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 _____

Par : Léa MARCIANO

Titre: Modélisation de la dérive des soins de santé à court terme

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 : AG2R La Mondiale

Signature :

Directeur de mémoire en entreprise :

Nom : KUSNIK Odile

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 Cedex 16

Résumé

Le marché de l'assurance complémentaire santé intervient de plus en plus dans le système de santé français. L'anticipation de la dérive sur l'exercice à venir joue un rôle majeur pour l'indexation tarifaire dans le pilotage technique du risque santé. Il apparait donc, dans l'intérêt des assurances complémentaires, de quantifier cette dérive.

Ce mémoire traite de l'assurance complémentaire santé, en particulier d'un portefeuille sur le segment individuel. L'objectif de ce mémoire est l'estimation à court terme de la dérive des soins de santé, qui est l'évolution des dépenses des soins. Pour ce faire, ce mémoire s'articule en quatre parties.
Après avoir présenté le système de santé français, les acteurs de ce marché et en particulier AG2R La Mondiale, une première approche macroéconomique de l'estimation de la dérive à court terme est étudiée. Cette dernière nous oriente vers la construction d'une base de données nécessaire à une approche de modélisation individuelle prenant en compte les caractéristiques propres à chaque individu. Dans un esprit de recherche, l'étude sera consacrée à un poste de soins. Deux types de modélisation sont mises en œuvre par l'utilisation de modèle linéaire généralisé. Une première modélisation consiste en la modélisation de la consommation des soins de santé sur deux années successives pour ainsi obtenir la dérive. Suite aux résultats obtenus par cette approche, une deuxième approche est mise en œuvre en modélisant l'évolution des dépenses de santé sur des profils d'individus.

Mots clés : complémentaire santé, portefeuille individuel, dérive des soins de santé, modèles linéaires généralisés.

Abstract

The complementary health insurance market intervenes more and more in the French health care system. The anticipation of the drift on the coming year plays a major role for the tariff indexation in the technical management of the health risk. It is therefore, in the best interest of the complementary health insurances, to quantify this drift.

This thesis deals with the complementary health insurance, in particular with the portfolio of the individual segment. The purpose of this dissertation which is divided into four parts is the short- term estimation of the health care drift, which is the evolution of the health care expenditures. After introducing the French health care system, the actors of this market and in particular AG2R LA Mondiale, a first macroeconomic approach to the estimation of short term drift will be presented. This macroeconomic study will take us towards the construction of a necessary database for an individual modeling approach taking into account the characteristics specific to each individual. In a spirit of research, the study will be devoted to a health post. Two types of modeling are implemented by the use of generalized linear model. A first modeling consists of the modeling of the consumption of health care during two successive years to then obtain the drift. Following the results obtained by this approach, a second perspective is implemented by modeling the evolution of health expenditures on individual profiles.

Keywords: complementary health, individual portfolio, health care drifts, generalized linear models.

Note de synthèse

Contexte :

Depuis quelques années, le poids important des complémentaires santé ne cesse de croître en ce qui concerne les remboursements des frais de santé des français. En effet, le déremboursement de la Sécurité Sociale entraîne un transfert des coûts vers l'assurance maladie complémentaire. De plus, la consommation médicale est en constante évolution. Celle-ci est due, en partie, à la création des dispositifs favorisant l'accès aux soins, le progrès technique et l'inflation. Pour maîtriser cette augmentation, de nombreuses mesures prennent forme.

Les organismes complémentaires doivent répercuter cette évolution dans leurs tarifs. Par conséquent, au moins une fois par an, l'assureur procède à l'indexation tarifaire (indexation des cotisations des adhérents) afin de faire face à la dérive naturelle des frais de santé et aux modifications des législations. La dérive naturelle des prestations de santé est liée au vieillissement de la population, aux développements des maladies chroniques et aux coûts de soins plus importants. Par ailleurs, de nombreuses réformes impactent les frais de santé, ce qui nécessite donc de mesurer ces changements sur les portefeuilles de santé.

