place d'un modèle de quasi-poisson, plus pertinent que le modèle de poisson en cas de sur ou sous dispersion, pourra donc amener à considérer comme non significatives certaines variables sensée être pertinentes aux vues du modèle de poisson.

Une seconde solution permettant de prendre en compte un phénomène de sur ou sous dispersion est l'utilisation d'un modèle linéaire généralisé classique à partir de la loi binomiale négative, que nous n'avons pas encore abordé mais qui est souvent utilisée pour la modélisation de fréquences. Cette loi est en règle générale exprimée en fonction de deux paramètres $(r, p) \in \mathbb{R}^+ \times [0; 1]$, de loi, espérance et variance indiquées ci-dessous ;

$$P(Y = y \,|X = x) = \frac{\Gamma(y + r)}{\Gamma(r)\Gamma(y + 1)} \times p^r \times (1 - p)^y \,, Espérance : r\frac{1 - p}{p}, Variance : r\frac{1 - p}{p^2}$$

En pratique dans les modèles linéaires généralisés, nous effectuons un changement de paramètre qui permet de faire en sorte que l'espérance corresponde à un paramètre μ. Nous réécrivons alors la loi, l'espérance et la variance dans ce cadre particulier ;

$$P(Y = y \,|X = x) = \frac{\Gamma(y + r)}{\Gamma(r)\Gamma(y + 1)} \times \left(\frac{r}{r + \mu}\right)^r \times \left(\frac{\mu}{r + \mu}\right)^y \,, Espérance : \mu \,, Variance : \mu + \frac{1}{r}\mu^2$$

Dans ce cas, la variance est bien différente de l'espérance. Si r est négatif, nous corrigeons la sous dispersion, si r est positif, nous corrigeons la sur dispersion et si le terme 1/r est égal à zéro, nous retombons sur le modèle de poisson. En pratique, les estimations des espérances et des paramètres seront légèrement différentes entre le modèle de poisson et le modèle binomiale négatif mais les écarts types seront notablement augmentés. La première introduction de la loi binomiale négative remonte à Blaise Pascal en 1679, elle fut ensuite assimilée à un mélange de loi Gamma-Poisson et de nombreuses écritures et applications de cette loi existent aujourd'hui, nous invitons ainsi le lecteur intéressé à se rapporter à des ouvrages de A. Cameron et P.K. Trivedi [19] ou encore de M. Greenwood et G.U. Yule [38].

Malgré l'utilisation d'une loi binomiale négative, il est possible qu'une sur ou sous dispersion résiduelle persiste. De la même façon que pour la loi de poisson, il est alors possible de mettre en place un modèle quasi-binomiale négatif, en tout point semblable à un modèle binomiale négatif, mais contenant un paramètre de sur dispersion correctif ϕ tel que la variance vaut $\phi\left(\mu + \frac{1}{r}\mu^2\right)$.

2.2.3.3 Les modèles « Zero Inflated » poisson et binomial négatif.

En dépit des solutions mises en avant ci-dessus, il est possible qu'un phénomène de sur ou sous dispersion demeure. Les insuffisances des modèles précités peuvent alors provenir de l'absence de variables ayant un poids important dans la détermination de la fréquence de sinistres. Une autre explication peut résider dans l'inadéquation de certaines hypothèses relatives aux modèles. En effet, au regard de la réalité des données, il est possible que la population des assurés n'ayant déclaré aucun sinistre soit scindée en deux sous populations.

La première population pourrait représenter des individus participant à l'expérience et déclarant un sinistre dès lors que celui-ci survient. A contrario, la deuxième population serait composée d'individus ne déclarant pas automatiquement chaque sinistre survenu. En effet, certains assurés sont amenés à ne pas déclarer des sinistres de faible gravité afin d'éviter un malus et donc une augmentation de leur prime d'assurance (le terme de « Bonus Hunger » est souvent employé). Les pénalités étant indépendantes du montant de sinistre, un assuré responsable a tout intérêt à dédommager lui-même la partie adverse pour les « petits » sinistres. Un deuxième cas de figure concerne les conducteurs qui ne s'arrêtent pas pour faire un constat lors d'un accrochage avec une voiture en stationnement.

Ainsi, pour la première population, la valeur zéro indique que le conducteur n'a pas eu un seul sinistre. Il n'en est pas de même pour la deuxième population, qui malgré l'absence de déclaration de sinistre, est porteuse d'un risque certain. En effet, il peut paraitre justifié de penser que l'absence de déclaration de sinistre dans ce cas est principalement liée à la non gravité du sinistre et non à la réalité objective du risque.

Les modèles précités ne permettent pas de distinguer ces deux populations. Nous allons donc introduire le modèle de poisson à inflation de zéro ou ZIP (Zero Inflated Poisson) et le modèle binomial négatif à inflation de zéro ou ZINB (Zero Inflated negative Binomial), développés par Lambert [46] puis par Greene [37].

Ces deux modèles, contrairement aux précédents, ne supposent pas que les valeurs nulles et les valeurs positives strictement sont issues du même modèle. Cette approche va en effet générer deux modèles séparés, puis les combiner. La variable à expliquer Y est ici considérée comme le produit de deux variables aléatoires B et Y^*. B peut alors prendre les valeurs 0 ou 1 et sera modélisée par une régression logistique tandis que Y^* correspond au modèle de poisson (ou au modèle binomial négatif) pour prédire la valeur de Y associée à des assurés ayant déclaré un sinistre. Les modèles ZIP et ZINB comportent donc deux parties : celle relative au modèle de comptage (qui rend compte de la sinistralité des assurés appartenant à la population déclarant tous les sinistres) et celle relative à l'inflation de zéro (qui rend compte de la probabilité de non déclaration de sinistre et d'appartenance à la seconde population). Le modèle à inflation de zéro correspond alors à un mélange d'une loi de poisson (ou d'une loi binomiale négative) et d'une loi de Dirac en 0, avec les expressions de probabilités suivantes ;

ZIP	**ZINB**
$\begin{cases} P(Y=0 \mid X=X_i) = p_i + (1-p_i)e^{-\lambda_i} \\ P(Y=y_i \mid X=X_i) = (1-p_i)\dfrac{{\lambda_i}^{y_i}}{y_i!}e^{-\lambda_i} \end{cases}$ $p_i = P(B=0)$ $\lambda_i \; paramètre \; de \; la \; loi \; de \; poisson$	$P(Y=y_i \mid X=X_i) = p_i 1_{\{y_i=0\}} + (1-p_i)\dfrac{\Gamma(y_i+r)}{\Gamma(r)\Gamma(y_i+1)} \times \left(\dfrac{r}{r+\lambda_i}\right)^r \times \left(\dfrac{\lambda_i}{r+\lambda_i}\right)^{y_i}$ $p_i = P(B=0)$ $(r, \lambda_i) \; paramètres \; de \; la \; loi \; binomiale \; négative$

Illustration 39 : Loi de probabilité pour les modèles ZIP et ZINB

En pratique, ce type de modèle est ajusté de façon itérative (voir Lambert [46]). En premier lieu, nous considérons que l'absence de sinistre est synonyme d'appartenance à la seconde population (inflation de zéro) et nous réalisons ainsi une régression logistique sur les observations sans sinistres et une régression de poisson sur les observations avec sinistres. Cette première étape nous fournit ainsi des estimateurs des p_i et des λ_i. Dans un deuxième temps, nous estimons la probabilité d'appartenance à la seconde population conditionnellement aux valeurs observées y_i à partir des estimations de l'étape précédente et en utilisant le théorème de Bayes ($P(A|B) \times P(B) = P(B|A) \times P(A)$). Les estimations obtenues nous permettent ainsi dans un troisième temps de reconsidérer les observations supposées appartenir à la seconde population (Si la probabilité d'appartenance est supérieure à 0,5 par exemple). La seconde population correspondant initialement à toutes les observations sans sinistres est ainsi modifiée (de même que la population des individus déclarant tous leurs sinistres). Nous réalisons ainsi une nouvelle régression logistique et une nouvelle régression de poisson sur les deux nouvelles populations obtenues. Nous réitérons ainsi les deux dernières étapes jusqu'à ce que les estimateurs des p_i et des λ_i convergent.

Cette approche mise en place par Lambert correspond à un algorithme EM (Estimation-Maximisation). Cette dernière possède l'avantage d'être aisée à mettre en place, mais la méthode employée par Greene [37] qui constitue une approche différente est plus rapide. Nous invitons ainsi le lecteur intéressé à se reporter à un de ses ouvrages pour de plus amples informations sur le sujet ([37] p11-13).

Afin de savoir si la mise en place d'un modèle ZIP ou ZINB est pertinente, une méthode simple peut être de comparer le nombre de valeurs nulles effectives avec la probabilité de non sinistralité fonction des paramètres estimés par le modèle de poisson ou le modèle binomial négatif. Une trop grande présence de zéro en excès sera alors révélatrice de la nécessité de mettre en place un modèle ZIP ou ZINB. Il est cependant à noter qu'une faible présence de zéro en excès ne rend pas illégitime l'utilisation d'un modèle à inflation de zéro.

Il existe tout de même un test statistique permettant de comparer un modèle poisson (respectivement binomial négatif) à son homologue à inflation de zéro. Il s'agit du test de Vuong, qui a proposé une statistique permettant de tester des modèles non emboités dans le cadre de l'estimation par maximum de vraisemblance. Ce test repose sur la comparaison de la probabilité d'observer y_i étant donné une loi ZIP (ou ZINB) à la probabilité d'observer y_i étant donné un modèle de poisson (ou un modèle binomial négatif). Nous inscrivons ci-dessous la statistique de Vuong employée pour tester un modèle à inflation de zéro contre son homologue classique ;

$$V = \frac{\sqrt{n}\overline{m}}{s_m}$$

$avec\ m_i = \log\left(\frac{f_1(y_i)}{f_2(y_i)}\right)$

$f_1\ densité\ du\ modèle$ à $inflation\ de\ zéro, f_2\ densité\ de\ son\ homologue$

$n\ nombre\ d'observations, \overline{m}\ et\ {s_m}^2\ moyenne\ et\ variance\ empirique\ des\ m_i$

Vuong montre alors que cette statistique suit une loi de Student et peut être approximée asymptotiquement par une loi normale centrée réduite (voir [62]). Si les deux modèles sont équivalents, la valeur moyenne du rapport de vraisemblance devrait être égale à zéro. Si le modèle à inflation de zéro est préférable, la valeur moyenne du rapport de vraisemblance devrait alors être sensiblement plus grande que zéro et inversement si le modèle classique est préférable, ce rapport devrait être inférieur à zéro de façon significative. Nous pouvons ainsi noter, en utilisant un seuil de 5% que ;

$|V| < 1{,}96 \Rightarrow Les\ deux\ modèles\ sont\ équivalents.$

$V > 1{,}96 \Rightarrow Le\ modèle$ à $inflation\ de\ zéro\ doit$ être $préféré.$

$V < -1{,}96 \Rightarrow Le\ modèle\ classique\ (Poisson\ ou\ Binomial\ négatif)doit$ être $préféré.$

Il est enfin à noter que lorsque le nombre de variables explicatives divergent d'un modèle à l'autre, les logarithmes de vraisemblance utilisés dans le test de Vuong sont affectés. Vuong propose alors d'employer une correction fonction du nombre de degré de liberté correspondant au critère AIC.

Pour de plus amples informations sur le test de Vuong ou sur les modèles à inflation de zéro de manière générale, le lecteur intéressé est invité à prendre connaissance des ouvrages de W.H. Greene [37], D. Lambert [46], V. Shankar et al. [58], Q.H. Vuong [62] et bien d'autres ([61], [64], [65]).

Chapitre 3 : Un exemple de modélisation.

Les aspects théoriques ayant été abordés, nous allons à présent mettre en avant un exemple de modélisation. Pour des raisons de clarté et de confidentialité, l'intégralité des résultats ne sera pas étudiée, mais seulement les points principaux utiles à la compréhension. Dans un premier temps, nous présenterons brièvement le jeu de données, ainsi que la méthodologie de construction de la modélisation qui sera retenue et les objectifs de cette dernière. Puis nous nous intéresserons plus en détails à la modélisation des montants de sinistres, des sinistres dits « graves » et de la fréquence de sinistres. Enfin nous analyserons d'éventuels segments de tarification sur ou sous tarifés en émettant la possibilité d'une évolution des tarifs en vigueurs. Nous faisons remarquer que l'ensemble des résultats mis en avant ici ont été obtenus à partir du logiciel R (dont nous décrirons rapidement les principales fonctions utilisées) et d'Excel.

3.1 Présentation des données, méthodologie et enjeux.

Cet exemple porte sur le portefeuille d'assurance automobile 2008 de « Mutant Assurances ». Comme nous l'avons vu précédemment dans un exemple d'extraction de données (voir *pages 53-57*), nous disposons d'une table reprenant l'ensemble des contrats justifiant d'au moins un jour de garantie durant l'année 2008, d'une table contenant l'ensemble des sinistres survenus en 2008, et d'une table contenant des informations sur les conducteurs assurés en 2008 (qui s'est malheureusement révélée inutilisable). Le portefeuille étudié recense 122 859 souscripteurs, pour 139 432 contrats et 219 335 avenants. L'exposition réelle cumulée de l'ensemble des polices d'assurances correspond à 106 675 contrats acquis (soit 106 675 années d'assurances). Le nombre de sinistres déclarés est de 17 870, mais il est à noter que seulement 14 405 sont clos. Parmi ces sinistres, 16 392 (91,7 %) sont purement matériels, et 1 478 (8,3 %) comprennent une composante corporelle.

Lors des différentes modélisations, la variable à expliquer ne sera pas toujours la même (montant ou fréquence par exemple), les variables explicatives seront cependant identique d'une modélisation à une autre et nous les décrivons à travers le tableau ci-dessous ;

Code Variable	Description
chaChoixTarif	*Correspond au type de contrat, Eco (RC seule), Eco Intégrale (RC+Dommages), Eco Plus (Intermédiaire).*
Classe	*Classe ou valeur à neuf du véhicule (A à M +X,Y,Z). A désigne un véhicule dont la valeur à neuf est faible, inversement M désigne un véhicule dont la valeur à neuf est élevée. Y et Z sont assimilables à A. X correspond à des véhicules pas ou partiellement assurés (principalement des véhicules avec toit ouvrant ou des contrats suspendus).*
Usage	*Type d'usage du véhicule (1 à 4).Les usages 2 et 3 ne possèdent pas de différence au niveau de la tarification. L'usage 4 correspond à une utilisation professionnelle du véhicule au sein d'une entreprise.*
Zone	*Zone du véhicule (2,4,5,6,7,9). La zone 9 correspond à la zone 5/6, la zone 7 aux DOM. Plus la zone est élevée, plus le risque sous jacent est grand.*
AnneeDepart	*Date de première mise en circulation du véhicule.*
fltBonusAuto	*Coefficient de réduction majoration (CRM).*
intCoefTarifaire	*Coefficient tarifaire de réduction majoration supplémentaire selon certaines conditions.*
Groupe	*Groupe ou puissance du véhicule (5 à 18 +99). 99 combiné avec une classe X correspond normalement à des contrats suspendus. 5 correspond à un véhicule peu puissant, inversement 18 correspond à un véhicule puissant.*
Options	*Option de tarification du contrat (BON, CLE, ESS, REL, C07, C10, C12, C20). Les options commençant par un C caractérisent un assuré dont l'ancienneté d'assurance révèle la présence d'un sinistre. Les autres options révèlent l'absence de sinistres responsables durant l'ancienneté d'assurance (<1 pour Relais, 1 pour Essai, 2 pour Sélectif et 3 pour Bonus).*

Illustration 40 : Description des principales variables explicatives

Sous le logiciel R et afin d'éviter toute confusion possible, les occurrences de l'usage, semblables à des valeurs numériques, seront précédées d'un U, celles de la zone d'un Z, et celles du groupe d'un G. De plus, la variable AnneeDepart est transformée en variable qualitative (« Très vieux » pour les véhicules de plus de 18 ans, « Vieux » pour les véhicules ayant entre 13 et 17 ans, « Agé » pour les véhicules allant de 8 à 12 ans, « Récent » pour les véhicules de 3 à 7 ans et enfin « Très récent » pour les véhicules de moins de 2 ans). Nous effectuons ensuite des regroupements de catégories. Les classes Y et Z sont associées à la classe A, la classe D et la classe E sont regroupées de même que les classes F à M et la classe X sera considérée comme une valeur manquante. Les groupes 5 à 11 sont regroupés de même que les groupes 12 à 14 et les groupes 15 à 18, les groupes 99 étant considérés comme des valeurs manquantes.

Par la suite, nous allons effectuer différentes modélisations. Ces dernières reposent sur l'hypothèse d'indépendance du coût moyen et de la fréquence de sinistre, classiquement retenue. Nous vérifions cette dernière à partir d'un regroupement par classes de risques en utilisant la fonction « cor.test » de R permettant de calculer le coefficient de corrélation de Pearson (0,0023 ici) , le Tau de Kendall (0,033 ici) ou le rho de Spearman (0,050 ici) d'un couple de variables. Premièrement nous allons modéliser les montants de sinistres standards, en prenant le soin d'effectuer une distinction entre les sinistres corporels et les sinistres matériels. Puis nous modéliserons séparément les sinistres dits « graves », et plus particulièrement leur montant moyen, ainsi que la probabilité d'occurrence d'un sinistre grave, en prenant une fois de plus soin de modéliser séparément les sinistres matériels et les sinistres corporels. Il nous restera alors à modéliser la fréquence de sinistre. Nous obtiendrons enfin le montant de la prime pure en calculant le produit entre la fréquence de sinistres et le montant moyen (en pondérant par le pourcentage de sinistres corporels et matériels) auquel nous rajouterons le produit entre les montants moyens des sinistres graves et leurs probabilités d'occurrence.

En ce qui concerne la modélisation des montants moyens de sinistres, la loi Gamma est celle la plus classiquement utilisée (distribution positive et continue avec augmentation de la variance avec l'espérance). Pour ce qui est de la probabilité d'occurrence des sinistres graves, nous utiliserons une régression logistique (modélisation de l'absence ou de la présence). Enfin, la fréquence sera modélisée à partir d'une régression de poisson ou par un modèle dérivé prenant en compte la sur dispersion (Binomial négatif, quasi-poisson, quasi-Binomial négatif, ZIP ou ZINB).

Les objectifs de ces modélisations sont multiples. Le premier d'entre eux et celui le plus clairement affiché est l'analyse de segments sur ou sous tarifés, qui implique donc une comparaison avec la prime pure effectivement payée par l'assuré aux vues de la tarification en vigueur. Cette analyse pourra éventuellement déboucher sur une évolution des tarifs en vigueur à venir. Dans cette optique, nous nous intéresserons particulièrement à la distinction entre les usages 2 et 3, qui est inexistante en termes de tarification actuellement, aux coefficients associés à la variable zone, car comme nous l'avons vu précédemment, le zonier actuel ne rend pas parfaitement compte de la réalité du risque, et enfin nous nous intéresserons à la variable AnneeDepart, qui n'est actuellement pas prise en compte dans la tarification, qui se base principalement sur la classe et le groupe (pour ce qui est des caractéristiques du véhicule). Cette étude pourra également nous donner des informations concernant les provisions d'ouvertures (celles inscrites suite à une déclaration de sinistre) et nous indiquer si celles-ci sont conformes avec la réalité.

