MASTER ACTUARIAT DE DAUPHINE

Mémoire présenté devant l'Université Paris Dauphine pour l'obtention du diplôme du Master Actuariat et l'admission à l'Institut des Actuaires

le 18 janvier 2016

Par : ATIA Rachel

Titre: Mise en place de modèles de tarification alternatifs face à la suppression règlementaire d'une variable tarifaire en automobile

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

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

Membre présent du jury de l'Institut des Actuaires :

Signature :

Entreprise :

Nom : Optimind Winter

Signature :

Directeur de mémoire en entreprise :

Nom : Mélanie Massias
Matthieu Lagadec

Signature :

Membres présents du jury du Master Actuariat de Dauphine :

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

Signature du responsable entreprise :

Secrétariat :

Bibliothèque :

Signature du candidat :

Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 PARIS C

RESUME

Depuis décembre 2012, la *Gender Directive* impose aux assureurs la suppression du genre de l'assuré de leur modèle de tarification. Cette contrainte règlementaire amène l'assureur à revoir sa segmentation. L'enjeu est de développer des modèles tarifaires reflétant au mieux le risque auquel il est exposé, tout en limitant les hausses tarifaires engendrées chez certaines catégories d'assurés, notamment les jeunes conductrices.

En effet, cette directive s'inscrit dans un contexte hyperconcurrentiel, accentué par la publication de la loi Hamon (2014), qui offre à l'assuré la possibilité de résilier à tout moment son contrat, après un an de souscription.

L'objet de ce mémoire est de proposer des modèles alternatifs afin de limiter l'impact de la suppression de la variable genre.

Après avoir rappelé les notions de tarification utiles à notre étude et présenté la base de données sur laquelle nous travaillerons, nous développerons un modèle non contraint par la *Gender Directive*, qui servira de base comparative aux autres modèles développés par la suite. Puis, nous mettrons en place différents modèles afin de pallier à la suppression de la variable genre.
Suite à ces développements, des analyses de qualité des modèles et des impacts tarifaires seront effectuées. Enfin nous testerons la robustesse de ces nouveaux modèles tarifaires face à des scénarios de modification de portefeuille de l'assureur.

Dans une dernière partie, nous évoquerons la prise en compte de nouveaux critères de segmentation tarifaires, et présenterons le principe du *Pay How You Drive*. La forte concurrence et les évolutions réglementaires poussent en effet les assureurs à s'orienter vers des modèles tarifaires innovants.

Mots clés : Tarification, segmentation, variable tarifaire, *Gender Directive*, Modèles Linéaires Généralisés, prime pure, *Data mining*, classification, régression logistique, Stress Test, *Pay How You Drive*

ABSTRACT

Since december 2012, the Gender Directive requires the insurers to remove the gender of their pricing model. This regulatory constraint leads the insurer to reconsider its segmentation. The challenge is to develop pricing models that best reflect the risk to which he is exposed, limiting the increase of premiums generated for some categories of insured, especially young female drivers.

Indeed, this directive is part of a hypercompetitive environment, enhanced by the publication of the Hamon law (2014), which offers the insured the right to rescind his contract at any time after one year subscription.

The purpose of this thesis is to propose alternative models in order to limit the impact of the removal of the gender variable.

After recalling pricing concepts that will be useful in our study and presenting the database on which we will work, we will develop a model that is not constrained by the Gender Directive, which will be used as a comparative base for the other models developed thereafter. Then we will implement different models to offset the elimination of the gender variable.
Following these developments, quality analysis of the models and tariff impacts will be performed. Finally we will test the robustness of these new pricing models face to portfolio changing scenarios.

In the last part, we will discuss the inclusion of new segmentation criteria in the pricing model and introduce the principle of the Pay How You Drive. Indeed, the strong competition and the growth of regulatory changes forced the insurers to turn to innovative pricing models.

Keywords: Pricing, segmentation, risk factor, *Gender Directive*, Generalized Linear Models, pure premium, Data mining, classification, logistic regression, Stress Test, Pay How You Drive

SYNTHESE

Afin de mieux maitriser le risque auquel il est exposé et de proposer un tarif au plus juste, l'assureur doit procéder à une fine analyse de son risque. Cette analyse repose sur la segmentation de son portefeuille d'assurés en catégories homogènes de risque, c'est-à-dire en sous-ensembles d'individus partageant des comportements similaires en termes de sinistralité. Pour définir ces sous-ensembles, les assureurs disposent de nombreuses informations, concernant l'assuré, le véhicule, mais aussi le passé de sinistralité de l'assuré.

Parmi ces critères, le genre de l'assuré est une donnée déterminante dans la tarification automobile, les données historiques ayant démontré une différence de sinistralité entre homme et femme, particulièrement accentuée chez les jeunes conducteurs.

Depuis décembre 2012, la *Gender Directive* impose aux assureurs de ne plus effectuer de distinction par genre dans l'établissement de leurs tarifs. La loi se veut en effet « plus juste » et veut instaurer un principe d'égalité entre homme et femme dans l'accès aux biens et aux services.

L'assurance automobile fait partie des branches les plus impactées par cette directive, le critère genre étant utilisé dans les modèles tarifaires de la plupart des assureurs. Cette directive amène donc les assureurs à revoir leur segmentation et impacte également les assurés : de fortes augmentations tarifaires sont à prévoir. C'est le cas notamment chez les jeunes conductrices qui payaient en moyenne 20 % moins cher leur prime d'assurance que les jeunes conducteurs.
L'assureur doit donc trouver un modèle de tarification alternatif proposant un tarif ne pénalisant pas les bons risques, sous peine d'assister à un changement de structure de son portefeuille. Le marché automobile est en effet très concurrentiel, d'autant plus depuis la publication de la loi Hamon (2014), qui facilite la résiliation de l'assuré puisqu'il peut désormais résilier son contrat à tout moment dans l'année, après un an de souscription.

La problématique de notre mémoire est de présenter différents modèles de tarifications conformes à la nouvelle règlementation.

Avant toute chose, le lecteur est sensibilisé à l'importance d'une bonne segmentation pour l'assureur (Chapitre 1) et averti des modalités d'application de la *Gender Directive* (Chapitre 2). La théorie des Modèles Linéaires Généralisés, modèles traditionnellement utilisés en assurance automobile, fait également l'objet d'une présentation théorique avant sa mise en application sur notre portefeuille d'assurés (Chapitre 3).

Notre étude s'est basée sur un portefeuille de contrats de garantie Responsabilité Civile, provenant de données publiques Australiennes. Dans un premier temps, la qualité des données a été vérifiée et la base de données a subi différents traitements nécessaires à la mise en place des modèles.

Il a ainsi fallu (liste non exhaustive) :

- regrouper les modalités d'une variable (Chapitre 4)
- partitionner une variable quantitative continue en classes
- s'assurer que les variables disponibles n'étaient pas corrélées (Chapitre 5)

Après avoir réalisé l'ensemble de ces analyses, notre modélisation commence par la mise en place d'un modèle type ayant pu être utilisé par les assureurs automobiles avant l'adoption de la *Gender Directive*.

Les modèles alternatifs devant pouvoir remplacer le modèle d'origine, qui prend en compte le genre de l'individu, il nous servira d'outil comparatif tout au long du mémoire et est appelé **modèle B** (Chapitre 6).

Puisque la *Gender Directive* interdit l'utilisation du critère genre, nous avons décidé de mettre en place, comme premier modèle règlementaire, un modèle différant du premier simplement par la non intégration de la variable « genre » dans la modélisation. C'est le **modèle S**ans **G**enre (Chapitre 7).

S'en est suivi une comparaison de ces deux modèles, aussi bien en termes de qualité des Modèles Linéaires Généralisés sous-jacents que d'impacts sur les tarifs et sur le ratio des Sinistres sur Primes (Chapitre 8). Ces comparaisons ont mis en avant la mauvaise mutualisation engendrée par le modèle qui consiste simplement à supprimer la variable « genre » du modèle de tarification : la perte d'information entraine la création de classes tarifaires plus hétérogènes qu'elles ne l'étaient avec la prise en compte du genre de l'assuré.

Suite à cela, plusieurs modèles alternatifs ayant pour objet de tenter de reproduire l'effet de la variable genre ont étés mis en place (Chapitre 9) :

- un modèle qui mutualise le risque entre les deux genres en pondérant les primes obtenues dans le modèle initial (B) par la proportion Homme/Femme du portefeuille. Ce modèle, appelé **modèle P**, présente l'avantage d'observer un niveau de primes moyen par genre similaire à celui d'avant la mise en place de la *Gender Directive* mais présente le défaut d'entrainer de forts changements tarifaires par sexe. Malgré tout, c'est le modèle qui s'est avéré être le plus concluant en terme d'analyse du S/P.

- un modèle prédictif du genre de l'assuré, avec comme variables de prédiction, les différentes variables disponibles dans la base. Cette fois ci il n'est plus question de reproduire l'effet de la variable genre, comme fait dans le modèle P par pondération, mais de définir un proxy de la variable genre à partir des autres variables (âge de l'assuré, valeur du véhicule…). Cette deuxième alternative a fait l'objet de deux variantes, chacune d'entre elle reposant sur une méthode mathématique qui lui est propre. Ainsi, deux modèles de prédictions ont étés mis en place :

  - le premier repose sur l'utilisation d'un modèle de régression logistique binaire (**modèle PR_L**)
  - le second s'appuie sur un arbre de classification (**modèle PR_CART**).

  Ces deux modèles présentent des taux d'erreur de classement de l'ordre de 37 % : ils manquent de précision et entrainent des changements tarifaires asymétriques entre les deux genres.

Le choix de cette pluralité de méthodes est volontaire : il permet de mesurer les écarts de performance entre les différentes approches mais également de rappeler au lecteur que la résolution d'une problématique peut passer par de nombreuses voies, parfois même très éloignées du domaine de l'assurance.

Cependant, nous ne nous sommes pas arrêtés à la construction de modèles alternatifs : nous avons également souhaité tester leur robustesse. Ce suivi des modèles fait l'objet de la fin du mémoire et débute par la comparaison des effets de leur mise en place aussi bien en termes de politique tarifaire pour l'assureur que d'impact tarifaire pour l'assuré (Chapitre 10).

Suite à l'étude des impacts tarifaire de ces nouveaux modèles, nous avons mis en avant :

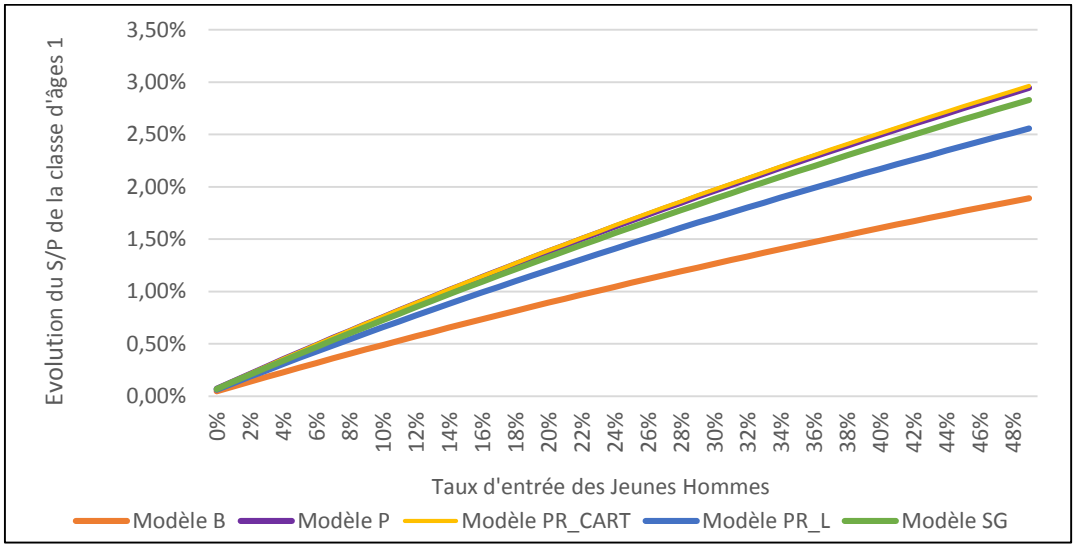
- une hausse généralisée de la prime pure pour les femmes de l'ordre de 4,84 % ;
- à l'inverse, une baisse moyenne de 7,95 % pour la prime pure des hommes.

Classe d'âges	PR_CART		PR_L		SG		P	
	H	F	H	F	H	F	H	F
1 (les plus jeunes)	▼ -10,3 %	▲ 4,45 %	▼ -9,41 %	▲ 5,54 %	▼ -7,61 %	▲ 7,63 %	▼ -8,14 %	▲ 7,01 %
2	▼ -9,6 %	▲ 3,11 %	▼ -8,72 %	▲ 4,13 %	▼ -7,14 %	▲ 5,93 %	▼ -7,12 %	▲ 5,96 %
3	▼ -9,2 %	▲ 2,93 %	▼ -8,46 %	▲ 3,80 %	▼ -7,50 %	▲ 4,88 %	▼ -6,79 %	▲ 5,69 %
4	▼ -10,3 %	▲ 3,94 %	▼ -9,71 %	▲ 4,66 %	▼ -7,06 %	▲ 6,57 %	▼ -7,93 %	▲ 6,73 %
5	▼ -8,7 %	▲ 2,97 %	▼ -7,20 %	▲ 4,67 %	▼ -6,30 %	▲ 5,69 %	▼ -6,40 %	▲ 5,58 %
6 (les plus âgés)	▼ -7,3 %	▲ 1,49 %	▼ -5,03 %	▲ 4,03 %	▼ -4,11 %	▲ 5,04 %	▼ -4,95 %	▲ 4,12 %
Total	▼ -9,4 %	▲ 3,2 %	▼ -8,4 %	▲ 4,4 %	▼ -7,1 %	▲ 5,9 %	▼ -7,0 %	▲ 5,9 %

Tableau 1 - Hausses / Baisse tarifaires (Base comparative : modèle B)

La population des jeunes conducteurs étant la plus concernée par ces changements, nous nous sommes alors intéressés aux effets que la modification de la composition du portefeuille aurait pour l'assureur. Les modèles précédemment créés ont donc subi des stress tests, consistant en la modification de la proportion homme/femme du portefeuille par le biais de simulations d'entrée/sortie de ce dernier (Chapitre 11).

Trois scénarios ont été testés parmi lesquels figure une entrée de jeunes conducteurs (hommes) dans le portefeuille de l'assureur. Les résultats de ce dernier sont résumés par le graphique ci-dessous :



Graphique 1 - Evolution du S / P sous l'hypothèse d'entrée de Jeunes Hommes dans le portefeuille

Nous pouvons observer que les courbes représentatives des différents modèles alternatifs mis en place sont bien éloignées de celle du modèle initialement utilisé par l'assureur (modèle B) : leur pente est plus forte, caractérisant une évolution importante du S/P de la classe des jeunes conducteurs en cas d'entrée de jeunes hommes dans le portefeuille.

Les deux autres Stress-Tests concernent une entrée de jeunes conductrices dans le portefeuille et une inversion de la proportion de jeunes Hommes/jeunes Femmes.
Leur conclusion est similaire à celle du premier scénario : la suppression d'une variable tarifaire a entrainé la création de modèles alternatifs moins robustes.

Malgré tout, c'est le modèle utilisant la prédiction du genre de l'assuré par le biais d'une régression logistique binaire qui s'avère être le plus robuste : c'est le modèle pour lequel le ratio S/P est le moins impacté par les changements de composition du portefeuille appliqués.

Les actuaires, bien conscients que la suppression de variables utilisées dans la segmentation de leur portefeuille entraine la création de classes tarifaires présentant de plus en plus d'hétérogénéité, ont décidé de pousser la segmentation à son paroxysme en proposant de nouveaux systèmes tendant à l'individualisation des tarifs : c'est le cas du *Pay How You Drive* (Chapitre 12).

La nouveauté introduite par cette nouvelle offre de contrat automobile est la prise en compte du comportement du conducteur, en plus des caractéristiques utilisées par les modèles précédents.
La présentation théorique de ce système de tarification en automobile, ainsi que les atouts de ce nouveau système seront exposés dans les toutes dernières pages de ce mémoire.

L'ensemble des travaux menés dans ce mémoire a été réalisé à l'aide du logiciel R et quelques lignes de code ont été renseignées pour permettre aux lecteurs intéressés de reproduire certaines manœuvres si besoin est, aussi bien dans le domaine de l'assurance que d'autres domaines nécessitant le traitement de données.

L'étude de ce mémoire repose sur la suppression de la variable sexe suite à la « Gender Directive » mais peut s'étendre plus généralement à la suppression d'autres variables tarifaires. La Directive Anti-discrimination étudie en effet également la possibilité d'interdire l'utilisation de certains critères tarifaires tels que l'âge de l'assuré.

SYNTHESIS

In order to better control the risk to which he is exposed and to propose a premium as fair as possible, the insurer must undertake a detailed analysis of its risk. This analysis is based on the segmentation of its portfolio in homogeneous risk classes, that is to say, the definition of subsets of individuals sharing similar behavior in terms of sinistrality. To define these subsets, insurers have a lot of information concerning the insured, the vehicle insured, but also the sinistrality past of the insured.

Among these criteria, the sex of the insured is a key information in car pricing: historical data have demonstrated a significant difference in sinistrality between men and women, particularly pronounced for young drivers.

Since december 2012, the Gender Directive requires insurers to no longer make any distinction by gender in setting their premiums. Indeed, the law is to be "fairer" and wants to establish a principle of equality between men and women in access to goods and services.

Car insurance is one of the sectors that is the most impacted by this directive because the gender of the insured was used in pricing models for most insurers. Therefore, this Directive leads insurers to review their segmentation, and also impacts the insured: strong increases of the premium level are expected. This is particularly the case for young female drivers who were paying on average 20% less for their insurance premium than young male drivers.
The insurer must find an alternative pricing model offering a fare level of premium and that is not penalizing good risks, otherwise he could assist to a change in the structure of its portfolio. Car insurance market is highly competitive, especially from the publication of the Hamon law (2014) which facilitated the rescission of an insurance contract, since the insured can now rescind any time in the year, after one year of subscription.

The issue of our theses is to present different pricing models compliant with the new regulation.

First of all, the reader will be aware of the importance of having a good segmentation for the insurer (Chapter 1) and warned about the modalities of application of the Gender Directive (Chapter 2). The theory of Generalized Linear Models, models traditionally used in car insurance, will also be the subject of a theoretical presentation before its implementation on our insured portfolio (Chapter 3).

Our study is based on an historical sinistrality portfolio of Civil Liability contracts from Australian public data. At first, the data quality has been checked and many treatments, necessary for the implementation of the models were applied on the database. We needed to (not exhaustive list):

- aggregate the modalities of a variable (Chapter 4)
- partition a continuous quantitative variable into classes
- ensure that the variables were not correlated (Chapter 5)

After making all of these analyzes, we have started our modelling with the establishment of a standard model that could be used by car insurers before the adoption of the Gender Directive. Knowing that the alternative models have to be able to replace the original model, which takes into account the gender of the individual, this fist model will serve as a comparative tool throughout the memory and is called Model B (Chapter 6).

Since the Gender Directive prohibits the use of such criteria, we then decided to set up as the first regulatory model a model differing from the first simply by the fact it doesn't integrate the variable "gender". This is the model *Sans Genre* (Chapter 7).

This was followed by a comparison of these two models as well as in terms of quality of Generalized Linear Models than in impact on the level of premiums and on the ratio of Claims on Premiums (Chapter 8). These comparisons have highlighted a poor pooling generated by the model that simply consisted in deleting the variable gender from the pricing model: the loss of information leaded to the creation of classes more heterogenic than they were with consideration of the gender of the insured. Following this, several alternative models designed to try to reproduce the effect of the gender variable have been set up (Chapter 9):

- a model that is pooling risk between the two genders by weighting premiums obtained in the initial model (B) by the Male / Female proportion of the portfolio. This model, called Model P, presents the advantage to give a similar average level of premium by gender than before the establishment of the Gender Directive, but leads to significant premium level changes. Whatever, if we consider that the rating factor is the analysis of the Claims on Premiums ratio, this model is the best one.

- a predictive model of the gender of the insured, that considers the variables available in the database as predictive variables. This time, there is no question of reproducing the gender variable, as done by weighting in the model, but to define a proxy of the variable gender thanks to other variables (insured age, vehicle value). This second alternative has been the subject of two variants, based on different mathematical methods. Thus, two prediction models were established:

    - The first one is based on a the use of a binary logistic regression model (model PR_L)
    - The second is based on a classification tree (Model PR_CART).

    Both models have a classification error rate about 37 %: they are not accurate and lead to asymmetric changes of the pure premium level between the two genders.

The choice of this plurality of methods is voluntary: we can measure the performance gaps between the different approaches and we also wanted to remind the reader that the resolution of a problem can go through many paths, sometimes far removed from the insurance field.

However, we did not stop after the construction of alternative models: we also wanted to test their robustness. Monitoring models is the subject of the end of the memory and begins by comparing their effects in terms of policy pricing (Chapter 10).

Following the study of the evolution of premiums due to the adoption of these new models, we have noticed:

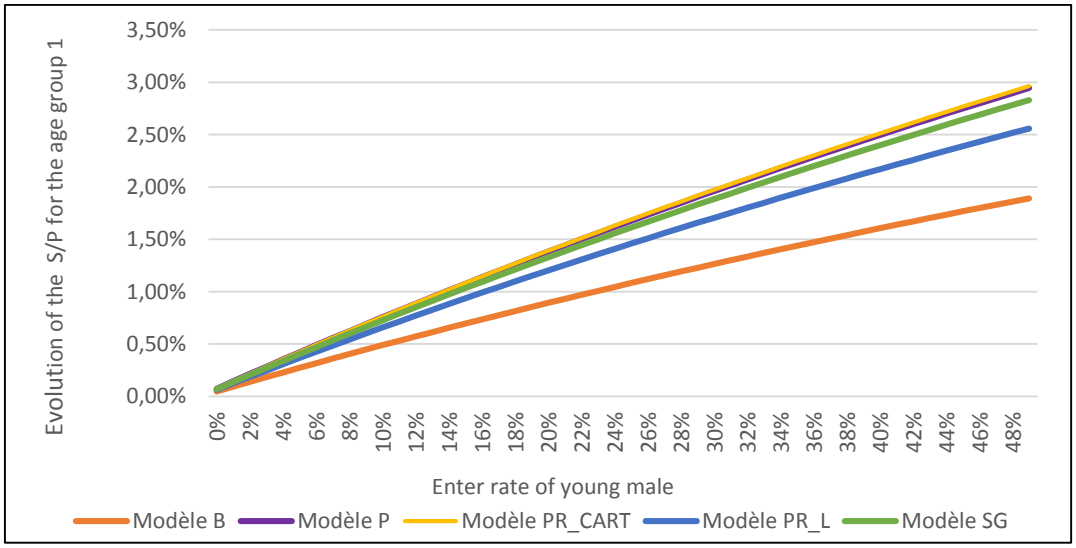
- A widespread increase of 4.84% in the women pure premium;
- Conversely, an average decrease of 7.95% for the men pure premium.

Age groups	PR_CART		PR_L		SG		P	
	H	F	H	F	H	F	H	F
1 (The youngest)	-10,3 %	4,45 %	-9,41 %	5,54 %	-7,61 %	7,63 %	-8,14 %	7,01 %
2	-9,6 %	3,11 %	-8,72 %	4,13 %	-7,14 %	5,93 %	-7,12 %	5,96 %
3	-9,2 %	2,93 %	-8,46 %	3,80 %	-7,50 %	4,88 %	-6,79 %	5,69 %
4	-10,3 %	3,94 %	-9,71 %	4,66 %	-7,06 %	6,57 %	-7,93 %	6,73 %
5	-8,7 %	2,97 %	-7,20 %	4,67 %	-6,30 %	5,69 %	-6,40 %	5,58 %
6 (The oldest)	-7,3 %	1,49 %	-5,03 %	4,03 %	-4,11 %	5,04 %	-4,95 %	4,12 %
Total	-9,4 %	3,2 %	-8,4 %	4,4 %	-7,1 %	5,9 %	-7,0 %	5,9 %

Table 1 - Increases / Decreases in pure premium level (Comparative base: model B)

Since the population of young drivers was the most affected by these changes, we then investigated what would be the effects of a modification of the portfolio composition for the insurer. The models previously created have been submitted to stress tests, consisting of modifying the man/woman portfolio proportion (Chapter 11).

Three scenarios were tested, including the input of young male drivers input in the insurer portfolio. The results are summarized in the graph below:



Graph 1 - Evolution of the S / P under the assumption of Young Men entry in the portfolio

We can observe that the curves representing the different alternative models are far removed from that of the initial model (Model B): their slope is stronger, characterizing an important evolution of the ratio of Claims on Premiums if young men enters into the portfolio.

The other two stress-tests consisted on the entry of young female drivers in the portfolio and a reversal of the proportion of young men / young women.
Their conclusion is similar to the first scenario: the removal of a pricing variable has resulted in the creation of less robust alternative model.

Nevertheless, this is the predictive model based on a binary logistic regression that is the most robust. The ratio of Claims on Premiums is the least impacted by the changes of the portfolio composition for this model.

Well aware that removing variables used in the segmentation process of their portfolios entails the creation of pricing classes more heterogenic, actuaries decided to push segmentation at its peak by proposing new systems seeking the individualization of prices: this is the case of the Pay How You Drive

(Chapter 12). This new car contract offers the possibility to take into account driver behavior, in addition to the features used by previous models.
The theoretical presentation of this automobile pricing system and the advantages of this new system will be exposed in the latest pages of this thesis.

All the work in this thesis was performed using the R software and a few lines of code have been indicated to allow interested readers to replicate some processes if necessary, as well as in the insurance field than in another field requiring data treatment.

The study of this thesis is based on the suppression of the variable sex following the "Gender Directive" but may be extended more generally to the removal of other pricing variables. The Anti - discrimination Directive also studies the possibility of prohibiting the use of criteria such as the age of the insured.

Remerciements

Je tiens tout d'abord à remercier Optimind Winter, pour m'avoir accueilli dans son équipe et offert la possibilité de continuer cette belle aventure.

Je remercie Christophe Eberlé, Président d'Optimind Winter ainsi que Tristan Palerm et Gildas Robert pour leur apport professionnel en tant que Directeurs métier.

Je remercie également mes maitres de stage professionnels, Mélanie Massias et Matthieu Lagadec pour leur encadrement et leurs précieux conseils ainsi que mon tuteur académique Idris Kharroubi pour ses remarques.

Je salue Adrien Suru, Executive Assistant du CFO d'Allianz France, qui a également été mon professeur d'assurance IARD à l'Université Paris Dauphine pour avoir réussi à me transmettre l'intérêt que je porte au secteur de l'assurance Non-Vie.

Enfin j'ai une pensée particulière pour mes parents, qui m'ont soutenue tout le long de la rédaction de ce mémoire et bien au-delà.

Introduction

Depuis 2012, la Commission Européenne a mis en place une directive qui oblige à revoir le système de tarification des assureurs européens. Cette directive, appelée *Gender Directive*, exige une tarification des contrats d'assurance égalitaire face au genre de l'individu. Ainsi, un homme et une femme, présentant les mêmes caractéristiques paieront la même prime d'assurance.

En assurance Non-Vie, et plus particulièrement en assurance automobile, les impacts de cette directive se font sentir. En effet, les statistiques montrent des différences significatives du coût des sinistres en fonction du genre, notamment chez les conducteurs les plus jeunes. Ce nouveau cadre réglementaire impose donc aux assureurs de réviser leur segmentation et leur tarification tout en assurant une bonne mutualisation du portefeuille dans un contexte hyperconcurrentiel.

Rappelons toutefois que si le critère « genre » ne peut plus être utilisé dans la tarification, il peut toujours être demandé à la souscription et peut être utilisé dans la cotation. Autrement dit, il est possible d'établir une prime pure différente selon le genre de l'individu, l'essentiel étant de proposer une prime commerciale identique aux deux genres à la fin.
Par ailleurs, le provisionnement de l'assureur n'est pas concerné par cette directive. Ce décalage entre tarification et provisionnement est l'un des impacts techniques de la *Gender Directive*.

La mise en place de cette directive a un coût, aussi bien pour l'assuré, qui verra son montant de prime évoluer, mais aussi pour l'assureur, qui devra déployer une stratégie marketing, mettre en place un modèle de tarification alternatif mais également faire face à des coûts de gestion et d'administration supplémentaires. De plus, il risque de subir une modification de la composition de son portefeuille en termes de répartition homme/femme mettant en péril sa mutualisation.

L'objet du mémoire est de **proposer des modèles alternatifs à la suppression d'une variable tarifaire en assurance automobile.**

L'étude est axée sur la *Gender Directive*, c'est-à-dire qu'elle porte sur la suppression de la variable « genre » de l'assuré. Cependant, elle peut également être étude étendue au cadre général de la suppression d'autres variables tarifaires, telles que l'âge, variable dont la suppression est en cours d'étude au sein du projet Anti-discrimination. Elle est construite à partir d'un portefeuille d'assurés automobile.

Dans un premier temps, les méthodes de tarification traditionnelles en IARD, notamment les Modèles Linéaires Généralisés, sont appliquées sur la base construite. Ensuite, l'objectif est de proposer des méthodes alternatives pour établir un nouveau tarif indépendant du genre de l'assuré.
Pour cela, trois solutions sont envisagées :

- un modèle tarifaire qui ne tiendrait pas compte du genre de l'assuré
- une mutualisation du risque en pondérant les deux primes obtenues à l'aide du modèle non contraint par la composition Homme/Femme du portefeuille
- l'utilisation d'un modèle prédictif pour le genre de l'assuré

Toutes les études présentées sont effectuées en séparant le traitement de la fréquence et du coût des sinistres. Pour chaque scénario, une étude comparative est effectuée avec le modèle précédant la mise en place de l'interdiction d'utilisation du critère genre. Enfin, les impacts tarifaires pouvant être importants, une simulation du comportement des assurés est mise en place pour s'assurer de la robustesse du modèle.

Sommaire

Si besoin, un rappel de la signification des acronymes utilisés est disponible en Annexe 6.
Au fil du mémoire, les références utilisées seront citées et les initiaux suivant correspondent à :

[O.] Ouvrage [R.] Revue [A.] Article [M.] Mémoire [P.] Polycopié [S.] Site internet [I.] Illustration

Partie I - Cadre théorique

Cette première partie concentre les fondements théoriques sur lesquels sera basée la suite du mémoire.

Nous commencerons par présenter les principes fondamentaux de la tarification, ainsi que la *Gender Directive*, directive qui nous servira d'appui pour illustrer la suppression d'une variable tarifaire, ici, le genre de l'individu. Puis, nous évoquerons les méthodes mathématiques généralement utilisées pour mettre en place un système de tarification, à savoir, les principes de la régression.

Les principes théoriques (tarification et régression) seront évoqués dans un cadre général avec quelques remarques spécifiques à leur application dans le cadre de l'assurance automobile, branche de l'assurance qui fera l'objet de notre mémoire.

Chapitre 1 - Principe de la tarification

Dans ce premier chapitre seront soumises au lecteur quelques généralités concernant la tarification d'un risque (§ 1.1).

Une fois ceci établi, nous présenterons la théorie Fréquence-Coût (§ 1.2), méthode qui consiste à séparer l'évaluation de la charge des sinistres selon ses deux composantes, à savoir le coût et la fréquence des sinistres.

Enfin, nous parlerons de la segmentation tarifaire effectuée par les assureurs dans le but de proposer un tarif au plus juste pour chacun des assurés ainsi que du principe de mutualisation au sein de chaque classe tarifaire sans oublier l'existence d'une asymétrie d'information entre assureur et assuré lors de la souscription d'un contrat (§ 1.3).

1.1 Généralités

1.1.1 La tarification : en tête du processus de conception d'un produit d'assurance

Pour toujours rester concurrentiel et rentable, chaque assureur doit régulièrement mettre au point de nouveaux produits afin de répondre à la demande du marché.
Leur conception repose sur les échanges entre les différentes directions de la compagnie d'assurance.

Dans un premier temps, des études sont lancées afin de définir le besoin des clients ainsi qu'une population cible : c'est la vision marketing. L'étape suivante est la tarification, que nous détaillerons dans ce mémoire. Elle est réalisée par le pôle Actuariat. Ce dernier s'intéresse également au besoin en réassurance et mène des études de rentabilité du produit. Suite à cela, des documents contractuels doivent être rédigés et l'agrément des autorités obtenu. C'est le pôle juridique qui s'assure que les conditions contractuelles représentatives du risque couvert sont claires et conformes à la législation.

Une fois tarifé, le produit d'assurance peut être lancé. Pour sa distribution, le choix du réseau utilisé, réalisé en amont de la tarification, est large : courtiers, agents généraux, réseau bancaire ou encore internet.

Des échanges réguliers entre l'actuaire, qui établit le tarif, et le distributeur ont lieu en amont et en aval de ce processus de conception du produit. Le plus souvent, ce sont les courtiers qui remontent les demandes particulières de leurs clients. Cela entraine parfois la modification de certaines conditions contractuelles. Ces modifications, en fonction de leur ampleur, peuvent être accompagnées ou non d'un ajustement du tarif.

Même après son lancement, le produit est suivi et un *reporting* est réalisé : des rapports actuariels sont établis pour permettre d'analyser la collecte, la satisfaction des clients, mais aussi, d'identifier de nouveaux besoins.

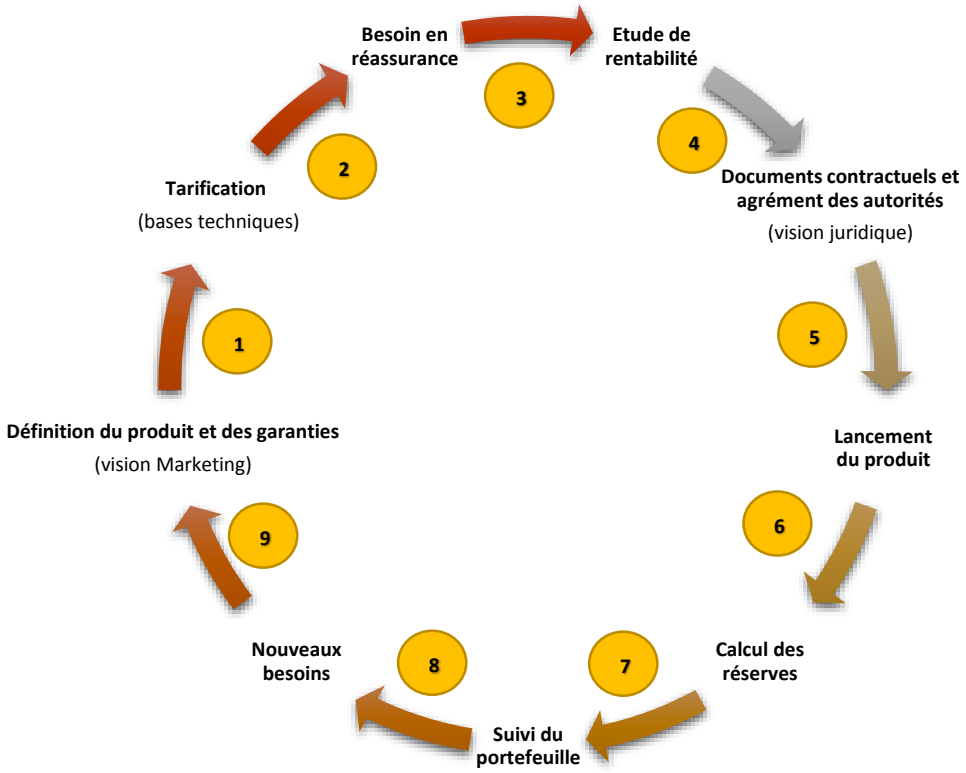


Figure 1- Les étapes du lancement d'un produit d'assurance

1.1.2 Engagements réciproques : prime contre couverture

Une police d'assurance matérialise un contrat passé entre un assureur et un assuré pour une durée déterminée appelée période de garantie ou de couverture.
En assurance automobile, celle-ci est le plus souvent d'un an avec tacite reconduction[1].
Depuis la souscription et jusqu'à échéance du contrat, l'assureur et l'assuré sont engagés réciproquement :

- L'assuré doit payer la prime d'assurance : en totalité à la souscription dans le cas d'une prime annuelle, et chaque mois dans le cas d'un fractionnement mensuel de la prime.
- L'assureur quant à lui, s'engage à payer tout sinistre garanti par le contrat survenant à l'assuré durant la période de couverture.

Ainsi, en contrepartie d'une prime d'un montant connu à la souscription, l'assureur s'engage à couvrir un risque de montant inconnu. C'est l'inversion du cycle de production, caractéristique du secteur de l'assurance.

[1] Clause d'un contrat qui indique que l'accord est reconductible automatiquement d'une période à une autre si aucune des parties ne se manifeste.

Il est donc fondamental pour la survie de l'activité de l'assureur d'estimer au mieux les futures dépenses auxquelles il va devoir faire face.

1.1.3 Les composantes de la Prime Commerciale

La prime d'assurance, appelée également Prime Commerciale (PC) est la prime réellement versée par l'assuré : c'est le prix que doit payer ce dernier pour bénéficier d'une couverture d'un risque en cas de sinistre.

Cette prime admet cinq composantes et se décompose comme suit :

$$PC = PP + CH + CS + T + M$$

Avec

La Prime Pure (PP) :

La Prime Pure correspond à la tarification du risque, c'est-à-dire, au coût probable des sinistres sur la période de couverture (l'espérance des sinistres). L'actuaire procède à une analyse aussi fine que possible afin de la déterminer. Elle est à usage interne.

Les Chargements pour frais (CH) :

Ils permettent de financer les différents coûts supportés par l'assureur qui ne sont pas directement liés à la sinistralité. Ces frais sont liés aux coûts d'acquisition des contrats (commissionnement du réseau de distribution des contrats) à la gestion et à l'administration des contrats ainsi qu'aux divers frais de fonctionnement de la compagnie d'assurance (loyers, matériel, informatique).

Les Chargements de Sécurité (CS) :

La volatilité naturelle des sinistres oblige l'assureur à prendre des précautions. Ces chargements de sécurité permettent à l'assureur de se protéger contre une faillite éventuelle en cas de sur-sinistralité imprévue.

Les Taxes (T) :

Les produits d'assurance sont assujettis à une taxe fiscale plus ou moins importante selon les garanties. Celle-ci est règlementée par le Code des Assurances[1]. Les montants collectés sont alors versés par les compagnies d'assurance au Trésor Public.
En plus de la taxe fiscale, la prime d'assurance comprend également une contribution pour financer des fonds ou organismes de solidarité nationale tels que la Sécurité Sociale.

La Marge (M) :

Elle comprend la rémunération des fonds propres demandée par les actionnaires et la marge bénéficiaire de l'assureur sur le produit.

[1] Sauf pour les mutuelles, régies par le Code de la Mutualité.

1.1.4 Lien entre Prime Pure et Prime Commerciale

Il faut bien garder à l'esprit que la prime commerciale, réellement payée par l'assuré, peut totalement différer de la Prime Pure. En principe, cette dernière constitue la base de la Prime Commerciale mais certains facteurs tendent à les éloigner.

C'est le cas notamment de **la règlementation**, qui contraint l'assureur à l'utilisation de certaines variables tarifaires. En effet, pour établir la prime pure de chaque assuré, l'assureur va devoir segmenter son portefeuille, c'est-à-dire le découper en classes d'individus présentant un risque similaire, à l'aide de critères de segmentation renseignés par l'assuré lors de la souscription du contrat. Cependant, certains critères de segmentation peuvent être utilisés pour la tarification technique mais ne sont pas autorisés pour la tarification commerciale (segmentation par genre par exemple, encadrée par la *Gender Directive*).

La stratégie commerciale est également un facteur d'éloignement de la prime pure et commerciale car souvent, pour des raisons de positionnement marketing, l'assureur fait le choix d'appliquer un tarif commercial unique même si le risque diffère selon les individus.

Enfin **la concurrence** est un levier important sur le niveau de la prime commerciale puisque la compagnie d'assurance doit prendre en considération ce que propose le marché afin de ne pas compromettre la distribution de son produit.

Ces éléments amènent les assureurs à revoir leur tarification en apportant des correctifs, souvent individualisés, à la Prime Commerciale.

1.2 La théorie Fréquence - Coût de sinistres

Les différentes composantes de la prime commerciale peuvent être réparties en deux groupes :

- celles qui ne dépendent pas de la réalisation du risque assuré (ex : niveau de Taxe, car imposé par la règlementation)
- celles qui traduisent le risque porté par l'assureur (ex : la Prime Pure)

L'enjeu principal pour l'assureur est d'estimer l'espérance et la variance des sinistres, autrement dit, combien il prévoit de payer et avec quel risque. En d'autres termes, il cherche à déterminer la **Prime Pure**.

1.2.1 Un problème de minimisation

Rappelons que l'une des particularités de l'activité d'assurance, est qu'elle présente une inversion de son cycle de production. L'assureur va alors déterminer au préalable (tarification à priori) une prime P minimisant l'écart par rapport à la charge de sinistre S, pour une police donnée sur une période d'assurance donnée. Cette prime P est appelée Prime Pure de l'assuré.

Sous l'hypothèse que cette distance est mesurée par l'écart quadratique moyen[1], l'assureur doit résoudre le problème de minimisation suivant :

$$\min_{P \text{ sc } P>0} \; \mathrm{d}\left(\mathrm{P},\, \mathrm{S}\right) := \mathrm{E}[(\mathrm{S}-\mathrm{P})^2]$$

[1] Ce choix permet de pénaliser symétriquement {P < S} et {P > S}.

Or

$$\begin{aligned} \mathrm{E}[(\mathrm{S}-\mathrm{P})^2] &= \mathrm{E}\left[\mathrm{S}^2 - 2\times \mathrm{PS} + \mathrm{P}^2\right] \\ &= \mathrm{E}[\mathrm{S}^2] - \mathrm{E}[\mathrm{S}]^2 + \mathrm{E}[\mathrm{S}]^2 - 2\mathrm{P}\times \mathrm{E}[\mathrm{S}] + \mathrm{P}^2 \\ &= \mathrm{Var}(\mathrm{S}) + (\mathrm{E}[\mathrm{S}] - \mathrm{P})^2 \end{aligned}$$

La solution à ce problème est donnée par :

$$\boxed{P = E[S]}$$

1.2.2 Modélisation

La Prime Pure recherchée correspond donc à l'espérance de la charge de sinistre : il nous faut à présent modéliser cette dernière.
La sinistralité d'un ensemble de contrats peut être modélisée de plusieurs façons, selon les problèmes à traiter. Deux d'entre elles se distinguent par leur caractère opérationnel : il s'agit du modèle individuel et du modèle collectif.
L'objet de ce paragraphe est de présenter ces deux modèles, en mettant un accent particulier sur le modèle collectif, ce dernier étant le modèle de base de l'assurance automobile.

Modèle individuel

Soit S_i le montant cumulé, éventuellement nul, des sinistres payés à l'$i^{ème}$ assuré, pour la période d'assurance. La charge totale ($S^{individuel}$) pour un portefeuille de n assurés est donc :

$$\mathrm{S}^{\mathrm{individuel}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{S}_{\mathrm{i}}$$

L'appellation « modèle individuel » est donc justifiée par le fait que la charge de sinistre soit calculée pour **chaque** individu. Ces dernières sont ensuite sommées pour obtenir la charge de sinistre totale du portefeuille.

Ici, les charges individuelles de sinistre S_i sont des variables indépendantes, mais pas forcément de même loi. De plus, $\boldsymbol{P(S_i = 0)}$, où $\{S_i = 0\}$ correspond à un individu qui n'aurait pas eu de sinistres durant la période d'observation, est strictement positive.
En particulier, les S_i n'ont pas de densité.

Modèle collectif

Dans ce modèle, soit N le nombre aléatoire de sinistres au cours de la période d'assurance. Soit C_i le coût du $i^{ème}$ sinistre, sans tenir compte de l'assuré concerné.

La charge totale de sinistre ($S^{collectif}$) est donc :

$$\mathrm{S}^{\mathrm{collectif}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{C}_{\mathrm{i}}$$

Cette décomposition permet de mieux cerner l'influence des différents facteurs susceptibles d'intervenir sur la loi de S : **la décomposition Fréquence – Coût** apparait.

La mise en place de ce modèle Fréquence-Coût nécessite la vérification des hypothèses suivantes :

- $N \in \mathbb{N}$, le nombre de sinistres ou fréquence de sinistres observés sur la période de couverture
- les variables aléatoires C_i sont indépendantes et de même loi,
- $\forall i, (C_i)_{i\geq 1}$ est indépendant de N (indépendance Coût-Fréquence).

Enonçons alors quelques propriétés[1] élémentaires découlant du modèle.

Propriété 1 :

Si N et C admettent des moments d'ordre un, il en résulte :

$$\mathbf{E[S] = E[N] \times E[C]}$$

Cette formule traduit une idée simple : **la Prime Pure est égale à la fréquence moyenne par coût moyen**.

Propriété 2 :

Si C et N admettent des moments d'ordre 2, il en résulte aussi :

$$\mathbf{Var[S] = E[N] \times Var[C] + Var[N] \times E[C]^2}$$

Remarque :

La variance de la charge de sinistre est souvent utilisée pour déterminer le montant des chargements de sécurité qui sont liés à la volatilité des sinistres.

1.2.3 Conditions d'application du modèle collectif

Certaines hypothèses du modèle collectif sont contraignantes et ne sont pas vérifiées sur le portefeuille d'assurés automobiles considéré dans son ensemble. Nous allons dans un premier temps les expliciter puis introduire le travail préalable à effectuer par l'actuaire afin de pouvoir utiliser le modèle collectif et ses propriétés dans sa quête de détermination de l'espérance de la charge de sinistres.

Portefeuille de grande taille

La méthode de tarification choisie dépend de **la nature des sinistres et de la taille du portefeuille**. En effet, la tarification d'une prime d'assurance pour couvrir les dommages aux matériels transportés dans une fusée est différente de celle d'une simple assurance auto.

En automobile, du fait du caractère obligatoire de la garantie Responsabilité Civile, nous sommes dans le cas d'un portefeuille de grande taille. C'est cette configuration qui nous permet d'utiliser l'espérance comme estimateur de la charge moyenne de sinistre.

[1] Les preuves de ces deux propriétés figurent en Annexe 1.

Notre méthode de tarification s'appuie donc sur un théorème présentant des résultats asymptotiques (dû au grand nombre d'assurés) : la Loi des grands nombres.

Enoncé de la Loi des grands nombres (LGN) :

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles (var) indépendantes et **identiquement distribuées** (i.i.d) définies sur le même espace de probabilité. Soit $\mu = E[X_i] < \infty$.
Alors la variable $X^{(n)} = \frac{\sum_{i=1}^{n} X_i}{n}$ vérifie :

$$\mathrm{X}^{(\mathrm{n})} \xrightarrow{\mathrm{ps}} \mu \quad \text{quand } \mathrm{n} \to \infty$$

Application de la LGN au portefeuille

Notons $S^{(n)}$ la charge moyenne de sinistre dans un portefeuille de taille n.
Dans le cas où le groupe des assurés est **homogène** et que leur **nombre** est **suffisant**, la loi des grands nombres s'applique et de ce fait :

$$\lim_{\mathrm{n}\to\infty} \mathrm{S}^{(\mathrm{n})} = \mathrm{E}[\mathrm{S}]$$

Où $S^{(n)}$ est la charge de sinistre moyenne pour une police donnée.

En pratique la loi des grands nombres ne s'applique pas directement au portefeuille des assurés. En effet, il faut considérer une remise en cause de :

- la **mutualisation** : le résultat obtenu est un résultat asymptotique tandis que le nombre d'assurés est fini.
- l'**homogénéité** : les assurés ne sont pas identiques.

Ainsi, il est nécessaire de segmenter le portefeuille et d'utiliser des espérances conditionnelles. De ce fait, l'assureur ne calcule pas $E[S]$ pour déterminer la prime pure de chaque assuré, mais $E[S|X]$ pour déterminer la prime pure d'un assuré appartenant au groupe X, où X est un **groupe homogène de risques**.
Il veille également à constituer des groupes présentant un **effectif suffisant** afin d'utiliser le caractère asymptotique du modèle.

Les relations précédentes sur l'espérance et la variance de sinistres en fonction de la fréquence et du coût deviennent alors :

$$\mathbf{E[S|X] = E[N|X] \times E[C|X]}$$

$$\mathbf{Var[S|X] = E[N|X] \times Var[C|X] + Var[N|X] \times E[C|X]^2}$$

1.3 Segmentation tarifaire, mutualisation et asymétrie d'information

Pour s'inscrire dans le cadre de la Loi des grands nombres, l'assureur doit :

- Disposer d'un portefeuille d'assurés de **taille suffisante**
- **Créer des groupes homogènes** au sein de son portefeuille d'assurés de façon à respecter le caractère identiquement distribué des variables représentant le montant des sinistres pour chaque individu appartenant au même groupe.

Dans le but de satisfaire ce deuxième point, il est donc nécessaire de déterminer, lors de l'établissement d'un tarif sur une population d'individus, le bon équilibre entre :

- La **segmentation** en classes de risques
- La **mutualisation** entre les individus d'un même groupe homogène de risques

Ces deux notions sont étroitement liées à la notion d'asymétrie d'information. Elles sont toutes trois présentées dans les paragraphes qui suivent.

1.3.1 La segmentation tarifaire

Le principe de la segmentation repose sur le fait que chaque individu doit payer une prime qui est fonction de son propre risque. Un tarif individuel ne pouvant être établi, l'assureur partitionne son portefeuille en classes tarifaires, regroupant les individus présentant un profil de risque proche. Au sein de chacune d'elles (cf. Figure 2), les assurés paient pour un risque moyen des individus qui la compose.
Le principe sous-jacent est qu'il ne serait pas raisonnable que certains assurés dits « prudents », aient à payer pour les risques pris par d'autres individus qui le seraient moins. Il est donc juste et normal que les tarifs d'assurance varient selon un certain nombre de critères, reflétant la responsabilité prise individuellement par l'assuré. Ces critères doivent être raisonnables et fondés sur les différences objectives et pertinentes.

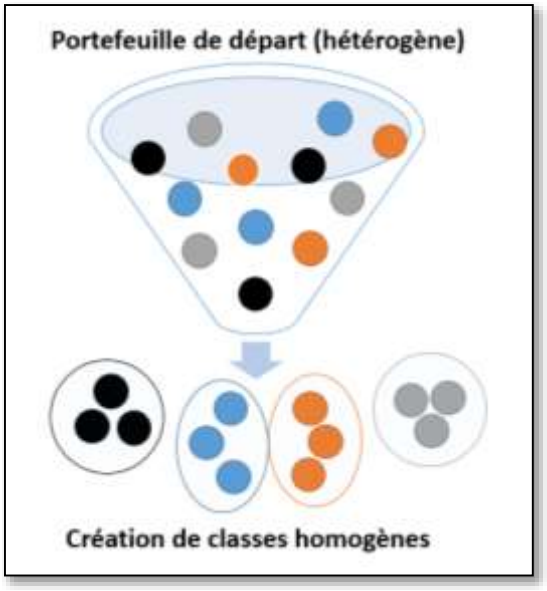


Figure 2- Représentation des effets et objectifs de la segmentation

La nécessité de segmenter

Prenons un exemple simple afin de comprendre la nécessité de segmenter.

Soit un duopole avec les caractéristiques suivantes :

- Un assureur A, qui décide de n'effectuer aucune segmentation
- Un assureur B, qui lui segmente en fonction de la « qualité » du conducteur

Chaque assureur propose un contrat C proposant les mêmes garanties.

Le Tableau 1 ci-dessous présente les différentes primes proposées par chacun des assureurs :

	Mauvais conducteurs	Bons conducteurs
Assureur A	700 €	700 €
Assureur B	1 500 €	500 €

Tableau 1 - Primes proposées par les assureurs A et B pour deux profils de conducteurs

L'assureur A propose un tarif unique pour l'ensemble de ses assurés. Dans ce contexte, la prime perçue pour les « bons conducteurs » sert, en partie, à couvrir les sinistres des « mauvais conducteurs ».
L'assureur B propose quant à lui deux tarifs différents en fonction du profil de risque. Ainsi, les « mauvais conducteurs » payent une prime supérieure du fait de leur exposition au risque. Les « bons conducteurs » payent, quant à eux, une prime plus faible, en accord avec leur profil de risque.

La mise en concurrence de ces deux assureurs entraîne des migrations d'assurés d'un portefeuille vers l'autre. En effet, les assurés s'orienteront naturellement vers l'assureur leur proposant le meilleur tarif. Ainsi, les « mauvais conducteurs » souscriront leur contrat chez l'assureur A et les « bons conducteurs » chez l'assureur B. L'assureur A ne disposera donc plus de « bons conducteurs » pour couvrir les mauvais et son modèle économique ne sera plus viable.

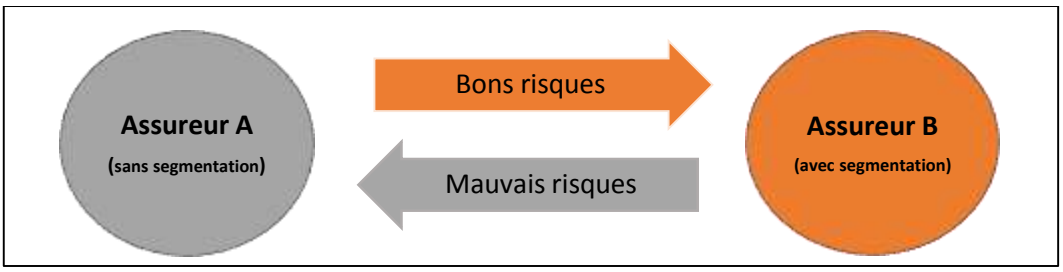


Figure 3 - Flux d'assurés provoqués par l'utilisation (ou non) d'une segmentation

Deux solutions se présentent alors pour cet assureur :

- réduire le tarif unique à 500 € pour capter à nouveau les bons conducteurs ;
- segmenter le tarif pour proposer un tarif différent en fonction de la « qualité » du conducteur.

Une réduction du tarif ne permettra plus à l'assureur A de collecter suffisamment de primes pour couvrir l'ensemble des sinistres. Ainsi, la seule solution pour cet assureur est de segmenter son tarif.

Limites de la segmentation

L'exemple précédent, certes très simpliste, exprime bien la nécessité pour l'assureur de segmenter. Néanmoins, la segmentation présente des limites et ne peut être poussée à l'extrême.

En effet, lorsqu'elle est trop fine, c'est-à-dire lorsque l'on tend à avoir un tarif individualisé, l'estimation de la Prime Pure est basée sur une classe tarifaire d'effectif trop faible pour assurer la validité du théorème de la loi des grands nombres. L'estimation de la charge moyenne de sinistre est alors biaisée conduisant à de mauvais résultats pour l'assureur.

Il convient donc de rester prudent et de **ne pas segmenter à l'extrême**.

Ceci nous conduit à parler de l'importance du **choix des critères de segmentation**. L'assureur choisit les critères les plus influents sur le niveau de sinistralité, en distinguant les critères influant respectivement sur la fréquence et le coût.
Parfois ce choix peut s'avérer contraint, que ce soit par la règlementation ou par un manque d'information (information indisponible due au coût trop important de son obtention, information biaisée, etc.).
Puisque la segmentation n'est pas parfaite, les individus regroupés dans une même classe tarifaire présentent tout de même une certaine hétérogénéité : c'est au sein de chacune des classes que s'applique le principe de mutualisation.

Les variables tarifaires fréquemment utilisées en assurance automobile

Les sociétés d'assurance disposent, en matière de tarification automobile, d'une grande liberté quant au choix des variables à renseigner par l'assuré ou à utiliser pour la tarification, exception faite lorsque l'usage de la variable est interdit suite à une règlementation.

Chaque assureur retient pour la tarification de son portefeuille les variables les plus pertinentes et les plus adaptées à son mode de gestion et à ses objectifs. Il en résulte une certaine diversité des méthodes de tarification utilisées sur le marché. Cependant, un certain nombre de variables reviennent dans l'établissement de la plupart des tarifs proposés par les sociétés d'assurance.

C'est particulièrement à ces facteurs « universels » que nous allons nous intéresser ici.

Lors de la signature d'un contrat d'assurance automobile, l'assureur dispose de deux types de variables (appelées critères de tarification) :

- Les **variables exogènes**, qui apportent des informations relatives au risque (âge ou genre de l'assuré, activité professionnelle, catégorie de véhicule, zone géographique du risque, etc.)
- Les **variables endogènes**, qui apportent des informations sur les réalisations individuelles passées du risque.

Pour chaque assuré, ces deux types de variables peuvent eux même être répartis en quatre groupes de variables tarifaires, selon l'objet concerné par l'information fournie :

- Les **caractéristiques du conducteur** (variables exogènes)
- Les **caractéristiques du véhicule** (variables exogènes)
- Le **type de contrat** (variables exogènes)
- Le **passé d'assurance** / la sinistralité (variables endogènes)

Nous verrons en dernière partie que d'autres variables, telles que la fréquence d'utilisation de la voiture, le comportement de l'assuré ou encore l'entretien/l'état général de la voiture, peuvent également être prises en compte.

Les variables tarifaires fréquemment utilisées en assurance automobile

Le type de garanties souscrites
Responsabilité Civile (RC, assurance minimale obligatoire) Attentats, Catastrophes naturelles, Bris De Glace, Vol, Dommages corporels du conducteur et Assistance, Dommages au Véhicule

La franchise (en cas d'accident ou de bris de glace)

La fréquence de payement de la prime
Le fractionnement de la prime peut traduire une position fragilisée, donc un véhicule moins entretenu.

Le genre
Il s'agit du genre du conducteur principal déclaré.

Le type du conducteur
Conduite exclusive / Conducteur secondaire.
Le tarif est généralement plus attractif lorsque vous êtes le seul conducteur du véhicule.

L'âge du conducteur, exprimé en années.

La situation de famille
Un célibataire paiera généralement plus cher qu'une personne mariée, considérée comme plus prudent.

La profession du conducteur.

La zone géographique, ou de façon équivalente le lieu de résidence (rural / urbain).

Le lieu de garage (garage/ rue)

L'ancienneté du permis
Majoration pour les conducteurs novices (moins de 3 ans de permis)

Le coefficient Réduction Majoration (CRM) ou Bonus-Malus (voir § 1.3.3)

La période de couverture (l'exposition)
Période, au cours de laquelle l'assuré est couvert par la police qu'il a souscrit, le plus souvent cette période est d'une année.

Responsabilité du conducteur : variable binaire qui indique si la responsabilité du conducteur est engagée en cas de sinistre.

Le nombre d'années d'assurance

Les infractions, suspensions de permis

L'ancienneté de véhicule
Elle sert à évaluer le coût des pièces en cas de dommage matériel.

Le carburant
L'acquisition d'un véhicule diesel se justifie souvent par un usage plus fréquent et pour de plus longues distances puisque le prix d'achat est certes plus élevé, mais le prix du litre de carburant est moins élevé.

La puissance réelle du véhicule / le poids du véhicule
L'idée est que plus un véhicule est puissant, plus il est difficile à manier pour l'automobiliste. Les véhicules puissants sont à l'origine d'un plus grand nombre d'accidents).

La valeur du véhicule : son prix. Pour les garanties dommages à la voiture, la valeur à neuf de celle-ci (les réparations d'une voiture chère sont plus onéreuses).

La marque du véhicule
Une Renault Clio se vole beaucoup plus souvent qu'une Dacia Logan, par exemple.

L'âge du véhicule

La date de 1e mise en circulation ou la date d'acquisition.