Problématique :

Il est crucial pour l'assureur de prévoir l'évolution de la sinistralité de son portefeuille pour une survenance future. L'indicateur sur lequel il se base est la dérive de la consommation de soins.

Ainsi, ce mémoire vise à estimer à court terme la dérive des soins de santé sur un portefeuille individuel et d'observer l'apport d'une approche individuelle dans l'estimation de cette fameuse dérive.

Démarche adoptée :

La première partie de ce mémoire apporte une vision globale sur le système de santé français. Dans cette partie sont présentés les principaux intervenants de ce marché ainsi que les différents mécanismes de remboursements des frais de santé proposés par les acteurs propres à ce marché. D'autre part, sont exposées les évolutions réglementaires récentes (comme par exemple le contrat responsable, la convention médicale) impactant le marché de la santé et en particulier les complémentaires santé.

La deuxième partie de ce mémoire est consacrée à une étude macroéconomique pour une estimation de la dérive à court terme, c'est-à-dire pour les années 2017 et 2018. La première étape de cette étude a été de construire une base de données sur un périmètre fermé regroupant les bénéficiaires présents du 1er janvier 2014 au 31 décembre 2016 du portefeuille individuel d'AG2R La Mondiale. A ces bénéficiaires sont associées les prestations versées par AG2R La Mondiale, c'est-à-dire le montant engagé par l'assureur. Le calcul de l'âge moyen de chaque gamme du portefeuille permet de scinder celui-ci en deux catégories : d'une part les gammes d'actifs, et d'autre part, les gammes de seniors.

Dans cette approche macroéconomique, la méthodologie utilisée a été de calculer une dérive pour deux types de bénéficiaires, les actifs et les seniors, et de calculer une tendance de dérive sur les années précédentes. En effet, en raison d'importantes modifications apportées aux garanties en 2016 dans le cadre de la mise en conformité au contrat responsable, il n'est pas possible de dégager une tendance 2016 pure. La construction d'une tendance sur des années antérieures est donc nécessaire. Ainsi, pour quantifier la dérive des années futures sont ajoutées l'estimation des mesures des impacts réglementaires (comme la convention médicale) à la tendance estimée.

Cette approche permet d'appréhender une dérive globale sur le portefeuille, mais ne met pas en relief les différences de dérive par rapport aux caractéristiques propres des individus, comme par exemple le fait

que les nouveaux adhérents ont une dérive plus importante. C'est pourquoi, l'apport d'une modélisation de la dérive par une approche individuelle peut être nécessaire pour capter des effets individuels.

La troisième partie est donc consacrée à la création d'un portefeuille servant par la suite à une étude statistique. Ce portefeuille regroupe les consommations de soins des années 2013 et 2014 de deux gammes du portefeuille individuel d'AG2R La Mondiale. Dans cette partie sont présentées les différentes étapes d'extraction des données pour créer une base de données finale fiable. En effet, la qualité des données est primordiale afin de ne pas fausser les résultats des différentes études statistiques.
Sur le périmètre ainsi constitué, une étude descriptive du portefeuille considéré permet une connaissance des différentes variables étudiées. Après avoir décrit le périmètre étudié, l'étape suivante est la modélisation.