3.2 Modélisation des montants de sinistres.

Avant toute chose, il est important de préciser que le terme de « montants de sinistres » est un abus de langage. Dans le cadre de cette modélisation, la variable à expliquer est en réalité la charge de sinistres, soit le montant réellement déboursé par l'assureur, qui ne correspond pas parfaitement au coût réel du sinistre. En effet, en raison de conventions entre différents assureurs, il est peu fréquent d'observer directement le coût réel du sinistre. Nous inscrivons ci-dessous la logique de la convention IDA (Indemnisation Directe de l'Assuré), pour la garantie en responsabilité civile, afin de comprendre la différence entre la charge pour l'assureur et le montant de sinistre ;

- Soit X et Y deux assureurs conventionnés.
- Un assuré de l'assureur X déclare un sinistre avec un tiers assuré de l'assureur Y.
- Si l'assuré de X est non responsable (RC0), alors l'assureur Y versera un montant forfaitaire de 1 204€ (on parle alors de recours) et l'assureur X règlera son assuré.
- Si le sinistre est en tort partagé (RC50), les assureurs X et Y se verseront tout deux 602€, supposés s'annuler, et régleront leurs assurés (dont le droit à indemnisation est réduit).
- Si l'assuré de X est responsable à 100% (RC100), alors l'assureur X versera un montant forfaitaire de 1 204€ à l'assureur Y qui règlera son assuré.

La responsabilité est déterminée par un barème que le lecteur intéressé pourra trouver sur Internet (http://www.lepermis.com/Assurances/barIRSA.htm). De plus, le lecteur intéressé trouvera plus d'informations sur Internet (http://btsassurance.canalblog.com/tag/Convention%20IRSA) ou encore pourra commander l'ouvrage publié par Seddita sur le sujet en se rendant sur le site http://www.seddita.com/detail_produit.asp?ID=1701.

Suite à cette parenthèse, nous allons nous intéresser à la modélisation de la charge moyenne des sinistres matériels standards (la logique étant la même pour les sinistres corporels). Pour commencer, nous allons représenter rapidement la répartition de la charge de sinistre dans différentes classes de montants.

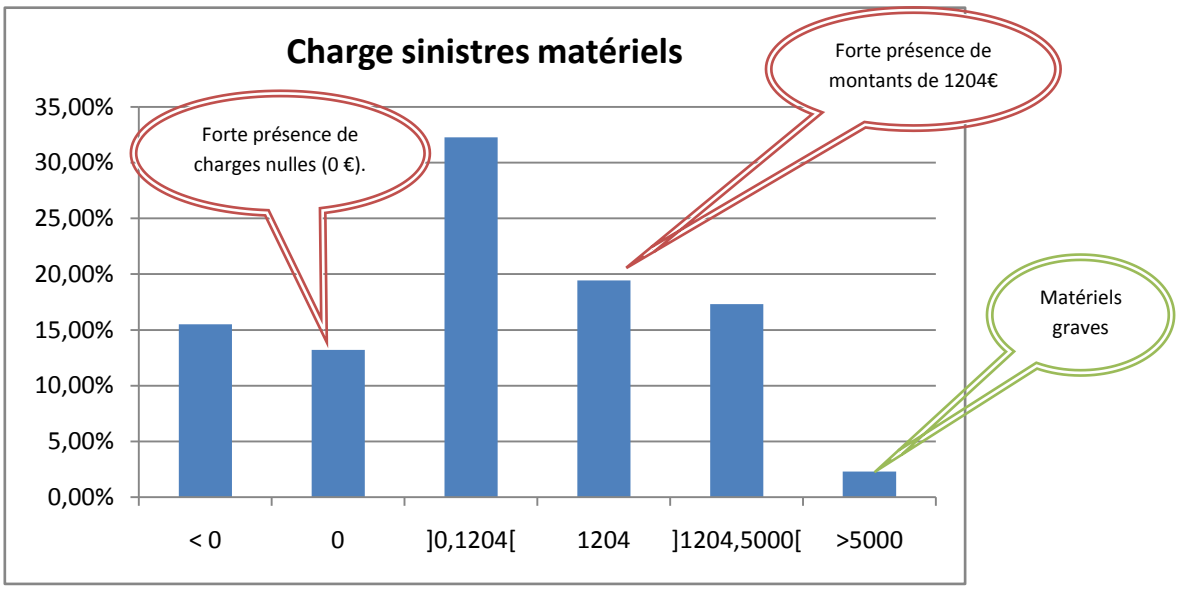


Illustration 41

Nous remarquons immédiatement les deux pics associés aux valeurs 0 et 1 204 (prises individuellement). Nous pouvons également convenir que la particularité de ces montants ne saurait être prise en compte par un modèle linéaire généralisé de loi Gamma. Nous prenons ainsi la décision de retirer ces valeurs de l'étude. Nous les y réintègrerons par la suite lors du calcul du montant moyen, en effectuant une pondération fonction du nombre de sinistres d'un montant de 0 € ou de 1 204 €. Il est à noter que la valeur 0 correspond aux sinistres de responsabilité nulle dans le cas où un des deux assureurs engagés est hors convention (c'est alors le droit commun qui est applicable). La valeur 1 204 (ainsi que la valeur 602 non significative ici) correspond au montant du recours de convention. Les valeurs 1 140 et 570 pourraient éventuellement être remarquables, elles correspondent en réalité aux provisions d'ouvertures respectivement pour les sinistres avec responsabilité totale et partielle. En pratique, ces valeurs ne seront pas retenues étant donné le fait que nous nous intéresserons aux sinistres clos (aucune provision ne pourra ainsi intervenir dans la charge de sinistre). Nous faisons ensuite remarquer la présence de sinistres matériels graves. Dans le cadre de notre étude, nous considèrerons qu'un sinistre matériel inférieur à 5 000 € est standard et inversement un sinistre matériel d'un montant supérieur à 5 000 € est considéré comme grave. Enfin, nous notons la présence de charges négatives, rendue possible par la notion de recours, mais étant un obstacle à la réalisation d'un modèle linéaire selon une loi Gamma. En pratique nous effectuerons un décalage de la charge de sinistre pour un montant de 1 300 €. Ce décalage ne devra cependant pas être oublié d'ici la fin de la modélisation, sous peine de créer une erreur non négligeable.

Nous allons alors pouvoir commencer la modélisation. Suite à une confrontation de la fonction de lien log et la fonction de lien inverse pour le modèle complet (avec toutes les variables), nous décidons de choisir comme fonction de lien la fonction inverse correspondant à un AIC plus intéressant (149 946 contre 150 356) et à une déviance plus faible (2600,5 contre 2651,3). La fonction canonique est donc ici celle qui est choisie, il en sera de même dans toute la suite. En ce qui concerne le choix des variables explicatives, nous commençons par réaliser une modélisation tenant compte de toutes les variables à l'aide de la fonction « glm », puis nous ne sélectionnons que les variables significatives, à partir d'une méthode descendante fonction du critère AIC en utilisant la fonction « step ». Nous effectuons alors un premier contrôle de la pertinence du modèle et des variables utilisées à partir des fonctions « summary », « anova », « drop1 » et « plot ». Enfin, nous réitérons ces différentes étapes en incluant les interactions entre les variables précédemment significatives, ce qui signifie que les occurrences des différentes variables ne sont pas analysées séparément mais par groupe de deux variables (un coefficient est ainsi attribué à l'occurrence ZoneZ2 : UsageU2 ou ZoneZ2 : UsageU3 par exemple). Nous inscrivons ci-dessous le code de chacune de ces étapes ainsi que les résultats obtenus à partir des différentes fonctions et leur signification.

(1) Nous réalisons une modélisation Gamma avec toutes les variables explicatives.

```
GLMA <- glm(Charge~. ,family=Gamma(link='inverse'), data=X,na.action=na.exclude)
```

(2) Nous ne gardons que les variables les plus pertinentes à partir de la fonction « step », décrite plus bas, et utilisant ici une méthode descendante.

```
GLMA2=step(GLMA,dir='backward')
```

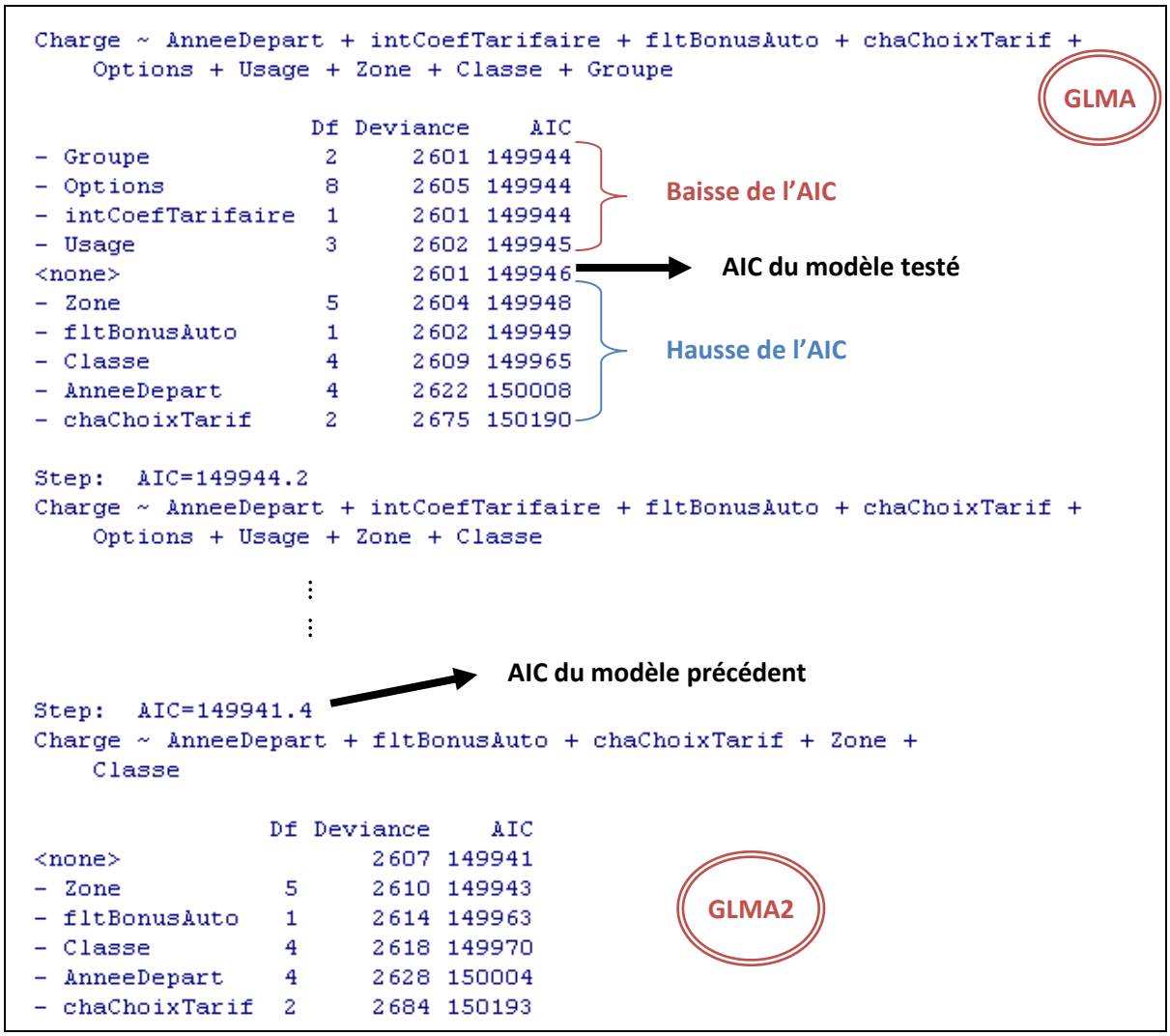


Illustration 42 : Principe de la fonction « step » avec une méthode descendante

Comme son nom l'indique, la fonction « step » procède par étape. La première étape consiste à calculer l'AIC du modèle de référence (ici GLMA), ainsi que l'impact du retrait de chaque variable sur l'AIC. Nous enlevons ainsi la variable dont le retrait permet une baisse maximale de l'AIC, jusqu'à ce que le retrait d'une des variables restantes soit synonyme de l'augmentation de l'AIC. Ici la variable Groupe est la première à être retirée. Nous utilisons ici une méthode descendante (dir='backward'), mais il est également possible d'utiliser une méthode ascendante (on teste alors l'impact du rajout de variables à préciser) ou encore une méthode hiérarchique, combinaison des deux méthodes précédentes.

(3) Contrôle de la pertinence du modèle et des variables explicatives, la fonction « summary ».

```
summary(GLMA2)
```

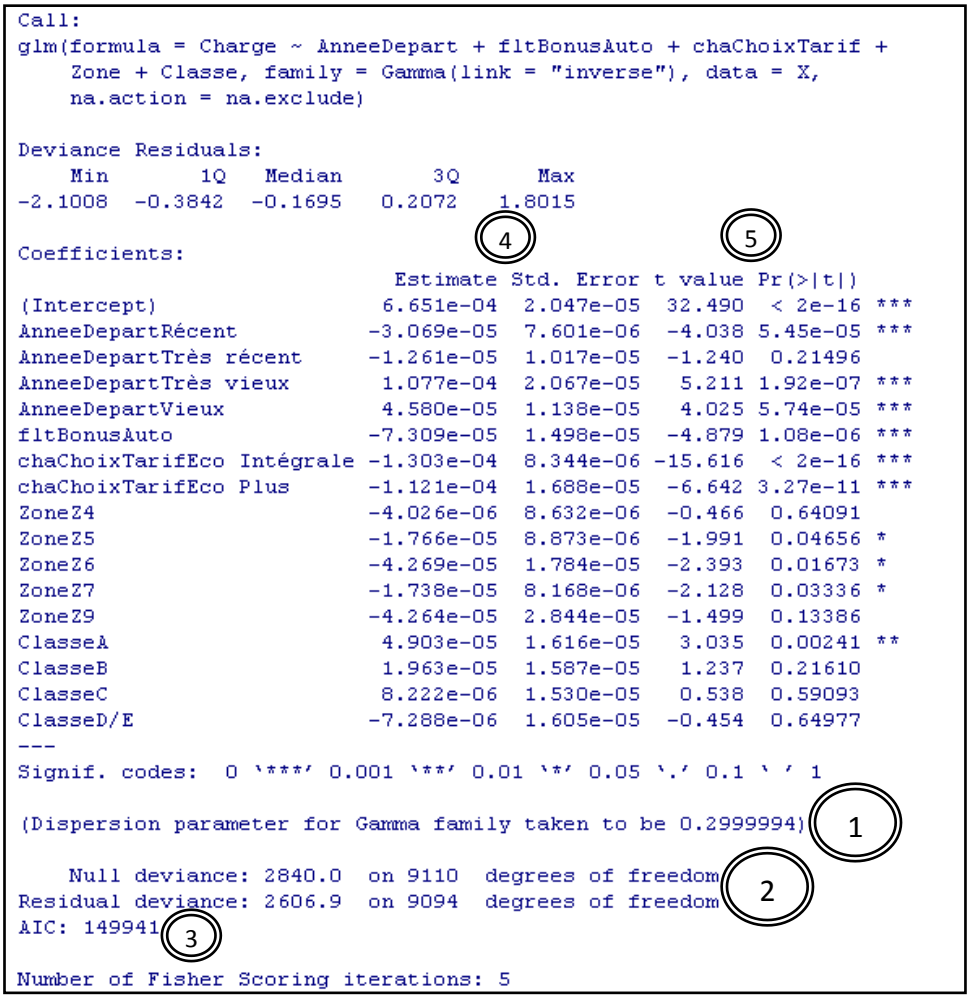


Illustration 43 : Eléments de la fonction « summary » d'un GLM

En ce qui concerne le contrôle de la légitimité du modèle, nous devons nous intéresser à la déviance standardisée et la comparer au nombre de degrés de liberté des résidus. Nous trouvons la déviance du modèle sur la deuxième ligne au niveau du cercle 2 (Residual deviance), de même que le nombre de degrés de liberté. Il est cependant nécessaire de la standardiser en divisant par l'estimation du paramètre de dispersion que l'on trouve au niveau du cercle 1. Nous obtenons ici que la déviance standardisée vaut 8 689,68 ce qui est inférieur au nombre de degrés de liberté. Nous pouvons donc admettre que le modèle est pertinent (on rappelle que le modèle est acceptable si le rapport de la déviance standardisée sur les degrés de liberté n'est pas grand devant 1), sans prendre le soin de réaliser un test de déviance (le quantile du Khi² étant automatiquement supérieur au nombre de degrés de libertés). Nous remarquons ensuite que la fonction « summary » fournit le critère AIC (cercle 3), ainsi que les coefficients et écarts types estimés des paramètres associés à chaque occurrence des prédicteurs (cercle 4). Les deux colonnes suivantes (cercle 5) sont issues d'un test de student visant à tester l'hypothèse de nullité d'un coefficient. Lorsque la p-value est faible, le coefficient associé est significativement non nul et un certain nombre d'astérisques apparaissent. Nous pourront cependant noter que ceci ne nous permet pas de juger de la significativité d'une variable, car le test de significativité mis en œuvre ici porte sur les occurrences de la variable, et non pas sur la variable elle-même. Nous remarquons également que toutes les occurrences des variables ne sont pas apparentes, en effet, pour chaque variable, une occurrence (généralement la plus présente) correspond à une occurrence de référence et l'ensemble forme un individu de référence

représenté par l'intercept. Ici l'individu de référence possède un contrat en garantie Eco, pour un véhicule agé de classe supérieure à F utilisé en zone 2. En ce qui concerne l'interprétation des coefficients, la fonction de lien étant la fonction inverse, un coefficient négatif indique un risque plus fort que pour l'individu de référence et inversement un coefficient positif indique un risque plus faible. Ainsi nous constatons à titre d'exemple que les véhicules récents sont associés à des coûts moyens matériels plus élevés que les véhicules âgés et les véhicules anciens à des montants plus faibles. De même, l'espérance de montant de sinistres matériels augmente avec le CRM. Une fois la méthodologie mise en avant, nous reviendrons en détails sur les résultats obtenus et l'interprétation des coefficients plus loin dans le document. Nous faisons enfin remarquer que les fonctions « confint » et « confint.standard » permettent d'obtenir des intervalles de confiances des coefficients représentés ici grâce respectivement à la loi du Khi² et à la loi normale (voir *page 95*).