L'usage du véhicule

Le kilométrage annuel

Le mode d'acquisition du véhicule
Prêt / comptant, occasion / neuf

Les Caractéristiques du conducteur - Les Caractéristiques du véhicule
Les caractéristiques du contrat - Les Antécédents d'assurance et sinistres

1.3.2 Le principe de la mutualisation

L'assureur offre un ensemble de garanties pour lesquelles son portefeuille est constitué d'un nombre important de contrats. Ceux-ci sont segmentés par l'assureur en classes de risques « homogènes » et indépendants. Chaque classe constitue une « mutualité » de risques.

Le principe de la mutualisation consiste à répartir le coût de la réalisation du risque couvert entre les membres d'un groupe soumis potentiellement au même risque et qui pourrait frapper certains d'entre eux. S'opère alors une **compensation** entre les risques sinistrés et ceux pour lesquels l'assureur reçoit une prime sans avoir à régler de prestation. La masse des primes collectées permet donc de verser une indemnité aux sinistrés.

Pour l'assureur, la **mutualisation des risques** entraîne deux conséquences :

- **la sélection des risques** de manière à rechercher **un équilibre** entre risques sinistrés et risques non sinistrés ;
- **le calcul du montant des primes** à verser par les assurés, déterminé par un calcul de probabilités.

Remarque :
Il semble important de souligner la distinction entre les deux notions suivantes : solidarité entre assurés et mutualisation.
En effet, dans le cas d'un principe de solidarité, tous les assurés paieraient une prime identique à l'assureur quel que soit leur profil de risque. Là n'est pas le fondement de l'assurance.

1.3.3 Asymétrie d'information lors de la détermination de la Prime Pure

Dans de nombreux cas, l'opération d'assurance se noue dans un contexte d'asymétrie d'information entre le preneur d'assurance et l'assureur : l'assureur ne dispose pas de la totalité des informations nécessaires à son évaluation du risque. Il doit donc recueillir un maximum d'informations aussi bien sur le risque assuré que l'assuré lui-même afin d'évaluer et de tarifer au mieux le risque à assurer.

Pour atteindre cet objectif, il dispose de moyens divers :

- un questionnaire plus ou moins élaboré auquel doit répondre le client lors de la souscription ;
- une prise en compte de statistiques d'antécédent de sinistralité au sein de son portefeuille sur une période antérieure ou sur le portefeuille d'un autre assureur ;
- une prise en compte de statistiques ou d'études de comportement à l'échelle nationale (étude INSEE).

Malgré tout, la qualité de l'information disponible par l'assureur reste partielle, et ce pour plusieurs raisons :

- Tout d'abord, et ce qui est le plus simple à appréhender, **certains renseignements concernant l'assuré ne peuvent être recueillis** pour des raisons éthiques, législatives ou règlementaires, voire constitutionnelles, par exemple, la religion ou la race ;

- De plus, **leur lien avec la quantification du risque doit être connu** et présenter un degré certain de permanence dans le temps. Ceci explique que dans la plupart des cas, des renseignements de type purement comportemental ne sont pas demandés : habitudes familiales, comportement au volant, etc. ;
- Enfin, il ne faut pas oublier que **l'acquisition d'informations et la gestion d'informations complexes ont un coût**. Ces coûts limitent le volume du système d'information.

Les données une fois recueillies, l'assureur effectue des analyses statistiques et actuarielles sur ces dernières. Il ne possède en réalité qu'une connaissance de nature statistique du risque.

Les informations dont il dispose, incomplètes de par l'existence d'une asymétrie d'information, ont pour conséquence la constitution de classes tarifaires présentant une hétérogénéité résiduelle.

Du point de vue de l'assuré, il arrive que ce dernier ne communique pas l'ensemble de l'information le concernant. Que cela se fasse volontairement ou non, de tels comportements existent et doivent être pris en compte par l'assureur.

Après avoir choisi un système de tarification, l'assureur se place dans une situation d'asymétrie d'information qui est d'autant plus accentuée que le marché de l'assurance est transparent et ouvert à la concurrence entre les sociétés d'assurance.

Cette situation l'expose à deux grands types de risque : le risque d'anti sélection et le risque moral.

Le **risque d'anti sélection** résulte de l'hétérogénéité résiduelle des catégories tarifaires.
Tous les assurés appartenant à une même mutualité vont devoir s'acquitter de la même prime mais il peut arriver que certains assurés aient conscience qu'ils présentent un risque inférieur au risque de cette mutualité.
Ils vont alors « profiter » du caractère concurrentiel du marché et chercher un autre assureur qui aurait des critères de segmentation différents pour obtenir un meilleur tarif.
En quittant leur premier assureur, ils déséquilibrent la mutualité à laquelle ils appartenaient : l'ancien assureur doit donc revoir son modèle de tarification pour reconstituer des classes homogènes à chaque vague d'entrées et sorties dans son portefeuille.

Quant au **risque moral** (ou *moral hazard* en anglais), définit par Adam Smith comme « la maximisation de l'intérêt individuel sans prise en compte des conséquences défavorables de la décision sur l'utilité collective », il provient du fait que la souscription du contrat peut modifier à postériori le comportement de l'assuré et donc rendre erronée la tarification du risque.
Concrètement, il s'agit du fait, pour un individu, d'adopter un comportement plus risqué parce qu'il se sait bien protégé.
En assurance automobile, ce phénomène peut s'illustrer par le fait qu'un assuré « tout risque » sera moins vigilant sur la route que s'il ne disposait que d'une assurance partielle.

L'aléa moral peut apparaitre avant et après la réalisation du sinistre.
Il convient alors de distinguer :

- L'aléa moral « ***expost*** » : le dommage étant déjà réalisé, l'assuré optimise son contrat en conséquence ;
- L'aléa moral « ***exante*** » : l'assuré adapte son comportement à son niveau de couverture.

Réduction de l'asymétrie d'information

L'asymétrie d'information ne pouvant être évitée par l'assureur, des mécanismes de surveillance ou d'incitations ont été mis en place pour réduire ses effets.
En pratique il s'agit :

- D'amener le souscripteur à révéler une partie de l'information inobservable par l'assureur en lui proposant différent contrats. Ainsi, le choix d'un assuré de souscrire un contrat avec **franchise**, alors que la société proposait également un contrat ne comportant pas de franchise peut être révélateur d'un comportement moins risqué.

- D'introduire dans la tarification la prise en compte des historiques de sinistralité qui sont souvent révélateurs du comportement et permettent une personnalisation de la prime : c'est le système « Bonus-Malus » en automobile. Lors de cette la tarification « **a postériori** », la prime d'assurance est multipliée par un coefficient de Bonus-Malus, également appelé Coefficient de Réduction-Majoration (CRM). Le mode de fonctionnement de ce coefficient est fixé par le Code des Assurances[1].

[1] Il est défini par l'article A 121-1 du code des assurances.

Chapitre 2 - *Gender Directive* et Directive générale Anti-discrimination

Nous avons vu que la segmentation du portefeuille d'assurés était importante.
Néanmoins, pour rester dans le cadre d'application de la Loi des Grands Nombres, il faut s'assurer que chaque groupe homogène de risque comporte un nombre d'individus suffisant, sans quoi, les observations ne seraient plus significatives.
Il faut donc trouver le point d'équilibre entre une segmentation trop fine, qui empêcherait la mutualisation, et une segmentation insuffisante qui augmenterait le risque porté par l'assureur.

Aujourd'hui, ces deux phénomènes sont illustrés sur le marché avec :

- D'un côté une **sur-segmentation** due à la concurrence et au souhait de l'assureur de proposer un tarif le plus proche du profil de risque de l'assuré pour l'inciter à souscrire. Certains assureurs évoquent l'idée d'un tarif individuel, semblant remettre en cause l'idée même de mutualisation des risques.
- De l'autre, une législation qui tend à **supprimer certaines variables tarifaires,** jugées discriminantes, des modèles de tarification, réduisant ainsi la qualité de la segmentation des assureurs.

L'objet de ce mémoire s'attache tout particulièrement à ce deuxième point : nous allons mettre en place des modèles alternatifs à la suppression d'une variable tarifaire en automobile. Pour ce faire, nous nous appuyons sur un exemple concret, puisqu'actuellement déjà en application, la ***Gender Directive*.** Cette dernière s'inscrit dans le cadre d'une Directive Anti-discrimination qui englobe bien plus que le critère genre.

2.1 La *Gender Directive* : contexte

Dès lors que l'assuré n'a pas de maîtrise sur la variable de différenciation tarifaire, l'Union européenne considère qu'il y a un risque de discrimination, incompatible avec le principe de non-discrimination proclamé par la Charte des droits fondamentaux de l'Union. L'âge et le genre sont deux variables particulièrement concernées puisqu'elles ne résultent pas d'un choix de l'assuré.

Le 1er mars 2011, la Cour de justice de l'Union européenne (CJUE) a décidé de mettre un terme aux différences de tarifs fondées sur le genre de l'assuré. Cet arrêté a pris effet le 21 décembre 2012.
Plus question donc pour les nouveaux contrats de différencier les cotisations ou les prestations en fonction du genre de l'adhérent, même si les données actuarielles apportent la démonstration que la femme n'est pas l'égale de l'homme devant le risque ou inversement.

En conséquence, certains assureurs doivent revoir leurs modèles et leurs méthodes de tarification.
Pour apprécier les changements ainsi engendrés, les origines et les conditions d'application de la *Gender Directive* seront développées.

2.1.1 Les étapes clés de la *Gender Directive*

L'adoption de la *Gender Directive* n'a pas été simple : il a fallu passer par plusieurs phases durant lesquelles les assureurs auront tenté de justifier tant bien que mal la nécessité de l'utilisation du genre de l'assuré dans la tarification, sans succès.

- 13 décembre 2004 : **Directive 2004/113/EC[1] du Conseil de l'Union européenne instaurant le principe général de l'égalité de traitement entre homme et femme** (appelée également Directive genre) dans l'accès et la fourniture de biens et de services.

Cette directive interdisait que la prise en compte du genre de l'assuré ne conduise à une différence en matière de prime ou de prestation dans les contrats d'assurance conclus après le 21 décembre 2007 ;

« Article 5-1: les États membres doivent s'assurer que, pour les contrats souscrits après le 21 décembre 2007, l'utilisation du sexe comme facteur de calcul des primes et des prestations en assurance ne doit pas déboucher sur des différences dans celles-ci. »

Néanmoins, la directive autorise à titre dérogatoire les Etats à permettre l'utilisation du genre en tant que facteur actuariel, quand les données actuarielles et statistiques le justifiaient.

« Article 5-2: les États ont la possibilité de déroger au principe d'égalité quand l'utilisation du facteur du sexe est déterminante pour l'analyse du risque au plan statistique et actuariel. »

La France décide alors d'utiliser cette autorisation, notamment en **assurance automobile[2]** où, parmi les sinistres graves, les jeunes hommes sont surreprésentés[3] par rapport aux jeunes femmes.
D'autres pays européens (Belgique, Irlande, Royaume-Uni...) ont également adopté ces dérogations.

- Juillet 2008 : **Recours en annulation de la dérogation accordée par l'article 5.2 par l'association de consommateurs belge « Test-Achats».**

L'association de consommateurs belges Test-Achats[4] considérant cette dérogation comme anticonstitutionnelle, demande l'abrogation de cette dérogation à la justice belge qui renvoie l'affaire à la Cour de Justice Européenne.

- 1er mars 2011 : **Arrêt « Test Achats » de la Cour de Justice de l'Union européenne (CJUE)**

La Cour de Justice Européenne prononce l'arrêt de cette dérogation à partir du 21 décembre 2012. Cette décision est motivée par le fait que la dérogation présente dans la directive 2004/113/CE n'avait aucune limite dans le temps.
Le risque était qu'elle soit «indéfiniment permise par le droit de l'Union. Dès lors, une telle disposition est contraire à la réalisation de l'objectif d'égalité de traitement entre les femmes et les hommes, et doit être considérée comme *invalide à l'expiration d'une période de transition adéquate ».*

- Le 21 décembre 2012 **marque donc la fin de la dérogation au principe d'égalité homme/femme.**

A compter de cette date, les primes d'assurance et les prestations des contrats d'assurance devront être aménagées afin qu'il ne soit aucunement tenu compte du genre de l'assuré.
La décision de la Cour s'applique à tous les contrats d'assurance commercialisés par les assureurs de l'UE auprès des particuliers résidant dans l'UE.

[1] Source : [R8]
[2] Article A. 111- 4 du Code des assurances
[3] Le taux de tués parmi les personnes de 18 à 24 ans était cinq fois plus élevé chez les hommes que chez les femmes (rapport ONISR 2010).
[4] Association dont l'action est basée sur les tests comparatifs de produits dans le but de promouvoir la défense des intérêts des consommateurs.

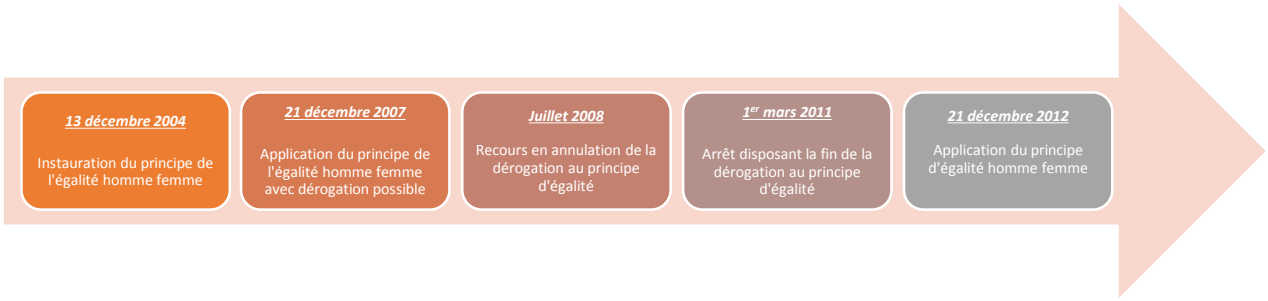


Figure 4 - Représentation des différentes étapes de mise en place de la *Gender Directive*

2.1.2 Portée d'application de la *Gender Directive*

Contrats concernés et portée temporelle

Cet arrêt touche aussi bien les offres individuelles des Sociétés d'Assurance, les Mutuelles que les Institutions de Prévoyance. De plus, il concerne en particulier :

- tous les **contrats d'assurance privés, volontaires et non liés au travail[1].**
- tous les accords contractuels nécessitant l'accord de l'ensemble des parties (nouveaux contrats, avenants[2], etc.)

Il faut également que le dernier consentement de l'une des parties intervienne après le 21 décembre 2012 : les contrats bénéficiant d'un consentement antérieur ne sont pas concernés, ce qui implique que la tarification des contrats bénéficiant de clauses de tacite reconduction reste inchangée.
Les contrats d'assurance automobile font donc partie intégrante de l'arrêt car ce sont des contrats d'assurance privé, non liés au travail. Ils présentent également une clause de tacite reconduction à laquelle il faudra porter attention.

Lignes directrices : les pratiques qui restent autorisées

Il n'est certes plus règlementaire de proposer un tarif différencié mais certaines pratiques restent encore permises aux organismes d'assurance, à savoir, le recueil, le stockage et l'utilisation de la variable genre pour :

- l'évaluation des risques (Provisions Techniques / calcul des Provisions réglementaires) et la tarification interne, consistant au calcul de la Prime Pure ;
- la tarification de la réassurance (la directive et l'arrêt s'appliquent uniquement aux primes et aux prestations) ;
- le Marketing et la publicité, qui offrent la possibilité d'influencer la composition du portefeuille.

[1] Article 5-1 de la directive 2004/113/CE – [R8]
[2] La prorogation d'un contrat et les changements de situation de risque de l'assuré nécessitant d'avertir l'assureur sont considérés comme des avenants.

2.2 La suppression de variables tarifaires, un sujet d'actualité ?

Le genre de l'individu est d'ores et déjà une variable tarifaire règlementée. Mais cette règlementation ne cesse d'évoluer et un projet de loi, la directive Anti-discrimination a déjà été proposé et est actuellement en cours d'étude. En effet le genre n'est pas la seule variable de différenciation tarifaire sur laquelle l'assuré n'a pas de maîtrise, l'âge est aussi concerné.

L'Union européenne considère qu'il y a là encore un risque de discrimination, ce qui est contraire à la directive sur l'égalité de traitement des personnes. L'âge pourrait donc être le prochain critère de différenciation tarifaire amené à disparaître.

Actuellement l'arrêt prévoit une proposition de directive relative à l'égalité de traitement (couvre d'autres motifs, tels que l'âge et le handicap dans l'accès aux biens et services).
À ce jour, le projet de Directive ne s'est pas concrétisé. Cependant, les travaux préparatoires laissent à penser que les assureurs pourraient bénéficier d'une dérogation, comme lors de l'entrée en vigueur de la *Gender Directive*.
Lors de la précédente décision sur le genre, malgré la volonté initiale du législateur de reconnaitre la spécificité du métier d'assureur, le critère a finalement été interdit. Les assureurs peuvent donc redouter qu'un tel retournement de situation ne se reproduise avec le critère de l'âge, d'autant plus que ce critère est tout autant utilisé par les assureurs que ne l'était le genre dans la tarification.

Catégorie de produit	Produit d'assurance	Facteurs		
		Genre	âge	Handicap
Assurance automobile	Assurance automobile - Assurance pour les voitures privées couvrant au minimum la Responsabilité Civile	++	++	+

Tableau 2 - Utilisation des facteurs d'évaluation des risques, en assurance automobile (sur la base de la fréquence d'utilisation signalée par les parties prenantes)[1]

++ : *Facteur signalé comme étant fréquemment utilisé (par 50 % ou plus de l'ensemble des associations professionnelles, des associations d'actuaires, des autorités compétentes et des organismes chargés des questions d'égalité interrogés)*

+ : *Facteur signalé comme utilisé occasionnellement (par 10 % à 50 % de l'ensemble des associations professionnelles, des associations d'actuaires, des autorités compétentes et des organismes chargés des questions d'égalité interrogés.*

Remarque :
Si l'âge est perçu comme un critère discriminant durant l'étape de la tarification, cela va encore plus loin : il l'est même lors de la souscription.
En effet, le bureau central de tarification (BCT) qui enregistre les plaintes dues à un refus d'assurance de responsabilité civile (RC-obligatoire) a recensé 33 % de refus ayant pour motif l'âge de l'assuré sur 712 décisions en 2013 (Source : AGIRA).

[1] Source : Annexe 2 du journal officiel de l'Union européenne (2012/C 11/01) - [R10]

Chapitre 3 - Principe de la régression

Dans les paragraphes qui suivent, nous présentons certains aspects d'une méthode qui joue un rôle essentiel pour la tarification en assurance Non-Vie : la régression linéaire.
En effet, c'est grâce à une généralisation de cette méthode que les assureurs déterminent la Prime Pure pour chacun des assurés.

Les modèles linéaires permettent d'établir une relation entre une variable à expliquer, notée Y, et des variables explicatives, notées X.
Ici, c'est la relation entre la Prime Pure, payée par l'assuré, et les informations dont l'assureur dispose sur l'assuré que nous souhaitons établir.

Dans un premier temps, nous présentons les fondements théoriques du Modèle Linéaire (§ 3.1), ainsi que les limites de ce dernier, notamment dans son utilisation pour la représentation de la charge de sinistre en assurance non vie.

Suite à cela, nous détaillons les étapes de la mise en place d'un modèle linéaires généralisés (GLM) (§ 3.2) qui est une extension du modèle linéaire classique. Nous évoquons notamment les différentes lois qui s'y prêtent.

Enfin, ce chapitre est illustré d'un cas particulier de GLM : le modèle de régression logistique (§ 3.3). Ce modèle est utilisé lors de la mise en place du modèle prédictif.

3.1 Présentation du modèle linéaire

Soient n individus observés pour lesquels nous disposons :

- d'une **variable réponse** $(Y_i)_{i=1,\dots,n}$, aussi appelée **variable à expliquer**
- de p variables explicatives $(x_i^1, \dots, x_i^p)_{i=1,\dots,n}$

L'objectif est d'établir une relation entre la variable à expliquer Y_i et les p variables explicatives $(x_i^1, \dots, x_i^p)$. Supposons l'existence d'une relation linéaire entre la variable à expliquer et les variables explicatives du type :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \times x_i^j + \varepsilon_i$$

Le modèle peut donc s'écrire sous la forme matricielle suivante :

$$\boxed{Y = X\beta + \varepsilon}$$

Avec :

- $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}$, un vecteur de $\mathbb{R}^n$ contenant les variables à expliquer

- $X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_i^1 & \vdots & x_i^j & \vdots & x_i^p \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{pmatrix}$, une matrice $(n, p+1)$, contenant les valeurs observées des p variables explicatives pour les n individus. On notera $x^j \in R^n$ la $j^{ème}$ colonne de X qui représente les observations de la $j^{ème}$ variable sur n individus,

- $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{pmatrix}$, un vecteur de $\mathbb{R}^{p+1}$ contenant les p+1 **paramètres de la régression**,

- $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$, un vecteur de $\mathbb{R}^n$ qui contient les erreurs du modèle, aussi appelés résidus.

Les termes d'erreur $(\varepsilon_i)_{i=1,\dots,n}$ sont supposés être des variables aléatoires indépendantes et identiquement distribuées suivant une loi Normale tels que :

$$\begin{cases} E[\,\varepsilon_i\,] = 0 \\ Var[\varepsilon_i] = \sigma^2 > 0 \end{cases}$$

L'hypothèse selon laquelle la variance des erreurs (σ^2) est constante est appelée « **hypothèse d'homoscédasticité »**.

Ainsi, la distribution de la variable réponse aléatoire Y_j est aussi Normale et il en découle :

$$E[Y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j \times x_i^j = \mu_i$$

$$Var[Y_i] = \sigma^2$$

Le but du modèle est d'estimer les paramètres $(\beta_j)_{\,0 \le j \le p}$ de la régression.

Remarque :

Le terme « **modèle linéaire »** recouvre plusieurs types de modèles et le contenu de X dépend de la forme des variables explicatives :

- on parle de **modèle de régression multiple** dans le cas où les variables sont quantitatives : x_i^j correspond alors à la valeur observée de la $j^{ème}$variable explicative chez l'individu i ;

- on parle de **modèle d'analyse de variance (ANOVA)** dans le cas où les variables sont qualitatives. Dans ce cas, les variables sont appelées « facteurs » et X est composé des variables indicatrices associées aux niveaux du (ou des) facteur(s).

Le cas de variables mixtes (quantitatives et qualitatives) peut aussi se présenter : on parle alors de **modèle d'analyse de covariance (ANCOVA)** et X prend respectivement les valeurs décrites précédemment en fonction de la nature de la variable.

Le modèle linéaire est facile à mettre en œuvre mais **n'est pas entièrement adapté à la tarification en assurance Non-Vie** : il suppose que la moyenne est une fonction linéaire des variables, ce qui n'est pas le cas en assurance automobile par exemple.

De plus, l'hypothèse selon laquelle les erreurs (ε_i) sont **distribuées** aléatoirement **de façon Normale** et l'**homoscédasticité** posent aussi problème.

En effet, le nombre de sinistres (fréquence) suit une loi de probabilité discrète positive qui prend ses valeurs dans $\mathbb{N}$: la loi Normale étant définie sur $\mathbb{R}$ risquerait de rendre le nombre de sinistres négatif. De même, les coûts des sinistres sont positifs et admettent souvent une queue de distribution plus importante vers la droite : ils ne peuvent donc pas être modélisés par une loi Normale.

Pour pallier à ces restrictions, le modèle a été élargi aux Modèles Linéaires Généralisés (GLM). Ces derniers permettent de modéliser des variables dont la distribution fait partie d'une famille de lois particulière : la **famille exponentielle**. Dans la suite de ce mémoire, nous utilisons ces modèles pour expliquer la fréquence et le coût des sinistres puisqu'ils rendent mieux compte de la réalité que ne le fait le modèle linéaire.

3.2 Les Modèles Linéaires Généralisés (GLM)

Au même titre que le modèle linéaire, le modèle linéaire généralisé consiste à étudier la liaison entre une variable dite « réponse » Y et un ensemble de variables explicatives x_1,…, x_p.
Il s'agit, dans ce cas, de faire une hypothèse sur la loi suivie par les observations de la variable réponse. Dans ce mémoire, l'étude portant sur l'assurance automobile, les lois testées appartiennent à la famille exponentielle, qui regroupe de nombreuses lois classiques (Poisson, Gamma, Normale, Binomiale, etc.).

L'idée reste d'utiliser une transformation mathématique sur la variable à expliquer mais en tenant compte cette fois-ci de la véritable distribution des erreurs.

La mise en place d'un GLM nécessite alors quatre étapes :

- Le choix d'un modèle (§ 3.2.1) ;
- L'estimation des coefficients de la régression (§ 3.2.2) ;
- La procédure de sélection des variables (§ 3.2.3) ;
- La vérification de la validité du modèle (§ 3.2.4).

Appliquer à la lettre les quatre étapes précédentes ne garantit aucunement l'obtention d'un modèle parfait, en effet, un risque de modèle (§ 3.2.5) demeure. Comme tout risque, il se doit d'être évalué et l'assureur doit mettre en place un moyen de s'en protéger, même partiellement.

Enfin, les GLM présentent également des limites dont on se doit d'être avertis avant application (§ 3.2.6).

3.2.1 Etape 1 : le choix d'un modèle

La première étape consiste à **faire le choix d'un modèle**, en définissant :

- la variable réponse (Y)
- la loi de probabilité de Y
- la fonction lien g
- les variables explicatives: $x^1, \ldots, x^p$.

Variable réponse et loi de probabilité

La variable réponse Y est la variable aléatoire à expliquer et pour laquelle l'espérance est déterminée. Par exemple, le montant des sinistres peut être considéré comme une variable réponse.

Soit n observations $Y_1, \ldots, Y_n$ considérées comme des réalisations de la variable Y. Cette dernière peut être binaire, discrète ou encore continue. La loi de Y doit appartenir à la famille exponentielle, c'est-à-dire que sa densité s'écrit sous la forme :

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

Où,

- $a(.), b(.), c(.)$ sont des fonctions connues et dérivables
- ϕ est le **paramètre de dispersion**
- θ est le **paramètre naturel**. Il est lié aux deux premiers moments de la loi.

En effet, il est possible de montrer[1] que :

$$E(Y) = \boldsymbol{\mu} = b'(\theta)$$

$$Var(Y) = \sigma^2 = b''(\theta) \times a(\phi)$$

[1] La démonstration est disponible dans [M2] - p87

Remarque :

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.

Distribution	$\boldsymbol{\theta_i}$	$\boldsymbol{b(\theta_i)}$	$\boldsymbol{a(\phi)}$	$\boldsymbol{\mu_i}$
Bernouilli $(1,\pi_i)$	$\ln(\frac{\pi_i}{1-\pi_i})$	$\ln(1+\exp(\theta_i))$	1	$\pi_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$
Binomiale négative (r,p) (r constant)	$\ln(p)$	$r \times \ln(1+\exp(\theta_i))$	1	$\exp(\theta_i) = \frac{r(1-p)}{p}$
Gamma $\Gamma(V, \frac{\upsilon}{\mu_i})$	$-\frac{1}{\mu_i}$	$-\ln(-\theta_i)$	$\frac{1}{\upsilon}$	$-\frac{1}{\theta_i} = \mu_i$
Poisson (λ_i)	$\ln(\lambda_i)$	$\exp(\theta_i)$	1	$\exp(\theta_i) = \lambda_i$
Normale $N(\mu_i, \sigma^2)$	μ_i	$\frac{{\theta_i}^2}{2}$	σ^2	$\theta_i = \mu_i$

Tableau 3 - Synthèse des principaux facteurs des Modèles Linéaires Généralisés par distribution

Critères de sélection de la loi

Il arrive souvent que plusieurs lois soient pertinentes pour la modélisation de la variable à expliquer. Dans un tel cas, de nombreux critères sont utilisés pour sélectionner le meilleur modèle et ainsi les départager.

Les deux critères les plus utilisés sont :

- le critère **AIC** ;
- le critère **BIC**.

Ces critères permettent de choisir dans un premier temps le meilleur modèle en termes d'ajustement aux données. Ils permettent également de choisir le meilleur modèle en termes de variables à retenir. Ils font tous deux intervenir dans leur calcul le nombre de paramètres du modèle ainsi que la Log Vraisemblance LL maximisée de ce dernier donnée par :

$$\max_{\beta} \mathrm{LL}\,(y, \theta(\beta), \phi) = \max_{\beta}\ \ln\left(\prod_{i=1}^{n} f(y_i|\theta, \phi)\right)$$

Critère AIC :

Une méthode de comparaison de modèles s'appelle le critère AIC (Critère d'Information d'AKAIKE) qui s'applique aux modèles estimés par Maximum de Vraisemblance. Le principe consiste à calculer pour chacun des modèles le critère :

$$\mathbf{AIC} = -\mathbf{2}\ \mathbf{LL} + \mathbf{2}(\mathbf{p}+\mathbf{1})$$

Où

- LL est la Log Vraisemblance maximisée du modèle ;
- $p+1$, correspond au nombre de paramètres du modèle.

C'est un compromis entre le biais, qui diminue lorsque le nombre de paramètres augmente, et la parcimonie, qui correspond à la description des données avec le plus petit nombre de paramètres possible.
Pour un GLM, chaque combinaison de variables explicatives choisies pour être incluses dans le modèle constitue un modèle en soi, et il faut décider quelles variables on conserve dans le modèle final. **Le meilleur modèle est celui qui présente le critère AIC minimal.**

Critère BIC

Une deuxième méthode de comparaison de modèle s'appelle le critère BIC (Bayesian Information Criterion)

$$\mathbf{BIC} = -\mathbf{2\ LL} + (\mathbf{p}+\mathbf{1})\ \mathbf{log}(\mathbf{n})$$

Où

- LL est la log vraisemblance des paramètres associée aux données ;
- $p+1$ correspond au nombre de paramètres du modèle ;
- n est le nombre d'individus composant l'échantillon.

Ce critère pénalise de manière plus importante le nombre de variables présentes dans le modèle que le critère AIC. **Le meilleur modèle est celui qui présente le critère BIC minimal.**

La fonction de lien

Soit une fonction de lien monotone et dérivable g.
Cette fonction matérialise la relation entre l'espérance de notre variable réponse et la composante déterministe.

$$g(E[Y]) = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p$$

Attention, ce n'est pas la variable réponse qui est transformée mais son espérance.

Les fonctions de lien les plus courantes sont les suivantes :

- **La fonction identité :** $g(x) = x$

Il en découle alors :

$$E[Y] = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p$$

C'est le modèle linéaire classique, aussi appelé modèle additif.

- **La fonction ln :** $g(x) = \ln(x)$

Il en découle alors :

$$E[Y] = \exp(\beta_0) \times \exp(\beta_1 x^1) \times \ldots \times \exp(\beta_p x^p)$$

C'est le modèle multiplicatif.

- **La fonction logit :** $g(x) = \ln(\frac{x}{1-x})$

Il en découle alors :

$$E[Y] = \frac{\exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p)}{1 + \exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p)}$$

C'est le modèle de régression logistique. Ce modèle est très adapté au cas où la variable à expliquer est comprise entre 0 et 1.

- **La fonction inverse :** $g(x) = \frac{1}{x}$

Il en découle alors :

$$E[Y] = \frac{1}{(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p)}$$

Choix de la fonction de lien

Chacune des lois de la famille exponentielle possède une fonction de lien spécifique, dite **fonction de lien canonique**, définie par $\theta = \eta$. Celle-ci permet de relier l'espérance μ au paramètre naturel θ.

Rappelons que $E[Y_i] = \mu_i$ pour $i \in \{1, \dots, n\}$.

Le lien canonique est tel que :

$$g(\mu) = \theta = \eta$$

Or,

$$\mu = b'(\theta)$$

D'où

$$\boxed{g^{-1} = b'}$$

Lors de la mise en place du GLM, rien ne nous oblige à choisir la fonction de lien canonique comme fonction de lien. **Ce choix dépend également de la fonction de réponse (Y).**

Il faut s'assurer que les valeurs prédites par le modèle par le biais de cette fonction de lien respectent la nature des valeurs de la variable Y.

Ainsi, dans le cas où la fonction réponse est une loi Gamma, pour éviter d'avoir des résultats négatif, il est préférable d'utiliser la fonction de lien « log » à la fonction « inverse ».
En pratique, la fonction de lien log est souvent utilisée car elle permet d'avoir un tarif multiplicatif.

Loi de probabilité	Fonction de lien canonique	
Bernoulli	$\eta = \ln(\frac{\mu}{1-\mu})$	Logit
Binomiale négative (r constant)	$\eta = \ln(p)$	Log
Gamma	$\eta = \frac{1}{\mu}$	Inverse
Poisson	$\eta = \ln(\mu)$	Log
Normale	$\eta = \theta$	identité

Tableau 4 - Liens canoniques associés aux lois de probabilité usuelles de la famille exponentielle

Les variables explicatives

Comme pour le modèle linéaire classique, considérons une combinaison linéaire des variables explicatives : $\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p$.

Le choix des variables explicatives n'est pas anodin. Il faut choisir un nombre limité de variables pour que le modèle soit utilisable en pratique, mais suffisant pour qu'il soit pertinent.
Il faut aussi choisir les « meilleures » variables au sens explicatif du terme, c'est-à-dire, celles qui ont le plus d'impact sur la variable à expliquer.

Pour ce faire, une **sélection individuelle des variables peut être réalisée au préalable**. Cette approche consiste à tester la pertinence des variables une à une séparément mais elle ne permet pas d'observer les effets croisés.

3.2.2 Etape 2 : estimation des coefficients de la régression

Les coefficients de la régression $\beta_0, \beta_1, \dots, \beta_p$ ainsi que le paramètre de dispersion ϕ sont inconnus et doivent être estimés. On se focalisera uniquement sur l'estimation des coefficients de la régression grâce à la méthode du Maximum de Vraisemblance : il s'agit donc de maximiser la Log-Vraisemblance.

Vraisemblance d'un échantillon

La vraisemblance d'un échantillon est la probabilité d'observer cet échantillon. L'échantillon doit être composé d'un groupe ordonné de réalisations y_i de n variables Y_i .

Supposons n variables indépendantes, la probabilité d'observer cet échantillon est le produit des probabilités d'observer chacune des réalisations. La vraisemblance de l'échantillon, notée L, est alors:

$$L(y, \theta, \phi) = \prod_{i=1}^{n} f(y_i|\theta, \phi)$$

Maximisation de la Log Vraisemblance

La Log Vraisemblance LL correspond au log de la fonction de vraisemblance :

$$\begin{aligned} \mathrm{LL}\,(\mathrm{y}, \theta(\beta), \phi) &= \ln\left(\prod_{i=1}^{n} f(y_i|\theta, \phi)\right) \\ &= \sum_{i=1}^{n} \ln\big(f(y_i|\theta, \phi)\big) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \end{aligned}$$

La recherche des estimateurs de vraisemblance revient à rechercher les $\beta_0, \beta_1, \dots, \beta_p$ qui vérifient les équations :

$$A_j = 0 \; pour \; j \in \{0,1, \dots, p\}$$

Où

$$\begin{aligned} A_j &= \frac{\partial LL\,(y, \theta(\beta), \phi)}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial \ln\big(f(y_i|\theta, \phi)\big)}{\partial \beta_j} \end{aligned}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right)$$

A_j est déterminé à partir de la formule suivante :

$$\frac{\partial \ln\left(f(y_i|\theta_i, \phi)\right)}{\partial \beta_j} = \frac{\partial \ln(f(y_i|\theta_i, \phi))}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$$

Comme $\mu_i = b'(\theta_i)$, il en découle :

$$\frac{\partial \ln(f(y_i|\theta_i, \phi))}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$$

et

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

De plus,

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = x_i^j \frac{\partial \mu_i}{\partial \eta_i}$$

Par conséquence,

$$\frac{\partial \ln\left(f(y_i|\theta_i, \phi)\right)}{\partial \beta_j} = \frac{\frac{\partial \ln(f(y_i|\theta_i, \phi))}{\partial \theta_i} \frac{\partial \mu_i}{\partial \beta_j}}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{(y_i - \mu_i) x_i^j \frac{\partial \mu_i}{\partial \eta_i}}{a(\phi) b''(\theta_i)} = \frac{(y_i - \mu_i) x_i^j}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

Et finalement

$$\boxed{A_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_i^j}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}}$$

Le paramètre ϕ n'apparait plus dans l'expression de $(A_j)_{0 \le j \le p}$. Les équations de vraisemblance peuvent être résolues sans s'en préoccuper.

Les équations obtenues ne sont pas linéaires en β et leur résolution passe par l'utilisation de méthodes itératives. La détermination des coefficients de la régression se fera à l'aide du logiciel R.

3.2.3 Etape 3 : Procédure de sélection des variables

Même après une première sélection des variables, via la **sélection individuelle des variables**, il est possible que certaines des variables explicatives utilisées dans le modèle ne soient pas significatives. Elles n'apportent pas d'information sur la variable à expliquer et ne doivent donc pas figurer dans le modèle. Il faut alors procéder à une deuxième sélection : la **sélection conjointe des variables**, qui consiste à tester la pertinence d'une variable en considérant les variables dans leur ensemble.
La sélection conjointe des variables est fondée sur l'utilisation d'algorithmes appliquant la régression de façon répétée. Trois méthodes sont habituellement utilisées :

- **La procédure Ascendante (Forward Selection) :** Elle consiste à utiliser un modèle linéaire avec seulement une variable explicative et à ajouter au fur et à mesure la variable explicative qui permet de faire diminuer l'AIC du modèle de manière la plus significative. Le processus est alors arrêté, lorsque l'ajout d'une variable supplémentaire entraine une hausse de l'AIC.

- **La procédure descendante: (Backward Selection) :** Elle consiste, à l'opposé de la procédure Forward, à calculer l'AIC du modèle complet (avec toutes les variables explicatives disponibles) ainsi que l'impact du retrait de chaque variable sur l'AIC. Le processus supprime ainsi du modèle la variable dont le retrait permet une baisse significative de l'AIC.

- **La procédure pas à pas mixte (STEPWISE Selection) :** Cette méthode est une combinaison des deux méthodes précédentes. Les variables jugées les plus significatives pour le modèle sont sélectionnées. A chaque étape, la variable la plus significative restante est ajoutée, puis les variables qui ne sont plus significatives sont retirées. En effet, en raison des corrélations, l'ajout d'une variable explicative peut diminuer la significativité d'une autre variable explicative déjà présente. La procédure Stepwise consiste donc à effectuer une procédure ascendante et descendante à chaque étape.

3.2.4 Etape 4 : Validité du modèle

La dernière étape consiste à **s'assurer de la validité du modèle** mis en place. Pour cela, il est nécessaire de :

- s'assurer de la cohérence des coefficients calculés par rapport aux études descriptives effectuées en amont,
- examiner les écarts entre les valeurs théoriques (obtenues avec le modèle) et les valeurs observées,
- s'intéresser aux déviances du modèle, qui permettent de légitimer l'utilisation du modèle mais également de comparer plusieurs modèles.

Cohérence des coefficients calculés

Même si les coefficients de la régression sont déterminés de manière automatique, les résultats obtenus doivent être contrôlés. Ainsi, il est nécessaire de faire attention :

- **Au signe des coefficients:**
  Dans le cas du modèle linéaire, un signe positif (respectivement négatif) pousse à la hausse (à la baisse) la variable à expliquer.
  Cependant, il faut prêter attention à la fonction de lien dans le cas des GLM. Ainsi par exemple, avec la **fonction de lien log** et d'une **variable à expliquer positive**, l'interprétation du signe des coefficients est identique à celle évoquée pour les modèles linéaire. En effet, un coefficient négatif revient à multiplier l'espérance de la variable à expliquer par un coefficient inférieur à $\exp(0) = 1$.

- **A la valeur des coefficients :**
  Dans le cas de variables qualitatives, plus la valeur absolue attribuée au coefficient β_j est grande, plus la modalité de la variable explicative a de poids dans l'explication de la variable réponse.

Écarts entre les valeurs du modèle et les valeurs observées

Un modèle est d'autant meilleur que les valeurs qu'il retourne sont proches des valeurs observées. Pour mesurer les écarts entre les valeurs théoriques et observées, il est intéressant d'observer :

- le rapport des sommes des valeurs observées et des valeurs théoriques dans le cas d'une variable réponse quantitative,
- le taux d'individus bien classés dans le cas d'une variable qualitative.

Rapport des sommes entre valeurs observées et valeurs théoriques

Ce quotient Q nous permet d'évaluer la qualité du modèle en « moyenne ». Il ne s'agit pas ici de voir si chaque valeur théorique est proche de la valeur réellement observée mais plutôt de savoir si la moyenne des valeurs théoriques est proche de la moyenne des valeurs observées.

Ainsi ce quotient vaut :

$$Q = \frac{\frac{\sum_{i=1}^{n} y_i}{n}}{\frac{\sum_{i=1}^{n} \hat{y}_i}{n}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{y}_i}$$

Où,

- $\frac{\sum_{i=1}^{n} y_i}{n}$ correspond à la moyenne des valeurs réellement observées$(y_i)_{1 \leq i \leq n}$
- $\frac{\sum_{i=1}^{n} \hat{y}_i}{n}$ correspond à la moyenne des valeurs théoriques obtenues par le modèle $(\hat{y}_i)_{1 \leq i \leq n}$

Dans le cadre de la modélisation d'une charge de sinistre, les valeurs observées correspondent aux montants des sinistres réalisés par individu et les valeurs théoriques à la prévision du montant de charge qui sera affecté à chaque assuré, c'est-à-dire sa Prime Pure. Le rapport Q est alors appelé **ratio des Sinistres sur Primes** (S/P).

La valeur de ce ratio permet de définir la politique tarifaire du modèle :

- Si $Q < 1$ le modèle surévalue le risque ;
- Si $Q > 1$ le modèle sous-évalue le risque.

Remarque :
Le rapport calculé ici est proche de ce que l'on appelle le **Ratio Combiné** (*Combined ratio*, noté CoR) donné par :

$$CoR = \frac{\mathrm{S} + \mathrm{F}}{\mathrm{P}}$$

Où

- S correspond à la charge de sinistre totale de l'assureur sur la période considérée
- P désigne l'ensemble des primes acquises par ce dernier
- F désigne les coûts supportés par l'entreprise sur cette même période

L'étude de ce ratio permet de mesurer la suffisance du tarif pour honorer le versement des prestations en cas de sinistre.
Attention, ce calcul est effectué en aval de la mise en place de la tarification, c'est-à-dire sur les sinistres observés et non pas l'historique ayant servi à mettre en place le tarif comme c'est le cas pour notre ratio de Sinistres sur Primes.

Taux de bien classés

Dans le cas où la variable à expliquer est qualitative, l'analyse du quotient entre la somme des valeurs observées et celle des valeurs théoriques n'a plus de sens.

Cette analyse est ici remplacée par le calcul d'un taux de bien classés τ_{BC} donné par :

$$\tau_{BC} = \frac{\sum_{i=1}^{n} 1_{\{y_i = \hat{y}_i\}}}{n}$$

Avec

$$1_{\{y_i = \hat{y}_i\}} = \begin{cases} 1, & \text{si l'individu i est bien classé par le modèle} \\ 0, & \text{sinon} \end{cases}$$

Les déviances [1]

Déviance du modèle

L'écart entre les valeurs théoriques et observées peut également être mesuré par la **déviance**, notée D, du modèle. Celle-ci représente l'écart entre la Log-Vraisemblance obtenue en β, et celle obtenue avec un modèle parfait, dit **saturé**.
Le modèle saturé correspond au modèle où la moyenne de la variable est définie par l'observation elle-même :

$$E[Y_i] = y_i$$

D'où

$$\begin{aligned} D &= \sum_{i=1}^{n} d_i \\ &= 2 \sum_{i=1}^{n} \ln\big(f(\hat{\mu}_i|\theta,\phi)\big) - \ln\big(f(y_i|\theta,\phi)\big) \\ &= 2\left(LL\,(\hat{\mu}, \theta(\beta), \phi) - LL\,(y, \theta(\beta), \phi)\right) \\ &= 2\,(LL_{sat} - LL\,) \end{aligned}$$

Avec,

- d_i la déviance due à la $i^{ème}$ observation
- LL la log vraisemblance du modèle
- LL_{sat} la log vraisemblance du modèle saturé

L'objectif lors de l'ajustement d'un GLM est de **minimiser D**. Toutefois, il ne faut pas conclure hâtivement que si deux modèles M_1 et M_2 ont respectivement des déviances D_1 et D_2 tels que $D_1 < D_2$ alors le modèle M_1 est meilleur. En effet, la déviance dépend des données et de la spécificité du modèle. Cette comparaison n'est valable que dans le cas de **modèles emboités**.

Deux modèles M_1 (modèle initial) et M_2 (modèle complet) sont emboîtés s'ils concernent les mêmes données et possèdent respectivement q_1 et q_2 variables telles que $q_1 < q_2$ et l'ensemble des variables utilisées dans M_1 est également utilisé dans M_2.

[1] Source : [P2]

Nous nous intéressons alors à la différence de leurs déviances, c'est-à-dire à **la décroissance de la déviance** lors du passage d'un modèle initial à un modèle plus complet donnée par :

$$\begin{aligned}\Delta_{1,2} &= D_2 - D_1 \\ &= 2(LL_{sat} - LL_2) - 2(LL_{sat} - LL_1) \\ &= 2(LL_1 - LL_2)\end{aligned}$$

Où

- D_1 est la déviance du modèle initial et D_2 la déviance du modèle complet
- LL_1 (respectivement LL_2) la Log Vraisemblance de M_1 (respectivement M_2)

Cette décroissance de la déviance est nommée **Test du rapport de vraisemblance.**

Les hypothèses du test sont les suivantes :

- H_0: Le sous modèle convient : $\beta_{q_1+1} = \beta_{q_1+2} = \cdots = \beta_{q_2} = 0$
- H_1 : Le modèle complet est meilleur : $(\beta_s)_{q_1<s<q_2} \neq 0$

Dans le cas où la loi suivie par les données admet deux paramètres (loi Gamma, loi Normale), $\Delta_{1,2}$ s'interprète comme une statistique de Fisher. Ainsi, pour un modèle initial contenant q_1 variables explicatives et un modèle complet en contenant q_2 , la décroissance de la déviance suit la distribution de Fisher $F(q_2 - q_1, n - p)$.

Si la valeur calculée de la statistique ($\Delta_{1,2}$) est supérieure à la valeur critique de Fisher, valeur lue dans la table pour un seuil de précision donné et un nombre de paramètres fixé, H_0 est rejetée et le modèle complet est meilleur que le modèle restreint.

En terme de p-value, si le seuil d'erreur est fixé à α, on rejette H_0 si la p-value est inférieure à α .

Déviance résiduelle

Il existe également une autre mesure de déviance, appelée **déviance résiduelle** D_0 du modèle.
Cette dernière est donnée par le double de la variation de Log Vraisemblance entre le modèle nul et le modèle saturé.

Le modèle nul correspond au cas où $E[Y_i] = c$ avec c une constante qui est estimée par la méthode de maximum de vraisemblance.

Soit :

$$D_0 = 2(LL_{sat} - LL_0)$$

Où

$$LL_0 = LL\ (c, \theta(\beta), \phi)$$

Déviance Standardisée

En pratique, la comparaison de deux modèles repose sur l'étude de leur **déviance Standardisée**, aussi appelée déviance normalisée, D_S qui permet de tenir compte du fait que les deux modèles testés n'admettent pas les mêmes paramètres de loi.

Cette dernière est donnée par :

$$\mathrm{D_S} = \frac{\mathrm{D_0}}{\phi}$$

Où ϕ est le paramètre de dispersion du modèle.

La déviance Standardisée est une statistique de test qui vérifie :

$$D_S \sim \chi^2_{n-p}$$

Avec,

- n le nombre d'observations utilisées pour construire le modèle
- p le nombre de variables explicatives du modèle.

De ce fait, elle peut être utilisée pour mesurer la qualité du test d'adéquation.

La moyenne d'une distribution du Chi-2 est égale à son degré de liberté, d'où :

$$E[D_S] = n - p$$

Si le modèle est en bonne adéquation avec les données, la déviance Standardisée doit être proche de la valeur n-p. Malheureusement la précision du test est douteuse dès lors que l'échantillon utilisé pour réaliser le test est de petite taille.
En connaissant les aspects approximatifs des tests construits, l'usage est souvent de comparer D_S avec le nombre de degrés de liberté (ddl) donné par :

$$ddl = n - p$$

Un modèle est jugé satisfaisant si $\frac{D_S}{ddl} < 1$.

3.2.5 Le risque de modèle

Malgré toutes les vérifications préalables sur la validité du modèle, un risque de modèle persiste. Il correspond au risque de perte pour l'assureur résultant du fait que le modèle utilisé ne reflète pas le risque réel observé.

Ce risque peut être scindé en :

- **un risque d'erreur de spécification,** lié à la mauvaise adéquation du modèle. Cela peut se traduire par la sélection d'un modèle ou d'une loi inappropriée, ou encore par un mauvais choix des variables explicatives retenues.
- **un risque d'estimation,** lié à l'approximation des paramètres du modèle.

Le risque du modèle peut être évalué :

- en analysant les résidus (termes d'erreur) du modèle,
- en effectuant des *Backtesting* et des *Stresstesting* qui permettent de construire des ratios de Sinistres sur Primes dynamiques,
- en calculant l'intervalle de confiance des coefficients obtenus lors de la régression : plus l'intervalle de confiance est grand, moins le modèle est précis.

Conscients de ce risque, les assureurs souhaitent s'en protéger. Pour cela, l'assureur applique un chargement de sécurité à son tarif. Soit S le montant de la charge totale de sinistres et α une constante, le chargement peut être de la forme :

- $\alpha\ \mathrm{E}(S)$: cette forme est très utilisée en pratique mais est parfois critiquée car elle ne dépend pas des fluctuations de S ;
- $\alpha\ \mathrm{Var}(S)$: fonction de la variance ;
- $\alpha\ \sqrt{\mathrm{Var}(S)}$: fonction de l'écart-type.

Analyse des résidus

Avant toute chose, nous définissons la matrice H $(n \times n)$(« *Hat matrix* ») donnée par :

$$H = X(X'X)^{-1}X'$$

Sur la diagonale de la matrice H en $i^{ème}$ position se trouve le « levier » h_{ii} de la $i^{ème}$ observation :

$$h_{ii} = h_i = x_i(X'X)^{-1}x_i'$$

Où x_i représente la $i^{ème}$ ligne de la matrice X.

Le rôle de la matrice H est important puisqu'elle permet de passer des Y vers les projections $\hat{Y}$ mais également des erreurs théoriques ε vers les résidus observés $\hat{\varepsilon}$ selon :

$$\hat{Y} = HY = X\hat{\beta}$$

$$\hat{\varepsilon} = (I - H)Y$$

De plus, H admet une propriété intéressante : elle est idempotente[1], c'est-à-dire que $\mathrm{H^m} = \mathrm{H}$, $\forall m \in \mathbb{N}$. En effet,

$$\begin{aligned} H.H &= X(X'X)^{-1}X'.X(X'X)^{-1}X' \\ &= X(X'X)^{-1}(X'X)(X'X)^{-1}X' \end{aligned}$$

Or

$$(X'X)(X'X)^{-1} = \mathrm{I}$$

Donc

$$\begin{aligned} &= X(X'X)^{-1}.I.X' \\ &= X(X'X)^{-1}X' \end{aligned}$$

De plus, ceci implique que $I - H$ est aussi idempotente car :

$$\begin{aligned} (\mathrm{I} - \mathrm{H}).(\mathrm{I} - \mathrm{H}) &= \mathrm{I}(\mathrm{I} - \mathrm{H}) - \mathrm{H}(\mathrm{I} - \mathrm{H}) \\ &= \mathrm{I} - \mathrm{H} - \mathrm{H} + \mathrm{H}.\mathrm{H} \\ &= \mathrm{I} - 2\mathrm{H} + \mathrm{H}.\mathrm{H} \\ &= \mathrm{I} - 2\mathrm{H} + \mathrm{H} \\ &= \mathrm{I} - \mathrm{H} \end{aligned}$$

[1] Source : [P3].

Résidus bruts, Standardisés et Studentisés

Les résidus bruts du modèle sont donnés par :

$$\boxed{\hat{\varepsilon}_i = y_i - \hat{y}_i}$$

Si $\hat{\varepsilon}_i$ est élevé, ceci signifie que la valeur y_i a mal été reconstituée par le modèle.
Ces résidus sont souvent étudiés dans le cadre d'un modèle linéaire mais moins dans le cadre d'un GLM. En effet, même si une hypothèse d'homoscédasticité a été faite sur les erreurs[1], il faut remarquer que les résidus observés sont **hétéroscédastiques :**

$$\sigma^2_{\hat{\varepsilon}_i} = Var(\hat{\varepsilon}_i) = \sigma^2_\varepsilon(1 - h_i)$$

car

$$Var(\hat{\varepsilon}_i) = Var\big(Y_i - \hat{Y}_i\big) = Var\big((I - H)Y_i\big) = \sigma^2(1 - h_i)^2 = \sigma^2(1 - h_i)$$

Par idempotence de la matrice $I - H$.

Nous devons donc normaliser le résidu par son écart type pour rendre les écarts comparables.
Lorsque nous travaillons sur un échantillon, nous ne disposons pas de la vraie valeur de σ^2_ε(variance des résidus) mais nous l'estimons par :

$$\hat{\sigma}^2_{\hat{\varepsilon}_i} = \hat{\sigma}^2_\varepsilon(1 - h_i)$$

Où

- h_i est lue dans la *Hat Matrix* H,
- $\hat{\sigma}^2_\varepsilon = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}$ est l'estimateur de la variance de l'erreur[2].

Le résidu Standardisé, également appelé résidu de Pearson est alors donné par le rapport :

$$\boxed{t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{\hat{\varepsilon}_i}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon\sqrt{(1 - h_i)}}}$$

Résidu Studentisé [3]

Le résidu Standardisé est un indicateur certes intéressant mais il présente un inconvénient fort : nous évaluons l'importance du résidu $\hat{\varepsilon}_i$ d'une observation qui a participé à la construction de la droite de régression. De fait, le point est juge et partie dans l'évaluation.

Une mesure objective devrait ne pas faire participer le point i dans la construction du modèle utilisé pour prédire la valeur $\hat{y}_i$.

Pour tester l'adéquation du modèle, le résidu Studentisé est utilisé. Il est défini par :

$$\boxed{t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_i}}}$$

Où $\hat{\sigma}_{(i)}$ est l'estimation de la variance sans tenir compte de la $i^{ème}$ observation. L'idée est de se rapprocher des méthodes dites, de validation croisées, qui consistent à sortir de l'apprentissage les données que l'on souhaite tester.

---

[1] Variance de l'erreur par hypothèse d'homoscédasticité : $\sigma^2_{\varepsilon_i} = Var(\varepsilon_i) = \sigma^2_\varepsilon$
[2] Preuve disponible en [P1] - p15
[3] Source : [P4] - p34

Nous pouvons alors montrer que[1] :

$$t_i^* \sim S(n-p-1)$$

Où $S(k)$ symbolise la loi de Student à k degrés de liberté.

Asymptotiquement, cette loi tend vers la loi Gaussienne standard N(0,1) : il faut donc s'attendre à ce que 95 % des valeurs soient comprises entre -2 et 2, ce qui correspond respectivement à une approximation des quantiles de la loi gaussienne centrée réduite à 2,5 % et 97,5 %.

Concrètement, les résidus Studentisés servent à détecter les données aberrantes, c'est-à-dire les données telles que l'erreur de prédiction est grande devant $\sigma\sqrt{1-h_i}$.

Attention à ne pas confondre **donnée aberrante** et **donnée influente** : la suppression d'une donnée influente entraine un changement significatif des paramètres du modèle tandis qu'une donnée aberrante présentera une valeur de résidu ($|t_i^*|$) significativement grand.

Résidus de déviance

Il existe également un autre type de résidus, appelé **résidus de déviance,** donné par :

$$\hat{\varepsilon}_{D_i} = \text{signe}\,(\hat{\varepsilon}_i) \times \sqrt{d_i}$$

Avec d_i la déviance associée à la $i^{ème}$ observation.

Vérifications graphique

Une fois ces définitions établies, nous allons maintenant décrire un ensemble d'outils visuels qui permettent de vérifier que les résidus obtenus possèdent bien toutes les propriétés auxquelles on s'attend.

La nullité de l'espérance des résidus et leur caractère homoscédastique

Ces propriétés peuvent être observées sur le tracé de la racine des valeurs absolues des résidus en fonction des prédictions du modèle (« *residuals versus fitted* »). Le nuage de points doit être réparti aléatoirement. L'absence de tendance et la constance de la variabilité de l'erreur viennent confirmer ces deux hypothèses.

L'adéquation des résidus à une loi Normale

L'adéquation des résidus à une loi normale peut être vérifiée à l'aide d'un QQ-plot (Quantile to Quantile Plot).

Ce dernier correspond à la représentation graphique des quantiles des résultats théoriques (en abscisse) en fonction des quantiles des données observées (en ordonnée).
Si l'hypothèse de normalité des résidus est vérifiée, alors les points de la représentation graphique sont approximativement alignés autour de la première bissectrice.

[1] Preuve disponible dans [O4] - p54 (Chapitre 3) / p69 (Chapitre 4)

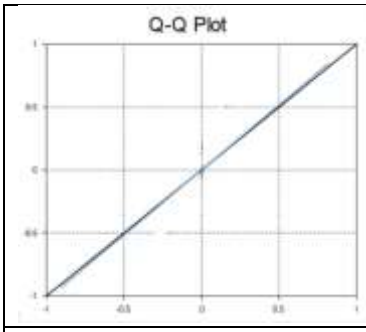
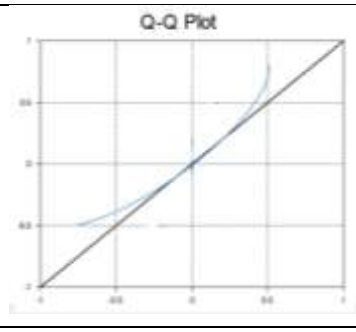
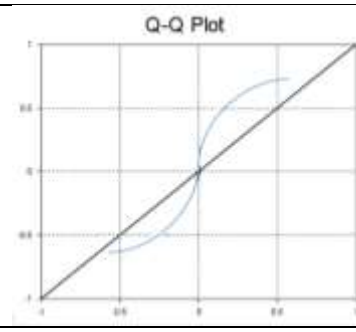
		
Bonne adéquation	Distribution asymétrique des résidus	Grande influence des valeurs extrêmes

Illustration 1 - Allures possibles des QQ-plot et interprétation

Mesurer l'influence d'une observation sur la modélisation : distance de Cook

Nous avons vu que les résidus Studentisés permettent de mesurer l'influence d'une observation sur la modélisation.

Intéressons-nous à présent non pas à l'influence de la $i^{ème}$ observation sur la valeur prédite mais à l'influence de la $i^{ème}$ observation sur l'ensemble des valeurs prédites. Il s'agit donc de 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 la distance de Cook de la $i^{ème}$ observation (DC_i) qui peut être exprimée en fonction des résidus Standardisés où p représente le nombre de variables explicatives :

$$DC_i = \frac{1}{p}\frac{h_i}{(1-h_i)}t_i^2$$

Le seuil de tolérance associé à cette distance de Cook le plus souvent employé est 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.

3.2.6 Limites du GLM

Les Modèles Linéaires Généralisés ont certes l'avantage de permettre le choix de la loi suivie par les données mais présentent tout de même certaines limites. Elles concernent notamment la détection et la modélisation d'interactions entre les variables quantitatives ou qualitatives ainsi que l'utilisation de variables explicatives quantitatives continues.