Enfin, nous nous sommes tournés vers la modélisation de la dérive par une approche statistique.
Le processus de modélisation statistique nous a conduits à revoir les aspects théoriques du modèle linéaire généralisé. Nous nous sommes ici concentrés sur une garantie en particulier, la garantie « Consultations et visites » qui représente un poids important sur le portefeuille.
Une première approche de la modélisation de la dérive par le biais de la modélisation de la consommation n'a pas été satisfaisante. Cette difficulté provient en partie de la prise en compte des non consommants dans la modélisation. De ce fait, il a donc été nécessaire de limiter ce nombre de montants nuls en travaillant sur des profils d'individus. Ainsi, a été calculé un coût moyen pour chaque profil d'individu.
Une seconde approche a ensuite été mise en place, en modélisant cette fois le taux d'évolution des dépenses de santé sur des profils d'individus par un modèle linéaire. Afin de procéder à la validation de ce modèle, les consommations de 2015 ont été utilisées. En pratique, les résultats de cette modélisation ne sont pas tout à fait satisfaisants. En effet, le but de cette modélisation était de trouver une mesure précise future de la dérive. Pour autant, cette modélisation a permis de donner l'ordre de grandeur de la dérive sur cette garantie. De plus, ce modèle permet de quantifier les variables influentes sur la dérive. Les résultats ont montré des différences de comportements en terme de dérive entre les individus, comme par exemple une plus grande évolution des dépenses chez les femmes.
Quant à une estimation précise de la dérive pour les années futures, celle-ci est difficile à obtenir car des effets inobservables ou non quantifiables impactent la dérive des soins de santé. L'estimation de la dérive est donc difficilement modélisable.

Les résultats d'une approche individuelle sur un portefeuille individuel ne fournissent pas en pratique des résultats assez riches permettant d'obtenir une mesure précise de la dérive. C'est pourquoi, une des voies futures de ce mémoire serait de prendre en compte un portefeuille collectif, puisque la structure de celui-ci est différente et possède d'autres variables. Il serait aussi intéressant de s'orienter vers une approche d'apprentissage statistique.

Executive summary

Context:

For years, the importance of complementary health in the French health care system has been increasing with regard to the reimbursement of health costs. Indeed the disbursement applied by the Social Security leads to a transfer of costs to the complementary health insurance. In addition, the medical consumption of the French is constantly evolving. This is due, in part to the creation of protocols promoting access to health care, to technical progress and inflation. To control this increase, several measures are taking shape.

Complementary organizations must reflect this change in their rates. Therefore, at least once a year, insurers proceed with tariff indexation (indexation of membership fees) in order to cope with the natural drift of health costs and changes in legislation. The natural drift in health benefits is linked to an aging population, to an increase of chronic diseases and to higher costs of health care. In addition, many reforms impact a health cost, which makes it necessary to measure these changes in the health portfolios.

Issue:

It is crucial for the insurer to predict the evolution of the loss ratio of its portfolio for a future occurrence. The indicator on which it is based is the drift of consumption of care.

Thus, this thesis aims to estimate in the short term the drift of health care on an individual portfolio and to observe the contribution of an individual approach in the estimation of this famous drift.

Approach taken:

The first part of this thesis provides a global vision of the French health care system. This section presents the main players in this market as well as the different reimbursement mechanisms for health care costs proposed by this specific market's players. On the other hand, are exposed the recent regulatory evolutions (like for example the responsible contract, the medical convention) impacting the market of the health and in particular the complementary health care system.

The second part of this thesis is devoted to a macroeconomic study for an estimate of the short-term drift, that is to say for the years 2017 and 2018. The first step in this study was to build a database on a closed perimeter bringing together the beneficiaries present from January 1, 2014 to December 31, 2016 of AG2R La Mondiale's individual portfolio. To these beneficiaries are associated benefits paid by AG2R La Mondiale, that is to say the amount incurred by the insurer. The calculation of the average age of each range of the portfolio divides the portfolio into two categories: on the one hand the asset ranges, and on the other hand, the senior ranges.

In this macroeconomic approach, the methodology used was to calculate a drift for two types of beneficiaries, assets and seniors, and calculate a drift trend over previous years. Indeed, due to significant changes to the guarantees in 2016 as part of the compliance with the responsible contract, it is not possible to identify a pure 2016 trend. The construction of a trend from previous years is therefore necessary. Thus, to quantify the drift of future years are added the estimated measures of regulatory impacts (such as medical convention) to the estimated trend.