(4) Contrôle de la pertinence du modèle et des variables explicatives, la fonction « anova ».

```
anova(GLMA2,test='F')
```

```
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: Charge

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev        F    Pr(>F)
NULL                         9110    2839.98
AnneeDepart    4   144.42      9106    2695.56 120.3465 < 2.2e-16 ***
fltBonusAuto   1     0.23      9105    2695.33   0.7572    0.3842
chaChoixTarif  2    75.10      9103    2620.23 125.1676 < 2.2e-16 ***
Zone           5     2.45      9098    2617.78   1.6333    0.1474
Classe         4    10.87      9094    2606.91   9.0611 2.662e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustration 44 : Résultats issus de la fonction « anova »

La fonction « anova » permet de réaliser un test de significativité des variables explicatives. Selon la même logique que précédemment, une p-value (dernière colonne) faible est synonyme de significativité de la variable. Il est à noter que la fonction « anova » tient compte de l'ordre d'introduction des variables (« Terms added sequentially, first to last »). En effet, le premier test confronte le modèle NULL (aucune variable explicative) au modèle avec l'AnneeDepart comme seule variable, le second confronte ce dernier au modèle avec AnneeDepart et fltBonusAuto et ainsi de suite. Nous constatons ici que le rajout de la variable fltBonusAuto au modèle avec la variable AnneeDepart n'est pas pertinent, ce qui ne signifie pas pour autant que la variable n'est pas significative dans le modèle complet. Dans le but de valider la significativité des variables d'un modèle, nous préfèrerons utiliser un test de significativité indépendant de l'ordre d'introduction des variables, testant le modèle complet contre le modèle sans une variable. Pour ce faire, nous allons utiliser la fonction « drop1 », plus pertinente ici que la fonction « anova » et décrite ci-dessous. Nous faisons enfin remarquer que pour les modélisations de données binaires ou de comptage, l'option « test='Chi' » sera préférée à l'option « test='F' ».

(5) Contrôle de la pertinence du modèle et des variables explicatives, la fonction « drop1 ».

```
drop1(GLMA2,test='F')
```

```
Single term deletions

Model:
Charge ~ AnneeDepart + fltBonusAuto + chaChoixTarif + Zone +
    Classe
              Df Deviance    AIC  F value     Pr(F)
<none>              2607 149941
AnneeDepart    4     2628 150004  18.4744 3.926e-15 ***
fltBonusAuto   1     2614 149963  24.5284 7.454e-07 ***
chaChoixTarif  2     2684 150193 133.8142 < 2.2e-16 ***
Zone           5     2610 149943   2.3869   0.03578 *
Classe         4     2618 149970   9.4826 1.200e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustration 45 : Résultats issus de la fonction « drop1 »

Comme indiqué ci-dessus, la fonction « drop1 » permet de contrôler la significativité des variables indépendamment de leur ordre d'introduction (« Single term deletions »). Nous constatons alors que les variables fltBonusAuto et Zone sont ici significatives, contrairement à ce que laissait paraitre la fonction « anova », bien que la significativité soit ici moins prononcée que pour les autres variables.

(6) Contrôle de la pertinence du modèle et des variables explicatives, la fonction « plot ».

```
par(mfrow=c(2,2))
plot(GLMA2,which=1:4,ask=F)
```

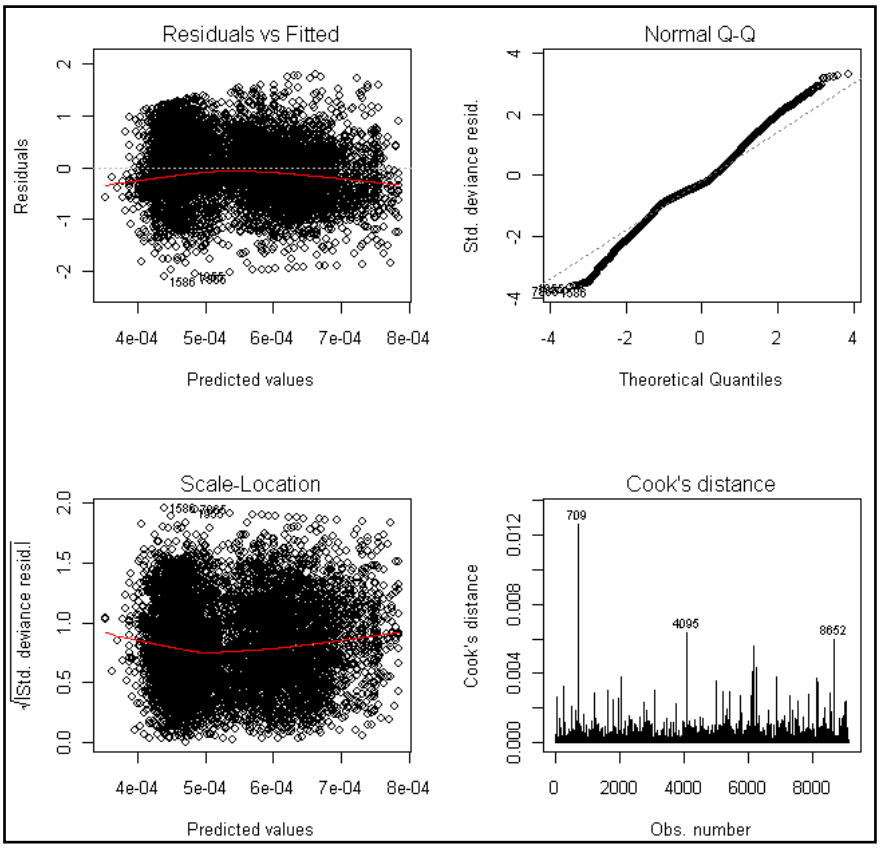


Illustration 46 : Graphiques d'analyse des résidus

Comme le titre de l'illustration 46 l'indique, l'utilisation de la fonction « plot » appliquée aux modèles linéaires généralisés nous permet d'effectuer une analyse des résidus, et plus particulièrement des résidus de déviance. Le premier graphique (en haut à gauche) est une représentation des résidus en fonction des valeurs prédites, l'absence de tendance significative et l'équidispersion des points autour de l'ordonnée 0 (comme c'est le cas ici) indique une bonne adéquation du modèle au problème. Le second graphique permet de contrôler l'adéquation des résidus à une loi normale. Le troisième est une représentation de la racine des résidus (en valeurs absolues) en fonction des valeurs prédites. Comme pour le premier graphique, l'absence de tendance est la preuve d'une bonne adéquation. Enfin le dernier graphique est celui des distances de Cook. Comme nous l'avons vu précédemment, une distance supérieure à 1 sera considérée comme anormale, nous remarquons que toutes les observations de l'étude sont convenables ici.

(7) Prise en compte des interactions.

```
GLMA3 <- glm(Charge~(AnneeDepart+Classe+chaChoixTarif+fltBonusAuto+Zone)^2
,family=Gamma(link='inverse'), data=X,na.action=na.exclude)
```

Une fois les interactions prise en compte, nous réitérons les étapes précédentes. Nous inscrivons ci-dessous le résultat issu de la fonction « drop1 », afin d'analyser les interactions significatives de cette étude.

```
Single term deletions

Model:
Charge ~ AnneeDepart + Classe + chaChoixTarif + fltBonusAuto +
    Zone + AnneeDepart:Classe + AnneeDepart:chaChoixTarif + Classe:chaChoixTarif +
    chaChoixTarif:fltBonusAuto + chaChoixTarif:Zone
                          Df Deviance    AIC scaled dev.   Pr(Chi)
<none>                          2511 149671
AnneeDepart:Classe        16     2526 149690          52 1.308e-05 ***
AnneeDepart:chaChoixTarif  8     2529 149719          64 6.521e-11 ***
Classe:chaChoixTarif       8     2532 149728          74 8.525e-13 ***
chaChoixTarif:fltBonusAuto 2     2514 149680          13  0.001529 **
chaChoixTarif:Zone        10     2518 149675          24  0.006768 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustration 47 : Exemple de résultats issus de la fonction « drop1 » avec interactions

Nous pouvons alors remarquer qu'en présence d'interactions, les termes individuels n'apparaissent pas dans l'étude de la significativité des variables. Nous pouvons ici remarquer que la variable chaChoixTarif est très présente en termes d'interaction. Ceci vient du fait que l'évolution du risque en fonction des critères d'évaluation n'est pas la même en fonction des garanties souscrites. En clair ici, nous savons que les véhicules plus récents possèdent un montant de sinistres matériels standards moyen plus élevé que la normale, mais les interactions nous indiquent également que la quantification du risque supplémentaire n'est pas la même en fonction du type de tarif choisi. Nous remarquons également qu'au-delà de la variable chaChoixTarif, seule l'interaction entre les variables AnneeDepart et Classe est prise en compte. Pour mémoire, un des objectifs affichés était de s'intéresser à la variable AnneeDepart, la tarification actuelle se basant sur les variables Groupe et Classe de façon conjointe. En ce qui concerne les montants de sinistres matériels, nous pouvons d'ores et déjà signaler que le Groupe n'apparait pas, et que l'interaction entre AnneeDepart et Classe semble la plus adaptée. Nous notons enfin que l'AIC, ici de 149 671, est bien inférieur à celui du second modèle qui était de 149 941. La prise en compte des interactions est donc réellement pertinente ici.

Une fois les coefficients obtenus, nous pouvons soit les exporter sous Excel afin d'effectuer des prédictions fonctions des différentes catégories de risques, soit effectuer des prédictions directement avec le logiciel R à partir de la fonction « predict ». A ce stade de la modélisation, il faut garder en mémoire d'éventuelles transformations de la variable à expliquer. Ici, nous avions retiré de l'étude les montants de valeurs 0 et 1 204, et nous avions effectué une translation afin de considérer des valeurs positives. Pour des questions de clarté et de volume, nous n'exposons pas ici tous les coefficients obtenus, mais simplement les résultats et interprétations.

Nous pouvons conclure que de manière générale, le montant moyen de sinistres matériels augmente avec la date de première mise en circulation et la classe du véhicule. L'impact de la zone est plus complexe, mais nous remarquons d'ores et déjà que la zone 7 (spécifique aux DOM-TOM), n'est pas porteuse d'un risque significatif, contrairement aux zones 5, 5/6 et 6 (agglomérations et zones côtières). Nous remarquons également que le montant moyen dépend de façon non négligeable du CRM, particulièrement pour les contrats de type Eco (garantie RC seulement). Ceci indique que les individus à faible CRM, non seulement ont une fréquence de sinistre plus faible (a priori) mais en plus possèdent en moyenne des montants de sinistres plus faibles que la moyenne. En ce qui concerne les sinistres corporels standards (< 30 000 €) que nous n'avons pas encore abordés, nous pouvons signaler que seul le type de contrat est retenu comme variable significative, ce qui indique que le montant de sinistres ne dépend pas sensiblement de l'assuré, et que la seule variation observable et identifiable est due aux indemnisations de type matériel (et particulièrement à la distinction des garanties RC et Dommages) contenues dans les sinistres à composante corporelle.

Lors de la définition des objectifs, nous avons émis la possibilité de faire une étude spécifique en parallèle afin de contrôler les provisions d'ouvertures. Nous savons en effet que lors de la déclaration d'un sinistre en responsabilité totale et hors convention par exemple, les provisions d'ouvertures sont portées à 1 140 € pour les sinistres matériels et 6 748 € pour les sinistres corporels. Nous pouvons ainsi établir une modélisation spécifique afin de calculer un montant moyen en fonction des garanties. N'ayant pas accès directement à la garantie touchées ici, nous effectuons une distinction entre les types de contrats Eco et Eco Plus associés à la garantie RC et les contrats de type Eco Intégrale contenant une composante garantie dommage. Nous inscrivons ci-dessous les résultats obtenus ;

	Charge <5000€		Charge >5000€	
	Eco+Eco Plus	Eco intégrale	Eco+Eco Plus	Eco intégrale
Nbr Matériels RC100	4171	1569	61	221
Montant moyen	1 365 €	2 234 €	9 297 €	8 206 €
Espérance	**1 479 €**	**2 971 €**		

	Charge <30000€		Charge >30000€	
	Eco+Eco Plus	Eco intégrale	Eco+Eco Plus	Eco intégrale
Nbr Corporels RC100	633	167	30 / 12	
Montant moyen	5 837 €	8 554 €	95 040 €	
Espérance	**9 873 €**	**14 352 €**		

Illustration 48 : Montants moyens de charge sinistre en responsabilité totale

Nous pouvons ainsi constater que les provisions d'ouvertures paraissent réellement sous-évaluées au regard des montants moyens réels, et ce, même dans l'hypothèse où seul les sinistres standards en responsabilité civile sont considérés. Bien entendu, un tel écart ne peut être mis en avant sans être discuté, comparé et analysé plus en détails. Cependant, les résultats exposés ci-dessus indiquent la nécessité de réaliser une étude plus poussée concernant les provisions d'ouvertures, portant sur un plus grand nombre d'années d'exercice, et cherchant à expliquer l'origine de tels écarts.

3.3 Modélisation des sinistres graves.

Nous avons entrevu ci-dessus un exemple de modélisation de montants de sinistres standards. Cette modélisation permet d'effectuer des prédictions et des analyses autour de la grande majorité des sinistres et pourrait à ce titre paraitre suffisante. Cependant, en assurance automobile, il arrive bien souvent qu'un faible nombre de sinistres soit responsable d'une très grande partie de la charge totale de sinistres, d'où la nécessité de modéliser séparément les sinistres d'un montant exceptionnel. Afin de mettre en évidence ce phénomène, nous exposons ci-dessous deux tableaux comparant le nombre de sinistres graves avec le nombre total de sinistres ainsi que la charge des sinistres graves avec la charge totale des sinistres.

	Nombre de sinistres	Charge totale des sinistres
Sinistres matériels	16 392 (100 %)	13 928 034.68 (100 %)
Sinistres matériels >5000€	382 (2,3 %)	3 199 314,34 (23 %)

	Nombre de sinistres	Charge totale des sinistres
Sinistres corporels	1 478 (100 %)	11 699 323,11 (100 %)
Sinistres corporels >30000€	45 (3 %)	5 428 975,26 (46 %)

Illustration 49 : L'impact des sinistres graves sur la charge totale

Nous pouvons ainsi remarquer que les sinistres matériels graves, qui représentent moins de 3 % des sinistres matériels, sont responsables d'environ un quart de la charge totale. De même, les sinistres corporels graves, qui représentent 3 % du nombre total de sinistres corporels, correspondent à environ la moitié de la charge totale des sinistres corporels. Nous comprenons ainsi la nécessité de modéliser séparément ces sinistres rares responsables en grande partie du coût total pour l'assureur.

Comme nous l'avons signalé précédemment, il convient ici de modéliser d'une part le montant moyen et d'autre part la probabilité de survenance d'un sinistre grave. La première modélisation, en tout point semblable au point précédent utilisera ainsi la loi gamma tandis que la deuxième modélisation utilisera la régression logistique.

Nous devons alors faire remarquer que lors de la modélisation de la probabilité d'occurrence d'un sinistre grave, il est nécessaire de prendre en compte l'exposition réelle au risque de chaque observation. En effet, un contrat présent en portefeuille pendant 5 jours n'a pas la même signification et le même impact qu'un contrat présent durant un an. Il est possible de prendre en compte ce phénomène en divisant le nombre de sinistres graves (qui est la variable à expliquer) par l'exposition et en indiquant dans la fonction « glm » l'option « weight=Exposition ».

Pour ce qui est de la modélisation des sinistres matériels graves, nous constatons que le montant moyen varie en fonction de la classe et de la date de première mise en circulation du véhicule, les coûts étant plus élevés pour les véhicules récents ou de plus de 13 ans, et de classe élevée. Il peut paraitre à première vu étonnant que les véhicules très âgés soient associés à un coût plus élevé, mais ceci peut alors provenir du fait qu'en cas de sinistre important, un véhicule âgé moins résistant implique et accuse des réparations en chaine. En ce qui concerne la probabilité de survenance, nous remarquons un fort impact du CRM ainsi que de l'option. Nous remarquons ensuite que la classe et l'ancienneté du véhicule joue également un rôle significatif, les véhicules récents et de classe élevée étant plus enclin à avoir un sinistre grave. Ceci nous indique que les véhicules récents et de valeurs à neufs élevées couteraient non seulement plus chère, mais seraient également plus souvent à l'origine de sinistres matériels de montants élevés, contrairement à des véhicules peu récents et peu coûteux (les réparations éventuellement en partie à la charge de l'assuré, peuvent parfois ne pas être supportées par lui et par conséquent ne pas être déclenchées).

Pour ce qui est des sinistres corporels graves, nous constatons comme précédemment qu'il est beaucoup plus difficile d'obtenir des variables significatives. En effet, aucune variable n'est significative dans la détermination du montant moyen (trop peu de données pour des variations importantes) et seul le CRM est significatif dans la détermination de la probabilité d'occurrence d'un sinistre corporel grave. Ceci nous laisse ainsi penser que les sinistres corporels graves sont en grande partie issus du hasard, aussi bien dans leurs montants que dans leurs probabilités de survenance. Un individu prudent à faible CRM sera cependant moins enclin qu'un autre à être impliqué dans un sinistre corporel grave, mais seule cette notion de prudence liée au CRM possède un pouvoir prédictif. Le type de contrat n'apporte aucune indication supplémentaire ainsi que les caractéristiques du véhicule.

3.4 Modélisation de la fréquence de sinistre.

Comme indiqué plus haut, la modélisation de la fréquence de sinistre peut être opérée à partir d'une régression de poisson ou dérivé (modèle Binomial négatif, quasi poisson, quasi Binomial négatif, ZIP, ZINB). De la même façon que pour la modélisation de la probabilité d'occurrence des sinistres graves, il est nécessaire de tenir compte de la durée d'exposition au risque.

Nous avons vu précédemment de façon théorique que l'exposition est prise en compte en considérant le logarithme de l'exposition avec un coefficient fixé à 1 (qui ne sera donc pas estimé). En pratique sous le logiciel R, on rajoute en tant que variable explicative le terme « offset(log(Exposition)) » pour réaliser ceci (le terme « offset » est récurent sous d'autres logiciels également).

Nous pouvons ainsi commencer par réaliser une régression de poisson sans considérer les interactions puis par effectuer de la même manière une régression quasi-poisson et contrôler ainsi la présence de sur ou sous dispersion ainsi que l'impact sur le choix des variables explicatives;

```
GLME <- glm(NombreSinistre~.-Exposition+offset(log(Exposition)),data=X,family=poisson)
GLME2<- glm(NombreSinistre~.-Exposition+offset(log(Exposition)),data=X,family=quasipoisson)
summary(GLME) ;summary(GLME2)
drop1(GLME,test='Chi') ; drop1(GLME2,test='Chi')
```

```
Call:
glm(formula = NombreSinistre ~ . - Exposition + offset(log(Exposition)),
    family = poisson, data = X, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4708  -0.5048  -0.3880  -0.2287   4.8892

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -9.22750    0.28896 -31.934  < 2e-16 ***
AnneeDepartRécent          0.11232    0.02308   4.868 1.13e-06 ***
AnneeDepartTrès récent     0.15495    0.03422   4.528 5.94e-06 ***
AnneeDepartTrès vieux     -0.34186    0.03558  -9.609  < 2e-16 ***
AnneeDepartVieux          -0.16903    0.02431  -6.952 3.61e-12 ***
intCoefTarifaire           0.15622    0.05402   2.892  0.00383 **
fltBonusAuto               1.05127    0.04852  21.667  < 2e-16 ***
chaChoixTarifEco Intégrale 0.87364    0.02316  37.725  < 2e-16 ***
chaChoixTarifEco Plus      0.26251    0.05021   5.229 1.71e-07 ***
OptionsCO7                 0.04099    0.03799   1.079  0.28060
OptionsC10                 0.02790    0.03789   0.736  0.46156
OptionsC12                 0.11385    0.04569   2.492  0.01271 *
OptionsC20                 0.48953    0.06826   7.171 7.43e-13 ***
OptionsCLE                 0.02942    0.06262   0.470  0.63848
OptionsESS                 0.19979    0.02848   7.015 2.31e-12 ***
OptionsREL                 0.41395    0.03398  12.183  < 2e-16 ***
OptionsSEL                 0.09845    0.02487   3.959 7.54e-05 ***
UsageU2                    0.34213    0.28081   1.218  0.22307
UsageU3                    0.44386    0.28112   1.579  0.11435
UsageU4                    0.54046    0.29410   1.838  0.06611 .
ZoneZ4                     0.04807    0.02386   2.015  0.04395 *
ZoneZ5                     0.22553    0.02421   9.316  < 2e-16 ***
ZoneZ6                     0.27952    0.04650   6.011 1.84e-09 ***
ZoneZ7                    -0.18418    0.02376  -7.752 9.08e-15 ***
ZoneZ9                     0.18396    0.07008   2.625  0.00867 **
ClasseA                   -0.24927    0.05926  -4.207 2.59e-05 ***
ClasseB                   -0.12946    0.05777  -2.241  0.02502 *
ClasseC                   -0.01132    0.05428  -0.209  0.83482
ClasseD/E                  0.01649    0.05351   0.308  0.75791
GroupeG12-14              -0.13366    0.05945  -2.248  0.02455 *
GroupeG5-11               -0.15280    0.06405  -2.386  0.01705 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 72580  on 137720  degrees of freedom
Residual deviance: 68167  on 137690  degrees of freedom
AIC: 96201