Modélisation d'interactions entre les variables

Dans un modèle GLM, les interactions entre variables explicatives doivent être spécifiées à priori par le statisticien. Par exemple, si deux variables explicatives sont fortement corrélées, leur insertion dans le modèle pourrait avoir un effet redondant sur la variable à expliquer. Il faut donc tester les interactions entre variables explicatives au préalable. Ces tests peuvent s'avérer fastidieux si le nombre de variables explicatives est important.

L'utilisation de variables explicatives quantitatives continues

Les valeurs prises par les variables quantitatives continues sont souvent uniques et nombreuses.
Si la $j^{ème}$ variable du modèle est de ce type, l'estimation du coefficient β_j prendra en compte l'ensemble des valeurs prises par la variable. Il arrive souvent que certaines valeurs soient extrêmes et non représentatives. Elles faussent l'estimation du coefficient.
De plus, au vu du nombre de valeurs prises par cette variable, le GLM ne donnera pas un poids significatif à chacune d'entre elles.

Hormis les problèmes d'estimations, il existe un problème d'ordre Marketing : une prime ne peut pas être attribuée pour chaque valeur possible de la variable quantitative.

Il est alors préférable de **transformer les variables quantitatives continues en variables quantitatives discrètes** ou mieux encore, de les **découper en classes de variables qualitatives**.
Il faut alors choisir un découpage optimal. Ce choix repose en grande partie sur le jugement de l'actuaire qui les réalise.

3.3 La régression logistique : un cas particulier de GLM

La régression logistique est un cas particulier de GLM. Elle correspond à l'utilisation d'une **loi Binomiale** avec pour fonction de lien la fonction **logit (**cf. § 3.2.1).

La régression logistique est surtout employée comme une technique prédictive. Elle vise à construire un modèle permettant de prédire les valeurs prises par une variable cible qualitative à partir d'un ensemble de variables explicatives.

Cette technique très populaire est exploitée dans différents domaines allant du marketing à l'épidémiologie. Nous présentons ici les fondements théoriques de cette méthode utilisée par la suite pour la mise en place d'un modèle prédictif du genre de l'assuré.

3.3.1 Modélisation des coefficients $\boldsymbol{\beta}$ de la régression logistique

On considère donc une variable binaire Y pouvant prendre deux valeurs ($\mathrm{Y} = 0$ ou $\mathrm{Y} = 1$) et l'on désire expliquer Y en fonction d'un vecteur $\mathbf{x}$ composé de *p* variables explicatives $\mathbf{x} = (x^1, x^2, \dots, x^p)$.

Soit :

$$\pi(\mathbf{x}) = \mathrm{E}(\mathrm{Y}|\mathbf{x}) = \mathrm{P}(\mathrm{Y} = 1|\mathbf{x}) \quad (1)$$

$$\mathrm{Y} = \pi(\mathbf{x}) + \varepsilon \quad (2)$$

L'erreur $\varepsilon = Y - \pi(\boldsymbol{x})$ ne peut prendre que deux valeurs, à savoir :

$$\varepsilon = \begin{cases} 1 - \pi(\mathrm{x}), & \text{si } \mathrm{Y} = 1 \text{ avec la probabilité } \pi(\mathrm{x}) \\ -\pi(\mathrm{x}), & \text{si } \mathrm{Y} = 0 \text{ avec la probabilité } 1 - \pi(\mathrm{x}) \end{cases}$$

L'espérance de ε est nulle, comme il est immédiat de le vérifier.
Nous nous plaçons ici dans le cadre de la régression logistique où $\pi(\boldsymbol{x})$ est modélisé comme défini en (1).

Rappelons que la fonction **logit** est la fonction : $x \to \ln(\frac{x}{1-x})$.
La régression logistique est alors une modélisation de $\pi(\boldsymbol{x})$ de telle sorte $H(x) := \ln(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})})$ soit une fonction linéaire des $(x^j)_{1\leq j\leq p}$, d'où :

$$\mathrm{H(x)} = \beta_0 + \beta_1 \mathrm{x}^1 + \cdots + \beta_\mathrm{p} \mathrm{x}^\mathrm{p}$$

La régression logistique est un cas particulier du modèle plus général suivant, où $\pi(x)$ est modélisé sous la forme :

$$\pi(\mathrm{x}) = \mathrm{G}(\beta_0 + \beta_1 \mathrm{x}^1 + \cdots + \beta_\mathrm{p} \mathrm{x}^\mathrm{p})$$

avec

- G une fonction connue, croissante, prenant ses valeurs dans [0,1] ;
- $(\beta_j)_{0\leq j\leq p}$ les $p+1$ paramètres de la régression logistique à estimer.

Le modèle logistique correspond au cas où :

$$\mathrm{G(u)} = \frac{\mathrm{e}^\mathrm{u}}{1+\mathrm{e}^\mathrm{u}}$$

La fonction $\mathbf{S(x)} = \boldsymbol{\beta_0} + \boldsymbol{\beta}'\mathbf{x}$ correspond au **score** et il en découle la relation :

$$\pi(\mathrm{x}) = \mathrm{G}(\mathrm{S(x)})$$

3.3.2 Règle d'affectation à une classe

Pour une observation $\mathbf{x}$, l'**affectation** à l'une ou l'autre des classes (Y=0 ou Y=1) est **fonction de la valeur de $\boldsymbol{\pi(x)}$** :

- si $\pi(\mathrm{x}) = \mathrm{G}\big(\mathrm{S(x)}\big) > 0{,}5$: l'individu est associé à la classe Y=1 ;
- si $\pi(\mathrm{x}) = \mathrm{G}\big(\mathrm{S(x)}\big) < 0{,}5$: l'individu est associé à la classe Y=0 ;
- si $\pi(\mathrm{x}) = 0{,}5$, l'affectation est indifférente.

De manière équivalente, elle est aussi **fonction** du **score** de l'individu :

- si $\mathrm{S(x)} > 0$: l'individu est associé à la classe Y=1 ;
- si $\mathrm{S(x)} < 0$: l'individu est associé à la classe Y=0 ;
- si $\mathrm{S(x)} = 0$: l'affectation est indifférente.

Dans le cas d'un ***score*** nul, pour départager le choix du classement nous décidons de procéder au tirage aléatoire d'une variable Z suivant une loi de Bernoulli de paramètre $p = 0{,}5$ ($Z \sim \mathcal{B}(0{,}5)$) puis d'appliquer le critère d'affectation suivant:

- si Z = 1 : affectation de l'individu à la classe 1 ;
- si Z = 0 : affectation de l'individu à la classe 0.

Partie II - Préparation et étude des données disponibles

Après avoir posé le cadre théorique nécessaire à la compréhension de ce mémoire, nous entrons dès cette deuxième partie dans la mise en pratique.

Notre étude se porte sur la tarification d'un contrat d'assurance automobile qui, comme beaucoup de modélisations, doit s'appuyer sur un historique de sinistralité.
Pour commencer une présentation du portefeuille utilisé et des différents traitements qui y seront appliqués avant utilisation (Chapitre 4) est proposée.

Suite à cela, nous analyserons les liens entre les données disponibles concernant la sinistralité (fréquence et coût des sinistres) et les différentes variables explicatives à disposition de l'assureur (Chapitre 5).

Chapitre 4 - Présentation du portefeuille et traitements

Ce chapitre est réservé à la présentation du portefeuille et des divers traitements qui lui seront appliqués pour le rendre utilisable.

Tout d'abord, nous commençons par présenter la base utilisée ainsi que les différentes informations (variables) qui la compose (§ 4.1). Il faudra ensuite s'assurer de la qualité des données ainsi que de leur compatibilité avec une utilisation en pratique (§ 4.2).

Enfin, les variables disponibles subiront elles aussi un traitement (§ 4.3), notamment un regroupement de leurs modalités s'appuyant sur plusieurs méthodes de classifications.
Ces regroupements détermineront le nombre de classes tarifaires proposées par l'assureur.

4.1 Présentation de la base et des variables

L'établissement d'un tarif exige un appui statistique. Ce dernier provient souvent d'un relevé de sinistralité des années antérieures sur un ensemble de contrats proposants des garanties identiques au contrat tarifé.
Nous avons donc appuyé la mise en place de nos modèles de tarification sur un historique de sinistralité automobile.

Pour commencer, présentons les origines de la base de données qui sera utilisée ainsi que les différentes informations disponibles pour chacun des contrats répertoriés.

4.1.1 La base de données

Notre base de données, *Car*, provient de données publiques disponibles à l'adresse suivante :
http://www.businessandeconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/research/books/GLMsforInsuranceData/data_sets.

Elle contient 67 856 lignes : chacune d'entre elles correspond aux données relatives à un contrat d'assurance.

Sur les 67 856 contrats recensés, 4 624 ont au moins un sinistre, soit 6,8 % du portefeuille.
Au total, 4 937 sinistres ont été relevés.

Les données étudiées proviennent d'Australie et concernent la garantie de Responsabilité Civile[1], obligatoire, au même titre qu'en France, pour tous les propriétaires d'un véhicule.
Elle assure le propriétaire contre les dégâts subis par les passagers du véhicule, les piétons, ou les autres conducteurs lors d'un accident.

Cette base contient les données concernant toutes les polices d'assurance et sinistres qui ont eu lieu en 2004 ou 2005 : la période d'observation est donc comprise entre 0 et 1 an.

4.1.2 Les notations et variables de la base

Pour chaque contrat, nous disposons de variables qui caractérisent l'assuré, son contrat ainsi que son véhicule, à savoir :

- **Gender** : le genre de l'assuré (M : homme, F : femme)
- **Agecat** : l'âge de l'assuré sous forme de classes allant de 1 à 6 avec pour correspondance les intitulés suivant :

1. Très jeunes	**4.** Actifs
2. Jeunes	**5.** Personnes âgées
3. Jeunes actifs	**6.** Personnes très âgés

- **Area :** la zone géographique de résidence de l'assuré (codée de A à F)
- **Veh_value** : la valeur du véhicule (en $10,000s).

- **Exposure :** la période durant laquelle l'individu a été observé. Cette exposition est exprimée en année et est comprise entre 0 et 1.
- **Veh_body :** le type de véhicule
  Cette variable possède treize modalités :
  - BUS
  - CONVT = *convertible*
  - COUPE
  - HBACK = *hatchback*
  - HDTOP = *hardtop*
  - MCARA = *motorized caravan*
  - MIBUS = *minibus*
  - PANVN = *panel van*
  - RDSTR = *roadster*
  - SEDAN
  - STNWG = *station wagon*
  - TRUCK = *camion*
  - UTE = *utilitaire*

[1] Le périmètre de sinistres recouvert par cette assurance automobile est important car la Prime Commerciale varie de façon significative en fonction des garanties souscrites.

- **Veh_age :** l'âge du véhicule sous forme de classes allant de 1 à 4 avec pour correspondance les intitulés suivants :

    1. Véhicule très récent
    2. Véhicule récent
    3. Véhicule ancien
    4. Véhicule très ancien

Nous disposons également d'informations sur l'historique de sinistralité de l'assuré durant la période d'observation à travers les variables :

- **Clm** : une variable binaire représentant l'occurrence de sinistre codifié de la sorte :
    - 0 : Aucun sinistre
    - 1 : Au moins un sinistre

- **Numclaims** : le nombre de sinistres observés. Cette variable prend ses valeurs dans {0,1,2,3,4}.

- **Claimcst0** : le montant cumulé des sinistres observés (porté à 0 en cas d'absence de sinistre)

Le nombre de sinistres observé par individu doit être rapporté à la durée d'exposition au risque.
Pour cela, on calcule la **fréquence annuelle de sinistres** définie par :

$$\text{fréquence} = \frac{\text{Numclaims}}{\text{Exposure}}$$

Nous introduisons également une nouvelle variable désignant le coût unitaire d'un sinistre (et non plus celui des sinistres cumulés) donné par :

$$\text{CoûtUnitaire} = \begin{cases} \frac{\text{Claimcst0}}{\text{Numclaims}}, & \text{si Numclaims} > 0 \\ 0, & \text{sinon} \end{cases}$$

4.2 Traitement des données

Cette étape préalable à la mise en place d'un modèle est primordiale.
En effet, une erreur opérationnelle lors du recueil des données est toujours possible (faute de frappe, oubli de renseignement, etc.).

Il faudra donc veiller à la qualité des données récoltées mais aussi à définir le périmètre du modèle que nous allons mettre en place, autrement dit, préciser les modalités du contrat tarifé.

4.2.1 Valeurs manquantes et données erronées

Pour chacun des contrats, toutes les variables demandées sont renseignées : notre base n'admet aucune information manquante.
Cependant, nous avons tout de même vérifié l'exactitude des valeurs renseignées :

- Pour les variables qualitatives, il a fallu s'assurer que la modalité renseignée fasse bien partie de la liste des modalités prévues

- Pour les variables quantitatives, nous avons veillé à ce que la valeur renseignée appartienne bien au domaine de définition de la variable

Après vérification, on remarque que pour la variable ***Veh_value***, qui correspond à la valeur du véhicule, certaines polices ont pour valeur « 0 ».
Il s'agit là d'une valeur par défaut qui témoigne d'un manque d'information (à moins qu'il n'existe des véhicules gratuits, auquel cas un petit voyage en Australie pourrait vite devenir intéressant !).

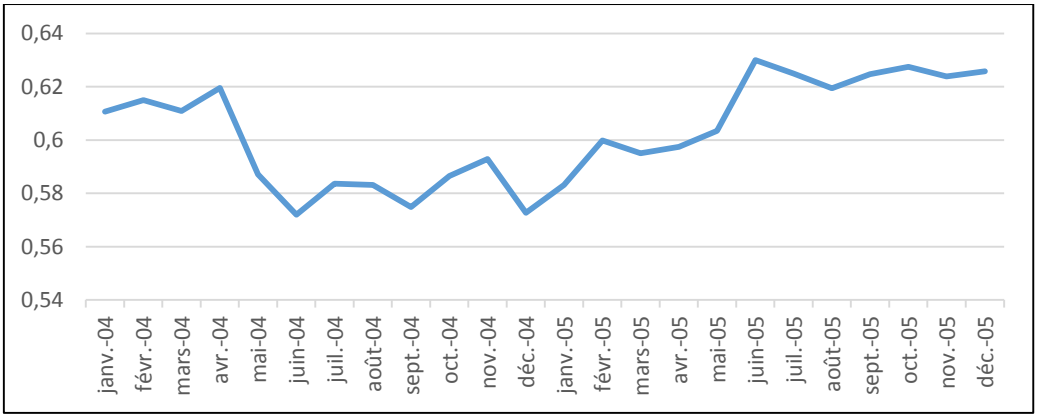
Pour éviter de biaiser les informations disponibles, nous ne cherchons pas à retrouver les valeurs de ces véhicules, en les définissant par exemple comme égales à la valeur moyenne d'un véhicule du même type : nous décidons de les **supprimer** du portefeuille étudié.

Cinquante-trois polices sont concernées, dont six polices présentant au moins un sinistre, ce qui amène notre base à 67 803 contrats.

4.2.2 Redressement des coûts

Les montants indiqués sont exprimés en dollar Australien (AUD).
Nous serons amenés à travailler avec des montants de Prime Pure dans la suite du mémoire, nous ramenons donc les montants présents dans la base à notre environnement actuel, à savoir, la France en 2015, pour faciliter l'analyse des résultats.

On commence par appliquer un changement de devise : les valeurs étant observées sur les années 2004 et 2005, nous prenons la moyenne mensuelle du cours AUD/€ sur cette période comme facteur multiplicatif, soit 0,60.



Graphique 1 - Cours Mensuel AUD / € sur la période 2004/2005 [1]

Notre travail de redressement sur les coûts des sinistres ne s'arrête pas là.

Les coûts des sinistres correspondent aux coûts effectivement réglés par l'assureur lors de l'occurrence du sinistre à la date de survenance de ce dernier.

Cependant, l'étude d'un historique dont la période d'observation est éloignée de la date d'étude pour la construction d'un tarif technique nécessite le redressement des coûts : il faut « projeter » les coûts des sinistres à la date d'étude du portefeuille, autrement dit, répondre à la question : « quel serait le coût d'un sinistre présentant les mêmes caractéristiques mais dans l'environnement présent ? ».

[1] Source : [S11]

Si les coûts des sinistres évoluent, c'est parce que les coûts de la main d'œuvre et du matériel évoluent également, cette évolution étant liée à l'inflation de manière plus générale.
Nous allons donc remettre les coûts en « *AS IF* », c'est-à-dire, retrouver le montant correspondant à un sinistre de nature identique qui se serait passé à l'heure de l'analyse.

Pour ce faire, nous utilisons un indicateur qui tient compte de l'inflation ainsi que de l'évolution du coût des réparations.

Dans ce mémoire, nous choisissons l'indice INSEE concernant les Pièces détachées et accessoires pour véhicules personnels.
Nous fixons également la date d'observation de l'indice en mileu de période d'observation , à savoir janvier 2005, et utilisons l'indice de ce début d'année pour effectuer le redressement.

Date d'observation de l'indice	Valeur de l'indice
Janvier 2005	109,50
Janvier 2015	145,13

Tableau 5 -Valeur de l'indice INSEE Pièces détachées et accessoires pour véhicules personnels[1]

Nous donnons ci-dessous la formule générale à appliquer pour rétablir un montant de sinistre en ***AS IF*** après avoir choisi un indice de redressement.

Définition :

Soient t_1 et t_2 deux dates, correspondant respectivement à la date de l'occurrence du sinistre et la date à laquelle nous souhaitons ramener le coût du sinistre.
Soient I_1 et I_2 , les valeurs de l'indice aux dates respectives t_1 et t_2.

Alors , pour un sinistre de coût C_1 dont la survenance est observée à la date t_1 , le montant C_2 du sinistre correspondant s'il s'était produit à la date t_2 est donné par :

$$\frac{C_1}{I_1} = \frac{C_2}{I_2}$$

d'où

$$\boxed{C_2 = C_1 \times \frac{I_2}{I_1}.}$$

Dans le cas présent , il faudra donc multiplier nos coûts de sinistre par 1,33[2].

Remarque :

Cette méthode est souvent appliquée en provisionnement, lorsque l'on étudie des triangles de déroulement des coûts après l'année de survenance et ne s'applique pas uniquement dans le cadre de l'assurance automobile. Des indicateurs « classiques » par secteur d'assurance y sont même dédiés (par exemple, l'indice de construction en Multirisque Habitation).

Les coûts des sinistres ne sont pas les seules valeurs monétaires dont on dispose : nous avons également la variable « ***Veh_value*** » qui correspond au prix du véhicule et qui est exprimée en AUD.

[1] Source : [S11]
[2] $1{,}33 = \frac{145{,}13}{109{,}50}$

Un facteur multiplicatif tenant compte de l'évolution du prix des véhicules ainsi que du changement de devise doit aussi lui être appliqué.

Nous décidons de conserver le même indice que précédemment pour rendre compte de l'évolution du prix des véhicules.

Au final nous avons :

$$\text{Veh\_value} = \text{Veh\_value} \times 0{,}7988^{1}$$

$$\text{Claimcst0} = \text{Claimcst0} \times 0{,}7988$$

4.2.3 Séparation des sinistres attritionnels et graves

Lors de l'occurrence d'un sinistre couvert par le contrat d'assurance, l'assureur va devoir indemniser son assuré en lui versant une partie ou la totalité du montant du sinistre.
Cependant, certains sinistres peuvent avoir un montant très élevé et l'assureur se doit de se protéger contre ces évènements, certes rares mais ayant un poids important dans sa charge totale de sinistre.

Ces sinistres, n'ayant ni une fréquence ni un coût semblable aux sinistres classiques, doivent être traités de façon différente lors de la tarification.

Pour ce faire, l'assureur doit pouvoir distinguer ce qu'il considère comme étant un sinistre grave. Une analyse des coûts des sinistres est alors nécessaire. Les sinistres n'entrant pas dans la catégorie des sinistres graves sont qualifiés d'attritionnels.

Analyse descriptive des coûts

Notre base de données est stockée dans la variable « *data* » et nous faisons volontairement apparaitre certaines lignes de code pour permettre à nos lecteurs de réutiliser certaines méthodes présentées dans le cadre de leurs travaux personnels ou simplement d'enrichir leur connaissance des outils proposés par le logiciel R.
Pour les 4 618 contrats qui ont présentés au moins un sinistre, nous dénombrons un total de 4 929 sinistres[2].

Nous commençons par calculer les paramètres de position de la variable des coûts unitaires : ils définissent la tendance générale de la distribution.

```
> summary(data$coûtUnitaire)

   Min.    1st Qu.    Median      Mean      3rd Qu.     Max.
  159.8     282.6     666.96    1662.98    1778.0    44670.0
```

Parmi ces six paramètres, notre attention se porte sur la médiane (666,96 €) et la moyenne (1 662,98 €) des coûts.
L'écart entre ces deux valeurs témoigne d'une distribution du coût des sinistres asymétrique.

[1] 0,7988 = 0,6 × 1,33
[2] Correspondants aux 4 624 initiaux auxquels on a soustrait six sinistres causés par des polices ayant une valeur de véhicule nulle.

La médiane est de l'ordre du tiers de la moyenne, ce qui laisse sous-entendre deux possibilités :

- La présence de valeurs extrêmes qui pèsent à la hausse sur la moyenne
- Un nombre important de coûts faibles qui pèsent à la baisse sur la médiane

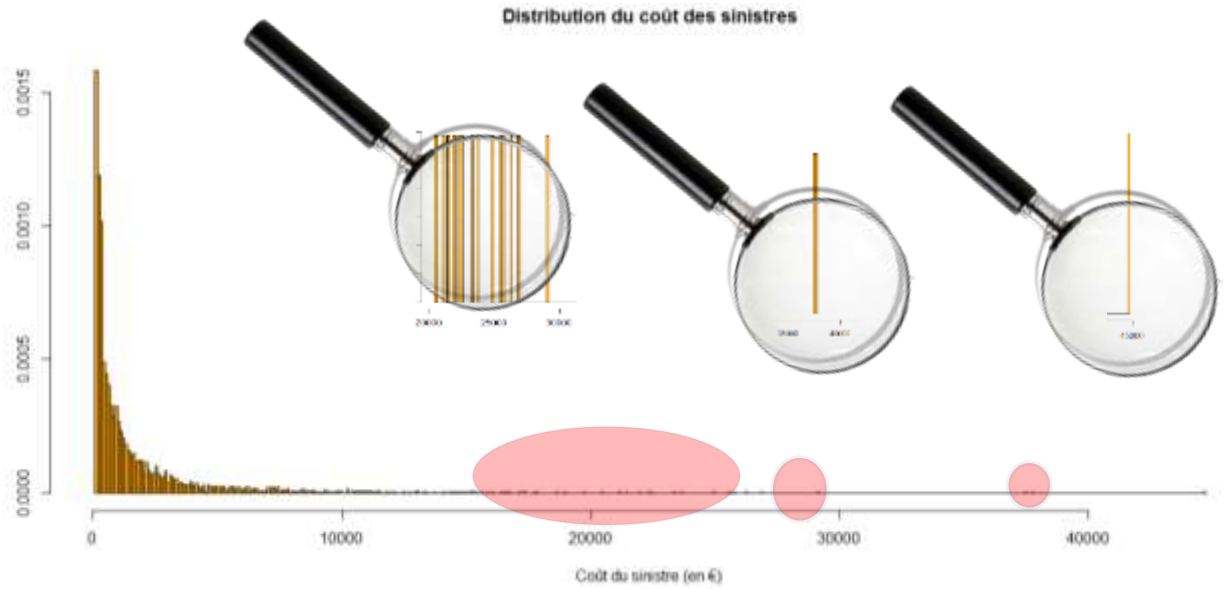
Nous devons entrer plus en détails dans la répartition du coût des sinistres pour pouvoir définir dans quel cas de figure nous nous trouvons.

Détection de valeurs extrêmes et traitement

Détection de valeurs extrêmes

La visualisation de la distribution du coût des sinistres est un premier recours pour déterminer l'origine de cet écart.

```
hist(data$coûtUnitaire,prob=TRUE, main="Densité du coût des sinistres", xlab="Coût du sinistre (en €) ",col="orange" , breaks=500)
```



Graphique 2 - Distribution du coût des sinistres

Dans le cadre de la tarification (et du provisionnement) en assurance Non-Vie, une hypothèse classique est celle selon laquelle le portefeuille est constitué de risques similaires. Un problème pour que cette hypothèse soit vérifiée est le poids important des sinistres graves.

On observe ici la présence de montants élevés (> 20 000 €) : leur poids est important dans la charge totale puisqu'ils influent de façon significative sur la moyenne (cf. Cercles sur le Graphique 2).

Pour résoudre ce problème, les sinistres observés sont souvent écrêtés : ils sont plafonnés à un niveau maximum, appelé seuil des graves, que nous noterons **M**.

Le choix de ce plafond peut être délicat et peut conduire à une sous-estimation ou à une surestimation des sinistres ordinaires. Ce seuil doit être suffisamment grand pour pouvoir utiliser les résultats précédents, mais pas trop afin de disposer d'un nombre suffisant d'observations pour obtenir des estimations de qualité.

Il existe de nombreuses pistes pour orienter le choix du seuil de l'écrêtement[1]. Dans ce mémoire, nous appliquons une technique classique qui consiste à choisir un seuil tel que la sur-crête représente un certain pourcentage de la charge totale de sinistres. **Le seuil est fixé à 5 % de la charge totale de sinistres.**

En rangeant les coûts des sinistres par ordre décroissant et en calculant le pourcentage cumulé de la charge de sinistre totale qu'ils représentent, nous obtenons un seuil de M=**21 400 €**, soit 15 polices concernées (cf. Tableau 6).

Nous distinguons alors :

- Les sinistres attritionnels, sinistres dont le coût est inférieur à 21 400 €
- Les sinistres graves, sinistres dont le coût est supérieur à 21 400 €

Coût du sinistre	Pourcentage cumulé de la charge totale de sinistres
44 668,83 €	0,54 %
37 779,03 €	1,01 %
37 436,82 €	1,46 %
29 156,75 €	1,82 %
26 872,38 €	2,15 %
26 211,42 €	2,47 %
25 717,00 €	2,78 %
25 540,43 €	3,09 %
24 956,45 €	3,40 %
23 671,20 €	3,68 %
23 383,94 €	3,97 %
22 404,23 €	4,24 %
22 375,76 €	4,52 %
21 904,38 €	4,78 %
21 469,37 €	**5,05 %**
21 173,19 €	5,30 %

Tableau 6 - Pourcentage cumulé de la charge de sinistre par coût décroissant

Traitement des graves

Dans le cadre de la modélisation de la charge de sinistre de notre portefeuille d'assurés automobile, nous utilisons un modèle collectif (cf. § 1.2.2) du type :

$$S = \sum_{i=1}^{N} C_i$$

Avec :

- S la charge totale de sinistre,
- N le nombre (fréquence) de sinistres
- C_i le coût du $i^{ème}$ sinistre

[1] Le lecteur intéressé pourra trouver de plus amples détails en [O6].

Au vu de la présence de sinistres graves, nous allons à présent adapter le modèle fréquence-coût à la séparation de ces derniers des sinistres attritionnels en définissant la charge de sinistre S comme suit :

$$S = \sum_{i=1}^{N} C_i + N_G \times G$$

Avec cette fois-ci :

- N le nombre de sinistre **attritionnels**
- C_i le coût du $i^{ème}$ sinistre **attritionnel**
- N_G le nombre de sinistres **graves**
- G le coût moyen d'un sinistre **grave**

N_G *et* G étant connus, le problème reste identique, à savoir, modéliser S pour déterminer son espérance via l'étude des variables fréquence (N) et coût (C) de sinistre.

Nous disposons d'un nombre de polices dites « graves » trop faible (15) pour permettre l'adéquation d'une loi à ces données. Nous décidons donc de répartir la charge de sinistres graves sur l'ensemble du portefeuille.

Chaque police est alors affectée d'une Prime Pure du type :

$$PP = PP_{attri} + PP_{grave}$$

*o*ù

- PP_{attri} correspond au montant de Prime Pure servant à couvrir les sinistres attritionnels
- $PP_{grave} = \frac{N_G \times G}{n}$, correspond au montant de Prime Pure servant à couvrir les sinistres graves (**n** étant le nombre de polices totales du portefeuille)

Nous pouvons d'ores et déjà calculer le montant de la Prime Pure des graves :

$$PP_{grave} = \frac{15 \times 27\,569{,}87}{67\,788} = 6{,}10\ €$$

Dans la suite du mémoire, l'ensemble des résultats obtenus concernera uniquement la base des attritionnels puisque le traitement des graves a déjà été effectué. Ces polices ne seront donc plus comptabilisées ni dans les coûts ni dans la fréquence de sinistre.

L'un des deux cas évoqués précédemment vient d'être soulevé, à savoir, que l'écart entre la médiane et la moyenne provenait de la présence de sinistres graves dans le portefeuille.
Renouvelons à présent l'analyse descriptive du coût sur les sinistres jugés comme **attritionnels** afin de pouvoir observer l'impact de la suppression des graves.

Nous obtenons alors les résultats suivants :

```
> summary(data$coûtUnitaire)

   Min.   1st Qu.  Median    Mean     3rd Qu.    Max.
  159.8   282.6    664.0     1584.0   1763.0    21173.19
```

L'écart entre la médiane et la moyenne a certes diminué, mais il reste extrêmement important. Intéressons-nous alors à la deuxième source possible d'écart : le poids important des valeurs faibles.

Présence de franchise

Sur le graphique de distribution des coûts (cf. Graphique 2), nous observons également un pic pour les valeurs faibles. Pour affiner l'analyse, nous étudions les montants de sinistre présentant plus de 10 occurrences.

Nous remarquons une redondance importante du montant 200 AUD, soit environ 160 euros : 713 sinistres sur 4 914 sinistres étudiés, soit plus de 14 % des sinistres, ont pour montant 159,75 €.
C'est aussi le coût minimal des sinistres répertoriés.

Coût du sinistre (en AUD)	Coût du sinistre	Nombre d'occurrences
200,00	159,75 €	713
345,00	275,58 €	40
353,77	282,58 €	219
353,80	282,60 €	40
367,73	293,73 €	42
369,15	294,87 €	29
389,95	311,48 €	94
390,00	311,52 €	55
408,95	326,66 €	15

Tableau 7– Les montants de sinistres présents plus de 10 fois

Nous pouvons donc affirmer avoir détecté la **présence d'une franchise atteinte** (*deductible*) d'une valeur de 200 AUD[1] : si le coût engendré par le sinistre survenu est inférieur au montant de la franchise, le coût reste à la charge de l'assuré. Le cas échéant, l'assureur prend à la charge la totalité du coût du sinistre.

Ce type de franchise permet à l'assureur de réaliser des économies sur la gestion : il n'a plus à traiter les petits sinistres. Par contre, cette formule peut avoir, dans certains cas, pour conséquence une aggravation du risque moral. En effet, l'assuré a certes intérêt à ce qu'il n'y ait pas de sinistre, mais, dès lors que le sinistre est survenu, il n'a plus nécessairement avantage à ce que son coût reste faible. Bien au contraire, il a souvent intérêt à ce que le montant soit suffisamment élevé pour atteindre le niveau de la franchise.

Nous relevons un total de 740 valeurs entre 200 et 205 AUD. Une partie d'entre elles pourraient bien traduire cet aléa moral : en en cas d'accident, de nombreux assurés gonflent leur facture pour arriver au seuil de la franchise, quitte à le dépasser quelque peu pour ne pas éveiller les soupçons.

Seuls les sinistres responsables et dont le montant dépasse la franchise ont été relevés. Nous n'avons donc aucune information concernant les sinistres non responsables et ceux d'un montant insuffisant pour être déclarés, que cela concerne leur fréquence ou leur coût.

[1] Nous revenons dans ce paragraphe aux montants exprimés en dollars australien avant leur redressement en « *As if* » car le contrat a été établi avec une franchise établie dans cette devise. Ainsi le montant de franchise relevé est bien entier.

Nous tarifons donc un contrat qui présente une franchise d de 200 AUD (soit 159,75 € à l'heure de l'étude) et cherchons à estimer au mieux :

$$PP_{attri} = \mathrm{E}[\mathrm{S}|\mathrm{C} \geq \mathrm{d}\,] = \mathrm{E}[\mathrm{N}|\mathrm{C} \geq \mathrm{d}\,] \times \mathrm{E}[\mathrm{C}|\mathrm{C} \geq \mathrm{d}\,]$$

4.3 Traitement des variables : classification

Rappelons que seules les polices d'assurance qui appartiennent à la même classe pour chacun des facteurs utilisé dans la tarification se verront affectées du même montant de prime.
Dans le but d'une démarche marketing mais aussi actuarielle, il est préférable de regrouper certaines modalités présentant des caractéristiques communes en classes :

- D'un point de vue marketing, il est bon de ne pas se retrouver avec un nombre trop important de tarifs différents ;
- D'un point de vue actuariel, il est vrai qu'il serait beaucoup plus fiable de disposer d'un codier le plus précis possible afin d'assurer une bonne segmentation, mais :

  – la multiplicité des modalités des variables laisserait trop de zones sans observations dans les estimations ;
  – dans le cadre de l'utilisation de certains modèles, notamment les GLM, les variables de type quantitatives continues doivent être transformées en classes de valeurs pour que l'information soit exploitable (cf. § 3.2.6).

La création de classes de modalités présente également d'autres avantages :

- Elle joue un effet de lissage lors de la recherche d'une relation linéaire entre les variables et le tarif.
- Elle permet de réduire le risque de corrélation entre les variables. En effet, ce sont les modalités de chacune des variables qui sont corrélées une à une, ou une à plusieurs. Le fait de regrouper ces modalités affaiblira donc ces corrélations.

Deux méthodes de classifications sont utilisées dans ce mémoire afin de répondre aux deux problématiques soulevées pour le point de vue actuariel :

- La méthode des K-means, qui entre dans la classe des méthodes de partitionnement.
- La Classification Hiérarchique Ascendante (CAH), faisant partie des méthodes de classification hiérarchique.

Le choix de la méthode de classification à utiliser est fonction du type de variable et de l'objet de la classification.

4.3.1 Rappel du type des variables disponibles

Variable	Type	Nature	Nombre de modalités
Veh_value	quantitative	continu	-
Veh_body	qualitative	nominale	13
Agecat	qualitative	ordinale	6
Area	qualitative	nominale	6
Veh_age	qualitative	ordinale	4
Gender	qualitative	binaire	2

Tableau 8 - Type et nature des variables de la base « Car »

Le tableau précédent nous permet d'identifier les variables nécessitant une classification :

- **La valeur du véhicule** (*veh_value*) : c'est une variable quantitative continue qui ne peut être utilisée comme telle dans le GLM.
- **Le type de véhicule** (*veh_body*) : cette variable qualitative présente un nombre de modalités trop important (13).

Remarque :

Ce travail de classification a déjà été fait pour certaines variables présentes dans la base. C'est le cas notamment de la variable ***Veh_age*** ou ***Agecat***. Nous n'affirmons pas que les méthodes utilisées ont étés celles que nous allons présenter. Bien au contraire, ce sont certainement des méthodes d'analyse de correspondance (ACP, ACM) qui ont été appliquées. Ces dernières ne seront pas évoquées dans ce mémoire et nous invitons le lecteur curieux à se référer aux nombreux ouvrages disponibles à ce sujet.

4.3.2 Clustering par K-means : regroupement d'observations autour de centres mobiles

Le data clustering est une méthode statistique qui permet de diviser un ensemble de données en groupes homogènes. Cette homogénéité se traduit par des critères de proximité définis en introduisant des mesures et des classes de distances entre les observations.
Cette méthode peut être utilisée pour classifier une variable quantitative continue. Il existe de nombreux algorithmes aboutissants à la partition des données. Nous allons nous concentrer ici sur l'algorithme des K-means.

Son principe consiste à partir d'une partition aléatoire d'un ensemble E en K classes puis de reformer pas à pas K nouvelles classes qui contiennent des individus de plus en plus proches.

Modélisation mathématique[1]

Considérons un ensemble fini E de n individus représentés par p caractéristiques quantitatives. Lorsque p est supérieur à 2, la représentation des individus ne peut plus se faire sur un simple repère mais sur un **nuage de** n **points**.
Les variables n'étant pas forcement exprimées dans la même unité, elles doivent être centrées et réduites afin que chacune d'entre elles ait un impact visible.

[1] Source : [O1]

Ainsi la variable x_i^j est transformée en $\frac{x_i^j - \bar{x}_j}{\sigma_j}$ avec σ_j l'écart-type de la variable considérée et $\bar{x}_j$ sa moyenne.

L'objectif est de regrouper les n individus en K classes homogènes. Ces individus seront représentés par n vecteurs $x_1, \dots, x_n$ de $\mathbb{R}^p$.

Commençons par quelques définitions.

Définition :
La **distance** sur un ensemble E est une application d définie sur le produit $E^2 = E \times E$ à valeurs dans l'ensemble $\mathbb{R}^+$ des réels positifs, qui vérifie les trois propriétés suivantes :

- Symétrie : $\forall\ (x, y) \in E^2, d(x, y) = d(y, x)$
- Séparation: $\forall\ (x, y) \in E^2, d(x, y) = 0 \Leftrightarrow x = y$
- Inégalité triangulaire : $\forall\ (x, y, z) \in E^3, d(x, z) \leq d(x, y) + d(y, z)$

Il existe notamment :

- La distance Euclidienne :

$$d(x, y) = \sqrt{\sum_{j=1}^{p} (x^j - y^j)^2}$$

- La distance de Manhattan :

$$d(x, y) = \sum_{j=1}^{p} |x^j - y^j|$$

Définition :
L'**inertie d'un nuage de n points** est la moyenne des carrés des distances de ces points à leur centre de gravité noté **g**.

Si $d(x_i, g)$ est la distance du point i au centre de gravité, alors l'inertie totale vaut :

$$I_T = \frac{1}{n} \sum_{i=1}^{n} d^2\,(x_i, g) \text{ où } g = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Pour un ensemble de points donné cette inertie est constante.

Définition :
Un **cluster** (une classe) k est un ensemble non vide d'observations.
Notons $\pi_k = \{\ i\ |x_i \in\ cluster\ k\ \}$.

Définition :
Le **centre de gravité** du cluster k est $g_k = \frac{1}{n_k} \sum_{i \in \pi_k} x_i$ où $n_k = Card\ \pi_k$

Définition :
Supposons que nous ayons K classes, la $k^{ième}$ classe π_k de centre de gravité g_k est formée de n_k $(k = 1, \dots, K)$ observations, son **inertie** vaut :

$$I_k = \frac{1}{n_k} \sum_{i \in \pi_k} d^2(x_i, g_k)$$

Elle est d'autant plus homogène que ses éléments sont proches de son centre de gravité : son inertie est alors faible.

Une mesure globale de l'homogénéité des classes appelée inertie intra-classes (d'où le symbole W comme *Within*) est donc la moyenne des inerties des K classes, chacune pondérée par son importante relative :

$$I_W = \sum_{k=1}^{K} \frac{n_k}{n} I_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \pi_k} d^2(x_i, g_k)$$

L'inertie du nuage des K centres de gravité, appelée **inertie inter-classes** (d'où le symbole B comme *Between*), est :

$$I_B = \sum_{k=1}^{K} \frac{n_k}{n} d^2(g_k, g)$$

Le théorème de Huygens nous permet d'écrire :

$$\boxed{I_T = I_W + I_B}$$

Le critère de classification choisit est alors celui qui rend **l'inertie intra-classes minimale** ou **l'inertie inter-classes maximale**.
Le problème se résout par itérations successives.

Algorithme de classification

A chaque étape de l'algorithme, chaque individu est associé à un numéro de cluster entre 1 et K. Soit $L_t^{(i)}$le numéro du cluster de l'individu x_i à l'issue de l'étape t et g_k^t le centre de gravité de la $k^{ième}$ classe à l'étape t.

Etape 1 : Initialisation

K centres de gravité $(g_k^0)_{1 \leq k \leq K}$ sont tirés aléatoirement.
Puis chaque individu est affecté au cluster dont le centre de gravité est le plus proche.
Autrement dit, chaque individu x_i se voit attribuer un numéro aléatoire entre 1 et K selon la règle suivante :

$$L_0^{(i)} = n |\ d(x_i, g_n^0) = min_{k \in [\![1;K]\!]} \{d(x_i, g_k^0)\}$$

A l'issue de cette première étape, K groupes sont formés aléatoirement. Ils contiennent chacun des individus qui ne se ressemblent pas. Nous allons chercher, un peu comme avec des aimants, à attirer ceux qui se ressemblent.

Etape 2

Les barycentres g_1^t,…,g_K^t des clusters obtenus à l'étape précédente sont déterminés.
Chaque observation est affectée à son barycentre le plus proche. De nouveaux clusters se forment :

$$L_t^{(i)} = n |\ d(x_i, g_n^t) = min_{k \in [\![1;K]\!]} \{d(x_i, g_k^t)\}.$$

Etape 3

L'étape 2 est renouvelée jusqu'à l'obtention d'un critère de classification stable.

Remarque :

Le processus de classification est simple mais la classification finale dépend de la partition aléatoire initiale. Il convient alors de répéter un nombre suffisant de fois l'algorithme pour s'assurer de la robustesse de la classification choisie.

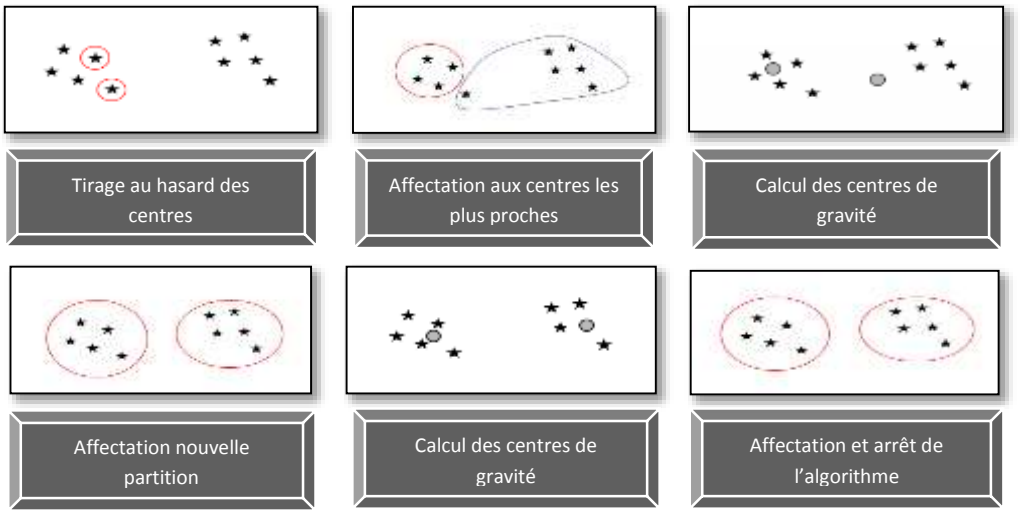


Figure 5 - Exemple de déroulement de l'algorithme dans le cas K=2

Le choix du nombre de classes

Afin de perdre le moins d'information, le nombre de classes K est déterminé de sorte que le rapport variance inter-classes sur variance totale soit supérieur à 95 %.
Néanmoins**,** en pratique, il arrive souvent que les classes obtenues ne soient pas utilisables telles quelle. Ceci peut être dû par exemple à des classes de taille trop hétérogènes. Il est donc nécessaire de procéder à des modifications spécifiques à la problématique étudiée pour les rendre exploitables.

Application : Partitionnement de la valeur des véhicules

La variable ***veh_value*** est une variable quantitative continue. Pour l'introduire en tant que variable explicative de la sinistralité en automobile dans nos modèles linéaire généralisés, il faut la transformer en variable qualitative par le biais de la création de classes de valeurs de véhicule (cf. § 3.2.6).

Pour ce faire, la méthode des K-means est utilisée en se plaçant dans le cas où p=1 puisque seule la valeur du véhicule caractérise ici l'individu.
La distance euclidienne (cf. § 4.3.2) est retenue pour mesurer l'écart entre les individus, qui n'est autre que l'écart entre la valeur de leurs véhicules.

Observons nos données :

```
> summary(data$veh_value)
```

```
   veh_value
Min.    : 0.1438
1st Qu.: 0.8068
Median : 1.1982
Mean    : 1.4204
3rd Qu.: 1.7174
Max.    :27.6054
```

Illustration 2 – Résultat de la fonction summary()

Nous disposons de n = 67 788 individus pour lesquels le coût des véhicules prend ses valeurs entre 1 438 € et 276 054 €. Pour choisir le nombre de classes en lequel va être partitionné cet intervalle, il faut :

$$\frac{\text{variance inter} - \text{classes}}{\text{variance totale}} > 95\ \%$$

Huit classes sont retenues pour satisfaire ce critère (cf. Tableau 9).

Nombre de classes	$\frac{\text{variance inter} - \text{classes}}{\text{variance totale}}$
6	91,8 %
7	93,9 %
8	95,1 %

Tableau 9 - Rapport variance inter-classes sur variance totale en fonction du nombre de classes

Cherchons alors la délimitation des classes de valeurs de véhicule à l'aide de la fonction ***kmeans()*** sous R. Cette fonction nous permet de spécifier via l'élément **nstart** le nombre de fois que l'algorithme doit tourner afin d'assurer sa robustesse (cf. Remarque précédente). Dans notre cas, il est relancé vingt-cinq fois.

```
x=matrix(c(sort(data$veh_value)), ncol=1)
cl <- kmeans(x, 8, nstart = 25)
plot(x,main="Partitionnement du coût des véhicules par K-Means", xlab=
"Indexation des contrats par valeur croissante de véhicule", ylab="Coût du
véhicule (en K€)", col = cl$cluster)
```

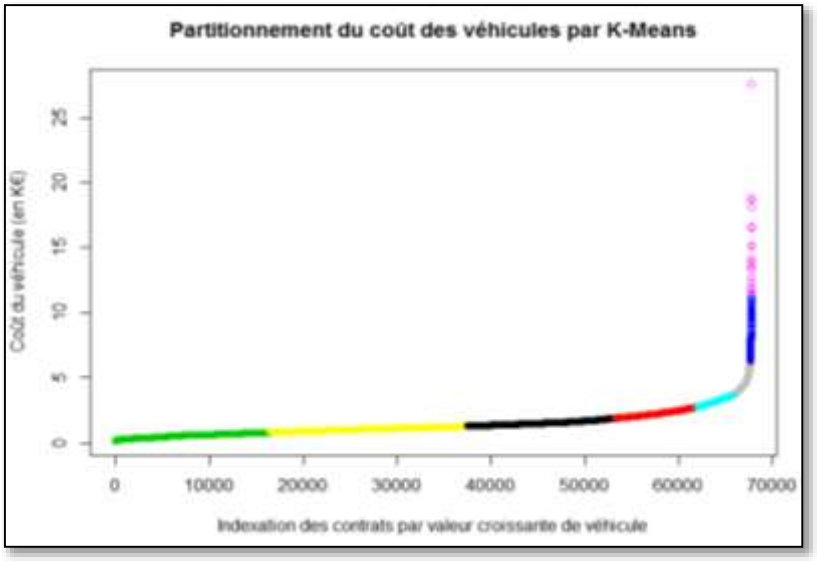


Illustration 3 – Découpage en classes de coûts de la variable *veh_value*

Huit classes de valeurs du véhicule sont retenues. Leurs limites ainsi que leur contenu sont détaillés dans le tableau ci-dessous :

Numéro de cluster	Borne inf (en K€)	Borne sup (en K€)	Effectif	Part du portefeuille
1	0,14	0,78	16 076	23,72 %
2	0,79	1,27	21 107	31,14 %
3	1,28	1,87	15 913	23,47 %
4	1,88	2,73	8 775	12,94 %
5	2,74	3,91	4 399	6,49 %
6	3,92	6,27	1 356	2,00 %
7	6,30	11,65	143	0,21 %
8	11,82	27,61	19	0,03 %
Total	-	-	67 788	100 %

Tableau 10 - Classes de coûts de véhicule

Le nombre de classes constituées est important et certaines classes (6, 7 et 8) ont un effectif qui représente moins de 5 % du portefeuille.
Les classes 5 à 8 sont regroupées afin de garantir un niveau d'effectif suffisant par classe.

Finalement, nous retenons les cinq classes de coûts suivantes :

Classe de coût du véhicule	Borne inf (en K€)	Borne sup (en K€)	Effectif	Part du portefeuille
VV1	0,14	0,78	16 076	23,72 %
VV2	0,79	1,27	21 107	31,14 %
VV3	1,28	1,87	15 913	23,47 %
VV4	1,88	2,73	8 775	12,94 %
VV5	2,74	27,61	5 917	8,73 %
Total	-	-	67 788	100,00 %

Tableau 11 - Les cinq classes finales de coûts de véhicule

4.3.3 Classification Hiérarchique Ascendante

Notre population est décrite par des variables quantitatives. Cette population doit être partitionnée en plusieurs sous-populations relativement homogènes.
Afin de répondre à cette problématique, une première façon de faire est de partir des individus un à un et de les regrouper par similarité. Une fois les premiers regroupements effectués, l'opération est renouvelée de manière à constituer des classes de classes, jusqu'à obtenir à l'extremum une unique classe regroupant l'ensemble des individus.

La proximité entre les individus mais aussi entre les classes formées est mesurée à l'aide d'une distance. Le fait de regrouper de cette façon en remontant jusqu'à un unique groupe s'appelle la **Classification hiérarchique ascendante (CAH)**.

Par exemple, une planète appartient à un système solaire, qui appartient à un amas d'étoiles, qui appartient à une galaxie, puis à un amas de galaxies, jusqu'à l'univers tout entier.
A chaque étape, on perd de l'information mais on sait que les éléments regroupés au sein d'une même classe sont relativement proches entre eux.

Algorithme de CAH

1ère étape : Préparation des données

Afin de pouvoir mesurer une distance entre individus, nous allons devoir « quantifier » les données qualitatives et mettre sur un même pied d'égalité toutes les variables considérées :

- Les variables qualitatives doivent être transformées en variables quantitatives. Une technique consiste à prendre toute les modalités existantes et à mettre 1 ou 0 selon la présence ou l'absence de cette modalité.
- Pour rééquilibrer les rôles des variables, l'usage est d'opérer leur centrage et leur réduction, en retirant la moyenne de la variable considérée et en divisant par l'écart-type de la variable considérée.

2ème étape : Application de l'algorithme

Initialisation :
Les classes initiales sont les n singletons-individus. Il faut alors calculer la matrice de leurs distances deux à deux.

Regroupement :
Les deux éléments les plus proches au sens de la distance entre groupes choisie sont regroupés. Il faut également mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes.

Itération :
Réitérer les deux étapes précédentes jusqu'à l'agrégation en une seule classe.

Choix d'une distance

La CAH nécessite donc de définir une distance entre groupes d'individus (appelée stratégie d'agrégation).

La méthode de Ward est la plus courante. Elle consiste à réunir les deux classes dont le regroupement fera le moins baisser l'inertie inter-classes (cf. § 4.3.2).

La distance de Ward $\Delta\,(A,B)$ qui représente la distance entre deux classes A et B est alors définie comme étant la distance entre leurs centres de gravité au carré, pondérée par les effectifs des deux classes[1].

$$\Delta\,(A,B) = \frac{\pi_A \times \pi_B}{\pi_A + \pi_B}\, d^2(g_A, g_B)$$

Avec

- π_A (respectivement π_B) le nombre d'individus dans la classe A (B)
- g_A (respectivement g_B) le centre de gravité de la classe A (B)
- d, la distance euclidienne (cf. § 4.3.2)

Cette méthode tend à regrouper les petites classes entre elles.

[1] On suppose tout de même l'existence de distances euclidiennes entre observations.

Remarque :

A noter que lorsque deux classes sont remplacées par leur réunion, la diminution de l'inertie inter-classes, et donc l'augmentation de l'inertie intra-classes, est égale à la distance de Ward. La méthode de Ward consiste alors à choisir à chaque étape le regroupement de classes tel que l'augmentation de l'inertie intra-classes soit minimale.

Représentation : un arbre de classification

Il est facile d'appréhender la notion de hiérarchie à travers une représentation visuelle appelée **arbre de classification** ou **dendrogramme.**

Le dendrogramme est une représentation graphique sous forme d'arbre binaire, d'agrégations successives jusqu'à réunion en une seule classe de tous les individus. La hauteur d'une branche est proportionnelle à la distance entre les deux objets regroupés.

Dans le cas du saut de Ward, la hauteur de la branche est proportionnelle à la perte d'inertie inter-classes.

Choix du nombre de classes

En ce qui concerne le choix du nombre de classes à retenir, la partition doit être définie en accord avec nos objectifs.
A chaque agrégation, la distance entre les groupes est réduite. Celle-ci est observable sur l'arbre de classification finalement obtenu. Généralement, la meilleure partition est celle qui précède une valeur de la distance inter-classes brutalement plus faible.

Application : regroupement des modalités pour la variable type de véhicule

Une tarification exhaustive prenant en compte un nombre trop important de modalités pour une même variable peut être complexe et peu réaliste.

Parmi les variables disponibles, l'une d'entre elles se distingue de par son nombre plus important de modalités : ***veh_body***, correspondant au type de véhicule, avec 13 modalités.
L'objectif est de regrouper par ensemble les modalités présentant des similarités au sein d'une même classe afin de n'étudier qu'un ensemble restreint de classes de modalités.
Les similarités étudiées sont fonction de la variable à expliquer. Il peut s'agir par exemple de similarité en terme de fréquence de réalisation du risque ou encore de coût moyen du risque.
Ainsi, selon la problématique envisagée, nous obtiendrons différents regroupements de modalités.

Les possibilités de regroupement sont nombreuses et quasi-uniques par assureur. Dans ce mémoire, la méthode de **Classification Hiérarchique Ascendante** est utilisée pour effectuer ces derniers**.**
Cette dernière permet initialement de regrouper des individus. Ceci n'est pas l'objet de notre recherche puisque nous voulons regrouper des modalités d'une variable.

Dans le même ordre d'idée, nous allons considérer les modalités comme des individus et donc leur attribuer des variables explicatives, à savoir :

- les fréquences moyennes de sinistre par type de véhicule dans le cas où la variable à expliquer est la fréquence des sinistres ;
- les coûts moyens de sinistre par type de véhicule dans le cas où la variable à expliquer est le coût moyen des sinistres.

Nous avons donc 13 individus, correspondant aux 13 modalités de ***veh_body*** et une variable explicative (p=1) par problématique de regroupement.

Afin d'éviter de se retrouver avec un nombre de classes différent pour la variable *veh_body* selon la variable à expliquer, le nombre de classes constituées dans les deux cas est identique même si leur contenu diffère.

Classes de fréquence pour la variable veh_body :

Considérons nos 13 individus (modalités) associés à leur fréquence moyenne de sinistres $(fréqM_i)_{1\leq i\leq 13}$.
Cette dernière ayant été obtenue par :

$$fréqM_i = \frac{1}{n_i}\sum_{k=1}^{n_i} fréquence_k^i$$

Avec

- n_i le nombre d'individus qui possèdent la modalité i pour le type de véhicule
- $fréquence_k^i$ la fréquence de sinistre du $k^{ième}$ individu qui possède la modalité i pour le type de véhicule

Veh_body	Fréquence moyenne de sinistre
BUS	38,63 %
CONVT	6,31 %
COUPE	26,89 %
HBACK	22,77 %
HDTOP	18,16 %
MCARA	28,83 %
MIBUS	13,02 %
PANVN	16,15 %
RDSTR	15,70 %
SEDAN	23,75 %
STNWG	17,82 %
TRUCK	26,30 %
UTE	15,69 %
Total général	**21,28 %**

Tableau 12- Fréquence moyenne de sinistre par type de véhicule

Nous utilisons la méthode de Wald et une distance euclidienne pour l'obtention du dendrogramme.

```
b=tapply(data$fréquence,data$veh_body,mean)
d=dist(b, method="euclidean")
hc=hclust(d, method="ward")
plot(hc, hang=-1)
```

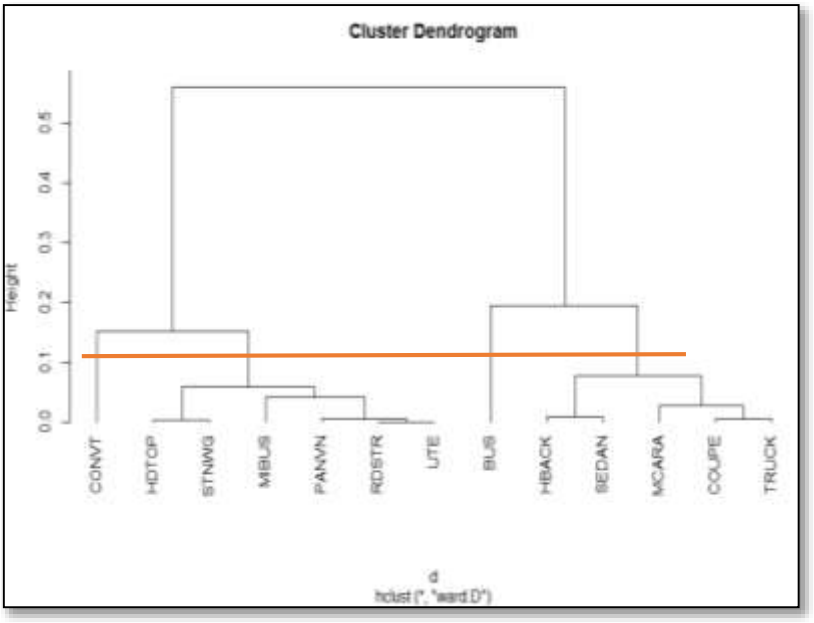


Illustration 4 – Dendrogramme présentant le découpage des types de véhicule

Une forte perte d'invertie inter-classes est observée en passant de quatre à trois classes. Les modalités sont donc fusionnées de manière à obtenir **quatre classes** pour la variable type de véhicule. Ces différentes classes sont délimitées en rouge sur le dendrogramme suivant.
Soit **VB_f** la variable représentative des classes de type de véhicule pour l'étude de la fréquence.

```
rect.hclust(hc, k=4, border="red")
```

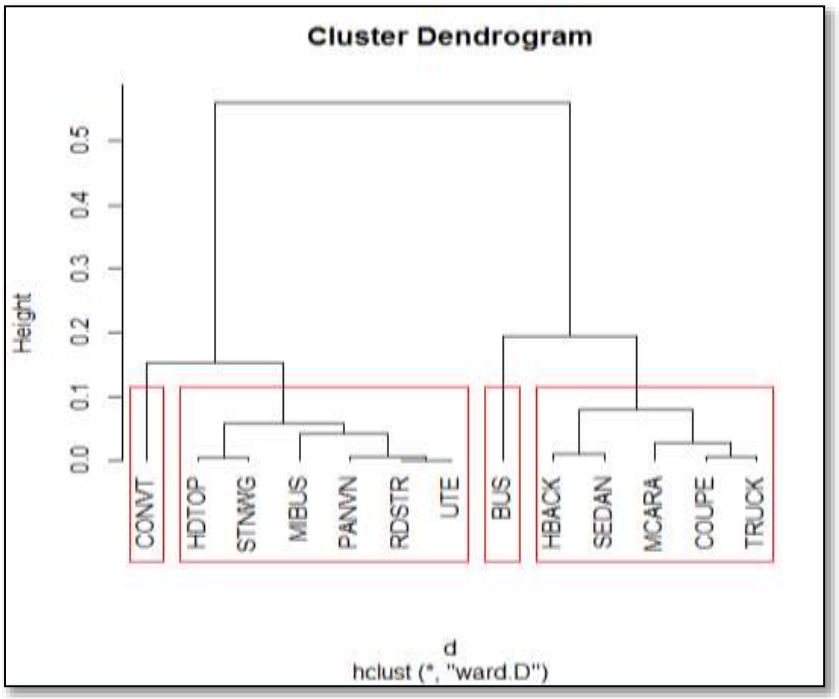


Illustration 5 - Découpage en 4 classes des types de véhicule (Critère de regroupement : fréquence)

Cette variable admet quatre modalités : VB_f1, VB_f2, VB_f3 et VB_f4 (de gauche à droite sur le dendrogramme).