This approach allows to apprehend a global drift on the portfolio, but does not highlight the differences in drift with respect to the characteristics of individuals, such as the fact that new members have a greater drift. Therefore, the contribution of drift modeling by an individual approach may be necessary to capture individual effects.

The third part is devoted to the creation of a portfolio which is then used for a statistical study. This portfolio includes the consumption of care of the years 2013 and 2014 of two ranges of the individual

portfolio of AG2R La Mondiale. In this part are presented the different steps of data extraction to create a reliable final database. Indeed, the quality of the data is essential so as not to distort the results of the various statistical studies.
On the scope thus constituted, a descriptive study of the portfolio considered allows an understanding of the various variables studied. After describing the studied perimeter, the following step is modeling.

Finally, we turned to modeling drift using a statistical approach.
The statistical modeling process led us to review the theoretical aspects of the generalized linear model. We focused on one guarantee in particular, the "Consultations and Visits" guarantee, which represents a significant weight on the portfolio.
A first approach to drift modeling through the modeling of consumption has not been satisfactory. This difficulty comes in part from the fact that non-consumers are taken into account in modeling. As a result, it was necessary to limit this number of zero amounts by working on individual profiles. Thus, an average cost is calculated for each individual profile.
A second approach has been put in place, modeling the rate of evolution of health expenditure on individual profiles by a linear model. In order to validate this model, 2015 consumption was used. In practice, the results of this modeling are not entirely satisfactory. Indeed, the purpose of this modeling was to find a precise future measure of drift. However, this modeling allowed to give the order of magnitude of the drift on this guarantee. Moreover, this model makes it possible to quantify the variables influencing the drift. The results showed differences in behavior in terms of drift between individuals, such as a greater change in spending among women.
As for an accurate estimate of drift for future years, it is difficult to obtain because unobservable or unquantifiable effects impact the drift of health care. The estimate of drift is therefore difficult to model.

The results of an individual approach on an individual portfolio do not provide in practice rich enough results to obtain an accurate measure of the drift. Therefore, one of the future paths of this thesis would be to take into account a collective portfolio, since the structure of this one is different and has other variables.
It would also be interesting to move towards a statistical learning approach.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à l'élaboration de ce mémoire, par leur soutien leur aide ou leur présence.
Pour commencer, je souhaite exprimer ma reconnaissance à ma maitre de stage, Odile KUSNIK, responsable du service actuariat santé d'AG2R La Mondiale pour m'avoir accueillie et conseillée tout au long de ces 6 mois. Je souhaite remercier de même les membres de l'équipe « Santé » pour leur bonne humeur et leurs conseils qui ont contribué au bon déroulement de mon stage, et plus particulièrement Omar ELJEBBARI pour son accueil et son suivi technique. Je remercie également Olivier SPILMAN pour ses conseils techniques et le temps qu'il m'a consacré.
Plus généralement, je remercie l'ensemble de la direction actuariat qui m'a accueillie chaleureusement tout au long de ce stage.

Je tiens en outre à remercier Antoine LY de l'université Paris Est pour ses conseils de modélisation.

Je remercie d'autre part mon tuteur académique Vincent RIVOIRARD pour ses conseils, son suivi et son soutien.

Enfin, je tiens à remercier toutes les personnes de mon entourage pour leur soutien constant et leurs encouragements et en particulier Déborah.

Table des matières

Introduction

L'assurance santé est un marché faisant intervenir plusieurs acteurs notamment l'Etat par l'intermédiaire de la Sécurité Sociale. D'ailleurs, chaque année est établie la loi de financement de la Sécurité Sociale qui dresse un bilan des dépenses et recettes de l'année pouvant conclure à de nouvelles réglementations. AG2R La Mondiale conçoit des produits en assurance santé sur les segments individuels et collectifs et dispose donc de gammes variées s'adressant à plusieurs types de bénéficiaires.