Number of Fisher Scoring iterations: 6
```

```
Call:
glm(formula = NombreSinistre ~ . - Exposition + offset(log(Exposition)),
    family = quasipoisson, data = X, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4708  -0.5048  -0.3880  -0.2287   4.8892

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -9.22750    0.38270 -24.112  < 2e-16 ***
AnneeDepartRécent          0.11232    0.03056   3.675 0.000238 ***
AnneeDepartTrès récent     0.15495    0.04532   3.419 0.000628 ***
AnneeDepartTrès vieux     -0.34186    0.04712  -7.255 4.04e-13 ***
AnneeDepartVieux          -0.16903    0.03220  -5.249 1.53e-07 ***
intCoefTarifaire           0.15622    0.07154   2.184 0.028994 *
fltBonusAuto               1.05127    0.06426  16.359  < 2e-16 ***
chaChoixTarifEco Intégrale 0.87364    0.03067  28.484  < 2e-16 ***
chaChoixTarifEco Plus      0.26251    0.06649   3.948 7.89e-05 ***
OptionsCO7                 0.04099    0.05032   0.815 0.415259
OptionsC10                 0.02790    0.05019   0.556 0.578263
OptionsC12                 0.11385    0.06051   1.881 0.059907 .
OptionsC20                 0.48953    0.09041   5.415 6.15e-08 ***
OptionsCLE                 0.02942    0.08294   0.355 0.722781
OptionsESS                 0.19979    0.03772   5.296 1.18e-07 ***
OptionsREL                 0.41395    0.04500   9.199  < 2e-16 ***
OptionsSEL                 0.09845    0.03294   2.989 0.002800 **
UsageU2                    0.34213    0.37191   0.920 0.357600
UsageU3                    0.44386    0.37231   1.192 0.233199
UsageU4                    0.54046    0.38952   1.388 0.165284
ZoneZ4                     0.04807    0.03160   1.521 0.128236
ZoneZ5                     0.22553    0.03206   7.034 2.02e-12 ***
ZoneZ6                     0.27952    0.06159   4.539 5.67e-06 ***
ZoneZ7                    -0.18418    0.03147  -5.853 4.84e-09 ***
ZoneZ9                     0.18396    0.09282   1.982 0.047500 *
ClasseA                   -0.24927    0.07848  -3.176 0.001492 **
ClasseB                   -0.12946    0.07651  -1.692 0.090624 .
ClasseC                   -0.01132    0.07189  -0.157 0.874892
ClasseD/E                  0.01649    0.07087   0.233 0.815974
GroupeG12-14              -0.13366    0.07873  -1.698 0.089573 .
GroupeG5-11               -0.15280    0.08483  -1.801 0.071651 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.754074)

    Null deviance: 72580  on 137720  degrees of freedom
Residual deviance: 68167  on 137690  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Illustration 50 : Confrontation de modèles poisson et quasi-poisson, fonction « summary »

Nous pouvons tout d'abord constater à travers les résultats de la fonction « summary » que le paramètre de sur dispersion du modèle quasi-poisson est de 1,754074. Ceci nous indique qu'il y a effectivement un phénomène de sur dispersion dans les données. Nous pouvons alors vérifier que comme nous l'avons signalé précédemment, les estimations des paramètres associées aux prédicteurs sont parfaitement identiques. De même nous constatons que les écarts types sont sensiblement amplifiés dans la modélisation quasi-poisson et nous pouvons vérifier que les écarts type du modèle quasi-poisson correspondent à ceux du modèle poisson multiplié par la racine du paramètre de sur dispersion. Nous notons enfin les légères différences dans la significativité des variables, plus parlantes dans les résultats de la fonction « drop1 » (voir ci-dessous).

Poisson

```
Single term deletions

Model:
NombreSinistre ~ (AnneeDepart + intCoefTarifaire + fltBonusAuto +
    chaChoixTarif + Options + Usage + Zone + Classe + Groupe +
    Exposition) - Exposition + offset(log(Exposition))
               Df Deviance   AIC   LRT   Pr(Chi)
<none>              68167 96201
AnneeDepart       4    68338 96364   171 < 2.2e-16 ***
intCoefTarifaire  1    68175 96208     8  0.004383 **
fltBonusAuto      1    68623 96656   456 < 2.2e-16 ***
chaChoixTarif     2    69592 97622  1425 < 2.2e-16 ***
Options           8    68378 96397   211 < 2.2e-16 ***
Usage             3    68200 96229    33 2.925e-07 ***
Zone              5    68465 96489   298 < 2.2e-16 ***
Classe            4    68244 96271    77 6.790e-16 ***
Groupe            2    68173 96203     6  0.060954 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quasi-Poisson

```
Single term deletions

Model:
NombreSinistre ~ (AnneeDepart + intCoefTarifaire + fltBonusAuto +
    chaChoixTarif + Options + Usage + Zone + Classe + Groupe +
    Exposition) - Exposition + offset(log(Exposition))
               Df Deviance scaled dev.   Pr(Chi)
<none>              68167
AnneeDepart       4    68338          97 < 2.2e-16 ***
intCoefTarifaire  1    68175           5 0.0314540 *
fltBonusAuto      1    68623         260 < 2.2e-16 ***
chaChoixTarif     2    69592         812 < 2.2e-16 ***
Options           8    68378         120 < 2.2e-16 ***
Usage             3    68200          19 0.0002831 ***
Zone              5    68465         170 < 2.2e-16 ***
Classe            4    68244          44 6.363e-09 ***
Groupe            2    68173           3 0.2029220
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustration 51 : Confrontation de modèles poisson et quasi-poisson, fonction « drop1 »

La fonction « drop1 » vient en effet confirmer le fait que la significativité des variables est différente d'un modèle de poisson à un modèle de quasi-poisson. Nous constatons que le Groupe qui était significatif (même faiblement) pour le modèle de poisson ne l'est plus pour le modèle de quasi-poisson. De même la significativité de la variable intCoefTarifaire est sensiblement réduite et l'on remarque que sa présence apporte un gain pour la déviance ou pour l'AIC relativement faible. Nous pourront donc négliger cette variable afin de privilégier la simplicité et la clarté du modèle.

Nous pourrions également indiquer les résultats issus de la modélisation binomiale négative fréquemment utilisée pour résoudre les problèmes de dispersion mais nous nous contenterons ici de signaler que les estimations obtenus sont très proches de celles issues d'une régression de poisson. Il est à noter que la fonction du logiciel R permettant de mettre en place un tel modèle est la fonction « glm.nb ». En ce qui concerne les modèles ZIP ou ZINB qui ne seront pas abordés ici, nous indiquons de nombreux packages du logiciel R permettant de les traiter, à savoir les packages « pscl », « gamlss », « VGAM » et « ZIGP ».

Ces quelques résultats nous invitent donc à garder une certaine prudence en ce qui concerne la modélisation de la fréquence et son pouvoir prédictif. Nous notons cependant que suite au retrait du groupe et de la variable intCoefTarifaire, les résultats obtenus semblent plus satisfaisant et parmi les interactions possibles, seule celle entre le CRM et l'option est retenue. Le rôle important de la classe du véhicule et de la date de première mise en circulation est à nouveau à signaler, la fréquence de sinistre diminuant sensiblement avec l'ancienneté du véhicule et augmentant avec la classe. Nous remarquons également que la zone 7 (DOM-TOM) possède une fréquence de sinistre plus faible que les autres, alors que les zones 5/6, 5 et 6 paraissent être les plus risquées par ordre croissant. Contrairement à ce qui était attendu, nous remarquons que les options de type CLE (associée à des antécédents de sinistres) sont associées aux fréquences de sinistres les plus faibles. Ceci vient en réalité du fait que d'une part les individus correspondant sont peu nombreux, et d'autre part ces individus sont sous le coup d'une « punition », ils ont connaissance du prix d'une erreur et ont tout intérêt à être vigilant et prudent. Une baisse des tarifs associés à ces options n'est ainsi pas réellement envisageable, car celle-ci limiterait la sanction, et donc l'impact sur la volonté de l'assuré d'user de prudence pour l'éviter à l'avenir et diminuer sa prime.

3.5 Analyse des segments sur ou sous tarifés.

Suite à l'ensemble des modélisations réalisées ci-dessus, il nous est possible de calculer une prime pure pour chaque catégorie de risque. Tout d'abord, résumons les différents types de modélisations réalisées ;

- Modélisation des montants moyens de sinistres à partir de la loi Gamma. En notant v la somme des coefficients fonction des caractéristiques de l'individu, nous obtenons le montant moyen en effectuant 1/ v. Ce montant doit alors être corrigé en fonction d'éventuels décalages d'échelles avant la modélisation ou d'éventuels retraits de certains montants spécifiques.
- Modélisation des probabilités de sinistres graves, en tenant compte de l'exposition exprimée en nombre de jours, à partir de la loi Binomiale (on parle de régression logistique). En notant v la somme des coefficients fonction des caractéristiques de l'individu, nous obtenons la probabilité d'occurrence d'un sinistre en effectuant $365 \times \frac{e^{v}}{1-e^{v}}$.
- Modélisation de la fréquence de sinistre à partir de la loi de Poisson, en tenant compte de l'exposition en nombre de jours. En notant v la somme des coefficients fonction des caractéristiques de l'individu, nous obtenons la probabilité d'occurrence d'un sinistre en effectuant $365 \times e^{v}$.

En utilisant les notations représentées ci-dessous, il nous est alors possible de calculer la prime pure issue du modèle, avec une pondération déterministe concernant la nature du sinistre (91,7 % de sinistres matériels et 8,3 % de sinistres corporels) ;

A : Montant moyen de sinistres matériels standards.
B : Montant moyen de sinistres matériels graves.
BB : Probabilité d'occurrence d'un sinistre matériel grave.
C : Montant moyen de sinistres corporels standards.
D : Montant moyen de sinistres corporels graves.
DD : Probabilité d'occurrence d'un sinistre corporel grave.
E : Fréquence de sinistre.
Prime Pure=Fréquence*(0,917*A+0,083*C)+BB*B+DD*D

Nous pouvons alors calculer cette prime pure pour l'ensemble des catégories de risques. Il est possible d'effectuer ceci à partir de la fonction « predict » du logiciel R, mais nous pouvons également réaliser ceci sous Excel à partir des coefficients de chaque modélisation. L'avantage de l'utilisation d'Excel réside dans le fait qu'il est possible de mettre en place un système de formulaire dans lequel on peut rentrer les caractéristiques de la catégorie de risque et voir instantanément les modifications occasionnées sur la prime pure issue du modèle et sur la prime pure issue de la tarification en vigueur. Nous exposons ci-dessous un tableau issu de cette possibilité contenant la fréquence de sinistre, le montant moyen des sinistres standards matériels et corporels ainsi que le produit entre les montants de sinistres graves et leurs probabilités d'occurrence. Un utilisateur quelconque peut alors modifier les paramètres (variables explicatives) et voir l'impact des modifications sur les différents résultats.

Caractéristiques de l'assuré	
Groupe	G5-11
AnneeDepart	Agés
intCoefTarifaire	1
fltBonusAuto	1
chaChoixTarif	Eco
Options	ESS
Usage	U2
Zone	Z2
Classe	A

Paramètres modifiables sous forme de listes déroulantes

Fréquence annuelle	**15,34%**
Montant moyen matériel standard	**351,27 €**
Montant moyen corpo standard	**2 406,79 €**
Matériel grave	**3,04 €**
Corporel grave	**53,64 €**
Prime pure théorique (modèle)	**136,76 €**
Prime pure réelle	**149,84 €**

Résultats fonctions des caractéristiques de l'assuré

Illustration 52 : Exemple d'interface utilisateur pour le calcul de la prime pure

Bien entendu, il est difficile de juger de la sur ou sous tarification à partir de ce seul tableau. Il est alors possible de regrouper dans un tableau l'ensemble des catégories de risque et de calculer un rapport S/P en divisant la prime pure théorique issue du modèle par la prime pure réelle. Dans l'hypothèse d'une bonne adéquation au risque de la tarification en vigueur, le S/P devra être proche de 1 en règle générale. En pratique, le S/P qui est visé doit inclure une notion de prudence, et il est préférable que ce dernier avoisine la valeur 0,8 ou moins. Un regard rapide sur la répartition des S/P nous permettra d'ores et déjà d'obtenir quelques résultats. Dans notre exemple, l'étude des coefficients et de ce tableau nous permet de remarquer que la zone 7 (DOM-TOM) est très fortement sur tarifée. Nous remarquons également la présence de variations importantes dues à la présence de la date de première mise en circulation dans la modélisation alors que cette variable est absente dans la tarification.

Afin de quantifier concrètement la sur ou sous tarification, et dans le but de comparer des données comparables, nous considérons la table de données comportant l'ensemble des variables explicatives utilisées jusqu'à présent ainsi que la prime pure théorique et la prime pure réelle pour chaque catégorie de risque. A partir de ce tableau de données, nous allons réaliser deux régressions linéaires utilisant les mêmes variables explicatives et visant à expliquer d'une part la prime pure théorique et d'autre part la prime pure réelle. Bien entendu nous assumons le fait de perdre une certaine quantité d'information, mais cette quantité se révèlera réellement négligeable.

Notre objectif étant de quantifier des évolutions de tarifs permettant de corriger la sur ou sous tarification, nous allons faire en sorte d'effectuer une modélisation permettant de reproduire la logique des tables tarifaires en vigueur. Cette logique utilise les interactions entre les variables chaChoixTarif, Groupe et Classe, puis entre chaChoixTarif, Zone et Usage, auxquels se rajoutent les Options, puis le CRM et le coefficient tarifaire (qui seront en réalité considérés comme tel, soit sans coefficient de modélisation particulier à estimer).

En effet, en ce qui concerne le coefficient tarifaire, nous avons pu remarquer que ce dernier consistait plus en un coefficient d'ordre commercial qu'en une réelle mesure du risque. Cette décision est cependant plus ou moins surprenante pour le CRM. En pratique, il convient d'analyser son impact sur la détermination du risque, afin de comprendre si ce coefficient peut être utilisé comme tel dans la tarification. Dans notre cas, bien que l'on constate suite à l'ensemble des modélisations une augmentation importante du risque pour les CRM élevés (constat à nuancer étant donné le faible effectif de CRM très élevés), une approximation linéaire nous permet de considérer que le CRM permet une bonne évaluation du risque en tant que tel, ce qui est ici avantageux.

Cette dernière modélisation nous permet ainsi de recréer une grille de tarification avec un fonctionnement totalement similaire à celle en vigueur et de quantifier de façon concrète les évolutions à mettre en place (à court terme). Afin d'illustrer l'interprétation des résultats obtenus, nous mettons en avant deux tableaux d'évolutions des tarifs issus de cette méthodologie. Le premier tableau correspond aux évolutions des tarifs de bases demandés en fonction du type de garantie demandée. Le tarif de base en responsabilité civile est déterminé par le groupe (première ligne) tandis qu'il dépend du groupe et de la classe pour les garanties supplémentaires. Le second tableau représente les évolutions des coefficients multiplicatifs attribués en fonction de la zone, de l'usage et du type de garantie.

	G5-11	G12-14	>G15
RC	+25%	+8%	-12%
A	+8%	-13%	-
B	+6%	0%	-
C	+8%	+8%	+6%
D/E	+7%	+15%	-2%
>F	-	-12%	-28%

Illustration 53 : Exemple d'évolution de tarifs (1)

R.C	Z2	Z4	Z5	Z9	Z6	Z7
2	0%	-5%	-1%	-21%	-24%	-45%
3	4%	0%	4%	-18%	-21%	-44%
4	-33%	-36%	-33%	-	-50%	-53%
DOM.	Z2	Z4	Z5	Z9	Z6	Z7
2	0%	-5%	2%	-17%	-18%	-27%
3	7%	2%	9%	-12%	-13%	-23%
4	-27%	-32%	-29%	-	-41%	-54%

Illustration 54 : Exemple d'évolution de tarifs (2)

Nous pouvons remarquer à travers le premier tableau que les groupes et classes faibles apparaissent comme étant sous tarifés (ils nécessitent une augmentation de tarif), contrairement aux classes et groupes élevés. Nous notons également que comme signalé précédemment, la zone 7, spécifique aux DOM-TOM, est réellement sur tarifée, de même que la zone 9 (5/6) et la zone 6. Nous remarquons également que la distinction entre l'usage 2 et l'usage 3, qui est inexistante dans la tarification en vigueur, est pourtant bien réelle, tandis que l'usage 4 parait sur tarifé.

Il ne faut pas perdre de vue que ces évolutions doivent être envisagées dans une globalité car ces dernières peuvent être le fruit d'une réallocation du risque. En effet, nous pouvons constater qu'un individu assuré en RC dont le véhicule est de groupe compris entre 5 et 11, circulant pour un usage 2 et en zone 5 aura une prime qui ne sera que faiblement modifiée (1,25 X 0,79 = 0,99). De plus et en pratique, il faut savoir que certaines sur ou sous tarifications possèdent une explication logique et ne feront pas l'objet d'évolutions de tarifs. Nous avons déjà fait remarquer ceci en ce qui concerne l'option qui vient pénaliser des conducteurs sinistrés, bien que ceux-ci se comporte prudemment par la suite. Nous pouvons également prendre l'exemple de la zone 7, pour laquelle la sur tarification est liée à une concurrence peu présente et à un positionnement stratégique en adéquation avec les prix du marché. De même il faut prendre garde à des évolutions tarifaires qui seraient éventuellement rendues incohérentes suite à une modification de la codification ou à une modification du risque. A titre d'exemple, les évolutions concernant les coefficients propres à la zone sont ici à considérer avec prudence, car le zonier, qui est en cours de modification, sera amené à changer. Enfin, nous faisons remarquer que dans certains cas la sur ou sous tarification peut être sous évaluée. A titre d'exemple, les usages 2 et 3 ne possèdent aujourd'hui aucune différence de tarif, ce qui incite les souscripteurs de contrats automobile à rentrer systématiquement et machinalement l'usage 2 comme caractéristique (pour des raisons de gain de temps ou par inattention). Ainsi la distinction entre l'usage 2 et l'usage 3 apparente dans les évolutions de tarifs préconisés par le modèle est probablement sous estimée (un certain nombre d'usage 3 étant référencés en usage 2).