Remarque :
Nous nous sommes intéressés à la composition des classes en termes d'effectif pour savoir si l'ensemble des polices du portefeuille était réparti de façon homogène parmi les quatre nouvelles classes formées ou non.

VB_f	Part du portefeuille (en % de l'exposition)
VB_f1	0,10 %
VB_f2	35,39 %
VB_f3	64,45 %
VB_f4	0,06 %
Total	**100,00 %**

Tableau 13 - Part du portefeuille par classe de véhicule (Critère de regroupement : fréquence)

Les effectifs de nos classes sont certes hétérogènes et parfois même insuffisants car inférieurs à 5 %. Cependant, il n'est pas question ici de regrouper les classes obtenues. En effet, celles-ci traduisent le réel écart de fréquence observée entre les différents types de véhicule.

Classes de coût pour veh_body

Regroupons à présent les modalités de la variable « ***veh_body*** » en considérant cette fois ci leurs similitudes par rapport au coût des sinistres.

Considérons nos 13 modalités associées à leur coût moyen de sinistre.

Veh_body	Charge cumulée de sinistres (a)	Nombre de sinistres (b)	Coût moyen par sinistre (a)/(b)
BUS	9 456,83 €	8	1 182,10 €
CONVT	5 502,56 €	3	1 834,19 €
COUPE	149 947,39 €	75	1 999,30 €
HBACK	2 005 384,55 €	1328	1 510,08 €
HDTOP	183 734,53 €	134	1 371,15 €
MCARA	8 526,01 €	15	568,40 €
MIBUS	92 740,90 €	45	2 060,91 €
PANVN	106 326,78 €	68	1 563,63 €
RDSTR	1 093,88 €	3	364,63 €
SEDAN	2 072 535,79 €	1595	1 299,40 €
STNWG	1 676 626,02 €	1236	1 356,49 €
TRUCK	245 617,56 €	129	1 904,01 €
UTE	454 655,86 €	275	1 653,29 €

Tableau 14 - Coût moyen des sinistres par type de véhicule

Ce dernier ayant été obtenu par :

$$Coût\ moyen\ _i = \frac{charge\ cumulée\ de\ sinistres_i}{n_i}$$

Avec

- n_i le nombre de sinistres impliquant un véhicule de type i
- $charge\ cumulée\ de\ sinistres_i$, la somme des coûts unitaires des sinistres impliquant un véhicule de type i

Nous utilisons là encore la méthode de Wald et une distance euclidienne pour l'obtention du dendrogramme. Un découpage en quatre classes est retenu pour avoir le même nombre de classes que lors du regroupement des modalités selon la fréquence de sinistres. Ces différentes classes sont délimitées en rouge sur le dendrogramme suivant (cf. Illustration 6).

Soit ***VB_c*** la variable représentative des classes de type de véhicule pour l'étude du coût. Cette variable admet quatre modalités : VB_c1, VB_c2, VB_c3 et VB_c4 (de gauche à droite sur le dendrogramme).

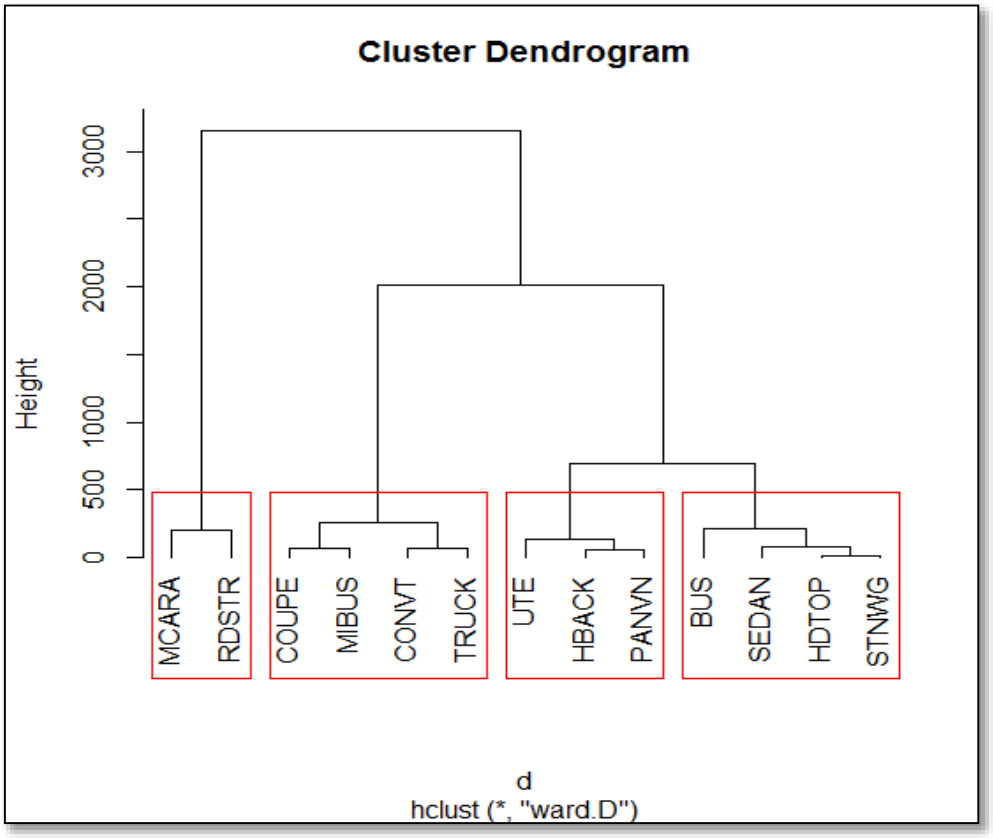


Illustration 6 - Découpage en 4 classes du type de véhicule (Critère de regroupement : coût)

Les classes de type de véhicule formées sont différentes de celles obtenues lors du regroupement précèdent. Ceci témoigne du fait que le type de véhicule influe de manière différente sur le coût et la fréquence de sinistres.

Remarque :

Nous nous sommes intéressés à la composition des classes en termes d'effectif pour savoir si l'ensemble des polices du portefeuille était réparti de façon homogène parmi les quatre nouvelles classes formées ou non.

VB_c	Part du portefeuille (en % de l'exposition)
VB_c1	0,21 %
VB_c2	59,38 %
VB_c3	35,65 %
VB_c4	4,75 %
Total	**100,00 %**

Tableau 15 - Part du portefeuille des classes de véhicule (Critère de regroupement : coût)

Les effectifs de nos classes certes hétérogènes et parfois même insuffisants car inférieurs à 5 %. Cependant, il n'est pas question ici de regrouper les classes obtenues. En effet, celles-ci traduisent le réel écart de coût observé entre les différents types de véhicule.

Chapitre 5 - Fréquence, coût et variables explicatives

L'étude de la composition du portefeuille (§ 5.1) est une étape cruciale à réaliser en amont de la mise en place d'un modèle de tarification.

L'un des intérêts de cette étude est de permettre de faire état du portefeuille actuel dans le but de la comparer avec la population cible. L'assureur peut ainsi adapter la tarification de façon à conserver la visée marketing de l'offre. Par exemple, dans le cas d'un produit d'assurance à destination des femmes, il faut veiller à ne pas sur-tarifer la gente féminine.

Une fois la composition du portefeuille connue, il faut également s'intéresser aux liens existants entre les différentes variables recueillies, que ce soit entre elles, mais également avec la sinistralité (§ 5.2).

5.1 Composition du portefeuille

5.1.1 Analyse globale

Suite aux traitements effectués (cf. § 4.2), nous disposons à présent d'un portefeuille de 67 788 polices. Pour chaque variable qualitative, nous étudions la proportion de contrat présentant les différentes modalités possibles donnée par :

$$Pourcentage_{modalité\ i} = \frac{\sum_{k=1}^{n_i} exposition_k^i}{\sum_{i=1}^{q}\sum_{k=1}^{n_i} exposition_k^i}$$

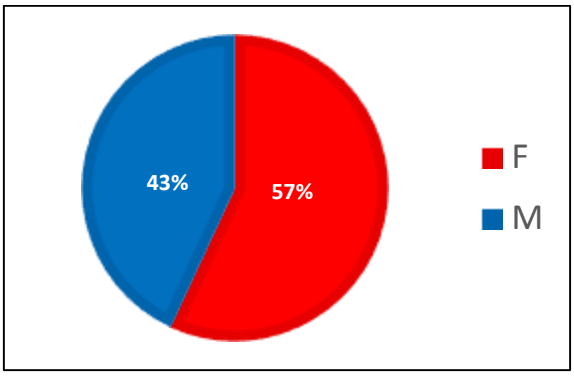
Avec

- n_i le nombre de polices admettant la modalité i pour la variable considérée
- q le nombre de modalités de la variable considérée
- $exposition_k^i$, la valeur de la variable « ***exposure*** » de la $k^{ième}$ police admettant la modalité i pour la variable considérée

Seule la variable quantitative ***veh_value*** est traitée différemment au vu du nombre de valeurs présentes trop important. Nous afficherons alors sa répartition.

Les modalités des variables **VV** (cf. Tableau 11), **VB_f** (cf. Illustration 5) et **VB_c** (cf. Illustration 6) correspondants respectivement aux nouvelles classes de valeurs du véhicule, types de véhicule pour l'étude de la fréquence et types de véhicule pour l'étude du coût ont déjà étés traitées précédemment et ne seront donc pas présentées de nouveau.

Gender

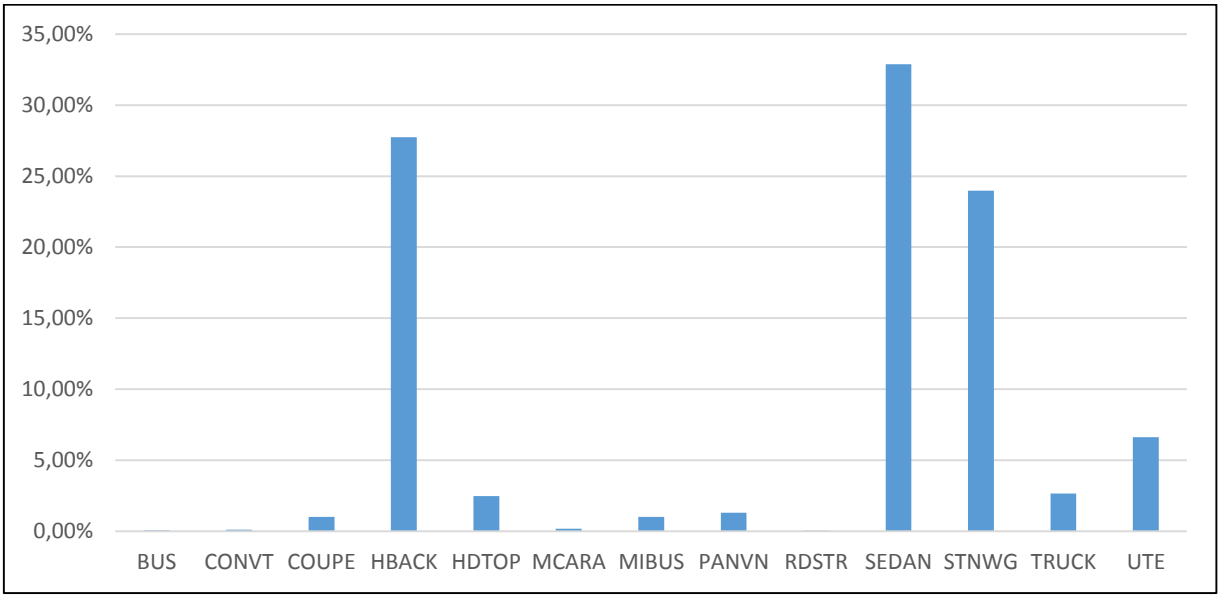


Graphique 3 - Composition du portefeuille par genre

Le portefeuille est à dominance féminine.

Il n'y a pas de majorité écrasante de l'un des deux genres.

Veh_body

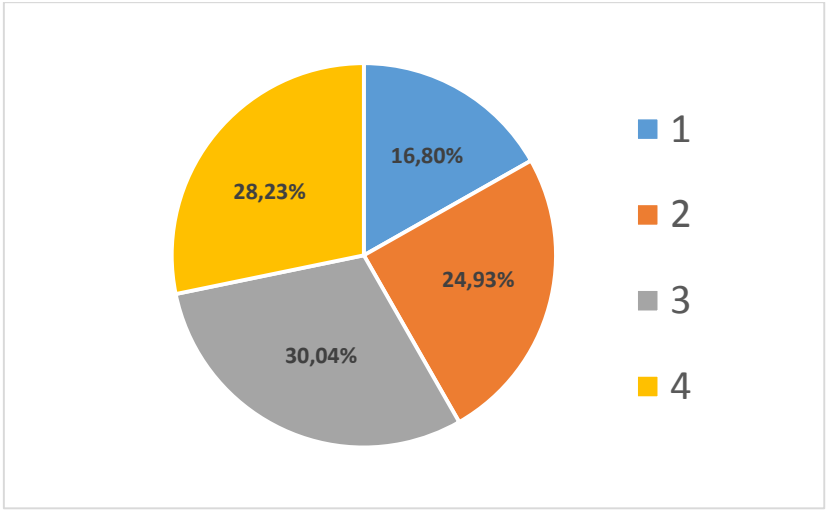


Graphique 4 - Répartition du portefeuille par type de véhicule

Sur les treize types de véhicules présents dans notre portefeuille, trois types se distinguent de par leur présence majoritaire: les HBACK, les SEDAN et les STNWG qui représentent à eux trois plus de 80 % du portefeuille.

Cette dominance a déjà été soulevée lors de la création de classes de véhicules via l'étude des effectifs par classe (cf. Tableaux 13 et 15).

Veh_age

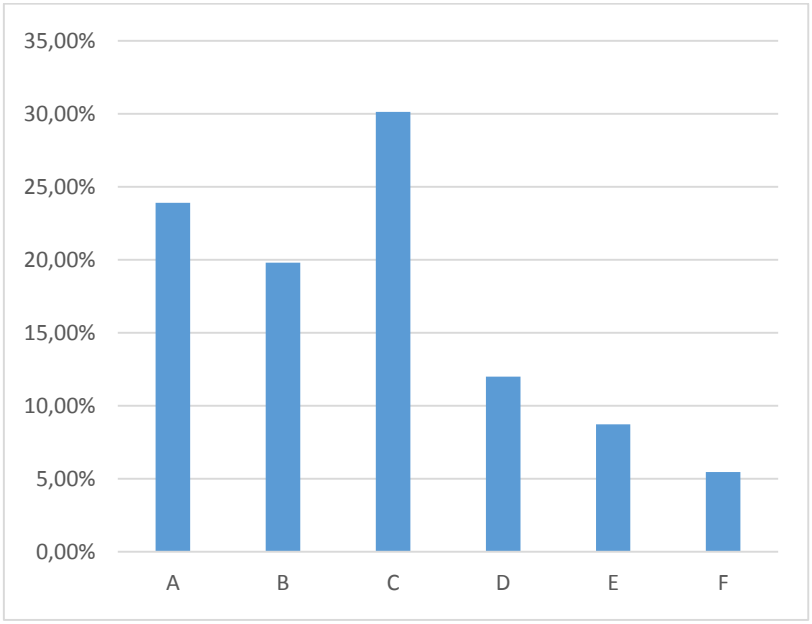


Les classes d'âges du véhicule sont assez homogènes.

Toutes les catégories d'âges sont bien représentées avec une légère dominance des véhicules les plus anciens.

Graphique 5 - Composition du portefeuille par classe d'âges de véhicule

Area



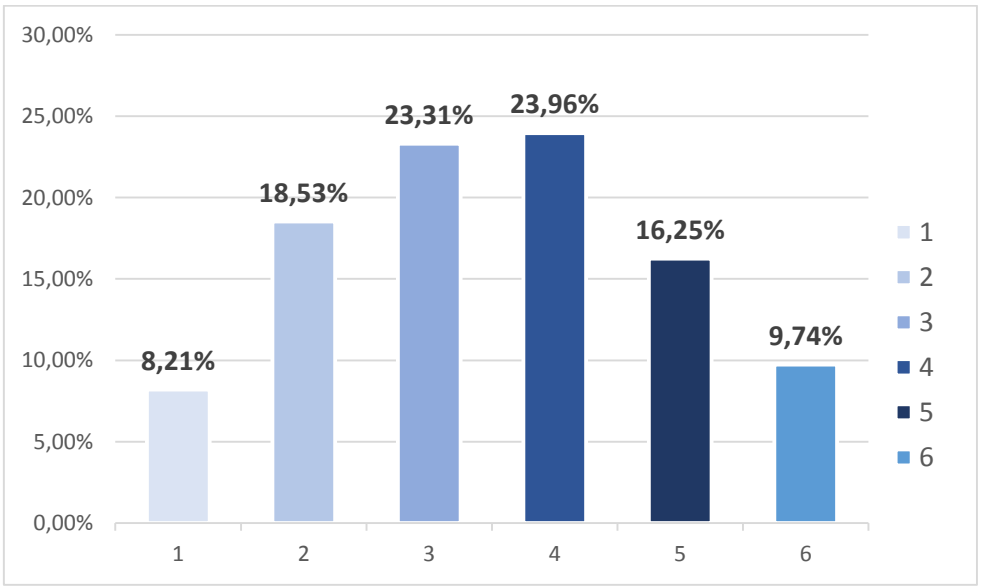
Graphique 6 - Répartition des contrats par Zone (Area)

La part de contrats pour chacune des six zones géographiques étudiées est supérieure à 5 %.

Il existe des disparités en termes d'effectif au sein des classes. Ces dernières pourraient s'expliquer si l'on avait connaissance de la signification de chacune des lettres.

Par exemple, si la lettre C correspond aux grandes villes et la zone F à la campagne, il est normal d'observer de tels résultats. N'ayant aucune information sur les intitulés des zones, nous ne ferons pas d'hypothèses là-dessus.

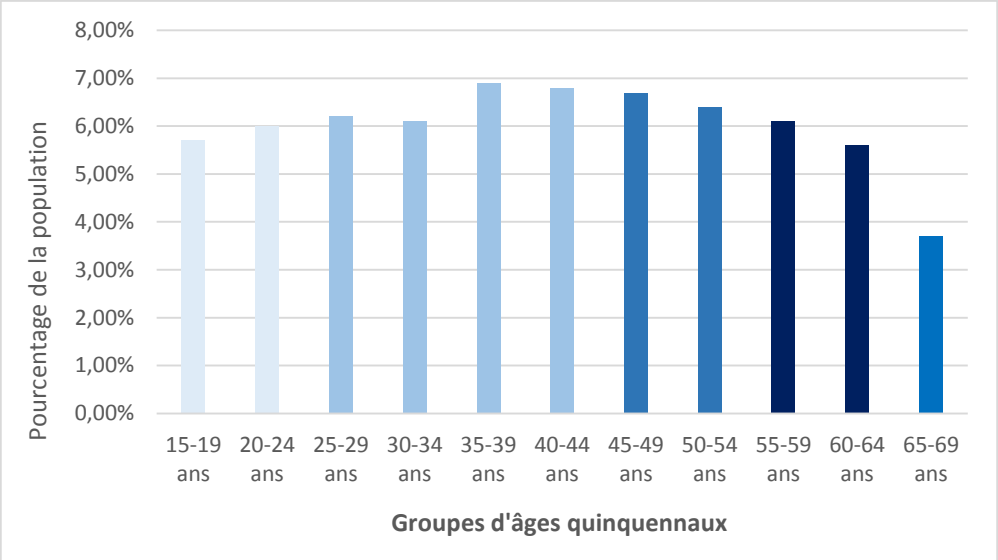
Agecat



Graphique 7 - Répartition du portefeuille par classe d'âges du conducteur

La part de contrats pour chacune des six classes d'âges du conducteur étudiées est supérieure à 5 %.

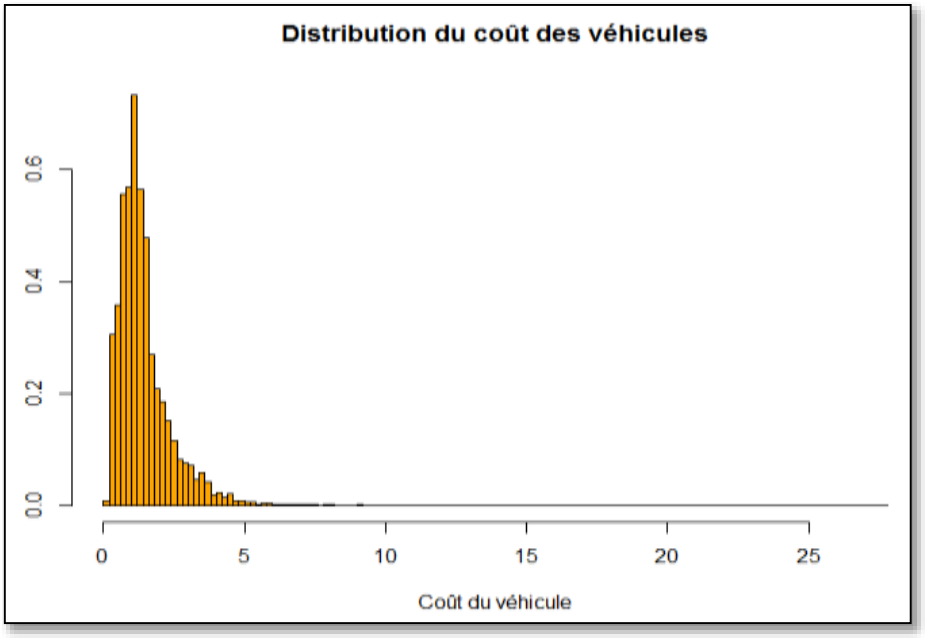
Il existe des disparités en termes d'effectif au sein des classes mais la répartition du portefeuille ne présente aucune structure particulière car ces dernières sont cohérentes avec la pyramide des âges observable en France (cf. Graphique 8).



Graphique 8 - Répartition de la population par groupes d'âges quinquennaux en France (Janvier 2015)[1]

Veh_value

La variable représentant la valeur des véhicules est quantitative continue. L'analyse se porte donc sur sa distribution pour avoir une idée de la composition du portefeuille.

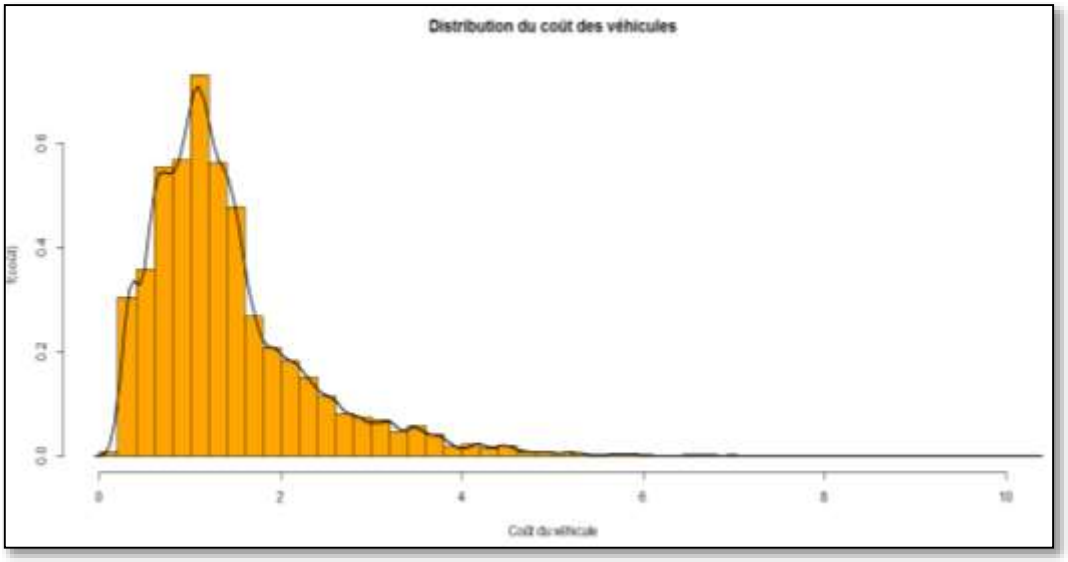


Graphique 9- Représentation de la distribution du coût des véhicules

La variable ***veh_value*** a déjà fait l'objet d'un traitement (cf. § 4.3.2). Elle a été transformée en classes de coûts de véhicules. Lors de ce découpage, nous avons mis en avant l'intervalle de valeurs prises par le coût des véhicules, à savoir , entre 1 438 € à 276 054 €.

La majorité des valeurs restent inférieures à 100 000 € (cf. Tableau 10).
Ci-dessous, la distribution tronquée à 100 000 euros permet d'avoir une meilleure vision de la composition du portefeuille en terme de coût des véhicules.

[1] Source : Insee, estimations de population (données provisoires arrêtées à fin 2014) - [S11]



Graphique 10- Représentation de la distribution du coût des véhicules inférieurs à 100 000 €

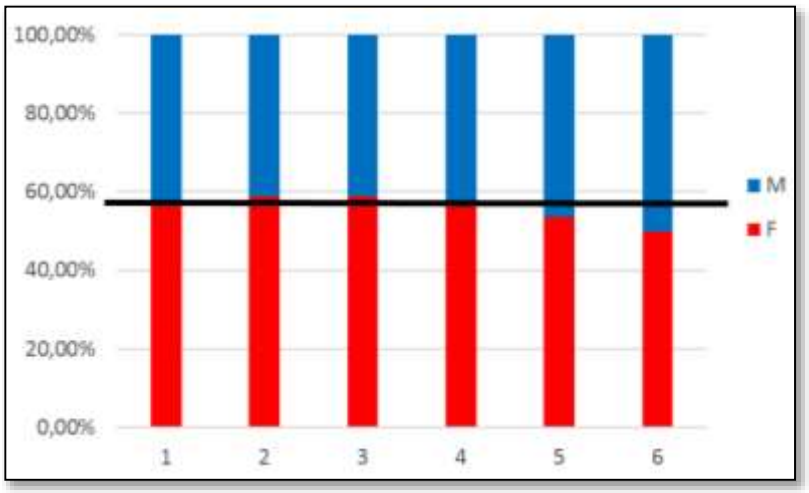
5.1.2 Composition du portefeuille en proportion par genre

Après avoir étudié la composition du portefeuille par variable, l'analyse se porte également sur la répartition du portefeuille par variable tarifaire mais cette fois ci, orientée selon le critère genre.

Ce critère nous intéresse tout particulièrement car il est amené à disparaitre. L'intérêt principal de cette étude reste d'évaluer l'influence des différentes variables disponibles sur la détermination du genre de l'assuré. Elle nous servira alors à valider les variables les plus significatives utilisées lors de la construction du modèle prédictif du genre.

Selon l'étude précédente, les femmes représentent 57 % du portefeuille (cf. Graphique 3) : nous devons donc retrouver ce taux dans la répartition intermodalités. Le cas échéant, cela signifie qu'il existe une différence significative selon le genre entre les modalités prises par chacune des variables. Sur chacun des histogrammes, **nous ferons figurer le niveau 57 % à l'aide d'une ligne continue en guise de base comparative.**
De même que précédemment, le cas de la variable quantitative ***veh_value*** sera traité séparément à l'aide d'un tracé de la distribution de la valeur du véhicule par genre.

Agecat

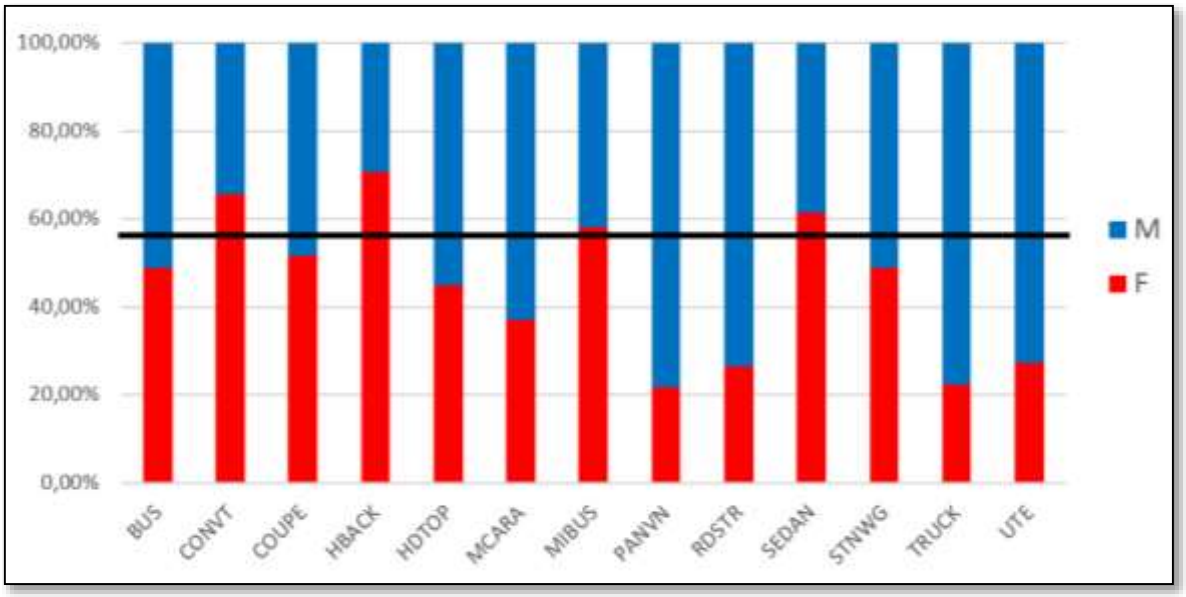


La répartition par classe d'âges respecte bien la répartition par genre.

Aucune spécificité n'est relevée si ce n'est une légère dominance masculine sur les âges élevés.

Graphique 11 - Répartition H/F par classe d'âges

Veh_body

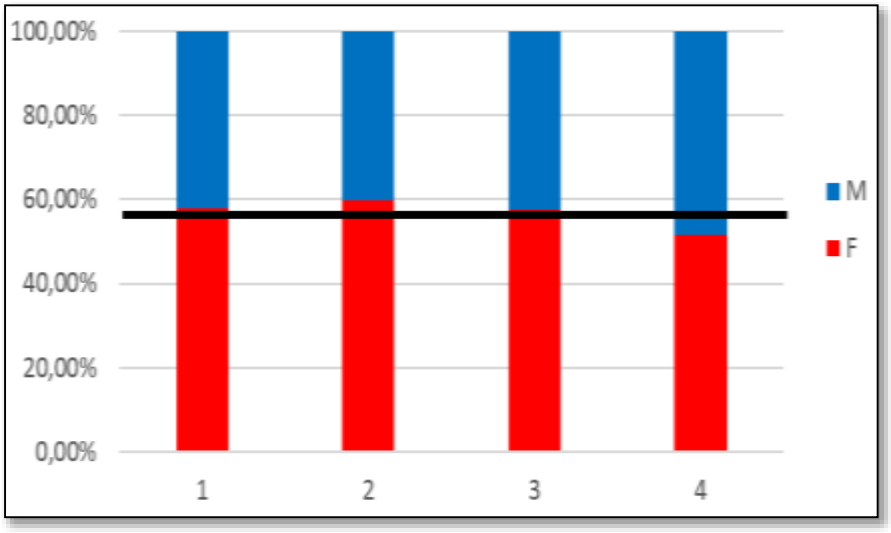


Graphique 12- Répartition H/F par type de véhicule

La proportion d'hommes et de femmes par type de véhicule n'est pas alignée avec la proportion Homme / Femme du portefeuille. Bien au contraire, les proportions alternent de part et d'autre de cette ligne.

Ce diagramme nous indique que ce sont majoritairement des femmes qui possèdent des HBACK et des SEDAN alors que les hommes préfèrent les RDSTR ou les UTE.

Veh_age

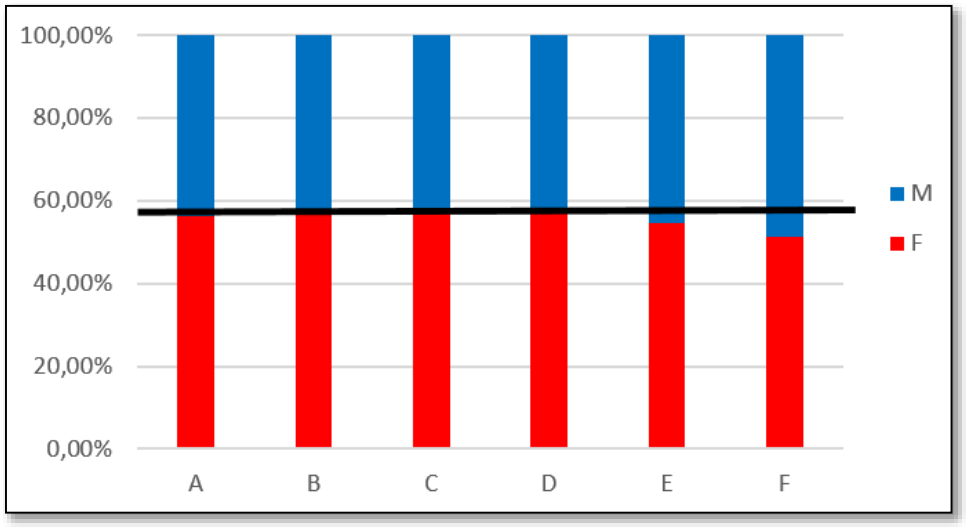


Graphique 13- Répartition H/F par classe d'âges du véhicule

La proportion d'hommes et de femmes par classe d'âges du véhicule est quasiment alignée avec la proportion H/F du portefeuille (57 % de femmes).

On relève une légère dominance féminine en classe 2 et masculine en classe 4.

Area



Graphique 14- Répartition H/F par zone

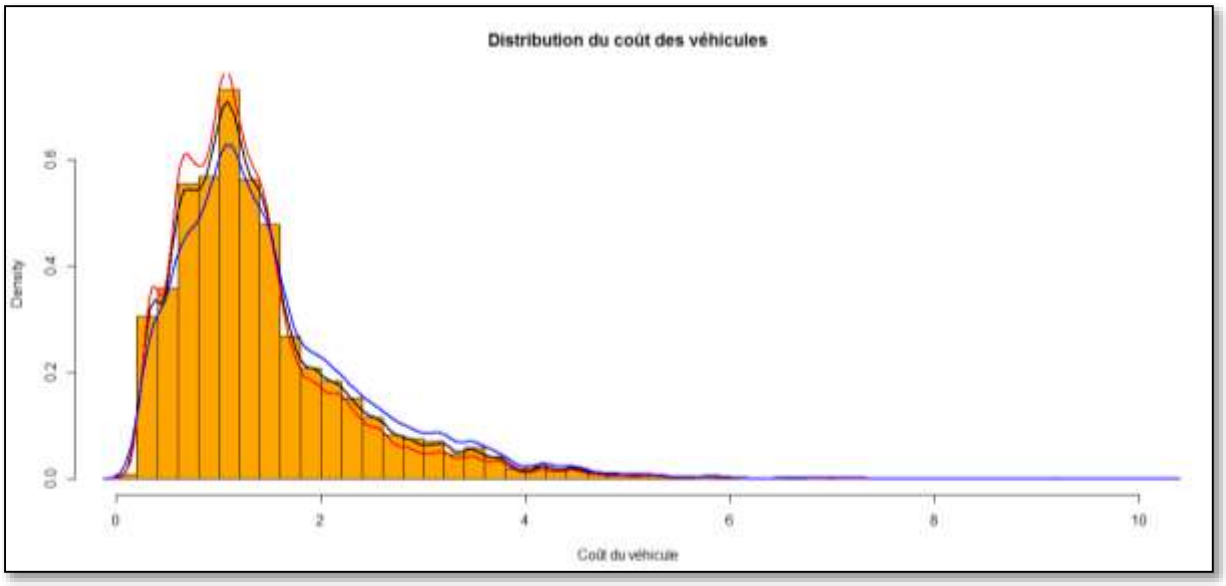
Là encore, la répartition par zone géographique respecte bien la répartition par genre. Aucune spécificité n'est relevée si ce n'est une présence masculine légèrement plus forte pour les zones E et F.

Veh_value

De même que précédement, la distribution tronquée à 100 000 € est utilisée pour avoir une meilleure vision de la composition du portefeuille en terme de coût des véhicules.

Cette fois ci , le graphique (cf. Graphique 15) fait apparaitre :

- la distribution du portefeuille tous genre confondus (**en noir**)
- la distribution des valeurs de véhicule appartenant à des hommes (**en bleu**)
- la distribution des valeurs de véhicule appartenant à des femmes (**en rouge**)



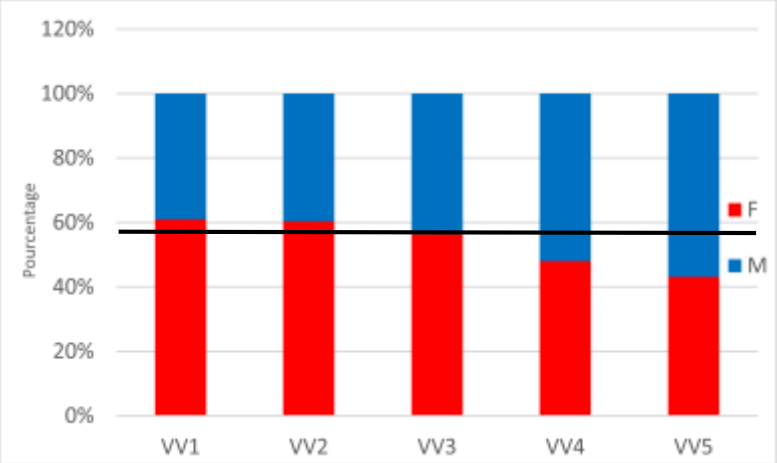
Graphique 15- Représentation de la distribution des valeurs de véhicule par genre

La courbe représentative de la distribution des valeurs des véhicules appartenant à des femmes est au-dessus de celle des hommes pour des valeurs inférieurs à 16 000 €.
Passée cette valeur, celle des hommes passe au-dessus, preuve que les hommes investissent généralement plus dans l'achat d'une voiture coûteuse.

Remarque :

La variable ***veh_value*** a été transformée en variable qualitative par découpage de cette variable continue en classes de valeurs de véhicule.

Il est intéressant de voir l'impact de la création de classes de valeurs de véhicule sur la répartition Homme / Femme.



Graphique 16- Répartition H/F par classe de valeurs du véhicule

Malgré le partitionnement, la distinction H/F apparait toujours, avec une dominance masculine sur les classes de valeurs les plus hautes. Néanmoins, la différence semble moins significative qu'en considérant la distribution complète de la valeur des véhicules.
Le découpage par classes induit la perte potentielle d'information.

Conclusion :

Suite à cette analyse de composition, le type de véhicule semble être le plus influencé par la variable genre. Nous nous attendons donc à un niveau de significativité important de cette variable dans le modèle prédictif.

5.2 Etude de corrélations

Parmi les variables explicatives candidates, on souhaite uniquement retenir celles qui sont les plus corrélées à la variable à expliquer mais également les variables explicatives les moins corrélées entre elles.

En effet, le modèle GLM ne tient pas compte des interactions entre les variables faisant partie du modèle : il faudra alors veiller à ne pas paramétrer le modèle avec des variables qui ont la même influence sur la sinistralité.

Pour répondre à ces deux problématiques, il est nécessaire d'étudier les corrélations :

- Des variables explicatives entre elles (§ 5.2.1) : si la corrélation est élevée entre deux variables, on peut soit en éliminer une, soit créer une variable explicative croisant les informations de ces deux variables ;

- De chaque variable explicative avec la variable à expliquer (§ 5.2.2): si la corrélation est faible on élimine la variable, si elle est élevée on la garde mais on ne sait pas si son apport est significatif.

Selon la nature des variables, plusieurs tests de corrélation peuvent être appliqués.

Figurent ci-dessous les trois tests que nous allons utiliser ainsi que leur cadre d'application.

Association entre 2 variables qualitatives	Test d'indépendance du χ^2 V de cramer
Association entre une variable quantitative et une variable qualitative à 2 modalités	Test t de Student pour groupes indépendants
Association entre une variable quantitative et une variable qualitative à plus de 2 modalités	Test H de Kruskal-Wallis

Tableau 16 - Tests à effectuer selon la nature des variables dont on étudie l'association

Les fondements théoriques de ces trois tests figurent en Annexe (cf. Annexe 2, 3,4) et nous proposons ici leur application directe à nos données.

Le seuil de confiance a été fixé à 95 %, soit $\alpha = 5\ \%$ pour l'ensemble de nos tests.

5.2.1 Etude des corrélations entre variables explicatives

Après transformation, toutes les variables potentiellement explicatives de la sinistralité dont nous disposons sont à présent qualitatives.

C'est donc le **test du χ^2 d'indépendance** qui est adapté à l'étude des corrélations entre ces dernières. Pour chaque paire de variables sur lesquelles le test est réalisé, on émet les hypothèses suivantes :

- Hypothèse nulle (H_0): les deux variables sont indépendantes
- Hypothèse alternative (H_1) : les deux variables sont corrélées

Puis on applique le processus suivant :

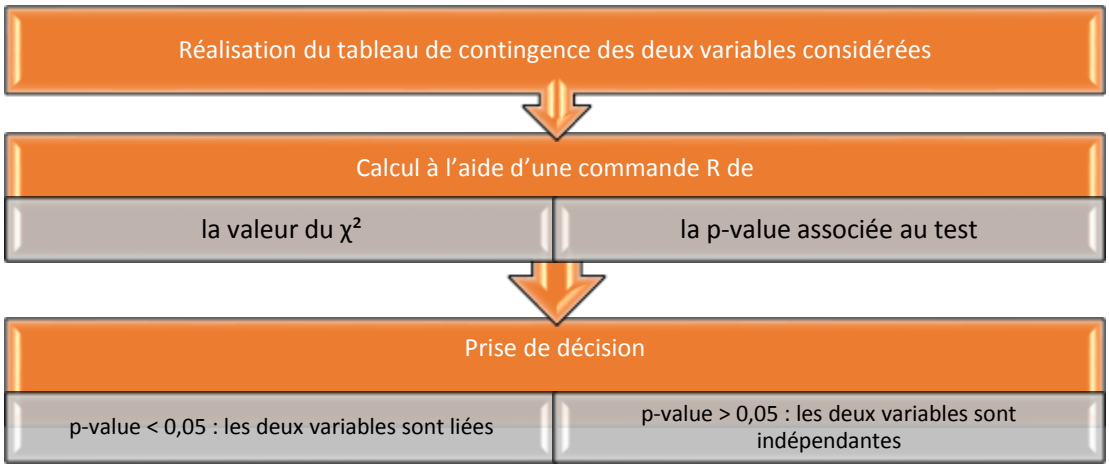


Figure 6 - Processus de test du Khi-2 d'indépendance

Même si la variable ***veh_body*** ne sera plus utilisée comme telle dans la tarification mais sous ses deux formes déclinées en classes (***VB_f*** pour la fréquence et ***VB_c*** pour le coût), nous l'avons volontairement inclue dans les tests afin d'observer l'effet du regroupement des modalités sur les corrélations.
Au total, vingt-sept tests du Khi-deux ont étés effectués.

Pour éviter la lecture d'un procédé redondant, on présente ici un exemple de sortie ainsi que le code utilisé pour la réalisation du test, l'ensemble des résultats étant synthétisé dans un tableau (cf. Tableau 17).

Remarque :

Si les conditions d'application du test ne sont pas vérifiées, notamment à cause d'un effectif insuffisant par case du tableau de contingence (cf. Annexe2), le logiciel R renvoie un message d'erreur pour en avertir l'utilisateur. On signalera l'apparition d'un message par la couleur **rouge** sur le tableau récapitulatif des résultats.
La robustesse du test nous permet cependant de conserver tous les résultats.

Exemple : la corrélation entre le type de véhicule et l'âge du véhicule

- **Réalisation du tableau de contingence entre *veh_body* et *veh_age***

```
tab_cont=table(data$veh_body,data$veh_age)
```

	BUS	CONVT	COUPE	HBACK	HDTOP	MCARA	MIBUS	PANVN	RDSTR	SEDAN	STNWG	TRUCK	UTE
1	2	21	87	4216	137	4	0	66	18	3537	3113	316	732
2	6	18	91	5417	269	16	47	152	9	5651	3776	310	814
3	16	29	167	5128	551	46	184	244	0	7277	4507	575	1335
4	14	13	435	4152	620	55	485	290	0	5764	4831	541	1704

Illustration 7- Tableau de contingence entre le type de véhicule et l'âge du véhicule obtenu sous R

- **Calcul à l'aide d'une commande R de la valeur du χ^2 et de la p-value du test**

```
> résultat <- chisq.test(tab_cont, correc=F)

Warning message:
In chisq.test(tab_cont, correc = F) :
  l'approximation du Chi-2 est peut-être incorrecte

> résultat
        Pearson's Chi-squared test
data:  tab_cont

X-squared = 2434.6, df = 36, p-value < 2.2e-16
```

- **Prise de décision :**

La p-value est inférieure à 0,05 : on rejette H_0 au profit de H_1.

Ci-dessous, le tableau récapitulatif des résultats obtenus :

	veh_age	gender	area	agecat	VV	VB_f	VB_c
veh_body	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16
veh_age		p-value < 2.2e-16	p-value = 5.609e-15	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16
gender			p-value = 1.706e-09	p-value < 2.2e-16	p-value < 2.2e-16	p-value = 8.168e-05	p-value < 2.2e-16
area				p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16
agecat					p-value < 2.2e-16	p-value = 1.024e-06	p-value < 2.2e-16
VV						p-value < 2.2e-16	p-value < 2.2e-16
VB_f							
VB_c							

Tableau 17- p-value des tests du Khi-deux d'indépendance

Toutes nos p values sont inférieures au seuil de 5 % : il existe donc un lien entre toutes les paires de variables. Pour quantifier ce lien, et savoir si l'on doit éliminer certaines variables du modèle, on étudie **l'indice de Cramer** (cf. Annexe 2) associé à chacun de ces tests.

Pour rappel, le V de Cramer est un indice appartenant à [0,1] : plus il est proche de 1, plus le lien entre les variables est important.

	veh_body	veh_age	gender	area	agecat	VV	VB_f	VB_c
veh_body	1,00	0,11	0,27	0,13	0,09	0,28	1,00	1,00
veh_age		1,00	0,07	0,02	0,03	0,40	0,05	0,06
gender			1,00	0,03	0,06	0,12	0,19	0,10
area				1,00	0,05	0,09	0,13	0,05
agecat					1,00	0,05	0,08	0,07
VV						1,00	0,27	0,18
VB_f							1,00	-
VB_c								1,00

Tableau 18- Indices de Cramer associés aux tests du Chi-deux par paire de variables

On relève les plus fortes valeurs de l'indice de Cramer qui témoignent des plus forts liens de corrélation entre :

- le type de véhicule et le genre de l'individu
- le type de véhicule et la classe de valeurs du véhicule
- l'âge du véhicule et la classe de valeurs du véhicule

L'analyse en composition du portefeuille par genre (cf. § 5.1.2) nous avait déjà permis de relever la corrélation entre le type de véhicule et le genre de l'assuré.

De plus, on observe que le regroupement des modalités de la variable ***veh_body*** a permis de réduire les corrélations entre les variables.

En effet, on relève un indice de Cramer de 0,27 entre la variable ***veh_body*** (13 modalités) et ***gender*** alors qu'il atteint seulement 0,19 (resp 0,10) lorsqu'il s'agit du croisement entre le genre et la variable ***VB_f*** (resp ***VB_c***).

Il en est de même pour la corrélation entre ***veh_body*** et ***VV*** avec un passage de 0,28 à 0,27 (resp 0,18) lorsqu'il s'agit du croisement entre ***VV*** et la variable ***VB_f*** (resp ***VB_c***).

Malgré les corrélations relevées, les liens demeurent suffisamment faibles pour que l'on puisse envisager la suppression d'une variable tarifaire de l'étude qui serait due à une corrélation trop importante avec d'autres variables (V de Cramer < 0,5).

On conservera donc l'ensemble de ces variables pour la modélisation.

5.2.2 Etude de corrélation entre variables explicatives et variable à expliquer

On souhaite à présent étudier le lien entre les variables explicatives dont nous disposons pour chaque assuré et les variables à expliquer observées.
La fréquence et le coût des sinistres (variables à expliquer) sont toutes deux quantitatives, quant à nos variables explicatives, elles sont qualitatives et admettent entre deux et six modalités.

La seule variable admettant deux modalités est la variable « ***gender*** », correspondant au genre de l'assuré.

Nous allons donc réaliser (cf. Tableau 16):

- pour la variable « ***gender*** » : un test de Student pour groupes indépendants pour déterminer son lien avec chacune des variables à expliquer.
- pour toutes les autres variables (qui admettent plus de deux modalités), ce sera le test de Kruskal-Wallis qui sera utilisé.

Avant tout test, on réalise une analyse descriptive sommaire qui nous permettra de contrôler mais également d'appuyer la cohérence des résultats de ces tests.

Les fréquences et coûts moyens par modalité ont été reportés dans le tableau suivant :

Modalité	Fréquence moyenne	Coût moyen
F (femme)	22,64 %	1 354,97 €
H (homme)	19,52 %	1 534,06 €
veh_age1	17,35 %	1 242,53 €
veh_age2	23,82 %	1 348,15 €
veh_age3	22,76 %	1 487,09 €
veh_age4	20,02 %	1 570,75 €
agecat1	37,64 %	1 839,44 €
agecat2	20,05 %	1 470,03 €
agecat3	20,90 %	1 413,44 €
agecat4	21,58 %	1 322,11 €
agecat5	17,48 %	1 295,57 €
agecat6	15,76 %	1 341,92 €
VB_c1/VB_f1	6,31 %	534,44 €
VB_c2/VB_f2	17,23 %	1 326,05 €
VB_c3/VB_f3	23,50 %	1 535,83 €
VB_c4/VB_f4	38,63 %	1 959,56 €
VV1	24,16 %	1 571,80 €
VV2	20,81 %	1 392,56 €
VV3	20,35 %	1 496,72 €
VV4	20,90 %	1 349,14 €
VV5	18,17 %	1 176,04 €
A	17,31 %	1 289,26 €
B	24,22 %	1 381,72 €
C	21,35 %	1 489,78 €
D	25,08 %	1 390,64 €
E	18,57 %	1 576,61 €
F	23,83 %	1 666,19 €

Tableau 19- Résultats de l'analyse descriptive : fréquence et coût de sinistre par modalité

Corrélation entre la sinistralité et le genre

L'étude va être effectuée en deux temps et la séparation réalisée est identique à celle utilisée pour le traitement du problème d'estimation de la charge de sinistre.

Tout d'abord, on s'intéresse au lien entre le **genre** et la **fréquence** de sinistre puis on se penchera sur le lien entre le **genre** et le **coût** des sinistres.
Pour chacune de ces étapes, on utilise le test de Student pour groupes indépendants.

Nos deux groupes étant les femmes (groupe 1) et les hommes (groupe 2), on dit que les deux groupes sont indépendants car un individu du groupe 1 ne peut se retrouver dans le groupe 2 et inversement.

On suivra ensuite le processus de test suivant :

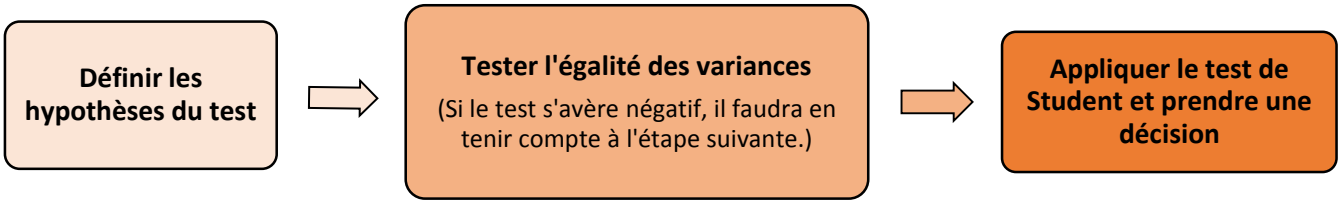


Figure 7 – Processus du test de Student

Remarque :
Nous ne testons pas la normalité des données car l'effectif de chacun des groupes (Homme/Femme) est bien supérieur à 30, borne à partir de laquelle ce test n'est plus nécessaire.

Corrélation entre le genre et la fréquence de sinistres

1) Hypothèses du test:

- Hypothèse nulle (H_0) : le genre de l'individu n'a aucune influence sur la fréquence de sinistre.
- Hypothèse alternative (H_1) : le genre de l'individu influe sur la fréquence de sinistre.

2) Test de l'égalité des variances

Le test d'égalité des variances s'effectue avec la fonction ***var.test()*** disponible sous R.
L'argument principal de ***var.test()*** est un modèle de la forme « var. réponse ~ var. explicative», soit dans notre cas : fréquence ~ genre.

```
var.test(data$fréquence~data$gender)

 F test to compare two variances

data:  data$fréquence by data$gender

F = 3.5632, num df = 38580, denom df = 29207,

p-value < 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

 3.487402 3.640624

sample estimates:

ratio of variances

          3.563247
```

Le ratio des variances vaut 3,56 et la p value est bien inférieure au seuil significatif de 5 % : **nous ne sommes donc pas en présence d'égalité des variances.**

Le test de Student est assez robuste pour assurer une réponse satisfaisante même si cette hypothèse n'est pas vérifiée. Il faudra néanmoins penser à l'indiquer lors de la réalisation du test pour que ceci soit pris en compte.

3) Test de Student pour groupes indépendants et prise de décision

Ce test est réalisé à l'aide de la fonction ***t.test()*** qui prend comme argument :

- **data$fréquence** ~ **data$gender**, qui correspond à notre modèle de données ;
- **alternative = "greater'', ** qui signifie que l'on effectue un test unilatéral et on s'attend à ce que la moyenne du groupe 1 (les femmes) soit plus élevée[1] que celle du groupe 2 (les hommes) ;
- **paired = FALSE**, qui indique que les groupes étudiés sont indépendants ;
- **var.equal=FALSE**, afin de préciser que les variances des deux groupes sont hétérogènes.

```
t.test(data$fréquence~data$gender, alternative="greater", paired = F)

Welch Two Sample t-test

data:  data$fréquence by data$gender

t = 1.5102, df = 61349, p-value = 0.06549

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

 -0.002780822          Inf

sample estimates:

mean in group F       mean in group M

      0.2263553             0.1951603
```

La p value renvoyée par le test est supérieure au seuil de 5 % : on ne rejette pas H_0 et on peut conclure que la fréquence de sinistre ne dépend pas du genre de l'individu.

Corrélation entre le genre et le coût des sinistres

1) Hypothèses du test:

- Hypothèse nulle (H_0) : le genre de l'individu n'a aucune influence sur le coût des sinistres.
- Hypothèse alternative (H_1) : le genre de l'individu influe sur le coût des sinistres.

2) Test de l'égalité des variances

Notre modèle de données est à présent : « **data$coûtUnitaire** ~ **data$gender** ».

[1] Cf. Tableau 19

```
var.test(data$coûtUnitaire~data$gender)

F test to compare two variances

data:  data$coûtunit by data$gender

F = 0.69235, num df = 2826, denom df = 2087, p-value < 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

 0.6389511 0.7498448

sample estimates:

ratio of variances

         0.6923497
```

Le ratio des variances vaut 0,69 et la p value est inférieure à 0,05 : nous ne sommes **donc pas en présence d'égalité des variances.**
On remarque tout de même que le ratio des variances est plus proche de 1 que pour l'étude des fréquences (0,69 contre 3,5).

De même que précédemment, on indiquera ce résultat lors de la réalisation du test pour que ceci soit pris en compte.

3) Test de Student pour groupes indépendants et prise de décision

A présent, les arguments du test deviennent :

- **data$coûtunit** ~ **data$gender**, qui correspond à notre modèle de données ;
- **alternative = "less"** qui signifie que l'on effectue un test unilatéral et on s'attend à ce que la moyenne du groupe 1 (les femmes) soit plus faible que celle du groupe 2 (les hommes) [1] ;
- **paired=FALSE**, qui indique que les groupes considérés indépendants ;
- **var.equal=FALSE**, pour tenir compte du fait que les variances sont hétérogènes.

[1] Cf. Tableau 19

```
t.test(data$coûtUnitaire~data$gender, alternative="less", paired = F)

 Welch Two Sample t-test

data:  data$coûtunit by data$gender

t = -2.6599, df = 3995.6, p-value = 0.003923

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

      -Inf -68.31894

sample estimates:

mean in group F mean in group M

       1354.970        1534.061
```

Notre p value est clairement inférieure au seuil de 5 %.
L'hypothèse selon laquelle le coût des sinistres chez les hommes est supérieur à celui observé chez les femmes est validée.

Corrélation entre les variables qualitatives à plus de deux modalités et la sinistralité

Le test de **Kruskal-Wallis** nous permet à présent d'étudier la corrélation entre la sinistralité et les variables qualitatives à plus de deux modalités.

Pour chaque variable considérée, une modalité sera considérée comme un groupe.

Par exemple, pour la variable ***veh_age*** qui correspond à la classe d'âges du véhicule, nous distinguons quatre groupes correspondant respectivement aux quatre modalités (1, 2, 3, 4) de cette variable.

Le but du test est de vérifier si l'un des groupes se comporte différemment des autres en termes de fréquence ou de coût de sinistre, autrement dit, de vérifier s'il existe une influence de la variable étudiée sur la sinistralité.

On procèdera selon les étapes suivantes :

1- Mise en place les hypothèses du test
2- Application du test et prise de décision

On soulève de nouveau le problème du nombre important de tests à effectuer.
Une synthèse des résultats sera présentée et un exemple détaillé d'application sur le lien entre la fréquence et l'âge du conducteur est fourni ci-dessous.

Exemple d'application : Corrélation entre la fréquence de sinistre et l'âge du conducteur

1) Hypothèses du test

- Hypothèse nulle (H_0) : l'âge de l'individu n'a aucune influence sur la fréquence de sinistres.
- Hypothèse alternative (H_1) : l'âge de l'individu influe sur la fréquence de sinistres.

2) Application du test et prise de décision

On utilise la fonction **kruskal.test()** qui retourne la p-value du test de Kruskal Wallis ainsi que la valeur.

```
kruskal.test(data$fréquence~data$agecat)

        Kruskal-Wallis rank sum test

data:  data$fréquence by data$agecat
Kruskal-Wallis chi-squared = 72.116, df = 5, p-value = 3.716e-14
```

La p-value retournée par le test est inférieure au seuil de 5 % : on peut donc affirmer que l'âge du conducteur influe sur la fréquence de sinistres.

R nous permet également d'analyser plus en détail l'origine des différences observées en donnant des résultats d'analyse entre les groupes deux à deux à l'aide de la fonction ***kruskalmc()***.

```
#Analyse des différences

kruskalmc(data$fréquence~data$agecat,probs = 0.05)

Multiple comparison test after Kruskal-Wallis

p.value: 0.05

Comparisons

       obs.dif critical.dif difference

1-2  483.67445     912.2164      FALSE
1-3  534.68446     886.0388      FALSE
1-4  620.39274     882.9869      FALSE
1-5  991.53063     939.8801       TRUE
1-6 1051.10963    1039.2210       TRUE
2-3   51.01001     682.5380      FALSE
2-4  136.71830     678.5715      FALSE
2-5  507.85618     751.1112      FALSE
2-6  567.43518     872.2234      FALSE
3-4   85.70829     642.9503      FALSE
3-5  456.84618     719.0924      FALSE
3-6  516.42517     844.8074      FALSE
4-5  371.13789     715.3286      FALSE
4-6  430.71688     841.6060      FALSE
5-6   59.57900     901.1157      FALSE
```

Les différences relevées concernent les groupes 1 et 5 ainsi que les groupes 1 et 6, soit les extrêmes en terme de fréquence de sinistralité.