Le monde de l'assurance santé est en constante évolution. De nouvelles réglementations impactent chaque année les complémentaires santé qui doivent donc s'adapter aux changements constants. De plus, depuis quelques années, de nombreuses mesures sont mises en place afin de maîtriser l'accélération continue de la consommation médicale. En effet, les dépenses de santé augmentent naturellement du fait de l'augmentation des maladies chroniques, du progrès technique et de l'inflation.
Dans ce contexte, les assureurs sont particulièrement touchés par cette augmentation puisqu'ils doivent la répercuter dans leurs tarifs. Ils cherchent donc à mesurer cette dérive. En effet, en terme de suivi technique, l'assureur a besoin de connaître la tendance de l'évolution à la hausse ou à la baisse des dépenses de soins sur les différents postes et d'anticiper cette dérive.

Afin de répondre à ses objectifs de rentabilité, l'indicateur central sur lequel l'assureur s'appuie pour décider de faire évoluer les cotisations de ses adhérents est la « dérive » des soins de santé.
Une mesure précise de la dérive est donc nécessaire dans un contexte d'indexation tarifaire. Ce mémoire a donc pour objet l'étude de la dérive des soins de santé sur le segment individuel. La problématique est alors de définir une mesure de la dérive à court terme. On se pose donc plusieurs questions :

- Comment obtenir une mesure à court terme de la dérive ?
- Doit-on seulement considérer une approche macroéconomique ?
- Une étude individuelle améliore t-elle la mesure de la dérive ?
- Les résultats obtenus sont-ils concluants ? Quelles sont les limites ?

Nous proposons dans le cadre de ce mémoire de modéliser la dérive des soins de santé représentant le taux d'évolution des dépenses de soins entre deux années consécutives. Ce mémoire s'organise donc en quatre parties successives :

- La première partie permet de se familiariser avec le contexte de l'assurance santé. Ainsi, plusieurs points seront abordés :
  - Le contexte général de l'assurance santé français dans lequel sont développés les mécanismes de remboursements des différents acteurs de ce marché.
  - Les évolutions réglementaires récentes et multiples qui montrent que le marché de la santé est en pleine évolution.
  - La présentation de l'assurance complémentaire d'AG2R La Mondiale.
  - La définition du risque santé, la dérive.
- La deuxième partie consiste à présenter dans un contexte d'indexation tarifaire une étude macroéconomique sur la dérive des soins de santé sur le portefeuille individuel, pour ainsi obtenir une mesure de la dérive à court terme.
- La troisième partie consiste en la mise en place d'une base de données nécessaire à une étude de modélisation mettant en place différentes variables étudiées et complétée par une étude statistique descriptive permettant de se faire une idée plus précise de la structure du portefeuille étudié.

- La quatrième partie est consacrée à la mise en place de différentes méthodes de modélisations se basant sur des méthodes classiques dans le monde de l'assurance, tels que les modèles linéaires généralisés.
- Enfin, nous conclurons sur la limite des modèles utilisés et proposerons des voies d'amélioration possible.

Partie 1 : Présentation de l'assurance santé

Cette première partie a pour but de positionner le cadre du mémoire. La présentation du système de santé français introduit par le fonctionnement de la Sécurité Sociale et de la complémentaire santé en particulier celle d'AG2R La Mondiale.

Dans un premier temps, nous présenterons le système de santé français en expliquant le mécanisme de remboursement des soins de santé propre à chaque intervenant de ce marché. Dans un second temps, nous évoquerons les évolutions réglementaires récentes impactant le système de santé ces dernières années. Puis, nous nous attarderons sur l'assurance complémentaire AG2R La Mondiale en présentant son portefeuille.
Enfin, nous terminerons cette première partie en présentant la problématique de ce mémoire et en expliquant la notion de dérive.