Il est enfin à noter que les modifications mises en avant ci-dessus concernent des actions de court terme, et ne prennent pas en compte l'ensemble des conclusions. En effet, nous avons vu à travers la modélisation que la date de première mise en circulation du véhicule est une variable beaucoup plus pertinente et significative que le groupe. Ainsi, l'introduction de cette variable dans la tarification pour les années à venir pourrait être bénéfique. Ceci impliquerait cependant une modification profonde de la grille tarifaire, et cette évolution ne peut ainsi s'inscrire que dans une logique de moyen ou long terme. Nous faisons également remarquer que la modélisation mise en avant à titre d'exemple ici porte sur une seule année d'exercice, ce qui peut paraitre insuffisant pour remettre en cause une grille tarifaire. Enfin, nous faisons remarquer que toute évolution des tarifs doit être discutée avec un certain nombre d'entités différentes de l'entreprise, et que dans cette logique, de nombreuses propositions doivent être étudiées, afin de contrôler la répartition d'un rapport S/P prévisionnelle en fonction des évolutions retenues.

CONCLUSION

Nous avons entrevu à travers ce document une méthodologie de l'analyse de la segmentation et de la tarification en assurance automobile. Nous avons tout d'abord compris que dans un contexte hyperconcurrentiel, la maitrise de la segmentation et de la tarification en découlant est primordiale pour préserver son portefeuille ou conquérir de nouveaux assurés au sein de la première source de chiffre d'affaires en assurances de biens et de responsabilité. Nous avons également constaté que dans un univers d'innovation constante, la segmentation des risques semble devenir de plus en plus poussée. A partir de ces constats, nous avons cherché à mettre en avant les différentes étapes de la modélisation du risque automobile. Nous avons ainsi entrevu de façon théorique et illustrée les notions d'extraction de bases de données, d'analyse descriptive et factorielle, puis de modélisation linéaire généralisée.

A travers la logique sous jacente à la mise en place d'une extraction de données, nous avons pu comprendre et préciser l'importance de la détection d'erreurs et de l'épurement des données à partir d'un périmètre d'étude clairement défini. Cette approche a également préciser le rôle de l'actuaire au sein de l'entreprise, qui ne s'arrête pas au seul établissement de résultats techniques, mais qui s'inscrit dans une logique de contrôle, de communication, et d'évolution des bases de données de l'entreprise ou de son fonctionnement en règle général dans une optique de long terme.

Nous avons ensuite mis en avant des techniques de base concernant l'analyse descriptive, qui reste un préliminaire à ne pas négliger de la modélisation du risque. En effet, cette dernière nous apporte un grand nombre d'intuitions concernant la modélisation et la situation du portefeuille et de l'entreprise, tout en permettant de détecter d'éventuelles erreurs ayant préalablement échappées à notre vigilance. Nous avons également entrevu la possibilité de mettre en place un système automatisé permettant un gain de temps considérable quant à cette étude. En complément de cette analyse descriptive, nous nous sommes intéressés plus en détail à l'analyse factorielle. Nous avons ainsi détaillé les notions d'analyses en composantes principales ou ACP, d'analyses factorielles des correspondances ou AFC et d'analyses des correspondances multiples ou ACM. La maitrise de ces différents aspects théoriques nous a alors permis d'obtenir une vision globale quant à l'étude en cours, tout en nous permettant d'affiner différentes intuitions. Ces techniques nous ont également permis de nous intéresser au contrôle de la bonne adéquation du zonier au risque d'assurance automobile, et de mettre en évidence un certain nombre d'améliorations à envisager et de problématiques à soulever dans le cadre d'une réévaluation du zonier en vigueur.

Nous en sommes alors enfin venus à la modélisation linéaire généralisée, en détaillant tout d'abord la théorie de la régression linéaire pour ensuite effectuer une généralisation. Nous avons alors pu constater que les aspects théoriques de ces méthodes sont relativement poussés et variés, à l'image des quantités de graphiques, tests et résidus qu'il est possible d'entrevoir. Nous avons tout de même pu mettre en avant l'ensemble des techniques classiques de la modélisation du risque automobile. Nous avons ainsi utilisé plus particulièrement les loi de Gamma et de Poisson, classiquement employées pour la modélisation de coûts moyens et de fréquences. Nous nous sommes alors intéressés plus particulièrement à la régression de poisson et aux phénomènes de sur ou sous dispersion, récurent dans ce type d'études. Des solutions d'améliorations telles que l'utilisation de la quasivraisemblance, de la loi Binomiale négative ou des modèles à inflation de zéros ont donc été abordées, afin d'affiner la modélisation en corrigeant la sur ou sous dispersion.

En pratique, nous avons constaté que les objectifs d'une telle modélisation étaient multiples. Bien que l'objectif principal ai été la détection de segments sur ou sous tarifés, nous avons remarqué que la modélisation du risque automobile possédait de nombreuses applications. Nous avons ainsi étudié de façon particulière le zonier et les évolutions à y apporter, puis nous nous sommes brièvement intéressés au contrôle des provisions d'ouvertures suite à une déclaration de sinistre. Nous avons également évoqué la possibilité de mettre en place des évolutions autour des tarifs en vigueur, d'une part à court terme en tenant compte de la sur ou sous tarification et en cherchant à reproduire un système de tarification identique, et d'autre part à moyen ou long terme avec l'introduction de variables jugées plus pertinentes à l'issu du modèle que certaines utilisées dans la tarification actuelle, impliquant ainsi une modification importante de la logique de calcul des tarifs.

Nous avons alors à nouveau constaté que les résultats techniques se devaient d'être interprétés et replacés dans le contexte de l'entreprise. En effet, nous avons remarqué à titre d'exemple que la sur tarification mise en évidence pour les DOM-TOM était justifiée par la présence limitée de la concurrence sur ce marché, et que toute réduction de tarif serait alors mal venue. Inversement, dans un contexte de concurrence, il pourrait s'avérer problématique de rehausser les tarifs de segments de populations associés à des baisses de tarifs chez la concurrence ou à une population étant le cœur de cible de la société étudiée. Les dimensions techniques et commerciales ne doivent alors pas être dissociées, et seule la discussion et l'ouverture d'esprit de l'actuaire aux métiers qui l'entourent lui permettront d'analyser et d'interpréter correctement les résultats d'une telle étude.

Enfin, nous pourront conclure que les méthodes évoquées dans ce document, bien que relativement élémentaires, permettent de répartir équitablement la charge de sinistre entre les différents assurés, en fonction de leurs caractéristiques et du risque qui leur est associé. Il est alors évident que les primes réellement payées par l'assuré s'écartent parfois significativement des primes théoriques, que ce soit en raison d'une position concurrentielle cohérente avec le marché, d'une volonté de conquérir ou de préserver un segment de tarification, d'obstacles techniques ou de dispositions réglementaires, ou encore à un niveau plus individuel de l'attribution de réduction de tarifs à titre commercial. Cependant et quelque soit le tarif finalement payé par l'assureur, l'actuaire doit être en mesure de comparer les primes réellement payées aux primes théoriques déterminées à partir des méthodes entrevues dans ce document. Ceci permet alors de détecter les segments sur ou sous tarifés, de les quantifier et éventuellement de les justifier. Ce type d'analyse peut alors mener à des évolutions de tarifs pour les segments sur ou sous tarifés sans justification apparente, mais également orienter les attributions de ristournes commerciales octroyées aux assurés ou encore les résiliations de polices associées à des primes trop éloignées de la réalité du risque sous jacent. Nous noterons enfin que l'étude de la tarification est une analyse en mouvement, qui doit sans cesse être remaniée dans le but de s'adapter aux évolutions du portefeuille et du risque automobile.

BIBLIOGRAPHIE

[1] Rapport annuel de la ffsa, **FFSA** (Fédération Française des Sociétés d'Assurances) ,2007.
Rapport annuel de la ffsa, **FFSA** (Fédération Française des Sociétés d'Assurances) ,2008.
« Assurance automobile : les résultats de l'année 2008 », www.ffsa.fr , rubrique Actualité, 01/09/2009.

[2] « Assurance des jeunes : la stratégie du moindre risque », **L'argus de l'assurance** N°7 017, 16/03/2007, p32-35.

[3] « Marketing : La guerre des bonus », **L'argus de l'assurance** N°7 043, 12/10/2007.

[4] « Pay as you drive : ce qu'il rapportera vraiment », **L'argus de l'assurance** N°7 046, 02/11/2007.

[5] « Le Top 20 des assureurs auto en 2007 », **L'argus de l'assurance** N°7 074, 09/05/2008.

[6] « Que valent les assurances auto low-cost ? », **Le Figaro**, 01/08/2008.

[7] « L'assurance automobile personnalise son offre », **Les Echos** N°19 292, 24/11/2004, p13.

[8] « Assurance auto : le temps du sur mesure », **Les Echos** N°19 672, 23/05/2006, p15.

[9] « Les assureurs jouent la carte 'low cost' sur Internet », **Les Echos** N°20 197, 19/06/2008, p32.

[10] « Assurance auto : stabilité des tarifs en 2009 », **Les Echos**, 05/12/2008.

[11] «''Pay as you drive'' Enjeux économiques et technologiques des nouveaux modèles de ''Paiement à l'usage'' dans l'assurance automobile », Livre Blanc **ITN SA**, Décembre 2008.

[12] « La sécurité routière en France. Bilan de l'année 2007 », **ONISR** (Observatoire National Interministériel de la Sécurité Routière), Conférence de presse-Hôtel de Roquelaure, 18/06/2008.
« La sécurité routière en France. Bilan de l'année 2008 », **ONISR** (Observatoire National Interministériel de la Sécurité Routière), Conférence de presse-Hôtel de Roquelaure, 19/06/2009.

[13] Baromètre prospectif 2008, **Observatoire de l'évolution des métiers de l'assurance**, 17/06/2008.

[14] « Explications concernant les systems "Pay as you drive" (PAYD) et l'utilisation de ''boîtes noires'' dans les véhicules automobiles », **PFPDT** (Préposé Fédéral à la Protection des Données et à la Transparence, confédération suisse).
Consultable sur http://www.edoeb.admin.ch/themen/00794/01154/01220/index.html?lang=fr

[15] « MMAbox : le Pay As You Drive selon MMA », www.zerotracas.com, 15/09/2008.

[16] **Allain E., Brenac T. (2001)**, « Modèles linéaires généralisés appliqués à l'étude des nombres d'accidents sur des sites routiers : le modèle de Poisson et ses extensions », Recherche Transports Sécurité N°72, p 3-18.
Consultable et téléchargeable sur http://www.inrets.fr/ur/ma/Equipe/Brenac/RTS-EA-TB-01.pdf

[17] **Benlagha N., Grun-Réhomme M. (2008)**, « Application de la théorie des valeurs extrêmes en assurance automobile », Euro-Mediterranean economics and finance review, vol. 3, N°1, p 40-59.
Consultable et téléchargeable sur http://ermes.u-paris2.fr/doctrav/trav0712.pdf

[18] **Benlagha N., Grun-Réhomme M., Vasechko O. (2008)**, « Les sinistres graves en assurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes », Modulad N°39, p 47-80.
Consultable et téléchargeable sur ;
http://www-roc.inria.fr/axis/modulad/archives/numero-39/Benlagha-39/Benlagha-et-al-39.pdf

[19] **Cameron A., Trivedi P.K. (1990)**, « Regression-based tests for overdispersion in the Poisson Model », Journal of Applied Econometrics, vol. 46, p 347-364.

[20] **Ciarlet P.G. (1998)**, « Introduction à l'analyse numérique matricielle et à l'optimisation », Ed. Dunod.

[21] **Collet D. (2003)**, « Modelling binary data », Ed. Chapman & Hall/CRC, 2nd Edition.

[22] **Confais J., Le Guen M. (2006)**, « Premiers pas en régression linéaire avec SAS », Modulad N°35, p 220-359.
Consultable et téléchargeable sur ;
http://www-roc.inria.fr/axis/modulad/numero-35/Tutoriel-confais-35/confais-35.pdf

[23] **Cook R.D. (1977)**, « Detection of influential observations in linear regression », Technometrics, vol. 19, p 15-18.

[24] **Cornillon P.A., Matzner-Løber E. (2007),** « Régression. Théorie et applications », Ed. Springer.
Consultable sur http://www.scribd.com/doc/14436842/regression

[25] **Dagnelie P. (1998)**, « Statistique théorique et appliquée », Ed. De Boeck Université, vol. 2.
Consultable partiellement en recherche sur http://books.google.fr/

[26] **Daudin J.J., Lebarbier E., Vuillet C. (2007)**, « Bases du modèle Linéaire », Polycopié AgroParisTech.
Consultable et téléchargeable sur http://www.agroparistech.fr/IMG/pdf/ModLin_2007.pdf

[27] **Delwarde A., Denuit M., Serant D.** (Préface) **(2005)**, « Construction de tables de mortalité périodiques et prospectives », Ed. Economica.
Consultable et téléchargeable partiellement sur ;
http://www.actu.ucl.ac.be/staff/denuit/ACTU2122_PartUn.pdf
http://www.actu.ucl.ac.be/staff/denuit/ACTU2122_PartDeux.pdf

[28] **Denuit M., Pitrebois S., Walhin J-F. (2001)**, « Méthodes de construction de systèmes bonus-malus en RC Auto », ACTU-L, vol. 1, p 7-38.
Consultable et téléchargeable sur ;
http://www.secura-re.com/secura/pdf/nopeer/sap-nopeer-2.pdf

[29] **Dodge Y. (2004)**, « Statistique. Dictionnaire encyclopédique », Ed. Springer.
Consultable partiellement en recherche sur http://books.google.fr/

[30] **Dodge Y. (2006)**, « Premiers pas en statistique », Ed. Springer.
Consultable partiellement en recherche sur http://books.google.fr/

[31] **Duby C. (2000)**, « Le modèle linéaire », Polycopié INA P-G Paris.

[32] **Escoffier B. (1965)**, « Analyse des correspondances », Thèse, Faculté des Sciences de Rennes.

[33] **Escoffier B., Pagès J. (1990)**, « Analyses factorielles simples et multiples », Ed. Dunod.

[34] **Escoufier Y. (1985)**, « L'analyse des correspondances, ses propriétés, ses extensions », In : Proceedings 45th session. Institut International de la statistique, 28.2.1-28.2.16.

[35] **Embrechts D., Kluppelberg C., Mikosch T. (1997)**, Ed « Modeling Extremal Events for Insurance and Finance »,. Springer.
Consultable partiellement en recherche sur http://books.google.fr/

[36] **Freund R.J., Wilson W.J., Sa P. (2006),** « Regression analysis: statistical modeling of a response variable », , Ed. Academic Press Inc, 2nd Edition.

[37] **Greene W.H. (1994)**, « Accouting for Excess Zeros and Sample Selection in Poisson and Negative Binomial regression Models », Working Paper EC-94-10, Department of Economics, Stern School of Business, New-York University.
Consultable et téléchargeable sur ;
http://www.stern.nyu.edu/eco/wkpapers/POISSON-Excess_zeros-Selection.pdf

[38] **Greenwood M., Yule G.U. (1920)**, « An inquiry into the nature of frequency distributions of multiple happenings », Journal of the royal statistical society, vol. 83, p 255-279.

[39] **Hauer E., Ng J.C.N., Lovell J. (1988)**, « Estimation of Safety at Signalized Intersections », Transportation Research Record 1185, Transportation Research Board.

[40] **Heiberger R.M., Holland B. (2004)**, « Statistical analysis and data display: an intermediate course with examples in S-Plus, R, and SAS », Ed. Springer.

[41] **Heyde C.C. (1997)**, « Quasi-likelihood and its applications », Ed. Springer.

[42] **Hoaglin D.C., Welsch R.E. (1978)**, « The hat Matrix in regression and ANOVA », The American Statistician, vol. 32, p 17-22.

[43] **Hoerl A.E. (1962)**, « Application of Ridge Analysis to Regression Problems », Chemical Engineering progress, vol. 58, N°3, p 54-59.

[44] **Hoerl A.E., Kennard R.W. (1970),** « Ridge Regression: Biased estimation for nonorthogonal Problems », Technometrics, vol. 12, p 55-67.

[45] **Hoerl A.E., Kennard R.W., Baldwin K.F. (1975)**, « Ridge Regression: Some Simulations », Communications in Statistics, vol. 4, p 105-123.

[46] **Lambert D. (1992)**, « Zero-inflated Poisson regression, with an application to defects in manufacturing », Technometrics, vol. 34, p 1-14.

[47] **Lecoutre E. (2003)**, « Méthodes de statistiques multifactorielles », UCL Institut de statistique. Consultable sur http://www.stat.ucl.ac.be/ISpersonnel/lecoutre/stats/ACP/index.html

[48] **Lejeune M. (2004)**, « Statistique. La théorie et ses applications », Ed. Spinger. Consultable partiellement en recherche sur http://books.google.fr/

[49] **McCullagh P., Nelder J.A. (1989)**, « Generalized Linear Models », Ed. Chapman & Hall/CRC, 2nd Edition.

[50] **Millot G. (2009)**, « Comprendre et réaliser les tests statistiques à l'aide de R : Manuel pour les débutants », Ed. De Boeck Université.
Consultable partiellement en recherche sur http://books.google.fr/

[51] **Pagès J. (2004)**, « Analyse factorielle de données mixte », Revue de statistique appliquée, vol. 52, N°4, p93-111.

[52] **Pitrebois S., Denuit M., Walhin J-F. (2003)**, « Tarification automobile sur données de panel », Bulletin des Actuaires Suisses, p 51-81.
Consultable et téléchargeable sur ;
http://www.secura-re.com/secura/pdf/withpeer/Pitrebois%5B3%5D.pdf

[53] **Rakotomalala R. (2009)**, « Pratique de la Régression Linéaire Multiple. Diagnostic et sélection de variables », Polycopié Université Lumière Lyon 2.
Consultable et téléchargeable sur ;
http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

[54] **Rouvière L. (2009)**, « Régression sur variables catégorielles », Polycopié Université Rennes 2.
Consultable et téléchargeable sur ;
http://www.sites.univ-rennes2.fr/laboratoire-statistique/ROUVIERE/ENSEIGNEMENTS/poly_freg_var_quali.pdf

[55] **Sakia R.M. (1992)**, « The Box-Cox transformation technique: a review », The Statistician, vol. 41, N°2, p 169-178.

[56] **Saporta G. (2006)**, « Probabilités, analyse des données et statistique », Ed. TECHNIP.
Consultable partiellement en recherche sur http://books.google.fr/

[57]**Schervish M.J. (1995)**, « Theory of statistics », Ed. Springer.
Consultable partiellement en recherche sur http://books.google.fr/

[58] **Shankar V., Milton J., Mannering F. (1997)**, « Modeling accident frequencies as zero-altered probability processes: an empirical inquiry », Accident Analysis and Prevention, vol. 29, N°6, p 829-837.

[59] **Shapiro S.S., Wilk M.B. (1965)**, « An analysis of variance test for normality (complete samples) », Biometrika, vol. 52, N°3/4, p 591-611.