Synthèse des résultats:

Le test précèdent a été effectué pour les cinq variables explicatives présentes dans le Tableau 19.
On rappelle que la variable « **gender** » a bénéficié d'un traitement différent (Test de Student) car elle possède uniquement deux modalités.
Cette dernière ne figure donc pas dans les résultats présentés ci-dessous.

	Kruskal Wallis	
	Coût	**Fréquence**
veh_age	1,16E-05	1,10E-05
agecat	0,026	3,72E-14
VB_c / VB_f	0,001438	0,01571
VV	0,04227	3,57E-08
Area	4,26E-05	0,005117

Tableau 20 - P-value associées aux tests

Toutes les p-values obtenues sont inférieures au seuil de 5 % : on s'attend donc à ce que toutes les variables présentant plus de deux modalités soient présentes aussi bien dans le modèle de coût que le modèle de fréquence.

Cependant, la comparaison des p-values nous permet de désigner les variables qui auront certainement le plus de significativité :

- Pour le coût : l'âge du véhicule et la zone
- Pour la fréquence : l'âge du conducteur et du véhicule ainsi que la valeur du véhicule

Conclusion

L'analyse des corrélations nous a permis d'établir les résultats suivants :

- le genre est un critère qui influence le coût mais pas la fréquence de sinistres
- l'âge du véhicule est un critère important dans l'explication de la sinistralité
- l'âge de l'assuré influe sur la fréquence de sinistres

Cependant, ces résultats doivent être nuancés car les tests ne permettent pas d'avoir une vision globale sur les interactions entre les variables : seuls des tests **par paire de variable** ont pu être réalisés.
Nous allons à présent passer à la modélisation de la sinistralité du portefeuille où l'élimination des variables non significatives se fera via la sélection conjointe des variables.

Partie III - Modélisation

Bien avant que la *Gender Directive* telle qu'elle est présentée aujourd'hui ne soit mise en place, les assureurs ont cherché le moyen de modifier leur méthode de tarification. Ils avaient alors en tête d'honorer au moins deux objectifs, à savoir : respecter la règlementation qui entrerait en vigueur sous peu mais aussi proposer un tarif juste aux assurés.

Les variables retenues pour la segmentation étaient déjà quasi-uniques pour chacun d'entre eux, la suppression de la variable « genre » n'a pas fait tendre leurs modèles vers une uniformisation, bien au contraire. En effet, chacun d'entre eux tente de faire preuve d'innovation tout en assurant un calcul juste du montant des primes à payer pour ses assurés afin d'accroître sa présence sur le marché.

Tous les assureurs n'ont donc pas optés pour la même méthode de tarification et cette troisième partie, qui constitue le cœur de la problématique du mémoire, est justement réservée à la mise en place de modèles de tarification alternatifs qui auraient pu (ou qui ont étés) utilisés par les assureurs après l'établissement de la *Gender Directive*.

Parmi les différentes méthodes envisagées, nous distinguons deux écoles, avec d'un côté la méthode qui consiste à ne plus utiliser la variable « genre » (quand bien même elle serait connue de l'assureur), et de l'autre, celles qui font intervenir le genre de l'assuré dans le calcul de la Prime Pure (mais pas dans la Prime Commerciale). Les conséquences sur la segmentation du portefeuille d'assurés sont alors différentes.

Pour tous les systèmes proposés, la modélisation de la charge de sinistres est fondée sur l'hypothèse d'un modèle collectif, dont l'une des propriétés est de permettre l'estimation de la charge moyenne des sinistres en séparant l'étude de cette dernière en l'étude de la fréquence moyenne et du coût moyen de sinistre selon la formule :

$$E[S] = E[N] \times E[C]$$

[Avec S la charge de sinistre, N la fréquence (le nombre) de sinistres et C leur coût].

Notre méthodologie suivra ensuite la logique suivante :

On commencera par établir un modèle de référence, dit modèle de base (**modèle B**) qui nous servira de base comparative. Il correspond au modèle existant avant la mise en application de la *Gender Directive* : c'est un modèle où l'utilisation de la variable « genre » n'est pas contrainte par la règlementation (Chapitre 6).

Puis, nous mettrons en place un modèle semblable au modèle de base, sauf qu'il ne considèrera plus la variable « genre » comme une variable tarifaire (Chapitre 7) : ce sera le modèle **S**ans **G**enre (noté **modèle SG**).

Nous comparerons alors ces deux modèles afin de soulever le besoin pour l'assureur de mettre en place des modèles alternatifs à cette suppression de variable et non pas de se contenter de la retirer de son précèdent modèle tarifaire (Chapitre 8).

Suite à celà, nous proposerons différents modèles alternatifs (Chapitre 9) qui pourront être utilisés par l'assureur pour être conforme aux exigences règlementaires à savoir :

- Un modèle de pondération (**modèle P**)
- Deux modèles prédictifs (**modèle PR_L** et **PR_CART**)

Le modèle de **B**ase et le modèle **S**ans **G**enre nécessiteront la mise en place de deux GLM, l'un permettant de modéliser le coût des sinistres et l'autre, leur fréquence.

Nous suivrons alors les quatre étapes nécessaires à leur mise en place décrites en Partie 1 (§ 3.2).
Le lecteur est invité à revisiter la partie théorique concernant les GLM en cas d'incompréhension de certains résultats.

Pour finir, nous proposerons une vue d'ensemble avec une synthèse des résultats obtenus dans le but de pouvoir comparer l'ensemble des modèles à la fois (Chapitre 10).

Chapitre 6 - Tarification sans contrainte sur l'utilisation du genre de l'assuré (modèle B)

Aujourd'hui, cette tarification n'est pas applicable telle qu'elle (car elle aboutit à la construction d'un tarif différencié selon le genre) mais servira de **base comparative** aux résultats obtenus par les méthodes alternatives suivantes.

C'est la méthode de tarification qui était utilisée avant la mise en place de la *Gender Directive* : l'utilisation du genre de l'assuré n'est pas règlementée.

6.1 Modèle pour le coût

La variable à expliquer dans ce modèle est le coût des sinistres.
Par « coût des sinistres » nous entendons le montant unitaire d'un sinistre considéré comme attritionnel.

6.1.1 Etape 1 : le choix du modèle

En règle générale, les montants de sinistres sont modélisés à partir d'une loi Gamma.

Bien que la fonction de lien canonique associée à la loi Gamma soit la fonction **inverse**, on met ici en application la remarque faite au sujet du choix de la fonction de lien (cf. § 3.2) : les valeurs estimées par le modèle doivent être positives tout comme les coûts. On souhaite également obtenir un modèle multiplicatif afin de pouvoir comparer l'influence de chacune des modalités par variable.
On utilise donc la fonction de lien « **log** » pour la loi Gamma.

Avant de commencer la modélisation, il va falloir adapter le Modèle Linéaire à nos données, en l'occurrence, tenir compte de la franchise et des nombreux sinistres dont le montant correspond à cette dernière.

La loi Gamma étant à support dans $[0\,,+\infty[$, on décide dans un premier temps de décaler les coûts des sinistres en leur soustrayant le montant de la franchise le temps de la modélisation.

Cette action permet de ne pas affecter volontairement un poids nul à la densité des coûts sur l'intervalle [0, d], où d correspond au montant de la franchise.
On utilisera la linéarité de l'espérance pour ensuite ajouter ce montant aux différents coûts obtenus pour le calcul du coût moyen par classe tarifaire.

De plus, le décalage des montants dans le but d'avoir une distribution qui débute en zéro a entrainé l'affectation d'un poids trop important sur la valeur nulle du fait des 713 coûts valant la franchise[1]: ceci n'est pas propice à l'adéquation d'une loi continue, pour laquelle le poids affecté à chaque valeur ponctuelle est nul par définition.
On ne tiendra donc pas compte des coûts égaux à la franchise dans la mise en place du Modèle Linéaire Généralisé.

Bien entendu, ils ne disparaissent pas de l'étude et seront réintégrés dans le calcul de la Prime Pure de la sorte:

$$PP = \boldsymbol{PP_{attri}} + PP_{grave}$$

$$= \boldsymbol{E[N|C > d] \times E[C|C > d] + d \times E[N|C = d]} + PP_{grave}$$

Avec

- $E[C|C > d]$, l'espérance des coûts supérieurs à la franchise (déterminée dans ce modèle)

- $E[N|C = d]$, la fréquence de sinistres dont le coût est égal à la franchise (déterminée sur la base des observations).

De ce fait, la modélisation de la fréquence des sinistres fera également l'objet d'une distinction entre l'occurrence des sinistres de coût égal au montant de la franchise et les autres.

En utilisant la fonction de lien log appliquée à la variable à expliquer « ***coûtUnitaire*** » on a alors la relation :

$$\log(E[C - d|C > d]) = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p$$

$$\Leftrightarrow$$

$$E[C - d|C > d] = \exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p)$$

$$\Leftrightarrow$$

$$E[C|C > d] = \exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p) + d$$

Remarque :
Le choix de la fonction de lien « log » est également appuyé par l'analyse des valeurs des critères de sélection de modèle : l'AIC et le BIC.
Pour rappel, le meilleur modèle est celui qui admet le plus petit AIC ainsi que le plus petit BIC.

Loi	AIC	BIC
Gamma inverse	68 945	69 090,8
Gamma log	68 941	69 086,98

Tableau 21- AIC et BIC du modèle avec la loi Gamma pour les fonctions de lien log et inverse

[1] Cf. Tableau 7

6.1.2 Etape 2 : Estimation des coefficients de la régression

Après le choix de la loi suivie par les coûts de sinistre, c'est au tour des coefficients attribués à chaque modalité de variable d'être estimés.
Cette estimation est réalisée à l'aide de la fonction ***glm ()*** disponible sous R.
La sélection des variables se faisant par la suite, on commence d'abord par intégrer l'ensemble des variables dans le modèle, à savoir :

Intitulé	Signification	Nombre de modalités
VV	Classe de valeurs du véhicule	5
VB_c	Classe de types de véhicule	4
Agecat	Classe d'âges du conducteur	6
area	Zone géographique	6
gender	Genre de l'assuré	2
Veh_age	Classe d'âges du véhicule	4

Tableau 22- Variables intégrées dans le modèle d'estimation du coût des sinistres

La fonction ***glm()*** retourne de nombreuses informations :

- Le modèle dont on cherche à estimer les coefficients

```
Call:
glm(formula = (data$coûtUnitaire - 159.75) ~ data$gender + data$agecat +
    data$veh_age + data$VB_c + data$area + data$VV, family = Gamma(link = "log"))
```

- Les indices de position des résidus de déviance du modèle

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5792  -1.5117  -0.7631   0.1563   5.4709
```

- L'estimation du paramètre β associé à chaque modalité (colonne Estimate)
- L'estimation de l'écart-type du paramètre associé à chaque modalité (colonne Std.error)
- **Intercept** : le terme constant dans la modélisation de la moyenne de Y

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.26966    0.42043  14.913  < 2e-16 ***
data$genderM     0.10896    0.05134   2.123 0.033849 *
data$agecat2    -0.22353    0.09393  -2.380 0.017372 *
data$agecat3    -0.26571    0.09167  -2.899 0.003767 **
data$agecat4    -0.34992    0.09163  -3.819 0.000136 ***
data$agecat5    -0.36008    0.10276  -3.504 0.000463 ***
data$agecat6    -0.35468    0.11712  -3.028 0.002473 **
data$veh_age2    0.07502    0.07739   0.969 0.332400
data$veh_age3    0.13143    0.08235   1.596 0.110575
data$veh_age4    0.10180    0.09923   1.026 0.304993
data$VB_cVB_c2   1.14640    0.39084   2.933 0.003374 **
data$VB_cVB_c3   1.28412    0.39519   3.249 0.001166 **
data$VB_cVB_c4   1.44783    0.40348   3.588 0.000337 ***
data$areaB       0.06269    0.07569   0.828 0.407565
data$areaC       0.13194    0.06854   1.925 0.054311 .
data$areaD      -0.01488    0.09010  -0.165 0.868863
data$areaE       0.09418    0.09728   0.968 0.333029
data$areaF       0.17437    0.11013   1.583 0.113411
data$VVVV2      -0.13993    0.08219  -1.703 0.088732 .
data$VVVV3      -0.05961    0.09374  -0.636 0.524874
data$VVVV4      -0.22297    0.11185  -1.993 0.046277 *
data$VVVV5      -0.34005    0.12772  -2.663 0.007785 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- La valeur observée de la statistique de test de Student (**t value**).

Ce test prend comme hypothèse : $H_0 : \beta_i = 0 \; contre \; H_1 : \beta_i \neq 0$, il permet ainsi de savoir si la modalité influe ou non sur la variable à expliquer. La t value est simplement obtenue en divisant la valeur du coefficient de régression par son écart-type (estimate/Std.error)

- La probabilité critique (p-value) donnant pour la statistique de test de Student, la probabilité de se situer au-delà d'un intervalle de confiance de la valeur estimée **(colonne Pr(>|t|)** sous H_0.

Cette valeur est lue dans la table du t à l'intersection entre la ligne (n-p) et la colonne représentant le niveau de confiance choisi (généralement α = 5 %)[1]. Si la valeur de la table est inférieure à la t value, on rejettera H_0.
Autrement dit, si l'on choisit comme seuil d'erreur α =5 %, une p-value supérieure à 5 % revient à accepter H_0 et donc à dire que la modalité en question n'est pas significative. ***Cependant l'interprétation de cette probabilité critique ne nous permet pas de juger de la significativité d'une variable puisque le test de Student est réalisé pour chaque modalité et non par variable. Ainsi on pourra pour une même variable trouver une modalité significative au sens du test de Student et une seconde modalité (toujours de la même variable) qui ne l'est pas. C'est le cas par exemple de la variable VV avec les modalités VV4 (significative) et VV3 (non significative).***

- La valeur du paramètre de dispersion ϕ du modèle (**dispersion parameter**)
- La déviance du modèle nul (**null deviance**) et du modèle étudié (**residual deviance**)
- Le degré de liberté de chacun des modèles donné par $ddl = n - p$

- La valeur de l'**AIC**
- Le nombre d'itérations nécessaires à l'obtention des coefficients (car la procédure est itérative)

```
(Dispersion parameter for Gamma family taken to be 2.544637)

    Null deviance: 7506.2  on 4178  degrees of freedom
Residual deviance: 7305.7  on 4157  degrees of freedom
AIC: 68941

Number of Fisher Scoring iterations: 8
```

Après analyse de toutes les valeurs retournées, on remarque qu'il manque le coefficient associé à une modalité pour chaque variable incluse dans le modèle.

Par exemple, pour la variable « ***gender*** », on a bien le coefficient associé à la modalité « M » (0,10896) mais nous ne disposons pas de celui associé à la modalité « F ».
En effet, la modalité absente pour chaque variable est appelée **modalité de référence** et son coefficient vaut 0 par construction : il n'apparait donc volontairement pas dans le résultat puisque sa valeur est fixe.

Un individu est appelé **individu de référence** si chacune des variables qui le caractérise prend comme modalité la modalité de référence.

Dans le cadre d'un modèle multiplicatif (fonction de lien log) comme c'est le cas ici, le coût moyen d'un sinistre pour l'individu i ($E[C_i]$) et le coût moyen d'un sinistre de l'individu de référence (C_{ref}) sont liées par la relation suivante :

[1] n représente le nombre d'individus ayant servi à la construction du modèle (nombre d'observations) et p le nombre de paramètres de ce dernier.

$$E[C_i - d | C_i > d] = C_{ref} \times \lambda_1 \times \lambda_2 \times \ldots \times \lambda_p$$

Où

- $(\lambda_j)_{j=1,\ldots,p}$ est le coefficient multiplicateur associé à la modalité prise par l'individu i de la variable j tel que :
  - $\lambda_j = \exp(\beta_j)$
- $C_{ref} = \exp(\beta_0) = \exp(\textbf{Intercept})$

R a choisi pour chaque variable la première valeur, dans l'ordre alphabétique, comme modalité de référence.

Habituellement, on choisit comme modalité de référence la modalité qui présente le plus de polices. Cependant, on ne renommera pas nos variables puisque le choix de la modalité de référence ne joue pas sur la qualité de la prédiction.

En effet, quelle que soit la modalité de référence adoptée pour chaque variable explicative catégorielle, nous obtiendrons exactement la même prédiction lorsque le modèle est appliqué sur un individu de la population.

L'interprétation des résultats se fera alors en fonction de cette modalité de référence.

6.1.3 Etape 3 : Procédure de sélection des variables

Parmi les trois méthodes de sélection conjointe des variables présentées (Ascendante, Descendante ou pas à pas Mixte), nous utiliserons la procédure descendante (« *Backward* ») pour éliminer les variables non significatives (cf. § 3.2.3).

Cette dernière peut être automatisée via la procédure ***step()*** sous R.

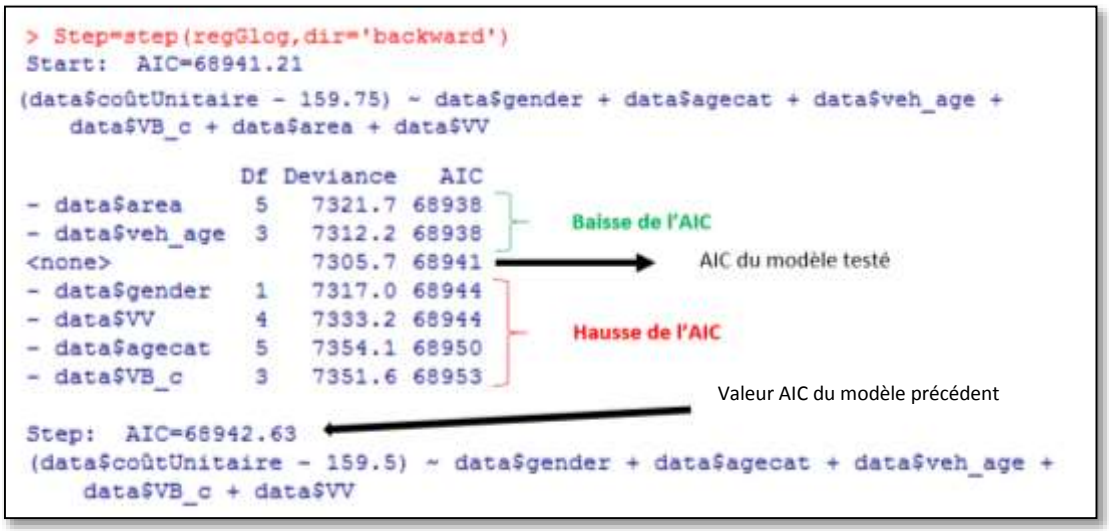


Illustration 8 – Résultat de la fonction step()

Comme son nom l'indique, la fonction « step » procède par étapes :

a. Calcul de l'AIC du modèle de référence (ici regGlog) ainsi que l'impact du retrait de chaque variable sur l'AIC.

b. Suppression de la variable dont le retrait permet une baisse maximale de l'AIC

c. Calcul de l'AIC du nouveau modèle (modèle sans la variable supprimée)

d. Renouvellement de l'étape 2 et 3 jusqu'à ce que le retrait d'une des variables restantes soit synonyme de l'augmentation de l'AIC.

Dans notre modèle, l'étape 2 a permis la suppression de la variable ***area*** (qui était à égalité avec la variable ***veh_age*** avec un AIC de 68 938 mais qui présente une déviance plus importante).
Puis lors du nouveau calcul de l'AIC du modèle sans la variable ***area***, toute suppression de variable impliquait une hausse de l'AIC : la procédure s'arrête donc ici.

Variable	Retenue/ Rejetée
VV	Retenue
VB_c	Retenue
Agecat	Retenue
area	Rejetée
gender	Retenue
Veh_age	Retenue

Tableau 23- Statut des variables entrées dans le modèle de coût des sinistres

Les résultats de cette procédure de sélection des variables soulignent la fragilité des tests de corrélation effectués par paire de variables puisque le test de Student avait par exemple bien souligné l'influence du critère « genre » sur le coût mais celui de Kruskal-Wallis avait également souligné l'influence de la zone sur ce dernier alors que la variable ***area*** vient d'être supprimée du modèle (cf. Tableau 23).

On affiche ci-dessous les nouveaux coefficients de la régression (obtenus après la procédure de sélection des variables) :

```
> summary(Step)

Call:
glm(formula = (data$coûtUnitaire - 159.75) ~ data$gender + data$agecat +
    data$veh_age + data$VB_c + data$VV, family = Gamma(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5820  -1.5137  -0.7591   0.1411   5.2601

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.32031    0.42011  15.044  < 2e-16 ***
data$genderM     0.11086    0.05139   2.157 0.031038 *
data$agecat2    -0.22553    0.09414  -2.396 0.016631 *
data$agecat3    -0.26711    0.09187  -2.908 0.003661 **
data$agecat4    -0.34567    0.09184  -3.764 0.000170 ***
data$agecat5    -0.35455    0.10291  -3.445 0.000576 ***
data$agecat6    -0.35512    0.11723  -3.029 0.002467 **
data$veh_age2    0.07906    0.07755   1.020 0.308000
data$veh_age3    0.13517    0.08203   1.648 0.099478 .
data$veh_age4    0.11755    0.09839   1.195 0.232223
data$VB_cVB_c2   1.14585    0.39164   2.926 0.003454 **
data$VB_cVB_c3   1.28495    0.39593   3.245 0.001182 **
data$VB_cVB_c4   1.44878    0.40428   3.584 0.000343 ***
data$VVVV2      -0.12830    0.08189  -1.567 0.117224
data$VVVV3      -0.03802    0.09292  -0.409 0.682463
data$VVVV4      -0.20090    0.10981  -1.830 0.067394 .
data$VVVV5      -0.32339    0.12497  -2.588 0.009695 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 2.56074)

    Null deviance: 7506.2  on 4178  degrees of freedom
Residual deviance: 7321.7  on 4162  degrees of freedom
AIC: 68943
```

Illustration 9 – Résultat de la procédure de sélection des variables

Ces coefficients nous permettent de retrouver le coût moyen d'un sinistre pour tous les individus par classe tarifaire crée.

Cinq variables tarifaires ont été retenues et la fonction de lien utilisée est la fonction log, on a alors :

$$\log[E(C-d|C>d)] = \beta_0 + \beta_1 \times gender + \beta_2 \times agecat + \beta_3 \times veh\_age + \beta_4 \times VB\_c + \beta_5 \times VV$$

D'où un coût moyen donné par :

$$E(C|C>d) = \exp(\beta_0 + \beta_1 \times gender + \beta_2 \times agecat + \beta_3 \times veh\_age + \beta_4 \times VB\_c + \beta_5 \times VV) + d$$

Avec

- $(\beta_i)_{i=1,\dots,5}$ la valeur du coefficient de régression associé à la modalité prise par la $i^{ème}$ variable
- β_0 , la valeur du coefficient associée à l'individu de référence donnée par **Intercept**

Par exemple, pour un homme de **classe d'âge 1**, ayant un véhicule de classe d'âge 1, dont le type de véhicule appartient à la classe VB_c2 et la valeur du véhicule à VV1, le coût moyen d'un sinistre est donné par :

$$\mathrm{E(C|C>d)} = \exp(6{,}32031 + \mathbf{0}, \mathbf{11086} + +0 + 0 + 1{,}14585 + 0) + 159{,}75 = \ 2\,112{,}55\ €$$

Pour une femme présentant les mêmes caractéristiques on aura un coût moyen de :

$$\mathrm{E(C|C>d)} = \exp(6{,}32031 + \mathbf{0} + 0 + 0 + 1{,}14585 + 0) + 159{,}75 = \ 1\,907{,}63\ €$$

6.1.4 Etape 4 : validité du modèle

Cohérence des coefficients calculés

Signe des coefficients

Dans le cadre d'un modèle multiplicatif (fonction de lien log), étudier le signe d'un coefficient revient à étudier sa position par rapport à zéro, qui est le coefficient associé à la modalité de référence.

Soit c le coefficient associé à la modalité de l'individu.

Par croissance de la fonction exponentielle on a :

$$\begin{cases} \exp(c) > \exp(0) = 1 \ si\ c > 0 \\ \exp(c) < \exp(0) = 1\, si\ c < 0 \end{cases}$$

Par positivité des coûts, on peut établir la règle suivante :

- Si le coefficient est positif, le coût moyen pour la modalité considérée est supérieur au coût moyen associé à la modalité de référence ;
- A l'inverse, si le coefficient est négatif, le coût moyen pour la modalité considérée est inférieur au coût moyen de la modalité de référence.

Le modèle ajusté aux coûts respecte bien la règle imposée par les signes : toutes les modalités pour lesquelles le coefficient associé est positif ont un coût moyen de sinistre inférieur à la modalité de référence et inversement.

Par exemple, la modalité « H » correspondant aux hommes pour lesquels la variable « ***gender*** » admet un coefficient de régression positif (0,11679). Cela témoigne du fait que le coût moyen des sinistres pour hommes (1 534,06 €) est plus élevé que celui relevé chez les femmes (1 354,97 €).

Modalité	Coefficient	Coût moyen	
data$genderM	0,11086	1534,06 €	F : 1 354,97 €
data$agecat2	-0,22553	1 470,03 €	
data$agecat3	-0,26711	1 413,44 €	
data$agecat4	-0,34567	1 322,11 €	Agecat1 : 1 839,44 €
data$agecat5	-0,35455	1 295,57 €	
data$agecat6	**-0,35512**	**1 341,92 €**	
data$veh_age2	0,07906	1 348,15 €	
data$veh_age3	0,13517	1 487,09 €	Veh_age1 : 1 242,53 €
data$veh_age4	0,11755	1 570,75 €	
data$VB_cVB_c2	1,14585	1 326,05 €	
data$VB_cVB_c3	1,28495	1 535,83 €	VB_c1 : 534,44 €
data$VB_cVB_c4	1,44878	1 959,56 €	
data$VVVV2	-0,1283	1 392,56 €	
data$VVVV3	-0,03802	1 496,72 €	VV1 : 1 571,80 €
data$VVVV4	-0,2009	1 349,14 €	
data$VVVV5	-0,32339	1 176,04 €	

Tableau 24 – Coefficients estimés et coût moyen des sinistres par modalité

Si l'on s'intéresse à la progression des coefficients en parallèle de la progression des coûts moyens univariés, on remarque une adéquation quasi-parfaite: plus le coût augmente, plus le coefficient associé à la modalité est grand.

Seule la **modalité 6** associée à la variable ***age_cat*** déroge à la règle.

Ci-dessous, un exemple de parallèle effectué en triant par ordre croissant les coefficients de régression de la variable **VV** :

Modalité	Coefficient	Coût moyen
data$VVVV5	-0,32339	1 176,04 €
data$VVVV4	-0,2009	1 349,14 €
data$VVVV2	-0,1283	1 392,56 €
data$VVVV3	-0,03802	1 496,72 €

Tableau 25- Coefficients de régression et coût moyen par classe de valeurs de véhicule

Dans la troisième colonne du tableau, on retrouve bien nos coûts moyens triés par ordre croissant.

Valeur des coefficients

Nous venons d'analyser la cohérence du signe des coefficients, mais il faut également s'attarder sur leur valeur.

La fonction de lien utilisée étant la fonction log, nos valeurs seront (par construction) toutes supérieures à la franchise. On vérifie alors la cohérence de l'étendue des valeurs et de leur moyenne.

A l'aide des coefficients obtenus (cf. Tableau 24), et après les avoir appliqués à l'ensemble de notre portefeuille[1], on détermine les indicateurs de position pour le coût moyen des sinistres :

	Minimum	Maximum	Moyenne
Coût moyen	441,46 €	3185,87 €	1649,92 €

Tableau 26- Indicateurs de position pour le coût des sinistres

L'intervalle de coûts obtenu avec ce modèle est cohérent : la moyenne des coûts moyens par classe tarifaire (1 649,92 €) est très proche du coût moyen observé sur le portefeuille (1 649,85 €).

Remarque :

Les coefficients associés aux modalités de la variable VB_c (classes de type de véhicule) se détachent des autres coefficients : leur ordre de grandeur est supérieur.

Ceci s'explique par l'analyse univariée des coûts moyens pour cette variable : en passant de la première à la dernière modalité, on voit le coût moyen presque quadrupler.

On peut alors dire que notre regroupement des types de véhicule selon le coût moyen des sinistres est efficace puisqu'il permet d'établir une différence significative entre les classes.

Déviance du modèle

On s'intéresse à présent à la déviance Standardisée afin de contrôler la légitimité du modèle.

Nous devons comparer cette dernière au nombre de degrés de liberté (ddl) des résidus et vérifier que le rapport déviance Standardisée sur ddl n'est pas grand devant 1 (cf. § 3.2.4).

La déviance de notre modèle vaut 7 321,7 et le nombre de degré de liberté 4 162.
On doit également standardiser la déviance en la divisant par le paramètre de dispersion, ici $\phi = 2{,}56$.

On obtient alors une déviance Standardisée (2 859,21) inférieure au nombre de degrés de liberté : on peut donc admettre que le modèle est pertinent.

Conclusion

Nous avons établi un modèle pour le coût en utilisant la loi Gamma et la fonction de lien log. Les variables retenues concernent l'assuré ainsi que son véhicule :

- Pour l'assuré : son genre ainsi que sa classe d'âges
- Pour son véhicule : sa classe d'âges du véhicule, sa classe de type, sa classe de coût.

Les résultats de validité du modèle sont concluants.

[1] Cette application a nécessité l'utilisation de la fonction ***predict()*** sous R

6.2 Modèle pour la fréquence

La seconde variable que l'on cherche à expliquer est la fréquence de sinistres.
Comme on s'intéresse à un montant de Prime Pure annuelle, notre fréquence doit également être exprimée à l'échelle d'une année.

Cette dernière est donc définie par :

$$fréquence = \frac{numClaims}{exposure}$$

Avec

- ***numClaims*** le nombre de sinistres observés durant la période d'exposition ;
- ***exposure*** la durée d'exposition exprimée en année.

On rappelle que l'on ne modélise que la fréquence des sinistres dont le coût est différent du montant de la franchise, ces derniers étant traités à part.

Dans la mise en place de ce modèle, certaines étapes sont similaires à celles réalisées pour le modèle d'estimation du coût et ne seront donc pas détaillées (les résultats seront tout de même présentés).
Le lecteur est invité à se référer au paragraphe précédent pour tout besoin d'explications supplémentaires.

6.2.1 Etape 1 : le choix du modèle

Deux approches sont généralement utilisées dans la modélisation des fréquences de sinistre :

- **Cas classique :** Utilisation de la loi de Poisson
- **Cas de sur-dispersion de données :** Utilisation de la loi Binomiale Négative

La fonction de lien est alors la fonction log, pour avoir un modèle multiplicatif et éviter de construire un modèle qui renvoie des valeurs négatives.

Remarque :
La sur-dispersion représente l'hétérogénéité qui peut exister au sein d'un même groupe d'individus malgré le travail d'analyse de données fait sur les variables explicatives: elle est due à une mauvaise segmentation à cause de l'indisponibilité de certaines variables pour l'assureur (conducteur réel du véhicule, agressivité…).

La loi de Poisson et la loi Binomiale Négative modélisent des processus de comptage et sont donc bien adaptés à la représentation de la variable ***numClaim*** (prenant ses valeurs sur [[0,4]]).
Elles ne sont cependant pas adaptées pour modéliser de façon directe la variable ***fréquence*** car cette dernière prend ses valeurs sur $\mathbb{R}^+$ et non sur $\mathbb{N}$.

Bien que l'on cherche à déterminer l'espérance de la fréquence, on utilisera la loi de comptage sur la variable ***numClaims,*** qui correspond au nombre de sinistres relevés durant la période d'observation.
Pour revenir au résultat attendu, c'est-à-dire l'estimation de la fréquence, le log de l'exposition de chacune des polices sera déclarée en variable « **offset** ».

Une variable est déclarée en offset si la variable à expliquer dépend linéairement de cette variable : la valeur du coefficient qui la précède est alors fixé à 1 et n'est donc pas estimé.

En utilisant la fonction de lien log appliquée à la variable à expliquer ***numClaims*** on a alors la relation :

$$\log[\mathrm{E}(\mathrm{numClaims}|\mathrm{C} > \mathrm{d})] = \beta_0 + \beta_1 \mathrm{x}^1 + \cdots + \beta_\mathrm{p} \mathrm{x}^\mathrm{p} + \mathbf{1} \times \log(\mathrm{exposure})$$

$$\Leftrightarrow$$

$$\mathrm{E}(\mathrm{numClaims}|\mathrm{C} > \mathrm{d}) = \mathrm{exposure} \times \exp(\beta_0 + \beta_1 \mathrm{x}^1 + \cdots + \beta_\mathrm{p} \mathrm{x}^\mathrm{p})$$

$$\Leftrightarrow$$

$$\mathrm{E}\left[\frac{\mathrm{numClaims}}{\mathrm{exposure}}|\mathrm{C} > \mathrm{d}\right] = \mathrm{E}[\mathrm{fréquence}|\mathrm{C} > \mathrm{d}] = \exp(\beta_0 + \beta_1 \mathrm{x}^1 + \cdots + \beta_\mathrm{p} \mathrm{x}^\mathrm{p})$$

Il faut à présent faire un choix entre la loi de Poisson et la loi Binomiale Négative.

Critère Espérance-Variance

L'un des critères de choix de loi est basé sur les moments de la variable de comptage : c'est le critère espérance-variance.

Soit N la variable de comptage dont on cherche la loi, on choisira :

- La loi de Poisson si $\mathrm{E}(\mathrm{N}) = \mathrm{Var}(\mathrm{N})$
- la loi Binomiale Négative si $\mathrm{E}(\mathrm{N}) < \mathrm{Var}(\mathrm{N})$

	Espérance	Variance
numClaims	0,062	0,068

Tableau 27- Espérance et variance de la variable numClaims

D'après ce critère, on devrait choisir la loi Binomiale Négative pour estimer notre fréquence de sinistres.

Test du Chi-2 d'adéquation

Le test du Chi-2 d'adéquation est un test non paramétrique : un grand avantage des procédures non paramétriques est qu'elles peuvent s'appliquer à des échantillons de très petite taille et qu'elles n'exigent aucune condition d'applicabilité particulière si ce n'est l'indépendance des mesures.
Cette dernière condition étant l'une des hypothèses de notre modèle (cf. § 1.2.2), nous pouvons l'utiliser.

Ce test permet de déterminer si un échantillon est distribué selon une loi, choisie par nos soins[1].

Les hypothèses du test sont les suivantes :

- Hypothèse nulle (H_0) : les données sont distribuées selon la loi indiquée
- Hypothèse alternative (H_1) : les données ne sont pas distribuées selon la loi indiquée

En pratique, ce test retourne une p-value qui peut être interprétée de la façon suivante :

- si p-value > α: la distribution des données est bien celle de la loi indiquée
- si p-value < α: la distribution des données n'est pas celle de la loi indiquée

[1] Pour en savoir plus sur le fonctionnement de ce test, cf. [S12]

où α correspond au seuil de tolérance (fixé) du test.

Ce test a été réalisé à l'aide de la fonction ***goodfit()*** sous R : elle renvoie la valeur de la statistique de test du Khi-2 ainsi que la p-value associée au test.
Nous indiquons ici que la méthode de décision à utiliser pour juger de l'adéquation ou non à la loi renseignée est celle qui minimise la valeur du Khi-2 (*Minimum Chi-squared*).

Une vision graphique de l'adéquation avec chacune des lois testées est disponible avec la fonction ***plot()*** : elle permet de retourner sur un même graphique les valeurs observées (en gris) et les valeurs prévues par la distribution (en rouge) de la loi avec laquelle on teste l'adéquation.

```
library(vcd)
gf=goodfit(data$fréquence,type="poisson",method="MinChisq")
plot(gf)

library(vcd)
gf=goodfit(data$fréquence,type="nbinomial",method="MinChisq")
plot(gf)
```

```
         Goodness-of-fit test for poisson distribution

                    X^2  df P(> X^2)
Pearson 2.920451e+104 364         0
```

```
         Goodness-of-fit test for nbinomial distribution

              X^2  df  P(> X^2)
Pearson 292.728 363 0.9972243
```

Illustration 10 – Test d'adéquation à la loi de Poisson (en haut) et à la loi Binomiale Négative (en bas)

Aussi bien visuellement qu'après interprétation de la p-value retournée par les deux tests, c'est sans hésitation que nous choisissons la loi Binomiale négative pour modéliser la fréquence de sinistre.

En effet, la p-value du test (0,997) est largement supérieure à 0,05 et les valeurs observées et théoriques semblent assez proche sur le graphique (cf. Illustration 11).

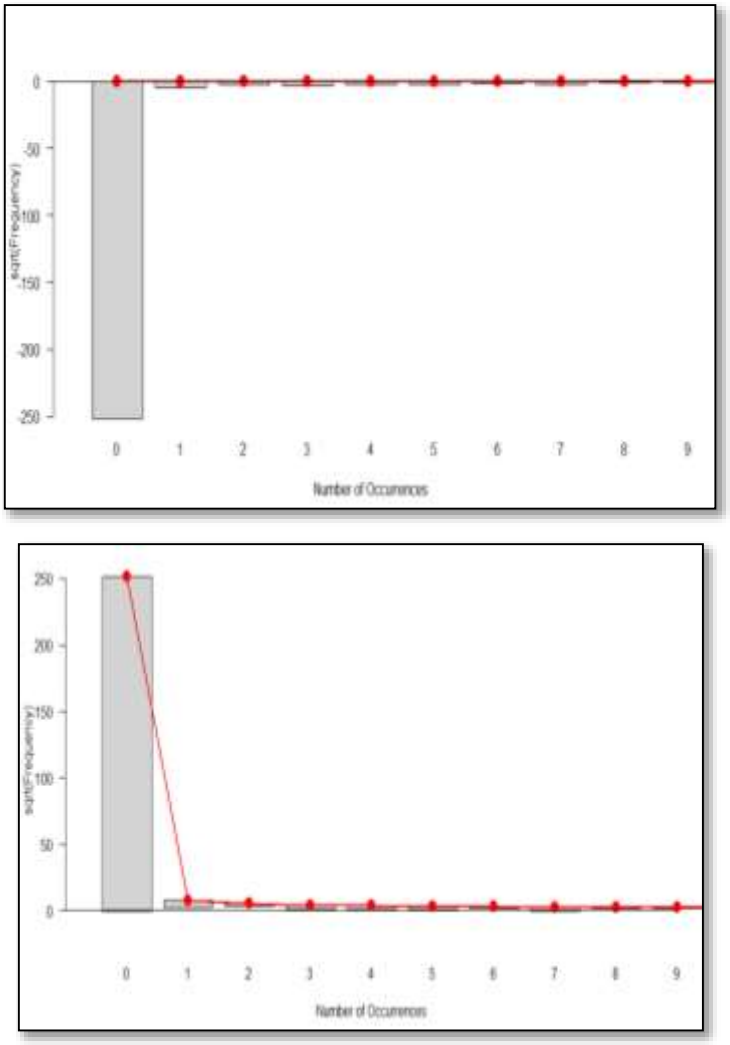


Illustration 11- Adéquation à la loi de Poisson (en haut) et à la loi Binomiale Négative (en bas)

Critères AIC / BIC

Les critères AIC et BIC sont généralement utilisés dans les procédures de sélection des variables mais peuvent également être utilisés dans le cadre de la sélection de la loi suivie par les données.

Loi	AIC	BIC
Poisson	30 716	30 916,02
Binomiale Négative	30 616	30 825,44

Tableau 28- AIC et BIC du modèle incluant toutes les variables pour les deux lois envisagées

Là encore, c'est la loi Binomiale Négative qui admet les plus faibles valeurs d'AIC et de BIC : elle sera donc utilisée pour modéliser la fréquence de sinistre.

6.2.2 Etape 2 : Estimation des coefficients de la régression

Après le choix de la loi suivie par notre fréquence de sinistre, c'est au tour des coefficients attribués à chaque modalité de variable d'être estimés.

L'estimation est réalisée à l'aide de la fonction ***glm.nb()*** disponible sous R.

La sélection des variables se faisant par la suite, on commence par intégrer l'ensemble des variables dans le modèle, à savoir :

Intitulé	Signification	Nombre de modalités
VV	Classe de valeurs du véhicule	5
VB_f	Classe de types de véhicule	4
Agecat	Classe d'âges du conducteur	6
area	Zone géographique	6
gender	Genre de l'assuré	2
Veh_age	Classe d'âges du véhicule	4

Tableau 29 - Les variables utilisées dans le modèle complet

Il ne faut également pas oublier d'intégrer l'exposition comme variable **offset**.

```
regBNlog=glm.nb(data$numclaims~data$gender+data$agecat+data$veh_age+data$VB_f+
data$area+data$VV+offset(log(data$exposure)))
summary(regBNlog)
```

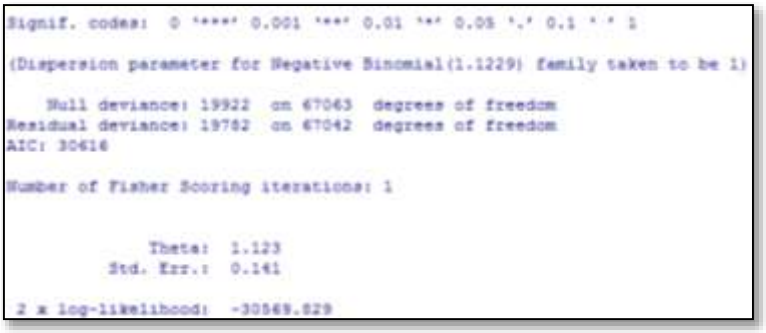
```
> summary(regBNlog)

Call:
glm.nb(formula = data$numclaims ~ data$gender + data$agecat +
    data$veh_age + data$VB_f + data$area + data$VV + offset(log(data$exposure)),
    init.theta = 1.122911158, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8919  -0.4124  -0.3191  -0.2084   3.9958

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.58609    0.60035  -4.308 1.65e-05 ***
data$genderM   -0.03325    0.03331  -0.998 0.318156
data$agecat2   -0.19564    0.06124  -3.195 0.001399 **
data$agecat3   -0.24562    0.05955  -4.125 3.71e-05 ***
data$agecat4   -0.26790    0.05947  -4.504 6.65e-06 ***
data$agecat5   -0.48653    0.06643  -7.324 2.41e-13 ***
data$agecat6   -0.44646    0.07566  -5.901 3.62e-09 ***
data$veh_age2   0.10145    0.05052   2.008 0.044614 *
data$veh_age3   0.07608    0.05437   1.399 0.161725
data$veh_age4   0.12201    0.06831   1.786 0.074087 .
data$VB_fVB_f2  0.54063    0.59231   0.913 0.361374
data$VB_fVB_f3  0.56684    0.59327   0.955 0.339351
data$VB_fVB_f4  1.60471    0.71097   2.257 0.024003 *
data$areaB      0.05108    0.04913   1.040 0.298446
data$areaC      0.02516    0.04448   0.566 0.571651
data$areaD     -0.04413    0.05844  -0.755 0.450210
data$areaE      0.04334    0.06319   0.686 0.492801
data$areaF      0.14271    0.07199   1.982 0.047453 *
data$VVVV2      0.16242    0.05410   3.002 0.002680 **
data$VVVV3      0.23164    0.06263   3.699 0.000217 ***
data$VVVV4      0.35480    0.07825   4.534 5.79e-06 ***
data$VVVV5      0.41304    0.09219   4.480 7.45e-06 ***
---
```

Illustration 12 – Coefficients de la régression retournés par R



Nous remarquons que le coefficient associé à la variable offset « ***exposure*** » ne figure pas dans la liste des coefficients retournés : il a bien été fixé à 1.

De plus, l'AIC de ce modèle est de 30 616, nous allons tenter de le réduire en appliquant au modèle une procédure de sélection des variables.

6.2.3 Etape 3 : Procédure de sélection des variables

On utilise la procédure descendante (« *Backward* ») comme méthodes de sélection conjointe des variables. Elle nous servira à éliminer les variables les moins significatives.

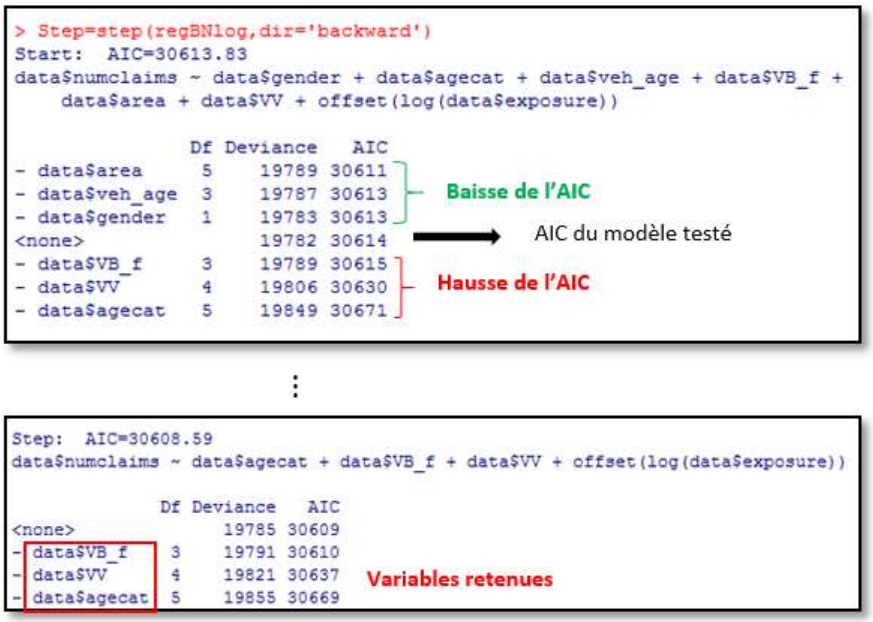


Illustration 13 – Procédure de sélection des variables du modèle

Après l'application de la procédure descendante, les variables ***gender, area,*** ainsi que la variable ***veh_age*** ne font plus partie du modèle d'estimation de la fréquence.

Variable	Retenue/ Rejetée
VV	Retenue
VB_f	Retenue
Agecat	Retenue
area	Rejetée
gender	Rejetée
Veh_age	Rejetée

Tableau 30 - Statut des variables entrées dans le modèle de fréquence de sinistre

De nouveau, ces résultats soulignent la fragilité des tests de corrélation effectués **par paire de variables.** En effet le test de Student avait par exemple bien prédit que le critère genre n'influerait pas sur la fréquence (cf. § 5.2.2) mais celui de Kruskal-Wallis avait souligné l'influence de l'âge du véhicule sur cette dernière alors que la variable ***veh_age*** vient d'être supprimée du modèle (cf. Tableau 30).

On remarque également que les variables retenues dans le modèle mis en place pour le coût sont différentes de celles retenues dans le modèle de fréquence, d'où l'intérêt d'avoir séparé l'étude de la sinistralité en deux, de façon à repérer les variables les plus influentes sur chacune des variables à expliquer.

Nous affichons alors les coefficients finaux de la régression:

```
> summary(Step)

Call:
glm.nb(formula = data$numclaims ~ data$agecat + data$VB_f + data$VV +
    offset(log(data$exposure)), init.theta = 1.117932537, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8701  -0.4132  -0.3199  -0.2090   3.9686

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.45083    0.59478  -4.121 3.78e-05 ***
data$agecat2   -0.18633    0.06113  -3.048  0.00230 **
data$agecat3   -0.24006    0.05940  -4.041 5.32e-05 ***
data$agecat4   -0.26658    0.05937  -4.490 7.11e-06 ***
data$agecat5   -0.48856    0.06627  -7.372 1.68e-13 ***
data$agecat6   -0.45337    0.07543  -6.010 1.85e-09 ***
data$VB_fVB_f2  0.54266    0.59206   0.917  0.35937
data$VB_fVB_f3  0.54381    0.59265   0.918  0.35883
data$VB_fVB_f4  1.62548    0.71059   2.288  0.02217 *
data$VVVV2      0.13622    0.04638   2.937  0.00331 **
data$VVVV3      0.19882    0.04884   4.071 4.68e-05 ***
data$VVVV4      0.29515    0.05875   5.024 5.06e-07 ***
data$VVVV5      0.34126    0.06699   5.094 3.50e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.1179) family taken to be 1)

    Null deviance: 19912  on 67063  degrees of freedom
Residual deviance: 19784  on 67051  degrees of freedom
AIC: 30611

Number of Fisher Scoring iterations: 1

                          Theta:  1.118
                      Std. Err.:  0.140

                 2 x log-likelihood:  -30582.591
```

Illustration 14 – Coefficients de la régression obtenus après sélection des variables

Ces coefficients nous permettent de retrouver la fréquence moyenne de tous les individus par classe tarifaire crée.
Trois variables tarifaires ont été retenues et la fonction de lien utilisée est la fonction log, on a alors :

$$\log[E(numClaims|C > d)] = \beta_0 + \beta_1 \times agecat + \beta_2 \times VB\_f + \beta_3 \times VV + \mathbf{1} \times \log(exposure)$$

$$\Leftrightarrow$$

$$E(numClaims|C > d) = exposure \times \exp(\beta_0 + \beta_1 \times agecat + \beta_2 \times VB\_f + \beta_3 \times VV)$$

Et une fréquence annuelle moyenne (obtenue en posant $exposure = 1$) donnée par :

$$\boxed{E(fréquence|C > d) = \exp(\beta_0 + \beta_1 \times agecat + \beta_2 \times VB\_f + \beta_3 \times VV)}$$

Avec

- $(\beta_i)_{i=1,\dots,3}$ la valeur du coefficient de régression associé à la modalité prise par la $i^{ème}$ variable.
- β_0 , la valeur du coefficient associée à l'individu de référence tel que :

$$E\left(fréquence_{ref}|C > d\right) = \exp(\beta_0)$$

Par exemple, pour un individu de **classe d'âge 1**, ayant un véhicule appartenant à VB_f2 et dont la valeur du véhicule appartient à VV2, la fréquence moyenne annuelle de sinistre est donnée par :

$$E(fréquence|C > d) = \exp(-2{,}45083 + \mathbf{0} + 0{,}54266 + 0{,}13622) = \ 17\ \%$$

Pour un individu présentant les mêmes caractéristiques mais appartenant à la **classe d'âge 2** on aura :

$$E(fréquence|C > d) = \exp(-2{,}45083 - 0{,}18633 + 0{,}54266 + 0{,}13622) = \ 14{,}11\ \%$$

6.2.4 Etape 4 : validité du modèle

Cohérence des coefficients calculés

Signe des coefficients

De même que précédemment (cf. § 6.1.4) par positivité des coûts on peut établir la règle suivante :

- Si le coefficient est positif, la fréquence moyenne pour la modalité considérée est supérieure à la fréquence de la modalité de référence ;
- A l'inverse, si le coefficient est négatif, la fréquence pour la modalité considérée est inférieure à la fréquence de la modalité de référence.

Modalité	Coefficient	Fréquence moyenne	
data$agecat2	-0,18633	20,05 %	
data$agecat3	-0,24006	20,90 %	
data$agecat4	-0,26658	21,58 %	
data$agecat5	-0,48856	17,48 %	
data$agecat6	-0,45337	15,76 %	agecat1 : 37,64 %
data$VB_fVB_f2	0,54266	17,23 %	
data$VB_fVB_f3	0,54381	23,50 %	
data$VB_fVB_f4	1,62548	38,63 %	VB_f1 : 6,31 %
data$VVVV2	0,13622	20,81 %	
data$VVVV3	0,19882	20,35 %	
data$VVVV4	0,29515	20,90 %	
data$VVVV5	0,34126	18,17 %	VV1 : 24,16 %

Tableau 31- Coefficients de la régression et fréquence moyenne de sinistre par modalité

Pour chacune des variables retenues, la règle des signes des coefficients est vérifiée :

- L'ensemble des coefficients obtenus pour la variable ***agecat*** sont négatifs, preuve que les individus appartenant à la classe d'âge 1 ont une fréquence de sinistre plus élevée que les autres, toute modalité égale par ailleurs.

- A l'inverse, pour les variables ***VB_f*** et ***VV***, nos coefficients sont tous positifs, preuve que les individus dont le véhicule n'appartient pas à la classe VB_f1 (respectivement VV1) ont une fréquence de sinistre plus élevée que les autres, toute modalité égale par ailleurs.

Si l'on revient à notre première variables explicatives, ***agecat***, et que l'on s'intéresse à la progression des coefficients en parallèle de la progression des fréquences univariées, on remarque que l'adéquation n'est pas bonne : la règle « plus la fréquence diminue, plus le coefficient associé à la modalité est faible » n'est pas vérifiée.

On soulève ici l'intérêt du GLM : la prise en compte de TOUTES les modalités à la fois.
En effet, les fréquences indiquées ont été observées en segmentant le portefeuille par rapport à une unique variable tandis que les coefficients du GLM tiennent compte d'une segmentation bien plus fine (croisement de trois variables dans le cas présent).

On ajoute également que lorsque pour une variable les coefficients obtenus pour chacune des modalités par le GLM suivent l'ordre des résultats obtenus par analyse descriptive univariée, cela indique que la variable en question a un poids important dans la segmentation : ce sont les coefficients des autres variables qui vont être « influencés » par les siens. C'est uniquement le cas de la variable ***VB_f***.

Valeur des coefficients

Nous avons analysé la cohérence du signe des coefficients, mais il faut également s'attarder sur leur valeur.

En effet, un modèle d'estimation de la fréquence qui renverrait des valeurs hors de l'intervalle [0,1] ne peut être qualifié de « bon modèle ».

A l'aide des coefficients obtenus (cf. Tableau 31), et après les avoir appliqués à l'ensemble de notre portefeuille, on détermine les indicateurs de position de la fréquence annuelle de sinistre :

	Minimum	Maximum	Moyenne
Fréquence annuelle	5,28 %	61,6 %	6,25 %

Tableau 32– Etendue et moyenne des valeurs obtenues pour la fréquence

L'intervalle de fréquence obtenu est cohérent avec l'analyse descriptive réalisée (cf. Tableau 19) et la fréquence moyenne sur l'ensemble du portefeuille est très proche de la fréquence moyenne observée sur le portefeuille (6,26 %).

Déviance du modèle

La légitimité du modèle peut être contrôlée par le biais de l'étude de la déviance Standardisée du modèle. Un modèle est pertinent si le rapport de la déviance Standardisée sur le degré de liberté des résidus n'est pas grand devant 1 (cf. § 3.2.4).

La déviance de notre modèle vaut 19 784, et le nombre de degré de liberté 67 051.
Pour obtenir la déviance Standardisée, il faut diviser cette déviance par le paramètre de dispersion, ici $\phi = 1$.

Notre déviance Standardisée (19 784) étant inférieure au nombre de degrés de liberté : on peut donc admettre que le modèle est pertinent.

Conclusion

Nous avons établi un modèle pour la fréquence en utilisant la loi Binomiale Négative et la fonction de lien log. Les variables retenues concernent l'assuré ainsi que son véhicule :

- Pour l'assuré : sa classe d'âges
- Pour son véhicule : sa classe de types, sa classe de coûts.

Les résultats de validité du modèle sont concluants.

6.3 Comparaison tarifaire Homme/Femme

La combinaison des résultats des deux Modèles Linéaires Généralisés (coût et fréquence) permet de déterminer une Prime Pure pour chacune des classes tarifaires crées. Ces dernières étant distinctes pour les deux genres, nous obtenons un tarif différencié.

On constate déjà qu'il y a un nombre assez conséquent de variables retenues et que le nombre de primes est considérable.
En effet, pour le modèle de fréquence **SG**, trois variables ont été retenues, soit 120[1] fréquences possibles.
Quant au coût, cinq variables ont été retenues, ce qui représente 960 coûts possibles.

L'obtention d'un nombre important de classes tarifaires est principalement due au manque de détail des variables de classe.
Si par exemple la variable ***veh_body***, correspondant au type de véhicule était subdivisée en sous-groupes (détail du modèle de véhicule), on aurait probablement pu éliminer la variable **VV**, correspondant à la classe de valeurs du véhicule puisque les sous-groupes en tiendraient compte indirectement.

Face à cela, nous ne pouvons présenter l'ensemble des primes par classe tarifaire. Nous choisissons dans ce mémoire de les présenter par classe d'âge.
Ce choix est motivé par l'importance de ce critère dans la tarification mais aussi par la Directive Anti-discrimination qui vise à le faire disparaitre des critères autorisés pour la segmentation.

Si l'on s'intéresse à la Prime Pure moyenne payée par classe d'âge et par genre, on constate facilement que le montant des Primes Pures chez les hommes sont plus élevés que ceux des femmes, et ceci quel que soit leur âge.

Ceci ne signifie bien entendu pas que TOUS les hommes paient plus cher que TOUTES les femmes puisqu'il s'agit ici d'une **moyenne**.

Classe d'âges / Genre	Homme	Femme	Ecart H/F
1	183,60 €	157,60 €	**16,50 %**
2	122,35 €	107,26 €	14,08 %
3	113,67 €	100,25 €	13,39 %
4	104,55 €	90,19 €	15,92 %
5	83,02 €	73,60 €	12,79 %
6	82,21 €	75,05 €	9,54 %

Tableau 33 - Ecarts de primes H/F par classe d'âge modèle B

Cependant, cette observation doit être nuancée : les écarts tarifaires tendent à s'amoindrir au fil des années avec un passage d'un écart de l'ordre de 17 % pour les jeunes conducteurs à environ 9,5 % pour les plus âgés (cf. Ecart H/F Tableau 33).

[1] $VV_{(5)} \times VB\_f_{(4)} \times agecat_{(6)}$

Chapitre 7 - La suppression totale du critère genre dans la cotation (Modèle SG)

Ce septième chapitre nous conduit à l'élaboration d'un modèle de tarification qui ne tient plus compte (de quelque manière que ce soit) du genre de l'individu.

Rappelons que dans le cas où un assureur est amené à supprimer une variable tarifaire de son modèle de tarification, c'est en premier lieu la segmentation de son portefeuille qui sera modifiée.
La qualité de la nouvelle segmentation dépend alors de l'importance de la variable supprimée dans la formation des classes tarifaires mais aussi de la capacité de l'assureur à remplacer cette information par une autre information qui influe également sur la sinistralité de l'assuré.

Dans le cas présent, nous nous contenterons de supprimer la variable « genre » sans tenter de la remplacer par une autre information.

7.1 Modèle pour le coût

La variable « ***gender*** » faisant référence au genre de l'individu a été retenue dans le modèle de coût de base : on s'attend donc à ce que sa suppression de la modélisation ait un effet sur les résultats obtenus pour le nouveau modèle d'estimation des coûts.

Le lecteur averti pourra s'il le souhaite se contenter d'observer les résultats fournis par les différentes illustrations (tableaux et sorties R) pour cette partie.

7.1.1 Etape 1 : le choix du modèle

La variable à expliquer est le coût des sinistres : pour le modéliser, nous utiliserons une loi Gamma et la fonction de lien log afin d'obtenir un modèle multiplicatif.

Ce choix est validé par la comparaison des critères de sélection AIC et BIC : les plus faibles valeurs sont obtenues avec l'utilisation de la fonction de lien « **log** ».

Loi	AIC	BIC
Gamma inverse	68 953	69 092,91
Gamma log	68 947	69 086,73

Tableau 34- AIC et BIC du modèle avec la loi Gamma pour les fonctions de lien log et inverse

Rappelons ici que nous avons, préalablement à l'application du modèle[1] :

- soustrait le montant de la franchise à tous nos coûts de sinistre de façon à respecter les bornes du support de la loi Gamma ($[0, +\infty[$)
- supprimé les valeurs nulles (correspondant aux montants égaux à la franchise) pour respecter le caractère continu de la densité de la loi Gamma qui l'empêche d'accorder un poids trop important à une seule et même valeur.

[1] Cf. § 6.1.1 pour plus de détails.

7.1.2 Etape 2 et 3 : Estimation des coefficients de la régression et procédure de sélection des variables

On doit à présent estimer les coefficients attribués à chaque modalité de variable.

On exécute alors les trois actions suivantes afin de les déterminer :

a. Détermination des coefficients de la régression comportant l'ensemble des variables avec la fonction ***glm() ;***

b. Application de la procédure descendante pour éliminer les variables non significatives ;

c. Résumé des coefficients de la régression associés aux modalités des variables finalement retenues.

Variable	Signification	Nombre de modalités	Retenue/Rejetée
VV	Classe de valeurs du véhicule	5	Retenue
VB_c	Classe de types de véhicule	4	Retenue
Agecat	Classe d'âges du conducteur	6	Retenue
area	Zone géographique	6	Rejetée
Veh_age	Classe d'âges du véhicule	4	Retenue

Tableau 35- Résultat de la procédure de sélection des variables du modèle SG

Après sélection des variables, on remarque que la variable ***area***, faisant référence à la zone géographique, n'a pas été retenue, tout comme dans le modèle de base.
On se retrouve donc avec un modèle **SG** où toutes les variables retenues sont aussi présentes dans le modèle B hormis le genre (par construction).