I. Le système de santé

1. Le fonctionnement du système de santé français

La couverture santé est le fruit d'une combinaison entre deux systèmes. Un système mutualisé, obligatoire et étatique, assuré par la Sécurité Sociale, et un système assurantiel privé. Cette couverture complémentaire est assurée auprès d'un organisme complémentaire d'Assurance maladie prenant la forme d'une mutuelle, d'une institution de prévoyance ou d'une société d'assurance. Pour autant, il peut exister des dépenses restant à la charge de l'assuré pour les prestations non prises en charge par la Sécurité Sociale. Ce sont ces restes à charge qui sont pris en compte pour une partie ou dans la totalité par les organismes complémentaires.

a) La Sécurité Sociale

Créée en 1945, l'Assurance maladie est l'acteur majeur du système de soins, il s'agit d'une couverture de base de différents risques.

La Sécurité Sociale *« est la garantie donnée à chacun, qu'en toutes circonstances, il disposera des moyens nécessaires pour assurer sa subsistance et celle de sa famille dans des conditions décentes »*[1].

La Sécurité Sociale a pour but de protéger les individus de certains risques appelés « risques sociaux ». Le risque social est le fait d'être exposé à des évènements impactant la position sociale d'un individu, provoquant une baisse de revenus, et ou une augmentation des dépenses. Les principaux risques pris en charge par la Sécurité Sociale sont la maladie, la maternité-famille, la vieillesse etc. Le risque santé dont les principales composantes sont les prestations du risque maladie, invalidité, accidents du travail et maladies professionnelles représente un poids important dans les dépenses de la Sécurité Sociale.

Les régimes de base de la Sécurité Sociale sont marqués par des logiques de distinctions professionnelles dont les principales sont :

[1] Exposé des motifs de l'ordonnance du 4 octobre 1945 portant création de la Sécurité sociale.

- Le régime général : concerne la plupart des salariés, les étudiants, les bénéficiaires de certaines prestations et les simples résidents ;
- Le régime agricole : assurant la protection sociale des exploitants et des salariés agricoles ;
- Les régimes des travailleurs non-salariés et non agricoles : assurant la protection sociale des artisans, commerçants, industriels, les professions libérales ;
- Les régimes spéciaux de salariés et de fonctionnaires : assurant la protection sociale des agents de la SNCF, EDF, fonctionnaires….

Le régime général, régime principal, couvre environ 80% de la population et représente plus de la moitié des dépenses de santé.
Avec son caractère universel, la Sécurité Sociale couvre donc quasiment toute la population résidant en France. D'abord réservée aux salariés, elle s'est étendue notamment avec l'apparition de la Couverture Maladie Universelle (CMU) depuis le 1er janvier 2000.

De plus, le système de Sécurité Sociale est organisé en séparant les différents risques :

- Au niveau de la maladie : Caisse Nationale d'Assurances Maladies (CNAM) ;
- Au niveau de la famille : Caisse Nationale d'Allocations Familiales (CNAF), CAF ;
- Au niveau de la vieillesse : Caisse Nationale d'Assurance Vieillesse (CNAV).

En ce qui concerne le périmètre étudié, à savoir la branche maladie gérée par une caisse connue sous le nom de CNAM, elle représente la part la plus importante des dépenses par rapports aux autres caisses pour le régime général.