[60] **Tufféry S. (2007)**, « Data mining et statistique décisionnelle: l'intelligence des données », Ed. Technip.
Consultable partiellement en recherche sur http://books.google.fr/

[61] **Vasechko O.A., Grun-Réhomme M., Benlagha N. (2009)**, « Modélisation de la fréquence de sinistres en assurances automobile », Bulletin Français d'Actuariat, vol. 9, N°18, p 41-63.
Consultable et téléchargeable sur ;
http://www.lynxial.fr/clients%5Cia%5Csitebfa.nsf/0/2E4F910EB1B2D2B4C125765D00782F40/$FILE/VASECHKO_GRUN%20REHOMME_BENLAGHA.pdf?OpenElement

[62] **Vuong Q.H. (1989)**, « Likelihood Ratio tests for Model Selection and Non-Nested Hypotheses », Econometrica, vol. 57, p 307-333.

[63] **Wedderburn R.W.M. (1974)**, « Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method », Biometrika 61, p 439-447.

[64] **Yang Z., Hardin J.W., Addy C.L., Vuong Q.H. (2007)**, « Testing approaches for Overdispersion in Poisson Regression versus the Generalized Poisson Model », Biometrica, vol. 49, p 565-584.

[65] **Yau K.K, Wang K., Lee A.H. (2003)**, « Zero-Inflated Negative Binomial Mixed Regression Modelling of Over-Dispersed Count Data with Extra Zeros », Biometrica, vol. 45, p 437-452.

Sites Internet de références

[A] Site internet de la ffsa : http://www.ffsa.fr

[B] Site internet de la sécurité routière : http://www.securiteroutiere.gouv.fr/

[C] Site internet de l'observatoire de l'évolution des métiers de l'assurance : http://www.metiers-assurance.org/

[D] Site de référence pour l'enseignement et la pratique du logiciel R : http://pbil.univ-lyon1.fr/R/enseignement.html

[E] Site sur les mathématiques appliquées : http://www.sciences.ch/

[F] Site de wikipedia et particulièrement : http://fr.wikipedia.org/wiki/Régression_linéaire_multiple ou encore http://fr.wikipedia.org/wiki/Régression_logistique

[G] Site concernant le Bulletin Français d'Actuariat : http://www.lynxial.fr/clients/ia/sitebfa.nsf

[H] Site de l'Argus de l'assurance : http://www.argusdelassurance.com/

[I] Site du quotidien Les Echos : http://www.lesechos.fr/

[J] Site de recherche d'ouvrage : http://books.google.fr/

[K] Site contenant un grand nombre de revue : http://economix.u-paris10.fr/

[L] Site de Modulad contenant des archives : http://www-roc.inria.fr/axis/modulad/index.htm

[M] Site de l'ASA (American Statistical Association), informations et possibilité de souscription pour Technometrics ou The American Statistician entre autre : http://www.amstat.org/index.cfm

[N] Site de The Econometric Society : http://www.econometricsociety.org/

[O] Site de Biometrika : http://biomet.oxfordjournals.org/

[P] Site du logiciel R téléchargeable gratuitement : http://www.r-project.org/

ANNEXES

Annexe 1 : Articles de loi concernant le coefficient bonus malus.

Article A121-1 En savoir plus sur cet article...
Modifié par Arrêté 2007-07-19 art. 1 1°, art. 2 1° 2° JORF 21 juillet 2007

Les contrats d'assurance relevant des branches mentionnées au 3 et au 10 de l'article R. 321-1 du code des assurances et concernant des véhicules terrestres à moteur doivent comporter la clause de réduction ou de majoration des primes ou cotisations annexée au présent article.

Sauf convention contraire, la clause visée au premier alinéa n'est pas applicable aux contrats garantissant les véhicules, appareils ou matériels désignés par les termes ci-après, tels que définis à l'article R. 311-1 du code de la route : cyclomoteur, engin de service hivernal, engin spécial, motocyclette légère, quadricycle léger à moteur, quadricycle lourd à moteur, véhicule de collection, véhicule d'intérêt général, véhicule d'intérêt général prioritaire, véhicule d'intérêt général bénéficiant de facilités de passage, véhicule et matériel agricoles, matériel forestier, matériel de travaux publics.

Article Annexe à l'article A121-1 En savoir plus sur cet article...
Créé par Arrêté du 31 octobre 2003 - art. Annexe, v. init.

Art. 1 er.-Lors de chaque échéance annuelle du contrat, la prime due par l'assuré est déterminée en multipliant le montant de la prime de référence, telle qu'elle est définie à l'article 2, par un coefficient dit coefficient de réduction-majoration », fixé conformément aux articles 4 et 5 suivants. Le coefficient d'origine est de 1.

Art. 2.-La prime de référence est la prime établie par l'assureur pour le risque présentant les mêmes caractéristiques techniques que celles présentées par l'assuré et figurant au tarif communiqué par l'assureur au ministre chargé de l'Économie et des Finances dans les conditions prévues à l'article R. 310-6.
Les caractéristiques techniques concernent le véhicule, la zone géographique de circulation ou de garage, l'usage socioprofessionnel ou le kilométrage parcouru, éventuellement la conduite exclusive du véhicule, ainsi que les réductions éventuelles figurant au tarif des entreprises d'assurance.
Cette prime de référence ne comprend pas les majorations éventuellement prévues pour les circonstances aggravantes énumérées à l'article A. 335-9-2 du Code des assurances. En revanche, pour l'application des dispositions de la clause, cette prime de référence comprend la surprime éventuellement prévue pour les conducteurs novices à l'article A. 335-9-1 du Code des assurances ainsi que les réductions éventuelles mentionnées à l'article A. 335-9-3.

Art. 3.-La prime sur laquelle s'applique le coefficient de réduction-majoration est la prime de référence définie à l'article précédent, pour la garantie des risques de responsabilité civile, de dommages au véhicule, de vol, d'incendie, de bris de glaces et de catastrophes naturelles.

Art. 4.-Après chaque période annuelle d'assurance sans sinistre, le coefficient applicable est celui utilisé à la précédente échéance réduit de 5 %, arrêté à la deuxième décimale et arrondi par défaut ; toutefois, lorsque le contrat garantit un véhicule utilisé pour un usage Tournées » ou Tous Déplacements », la réduction est égale à 7 %.
Le coefficient de réduction-majoration ne peut être inférieur à 0, 50. Aucune majoration n'est appliquée pour le premier sinistre survenu après une première période d'au moins trois ans au cours de laquelle le coefficient de réduction-majoration a été égal à 0, 50.

Art. 5.-Un sinistre survenu au cours de la période annuelle d'assurance majore le coefficient de 25 % ; un second sinistre majore le coefficient obtenu de 25 %, et il en est de même pour chaque sinistre supplémentaire.
Le coefficient obtenu est arrêté à la deuxième décimale et arrondi par défaut.
Si le véhicule assuré est utilisé pour un usage Tournées » ou Tous Déplacements », la majoration est égale à 20 % par sinistre.
La majoration est, toutefois, réduite de moitié lorsque la responsabilité du conducteur n'est que partiellement engagée notamment lors d'un accident mettant en cause un piéton ou un cycliste.
En aucun cas le coefficient de réduction-majoration ne peut être supérieur à 3, 50.
Après deux années consécutives sans sinistre, le coefficient applicable ne peut être supérieur à 1.

Art. 6.-Ne sont pas à prendre en considération pour l'application d'une majoration les sinistres devant donner lieu ou non à une indemnisation, lorsque :
1 º l'auteur de l'accident conduit le véhicule à l'insu du propriétaire ou de l'un des conducteurs désignés, sauf s'il vit habituellement au foyer de l'un de ceux-ci ;
2 º la cause de l'accident est un événement, non imputable à l'assuré, ayant les caractéristiques de la force majeure ;
3 º la cause de l'accident est entièrement imputable à la victime ou à un tiers.

Art. 7.-Le sinistre survenu à un véhicule en stationnement par le fait d'un tiers non identifié alors que la responsabilité de l'assuré n'est engagée à aucun titre, ou lorsque le sinistre mettant en jeu uniquement l'une des garanties suivantes : vol, incendie, bris de glace, n'entraîne pas l'application de la majoration prévue à l'article 5 et ne fait pas obstacle à la réduction visée à l'article 4.

Art. 8.-Lorsqu'il est constaté qu'un sinistre ne correspond pas à la qualification qui lui avait été donnée initialement, la rectification de la prime peut être opérée soit par le moyen d'une quittance complémentaire, soit à l'occasion de l'échéance annuelle suivant cette constatation.
Aucune rectification de prime ne sera, toutefois, effectuée si la constatation est faite au-delà d'un délai de deux ans suivant l'échéance annuelle postérieure à ce sinistre.

Art. 9.-La période annuelle prise en compte pour l'application des dispositions de la présente clause est la période de douze mois consécutifs précédant de deux mois l'échéance annuelle du contrat.
Si le contrat est interrompu ou suspendu pour quelque cause que ce soit, le taux de réduction ou de majoration appliqué à l'échéance précédente reste acquis à l'assuré mais aucune réduction nouvelle n'est appliquée, sauf si l'interruption ou la suspension est au plus égale à trois mois.
Par exception aux dispositions précédentes, la première période d'assurance prise en compte peut être comprise entre neuf et douze mois.

Art. 10.-Le coefficient de réduction-majoration acquis au titre du véhicule désigné au contrat est automatiquement transféré en cas de remplacement de ce véhicule ou en cas d'acquisition d'un ou plusieurs véhicules supplémentaires.
Toutefois, le transfert de la réduction n'est applicable que si le ou les conducteurs habituels du ou des véhicules désignés aux conditions particulières du contrat demeurent les mêmes, sauf en cas de réduction du nombre des conducteurs.

Art. 11.-Si le contrat concerne un véhicule précédemment garanti par un autre assureur, le coefficient de réduction-majoration applicable à la première prime est calculé en tenant compte des indications qui figurent sur le relevé d'informations mentionné à l'article 12 ci-dessous, et des déclarations complémentaires de l'assuré.

Art. 12.-L'assureur délivre au souscripteur un relevé d'informations lors de la résiliation du contrat par l'une des parties et dans les quinze jours à compter d'une demande expresse du souscripteur.
Ce relevé comporte notamment » les indications suivantes :
-date de souscription du contrat ;
-numéro d'immatriculation du véhicule ;
-nom, prénom, date de naissance, numéro et date de délivrance du permis de conduire du souscripteur et de chacun des conducteurs désignés au contrat ;
-nombre, nature, date de survenance et conducteur responsable des sinistres survenus au cours des cinq périodes annuelles précédant l'établissement du relevé d'informations, ainsi que la part de responsabilité retenue ;
-le coefficient de réduction-majoration appliqué à la dernière échéance annuelle ;
-la date à laquelle les informations ci-dessus ont été arrêtées.

Art. 13.-Le conducteur qui désire être assuré auprès d'un nouvel assureur s'engage à fournir à celui-ci le relevé d'informations délivré par l'assureur du contrat qui le garantissait précédemment, au souscripteur de ce contrat.

Art. 14.-L'assureur doit indiquer sur l'avis d'échéance ou la quittance de prime remis à l'assuré :
-le montant de la prime de référence ;
-le coefficient de réduction-majoration prévu à l'article A. 121-1 du Code des assurances ;
-la prime nette après application de ce coefficient ;
-la ou les majorations éventuellement appliquées conformément à l'article A335-9-2 du Code des assurances ;
-la ou les réductions éventuellement appliquées conformément à l'article A. 335-9-3 du Code des assurances.

Annexe 2 : Délibération de la CNIL (Commission Nationale Informatique et Liberté) au sujet de la géolocalisation de conducteurs.

Délibération CNIL du 17 Novembre 2005 :

Par une délibération du 17 novembre dernier, la CNIL a refusé la mise en œuvre d'un dispositif qui aurait conduit un assureur à pouvoir géolocaliser de façon permanente des jeunes conducteurs. Ce refus est basé, d'une part, sur le traitement systématique de données relatives aux dépassements de limitations de vitesse et, d'autre part, sur le caractère disproportionné du dispositif au regard de sa finalité.

La CNIL a été saisie d'un projet concernant une nouvelle offre d'assurance automobile à destination des jeunes conducteurs qui y auraient volontairement souscrit. Cette nouvelle offre repose principalement sur l'engagement pris par le jeune conducteur de respecter un certain nombre de règles parmi lesquelles figure le respect des limitations de vitesse et un temps de conduite limité.

Afin de vérifier le respect des ces engagements pouvant conduire à une baisse de la surprime appliquée aux jeunes conducteurs, la compagnie d'assurance demande aux assurés d'équiper leur véhicule d'un dispositif de géolocalisation de type GPS-GSM. En collectant les informations relatives aux déplacements du véhicule toutes les deux minutes, ce dispositif et le traitement, qui lui est associé, permettent à la compagnie d'assurance de déterminer la localisation du véhicule, les vitesses pratiquées, le type de route sur lequel roule le véhicule ainsi que les horaires et les durées de conduite.

Si on ne peut que souscrire à des actions de prévention routière, la CNIL se doit néanmoins de vérifier la conformité des traitements qui lui sont présentés avec les dispositions de la loi "informatique et libertés". Au cas présent, si la géolocalisation de véhicules n'est pas, en soi, contraire aux dispositions de la loi, la Commission a refusé la mise en œuvre du traitement présenté pour deux raisons.

En premier lieu, le traitement, qui a notamment pour objet de collecter de manière systématique les vitesses maximales pour les comparer aux vitesses autorisées, constitue un traitement portant sur des données relatives à des infractions, à savoir les éventuels dépassements des limitations de vitesse. Or l'article 9 de la loi du 6 janvier 1978 interdit à des personnes privées de faire de tels traitements.

En second lieu, la mise en œuvre d'un traitement permettant d'enregistrer l'intégralité des déplacements effectués par les assurés ne répond pas à l'exigence de proportionnalité posée par la loi. L'atteinte à la liberté d'aller et venir anonymement auquel pouvait conduire la mise en place du dispositif de géolocalisation est telle qu'elle ne peut être justifiée par la nécessité de contrôler le respect d'engagements pris par l'assuré.

Cette décision s'inscrit dans la réflexion menée par la CNIL sur les limites dans lesquelles peut s'effectuer le traitement de la donnée de géolocalisation des personnes qui, par nature, présente une sensibilité particulière. Elle souligne aussi que le consentement ne suffit pas à rendre légitime tout traitement de données personnelles.

Source : http://www.cnil.fr/

Annexe 3 : Démonstration de la propriété 3 : Théorème spectral en dimension finie.

On démontre ici le théorème énoncé comme suit ;

Soit A une matrice réelle symétrique, alors il existe une matrice P orthogonale et une matrice D diagonale dont tous les coefficients sont réels et telles que $P^{-1}AP = D$. Autrement dit, toute matrice symétrique réelle admet une base de vecteurs propres orthonormale dont les valeurs propres associées sont réelles.

Nous allons alors procéder en plusieurs étapes.

Etape 1 : Toutes les valeurs propres d'une matrice symétrique sont réelles.

Soit u un vecteur propre de A a priori complexe, associé à la valeur propre λ. On note $\bar{u}$ son vecteur conjugué. On a alors $\bar{u}^T Au = \bar{u}^T \lambda u = \lambda \|u\|^2$. On sait de plus que A étant symétrique et réelle, $A^T = A$ et $\bar{A} = A$, d'où $\bar{u}^T Au = \bar{u}^T \bar{A} u = \overline{(Au)}^T u = \overline{(\lambda u)}^T u = \bar{\lambda} \bar{u}^T u = \bar{\lambda} \|u\|^2$. u étant non nul, on en déduit que $= \bar{\lambda}$, les valeurs propres sont donc réelles.

Etape 2 : Deux vecteurs propres distincts de A sont orthogonaux.

Soit λ et μ deux valeurs propres distinctes de A associées aux vecteurs propres u et v. On a alors ;

$$\lambda v^T u = v^T \lambda u = v^T Au = (Av)^T u = \mu v^T u$$

Ainsi on a l'égalité $(\lambda - \mu) v^T u = 0$ ce qui implique $v^T u = 0$ et donc le fait que u et v soit orthogonaux.

Pour la suite, on pose V un sous espace vectoriel de $\mathbb{R}^n$ invariant par A ($\forall v \in V, Av \in V$).

Etape 3 : L'orthogonale de V notée $V^\perp$ est invariant par A.

Soit u un vecteur de V et v un vecteur de l'orthogonal de V. Alors $u^T(Av) = (Au)^T v = 0$ car $Au \in V$ par invariance. On en déduit donc que $Av \in V^\perp$ et donc que l'orthogonal de V est invariant par A.

Etape 4 : Si $(w_1 \quad \dots \quad w_k)$ est une base orthonormée de $V^\perp$ alors la matrice de A réduite à $V^\perp$ dans la base $(w_1 \quad \dots \quad w_k)$ est aussi symétrique.

On note b_{ij} les coordonnées de A dans la base en question. On a alors par construction $Aw_j = \sum_{i=1}^{k} b_{ij} \times w_i$ pour tout j=1…k. On en déduit alors $w_m^T Aw_j = b_{mj}$ car la base est orthogonale et on écrit de même $w_m^T Aw_j = (Aw_j)^T w_m = w_j^T Aw_m = b_{jm}$. On en déduit donc que la matrice de A réduite à $V^\perp$ est symétrique également.

Etape 5 : Toute matrice symétrique est diagonalisable dans une base orthonormale.

On raisonne alors par récurrence sur la dimension de la matrice symétrique A. Si A est de dimension 1, alors le résultat est trivial. On suppose alors le résultat vrai pour les matrices symétriques réelles d'ordre inférieur ou égal à k et l'on considère une matrice A symétrique d'ordre k+1.

Hypothèse de récurrence : Toute matrice A symétrique réelle d'ordre inférieur ou égal à k est diagonalisable dans une base orthonormale et il existe ainsi une matrice P orthogonale et donc inversible et une matrice D diagonale telle que $P^{-1}AP = D$. P contient les vecteurs propres de A qui composent la base orthonormale et D contient les valeurs propres de A.

On note W l'espace des vecteurs propres de A, invariant par A. Cette espace est de dimension non nulle (on admet ici que toute matrice réelle admet au moins un vecteur propre).On en déduit que $W^{\perp}$ est également invariant par A grâce à l'étape 3. On sait de plus que $\mathbb{R}^{k+1}$ se décompose en deux sous espace que sont W et $W^{\perp}$.

Si $W^{\perp}$ est de dimension nulle, on a alors $\mathbb{R}^{k+1} = W$, il suffit alors de prendre une base orthonormale de W qui diagonalisera nécessairement A par construction. En effet, soit S la matrice contenant les vecteurs de la base orthonormale en colonne, on a alors $S^{-1}AS = S^{-1}\lambda S = \lambda I_{k+1}$ avec λ le vecteur des k+1 valeurs propres nécessairement réelles d'après l'étape 1.

On suppose alors que la dimension de $W^{\perp}$ est strictement positive. On muni alors $W^{\perp}$ d'une base orthonormale $(u_1 \quad \cdots \quad u_m)$ avec m<=k et on note B la restriction de A à $W^{\perp}$ dans cette base. On sait alors d'après l'étape 4 que B est une matrice symétrique. On utilise ensuite l'hypothèse de récurrence pour dire qu'il existe une matrice H orthogonale telle que $H^{-1}BH$ soit diagonale.

On considère alors une base orthonormale $(w_1 \quad \cdots \quad w_{k+1-m})$ de W et on définit G la matrice définie par les vecteurs $(w_1 \quad \cdots \quad w_{k+1-m}, u_1 \quad \cdots \quad u_m)$. On peut alors écrire la relation (avec λ vecteur de valeurs propres) ;

$$G^{-1}AG = \begin{bmatrix} \lambda I_{k+1-m} & 0 \\ 0 & B \end{bmatrix}$$

On définit alors la matrice $L = \begin{bmatrix} I_{k+1-m} & 0 \\ 0 & H \end{bmatrix}$ et on pose S=GL.