```
> summary(Step)

Call:
glm(formula = (data$coûtUnitaire - 159.75) ~ data$agecat + data$veh_age +
    data$VB_c + data$VV, family = Gamma(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5990  -1.5182  -0.7681   0.1371   5.1001

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.375371   0.420419  15.164  < 2e-16 ***
data$agecat2   -0.233852   0.094275  -2.481 0.013158 *
data$agecat3   -0.283900   0.091902  -3.089 0.002021 **
data$agecat4   -0.353826   0.091979  -3.847 0.000121 ***
data$agecat5   -0.359963   0.103056  -3.493 0.000483 ***
data$agecat6   -0.348667   0.117339  -2.971 0.002981 **
data$veh_age2   0.073906   0.077656   0.952 0.341301
data$veh_age3   0.144875   0.081933   1.768 0.077099 .
data$veh_age4   0.142452   0.097582   1.460 0.144413
data$VB_cVB_c2  1.116113   0.392139   2.846 0.004446 **
data$VB_cVB_c3  1.258071   0.396458   3.173 0.001518 **
data$VB_cVB_c4  1.435224   0.404918   3.544 0.000398 ***
data$VVVV2     -0.112430   0.081759  -1.375 0.169162
data$VVVV3     -0.009251   0.092437  -0.100 0.920288
data$VVVV4     -0.163046   0.108714  -1.500 0.133751
data$VVVV5     -0.280967   0.123707  -2.271 0.023184 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 2.568923)

    Null deviance: 7506.2  on 4178  degrees of freedom
Residual deviance: 7333.6  on 4163  degrees of freedom
AIC: 68949
```

Illustration 15 - Sortie R des coefficients de la régression du modèle de coût des sinistres

Les valeurs des coefficients associées aux différentes modalités des variables retenues sont cependant différentes de celles obtenues précédemment (cf. § 6.1.3): l'impact du genre de l'individu doit être « compensé » par d'autres variables explicatives retenues dans le modèle.

7.1.3 Etape 4 : validité du modèle

Cohérence des coefficients calculés

Signe des coefficients

Pour éviter d'avoir à refaire une analyse du signe et de l'ordre des coefficients, nous juxtaposons les résultats obtenus avec le modèle B à ceux obtenus pour le modèle **SG** (cf. Tableau 36).

En rouge figurent les coefficients négatifs et en vert les positifs : signes et ordres inter-modalités sont identiques.

Nous invitons alors le lecteur à se référer au paragraphe 6.1.4 pour de plus amples détails sur l'analyse du signe des coefficients.

Modalité	Modèle SG	Modèle B
data$genderM	-	0,11086
data$agecat2	-0,233852	-0,22553
data$agecat3	-0,2839	-0,26711
data$agecat4	-0,353826	-0,34567
data$agecat5	-0,359963	-0,35455
data$agecat6	-0,348667	-0,35512
data$veh_age2	0,073906	0,07906
data$veh_age3	0,144875	0,13517
data$veh_age4	0,142452	0,11755
data$VB_cVB_c2	1,116113	1,14585
data$VB_cVB_c3	1,258071	1,28495
data$VB_cVB_c4	1,435224	1,44878
data$VVVV2	-0,11243	-0,1283
data$VVVV3	-0,009251	-0,03802
data$VVVV4	-0,163046	-0,2009
data$VVVV5	-0,280967	-0,32339

Tableau 36 - Coefficients de régression des modèles B et SG

Valeur des coefficients

Intéressons-nous à présent aux indicateurs de position des coûts théoriques retournés par le modèle. Les valeurs obtenues dans le modèle de base (B) seront rappelées en guise d'outil comparatif.

	Minimum	Maximum	Moyenne
Coût moyen SG	468,84 €	3 010,63 €	1 648,90 €
Coût moyen B	441,46 €	3 185,87 €	1 649,92 €
Coût moyen portefeuille	159,8 €	21 173,19 €	1 649,85 €

Tableau 37- Indicateurs de position pour le coût moyen théorique des sinistres

L'intervalle de coûts obtenu est cohérent : la moyenne des coûts reste toujours très proche du coût moyen observé sur le portefeuille (1 649,85 €).

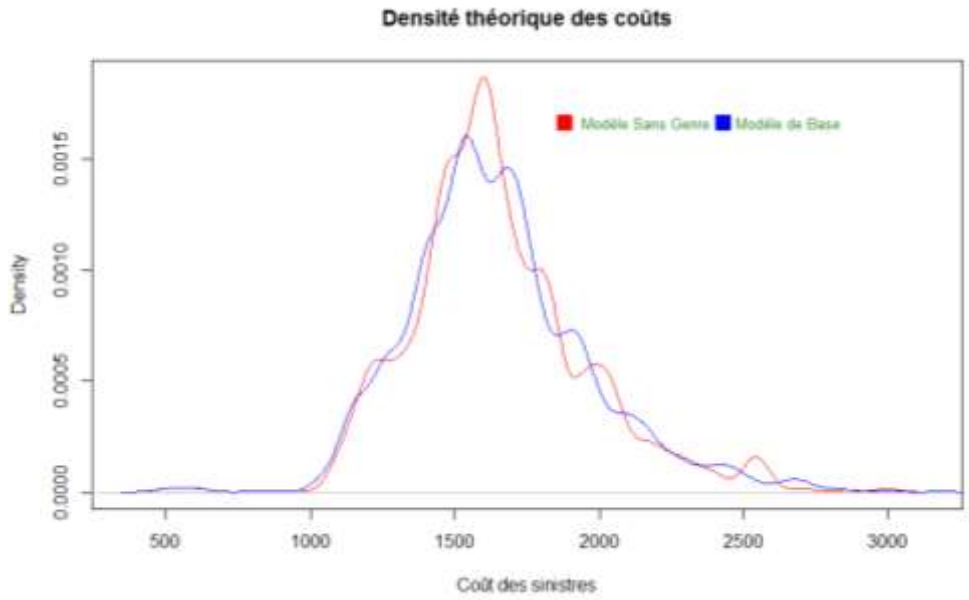
Si l'on s'intéresse à présent à nos deux modèles, le modèle **S**ans **G**enre retourne :

- un coût moyen (1 648,90 €) plus éloigné du coût moyen observé directement sur le portefeuille (1 649,85 €) que le modèle de base (1 649,92 €).
- un coût minimum (468,84 €) plus élevé que le coût minimum observé sur le portefeuille (159,8 €)

- un coût maximum (3 010,63 €) bien plus faible que le coût maximum relevé sur le portefeuille (21 173,19 €)

L'écart entre le modèle **S**ans **G**enre et les observations réalisées sur le portefeuille relevé par les deux derniers points est représentatif de l'effet des modèles tarifaires : la mutualisation du risque.
Les valeurs observées sur le portefeuille ne présentent aucune mutualisation : à chaque police est affecté un montant de sinistre, sans aucun partage de risque avec un quelconque autre assuré.
Par contre, dans les deux modèles, la création de classes tarifaires induit le partage du risque au sein de ces dernières.

Le tableau 37 laisse transparaitre une divergence dans la répartition de la charge de sinistres sur l'ensemble du portefeuille pour chacun des deux modèles mis en place, divergence affirmée par la représentation des densités théoriques des coûts obtenus.



Graphique 17 – Densité des coûts des sinistres pour le modèle de Base et Sans Genre

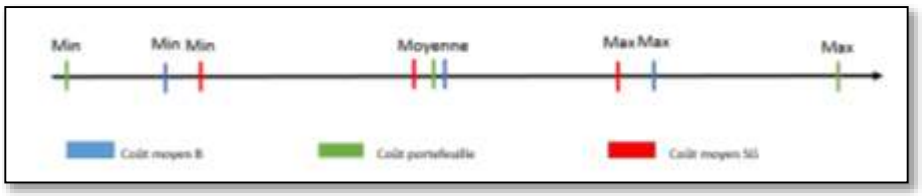


Figure 8 – Indices de position des coûts (schéma non à l'échelle)

La densité théorique des coûts obtenue avec le modèle **S**ans **G**enre présente un poids plus important autour de la moyenne que celle du modèle de **B**ase : la mutualisation du risque est réalisée sur un segment de coût moins important, traduisant une perte de finesse de segmentation due à la suppression du critère « **genre »**.

On observe également que les queues de distribution des coûts théoriques du modèle **B** sont plus épaisses que celles du modèle **S**ans **G**enre.

Ces queues de distribution traduisent la capacité du modèle à produire des classes tarifaires homogènes : plus elles sont épaisses, plus les classes tarifaires crées sont homogènes car le modèle arrive à se rapprocher d'autant plus de l'étendue des coûts individuels relevés sur le portefeuille.

Déviance du modèle

La déviance standardisée permet de contrôler la légitimité du modèle.

Nous devons comparer cette dernière au nombre de degrés de liberté (ddl) des résidus et vérifier que le rapport déviance Standardisée sur ddl n'est pas grand devant 1.

La déviance de notre modèle et son degré de liberté valent respectivement 7 333,6 et 4 163.
On doit également standardiser la déviance en la divisant par le paramètre de dispersion ϕ=2,57.

On obtient alors une déviance Standardisée inférieure au nombre de degrés de liberté (2 854,7) : on peut donc admettre que le modèle est pertinent.

Conclusion :
Nous avons établi un modèle pour le coût en utilisant la loi Gamma et la fonction de lien log. Les variables retenues concernent l'assuré ainsi que son véhicule :

- Pour l'assuré : sa classe d'âges
- Pour son véhicule : sa classe d'âges, sa classe de types, sa classe de coûts.

Les résultats de validité du modèle sont concluants.

7.2 Modèle pour la fréquence

Bien que le genre de l'assuré ne soit pas une variable retenue dans le modèle de fréquence de **B**ase, il est possible que la suppression d'une variable modifie la loi suivie par les données. Nous allons donc nous assurer que ceci n'est pas le cas pour notre portefeuille.

Mis à part cela, il se peut également que la variable « genre » n'ait pas été retenue dans le modèle précèdent car une autre variable apportait une information semblable, le fait de ne pas la renseigner pourrait alors modifier la significativité des autres facteurs.

Les quatre étapes de la mise en place du GLM ne seront pas détaillées car le modèle obtenu est identique au modèle de base.
Nous allons néanmoins expliquer la procédure qui nous a permis d'arriver à cette conclusion.

7.2.1 Etape 1 : le choix du modèle

La problématique concernant le choix de la loi reste identique à celle exposée pour le modèle de fréquence de **B**ase.

La seule différence réside dans l'initialisation du modèle : la variable « genre » ne fait à présent pas partie des variables renseignées.

Critères AIC/BIC

Nous utilisons les critères AIC et BIC comme critères de sélection de la loi suivie par la fréquence de sinistre.

	AIC	BIC
Loi Poisson	30 715	30 905,98
Loi Binomiale Négative	30 615	30 815,32

Tableau 38 - AIC et BIC par loi dans le modèle de fréquence Sans Genre

L'AIC et le BIC, de par leur définition, tiennent compte du nombre de variables considérées dans le modèle. Nous n'obtenons donc pas les mêmes valeurs que pour le modèle B.
Cependant, on conserve la loi Binomiale Négative pour laquelle ces deux indicateurs sont minimums.

Critère espérance variance et test du Chi-2 d'adéquation

Inclure ou non la variable « ***gender*** » dans le modèle ne modifie pas les conclusions du critère Espérance-Variance et du Test du Khi-2 d'adéquation pour le choix de la loi (cf. § 6.2.1).
En effet, ces deux critères de choix ne tiennent pas compte des variables tarifaires associées au modèle. Ils portent donc de nouveau notre choix vers la loi Binomiale Négative.

7.2.2 Etape 2 et 3 : Estimation des coefficients de la régression et procédure de sélection des variables

Pour déterminer les coefficients de la régression, on se sert de la fonction ***glm.nb()*** puis on utilise la **procédure descendante** afin d'éliminer les variables non significatives.

Seules les variables ***VV***, ***VB_f*** et ***agecat*** correspondant respectivement aux classes de valeurs de véhicule, aux classes de types de véhicules (crées par fréquences de sinistres similaires) et aux classes d'âges de l'assuré sont retenues.

On obtient alors des variables ainsi que des coefficients de régression identiques à ceux trouvés lors de la mise en place du premier modèle de fréquence.

Le modèle obtenu étant identique au modèle de fréquence de **B**ase, l'étape 4 consistant à s'assurer de la validité du modèle a déjà été traitée et ne sera pas présentée.

On retient que le modèle de fréquence obtenu sans inclure la variable genre comme variable du modèle est identique au modèle de fréquence qui inclue cette dernière mais dans lequel elle n'est pas considérée comme ayant une influence significative sur la fréquence de sinistralité automobile du portefeuille.

Chapitre 8 - Comparaison des modèles

Les deux chapitres précédents ont servis à la mise en place respective d'un modèle de tarification applicable avant la publication de la *Gender Directive* et d'un modèle en réponse à cette nouvelle règlementation, à savoir la simple suppression de la variable jugée discriminante du modèle initial.

L'intérêt de ce chapitre est alors d'observer l'impact de sa suppression aussi bien sur la qualité d'ajustement des Modèles Linéaires mis en place que sur la segmentation et la mutualisation au sein d'une même classe tarifaire.
Pour ce faire, nous analyserons les variations de niveau de Prime Pure engendrées ainsi que les ratios des Sinistres sur Primes.

De même que précédemment, on se restreindra à la présentation de ces analyses par classe d'âges au vu de l'existence d'un nombre important de classes tarifaires.

8.1 Comparaison des modèles linéaires généralisés mis en place

Dans ce paragraphe nous comparons la qualité des Modèles Linéaires Généralisés mis en place pour les modèles tarifaires B et SG :

- Leur modèle de fréquence étant identique, aucune comparaison n'est nécessaire
- Leur modèle de coût est pour sa part différent, c'est sur lui que nous nous pencherons

Les variables retenues pour ces deux modèles de coût étant identiques au genre près, on analyse dans un premier temps deux indicateurs de qualité du modèle qui tiennent compte du nombre de variables dont est constitué ce dernier : l'AIC et le BIC.

	AIC	BIC
Modèle B	68 943	69 056,71
Modèle SG	68 949	69 056,78

Tableau 39 - Critères AIC / BIC

Les valeurs obtenues permettent de conclure : aussi bien en termes d'AIC que de BIC minimum, c'est le modèle de base qui est meilleur.

Cependant, nous nous devons également de souligner que les écarts entre les deux modèles sont faibles avec un écart de seulement 6 points pour l'AIC et de moins de 0,1 points pour le BIC.

Nous réalisons donc une comparaison plus poussée des performances de ces deux modèles en analysant les résidus et la déviance du modèle (§ 8.1.1) ainsi que les intervalles de confiance des coefficients obtenus dans chacune des régressions (§ 8.1.2).

8.1.1 Résidus et déviance

Les résidus Studentisés

La présence de points aberrants perturbe la modélisation et les estimations des coefficients de la régression.

Il est donc indispensable de détecter les points atypiques de chacun des modèles pour évaluer la qualité de l'estimation des coefficients et donc de la modélisation.

Une donnée atypique est caractérisée par une grande valeur de résidu Studentisé en valeur absolue ($|t_i^*|_{1\leq i\leq n}$).
Ces derniers suivent une loi Normale centrée réduite : si la modélisation est de qualité, environ 5 % des résidus devraient être hors de l'intervalle [-2,2].

	Moyenne des résidus Studentisés	**Ecart type des résidus Studentisés**	$\boldsymbol{Card\{|t_i^*|\} > 2}$
Modèle B	-0,4058304	0,9153574	2,7 %
Modèle SG	-0,4063228	0,9151015	2,5 %

Tableau 40- Analyse des résidus Studentisés

```
plot(rstudent(Step))
```

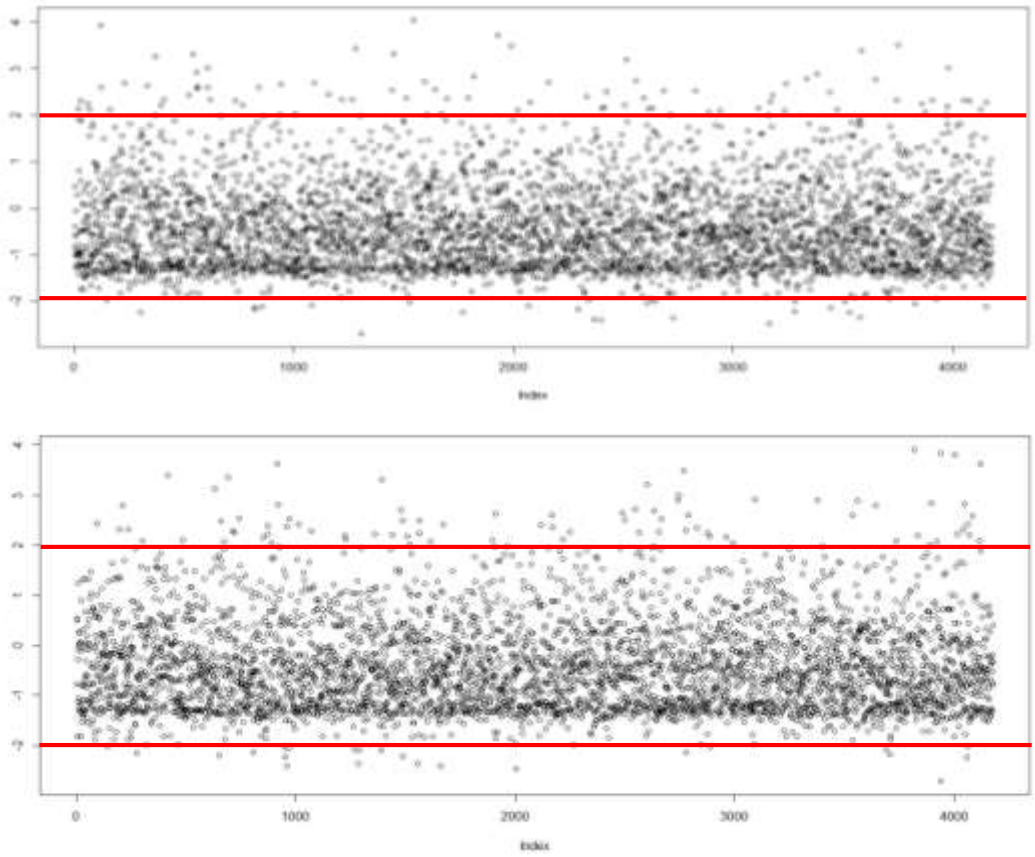


Illustration 16 - Résidus Studentisés du modèle de Base (en haut) et du modèle SG (en bas)

La comparaison des deux modèles via leurs résidus Studentisés met en avant le modèle B en terme de qualité :

- L'espérance de ses résidus est la plus proche de 0 ;
- L'écart-type de ses résidus est le plus proche de 1 ;
- Ses résidus ont les quantiles qui se rapprochent le plus des quantiles de la loi Normale.

Cependant, la différence de qualité entre les deux modèles n'est pas flagrante : ils semblent même assez proches.

Résidus du modèle et résidus de déviance

Nous présentons ici quatre graphes qui nous permettent d'effectuer une analyse des résidus de déviance.
Ces quatre graphiques sont illustrés pour chacun des deux modèles obtenus, à savoir le modèle de **B**ase et le modèle **S**ans **G**enre.

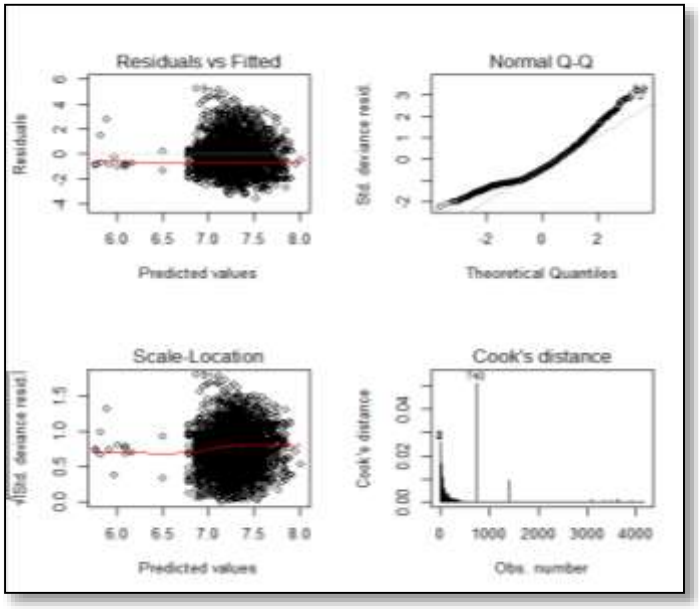


Illustration 17 - Résidus du modèle B

Le premier graphique (en haut à gauche) est une représentation des résidus en fonction des valeurs prédites. On observe l'équidispersion des points autour de l'axe des ordonnées, ce qui traduit une bonne adéquation du modèle.

Le second graphique correspond à un QQ-plot : il permet de contrôler l'adéquation des résidus Studentisés à une loi Normale centrée réduite. Ici, les points sont orientés sur la diagonale avec en abscisse les quantiles théoriques et en ordonnée les quantiles observés d'une loi Normale centrée réduite : l'adéquation est correcte.

Nous disposons également d'une représentation de la racine des résidus (en valeurs absolues) en fonction des valeurs prédites (cf. 3ème graphique).
Son interprétation est identique à celle du premier graphique.

Enfin une représentation des distances de Cook est disponible sur le dernier graphique.

La distance de Cook est une autre mesure de l'impact d'une observation sur l'équation de la régression.

Cette distance correspond à la différence entre les coefficients β calculés et les valeurs qui auraient été obtenues si l'observation correspondante n'avait pas fait partie de l'analyse. Aucune des observations n'admet une distance de Cook supérieure à 1, valeur à partir de laquelle la distance est considérée comme anormale.

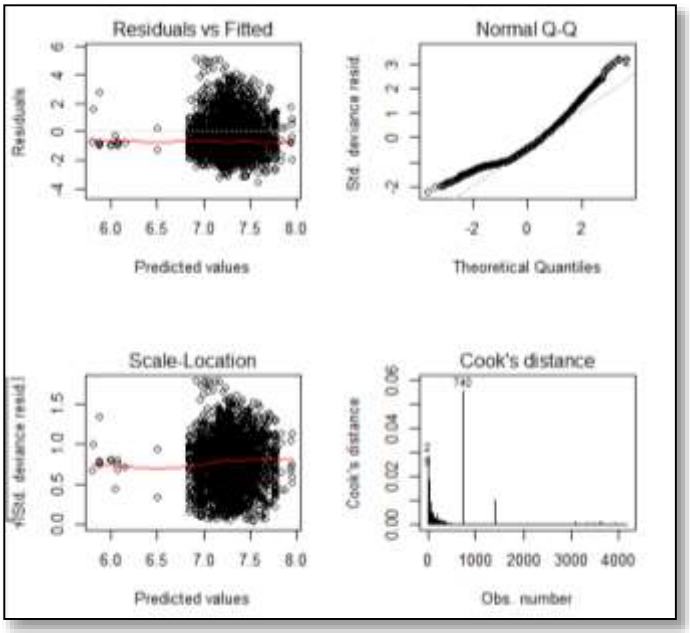


Illustration 18 - Résidus du modèle SG

Pour chacun des deux modèles analysés,
les quatre graphiques obtenus sont quasi-identiques : leur qualité « technique » est donc très proche.

Analyse des déviances : Test du rapport de vraisemblance

Le test du rapport de vraisemblance permet de comparer deux modèles emboités[1] : il est ici question de comparer le modèle complet (modèle **B**) avec le modèle privé de la variable « genre » (modèle **SG**). Ces modèles ont respectivement $q_2 = 5$ et $q_1 = 4$ paramètres.

La modalité « F » étant la modalité de référence de la variable ***gender***, il s'agit ici de tester si le coefficient associé à la variable genre (β_M) est significativement non nul.

Notre test est axé autour des deux hypothèses suivantes :

- Hypothèse nulle (H_0) : $\beta_M = 0$ (le sous modèle convient)
- Hypothèse alternative (H_1) : $\beta_M \neq 0$ (le modèle complet est meilleur)

Il est réalisé à l'aide de la fonction **anova()** sous R et prend comme paramètre : le modèle réduit (Stepp), le modèle complet (Step) ainsi qu'une indication concernant la loi suivie par la statistique de test, ici la loi de Fisher (F) .

```
> anova(Stepp,Step,test = "F")
Analysis of Deviance Table

Model 1: (data$coûtUnitaire - 159.75) ~ data$agecat + data$veh_age + data$VB_c +
    data$VV
Model 2: (data$coûtUnitaire - 159.75) ~ data$gender + data$agecat + data$veh_age +
    data$VB_c + data$VV
  Resid. Df Resid. Dev Df Deviance    F  Pr(>F)
1      4163     7333.6
2      4162     7321.7  1   11.831 4.62 0.03166 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustration 19 – Résultat du test du rapport de vraisemblance

Notre statistique de Test $\Delta_{1,2}$, qui correspond à la différence entre les déviances des deux modèles, vaut ici 11,831.

Cette dernière doit être comparée à la valeur du quantile d'ordre 5 % de la loi de Fisher de paramètres $(1\ , n-p)$, la valeur 1 correspondant à la différence du nombre de paramètres entre les deux modèles $(q_2 - q_1)$.

Ce quantile a pour valeur 4,62 : la statistique de test est donc largement supérieure à ce niveau critique.
On en conclu que H_0 est rejetée : le modèle complet (**B**) est meilleur que le modèle restreint (**SG**).

8.1.2 Analyse des intervalles de confiance des coefficients

Les coefficients de la régression $(\hat{\beta}_j)_{1\leq j\leq p}$ sont obtenus en maximisant la fonction de Log Vraisemblance du modèle :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, \mathrm{LL}\,(y, \theta(\beta), \phi) = \underset{\beta}{\operatorname{argmax}} \ln\left(\prod_{i=1}^{n} f(y_i|\theta, \phi)\right)$$

[1] Cette notion a été développée en 3.2.4 – Déviance du modèle

Leur détermination se fait à l'aide de méthodes itératives et un intervalle de confiance à 95 % de ces coefficients $(\hat{\beta}_j)_{1\leq j\leq p}$ peut être obtenu par la méthode du Maximum de Vraisemblance, ou encore la méthode de Wald[1].
En pratique, on utilise la fonction ***confint()*** du logiciel R pour obtenir les bornes de l'Intervalle de Confiance (**IC**) de chacun des coefficients de la régression.
C'est la taille des intervalles de confiance que nous analysons : plus leur étendue est grande, moins le modèle est précis.

Pour chaque coefficient sont indiquées la borne inférieure (2,5 %) et supérieure (97,5 %) de l'intervalle de confiance à 95 % de la valeur du coefficient de régression.

Modalité	2,5 %	Estimation	97,5 %
(Intercept)	5,56	6,32	7,25
data$genderM	0,01	0,11	0,21
data$agecat2	-0,41	-0,23	-0,04
data$agecat3	-0,45	-0,27	-0,09
data$agecat4	-0,53	-0,35	-0,17
data$agecat5	-0,56	-0,35	-0,15
data$agecat6	-0,58	-0,36	-0,12
data$veh_age2	-0,07	0,08	0,23
data$veh_age3	-0,02	0,14	0,29
data$veh_age4	-0,08	0,12	0,31
data$VB_cVB_c2	0,27	1,15	1,84
data$VB_cVB_c3	0,40	1,28	1,99
data$VB_cVB_c4	0,55	1,45	2,17
data$VVVV2	-0,29	-0,13	0,03
data$VVVV3	-0,22	-0,04	0,15
data$VVVV4	-0,42	-0,20	0,02
data$VVVV5	-0,57	-0,32	-0,08

Tableau 41 - IC des coefficients du modèle B

Modalité	2,5 %	Estimation	97,5 %
(Intercept)	5,61	6,38	7,30
data$agecat2	-0,42	-0,23	-0,05
data$agecat3	-0,47	-0,28	-0,11
data$agecat4	-0,54	-0,35	-0,17
data$agecat5	-0,56	-0,36	-0,16
data$agecat6	-0,58	-0,35	-0,12
data$veh_age2	-0,08	0,07	0,22
data$veh_age3	-0,01	0,14	0,30
data$veh_age4	-0,05	0,14	0,33
data$VB_cVB_c2	0,23	1,12	1,81
data$VB_cVB_c3	0,37	1,26	1,96
data$VB_cVB_c4	0,53	1,44	2,16
data$VVVV2	-0,28	-0,11	0,05
data$VVVV3	-0,19	-0,01	0,17
data$VVVV4	-0,38	-0,16	0,05
data$VVVV5	-0,52	-0,28	-0,04

Tableau 42 - IC des coefficients du modèle SG

Ces chiffres étant nombreux et difficiles à analyser à vue d'œil, le tableau ci-dessous offre un récapitulatif des comparaisons des Intervalles de Confiance pour chaque coefficient, et donc chaque modalité (hormis la modalité de référence pour chaque variable).

	Modèle B	Modèle SG
data$agecat2	×	
data$agecat3	×	
data$agecat4	×	
data$agecat5		×
data$agecat6	×	
data$veh_age2	×	
data$veh_age3	×	
data$veh_age4		×
data$VB_cVB_c2	×	
data$VB_cVB_c3	×	
data$VB_cVB_c4	×	
data$areaB	×	
data$areaC		×

× - Modèle présentant la plus grande étendue pour l'IC de la modalité considérée

Nous remarquons une alternance de précision dans les coefficients. De plus, les valeurs (des étendues) sont là aussi très proches.
Aucun modèle ne se distingue.

[1] La méthode de construction de ces intervalles de confiance est disponible au Chapitre 9 de [O5]

data$areaD	×	
data$areaE	×	
data$areaF	×	

Tableau 43 - Comparaison de l'étendue des IC

8.2 Comparatif des tarifs obtenus et des S/P

Le paragraphe précèdent nous a permis d'établir que les modèles **B** et **SG** étaient fort semblables en terme de qualité sur le plan théorique.

Nous arrivons à présent à la comparaison de l'aspect pratique des modèles : nous allons nous intéresser aux montants de Prime Pure obtenus et à la politique tarifaire de l'assureur induite par la conservation de chacun des modèles. Les résultats seront de nouveau présentés par classe d'âges faute de pouvoir présenter les résultats par classe tarifaire au vu de leur nombre trop important.

Ainsi, nous tiendrons compte des deux parties engagées dans un contrat lors de cette analyse comparative :

- D'un côté, l'assuré qui souhaite payer une Prime Commerciale « raisonnable » et représentative de son risque ;
- De l'autre, l'assureur, qui veut assurer la pérennité de son activité et se doit de proposer une prime qui couvre les pertes dues aux remboursements des sinistres.

8.2.1 Prime Pure moyenne par classe d'âges

La mise en place de deux GLM nous aura permis de déterminer le coût ($E[C|C > d]$) et la fréquence moyenne ($E[N|C > d]$) par classe tarifaire : on en déduit également la Prime Pure attritionnelle payée par les individus appartenant à chacune de ces classes à l'aide de la relation suivante :

$$\mathrm{PP_{attri}} = \mathrm{E[S_{attri}]} = \mathrm{E[C|C > d]} \times \mathrm{E[N|C > d]} + \mathrm{d} \times \mathrm{E[N|C = d]}$$

On commence par évaluer les écarts de primes inter-âges. Nous traitons alors les femmes et les hommes de manière indifférenciée en calculant une Prime Pure moyenne par classe d'âge pour chacun des deux modèles.

Les résultats sont présentés dans les tableaux suivants :

Classe d'âge	PP moyenne	Base 100
1	168,75 €	100
2	113,43 €	67,2
3	105,74 €	62,7
4	96,24 €	57,0
5	77,96 €	46,2
6	78,66 €	46,6

Tableau 44 – Prime Pure par classe d'âge modèle B

Classe d'âge	PP moyenne	Base 100
1	169,62 €	100
2	113,61 €	67,0
3	105,14 €	62,0
4	96,12 €	56,7
5	77,78 €	45,9
6	78,84 €	46,5

Tableau 45 - Prime Pure par classe d'âge modèle SG

C'est sans étonnement que nous observons un tarif dégressif (décroissant) avec l'âge de l'assuré : cette décroissance n'est pas progressive puisque la classe d'âge 1 paie plus de 30 % plus cher que la classe

d'âge 2 tandis que les baisses suivantes inter-classes d'âges sont plutôt de l'ordre de 5 % mais est commune aux deux modèles.

De plus, les Primes Pures moyennes par classe d'âge sont assez proches.

Suite à ces deux constats, nous devons à présent vérifier l'impact du nouveau modèle sur les Primes Pures moyennes **par genre** (§ 8.2.3).

8.2.2 Ratio des Sinistres sur Prime par classe d'âge

Si l'on s'intéresse à la répartition des ratios S/P par classe d'âge, on remarque que ceux obtenus pour le modèle de **B**ase sont plus resserrés autour du niveau de référence (100 %), avec une étendue de 1,13 % seulement contre 1,9 % pour le modèle **SG**.

Pour les deux modèles, la prime surévalue le risque des plus jeunes : leur ratio de Sinistres sur Primes est inférieur à 100 % ce qui signifie que la part de la prime commerciale servant à couvrir le risque encouru par l'assureur pour l'assuré est plus importantes que les remboursements des sinistres qui auraient été effectués pour cette classe d'âges si le niveau de sinistralité restait inchangé.

Classe d'âge	Charge cumulée des sinistres	Primes collectées	S / P
1	947 002,88 €	955 114,64 €	99,15 %
2	1 438 686,74 €	1 443 193,13 €	99,69 %
3	1 647 124,87 €	1 647 641,08 €	99,97 %
4	1 531 291,18 €	1 539 927,30 €	99,44 %
5	822 580,09 €	828 263,61 €	99,31 %
6	511 558,36 €	510 118,27 €	100,28 %
Total général	6 898 244,12 €	6 924 258,02 €	**99,62 %**

Tableau 46 - Ratio S / P par classe d'âge (modèle B)

Classe d'âge	Charge cumulée des sinistres	Primes collectées	S / P
1	947 002,88 €	960 048,54 €	98,64 %
2	1 438 686,73 €	1 445 472,72€	99,53 %
3	1 647 124,87 €	1 638 358,08 €	100,54 %
4	1 531 291,18 €	1 538 069,04€	99,56 %
5	822 580,09 €	826 380,79 €	99,54 %
6	511 558,36 €	511 250,97 €	100,06 %
Total général	6 898 244,12 €	6 919 580,14 €	99,69 %

Tableau 47 - Ratio S / P par classe d'âge (modèle SG)

Leur Prime Pure étant déjà plus élevée de par leur niveau de sinistralité plus important, il faut prêter particulièrement attention à ne pas accentuer ce constat par l'utilisation d'un modèle qui n'arriverait pas à mettre en adéquation la Prime Pure et le niveau de sinistralité observé.

C'est le modèle **B** qui répond le mieux à cette problématique avec un écart de 0,85 % au niveau de référence contre près de 1,5 % pour le modèle **SG**.

8.2.3 Prime Pure par classe d'âge et par genre

Nous avons vu que les Primes Pures moyennes obtenues pour chacune des classes d'âges dans les deux modèles sont semblables si l'on considère de manière indifférenciée les deux genres (cf. Tableaux 44 et 45). Ceci indique que notre deuxième modèle (modèle SG) reproduit bien la Prime **en moyenne**.
Intéressons-nous à présent à la comparaison des Primes Pures **différenciées** retournées par le modèle de base avec les primes obtenues si l'on ne prend pas en compte le genre comme critère de segmentation.

Classe d'âge	Prime Pure moyenne SG	B - Homme	B - Femme
1	169,62 €	183,60 €	157,60 €
2	113,61 €	122,35 €	107,26 €
3	105,14 €	113,67 €	100,25 €
4	96,12 €	104,55 €	90,19 €
5	77,78 €	83,02 €	73,60 €
6	78,84 €	82,21 €	75,05 €

Tableau 48 - Prime Pure moyenne par classe d'âge (modèle SG) et par genre (modèle B)

Le constat précèdent est alors bien nuancé par cette nouvelle analyse.

Dans un premier temps, on souligne que la Prime Pure obtenue avec le modèle SG est toujours située entre les deux niveaux de Prime H/F. Ceci traduit un effet de **mutualisation** induit par la suppression du critère genre de la tarification.

Avec ce modèle, les femmes (sur l'ensemble du portefeuille) paieront des primes plus importantes. A l'inverse, les hommes seront satisfaits puisqu'ils subiront une baisse tarifaire.

Imposer des changements tarifaires à ses assurés (surtout des hausses tarifaires) n'est pas chose facile pour l'assureur. En effet, il risque de voir certains de ses assurés résilier leurs contrats pour aller souscrire chez des assureurs qui proposeraient des offres plus alléchantes, quitte à même parfois, être moins couverts. La simple suppression du genre dans le système tarifaire en place avant la publication de la *Gender Directive* n'est donc pas une bonne réponse à cette nouvelle règlementation.

En supprimant le critère genre de sa tarification, il perd en qualité de segmentation et mutualise le risque sur des classes qui sont bien moins homogènes que celles créées avec le modèle de **B**ase, d'où les changements tarifaires induits. L'assureur va donc devoir mettre en place un modèle de tarification alternatif, qui lui permet de garder un certain niveau de segmentation afin de ne pas impacter le niveau des primes de manière trop importante.

L'objet du chapitre suivant sera de proposer ce type de modèles et d'étudier leur mise en place.

Conclusion :

Les deux derniers chapitres nous ont permis d'établir un modèle Sans Genre par GLM et de le comparer au modèle de base. Tout genre confondu, ce modèle se révèle avoir de primes quasi équivalentes au modèle de base. De plus la qualité du modèle Sans Genre est également aussi bonne que le modèle initial.

Toutefois si l'on compare les Primes Pures obtenues avec les Primes Pures Homme/Femme du modèle de base, on remarque une forte disparité en défaveur des femmes surtout pour les plus jeunes conductrices. Le modèle Sans Genre n'est donc pas une réponse adaptée à cette nouvelle règlementation car son utilisation risque d'entrainer des changements de composition du portefeuille. En effet, il ne serait pas étonnant que les jeunes conductrices quittent le portefeuille suite à des hausses tarifaires importantes.

Chapitre 9 - Modèles de tarification alternatifs

La *Gender Directive* autorise l'assureur à établir une Prime Pure différentiée à condition que la Prime Commerciale finalement proposée à l'assurée soit bien identique pour les deux sexes.

Dans cette partie vont être mis en place deux modèles qui intègrent le critère « genre » sans pour autant proposer des Primes Commerciales différentiées aux assurés.

Il s'agit du modèle qui mutualise le risque entre les deux genres en pondérant leur prime respective par la proportion Homme / Femme du portefeuille ainsi qu'une tarification basée sur l'utilisation d'un modèle prédictif du genre de l'assuré.

Ces modèles ne nécessitent pas l'estimation des coefficients associés à chaque modalité de variable via un GLM car ils s'appuient sur les résultats obtenus avec le modèle de **B**ase. L'utilisation de ces coefficients sera réadaptée en fonction de la solution envisagée.

De plus, les valeurs obtenues pour le modèle initial (modèle **B**) seront rappelées à chaque nouveau résultat en guise d'outils comparatifs pour juger de la qualité de ces alternatives.

9.1 Mutualisation du risque par pondération en proportion Homme / Femme (Modèle P)

9.1.1 Présentation du modèle

Ce modèle de tarification consiste à déterminer une nouvelle prime qui résultera de la pondération des deux primes (Homme / Femme) obtenues à l'aide du système de tarification déjà existant.

Se pose alors la question de savoir à quelle échelle est effectuée la pondération : relativement à la proportion H / F du portefeuille ou bien relativement à la proportion H / F par classe tarifaire ?

Illustrons les impacts de ce choix par un exemple.

Soit un portefeuille de 1 000 assurés dont le montant des primes et la répartition des assurés sont explicités dans le tableau ci-dessous.

Age		Homme	Femme	Total (effectif)
< 25 ans	Prime	1200 €	800 €	-
	Effectif	160	40	200
≥ 25 ans	Prime	600 €	700 €	-
	Effectif	400	400	800
Total (effectif)	-	500	500	1000

Tableau 49 – Prime et effectif par genre et par classe d'âge

Choix 1 : Pondération relative à la proportion H/F du portefeuille

Si l'assureur choisit de pondérer relativement à la proportion H/F du portefeuille, qui est dans notre exemple de 50 % / 50 %, nous obtenons la nouvelle grille tarifaire suivante :

Age/Genre	Homme	Femme
< 25 ans	1 000 €	1 000 €
≥ 25 ans	650 €	650 €

Tableau 50- Grille tarifaire obtenue avec le choix 1

Choix 2 : Pondération relative à la proportion H/F par classe tarifaire

Si l'assureur choisit de pondérer relativement à la proportion H/F du portefeuille par classe tarifaire, nous devons appliquer :

- une proportion de 80 % (H) / 20 % (F) pour les moins de 25 ans
- une proportion 50 % / 50 % pour les plus de 25 ans

Nous obtenons alors la nouvelle grille tarifaire suivante :

Age/Genre	Homme	Femme
< 25 ans	1 120 €	1 120 €
≥ 25 ans	650 €	650 €

Tableau 51- Grille tarifaire obtenue avec le choix 2

Nous observerons une différence significative du niveau de prime en fonction du choix réalisé pour les assurés âgés de moins de 25 ans : ceci est dû au fait que la proportion par classe tarifaire diffère significativement de celle observable dans la totalité du portefeuille.

Ce choix, à réaliser par l'assureur, dépendra notamment de l'impact sur sa rentabilité mais également de l'impact tarifaire pour les assurés.

Le modèle le plus simple reste tout de même l'utilisation de la proportion du portefeuille global.
Cette méthode, qui n'engendre à priori pas de coûts de mise en place d'un nouveau système de tarification présente l'inconvénient de réduire la corrélation entre la Prime Commerciale et le coût du risque (la Prime Pure).

9.1.2 Calcul de la Prime Pure

Le modèle de base nous a permis de déterminer le montant de Prime Pure attritionnelle par classe tarifaire pour chacun des deux genres :

$$PP^{i}_{attri} = \mathrm{E}\left[S^{i}_{attri}\right] = \mathrm{E}\left[C^{i} \middle| C^{i} > \mathrm{d}\right] \times \mathrm{E}\left[N^{i} \middle| C^{i} > \mathrm{d}\right] + \mathrm{d} \times \mathrm{E}\left[N^{i} \middle| C^{i} = \mathrm{d}\right]$$

Où $i \in \{F, H\}$

A présent, on reprend le montant de prime obtenu que l'on va pondérer en fonction de la **proportion Homme / Femme du portefeuille global**.

Pour rappel, le portefeuille est constitué à 57 % de femmes et 43 % d'hommes.
Notre nouvelle prime est alors donnée par :

$$PP = 57\,\% \times PP^{F}_{attri} + 43\,\% \times PP^{H}_{attri}$$

Or, comme le genre n'entre pas en jeu dans le calcul de la fréquence moyenne :

$$\mathrm{E}[N^{F}|C^{F} > \mathrm{d}] = \mathrm{E}[N^{H}|C^{H} > \mathrm{d}] = : E[N|C > d]$$

Et

$$\mathrm{E}[N^{F}|C^{F} = \mathrm{d}] = \mathrm{E}[N^{H}|C^{H} = \mathrm{d}] =: E[N|C = d]$$

D'où

$$\boxed{PP_{attri} = E[N|C = d] \times d + E[N|C > d] \times (\,57\,\% \times \mathrm{E}[C^{F}|C^{F} > \mathrm{d}] + 43\,\% \times \mathrm{E}[C^{H}|C^{H} > \mathrm{d}])}$$

9.1.3 Prime Pure et Ratio de Sinistres sur Primes par classe d'âge

Prime Pure moyenne par classe d'âges

Il n'est plus question ici d'analyser le montant moyen par genre dans le modèle **P** puisque les deux genres paient un montant identique. Nous allons donc uniquement nous intéresser à la Prime Pure moyenne par classe d'âges.

Classe d'âges	PP (modèle P)	PP Homme (modèle B)	PP Femme (modèle B)	Base 100 (modèle P)	Base 100 (modèle B)
1	168,65 € (-8,14 % pour H) (+7,10 % pour F)	183,60 €	157,60 €	100	100
2	113,65 € (-7,12 % pour H) (+5,96 % pour F)	122,35 €	107,26 €	67,4	67,2
3	105,95 € (-6,79 % pour H) (+5,69 % pour F)	113,67 €	100,25 €	62,8	62,7
4	96,26 € (-7,93 % pour H) (+6,73 % pour F)	104,55 €	90,19 €	57,1	57,0
5	77,71 € (-6,40 % pour H) (+5,58 % pour F)	83,02 €	73,60 €	46,1	46,2
6	78,15 € (-4,95 % pour H) (+4,12 % pour F)	82,21 €	75,05 €	46,3	46,6

Tableau 52-Prime Pure moyenne par classe d'âge (modèle P)

La pondération en proportion Homme / Femme n'a pas affectée l'écart relatif de Prime Pure inter-âges : les très jeunes conducteurs paient toujours plus de 30 % plus cher pour leur assurance RC automobile et le tarif est décroissant par la suite.

On peut donc affirmer que la proportion Homme / Femme pour chacune des classes d'âges est assez proche de celle observée dans le portefeuille global (ceci sera vérifié dans la remarque suivante).

De plus, cette fois ci par construction, la Prime Pure unisexe obtenue par classe d'âges est bien comprise entre les deux niveaux de Prime Pure différenciés.
De nouveau, une mutualisation est appliquée (cette fois ci volontairement par construction) entre les hommes et les femmes : des changements tarifaires se font sentir.

Remarque :

Nous avons ici choisi d'appliquer une pondération en fonction de la proportion du portefeuille total, mais nous aurions tout autant pu utiliser la proportion par classe d'âges par exemple.

Classe d'âges	F	M
1	56,71 %	43,29 %
2	58,81 %	41,19 %
3	58,82 %	41,18 %
4	57,10 %	42,90 %
5	53,47 %	46,53 %
6	49,80 %	50,20 %
Total général	**56,49 %**	**43,51 %**

Tableau 53- Structure H/F du portefeuille par classe d'âges

Etant donné que la répartition inter-classes d'âges est proche de la répartition du portefeuille considéré dans son ensemble, le choix d'échelle de pondération ne se fera pas sentir sur le montant des Primes Pures.

Classe d'âges	PP moyenne	Base 100
1	168,61 €	100
2	113,38 €	67,2
3	105,71 €	62,7
4	96,20 €	57,1
5	77,94 €	46,2
6	78,67 €	46,7

Tableau 54 – Prime Pure moyenne par âge obtenue après pondération par classe d'âges (modèle P)

Ratio de Sinistres sur Primes par classe d'âges

Nous avons vu dans le paragraphe précèdent que la proportion par classe d'âges d'hommes et de femmes était à peu près stable. On ne devrait donc subir qu'un impact très modéré de la pondération appliquée sur le ratio de Sinistres sur Primes.

Classe d'âges	Charge cumulée des sinistres	Primes collectées	S / P	S/P (modèle B)
1	947 002,88 €	954 554,80 €	99,21 %	99,15 %
2	1 438 686,73 €	1 445 905,38 €	99,50 %	99,69 %
3	1 647 124,87 €	1 650 977,40 €	99,77 %	99,97 %
4	1 531 291,18 €	1 540 318,47 €	99,41 %	99,44 %

5	822 580,09 €	825 551,17 €	99,64 %	99,31 %
6	511 558,36 €	506 782,81 €	100,94 %	100,28 %
Total général	6 898 244,12 €	6 924 090,04 €	**99,63 %**	99,62 %

Tableau 55- Calcul du S / P par classe d'âges (modèle P)

C'est en effet le cas : ce modèle permet à l'assureur de garder sa politique tarifaire, à savoir de légèrement surévaluer le risque des très jeunes conducteurs et de sous-évaluer celui des plus âgés, tout en restant dans une fourchette assez fine autour des 100 % pour faire payer à l'assuré une Prime Pure représentative de son niveau de risque.

Conclusion :

Le modèle de pondération permet de retrouver l'équilibre inter-classes d'âges que présentait le modèle initial. Cependant, l'utilisation de ce modèle n'est conseillée que si la proportion Homme /Femme reste inchangée. Cette hypothèse n'est pas toujours réalisable, surtout au vu des changements tarifaires induits par le modèle.

Dans notre cas, ils sont quasi-symétriques car la proportion Homme/Femme de notre portefeuille est proche de 50 %. Le risque est alors d'assister à une modification de la composition du portefeuille qui bouleverserait cet équilibre.

9.2 Une tarification unisexe basée sur un modèle prédictif

L'une des applications de l'analyse prédictive les plus connue est l'évaluation du risque client dans l'ensemble des services financiers : pour décider si l'on accorde ou non un prêt à un client, on regarde par exemple son FICO score[1]. Ce dernier permet de classer les individus selon la probabilité de rembourser leurs crédits en temps voulu et sa détermination s'appuie sur différents critères liés à la possibilité de remboursement (historique de paiement, niveau actuel d'endettement, etc.)

Dans le même ordre d'idée, nous mettons en place dans cette partie une tarification basée sur un modèle prédictif du genre qui s'appuie sur différents critères appelés **variables de prédiction**.

En effet, l'utilisation de la variable « genre » n'est plus autorisée mais l'assureur souhaite toujours établir un tarif qui reflète au mieux le niveau de risque de chacun des assurés.
Il peut alors tenter de retrouver cette information (le genre) à l'aide d'autres informations dont il dispose. Le genre de l'individu n'étant pas communiqué directement par l'assuré, l'utilisation indirecte de la variable « genre » ne rentre pas dans le cadre d'application de la *Gender Directive*.

Pour ce faire, il va mettre en place un autre modèle lui permettant de prédire, à l'aide de variables, c'est-à-dire d'informations fournies par l'assuré (qui peuvent être celles utilisées dans la segmentation tarifaire, ou non), le genre de l'individu.

Pour ce système de tarification, l'assureur conserve son modèle de tarification différentié mais n'apporte pas l'information concernant le genre du client à partir de l'information renseignée par ce dernier.

Plus le modèle de prédiction sera performant, moins il y aura de différences de tarifs avec le modèle précédant la mise en place de la *Gender Directive*.
On propose d'illustrer le fonctionnement de ce modèle par un exemple.

[1] FICO: Fair Isaac Corporation (faisant référence au créateur du score)

Soit un modèle de prédiction du genre prenant en compte deux variables tarifaires et retournant les résultats suivants.

Variables prédictives		Prédiction
Modèle de véhicule	**Age**	**Genre**
Ferrari	20-30 ans	Homme
Citroën	35-40 ans	Femme
Twingo	20-25 ans	Femme

Tableau 56 - Présentation du modèle de prédiction

Soit le portefeuille d'un assureur automobile comportant trois assurés pour lequel sont renseignés le modèle du véhicule, l'âge de l'assuré ainsi que son genre.

N° d'assuré	Modèle de véhicule	Age	Genre
1	Ferrari	22 ans	Homme
2	Citroën	35 ans	Homme
3	Twingo	20 ans	Femme

Tableau 57 - Portefeuille de l'assureur

L'assuré N°1 est un homme et est également prédit comme étant de genre masculin : il se verra donc appliquer le tarif Homme. Par contre, l'assuré N°2 qui est en réalité un homme, est considéré comme une femme lors de la tarification de son contrat (erreur du modèle prédictif) : le tarif appliqué à cet assuré sera donc le tarif féminin obtenu avec le modèle de tarification existant.

La mise en place de ce genre de modèle a été motivée par l'**intérêt Marketing** qu'il présente : après la mise en place de la *Gender Directive*, bien que rien n'interdise à l'assureur de demander à l'assuré de renseigner son genre, un individu averti ne comprendra pas l'intérêt de fournir ce renseignement s'il n'est pas utilisé dans le calcul de son tarif.

Ce modèle présente également un **intérêt actuariel**, à savoir, la possibilité pour l'assureur de proposer une Prime Commerciale plus proche de la Prime Technique que s'il appliquait une pondération de la proportion Homme / Femme du portefeuille sur le tarif.

Le **data mining**[1] propose de nombreuses méthodes d'élaboration d'un modèle prédictif, nous en avons sélectionné deux, adaptées à notre problématique:

- Un modèle de régression logistique binaire
- Un arbre de classification, appartenant à la famille des arbres de Classification et de Régression (CART : *Classification And Regression Trees*)

Nous comparerons ainsi la performance de ces deux méthodes prédictives.

9.2.1 Un modèle prédictif basé sur une régression Logistique binaire (Modèle PR_L)

Ce modèle fait partie de la grande famille des Modèles Linéaires Généralisés dont la présentation a fait l'objet du troisième chapitre. En effet, la régression logistique binaire désigne un Modèle Linéaire Généralisé pour lequel la loi suivie par la variable à expliquer est la loi Binomiale accompagnée de la fonction de lien « logit ».

[1] Le data mining consiste à extraire une information, une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

Il n'est donc pas ici question de commencer par introduire les fondements théoriques du principe de régression logistique, ceci ayant déjà été fait précédemment.

Mise en place du modèle

La variable à prédire (à expliquer) est ici le genre de l'individu. Cette dernière admet deux modalités « F » et « M » que nous renommerons afin d'adapter notre problème à un modèle Binomial :

- La valeur 1 sera attribuée aux femmes $(F \to 1)$
- La valeur 0 sera attribuée aux hommes $(M \to 0)$

Le choix de la numérotation n'a pas été laissé au hasard : prend la valeur 1 le groupe ayant la plus grande probabilité d'occurrence (cf. Graphique 5).

Afin de pouvoir évaluer la qualité de cette régression, nous allons séparer notre portefeuille d'individu en deux :

- 75 % du portefeuille servira à la mise en place du modèle (**échantillon d'apprentissage**[1])
- les 25 % restant serviront au calcul de la précision du modèle (**échantillon test**)

Le choix des individus appartenant à l'un ou l'autre des échantillons est effectué par un tirage sans remise de 16 947[2] polices parmi 67 788.

Etape 1 : choix de la loi

La variable à prédire prenant les valeurs {0,1}, c'est la loi Binomiale accompagnée de la fonction de lien « logit » qui nous servira à appliquer le GLM.

Pour rappel, la fonction « **logit** » est donnée par :

$$g(x) = \ln(\frac{x}{1-x})$$

D'où son inverse :

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$$

Etape 2 et 3 : sélection des variables et détermination des coefficients de la régression

Sont entrées dans le modèle les variables suivantes :

Intitulé	Signification	Nombre de modalités
Veh_age	Classes d'âges du véhicule	4
Agecat	Classes d'âges du conducteur	6
Veh-body	Type de véhicule	13
VV	Classes de valeurs du véhicule	5
Area	Zone géographique	6

[1] Cet échantillon sera désigné par « baseM » dans le code R.
[2] $25\ \% \times 67\ 788 = 16\ 947$

Tableau 58-Variables intégrées dans le modèle PR_L

La variable ***veh_body*** est intégrée dans le modèle sous sa forme brute, c'est-à-dire avant le regroupement de ses modalités sous forme de classe, et ce, pour deux raisons :

- Elle a subit des traitements orientés selon la problématique de fréquence et coût semblables, la problématique est ici toute autre puisqu'il s'agit de déterminer le genre de l'individu.
- L'analyse en composante du portefeuille par genre nous a permis de déceler l'importance de cette variable dans son rôle à jouer pour la détermination du genre de l'individu : plus le nombre de modalité sera important, plus le modèle de prédiction sera précis.

Pour sélectionner les variables les plus significatives du modèle et ainsi déterminer les coefficients de la régression logistique, on utilise la procédure descendante (*Backward selection*) comme fait lors de la mise en place du modèle **B** (modèle de base).