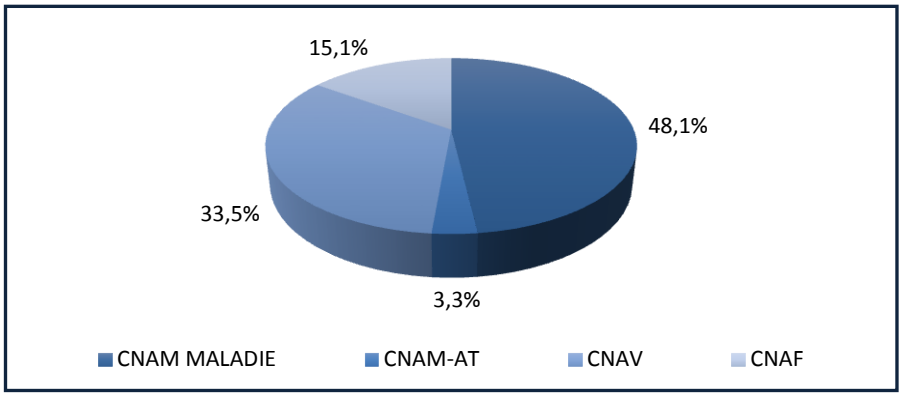

Figure 1 : Part de chaque branche dans le régime général en 2015
(Source : Comité des comptes de la Sécurité Sociale, juin 2016)

La couverture santé des français est à l'image d'une double couverture. En effet, l'Etat se préoccupe de l'assurance santé obligatoire couverte par la Sécurité Sociale et les organismes assurantiels privés prennent en compte la complémentaire santé afin de prendre en charge tout ou en partie du « reste à charge » de l'assuré. Cette double prise en charge garantit au citoyen français un fort taux de remboursement sur les différents postes de santé.

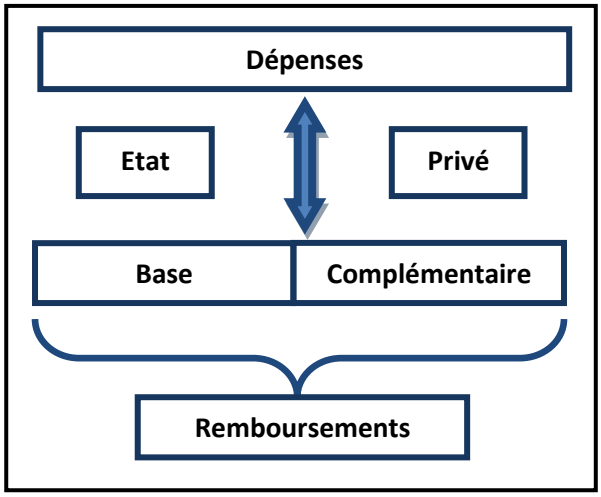

Figure 2 : Fonctionnement du système de santé français

b) Les organismes complémentaires

Agissant en complément de la Sécurité Sociale, ces organismes fournissent une protection supplémentaire aux risques assurés par la Sécurité Sociale. En effet, la couverture assurée par la Sécurité Sociale n'est pas intégrale. D'une part, dans la majorité des cas la Sécurité Sociale ne rembourse pas la totalité d'un acte et laisse une partie du coût de l'acte à la charge de l'individu : c'est ce qu'on appelle le ticket modérateur. Ce ticket modérateur est calculé sur un tarif de référence appelé « base de remboursement » (exemple 25€ pour une consultation de généraliste). D'autre part, certains praticiens proposent des tarifs supérieurs à la base de remboursement, et ces dépassements d'honoraires impliquent un reste à charge plus élevé pour le patient. Par ailleurs, la Sécurité Sociale rembourse peu les lunettes et les prothèses dentaires. C'est dans cette optique que les complémentaires interviennent, afin de permettre à la population de se protéger pleinement et contre un plus grand nombre de risques. Néanmoins, certaines réglementations imposent aux complémentaires des contraintes dans leur remboursement afin de limiter les abus de consommation en santé.

On distingue deux types de couverture : individuelle et collective (contrat d'entreprise). Un contrat collectif est souscrit par une entreprise ou une branche professionnelle au bénéfice d'une catégorie de personnes (les salariés d'une entreprise, un groupe de salariés relevant d'une branche professionnelle…). L'assurance complémentaire santé peut être également souscrite à titre individuel, c'est à dire directement par un particulier. Les contrats individuels étaient majoritaires avant l'ANI 2013 qui généralise la couverture d'entreprise depuis le 1er janvier 2016. En effet, depuis le 1er janvier 2016, les contrats collectifs d'entreprises sont à adhésion obligatoire.