On vérifie alors que S est orthogonale ;

$$S^T S = (GL)^T(GL) = L^T G^T GL = L^T I_{k+1} L = I_{k+1}.$$

On constate enfin que ;

$$\mathrm{S}^{-1}\mathrm{AS} = L^{-1}G^{-1}AGL = \begin{bmatrix} I_{k+1-m} & 0 \\ 0 & H^{-1} \end{bmatrix}\begin{bmatrix} \lambda I_{k+1-m} & 0 \\ 0 & B \end{bmatrix}\begin{bmatrix} I_{k+1-m} & 0 \\ 0 & H \end{bmatrix} = \begin{bmatrix} \lambda I_{k+1-m} & 0 \\ 0 & H^{-1}BH \end{bmatrix}$$

La matrice obtenue est donc bien diagonale, ce qui démontre alors la récurrence et par conséquent le théorème spectral en dimension réelle.

Ce résultat se généralise pour une matrice M-symétrique et une base M-orthogonale sans complexité supplémentaire majeure.

Annexe 4 : Cartes des départements de France.

```
library("ade4") #Pour effectuer des analyses factorielles
data(elec88) #Pour les tracés de cartes de France
area.plot(elec88$area, lab = elec88$lab$dep, clab = 0.6) ;x11()
#Importer table DepReg contenant numéros et noms de départements + régions triée par numéros
area.plot(area.util.class(elec88$area,DepReg$reg),lwdgraph=4,cpoint=1,clab=1)
#Voir l'annexe 5 pour la fonction area.util.class(,)
s.value(elec88$xy,elec88$tab[,1],csize=0,area=elec88$area,add.plot=TRUE,clegend=0)
text(elec88$xy,labels=DepReg$num,col="blue")
```



Illustration 55 : Carte des départements de France métropolitaine

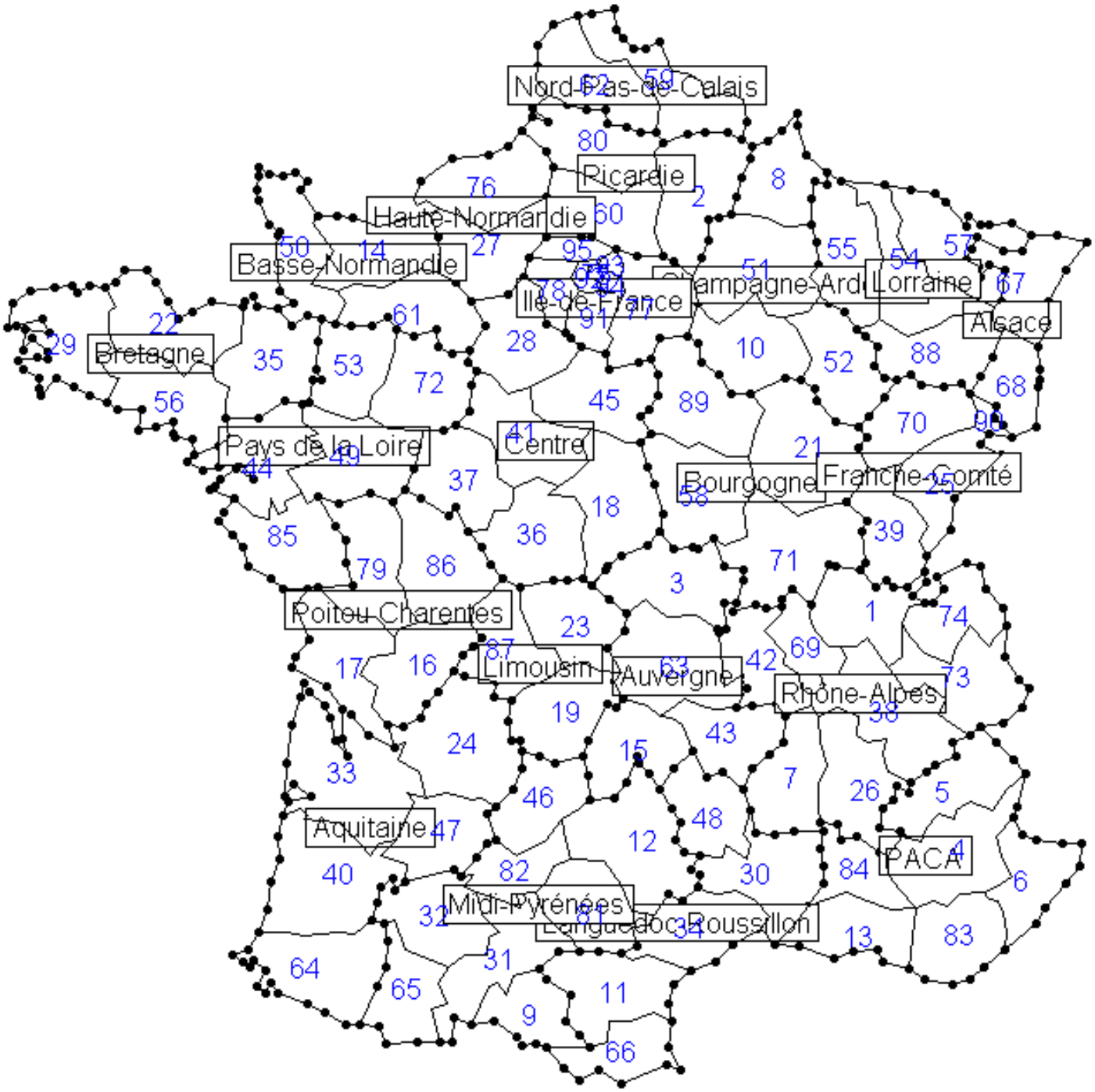


Illustration 56 : Carte des départements et régions de France métropolitaine

Annexe 5 : Représentation d'une ACP sur une carte des régions de France.

Nous avons vu précédemment comment réaliser une ACP et la représenter sur une carte des départements de France métropolitaine grâce au package « ade4 » de R et plus particulièrement à la table « elec88 ». On inscrit ci-dessous une fonction permettant d'obtenir la représentation des régions, que l'on utilise également pour le tracé des cartes des départements de France.

```
area.util.class<-function(area,fac){
if (nlevels(area[,1]!=length(fac)))
stop("non convenient matching")
lreg<-split(as.character(unique(area[,1])),fac)
"contour2poly"<-function(x) {
a=paste(x[,1],x[,2],sep="_")
b=paste(x[,3],x[,4],sep="_")
a=cbind(a,b)
points=a[1,1]
curr=a[1,1]
rowcur=1
colcur=1
npts=nrow(x)
for(k in (1:(npts-2))) {
colnew=3-colcur
curnew=a[rowcur,colnew]
points=c(points,curnew)
a<-a[-rowcur,]
coo=which(a==curnew,arr=TRUE)
rowcur=coo[1,1]
colcur=coo[1,2]
curr=a[rowcur,colcur]}
colnew=3-colcur
curnew=a[rowcur,colnew]
points=c(points,curnew)
return(matrix(as.numeric(unlist(strsplit(points,"_"))),ncol=2,byr=TRUE))}
"souscontour"<-function(k) {
sel=unlist(lapply(lreg[[k]],function(x) which(area[,1]==x)))
area.sel=area[sel,]
area.sel[,1]=as.factor(as.character(area.sel[,1]))
w=area.util.contour(area.sel)
w=contour2poly(w)
w=cbind(rep(k,nrow(w)),w)
return(w)}
lcontour<-lapply(1:nlevels(fac),souscontour)
w=lcontour[[1]]
for(k in 2:length(lcontour)) w<-rbind.data.frame(w,lcontour[[k]])
w[,1]<-as.factor(levels(fac)[w[,1]])
return(w)}
```

Le tracé des département fut lancé par le code « *area.plot(elec88$area,val=acp0$li[,1])* » avec acp0 l'ACP d'un tableau approprié. En prenant cette fois un tableau contenant les régions en individus et en utilisant un tableau DepReg contenant les noms de régions, le code deviendra;

« *area.plot(area.util.class(elec88$area,DepReg$reg),val=acp0$li[,1])* »

On donne enfin un exemple de représentation et d'interprétation succincte que l'on peut ainsi obtenir, sans rappeler le code R préalablement mis en avant dans le paragraphe 2.2.3.3;

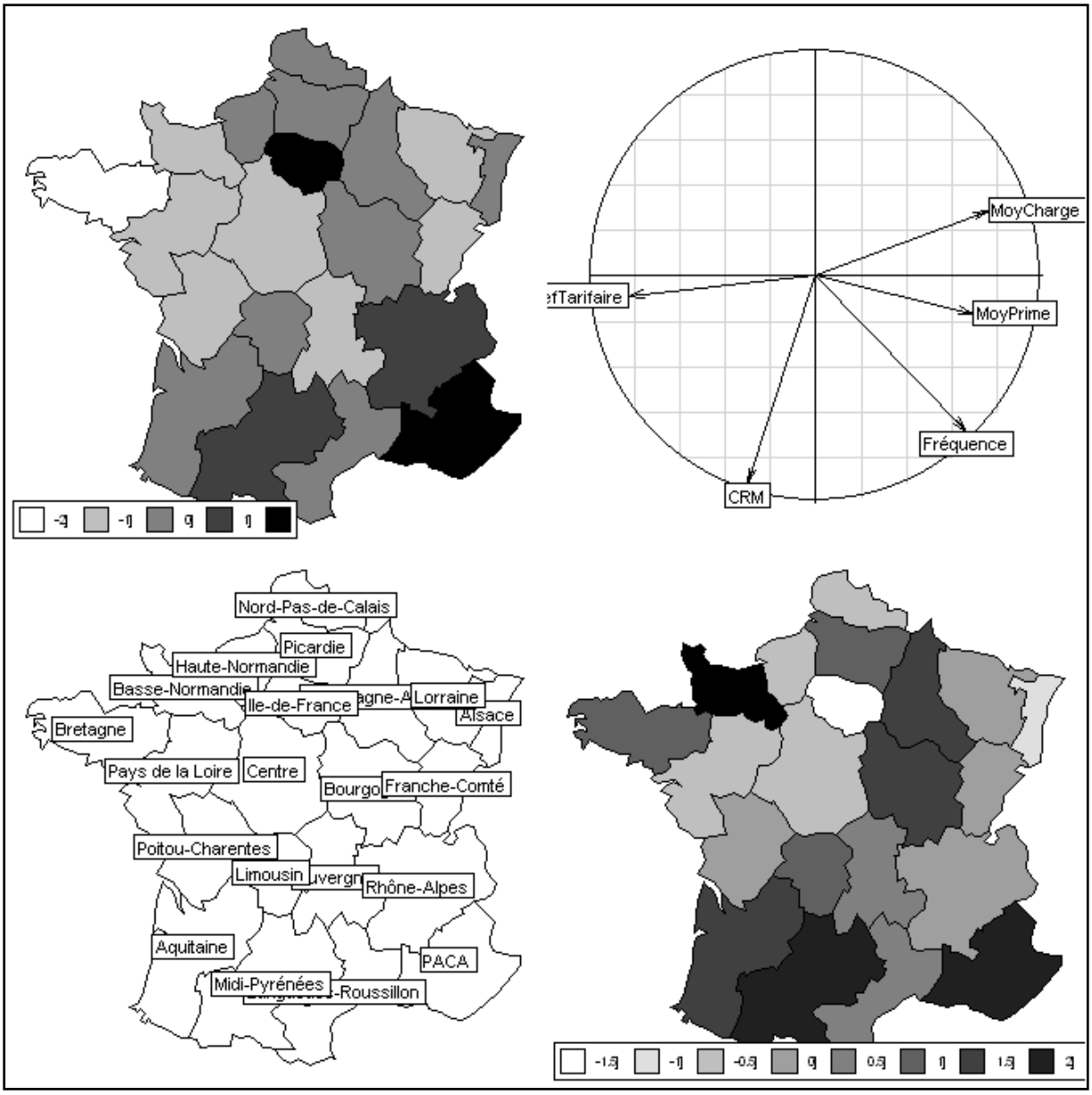


Illustration 57 : Représentation d'un exemple d'ACP sur une carte des régions de France

On constate ici que la région d'île de France, conformément à l'intuition donnée par l'étude des départements, possède une fréquence de sinistres élevée et plus particulièrement dans ce cadre une population à fort CRM (représentation en blanc sur la seconde carte de France). On note qu'elle possède également avec la région PACA une forte charge moyenne (représentation en noir sur la première carte de France). On note ensuite que la région Rhône Alpes est une région à risque (charge moyenne et fréquence élevée). On remarque en revanche que les régions du nord ouest sont porteuses d'un risque faible (représentation en clair sur la première carte et en plus foncé sur la seconde).

Annexe 6 : Exemple d'interface utilisateur pour l'analyse descriptive.

Nous avons entrevu précédemment la possibilité d'automatiser l'étape d'analyse descriptive. Bien qu'il ne soit pas envisageable d'expliciter la totalité de l'automatisation effectuée à l'occasion de l'étude pour des raisons évidentes de clarté et de volume du mémoire, nous représentons ci-dessous trois interfaces proposées à l'utilisateur du programme et réalisées sous EXCEL.

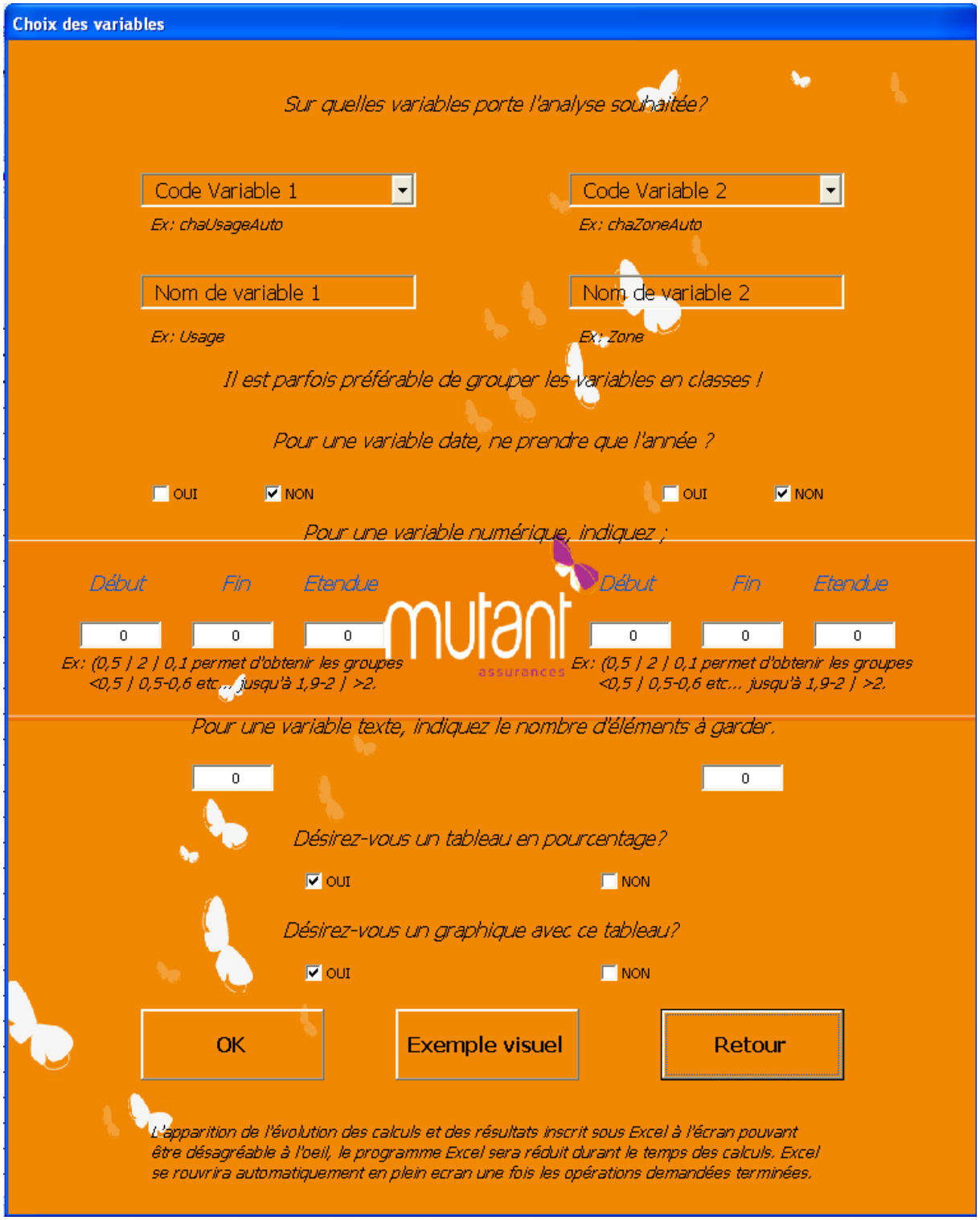


Illustration 58 : Exemple d'interface utilisateur pour l'automatisation de l'analyse descriptive

L'interface présentée ci-dessus permet à l'utilisateur de spécifier un certain nombre de renseignements afin de réaliser un tableau croisé de deux variables ou tableau de contingence reflétant la répartition de l'exposition réelle au risque entre les différentes occurrences des variables choisies. Les listes déroulantes contiennent l'ensemble des variables qui peuvent être choisies pour l'analyse et qui dépendent de la table qui a été téléchargée auparavant. Les zones de textes suivantes représentent le nom que l'on souhaite donner à la variable. Dans l'hypothèse où la variable choisie est une date, il est parfois préférable de n'en sélectionner que l'année, cette possibilité est donc offerte à l'utilisateur. Si la variable choisie est numérique, il peut alors être intéressant d'effectuer un regroupement (on préférera par exemple étudier des tranches de CRM que toutes les occurrences possibles du CRM), cette possibilité est donc également donnée à l'utilisateur qui peut renseigner un nombre de départ, un nombre de fin, et l'étendue des classes qui doivent être formées. Enfin, dans l'hypothèse ou la variable choisie seraient une variable texte contenant un grand nombre de possibilités (la marque de la voiture à titre d'exemple), la possibilité de ne sélectionner qu'une partie des occurrences est offerte à l'utilisateur (les 10 occurrences les plus importantes par exemple). L'utilisateur peut alors demander s'il le désire qu'un tableau supplémentaire en pourcentage par colonne ou un graphique soit réalisé. Le bouton « Exemple visuel » permet à l'utilisateur d'obtenir un aperçu de la forme que prendront les résultats renvoyés. Nous renvoyons ci-dessous l'image renvoyée par ce bouton (représentation très écrasée ici mais suffisante pour la compréhension). Le premier tableau obtenu contient la seconde variable choisie en ligne et la première en colonne, un dégradé de couleurs permet de repérer facilement les occurrences les plus significatives. Le second tableau est identique au premier à ceci près qu'il contient les pourcentages par colonne et que les occurrences associées à une valeur supérieure à 10% sont représentées en rose. Enfin le graphique est en réalité le tracé correspondant au second tableau.