On détermine alors les coefficients associés à chaque modalité pour les variables retenues dans la régression binaire.

```
glm(formula = baseM$BinaryGender ~ baseM$veh_body + baseM$VV +
    baseM$agecat + baseM$veh_age + baseM$area, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8387  -1.1821   0.7755   0.9847   2.1014

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.32396    0.35333   0.917  0.35921
baseM$veh_bodyCONVT  1.07949    0.43579   2.477  0.01325 *
baseM$veh_bodyCOUPE  0.40236    0.35812   1.124  0.26122
baseM$veh_bodyHBACK  0.94648    0.34889   2.713  0.00667 **
baseM$veh_bodyHDTOP  0.04568    0.35278   0.129  0.89697
baseM$veh_bodyMCARA  0.14591    0.41014   0.356  0.72203
baseM$veh_bodyMIBUS  0.71650    0.35907   1.995  0.04600 *
baseM$veh_bodyPANVN -0.99490    0.36188  -2.749  0.00597 **
baseM$veh_bodyRDSTR -0.38756    0.68860  -0.563  0.57356
baseM$veh_bodySEDAN  0.66230    0.34845   1.901  0.05734 .
baseM$veh_bodySTNWG  0.32510    0.34833   0.933  0.35065
baseM$veh_bodyTRUCK -1.05511    0.35414  -2.979  0.00289 **
baseM$veh_bodyUTE   -0.78528    0.34989  -2.244  0.02481 *
baseM$VVVV2         -0.25377    0.03137  -8.089 6.03e-16 ***
baseM$VVVV3         -0.44061    0.03817 -11.544  < 2e-16 ***
baseM$VVVV4         -0.60042    0.04942 -12.150  < 2e-16 ***
baseM$VVVV5         -0.86521    0.05887 -14.696  < 2e-16 ***
baseM$agecat2        0.10254    0.03972   2.582  0.00984 **
baseM$agecat3        0.09996    0.03864   2.587  0.00967 **
baseM$agecat4       -0.02544    0.03838  -0.663  0.50739
baseM$agecat5       -0.21239    0.04072  -5.216 1.83e-07 ***
baseM$agecat6       -0.55645    0.04471 -12.446  < 2e-16 ***
baseM$veh_age2      -0.01943    0.02994  -0.649  0.51636
baseM$veh_age3      -0.21514    0.03291  -6.537 6.27e-11 ***
baseM$veh_age4      -0.59137    0.04195 -14.096  < 2e-16 ***
baseM$areaB          0.05039    0.02846   1.771  0.07662 .
baseM$areaC          0.04136    0.02551   1.621  0.10491
baseM$areaD          0.32864    0.03377   9.733  < 2e-16 ***
baseM$areaE          0.31431    0.03807   8.256  < 2e-16 ***
baseM$areaF          0.28054    0.04641   6.045 1.50e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustration 20 - Résultats de la procédure descendante sous R

Même après la sélection des variables, on remarque que toutes les variables disponibles ont été retenues : il faut à présent retrouver le genre de l'individu à partir de ses caractéristiques.

Etape 4.3 : Analyse des coefficients de la régression

L'analyse de composition par genre du portefeuille n'avait pas soulevé l'importance de tous les critères dans leur capacité à expliquer la variable genre. Malgré tout , ils ont tous été retenus. Ceci ne signifie pas que leur niveau de significativité est identique.

En effet, plus l'ordre de grandeur des coefficients d'une même variable est grand, plus cette variable est significative. Les résultats précedents mettent également en avant la variable ***veh_body*** de par la valeur absolue de ses coefficients plus élevée que ceux des autres variables.

Mis à part l'ordre de grandeur des coefficients, on peut également s'intéresser aux signes des coefficients obtenus pour vérifier la cohérence avec notre analyse descriptive.
La fonction de lien utilisée n'est plus la fonction log et l'interprétation du signe doit passer par une petite étude de variation de la fonction réciproque de la fonction de lien utilisée.
Cette dernière, donnée par :

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}$$

admet comme dérivée la fonction :

$$(g^{-1})'(x) = \frac{e^x}{(1 + e^x)^2} > 0$$

La fonction logit est donc croissante. Un coefficient positif fait tendre la prédiction vers 1, donc vers la réponse « Femme » tandis qu'un coefficient négatif fera pencher la balance vers la prédiction d'un genre masculin.

Par exemple, pour la variable VV, correspondant à la valeur du véhicule, on remarque que les coefficients sont tous négatifs d'une part, et décroissent d'autant plus que la valeur du véhicule augmente. Ceci reflète bien le fait que les hommes ont tendance à être en posséssion de véhicules plus onéreux que les femmes.

Nous laissons le soin au lecteur de vérifier que ceci est également valable pour les autres variables considérées.

Prédiction du genre de l'individu et calcul de Prime Pure

En utilisant la fonction de lien logit appliquée à la variable à expliquer ***gender*** on obtient la relation :

$$\log\left(\frac{E(gender)}{1 - E(gender)}\right) = \beta_0 + \beta_1 x^1 + \cdots + \beta_5 x^5$$

Avec

- $(\beta_i)_{\ i=1,\dots,5}$ le coefficient de régression logistique associé à la $i^{ème}$variable du modèle
- β_0 la valeur (donnée par « intercept ») qui correspond au coefficient associé à l'individu de référence.

Soit

$$E[gender] = P(Y = 1|x^1, \dots, x^5) = \frac{\exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^5)}{1 + \exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^5)}$$

Une fois l'espérance du genre déterminée pour l'individu en question, l'affectation à l'un ou l'autre des genres (Homme/Femme) est fonction de la valeur obtenue :

- si $E[gender] > 0{,}5$: l'individu est associé à la classe $\mathrm{Y} = 1 \rightarrow$ c'est une femme

- si $E[gender] < 0{,}5$: l'individu est associé à la classe $Y = 0 \rightarrow$ c'est un homme

Remarque :
Le cas $E[gender]$=0,5 ne peut se présenter au vu des coefficients obtenus.

Ainsi, si un **jeune individu** souhaite souscrire une garantie RC automobile pour une **COUPE neuve** appartenant à **VV3** en zone **A**, il lui sera appliqué le tarif Femme car :

$$E[gender] = \frac{\exp(0{,}32396 + 0{,}40236 - 0{,}44061 + 0 + 0 + 0{,}05039)}{1 + \exp(0{,}32396 + 0{,}40236 - 0{,}44061 + 0 + 0 + 0{,}05039)}$$

$$= 0{,}73 > 0{,}5$$

Qualité du modèle :

Dans un premier temps, pour analyser la qualité du modèle on s'intéresse aux indices de position des valeurs retournées par ce dernier.

Minimum	Maximum	Moyenne
0,07	0,78	0,47

Tableau 59 - Indices de position de la fréquence théorique de sinistres (modèle PR_L)

Plus la valeur retournée est proche de 0 (respectivement 1), plus la probabilité de dire que l'individu est un homme (respectivement une femme) sans se tromper est forte.

L'étendue de nos valeurs est de 0,71 : **ce résultat n'est pas assez proche de 1 pour dire que le modèle est performant.**

De plus, si l'on analyse la moyenne, elle est de 0,47 alors qu'un modèle « parfait » retournerait 0,57 (correspondant à la proportion de femmes de 57 %) : nous en sommes donc loin.

On va également réaliser un tableau croisant appartenance et affectation aux différents genres pour déterminer le taux de mal classés du modèle.

Ce dernier, contrairement au résultat précèdent, a été obtenu sur la base de l'échantillon test et non pas du portefeuille complet.

Appartenance / Affectation	F	M	Total
F	7 926	1 612	9 538
M	4 429	2 980	7 409
Total	12 355	4 592	16 947

Tableau 60– Répartition des individus par appartenance et affectation

Nous obtenons un taux de mal classés τ_{MC} de $\frac{1\,612+4\,429}{16\,947} = 35{,}65\ \%$.

Remarque :
Il est intéressant d'évaluer le taux de mal classés par genre et par âge car les enjeux ne sont pas les mêmes.

En effet, dans le cas où la Prime Pure des hommes est supérieure à celle des femmes :

- Si un homme est pris pour une femme, il paiera moins cher que son dû: la rentabilité de l'assureur est en danger.
- A l'inverse si ce sont les femmes qui sont prises pour des hommes, elles paieront plus cher que leur dû : c'est la présence des femmes dans le portefeuille qui est menacée.

Classe d'âges	Taux de mal classés H	Taux de mal classés F
1	50,65 %	12,13 %
2	61,65 %	15,74 %
3	64,65 %	16,89 %
4	66,91 %	14,35 %
5	52,22 %	22,62 %
6	53,20 %	26,10 %
Total	60 %	17 %

Tableau 61- Taux de mal classés par genre et classe d'âges

Les femmes sont donc bien reconnues, et une grande partie des erreurs provient du classement des hommes.

Prime Pure et rentabilité de l'assureur

Prime Pure par classe d'âges

Nous avons appliqué les coefficients obtenus pour le modèle **B** à nos assurés en remplaçant leur genre par la modalité prédite par le modèle de régression logistique binaire.
Nous déterminons alors les Primes Pures moyennes par classe d'âges.

Là encore, il n'est pas question d'analyser une prime par genre puisque deux individus de genre différents qui présenteraient les mêmes caractéristiques se verraient affecter un montant de prime identique.

Classe d'âges	Prime Pure moyenne	B - Homme	B - Femme	Base 100	Base 100 (modèle B)
1	166,33 €	183,60 €	157,60 €	100,0	100
2	111,69 €	122,35 €	107,26 €	67,1	67,2
3	104,05 €	113,67 €	100,25 €	62,6	62,7
4	94,40 €	104,55 €	90,19 €	56,8	57,0
5	77,04 €	83,02 €	73,60 €	46,3	46,2
6	78,08 €	82,21 €	75,05 €	46,9	46,6

Tableau 62- Prime Pure par classe d'âges (modèle PR_L)

Le modèle prédictif, bien que non « ultra-performant », ne provoque pas de modification substantielle concernant le niveau moyen de Prime Pure inter-âges, tous genres confondus, comme l'indiquent les deux dernières colonnes du Tableau 62.

Les montants de Prime Pure unisexe obtenus sont là encore compris entre les deux niveaux de Prime Pure déterminés avec le modèle initial.
Cependant, on remarque que la Prime Pure moyenne par âge s'approche fortement plus de la Prime Pure pour femme que de celle des hommes du modèle **B**.

En effet, ce résultat prend sa source dans le taux de mauvais classement des hommes (près de 60 %) : les hommes sont pris pour des femmes, le tarif moyen qui leur est appliqué tend donc vers celui de ces dernières.

Ceci va créer un problème pour l'assureur : ce modèle ne permet pas de mutualiser le risque puisque le risque d'erreur n'est pas symétrique entre les deux genres. Ceci impactera forcement la rentabilité de l'assureur s'il commercialise le contrat d'assurance RC en mettant en place ce modèle.

Ratio de Sinistres sur Primes

Le modèle de prédiction n'étant pas précis, l'impact sur la politique tarifaire de l'assureur devrait se faire sentir.

Nous analysons ici les ratios Sinistres sur Primes par classe d'âges : ils sont tous au-dessus du niveau de référence (100 %).

Ceci n'est pas bon signe : l'assureur sous-évalue le risque présenté par toutes les classes d'âges.

Classe d'âges	Charge cumulée des sinistres	Primes collectées	S / P	S / P (modèle B)
1	947 002,88 €	941 434,85 €	100,59 %	99,15 %
2	1 438 686,73 €	1 420 982,75 €	101,25 %	99,69 %
3	1 647 124,87 €	1 621 364,64 €	101,59 %	99,97 %
4	1 531 291,18 €	1 510 420,55 €	101,38 %	99,44 %
5	822 580,09 €	818 466,24 €	100,50 %	99,31 %
6	511 558,36 €	506 330,70 €	101,03 %	100,28 %
Total général	**6 898 244,12 €**	**6 818 999,72 €**	**101,16 %**	99,62 %

Tableau 63 – S / P par classe d'âges avec le modèle PR_L

Cette politique tarifaire ne peut être mise en œuvre en pratique car même si l'assureur va certainement attirer bon nombre de clients en proposant des tarifs plus intéressants, il court à la faillite en ayant plus de dépenses (remboursement des sinistres) que d'entrées d'argent (primes acquises).

Ce modèle n'est donc pas une bonne solution, passons à présent à la mise en place d'un modèle prédictif qui s'appuie sur une toute autre méthode d'affectation à un genre.

9.2.2 Un modèle prédictif basé sur un arbre (Modèle PR_CART)

Les arbres de classification appartiennent à la famille des Arbres de Classification et de Régression (CART : Classification And Regression Trees).

Ces derniers sont des méthodes d'apprentissage automatique utilisées pour construire des modèles de prédiction à partir de données, le résultat étant représenté graphiquement par un **arbre de décision**.

Parmi ces méthodes, on distingue :

- **Les arbres de Classifications**, qui sont adaptés à la prédiction de variables prenant un nombre fini de valeurs non ordonnées (variables qualitatives nominales). L'erreur de prédiction est alors mesurée en termes d'erreur de classification (calcul du taux de mal classés).

- **Les arbres de Régression**, réservés à la prédiction de variables quantitatives (continues ou discrètes). Dans ce cas, l'erreur de prédiction est mesurée par l'écart des carrées entre la valeur observée et la valeur obtenue.

C'est la variable ***gender*** que nous cherchons à prédire, elle prend ses valeurs dans {''M'', ''F''} : ce deuxième modèle de prédiction sera donc établi à l'aide d'un Arbre de Classification[1].

Soulignons ici que les variables utilisées dans le modèle prédictif ne sont pas forcément celles utilisées dans le modèle de tarification. Ainsi, dans le souci d'obtenir un modèle précis, on conservera les treize modalités de la variable ***veh_body*** et nous n'utiliserons donc pas les classes de type de véhicule précédemment établies.

De plus, utiliser les classes de type de véhicule n'aurait aucun sens puisqu'elles ont été établies selon une problématique de fréquence et coût voisins tandis que la problématique étudiée ici est la détermination du genre de l'individu.

Prédiction du genre de l'individu

On dispose de n = 67 788 individus dont 38 581 femmes et 29 207 hommes.

Les variables prédictives utilisées sont : la classe d'âges du véhicule et de l'individu, la zone géographique, la valeur du véhicule ainsi que son type.

Pour ce modèle de prédiction, nous ne séparons pas notre échantillon en deux (échantillon d'apprentissage et échantillon test) car ceci est déjà pris en compte dans la **procédure d'élagage**[2] de l'arbre (procédure qui consiste à sélectionner un arbre plus simple que celui obtenu à l'arrêt de la construction de l'arbre, mais qui est tout aussi précis pour classer les nouvelles observations) lors de la validation croisée.

Pour la création de l'arbre de classification, on fait appel à la fonction **rpart()** du logiciel R.

Nous spécifions ici quelques paramètres :

- MINSPLIT = 10 indique qu'on ne doit pas segmenter un nœud avec moins de 10 observations.
- MINBUCKET = 1 nous permet de refuser toute segmentation où l'un des nœuds enfant aurait moins d'une observation.

[1] Il est fortement conseillé au lecteur de lire l'Annexe 5 concernant les arbres de classification avant de poursuivre sa lecture.
[2] Voir Annexe 5 (Elagage de l'arbre) pour plus de détails.

```
> arbre= rpart(gender~veh_body+VV+agecat+veh_age+area,data=data,
method= "class",control=rpart.control(minsplit=10,minbucket=1))
> print(arbre)
```

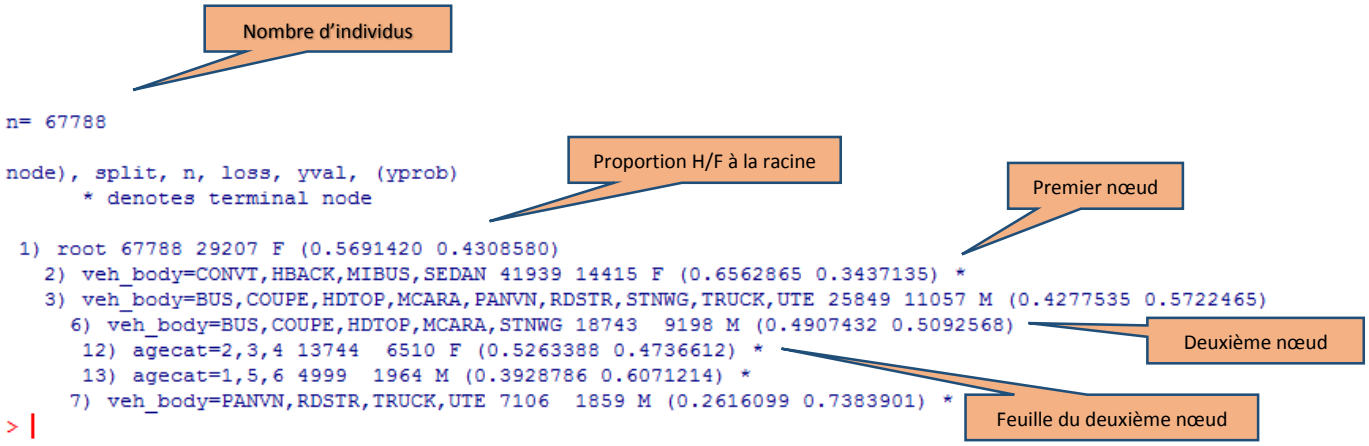


Illustration 21 - Résultat de l'arbre de classification

On peut lire que le nombre d'individus considéré est bien n = 67 788.

De plus, à chaque nœud de l'arbre apparait la proportion Femme / Homme des individus passant par ce chemin.

Par exemple à la racine, aucun chemin n'ayant encore été emprunté, on peut lire la proportion Femme/Homme du portefeuille global : 0,57 / 0,43.
Une fois le premier nœud passé, avec la règle de fractionnement concernant le type de véhicule possédé par l'assuré, la proportion Femme/Homme d'individus répondant « Oui » au premier test est alors de 0,65 / 0,34 et ainsi de suite.

Cependant, cette fonction ne renvoie pas l'arbre final puisque l'élagage de l'arbre n'a pas encore été effectué (il n'est pas automatique) : en pratique, nous devons réaliser nous même le post élagage en nous aidant du tableau « CPTABLE » obtenu avec la commande ***printcp()***.

```
> printcp(arbre)

Classification tree:
rpart(formula = gender ~ veh_body + VV + agecat + veh_age + area,
    data = data, method = "class", control = rpart.control(minsplit = 10,
        minbucket = 1))

Variables actually used in tree construction:
[1] agecat   veh_body

Root node error: 29207/67788 = 0.43086

n= 67788

        CP nsplit rel error  xerror      xstd
1 0.127880      0   1.00000 1.00000 0.0044144
2 0.012394      1   0.87212 0.87236 0.0043176
3 0.010000      3   0.84733 0.85130 0.0042961
```

Illustration 22 - Le tableau CPTABLE de RPART

Cette fonction retourne :

- Les variables utilisées dans la construction de l'arbre **: *agecat*** et ***veh_body***
- **CP** : les différents coefficients de pénalité
- **Nsplit** : le nombre de nœuds de chaque arbre donné par : $\text{nsplit} = : \text{nombre de feuilles} - 1$
- **rel error** : l'erreur calculée sur l'échantillon d'apprentissage (normalisée de manière à ce que l'erreur sur la racine soit égale à 1)
- **xerror** : l'erreur calculée en validation croisée (également normalisée)
- **xstd :** l'écart type de l'erreur calculée en validation croisée.

Par défaut, la validation croisée a été faite avec K = 10, K correspondant au nombre de portions en lequel sera découpé l'échantillon de départ pour réaliser la validation croisée.

On remarque que l'erreur calculée sur l'échantillon d'apprentissage (erreur de substitution) est plus faible que celle calculée en validation croisée ce qui confirme le caractère « optimiste » de cette première.

Pour choisir l'arbre définitif, on doit exploiter les résultats précédents : **l'arbre optimal est celui qui minimise l'erreur de validation croisée ($\boldsymbol{\varepsilon = 0,85130}$).**

Pour obtenir cet arbre, nous devons donc prendre un indice de complexité CP=0,01 (associé à l'erreur minimale) et faire appel à la fonction ***prune ()***.

Dans notre cas, l'arbre optimal est identique à celui obtenu avant élagage puisque c'est l'arbre maximal (nsplit=3).

```
> arbre.elag=prune(arbre,cp=0.01)
> print(arbre.elag)
n= 67788

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 67788 29207 F (0.5691420 0.4308580)
   2) veh_body=CONVT,HBACK,MIBUS,SEDAN 41939 14415 F (0.6562865 0.3437135) *
   3) veh_body=BUS,COUPE,HDTOP,MCARA,PANVN,RDSTR,STNWG,TRUCK,UTE 25849 11057 M (0.4277535 0.5722465)
     6) veh_body=BUS,COUPE,HDTOP,MCARA,STNWG 18743  9198 M (0.4907432 0.5092568)
      12) agecat=2,3,4 13744  6510 F (0.5263388 0.4736612) *
      13) agecat=1,5,6 4999  1964 M (0.3928786 0.6071214) *
     7) veh_body=PANVN,RDSTR,TRUCK,UTE 7106  1859 M (0.2616099 0.7383901) *
```

Illustration 23- Arbre optimal obtenu sous R après élagage

On peut à présent dresser l'arbre de classification pour une interprétation facilitée du résultat :

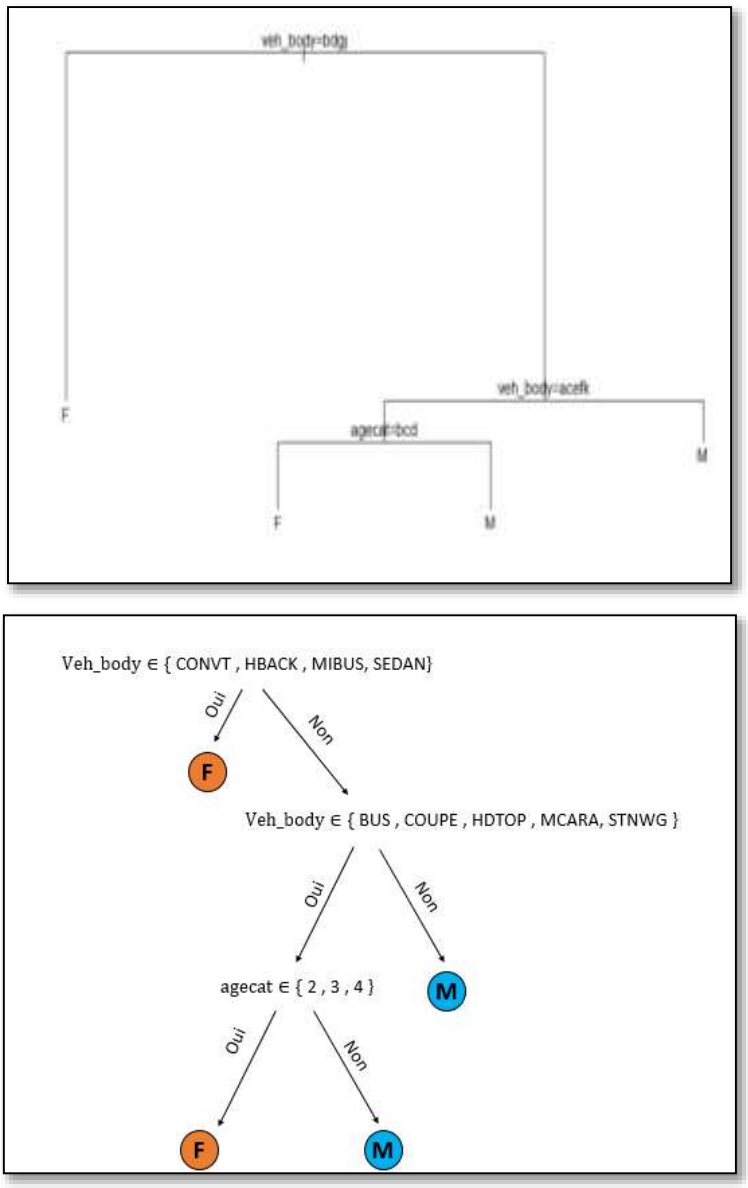


Illustration 24- Arbre de classification retourné par R (en haut) et retranscription (en bas)

On souligne que c'est en premier lieu le type de véhicule qui permet de prédire le genre de l'individu : il est impliqué dans deux des trois règles de fractionnement de l'arbre.
Ce résultat ne nous étonne pas car il avait déjà été souligné lors de l'analyse de la composition du portefeuille en proportion Homme / Femme que le type de véhicule influait significativement sur le genre (cf. Graphique 12).

Par contre, l'influence du critère âge sur la détermination du genre de l'individu était moins évidente. Ceci est souligné par le fait que l'âge n'intervient qu'en troisième règle de fractionnement (après les deux sur le type de véhicule) : sa corrélation avec le genre de l'individu n'était pas détectable par une simple analyse univariée, il aurait fallu croiser les variables ***agecat*** et ***veh_body*** pour s'en apercevoir.

Evaluation de la qualité de l'arbre

Les erreurs de resubtitution étant normalisées, nous ne pouvons lire leur valeur directement.

Ces dernières ont été recalculées à l'aide de la commande ***predict()*** qui applique les règles de partitionnement à chacun des individu pour le classer en tant qu'homme ou femme.