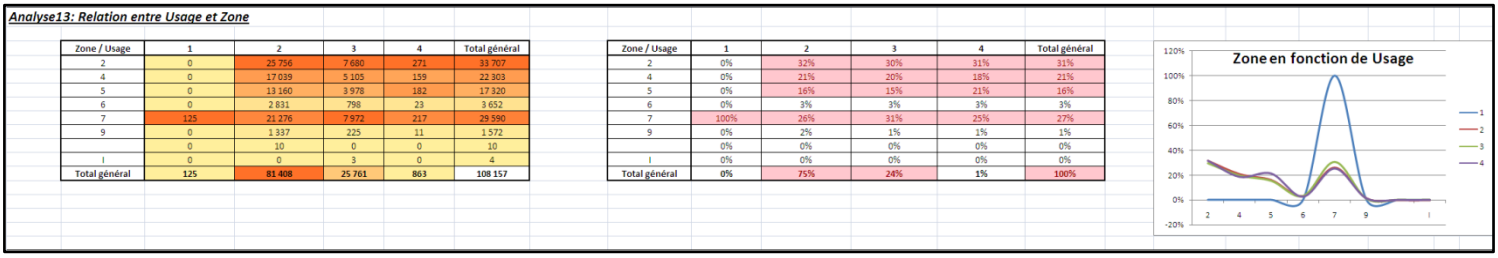


Illustration 59 : Exemple visuel de résultat issu de l'analyse descriptive

Comme nous l'avons vu plus haut, une analyse similaire doit être mise en place autour des sinistres. Contrairement au cas précédent, le contenu des tableaux ne sera pas l'exposition. En effet, dans l'interface représentée ci-dessous, on peut apercevoir qu'un choix devra être fait entre la charge totale, la prime totale, le rapport S/P (qui rappelons-le est calculé sur la base de la prime chargée et non de la prime pure), la fréquence ou la charge moyenne. Une autre particularité est la possibilité de ne sélectionner que les sinistres corporels ou matériels à condition que l'indicateur choisi préalablement soit la charge totale ou la charge moyenne. En dehors de ces particularités, l'interface et les résultats obtenus sont en tout points similaires à l'interface précédente, nous ne décriront donc pas plus cet objet de l'automatisation.

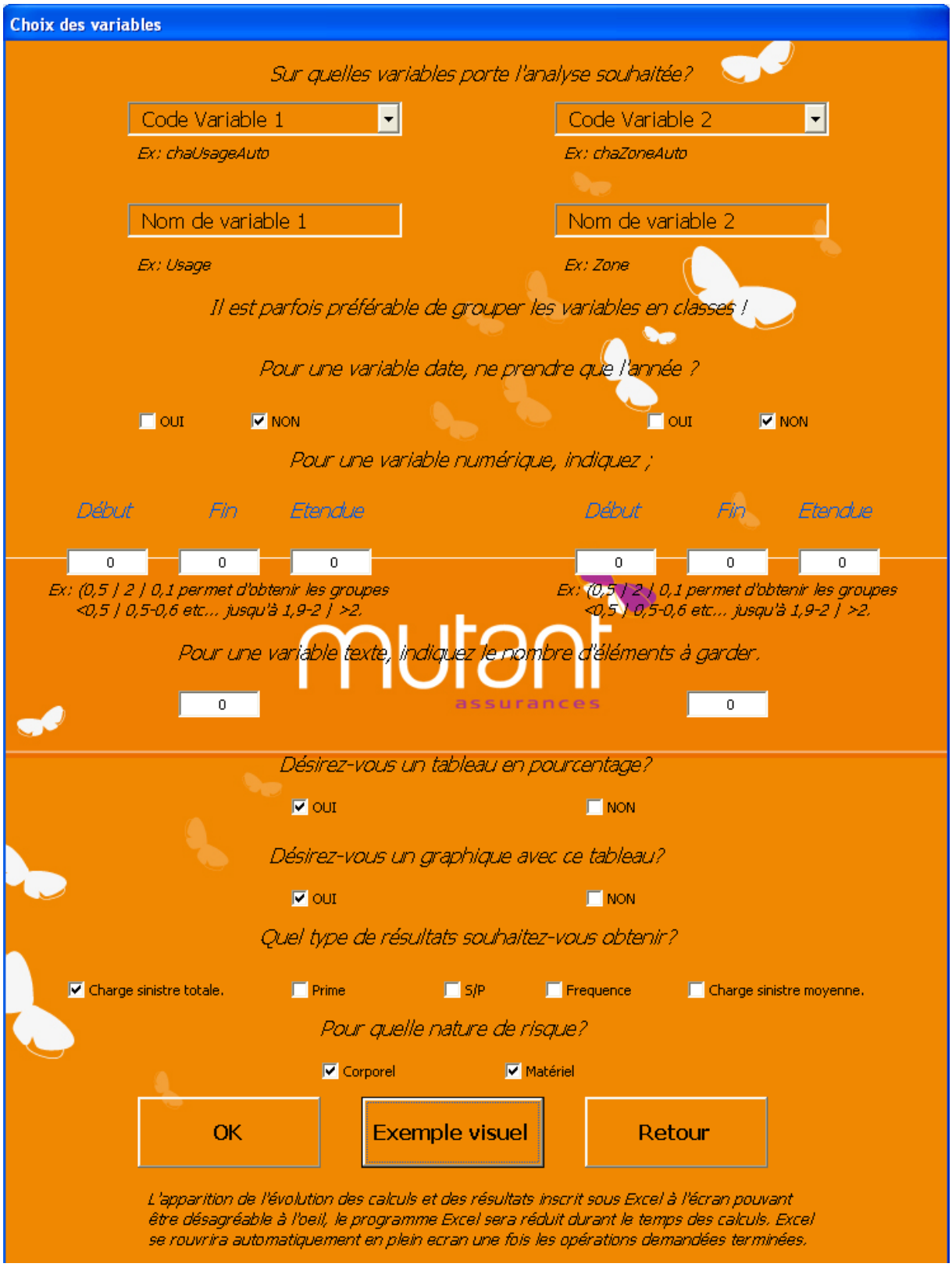


Illustration 60 : Exemple d'interface utilisateur pour l'automatisation de l'analyse descriptive

Nous allons enfin représenter une dernière interface, qui permet de construire un tableau récapitulatif d'une variable contenant la répartition entre les occurrences de la variable du nombre de contrats acquis, du nombre de sinistres, de la fréquence, de la charge totale, de la prime totale, du rapport S/P (qui rappelons-le encore une fois est calculé en fonction de la prime chargée et non de la prime pure), de la charge moyenne et de la prime moyenne. Ce tableau est représenté avec une mise en forme conditionnelle qui inscrit pour chaque colonne les deux plus grandes valeurs en rouge et les deux plus faibles en bleu. Nous représentons ci-dessous l'interface en question ainsi qu'un exemple visuel du résultat obtenu sans descriptions supplémentaires qui s'avèreraient inutiles.



Illustration 61 : Exemple d'interface utilisateur pour l'automatisation de l'analyse descriptive

Analyse12: Tableau récapitulatif pour Option

Option	Contrats acquis	Nombre Sinistres	Freq	Charge totale	Total Prime	S/P	Charge moyenne	Prime moyenne
BON	27 653,16	4 105	14,84%	4 656 874,69	12 366 983,19	0,38	1 134,44	447,22
C07	6 650,59	1 137	17,10%	1 566 941,48	4 839 347,09	0,32	1 378,14	727,66
C10	8 373,24	1 469	17,54%	2 172 267,28	6 203 586,68	0,35	1 478,74	740,88
C12	5 162,76	940	18,21%	1 260 204,89	3 953 468,09	0,32	1 340,64	765,77
C20	1 392,89	327	23,48%	1 620 168,98	1 620 841,29	1,00	4 954,65	1 163,65
CLE	2 142,33	395	18,44%	587 199,94	1 825 210,17	0,32	1 486,58	851,98
ESS	18 889,68	3 070	16,25%	4 387 944,81	9 267 992,34	0,47	1 429,30	490,64
REL	12 053,67	2 282	18,93%	3 544 012,12	10 261 360,21	0,35	1 553,03	851,31
SEL	25 838,85	4 145	16,04%	5 831 743,48	11 524 047,65	0,51	1 406,93	446,00
Total général	108 157,17	17 870	16,52%	25 627 357,67	61 862 836,72	0,41	1 434,10	571,97

Illustration 62 : Exemple visuel de résultat issu de l'analyse descriptive

Annexe 7: Enoncé du théorème de Cochran et démonstration.

Enoncé

Soit n variables aléatoires $Y_1 \quad \cdots \quad Y_n$, indépendantes et de même loi normale centrée réduite. En notant $Y = (Y_1 \quad \cdots \quad Y_n)$, nous avons $\|Y\|^2 = \sum_{i=1}^{n} Y_i^2$, et donc par construction, $\|Y\|^2 \sim \chi_n^2$.

Soit L un sous espace de $\mathbb{R}^n$, de dimension p. On note $P_L(Y)$ la projection orthogonale de Y sur L, ce qui conduit à la décomposition orthogonale suivante ;

$$Y = P_L(Y) \oplus (Y - P_L(Y))$$

Alors les variables aléatoires $\|P_L(Y)\|^2$ et $\|Y - P_L(Y)\|^2$ sont indépendantes et respectivement de loi χ_p^2 et χ_{n-p}^2.

Démonstration

Première étape ;

La démonstration de ce théorème passe par l'utilisation de deux lemmes. Soit $u = (u_1 \quad \cdots \quad u_n)$ et $v = (v_1 \quad \cdots \quad v_n)$ deux vecteurs unitaires de $\mathbb{R}^n$. Alors ;

- $c_u(Y) = \sum_{i=1}^{n} u_i Y_i \sim \mathcal{N}(0,1)\ en\ posant\ P_u(Y) = c_u(Y)u$
- $Cov\big(c_u(Y), c_v(Y)\big) \iff \langle u|v\rangle = 0$

Le premier point découle simplement du fait que $c_u(Y)$ est une combinaison linéaire de lois normales indépendantes. Le deuxième point se démontre très simplement en écrivant les quelques lignes suivantes ;

$$Cov\big(c_u(Y), c_v(Y)\big) = E[c_u(Y) \times c_v(Y)] = \sum_{i,j} E\big[Y_i Y_j\big] u_i v_j = \sum_i u_i v_i = \langle u|v\rangle$$

Deuxième étape ;

Nous munissons l'espace L d'une base orthonormée $(l_1 \quad \cdots \quad l_p)$. Le projeté de Y sur L devient alors la somme des projections sur les axes l_i et l'on écrit ;

$$P_L(Y) = \sum_{i=1}^{p} c_{l_i}(Y) l_i \ \ avec\ c_{l_i}(Y) \sim \mathcal{N}(0,1)$$

Les vecteurs unitaires l_i formant un système orthonormé, les variables $c_{l_i}(Y)$ sont indépendantes entre elles d'après la première étape, ce qui implique que ;

$$\|P_L(Y)\|^2 = \sum_{i=1}^{p} \big\|P_{l_i}(Y)\big\|^2 = \sum_{i=1}^{p} c_{l_i}(Y)^2 \ \ et\ donc\ \|P_L(Y)\|^2 \sim \chi_p^2$$

Troisième étape ;

On complète alors la base orthonormée $(l_1 \quad \cdots \quad l_p)$ en une base de $\mathbb{R}^n$ et on projette sur le sous espace F engendré par la base $(l_{p+1} \quad \cdots \quad l_n)$. Ceci nous permet alors d'écrire que Y= $P_L(Y) + P_F(Y)$ et que $\|Y - P_L(Y)\|^2 = \|P_F(Y)\|^2 \sim \chi^2_{n-p}$. L'orthogonalité des espaces F et L nous permet d'obtenir l'indépendance des vecteurs $P_L(Y)$ et $P_F(Y)$, et par conséquent de leurs normes, ce qui conclut cette démonstration.

Retour sur les modèles linéaires

(Pour le rappel des notations utilisées, le lecteur est prié de se reporter aux *pages 74-75* de ce document)

Le théorème de Cochran est en effet utile dans la théorie des modèles linéaires, car il permet de démontrer la formule fondamentale suivante ;

$$\boxed{SCR(\hat{\theta}) \sim \sigma^2 \chi^2(n-p)\ , de\ plus\ SCR(\hat{\theta})\ et\ \hat{\theta}\ \ sont\ indépendants}$$

Pour ce faire, il suffit de considérer non pas Y, mais $\frac{\varepsilon}{\sigma} \sim \mathcal{N}(0,1)$ et la projection sur l'espace L de dimension p engendré par les vecteurs colonnes de X. La projection sur L de ε correspondant à $\widehat{X\theta} - X\theta$, on obtient alors que $\left\|\frac{\varepsilon}{\sigma} - P_L(\frac{\varepsilon}{\sigma})\right\|^2 = \left\|\frac{Y-\widehat{X\theta}}{\sigma}\right\|^2 = \frac{SCR(\hat{\theta})}{\sigma^2} \sim \chi^2(n-p)$ d'après le théorème de Cochran. On obtient de même l'indépendance entre $SCR(\hat{\theta})$ et $\left\|\widehat{X\theta} - X\theta\right\|^2$ qui équivaut à l'indépendance entre $SCR(\hat{\theta})$ et $\hat{\theta}$, ce qui conclut la démonstration. En effet ;

$$SCR(\hat{\theta}) \perp \left\|\widehat{X\theta} - X\theta\right\|^2 \Leftrightarrow\ \hat{\varepsilon}\ \perp\ \widehat{X\theta} - X\theta \Leftrightarrow E\left[\left(\widehat{X\theta} - X\theta\right)\hat{\varepsilon}\right] = 0\ car\ (\hat{\varepsilon}, \hat{\theta}) gaussien$$

$$SCR(\hat{\theta}) \perp \left\|\widehat{X\theta} - X\theta\right\|^2 \Leftrightarrow E\left[\left(\hat{\theta} - \theta\right)\hat{\varepsilon}\right] = 0 \Leftrightarrow\ \hat{\varepsilon}\ \perp\ \hat{\theta} \Leftrightarrow\ SCR(\hat{\theta}) \perp\ \hat{\theta}$$

Index des illustrations

Illustration 1 : Structure du chiffre d'affaires directes en assurances de biens et de responsabilité en 2008 8

Illustration 2 : Cotisations 2008 en assurances de biens et de responsabilités 8

Illustration 3 : Classement des assureurs automobiles en 2007 par CA et nombre de contrats 9

Illustration 4 : Exemple d'évolution du CRM avec ou sans changement d'assureur 11

Illustration 5 : Schéma récapitulatif du système PAYD 16

Illustration 6 : Formules de calcul du rapport S/P, de la fréquence et autres 22

Illustration 7 : Régression affine d'un nuage de points et visualisation d'un triangle rectangle 28

Illustration 8 : Représentation des valeurs propres d'un exemple d'ACP 33

Illustration 9 : Représentation du nuage de points sur les deux premiers axes principaux d'un exemple d'ACP 34

Illustration 10 : Valeurs propres, nuages de points et cercle de corrélations d'un exemple d'ACP 35

Illustration 11 : Représentation d'un exemple d'ACP sur une carte de France 37

Illustration 12 : Tableau de contingence sur la classe et le groupe du véhicule (exemple d'AFC) 41

Illustration 13 : Représentation des valeurs propres et des nuages de points d'un exemple d'AFC 42

Illustration 14 : Représentation des deux nuages de points d'un exemple d'AFC en tenant compte de la qualité de représentation 43

Illustration 15 : Confrontation du profil lignes et du profil colonnes d'un exemple d'AFC 44

Illustration 16 : Tableaux de contingences et de Burt 45

Illustration 17 : Représentation des valeurs propres d'un exemple d'ACM 47

Illustration 18 : Représentation de chaque variable et de ses modalités avec un nuage de points des individus (exemple d'ACM) 48

Illustration 19 : Exemple d'ACM pour une quantité de variables restreinte 49

Illustration 20 : Représentation des valeurs propres et nuages de points pour le tableau disjonctif complet et le tableau de Burt d'un exemple d'ACM 50

Illustration 21 : Tableau de numérotation des contrats et clients 53

Illustration 22 : Tableau de numérotation des sinistres et vérification de la cohérence de la date d'accident 55

Illustration 23 : Exposition en fonction de Classe , exemple de résultat d'analyse descriptive..... 59

Illustration 24 : S/P, Fréquence et autres caractéristiques principales des occurrences d'une variable (exemple)..... 60

Illustration 25 : Analyse croisée de deux variables, exemple du groupe et de la classe du véhicule . 61

Illustration 26 : Représentation des départements sur le premier plan factoriel et interprétation des axes (exemple d'ACP) 64

Illustration 27 : Carte de France récapitulative d'un exemple d'AFC..... 65

Illustration 28 : Représentation des variables d'un exemple d'AFDM..... 67

Illustration 29 : Première représentation graphique (exemple d'AFDM) 68

Illustration 30 : Deuxième représentation graphique (exemple d'AFDM)..... 69

Illustration 31 : Troisième représentation graphique (exemple d'AFDM) 70

Illustration 32 : Quatrième représentation graphique (Exemple d'AFDM) 71

Illustration 33 : Représentation graphique des sommes de carrés dans le cas avec intercept 78

Illustration 34 : Tableau des différents types de résidus 81

Illustration 35 : Un exemple de courbe sigmoïde 84

Illustration 36 : Lois de la famille exponentielle classiquement utilisée en modélisation linéaire généralisée 88

Illustration 37 : Schéma de construction d'un modèle linéaire généralisé..... 89

Illustration 38 : Fonctions de lien usuelles 90

Illustration 39 : Loi de probabilité pour les modèles ZIP et ZINB 107

Illustration 40 : Description des principales variables explicatives..... 109

Illustration 41 : Répartition des montants de sinistres matériels 111

Illustration 42 : Principe de la fonction « step » avec une méthode descendante..... 113

Illustration 43 : Eléments de la fonction « summary » d'un GLM 114

Illustration 44 : Résultats issus de la fonction « anova » 115

Illustration 45 : Résultats issus de la fonction « drop1 » 116

Illustration 46 : Graphiques d'analyse des résidus..... 116

Illustration 47 : Exemple de résultats issus de la fonction « drop1 » avec interactions..... 117

Illustration 48 : Montants moyens de charge sinistre en responsabilité totale 118

Illustration 49 : L'impact des sinistres graves sur la charge totale..... 119

Illustration 50 : Confrontation de modèles poisson et quasi-poisson, fonction « summary »..... 121

Illustration 51 : Confrontation de modèles poisson et quasi-poisson, fonction « drop1 » 122

Illustration 52 : Exemple d'interface utilisateur pour le calcul de la prime pure..... 124

Illustration 53 : Exemple d'évolution de tarifs (1)..... 125

Illustration 54 : Exemple d'évolution de tarifs (2)..... 125

Illustration 55 : Carte des départements de France métropolitaine 139

Illustration 56 : Carte des départements et régions de France métropolitaine 140

Illustration 57 : Représentation d'un exemple d'ACP sur une carte des régions de France..... 142

Illustration 58 : Exemple d'interface utilisateur pour l'automatisation de l'analyse descriptive 143

Illustration 59 : Exemple visuel de résultat issu de l'analyse descriptive 144

Illustration 60 : Exemple d'interface utilisateur pour l'automatisation de l'analyse descriptive 145

Illustration 61 : Exemple d'interface utilisateur pour l'automatisation de l'analyse descriptive 146

Illustration 62 : Exemple visuel de résultat issu de l'analyse descriptive 146