```
> prediction=predict(arbre.elag,base,type="class")
> summary(prediction)
    F     M
55683 12105
> mat=table(base$gender,prediction)
> mat
   prediction
        F     M
  F 34758  3823
  M 20925  8282
> TauxErreur = 1-((mat[1,1]+mat[2,2])/sum(mat))
> TauxErreur
[1] 0.3650794
```

Illustration 25 – Erreur de prédiction du modèle PR_CART

Nous obtenons un taux d'erreur (de mal classés) de 36,5 % : **le modèle de prédiction ne peut être qualifié de « performant ».**

Remarque :
Il pourrait être intéressant d'utiliser un tel arbre pour l'évaluation de la fréquence de sinistre par exemple, en guise d'alternative aux Modèles Linéaires Généralisés.

Prime Pure et ratio de Sinistres sur Primes

Pour tarifer les contrats, nous utilisons les coefficients obtenus dans le modèle **B** suite à la mise en place des deux GLM.

Seules les données en entrée pour la variable ***gender*** seront modifiées : nous entrons les résultats prédits dans la variable « ***gender*** » au lieu d'insérer ceux renseignés par les assurés eux-mêmes.

On s'intéresse à la Prime Pure moyenne payée par classe d'âges si l'assureur utilise ce modèle prédictif (cf. Tableau 64) ainsi qu'à la politique tarifaire de son activité d'assurance (cf. Tableau 65).

Classe d'âges	Prime moyenne	B - Homme	B - Femme	Base 100	B - Base 100
1	164,61 €	183,60 €	157,60 €	100,0	100
2	110,59 €	122,35 €	107,26 €	67,2	67,2
3	103,19 €	113,67 €	100,25 €	62,7	62,7
4	93,75 €	104,55 €	90,19 €	57,0	57,0
5	75,79 €	83,02 €	73,60 €	46,0	46,2
6	76,17 €	82,21 €	75,05 €	46,3	46,6

Tableau 64– Prime Pure moyenne par classe d'âges (modèle PR_CART)

Classe d'âges	S/P	B- S/P
1	101,65 %	99,15 %
2	102,24 %	99,69 %
3	102,44 %	99,97 %
4	102,08 %	99,44 %
5	102,17 %	99,31 %
6	103,56 %	100,28 %
Total général	**102,26 %**	99,62 %

Tableau 65 - Ratio S / P par classe d'âges (modèle PR_CART)

Ce modèle reproduit quasi à l'identique la relativité des Primes Pures moyennes inter-âges. Cependant, tout comme le modèle prédictif précédemment mis en place, il sous-estime le risque encouru par l'ensemble des classes d'âges (et de manière encore plus importante que le modèle **PR_L**).

Ici le ratio des Sinistres sur Primes dépasse les 102 % pour la quasi-totalité des âges (sauf les très jeunes conducteurs) et le S / P global est bien au-dessus de celui obtenu pour le modèle **B**.

Cette sous-évaluation se fait également sentir dans le niveau des Primes Pures payées par classe d'âges : ils sont compris entre les Primes Pures différenciées du modèle **B** mais bien plus proches du tarif des femmes que de celui des hommes.

On en déduit donc que l'erreur du modèle de prédiction est majoritairement due à un mauvais classement des hommes, qui sont pris pour des femmes et qui, dans le cas où ce modèle serait mis en place, paieraient une prime inférieure au risque qu'ils présentent.

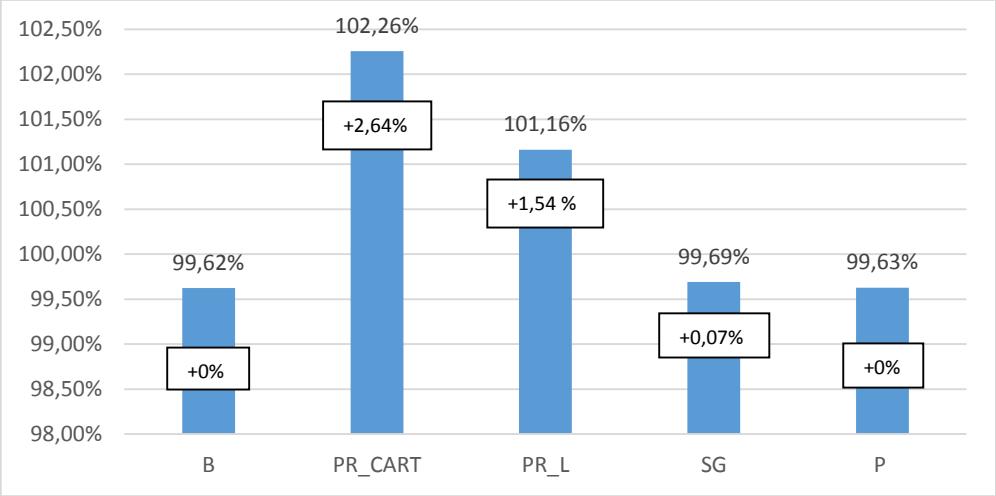
Avec un ratio de 102,26 %, le modèle mis en place sous-estime le risque présenté par l'ensemble du portefeuille et ne peut être qualifié de « bon modèle ».

Remarque :
La mise en place d'un modèle prédictif du genre de l'individu, quelle que soit la méthode utilisée, peut présenter un **risque de discrimination indirecte**.
En effet, si l'on arrive à réaffecter à chaque assuré son genre de manière assez précise et qu'il paie une prime ajustée en conséquence, la discrimination liée au genre de l'individu n'aura pas disparue.
Néanmoins, au vu des résultats obtenus (environ 37 % d'erreur), nos modèles prédictifs ne présentent pas une précision suffisante pour que les modèles puissent être considérés comme « des solutions de contournement » de la législation.

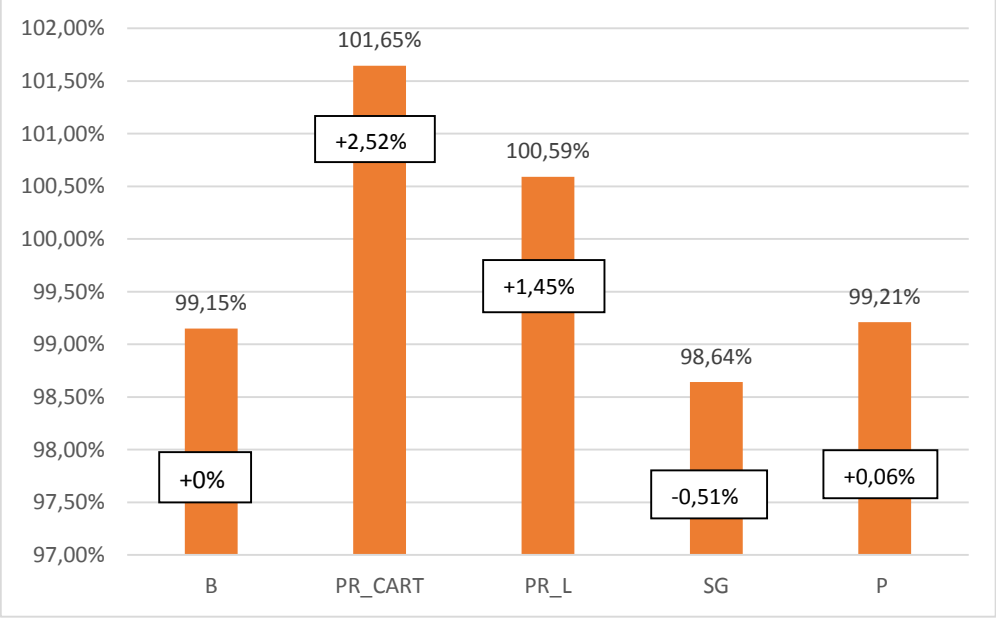
Chapitre 10 - Synthèse des résultats et conséquences de la mise en place de tels modèles

10.1 Les ratios de Sinistres sur Primes

Lors de la mise en place de nos modèles, nous avions calculé pour chacun d'eux un indicateur de la politique tarifaire à fin technique : le ratio de Sinistres sur Primes pour le portefeuille global (cf. Graphique 18) mais également par classe d'âges (cf. Graphique 19).



Graphique 18- Ratio S / P par modèle sur le portefeuille entier



Graphique 19 - Ratio S / P par modèle pour la classe d'âges 1

Les S / P globaux et ceux calculés uniquement sur la première classe d'âges semblent partager une tendance commune : les modèles de prédiction sont les moins performants.

Quant aux modèles **SG** et **P**, tous deux offrent une mutualisation du risque entre les deux genres. On soulève tout de même que le modèle de **P**ondération semble le plus adapté, preuve que l'idée la plus simple n'est pas pour autant la moins performante.

Finalement, si l'on devait classer les modèles en termes de **compétence de remplacement** en s'appuyant sur l'analyse des S/P obtenus, on aurait le classement suivant :

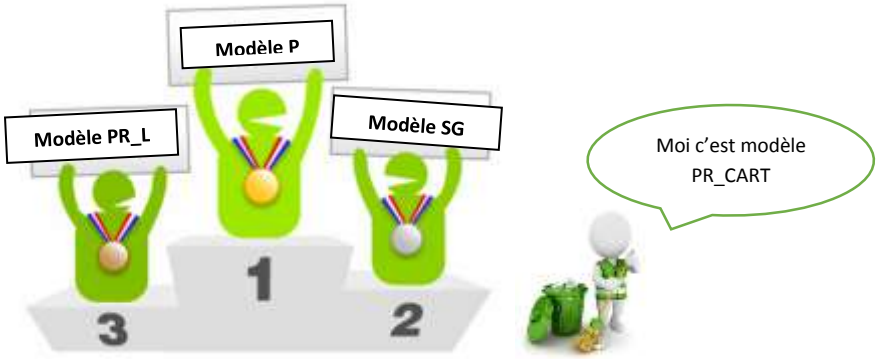


Illustration 26 - Classement des différents modèles mis en place
(Facteur de notation : S/P)

Cependant, on se doit de relativiser les résultats obtenus.
En effet, lors du calcul de ce ratio, une hypothèse importante a été réalisée : la constance de la composition du portefeuille. Autrement dit, nous avons calculé le montant total des primes et des sinistres en supposant que les assurés présents dans le portefeuille resteraient identiques et que leur niveau de sinistralité demeurerait également semblable.

Or il se peut que la mise en place d'un nouveau modèle de tarification ait pour effet indésirable d'entrainer un changement de composition du portefeuille. C'est ce que nous allons voir dans le paragraphe qui suit.

10.2 Comparaison des Primes Pures

10.2.1 Les changements tarifaires dus à la *Gender Directive*

On souhaite réaliser une étude concernant les variations de Prime Pure en fonction du modèle utilisé. Pour ce faire, nous resterons à l'échelle d'une analyse par classe d'âges et par genre et analyserons les hausses (▲) et les baisses (▼) tarifaires dues à la mise en place de chacun des modèles comparé au modèle de base.

Classe d'âges	PR_CART		PR_L		SG		P	
	H	F	H	F	H	F	H	F
1	▼ -10,3 %	▲ 4,45 %	▼ -9,41 %	▲ 5,54 %	▼ -7,61 %	▲ 7,63 %	▼ -8,14 %	▲ 7,01 %
2	▼ -9,6 %	▲ 3,11 %	▼ -8,72 %	▲ 4,13 %	▼ -7,14 %	▲ 5,93 %	▼ -7,12 %	▲ 5,96 %
3	▼ -9,2 %	▲ 2,93 %	▼ -8,46 %	▲ 3,80 %	▼ -7,50 %	▲ 4,88 %	▼ -6,79 %	▲ 5,69 %
4	▼ -10,3 %	▲ 3,94 %	▼ -9,71 %	▲ 4,66 %	▼ -7,06 %	▲ 6,57 %	▼ -7,93 %	▲ 6,73 %
5	▼ -8,7 %	▲ 2,97 %	▼ -7,20 %	▲ 4,67 %	▼ -6,30 %	▲ 5,69 %	▼ -6,40 %	▲ 5,58 %
6	▼ -7,3 %	▲ 1,49 %	▼ -5,03 %	▲ 4,03 %	▼ -4,11 %	▲ 5,04 %	▼ -4,95 %	▲ 4,12 %
Total	▼ -9,4 %	▲ 3,2 %	▼ -8,4 %	▲ 4,4 %	▼ -7,1 %	▲ 5,9 %	▼ -7,0 %	▲ 5,9 %

Tableau 66 - Hausses / Baisse tarifaires (Base comparative : modèle B)

Après analyse de ces résultats, nous pouvons d'ores et déjà prévoir que l'absence de distinction Homme/Femme dans la tarification comme le stipule la Directive Européenne va surtout poser problème aux assureurs pour la population des **jeunes conducteurs** (classe d'âges 1), en particulier pour **les jeunes femmes**.

En effet, ce sont les populations à risque, comme les jeunes conducteurs, qui sont les plus touchées : la Prime Pure des jeunes conductrices devrait augmenter de façon significative (+6,2 % en moyenne), à l'inverse celle des jeunes conducteurs devrait diminuer (-8,9 % en moyenne).

Identifier la population la plus touchée par la *Gender Directive* peut permettre d'envisager la mise en place d'une solution spécialement adaptée. Nous aurions par exemple pu poursuivre notre modélisation en supprimant du modèle la variable « genre » pour les populations les moins concernées par l'impact de la suppression de cette dernière et proposer un modèle innovant uniquement pour la classe des jeunes conducteurs.

Au global, les hommes devraient voir leur prime diminuer de 7,95 % en moyenne tandis que les femmes auront le droit à une hausse moyenne de 4,84 %.

Remarque :

On observe des Hausses / Baisses tarifaires quasi symétriques entre les deux genres pour les modèles P et SG. Cette symétrie témoigne de la mutualisation engendrée par la mise en place de ces modèles. A noter que la symétrie n'est pas parfaite car la proportion H/F n'est pas à 50 % non plus.
Par contre, pour les deux modèles prédictifs, aucune symétrie n'est visible : ces modèles favorisent les fortes baisses tarifaires pour les hommes et une hausse plus modérée pour les femmes. Ceci traduit les composantes de l'erreur de prédiction des modèles : l'erreur est due en majeure partie à une mauvaise prédiction du genre masculin.

10.2.2 Remise en question de la stabilité du portefeuille : analyse comportementale et loi Hamon

De tels changements de primes vont certainement mener à des **changements de comportement** de la part des assurés.

Par changements comportementaux, on entend :

- Des entrées / sorties du portefeuille
  - On pense notamment aux jeunes conductrices qui vont certainement se montrer frileuses quant à l'augmentation de leurs primes (puisqu'elles sont habituées à payer moins cher) et qui chercheront un assureur qui propose des tarifs plus compétitifs.
  - Mais également aux conducteurs (les hommes) qui vont tenter de faire jouer la concurrence pour obtenir la meilleure baisse, même si leur nombre sera certainement inférieur à celui des femmes ayant opté pour un changement d'assureur.

- Un changement du niveau de couverture du contrat d'assurance

  – Statistiquement parlant, les femmes sont plus averses au risque que les conducteurs masculins ce qui les pousse à se couvrir d'avantage.
  Le niveau de couverture étant le levier le plus évident pour faire baisser leur prime, elles pourraient opter pour des formules moins couvrantes (ne pas souscrire d'assurance vol de véhicule par exemple), ce qui impacterait le niveau d'encaissement de l'assureur.

  – A l'inverse, les jeunes hommes pourraient se permettre d'avoir une meilleure couverture (passage d'une assurance au tiers à une assurance tout risque par exemple).

Hausse tarifaire et sortie du portefeuille : loi Hamon

Dans le portefeuille étudié, pour les individus subissant une hausse de prime, il n'est pas question de réduire la garantie souscrite, puisque la garantie RC est obligatoire en France, mais plutôt de sortir du portefeuille de leur assureur actuel pour entrer dans celui d'un autre assureur plus attractif.

Ces mouvements inter-assureurs ont été facilités par la publication de **la loi Hamon** au Journal officiel le 18 mars 2014 : au 1er janvier 2015, il devient désormais possible pour l'assuré de résilier son contrat d'assurance automobile à tout moment après une première année pleine et ce, sans aucun préjudice.

Jusqu'alors, les contrats se renouvelaient automatiquement pour la plupart, et la résiliation de ces-derniers n'était possible qu'à échéance annuelle, sous réserve du respect du dépôt d'un préavis courant de 1 à 3 mois.

L'assuré va donc être particulièrement avisé de comparer son contrat d'assurance actuel avec ceux proposés par la concurrence et ainsi bénéficier des primes d'assurance les plus adaptées à ses besoins et à son budget.

Si l'on s'intéresse aux chiffres relevés suite à la publication de cette nouvelle loi, sur les trois premiers mois de l'année (2015), **8 %** des Français ont résilié leur assurance-automobile[1].

Les motivations sont majoritairement économiques avec 57 % des personnes interrogés qui ont souhaité souscrire un contrat moins coûteux alors que seuls 34 % ont cité l'envie d'obtenir de meilleures garanties.

A noter que plus d'un Français sur deux (58 %) a connaissance de la loi Hamon et de ses effets sur la résiliation d'assurance.

[1] Source : Sondage OpinionWay pour LesFurets.com [A9]

Partie IV - L'innovation, une nécessité

A ce stade du mémoire, la mise en place des différents modèles est terminée.
Ils nous ont permis de déterminer le niveau de Prime Pure de chacune des classes tarifaires crées.

Une étude comparative avec le modèle de base a suivi leur mise en place ainsi que la quantification de leurs impacts sur le tarif proposé aux assurés.

Toutes ces analyses ayant été menées avec une hypothèse de constance de composition du portefeuille, nous jugerons la robustesse des modèles face à une modification de cette dernière en appliquant divers Stress-Tests (Chapitre 11).

Pour finir, nous ouvrirons sur des perspectives d'amélioration de la tarification, avec l'utilisation de nouvelles variables tarifaires et ferons un point sur la situation actuelle du marché automobile en termes de méthodes de tarification (Chapitre 12).

Chapitre 11 - Robustesse des modèles

La mise en conformité des modèles de tarification à la règlementation a engendré des changements tarifaires qui posent la problématique d'une dérive de la composition du portefeuille face à des mouvements d'entrée-sortie du portefeuille de la part des assurés.

Dans le cadre de la suppression de la variable « genre », c'est à la dérive de la structure du niveau de mixité des contrats en stock que nous allons nous attacher tout particulièrement.

Au vu des résultats concernant les hausses / baisse tarifaires prévues par nos modèles (cf. Tableau 66), notre étude se **focalisera** sur la **classe d'âges 1**, à savoir les très jeunes conducteurs puisque c'est pour cette classe d'âges que les modifications tarifaires se font le plus sentir.

Dans les paragraphes qui suivent, nous allons tester la résistance des différents modèles mis en place ainsi que celle du modèle de base à un changement de composition du portefeuille.

Pour cela, trois Stress-Tests[1] seront réalisés, impliquant les scénarios suivants :

- Scénario 1 : des jeunes conducteurs entrent dans le portefeuille (§ 11.1)
- Scénario 2 : des jeunes conductrices sortent du portefeuille (§ 11.2)
- Scénario 3 : Inversion de la proportion H/F des jeunes conducteurs (§ 11.3)

Sera alors qualifié de « meilleur modèle » le modèle le plus robuste, c'est-à-dire le moins sensible aux différents chocs appliqués via ces tests.

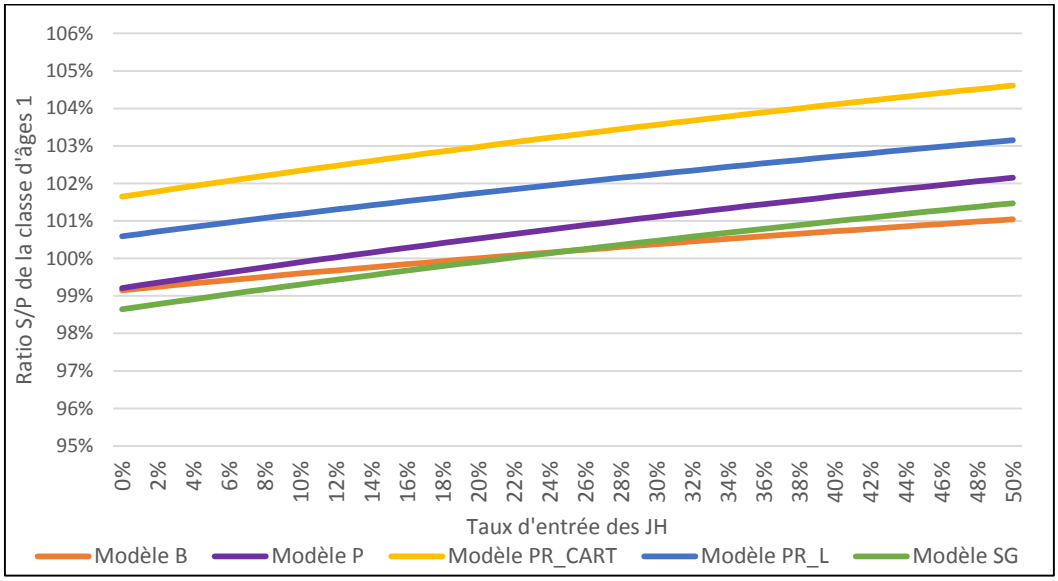
[1] Ces stress-tests ont été inspirés par ceux imposés par EIOPA aux assureurs (European Insurance and Occupational Pension Authority) en phase QIS5 de la mise en place de la Directive Solvabilité II.

11.1 Entrée de Jeunes Hommes dans le portefeuille

Conscients de la nature fortement concurrentielle du marché de l'assurance automobile, les Jeunes Hommes (JH) peuvent être tentés de changer d'assureur pour obtenir un tarif encore plus bas même après une baisse tarifaire annoncée par leur assureur.

Certains assureurs devront donc faire face à des « vagues » d'entrée de jeunes conducteurs dans leur portefeuille, vague dont l'ampleur sera plus ou moins importante en fonction du niveau de prime demandé par ce dernier.

On a donc souhaité quantifier l'impact d'un tel évènement sur l'équilibre tarifaire de l'assureur pour chacun des modèles mis en place.



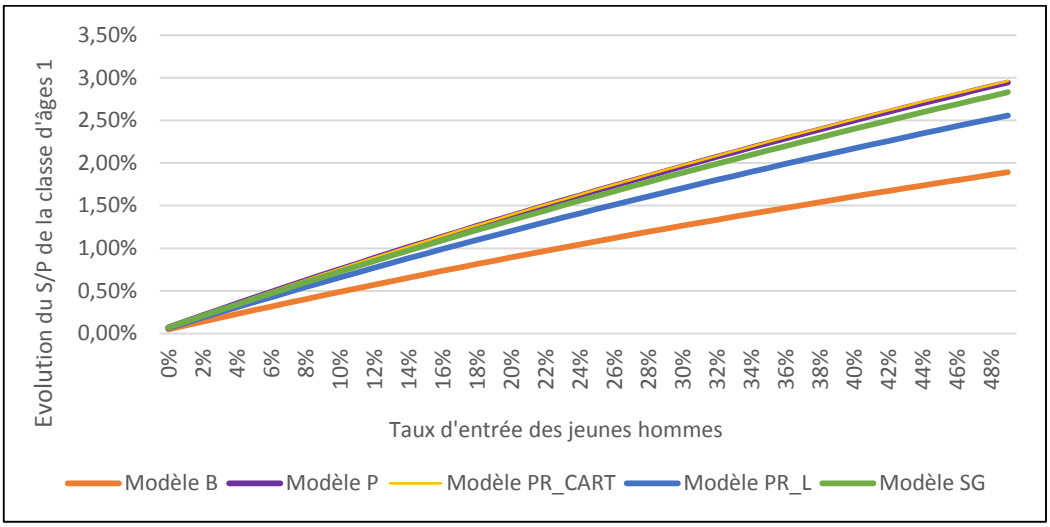
Graphique 20-Valeur du S / P sous l'hypothèse d'entrée de JH dans le portefeuille

Sur ce graphique, on observe une tendance à la hausse pour TOUS les modèles.
En effet, les jeunes conducteurs étant considérés comme des mauvais risques, leur arrivée dans le portefeuille ne peut être que défavorable à l'assureur si ce dernier ne fait pas évoluer son modèle tarifaire en conséquence.

Cependant, on peut observer que le modèle de base et celui sans genre sont les derniers à atteindre le palier des 100 % : il faut seulement un taux de 13 % d'entrée de Jeunes Hommes pour que le modèle pondéré ne soit plus rentable alors qu'il en faut près de 20 % pour le modèle de base.

Malgré tout, le graphique précédent ne permet pas de comparer efficacement les différents modèles car le Ratio de Sinistres sur Primes de chacun d'entre eux est à un niveau initial différent.

On privilégiera donc le tracé de l'**évolution** du S/P par rapport au taux d'entrée de Jeunes Hommes et non plus le **niveau** de ce dernier.



Graphique 21 - Evolution du S / P sous l'hypothèse d'entrée de Jeunes Hommes dans le portefeuille

Le constat est cette fois ci bien plus parlant : plus la pente de la fonction représentative de l'évolution du S/P en fonction du taux d'entrée des Jeunes Hommes est importante, plus l'impact de ce flux d'assuré sur la rentabilité de l'assureur est important.

C'est donc le modèle **PR**édictif construit à l'aide d'une régression **L**ogistique binaire qui est le moins impacté par cette entrée de JH dans le portefeuille puisqu'il admet la plus faible pente.

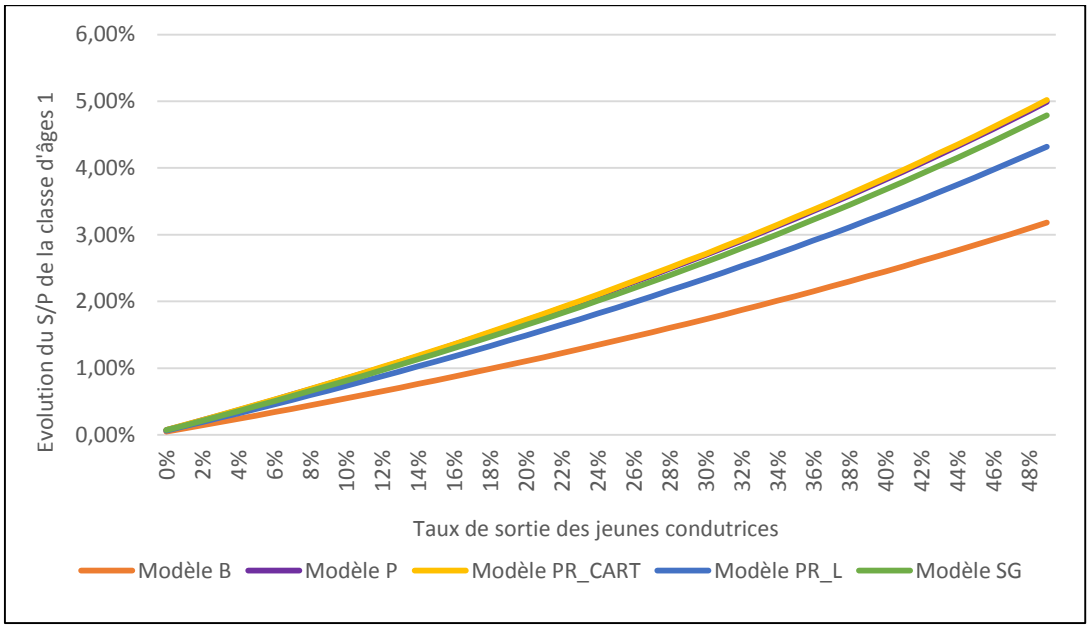
On souligne également que le modèle de **P**ondération arrive à la dernière place en termes de robustesse, suivi par le modèle **P**rédictif utilisant un arbre de **C**lassification.

On peut donc affirmer que la solution la plus simple pour l'assureur à priori n'est pas forcément la meilleure.

11.2 Sortie des jeunes conductrices du portefeuille

Cette fois ci, on suppose que ce sont les jeunes femmes qui, suite à de fortes hausses tarifaires, vont être tentées de résilier leur contrat actuel pour aller se nicher chez un assureur plus compétitif.

Comme les assureurs sont tous soumis à la même règlementation, le gain de compétitivité sera fonction du modèle tarifaire choisi par ces derniers.



Graphique 22 - Evolution du S / P classe d'âges 1 en fonction du taux de sortie des jeunes conductrices

De nouveau, on constate que le modèle de **B**ase est le plus robuste avec une hausse du S/P la plus modérée.

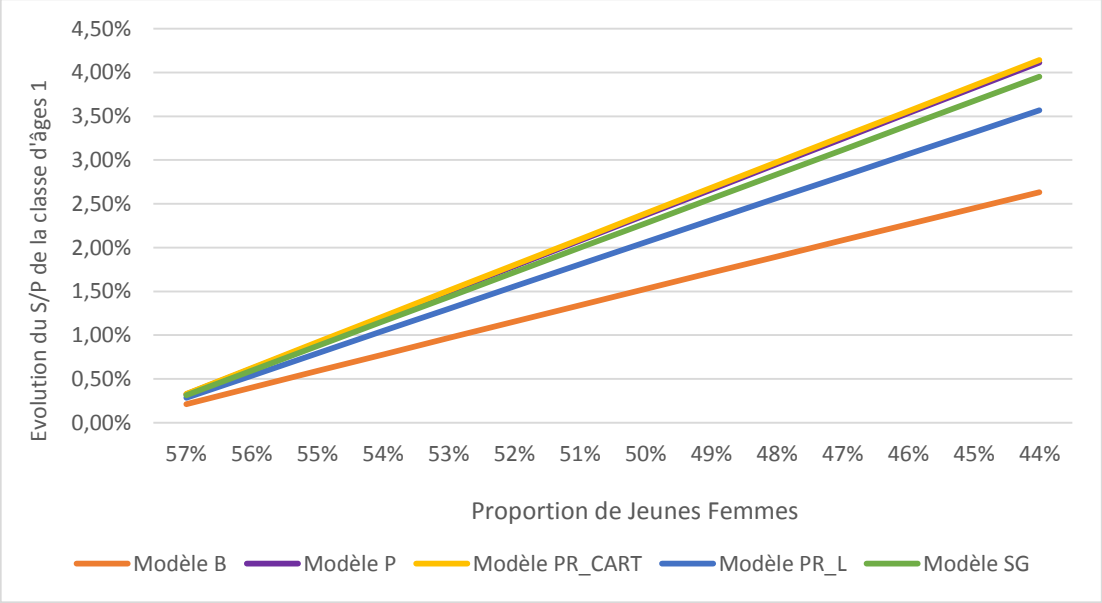
On remarque également que l'ordre de robustesse des modèles alternatifs est inchangé : les modèles les plus sensibles à une entrée de mauvais risques sont également les plus sensibles à une sortie des bons risques.

11.3 Vers une inversion de la proportion H/F des jeunes conducteurs

Pour ce dernier choc, nous décidons de considérer les entrées et sorties simultanément.
C'est le Stress-test qui est le plus fidèle à ce que l'on observe en pratique.

Notre classe d'âges 1 correspondant aux jeunes conducteurs, est composée à 56,71 % de jeunes conductrices et donc à 43,29 % de jeunes conducteurs.

Nous allons quantifier l'effet d'une inversion progressive de ces taux, de façon à amener le taux de jeunes conducteurs à 56,71 % et celui de jeunes conductrices à 43,29 %.



Graphique 23 - Evolution du S/P classe d'âges 1 en fonction de la proportion de femmes dans la classe d'âges 1

C'est sans grande surprise que l'on observe que le modèle de **B**ase est là encore, le moins sensible à l'inversion des proportions Homme/Femme et que le modèle prédictif fondé sur une régression logistique binaire est le plus robuste.

On peut également souligner la tendance linéaire de l'évolution du S/P pour chacun des modèles lorsque l'on considère les entrée/sorties simultanément.
On observait une tendance logarithmique pour les entrées (cf. Graphique 21) et une tendance exponentielle pour les sorties (cf. Graphique 22).

Enfin, on se doit de faire remarquer que si les impacts de ces tests sont tout de même modérés (maximum de +6 % sur le ratio S/P) et n'ont pas un impact bien plus important, c'est parce qu'aussi bien les départs que les arrivées dans le portefeuille ont été choisis de manière aléatoire.

En réalité, ce sont surtout les bons risques (ceux qui n'ont pas eu d'accident) qui vont se décider à sortir du portefeuille, auquel cas la rentabilité de l'assureur pourrait être bien plus menacée.

11.4 Conclusion : Comparaison des modèles et Stress-Tests

La suppression d'une variable tarifaire a finalement peu d'effet sur la **qualité** du modèle mis en place.

D'un point de vue technique, qu'elle que soit la variable supprimée, la théorie des Modèles Linéaires Généralisés étant basée sur une estimation de la moyenne de la variable à expliquer, il sera toujours possible d'adapter les valeurs des coefficients de la régression pour que le modèle mis en place reflète « en moyenne » la variable à expliquer, et , de but en blanc, permette l'obtention d'un ratio de Sinistres sur Primes très proche des 100 %, valeur attendue par l'assureur.
En effet, le Théorème Central Limite assure toujours la viabilité d'un tel système à un niveau de confiance fixé à l'avance.

En pratique, cela signifie que l'on pourra toujours ajuster une loi à notre charge de sinistre mais la mutualisation qui devra s'opérer au sein des nouvelles classes tarifaires formées devra être plus importante.
Cependant, l'introduction d'hétérogénéité dans le portefeuille ne change pas fondamentalement la logique de redistribution de la charge des sinistres au sein des assurés.

Mais attention à ne pas se méprendre, il faut bien garder à l'esprit que la création d'un modèle de tarification ne doit pas rester uniquement théorique (avec l'ajustement du meilleur modèle en termes de « conditions mathématiques à respecter ») mais bien prendre en compte l'environnement dans lequel ce modèle sera appliqué.

Face à cet enjeu, la **suppression d'une variable tarifaire** a cette fois ci un **impact important**.
En effet, cette dernière, en causant une modification de la segmentation du portefeuille de l'assureur, a aussi provoqué sur son passage des modifications tarifaires.

Ces modifications seront plus ou moins importantes en fonction de la variable supprimée et seront suivies de changements de comportement des assurés.

Est considéré comme « bon modèle » un modèle qui ne serait pas sensible aux changements de structure du portefeuille, autrement dit, un modèle **robuste**.

Les résultats des Stress-Tests témoignent de la perte de robustesse induite par la mise en place de modèles alternatifs : nous avons obtenu pour chacun des trois scénarios testés une courbe représentative de l'évolution du S/P du modèle de **B**ase bien éloignée de celles des modèles alternatifs.

Même si ces derniers, pour leur part, étaient assez proches, c'est le modèle de prédiction utilisant la mise en place d'un modèle de régression logistique qui s'est avéré être le plus robuste.

Bien que la qualité de la prédiction ne soit pas suffisante (37 % d'erreur), la présence du critère genre dans la segmentation est préférée à sa suppression totale (modèle SG) : l'assureur devra donc chercher des proxys pour remplacer cette variable, ou encore mettre en place un modèle de prédiction performant.

Illustration 27 - Classement des différents modèles mis en place
(Critère de notation : robustesse)

La suppression d'une variable tarifaire donne également un nouveau challenge à l'assureur : être capable de modéliser (d'anticiper) le comportement de ses assurés face aux Hausses / Baisses tarifaires de façon à construire un modèle résistant.

L'assureur pourrait ainsi par exemple mettre en place un modèle de « prédiction du taux de résiliation » des contrats automobiles qui tiendrait compte de l'ancien tarif ainsi que le nouveau, du tarif moyen proposé par les concurrents mais également de l'ancienneté d'assurance du client.

Outre le problème d'estimation des primes moyennes appliquées par ces concurrents, la mise en place d'un tel modèle nécessiterait un historique important pour assurer une certaine robustesse de la prédiction.
Face au caractère récent de la Directive Anti-Discrimination, cet historique est à l'heure actuelle indisponible.

Remarque :
Ce besoin de modélisation rejoint une problématique rencontrée en Assurance-Vie : la **modélisation du taux de rachat des contrats**.
L'enjeux est néanmoins différent puisqu'en Assurance-Vie, une hausse des taux de rachat est redoutée du fait que les primes soient placées tandis qu'en assurance automobile, il s'agit d'une question de rentabilité de l'assuré après étalement des frais de gestion et des coûts d'acquisition (campagne marketing, comparateur, etc.) sur la durée et non d'un problème lié à des placements financiers.

Chapitre 12 - Transformer les enjeux règlementaires en opportunités

La mise en application de la *Gender Directive* en 2012 a obligé les assureurs à utiliser des critères autres que le genre de l'assuré pour l'élaboration de leurs tarifs. Cette mesure Anti-discrimination pourrait dans un futur proche, s'étendre à une deuxième information jugée comme discriminante : l'âge de l'assuré.

L'assurance automobile est un secteur particulièrement touché par ces mesures car, comme nous l'avons vu, les comportements inter-âges et inter-genres sont très différents au volant et sont donc déterminants lors de la détermination du tarif.

L'ingéniosité et l'innovation des assureurs ont étés mises à l'épreuve face à ces nouvelles contraintes règlementaires : il a fallu trouver un moyen de segmenter tout aussi bien leur portefeuille sans pour autant déroger à la règle.

Les modèles mis en place dans ce mémoire, n'ayant pas témoignés d'une efficacité suffisante, ont sûrement étés mis en place par les assureurs dans les mois suivant la mise en application de la *Gender Directive* pour une durée déterminée, laissant le temps à d'autres pistes plus fructueuses d'être exploitées.

Actuellement, l'une des pistes explorée pour répondre à cet enjeu de suppression de variable est de se tourner vers l'étude de nouvelles variables tarifaires corrélées à la sinistralité afin de les appliquer dans le cadre d'une tarification individualisée.
La segmentation étant ainsi poussée à l'extrême, l'assureur aura une analyse plus fine de son risque mais met dans le même temps entre parenthèses l'un des principes fondateurs de l'assurance, à savoir, la mutualisation des risques.

On va donc à présent se pencher sur cette nouvelle piste avec dans un premier temps, l'exposition des variables innovantes, encore à l'étude, qui sont liées à la sinistralité en automobile.
Puis nous introduirons l'assurance automobile à l'usage, « *le Pay How You Drive* », en présentant son fonctionnement mais également ses atouts.
Enfin, nous terminerons en dressant un état des lieux des modèles de tarification en assurance automobile.

12.1 L'utilisation de variables innovantes en assurance automobile

Pour faire face à ces contraintes grandissantes en termes de législation, une alternative sérieusement envisagée pour les assureurs a été de se tourner vers le *Big Data*.

En effet, avec la capacité à construire des modèles complexes et de plus en plus personnalisés, les assureurs pourraient s'affranchir des contraintes règlementaires.

Certains assureurs ont donc fait le choix de totalement modifier l'esprit des variables utilisées pour la tarification : ces nouvelles variables seront présentées dans le paragraphe qui suit.

12.1.1 Les variables innovantes

Nous avons vu que le genre de l'individu impacte le niveau de sinistralité, notamment le coût des sinistres. En réalité, le genre est seulement un indicateur (proxy) sur la capacité à conduire.
Ce sont les comportements inter-genres qui sont différents et qui expliquent une différence de sinistralité.

Les assureurs ont bien compris que, s'ils peuvent surveiller le comportement du conducteur par le risque individuel, le genre est nul et non avenu.
L'idée est donc de se pencher sur de nouvelles variables, fortement corrélées non pas au véhicule ou à l'assuré, mais à ses habitudes de conduite à proprement parlé.

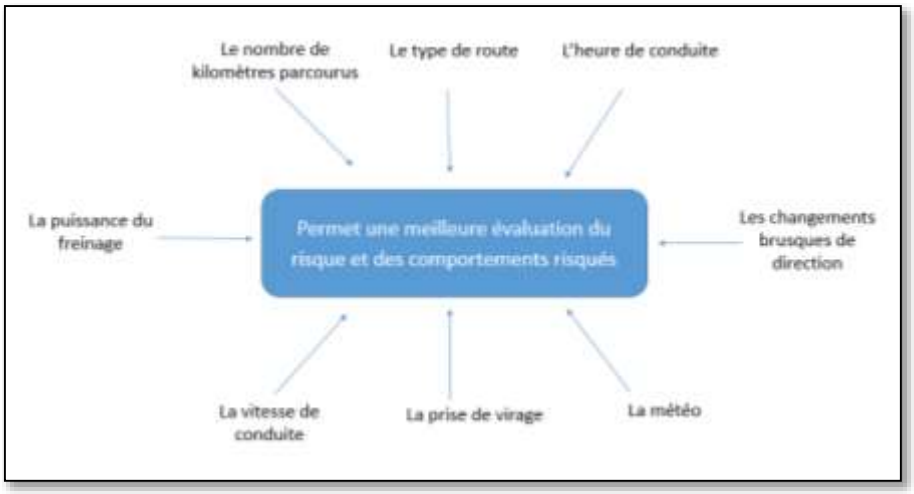


Figure 9 - Les variables tarifaires à l'étude

Il s'agit en premier lieu de **la fréquence de conduite**, mesurée par le biais du **kilométrage**.
Le nombre de kilomètres parcourus par an est fréquent dans le questionnaire de souscription en assurance et il est directement corrélé au tarif.
Plus on utilise son véhicule, plus le risque d'accident est important et plus la Prime Commerciale augmente.

Il ne s'agit pas ici d'indiquer une fourchette (comme il en était question précédemment) mais d'avoir une mesure exacte au kilomètre près de la distance parcourue par l'assuré : réduire son kilométrage permettrait alors de réaliser des économies.

Les assureurs se penchent également sur **les heures d'utilisation du véhicule**, une conduite de nuit étant statistiquement plus dangereuse qu'une conduite en plein jour, ainsi que sur les conditions météorologiques.

De nombreux autres critères sont en cours d'étude (cf. Figure 9) : leur rôle sur le niveau de sinistralité est irréfutable mais n'a pas encore été quantifié.

12.1.2 Une segmentation illimitée

Alors que les variables explicatives étaient limitées aux caractéristiques de l'assuré ou de son véhicule ainsi qu'aux évènements passés[1], il devient aujourd'hui possible de faire payer l'assuré en fonction de son comportement en temps réel.

[1] Lors de la tarification à postériori avec l'application du Bonus-Malus.

Il est ainsi envisageable 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 (le temps de réaction étant deux fois plus long au-delà de deux heures de conduite, il est conseillé de faire une pause toutes les deux heures).

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.

Selon cette nouvelle logique, un assuré conduisant essentiellement de nuit et sur des routes réputées dangereuses s'acquittera d'une prime supérieure à son voisin effectuant ses trajets principalement de jour et sur des autoroutes statistiquement moins accidentogènes.

Remarquons que les Modèles linéaires Généralisés (modèles traditionnellement utilisés) ne sont alors pas adaptés à l'intégration de telles variables : elles sont pour la plupart quantitatives continues, et le cas échéant, qualitatives avec un nombre important de modalités.

On rappelle que certaines variables présentes dans la base étudiée dans ce mémoire présentaient les mêmes limites en termes d'intégration dans un GLM (variable quantitative ou nombre de modalités trop important) et ont dues être retraitées en amont de la tarification.

Ces retraitements ont abouti à des regroupements qui permettent de rendre utilisable le modèle en pratique, mais font également perdre en terme de qualité de segmentation : le but ici est inverse et de tels traitements ne devront donc pas être réalisés.

Les assureurs vont devoir mettre au point des modèles innovants pour pouvoir utiliser des variables de ce type, il pourrait s'agir par exemple de l'utilisation d'Arbres de Régression.

S'ils y parviennent, ce sera le début d'une tarification individualisée et la fin de la mutualisation, puisque les classes tarifaires auront disparues.

Ceci entre dans l'état d'esprit ambiant de la société dans laquelle nous nous trouvons : payer uniquement pour ce que l'on consomme.

12.2 Les offres en vogue : le *« Pay How You Drive »*

Dans un premier temps, il n'a pas été question pour les assureurs de supprimer les anciennes variables tarifaires. Au contraire, le nouveau système de tarification devrait en contenir bien plus que le précédent système.

Le nouveau concept d'assurance automobile en vogue est le « *Pay How You Drive* » (PHYD), qui signifie littéralement « Payez comme vous roulez ». Son objectif est de faire baisser la prime d'assurance automobile en déterminant le niveau de prime en fonction du comportement de l'assuré au volant.
Ce comportement est relevé à l'aide d'un boitier installé dans le véhicule : un besoin en télématique[1] est né.

[1] La télématique est un terme qui recouvre les applications associant les télécommunications et l'informatique, apparu en France à l'occasion de la filière technologique qui allait donner vie au Minitel.

Ce besoin devrait facilement être comblé puisque le coût de la technologie télématique est réduit. De plus, la prochaine génération de conducteurs (génération iPhone) est plus encline à utiliser la technologie.

L'utilisation de cette nouvelle technologie devrait également apporter une aide en matière de détection de fraude et de prévention des accidents.

	Assurance traditionnelle	Pay How You Drive
Risque indexé sur le kilométrage	Impossible	Possible
Risque indexé sur le moment de la journée	Impossible	Possible
Risque indexé sur le type de routes	Impossible	Possible
Risque indexé sur les conditions météo	Impossible	Impossible
Intrusion dans la vie privée	Faible	Elevé
Coût des infrastructures	Faible	Elevé
Capacité de précision actuarielle et de maîtrise des risques	*	***

Tableau 67 - Méthodes de tarification et gestion du risque

Remarque:
Le *Pay How You Drive* succède au « *Pay As You Drive* » (litt. « Payez autant que vous roulez »). Le principe de ce dernier était similaire au PHYD à la différence près que seul le kilométrage est pris en compte : l'assuré paie une prime fonction du nombre de kilomètres parcourus et son comportement au volant (freinage, agressivité..) ainsi que l'environnement de conduite (type de route, conduite de nuit, etc.) n'est pas pris en compte.

12.2.1 Fonctionnement[1]

La mise en œuvre d'une offre *Pay How You Drive* nécessite différents éléments technologiques de collecte, de réception et d'émission de données :

- Le boîtier installé dans le véhicule demeure l'instrument majeur. Ce boitier, appelé odomètre, est un instrument de mesure qui permet de connaître la distance parcourue par le véhicule. Dans le cadre du *Pay How You Drive*, celui-ci est relié à un GPS (Géopositionnement Par Satellite) et transmet les informations via le réseau de téléphonie mobile (GSM).

- Un satellite GPS est utilisé pour localiser en permanence le véhicule. Les informations sur la localisation du véhicule et sa vitesse sont transmises et stockées en temps réel dans le boîtier. Celui-ci transmet ces informations à l'assureur qui pourra ensuite les convertir et chiffrer les primes personnalisées.

- Pour mettre en œuvre le *Pay How You Drive*, les assureurs doivent faire appel à différents acteurs. Des prestataires de services télématiques pour la transmission de l'information, des fournisseurs et des installateurs de boîtiers, des sociétés de services informatiques et des éditeurs de logiciels pour gérer les contrats *Pay How You Drive*. Enfin, le *Pay How You Drive* nécessite aussi des capacités informatiques de traitement de données très puissantes.

[1] Paragraphe issu du Livret blanc ITN [R11]

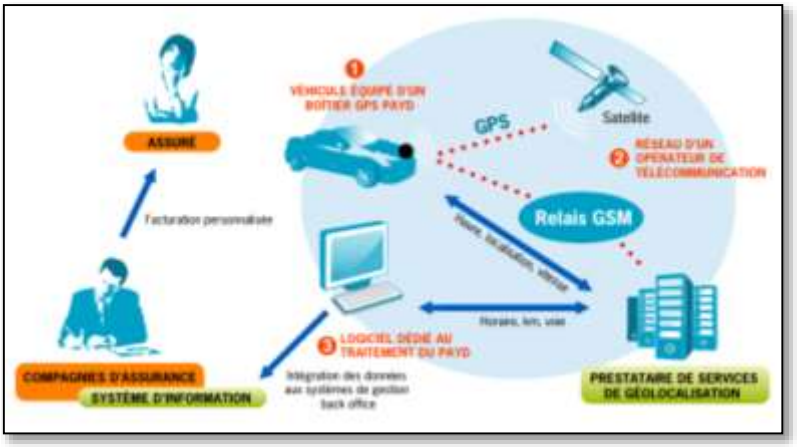


Illustration 28 - Fonctionnement du PAYD[1]

12.2.2 Une population cible : les jeunes conducteurs

Le choix du *Pay How You Drive* est motivé par l'envie de segmenter encore plus dans le but de proposer des tarifs au plus juste, pour toutes les catégories d'âges. L'adoption de la *Gender Directive* a alors été une motivation de plus pour les assureurs de creuser dans cette direction.

Néanmoins, le lancement de cette nouvelle offre n'est pas généralisé : le projet de développement a comme population cible les jeunes conducteurs.

En effet, ce sont ceux qui sont le plus familiarisé avec les nouvelles technologies mais également, la population la plus susceptible d'être intéressée par une ristourne sur prime au vu des tarifs dont elle doit s'affranchir.

Cette offre dispose déjà d'un petit historique de lancement puisque c'est Solly Azar, 2ème courtier grossiste en France, qui a lancé le premier PAYD français à destination des particuliers de 18 à 25 ans sous le nom d' « *Easy Drive »* en Juin 2008[2].

Plus récemment (Juin 2014), Direct Assurance, filiale d'**Axa** spécialisée dans la vente sur internet, a fait un premier pas dans l'assurance automobile tarifée en fonction du comportement du conducteur en lançant son contrat « *You Drive*[3]» pour les assurés de 18-24 ans.

Plus qu'une assurance, un éducateur de conduite

L'occurrence d'un sinistre pour les jeunes conducteurs résulte de la combinaison d'un manque d'expérience et d'une attitude insouciante, voir dangereuse, au volant.

Pour ce qui est de l'expérience, elle s'acquière avec le temps et l'assureur ne peut apporter de solution à ce facteur de sinistralité. Comme dans toute discipline, une pratique est nécessaire pour mettre en place les automatismes et devenir meilleur.

Par contre, par le biais de cette nouvelle offre, l'assureur touche le deuxième facteur de sinistralité : l'attitude du conducteur au volant.

[1] Source : [R11]
[2] Source : [R7]
[3] Source : [S13]

On conviendra tous qu'il est plus difficile de changer de mauvaises habitudes que de ne pas les prendre, l'assureur travaille donc dans ce sens en mettant l'assuré « à l'épreuve » à chaque prise de conduite.

L'aspect psychologique de cette méthode est non-négligeable : en donnant le plein pouvoir au conducteur de définir son niveau de prime, l'assureur est capable d'influencer les comportements, amenant les assurés à conduire moins, et plus prudemment. Ainsi, il ne sera plus question de « jeter la faute » sur l'assureur pour justifier un tarif élevé mais plutôt d'adapter son comportement au volant dans le but de le réduire.

Le modèle de tarification ici proposé est transparent : l'assuré a conscience des paramètres entrant en jeu dans l'évaluation de sa prime d'assurance, ce qui permet de le sensibiliser d'autant plus à ces facteurs de sinistralité.

Cette nouvelle offre permet donc la mise en place d'un cercle vertueux « parfait » aussi bien pour l'assureur que pour l'assuré.

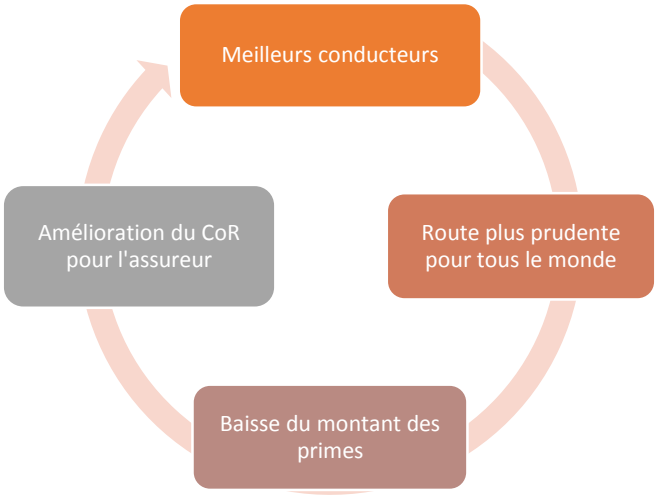


Figure 10 - Cercle vertueux permis par le PHYD

Conclusion

Détenir le maximum d'informations pertinentes sur le risque à assurer est essentiel pour l'assureur. En effet, c'est grâce à ces dernières qu'il peut évaluer le montant de la prime à faire payer à chacun des assurés.

En assurance automobile, il peut s'agir de renseignements concernant le type de contrat, l'assuré et son passé de sinistralité, tels que son âge, son lieu de domiciliation, son coefficient de Bonus-Malus ou encore son genre, mais également de renseignements concernant son véhicule, tels que sa marque, son prix ou encore sa puissance.
L'assureur est alors libre d'en disposer à sa guise pour tarifer son assuré et lui proposer un montant de Prime Commerciale représentative du risque qu'il représente.

Mais voilà, la *Gender Directive*, publiée en décembre 2012, oblige les assureurs à se détacher de l'utilisation du critère « genre » dans leur tarification des contrats en assurance automobile.
La suppression de cette variable du système de tarification pose problème aux assureurs automobiles puisque le genre est un facteur de risque reconnu statistiquement, notamment chez les conducteurs novices : les jeunes hommes sont plus dangereux que les jeunes femmes au volant.
Bien que cette différence de profil de risque soit avérée, le genre reste tout de même considéré comme une variable discriminante et son usage est proscrit. L'âge pourrait d'ailleurs être le prochain critère amené à disparaitre.

Face à cette suppression règlementaire, une problématique nouvelle est créé pour l'assureur, qui est l'objet de ce mémoire, à savoir, proposer des solutions de tarification alternatives face à cette suppression qui permettent de limiter au maximum les impacts de la Directive.

C'est muni d'une base de données de sinistralité automobile que nous avons souhaité illustrer le cas de la suppression de la variable « genre ».

Comme prérequis en amont de toute tarification, nous avons réalisé un travail de contrôle de qualité et traitement de nos données. Cette opération n'étant pas spécifique à la problématique développée, nous ne nous attardons pas sur ce sujet.

Pour commencer, nous avons mis en place un modèle tarifaire non contraint par la *Gender Directive* : il nous a servi de base comparative dans toute la suite du mémoire.
Ce modèle assure la segmentation du portefeuille en classes de risque les plus homogènes comparées à tous les autres modèles mis en place, car c'est celui qui prend en compte le plus d'informations sur l'assuré et son véhicule.

Puis, différents modèles alternatifs sont proposés afin d'appréhender cette suppression de variable. Ces derniers, dans le souci d'être conformes à la nouvelle règlementation, ne tiennent plus directement compte du genre de l'assuré. On distingue alors :

- un modèle consistant en la simple suppression de la variable « genre » des données en entrée ;
- un modèle de pondération par la proportion Homme / Femme du portefeuille, accentuant le principe de mutualisation du risque en Assurance.
- deux modèles de prédiction du genre de l'assuré, avec l'utilisation de techniques diverses, à savoir, une régression logistique binaire ainsi qu'un arbre de classification.

Ceci étant fait, nous nous sommes intéressés à la qualité des différents modèles, comparée au modèle de **B**ase, utilisé avant la mise en place de la *Gender Directive*.

Nos analyses des résidus du modèle, de la déviance, mais également du ratio de Sinistres sur Primes pour chacun des modèles ont révélé une proximité significative en terme de qualité des modèles.
Cette perte de segmentation n'a donc pas directement induit une politique de tarification différente sur la totalité du portefeuille de l'assureur puisque les modèles arrivent toujours à reproduire un effet « moyen », autrement dit, à mutualiser le risque, même au sein de populations non-homogènes.
C'est d'ailleurs le modèle de mutualisation du risque entre les deux genres par **P**ondération en proportion Homme / Femme du portefeuille qui est mis en avant en terme de capacité de reproduction de la charge **moyenne**, autrement dit, lors de l'analyse du ratio des Sinistres sur Primes.

Néanmoins, ceci nous éloigne du principe de l'assurance, qui consiste à **former des mutualités présentant le moins d'hétérogénéité possible**, et ceci dans le but, d'avoir un modèle robuste en cas de changement d'un des paramètres pris comme hypothèse dans le modèle tarifaire.
En effet, la mise en place de ces différents modèles a entrainé une hausse tarifaire de l'ordre de 5,6 % pour les femmes et une baisse de 7,4 % en moyenne pour les hommes : on prévoit alors des flux inter-assureurs de la part des assurés désireux de payer la prime la plus basse.

Nous avons a alors remarqué que le segment le plus touché par les modifications tarifaires était celui des très jeunes conducteurs, les novices, ce qui a orienté notre choix d'étude sur cette population spécifique.

Il a donc été légitime de s'intéresser à la robustesse des modèles face au comportement des jeunes assurés induit par la mise en place de ces nouvelles méthodes de tarification.
Pour ce faire, des Stress-Tests modifiant le paramètres de proportion constante de Jeunes Hommes / Jeunes Femmes dans le portefeuille ont ainsi étés réalisés mesurant la robustesse de l'équilibre tarifaire de nos modèles.

Trois scénarios ont été étudiés et pour les trois chocs appliqués, la conclusion est identique : **c'est le modèle de prédiction basé sur une régression logistique binaire qui est le plus robuste des modèles alternatifs.**

Cependant, la robustesse de ce modèle est encore loin de celle du modèle de Base car le modèle de prédiction utilisé admet un fort taux de mal classés (de l'ordre de 37 %).
Face à ce résultat, on comprend alors l'importance de réussir à construire de bons « proxys » de la variable tarifaire supprimée.

Il est bon de souligner que la variable « genre » n'admettait que deux modalités. La suppression d'une tout autre variable en admettant davantage présenterait une difficulté supplémentaire (et nouvelle) pour l'assureur. En effet, la probabilité de se tromper sur l'estimation de cette dernière est alors bien plus importante.

En réponse à ce nouveau challenge, une solution bien différente de la recherche de « proxys » a vu le jour chez les assureurs ayant le plus de moyens : ouvrir les modèles de tarification à l'innovation, avec l'utilisation de nouvelles variables tarifaires mais également de nouveaux principes.
Si aujourd'hui, les assureurs ont répondu à la problématique imposée par la *Gender Directive*, la suppression de la variable tarifaire « genre » les a d'autant plus motivé à développer une nouvelle méthode de tarification pour combler la suppression de variable clé et disposer d'une analyse plus fine du risque.

Ainsi, ils se penchent sur une individualisation des tarifs et du matériel technologique est à présent nécessaire au relevé d'informations émanant directement du véhicule de l'assuré : c'est le développement du « *Pay How You Drive* » (litt. « Payez comme vous roulez »), qui pourrait marquer un tournant dans l'assurance traditionnelle laissant place à l'assurance connectée.

Cette « révolution » est freinée par de nombreux obstacles, qu'ils soient d'ordre économiques, car il faut être en mesure de concevoir un business model efficient, ou encore culturels, puisqu'il faut réussir à convaincre les conducteur méfiants qu'ils doivent accepter de se faire « pister » en permanence.

Enfin, un enjeu de taille pour l'assureur est également la capacité à traiter l'ensemble des données reçues via le boîtier placé dans le véhicule. Cet obstacle technologique lié au volume croissant de données comportementales à traiter pour l'assureur trouve pour le moment une solution dans l'externalisation du traitement auprès de sociétés spécialisées.

Malgré tous ces obstacles, on peut néanmoins rester optimistes quant à l'avenir du *Pay How You Drive* car ces derniers ne sont pas immuables et peuvent tomber dès lors que le client identifie un bénéfice que ce soit en matière de services (informations, prévention, assistance...) et/ou de tarif.

L'apparition du *Pay How You Drive* marque le tournant d'une page pour les assureurs automobiles et un nouveau rebond devrait se faire connaitre avec l'apparition des voitures connectées, technologie qui soulève déjà des débats quant à la responsabilité des conducteurs en cas de sinistre…

Bibliographie

[O1] BELLANGER L. et TOMASSONE R. (2014), Exploration de données et méthodes statistiques ; Data analysis et Data mining Avec le logiciel R, *Ellipses*

[O2] CHARPENTIER A., DUTANG C., (2012), L'actuariat avec R. Version numérique ;
Consultable sur : http://cran.r-project.org/doc/contrib/Charpentier_Dutang_actuariat_avec_R.pdf

[O3] COOK R.D. (1977), « Detection of influential observations in linear regression », *Technometrics*, vol. 19, p 15-18.

[O4] CORNILLON P-A. et MATZNER-LOBER E. (2010) Régression avec R, *Springer Science & Business Media*

[O5] DENUIT M. et CHARPENTIER A. (2005) : Mathématiques de l'assurance non-vie, Tome II : Tarification et provisionnement *Economica*

[O6] EMBRECHTS P., KLÜPPELBERG C., Mikosch T. (1997) Modelling Extremal Events for Insurance and Finance, *Springer*.

[07] Luzi M. (2007), Assurance IARD, *Economica*.

[O8] PARTRAT C. et BESSON J-L. (2004) Assurance non vie, Modélisation, Simulation, *Ecomomica*

[R1] BOUCHER J.P, Segmentation du nombre de sinistres en assurance : de la loi de Poisson aux modèles gonflés à zéro et à barrière, *bibliothèque Université Catholique de Louvain*

[R2] Denuit M. et al. (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

[R3] LEVEILLARD P. et al. , Influence de la partition homme/femme et de l'expérience kilométrique dans l'assurance automobile. 2014. <hal-01081759>

[R4] PIET DE JONG et SYDNEYGILLIAN Z. HELLER, Generalized Linear Models for Insurance Data,
International Statistical Review, 2008, Vol.76(2), pp.315-315

[R5] SCOR Global Life SE (2012), Insurance without sex Actuarial Challenge Trends and chances for the Product World,
Consultable sur:
https://piu.org.pl/public/upload/ibrowser/120530-Gender/06_SCOR_2012-05
21%20Insurance%20without%20sex%20Warsaw.pdf

[R6] TEXIER L., Maison des Arts & Métiers (2015) Ethique de l'assurance et statistique: exemple du tarif unisexe
Consultable sur :
http://webcache.googleusercontent.com/search?q=cache:dG6_45-zNgUJ:fr.slideshare.net/Risk_and_Analysis/ethique-et-statistique-en-assurance-le-cas-du-tarif-unisexe+&cd=1&hl=fr&ct=clnk&gl=fr

[R7] DOSSIER DE PRESSE : Solly Azar Assurances lance le 1er « Pay As You Drive » français à destination des particuliers (Petit déjeuner presse, le 17 avril 2008)
Consultable sur :
www.sollyazar.com/dossiers-de-presse/dossiers-de-presse/dossier-de-presse-easy-drive-le-premier-pay-as-you-drive-a-destination-des-jeunes-conducteurs-/download.html+&cd=1&hl=fr&ct=clnk&gl=fr

[R8] Journal officiel de l'Union européenne, DIRECTIVE 2004/113/CE DU CONSEIL du 13 décembre 2004 mettant en œuvre le principe de l'égalité de traitement entre les femmes et les hommes dans l'accès à des biens et services et la fourniture de biens et services

[R9] L'institut des Actuaires, Journée d'Etude – Assurance de Personnes (15 Septembre 2011 à Deauville), Conséquences de la décision de la CJUE sur la différentiation par sexe en Assurance de Personnes

[R10] Journal officiel de l'Union européenne (2012/C 11/01)
Consultable sur ;
http://eur-lex.europa.eu/legal-content/FR/ALL/?uri=OJ:C:2012:008:TOC

[R11] «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.

[A1] « Pay As You Drive : ce qu'il rapportera vraiment », **L'Argus de l'assurance** N°7 046, 02/11/2007

[A2] Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. **Journal of Clinical Epidemiology**, 54, 979–985.

[A3] CHARPENTIER.A, La loi des grands nombres et le théorème central limite comme base de l'assurabilité ? , **Risques** n° 86 / Juin 2011
Consultable sur : https://www.ffsa.fr/webffsa/risques.nsf/html/.../Risques_86_0019.htm

[A4] Planchet.F et Leroy.G, Problématiques associées au partage des risques, **la tribune de l'assurance** • n° 160 • juillet-août 2011

[A5] TUFFERY S., La fin de la tarification selon le genre aura des conséquences négatives pour les consommateurs (CEA), Décembre 2011, **L'Argus de l'assurance**
Consultable sur :
http://www.argusdelassurance.com/a-la-une/la-fin-de-la-tarification-selon-le-genre-aura-des-consequences-negatives-pour-les-consommateurs-cea.52940

[A6] Article Newsletter du LAB - Décembre 2012- La Directive Gender : un renforcement des synergies entre actuariat et marketing rendu incontournable, **LAB**
Consultable sur :
http://www.cerclelab.com/les-newsletters-du-lab/107/809-2012-12-art-directive-gender.html

[A7] THEVENIN L. (2014), Auto, habitation, santé : les tarifs 2015 de l'assurance vont encore augmenter, **Les Echos**,
Consultable sur :
http://www.lesechos.fr/28/11/2014/LesEchos/21824-120-ECH_auto--habitation--sante---les-tarifs-2015-de-l-assurance-vont-encore-augmenter.htm

[A8] ACEDO S. (Août 2014), Assurance auto : hausses de tarifs à prévoir en 2015 (étude Xerfi), **l'Argus de l'assurance**
Consultable sur :
http://www.argusdelassurance.com/institutions/assurance-auto-hausses-de-tarifs-a-prevoir-en-2015-etude-xerfi.81258

[A9] Estimations de l'effet loi Hamon, (04/06/2015), **les Echos**
http://www.lesechos.fr/journal20150604/lec2_finance_et_marches/021110297884-loi-hamon-8-des-francais-ont-resilie-leur-assurance-auto-au-1-er-trimestre-1125039.php

[M1] GONNET G. (2010), Etude de la tarification et de la segmentation en assurance automobile, mémoire d'actuariat, *Université Claude Bernard – Lyon 1*

[M2] KOLASA S., PSAUME J. (2011), Tarification en IARD et nouvelles contraintes de rentabilité : Etude d'un produit Flotte Automobile, Mémoire d'actuariat, *CEA*

[M3] MAISONNEUVE B. (2012), Conséquences de l'interdiction de pratiquer des discriminations en assurance selon le sexe de l'assuré sur la tarification en Prévoyance, Mémoire d'actuariat, *Université Paris Dauphine*

[P1] DAUDIN J.J et al. (2007), « Bases du modèle Linéaire », *Polycopié AgroParisTech*

[P2] JOHNSON Paul E. (2006), Residuals and analysis of fit
Consultable sur : http://pj.freefaculty.org/guides/stat/Regression-GLM/GLM2-SigTests/GLM-2-guide.pdf

[P3] JOHNSON Paul E. (2006), *The Hat Matrix and Regression Diagnostics*
Consultable sur : http://pj.freefaculty.org/guides/stat/Regression/RegressionDiagnostics/OlsHatMatrix.pdf

[P4] RAKOTOMALALA R. (2015), Pratique de la Régression Linéaire Multiple Diagnostic et sélection de variables, *Université Lumière Lyon 2*
Consultable sur : http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

[P5] STILGENBAUER J-L. (2014), Informatique & Analyse de Données

[P6] SURU A. (2015), Principes de base de l'assurance dommage

[S1] www.securite-routiere.gouv.fr/.../Bilan+provisoire+ONISR+2014.pdf
(Bilan provisoire : L'accidentalité routière en 2014), site consulté le 13 Mai 2015

[S2] www.confused.com/~/media/docs/eu-gender-directive factsheet.pdf+&cd=1&hl=fr&ct=clnk&gl=fr
(Effects of the EU *Gender Directive* on Motor Insurance, Life Insurance & Annuities (5 Décembre 2012)), site consulté le 18 Mai 2015

[S3]www.assuralia.be/fileadmin/content/stats/03_Cijfers_per_tak/01_Auto/08_Evolutie_schadefrequentie/FR/Evolution_fr%25E9quence_des_sinistres_RCAutomobile_FR_v2.pdf+&cd=1&hl=fr&ct=clnk&gl=fr
(Evolution de la fréquence des sinistres en assurance RC automobiles, 2008, Assuralia), site consulté le 26 Juin 2015

[S4] http: //freakonometrics.blog.free.fr/
(CHARPENTIER A., Statistique de l'assurance, STT 6705, Statistique de l'assurance II (2010)), site consulté le 4 Mai 2015

[S5] http://www.insuranceeurope.eu/uploads/Modules/Publications/oxera-study-on-gender-use-in-insurance.pdf
(The impact of the ban on the use of gender in insurance, Oxera (2011)), site consulté le 10 Juillet 2015

[S6] http://www.agira.asso.fr/content/bureau-central-de-tarification
(Bureau Central de Tarification (BCT), AGIRA), site consulté le 15 Avril 2015

[S7] www.associationfrancaisedelassurance.fr
(Association Française de l'Assurance (2012), Données relatives aux différences entre les hommes et les femmes dans les prestations en assurance), site consulté le 15 Avril 2015

[S8] http://www.ffsa.fr/sites/jcms/fp_8801/les-etudes-et-statistiques
(Site de la FFSA GEMA), site consulté le 15 Avril 2015

[S9] http://grasland.script.univ-paris-diderot.fr/STAT98/stat98_8/stat98_8.htm
(Tableau de contingence et test du Chi-deux), site consulté le 22 Juillet 2015

[S10] http://fbegin.profweb.ca/ZEA/TestVarFisher.htm
(Test d'égalité de deux variances pour le test de Fisher), site consulté le 22 Juillet 2015

[S11] http://www.insee.fr/
Site de l'INSEE, site consulté le 20 Avril 2015

[S12] http://www.lmpt.univ-tours.fr/~gallardo/Stat2008-2.pdf
Test d'ajustement d'un échantillon à une loi théorique, site consulté le 9 Juin 2015

[S13] http://www.youdrive.fr/
(Offre You Drive de Direct Assurance), site consulté le 25 Août 2015

[I1] http://www.math.univ-metz.fr/~bonneau/STAT0607/table_khi2_complete.pdf
(Table du Khi-deux), site consulté le 22 Juillet 2015

Annexe 1 - Démonstrations

Propriété 1 :

Si N et C admettent des moments d'ordre un, il en résulte :

$$\boldsymbol{E[S] = E[N] \times E[C]}$$

Démonstration :

$$\begin{aligned} E(S) &= E[E(S|N)] \\ &= P(N=0)E(S|N) + \sum_{n=1}^{+\infty} P(N=n)\, E(S|N=n) \\ &= \sum_{n=1}^{+\infty} P(N=n)\, E[\sum_{j=1}^{n} C_j \,|N=n\,] \\ &= \sum_{n=1}^{+\infty} P(N=n)\, E\left(\sum_{j=1}^{n} C_j\right) \\ &= [\sum_{n=1}^{+\infty} nP(N=n)]\, E[C] = E[N]\, E[C] \end{aligned}$$

D'après l'hypothèse d'indépendance des coûts et de la fréquence des sinistres

Propriété 2 :

Si C et N admettent des moments d'ordre 2, il en résulte aussi :

$$\boldsymbol{Var[S] = E[N] \times Var[C] + Var[N] \times E[C]^2}$$

Démonstration :

(i) $E[S] = E[E(S|N)]$ avec $E(S|N=n) = \begin{cases} 0 & si\ n = 0 \\ n \times E(C) & si\ n \geq 1 \end{cases}$

D'où $E(S|N) = N \times E[S]$ et donc $E[S] = E[C] \times E[N]$

(ii) De plus, la formule de décomposition de la variance donne :

$$V(S) = E[V(S|N)] + V[E(S|N)]$$

Et $E(S|N) = N \times E[S]$ conduit à $V[E(S|N)] = V[N] \times E[C]^2$

(iii) Enfin, $V(S|N=n) = \begin{cases} 0 & si\ n = 0 \\ n\ \times V\,[C] & si\ n \geq 1 \end{cases}$

soit $V(S|N) = N \times V(C)$ et *et* $E[V(S|N)] = E[N] \times V[C]$

Annexe 2 - Test d'indépendance du Khi deux[1]

Le but du test

Ce test s'applique lorsque l'on souhaite démontrer l'indépendance ou la dépendance de deux critères de nature qualitative dans une expérience.

On considère alors un échantillon de polices pouvant être classées selon un certain nombre de colonnes (critère 1) et de lignes (critère 2).

Les critères 1 et 2 sont des variables dont la valeur est disponible pour l'ensemble des polices étudiées. On suppose que le critère 1 (respectivement critère 2) admet c (respectivement l) modalités.

Les hypothèses du test:

- Hypothèse nulle (H_0): les deux variables sont indépendantes
- Hypothèse alternative (H_1) : les deux variables sont corrélées

Méthode de calcul de la statistique de test et exemple d'application

Nous appuyons l'explication de la méthode de calcul de la statistique du test du Khi-deux sur un exemple.

On suppose que l'on dispose de n = 250 individus. On s'intéresse au lien existant entre leur genre (Femme/Homme) et le port de lunettes (Lunettes/Sans lunettes).

➢ Mise en place du tableau de contingence

Grâce aux données, il est possible de réaliser le tableau de contingence suivant :

	Hommes	Femmes	Total
Lunettes	30	40	70
Sans lunettes	60	120	180
Total	90	160	250

➢ Calculer les valeurs théoriques

On calcul les valeurs théoriques en se servant des valeurs expérimentales.

	Hommes	Femmes	Total
Lunettes	$\mathbf{25,2} = 90 \times \frac{70}{250}$	$\mathbf{44,8} = 160 \times \frac{70}{250}$	70
Sans lunettes	$\mathbf{64,8} = 90 \times \frac{180}{250}$	$\mathbf{115,2} = 160 \times \frac{180}{250}$	180
Total	90	160	250

[1] Source : [S9]

➢ Calculer la valeur du Chi-Deux puis interpréter

On cherche à présent à déterminer la valeur du Chi-deux pour chaque échantillon de l'expérience. Pour cela, on applique la formule suivante :

$$\chi^2 = \frac{(\text{fréquence}_{\text{observée}} - \text{fréquence}_{\text{théorique}})^2}{\text{fréquence}_{\text{théorique}}}$$

D'où les résultats suivants:

	Hommes	Femmes	Total
Lunettes	0,91	0,51	1,43
Sans lunettes	0,36	0,20	0,56
Total	1,27	0,71	**1,98**

Nous comparons alors la valeur du Chi-deux observée (1,98) à sa valeur théorique, disponible dans la table du Khi-deux, afin de prendre une décision.

Fonctionnement de la table

La valeur théorique du Khi-deux peut être lue sur la table associée à cette loi.
Cette table est à double entrée :

- **En ligne :** le degré de liberté, donné par $(c-1)\times(l-1)$, où c et l représentent respectivement le nombre de colonnes et lignes du tableau de contingence.
- **En colonne :** l'intervalle de confiance du test (ex : 0,05 pour un test à 95 %).

Loi de Khi-deux

Le tableau donne x tel que P(K > x) = p

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245
9	1,1519	1,7349	2,0879	2,5324	3,3251	4,1682	5,3801	12,2421	14,6837	16,9190	19,6790	21,6660	23,5894	27,8772
10	1,4787	2,1559	2,5582	3,0591	3,9403	4,8652	6,1791	13,4420	15,9872	18,3070	21,1608	23,2093	25,1882	29,5883

Table 1- Table du Khi-2[1]

Interprétation

La prise de décision quant à l'hypothèse à retenir peut se faire à l'aide de la comparaison entre les deux valeurs du Khi-2.

Ainsi, si $\chi^2{}_{observé} < \chi^2{}_{théorique}$, on accepte H_0. Le cas échéant, on rejette H_0 au profit de H_1.

Dans notre exemple, pour un test fiable à 95%, notre $\chi^2{}_{théorique} = 3{,}84 > \chi^2{}_{observé} = 1{,}98$: il existe donc une corrélation entre le port de lunettes et le genre de l'assuré.

En pratique, les logiciels de traitement statistique réalisent toutes les opérations précédentes automatiquement (à notre place) et fournissent directement une probabilité appelée **p-value**.

[1] Source: [I1]

La p-value est la probabilité, sous H_0, d'obtenir une statistique aussi extrême « grande » que la valeur observée sur l'échantillon : c'est le plus petit seuil de significativité pour lequel l'hypothèse nulle est acceptée.

Ainsi, pour un seuil de significativité α donné, on compare la p-value et le seuil α, afin d'accepter, ou de rejeter H_0 :

- Si p-value < α, on va rejeter l'hypothèse H_0 (en faveur de H_1)
- Si p-value > α, on va rejeter l'hypothèse H_1(en faveur de H_0)

Conditions de validité du test du Chi-2

Le test du Chi-2 est relativement simple à mettre en œuvre mais ne peut cependant, cependant être utilisé valablement que si certaines conditions sont remplies.

Les trois conditions principales sont les suivantes :

1) L'effectif total du tableau de contingence (correspondant au nombre d'individus observés) doit être supérieur ou égal à 20
2) L'effectif marginal du tableau de contingence (correspondant à la somme des lignes et la somme des colonnes) doit toujours être supérieur ou égal à 5.
3) L'effectif théorique des cases du tableau de contingence doit être supérieur à 5 dans 80 % des cases du tableau de contingence.

Ces conditions sont évidemment assez contraignantes et elles sont souvent violées lorsque l'on traite des populations de petite taille.

Remarque :

L**e test du Chi-2 est relativement robuste**, ce qui signifie que ses conclusions demeurent en général valides, même lorsque les hypothèses de base ne sont pas tout à fait respectées.

Le coefficient de Cramer

Le test du Khi-deux nous permet de déterminer l'existence ou non d'un lien entre deux variables mais il ne permet pas de quantifier ce lien : les variables sont-elles très liées ou légèrement seulement?

Afin d''évaluer le degré de relation entre les deux variables qualitatives, divers indices ont étés proposés.
Nous avons retenu l'indice de Cramer (V de Cramer) qui varie entre 0 et 1.
Si le coefficient est proche de 0, les variables ne sont pas liées. Si le coefficient est proche de 1, les variables sont liées.

Le V de Cramer est donné par :

$$V = \sqrt{\frac{\chi^2}{n \times \min(c-1, l-1)}}$$

Annexe 3 - Test de Student pour groupes indépendants

Le test de Student pour groupes indépendants est le test statistique paradigmatique lorsque la **variable réponse** est **quantitative** et **la variable explicative** est **qualitative à deux modalités**.
On appelle échantillon 1 (respectivement 2), l'ensemble des individus ayant pour valeur de la variable explicative la première (respectivement $2^{ème}$) modalité.

Hypothèses du test :

- Hypothèse nulle (H_0) : la variable explicative n'a aucune influence sur la variable à expliquer
- Hypothèse alternative (H_1) : la variable explicative influe sur la variable à expliquer.

La statistique de test de Student

La statistique du test de Student (**t**) est donnée par :

$$\boxed{t = \frac{\overline{X_1} - \overline{X_2}}{S_{com} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

Avec

- S_{com} l'écart type commun aux deux groupes donné par :

$$S_{com} = \sqrt{\frac{(n_1 - 1)S_{cor.1}^2 + (n_2 - 1)S_{cor.2}^2}{ddl}}$$

Où

- $S_{cor.1}^2$ et $S_{cor.2}^2$ sont respectivement les variances des échantillons 1 et 2
- n_1 et n_2 sont respectivement les tailles des échantillons 1 et 2
- $ddl = n_1 + n_2 - 2$, le degré de liberté

La statistique de test consiste à établir le rapport entre d'une part la différence des moyennes d'échantillons (numérateur) et d'autre part la variabilité de l'ensemble des données (dénominateur).

Remarque :

Lorsque le nombre d'observations dans les deux groupes est grand (c'est-à-dire ≥ 30), le dénominateur de la statistique de test se simplifie car le Théorème Central Limite s'applique.

Concrètement cela revient alors à calculer une valeur ***z*** au lieu d'une valeur de ***t*** avec :

$$z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_{cor.1}^2}{n_1} + \frac{S_{cor.2}^2}{n_2}}}$$

Interprétation de la statistique de Student

Cette statistique de test est appelée statistique de Student car elle suit la loi $S(n-1)$, où n correspond au nombre d'individus total.
La prise de décision quant à l'hypothèse à retenir peut se faire à l'aide de la comparaison entre la valeur critique de Student (lue dans la table de Student qui fonctionne comme la table de Fisher) et la statistique de test.

Ainsi, si $|t| < Student_{critique}$, on accepte H_0. Le cas échéant, on rejette H_0 au profit de H_1.

En pratique, les logiciels de traitement statistique renvoient également une probabilité appelée **p-value** (probabilité, sous H_0, d'obtenir une statistique aussi extrême « grande » que la valeur observée sur l'échantillon) qui permet de conclure.

Ainsi, pour un seuil de significativité α donné, on compare la p-value et le seuil α, afin d'accepter, ou de rejeter H_0 :

- Si p-value < α, on va rejeter l'hypothèse H_0 (en faveur de H_1)
- Si p-value > α, on va rejeter l'hypothèse H_1(en faveur de H_0)

Conditions d'application du test

Bien qu'il soit très robuste (comme tous les tests paramétriques), le test de Student doit néanmoins être utilisé conditionnellement à deux stipulations importantes.

Premièrement, la distribution des données doit être normale ou quasi normale.
Cette normalité doit être systématiquement testée pour $n_{1\ ou\ 2} < 30$ (Nombre d'observations du groupe 1 ou 2) mais ce test est facultatif lorsque $n_{1\ et\ 2} > 30$ car le TCL s'applique.

Pour tester la normalité des données, on utilise le **test de Shapiro-Wilk** (détails disponibles en [P5]).

Deuxièmement, il faut que les échantillons en présence aient une variance qui soit du même ordre de grandeur (et ce, quelle que soit la valeur prise par $n_1\ et\ n_2$).

Pour tester l'égalité des variances, on utilise le **test de Fisher**[1].

En pratique : Test de Student sous R

C'est avec la fonction ***t.test()*** que l'on réalise le test de Student avec R.
Quatre arguments doivent alors être spécifiés :

- Le premier est le modèle des données qui prend la forme :

  variable à expliquer ~ variable explicative

- Le second argument s'appelle alternative : il sert à préciser les modalités du test. En effet, l'hypothèse H_1 selon laquelle la variable explicative influe sur la variable à expliquer peut être affinée. Les possibilités sont les suivantes :

[1] Source : [S10]

- Alternative = « two sided » signifie que l'on veut réaliser un test bilatéral. Aucune indication n'est donnée sur le sens de l'influence.

- Alternative = « greater » signifie que le test est unilatéral et que l'on s'attend de plus à ce que la moyenne de l'échantillon 1 soit plus élevée que celle de l'échantillon 2.

- Alternative = « less » signifie signifie que le test est unilatéral et que l'on s'attend à ce que la moyenne de l'échantillon 1 soit plus faible que celle de l'échantillon 2.

- Le troisième argument s'appelle « paired ». Il s'agit d'un argument logique qui permet de spécifier le type du test :

  - Paired = TRUE, signifie que l'on effectue le test sur des conditions appariées, autrement dit, qu'il est possible que le même individu appartiennent aux deux échantillons

  - Paired = FALSE , signifie que l'on effectue le test sur des groupes indépendant : les deux modalités de la variable explicative sont incompatibles

- Le dernier argument s'appelle « var.equal ». il s'agit d'un argument logique qui sert à préciser s'il y a ou non homogénéité des variances entre les deux échantillons :

  - var.equal = TRUE, signifie que les variances sont homogènes.
  - var.equal = FALSE, indique que les variances sont hétérogènes

Annexe 4 – Test de Kruskal-Wallis

Ce test a été développé en 1952 par W.H. Kruskal et W.A Wallis. Il s'agit d'un test non paramétrique qui ne s'applique pas aux données elles-mêmes mais aux rangs[1] obtenus à partir des données.

Il est utilisé lorsqu'on dispose d'échantillons (de groupes) de mesures et que l'on cherche à savoir si au moins l'un d'entre eux diffère des autres échantillons. De manière équivalente cela revient à se demander si tous les échantillons sont oui ou non issus de la même population.

Hypothèses statistiques

Le test se fonde sur l'examen des moyennes des rangs obtenus à partir des données. Dans le cas où l'on étudie k échantillon, on teste donc les hypothèses suivantes :

- **Hypothèse initiale** $(\mathbf{H_0})$: $\overline{r_1} = \overline{r_2} = \cdots = \overline{r_k}$
  - Il n'y pas de différence entre les moyennes des rangs 1 à k. Toutes les observations sont donc tirées de la même population.

- **Hypothèse alternative** $\mathbf{H_1}$: $\exists \overline{r_i}, \overline{r_j}\ tel\ que\ \overline{r_i} \neq \overline{r_j}$
  - Parmi les $\overline{r_k}$ moyennes, il y en a au moins une qui diffère des autres. Les observations d'au moins un groupe sont donc tirées d'une population différente.

La statistique de test

La statistique du test de Kruskal-Wallis (H) est donnée par :

$$\boldsymbol{H} = \frac{12}{N(N+1)} \times \sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - 3(N+1)$$

Avec

- $N = \sum_{i=1}^{k} n_i$, avec n_i le nombre d'observations du groupe i
- k, le nombre de groupes indépendants étudiés
- R_i, la somme des rangs dans le groupe i obtenue par $R_i = \sum_{j}^{n_i} R(X_{ij})$, avec X_{ij} la $j^{ème}$ observation du groupe i avec $i = 1, \ldots, k$.

Lorsqu'il y a des ex-æquo dans les données, on définit un rang moyen pour les observations identiques. On calcule alors une statistique H corrigée, notée $\widetilde{H}$:

$$\widetilde{H} = \frac{H}{1 - \frac{\sum_{g=1}^{G}\left(t_g^3 - t_g\right)}{N^3 - N}}$$

Avec,

- G=nombre total d'observations distinctes dans le jeu de données
- t_g =nombre d'observations pour une valeur (si une valeur apparaît une fois, alors $t_g = 1$; si une valeur se répète deux fois $t_g = 2$ …etc).

[1] Classement des données par ordre croissant attribution d'un numéro

Remarque :
La correction est en général négligeable quand le nombre d'ex-æquo est faible, elle prendra de plus en plus d'importance lorsque les ex-æquo augmentent.

Interprétation de la statistique de Kruskal-Wallis

Si $n_k > 5, \forall k$, la statistique H de Kruskal-Wallis peut être approximée sous H_0 par une lois du χ^2 à (k-1) degré de liberté.

La région critique du test (RC) au risque α s'écrit :

$$RC: H > \chi^2_{1-\alpha}(\mathrm{k}-1)$$

Pour prendre une décision quant à la validité de l'une ou l'autre des hypothèses, on doit comparer la valeur de la statistique H à la valeur théorique du quantile d'ordre α d'une variable suivant une loi du Khi-deux à (k-1) degrés de liberté.

Ainsi, si :

- $H < \chi^2_{1-\alpha}(\mathrm{k}-1)$: on accepte H_0
- $H > \chi^2_{1-\alpha}(\mathrm{k}-1)$: on rejette H_0

De manière équivalente, en terme de p-value, si :

- $\mathrm{p}-\mathrm{value} > \alpha$: on accepte H_0
- $\mathrm{p}-\mathrm{value} < \alpha$: on rejette H_0

Annexe 5 - Les arbres de classification

Dans le cadre d'un problème de classification, on dispose d'un échantillon d'apprentissage de n observations d'une variable Y (variable à prédire) qui prend les valeurs 1,2,..,k et de p variables de prédiction.

Le but étant de trouver un modèle de prédiction des valeurs de Y à partir de nouvelles valeurs de X.
Cet échantillon va être divisé selon une règle de fractionnement et la qualité du critère de fractionnement.

Dans le cas où les variables de prédiction sont qualitatives, les règles de fractionnement sont des questions du type « est-ce que $x^j \in \{d_1,..d_r\}$? » où $(x^j)_{1\leq j\leq p}$ est la $j^{ème}$variable de prédiction et $(d_s)_{1\leq s\leq r}$une des modalités de cette variable.

Plus le nombre de variables de prédiction est important, plus le nombre de **règles de fractionnement** possibles est grand.

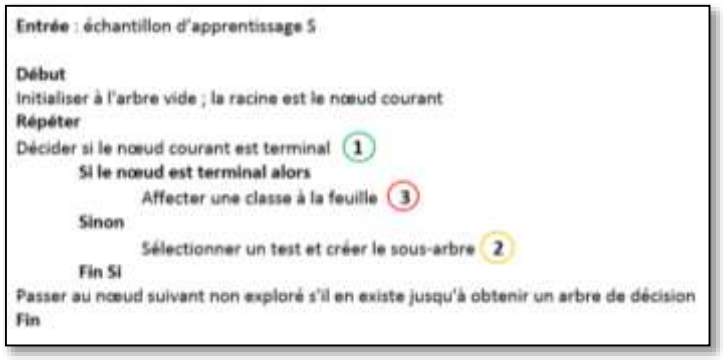
La « **qualité du critère de fractionnement** » compare différentes scission et détermine laquelle de ces dernières va produire les sous-échantillons les plus homogènes.

Une mesure de la qualité du critère de fractionnement est l'indice de Gini (g) donné par :

$$g(t) = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ i\neq j}}^{k} p(j|t)p(i|t)$$

Ou $p(j|t)$ est la probabilité de la modalité j d'appartenir au nœud t.
C'est l'une des plus utilisée pour mesurer la pureté des nœuds dans un contexte de classification.

Algorithme de création de l'arbre



1 Décider si le nœud courant est terminal
Un nœud *t* est terminal si $g(t) \leq i_0$ où i_0 est un paramètre à fixer.

2 Sélectionner un test à associer à un nœud
Soit *t* une position et soit *test* un test.
Si ce test devient l'étiquette du nœud à la position *t*, alors on appelle P_{gauche} (respectivement P_{droite}) la proportion d'éléments de l'ensemble des exemples associés à *t* qui vont sur le nœud en position t_1 (respectivement t_2).

La réduction d'impureté définie par le test *test* est identique au gain et définie par :

$$Gain(t, test) = g(t) - (P_{gauche} \times g(t_1) + P_{droite} \times g(t_2))$$

En position *t*, on choisit le test qui maximise la quantité *Gain*(*t*,*test*).

3 Affecter une classe à une feuille

On attribue la classe majoritaire.

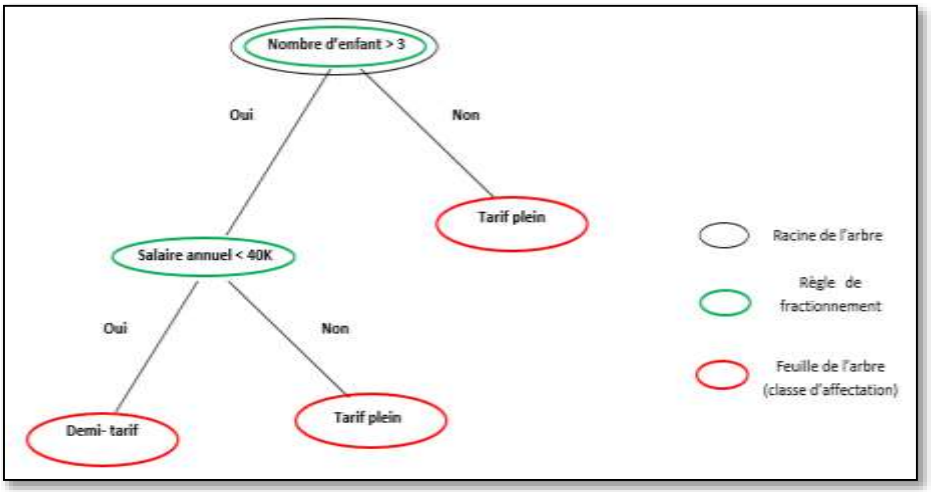


Figure 11- Arbre illustratif des termes employés

Elagage de l'arbre

Une fois l'algorithme de construction de l'arbre arrêté, il est toujours utile d'évaluer encore la qualité de la prédiction de l'arbre **à partir d'un échantillon d'observations qui ne participe pas à sa construction.**

Ces évaluations sont utilisées pour **élaguer** l'arbre, c'est-à-dire, sélectionner un arbre plus simple que celui obtenu à l'arrêt de la construction de l'arbre, mais qui est tout aussi précis pour classer les nouvelles observations.

C'est ce que l'on appelle la **validation croisée** (*Crossvalidation*).

Principe de la validation croisée

Tout d'abord, on construit l'arbre maximal avec la totalité de l'échantillon d'apprentissage.

On calcule ensuite les séquences d'arbres emboîtés en modulant le paramètre de complexité (CP), c'est-à-dire le nombre de nœuds de l'arbre. L'arbre élagué sera choisi parmi ces candidats.

Pour chaque valeur de CP est calculée :

- l'erreur en resubstitution
- l'erreur en validation croisée

Erreur en resubstitution :

C'est la proportion de cas qui sont mal classés par le classificateur construit à partir de l'ensemble de l'échantillon. On utilise ici le même échantillon pour construire et tester les données : c'est un estimateur optimiste.

Cette estimation est calculée de la manière suivante :

$$R = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ou x_i est une fonction indicatrice telle que :

$$x_i = \begin{cases} 1, & \text{si l'individu i est mal classé} \\ 0, & \text{sinon} \end{cases}$$

Cette erreur est calculée à titre purement indicatif : elle diminue quand la taille de l'arbre augmente.

L'erreur en validation croisée

Pour calculer l'erreur en validation croisée, on partitionne les données en K portions (généralement K=10).
Un arbre est alors établit sur les K-1 portions et la portion restante est utilisée pour estimer l'erreur. En faisant tourner les sous échantillons, on dispose donc de K arbres.

On utilise alors les valeurs des paramètres de complexité de la table précédente pour réduire chaque arbre et calculer l'erreur sur la fraction test qui leur est associée.

Au final, pour une valeur spécifiée de CP, on disposera de K arbres avec leurs taux d'erreur respectifs. Le taux d'erreur en validation croisée sera donné par la moyenne de ces taux.

Le calcul du taux d'erreur est cette fois ci pertinent puisqu'il fait intervenir des échantillons test n'ayant pas participé à la construction de l'arbre.

Au final, **l'arbre** sera composé du nombre de CP associé à **l'erreur en validation croisée la plus faible.**

Les avantages des méthodes de Classification et Régression par Arbre

De nombreuses autres méthodes existent pour résoudre des problèmes de classification ou de régression. C'est le cas notamment des Modèles Linéaires Généralisés qui ont été utilisées pour déterminer le coût et la fréquence moyenne de sinistre.

Les techniques de classification par arbre, quand elles « fonctionnent » et produisent des prévisions précises basées sur quelques conditions logiques (si-alors) présentent un certain nombre d'avantages par rapport à bon nombre de ces techniques alternatives.

La simplicité des résultats

Dans la plupart des cas, les résultats sont résumés dans un arbre de façon très simple.

On descend de la racine aux feuilles en passant par un chemin.
Le chemin est déterminé par la réponse aux règles de fractionnement situées à chaque nœud de l'arbre : oui, à gauche, non, à droite. Une classe d'affectation (le résultat) figure sur chaque bout de chemin possible, c'est-à-dire sur les feuilles.

Cette simplicité est utile non seulement à des fins de classification rapide de nouvelles observations (il est beaucoup plus facile d'évaluer seulement une ou deux conditions logiques, que de calculer les scores de classification pour chaque groupe possible, ou des valeurs prédites , sur la base de toutes les

variables) , mais également lorsque l'on doit expliquer à un tiers pourquoi les observations sont classées ou prédites d'une manière particulière.

Les méthodes d'arbres sont non paramétriques et non linéaire

Les résultats définitifs de l'utilisation de méthodes d'arbres de classification ou de la régression peuvent être résumés en une série de conditions logique « Si-Alors ».

Par conséquent, il n'y a aucune hypothèse implicite sur les relations sous-jacentes entre les variables prédictives et la variable dépendante : il n'y a ni relation linéaire ni fonction de lien comme dans les modèles linéaires généralisés.

Ainsi, les méthodes d'arbres sont particulièrement bien adaptées à des tâches d'exploration de données, où il y a souvent peu de connaissance a priori concernant les variables qui sont liées et la manière dont elles le sont.

Dans ces types de l'analyse des données, les méthodes d'arbres peuvent souvent révéler des relations simples entre quelques variables qui auraient pu être facilement passés inaperçus en utilisant d'autres techniques d'analyse.

Remarque :

Malgré sa simplicité et son élégance, l'approche de recherche exhaustive a une propriété indésirable. Une variable non ordonnée avec m valeurs distinctes a $(2^{m-1}-1)$ divisions de la forme $X \in S$.
Par conséquent, toute chose égale par ailleurs, les variables qui ont le plus de modalités ont le plus de chance d'être sélectionnées. Ce biais de sélection affecte l'intégrité des conclusions tirées à partir de la structure arborescente.

Annexe 6 – Acronymes

Variable	Référence	Signification
VV	Vehicule Value	Classe de valeurs du véhicule
VB_c	Vehicule Body cost	Classe de types de véhicules formée par coûts moyens semblables
VB_f	Vehicule Body frequency	Classe de types de véhicules formée par fréquences moyennes semblables

Tableau 68 – Acronymes des variables tarifaires transformées

Modèle	Signification
B	Modèle de base (sans contrainte)
SG	Modèle Sans Genre (suppression du genre de la tarification)
P	Modèle de Pondération par la proportion Homme/Femme
PR_L	Modèle Prédictif fondé sur une régression logistique binaire
PR_CART	Modèle prédictif fondé sur un arbre de classification

Tableau 69 – Acronymes des modèles tarifaires créés