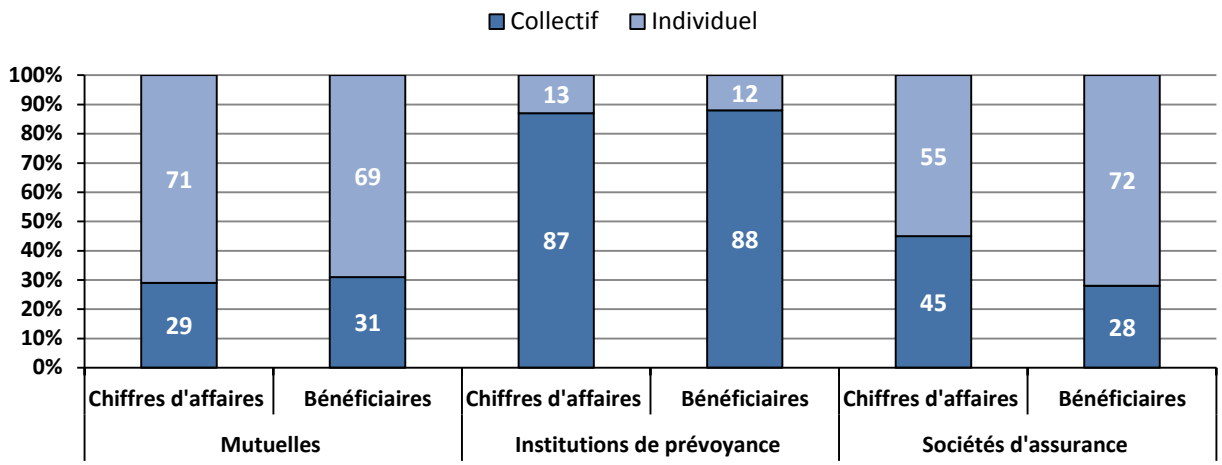

Figure 3 : Répartition des types de contrats selon les organismes complémentaires en 2014

La couverture complémentaire est assurée essentiellement par trois intervenants :

- *Les mutuelles :*

Ce sont des acteurs majeurs de la protection sociale française et la principale source de complémentaire en terme de nombre de personnes assurées.

Le mouvement mutualiste est né de la volonté des salariés de se prémunir contre les grands risques. La création en 1945 de la Sécurité Sociale a eu pour effet de concentrer l'action des mutuelles sur le risque maladie. Elles sont régies par le Code de la Mutualité et s'organisent autour du principe de solidarité et de mutualisation. Leurs fonds proviennent majoritairement des cotisations des assurés et leurs assemblées générales sont composées des assurés eux-mêmes. Le fonctionnement interne des mutuelles est basé sur un principe égalitaire, il n'est pas lié à l'apport de capital : chaque adhérent possède une voix dans les délibérations.

De plus, les mutuelles sont à but non lucratif, c'est-à-dire que tout excédent est réparti au sein de la mutuelle entre les membres sous la forme d'un gel de cotisation ou est tout simplement mis en réserve. Les cotisations versées aux mutuelles sont indépendantes du risque individuel de l'adhérent : il n'existe pas de sélection selon l'état de santé de l'adhérent (le questionnaire médical est interdit). Cependant, le risque est parfois partiellement maîtrisé par la catégorisation de la mutuelle (mutuelle d'enseignants, de cadres, d'étudiants...). La quasi totalité des mutuelles sont adhérentes à la Fédération Nationale de la Mutualité Française (FNMF).

Il existe trois types de mutuelles, distinguées selon le type de garanties ou selon leur champ de recrutement géographique ou/et professionnel :

- Les mutuelles professionnelles ;
- Les mutuelles de fonctionnaires ou assimilées au recrutement le plus souvent national ;
- Les mutuelles gérant un risque spécialisé (tels les risques sportifs ou scolaires).

Le portefeuille des mutuelles est composé majoritairement de contrats individuels qui représentent sept dixièmes de leurs bénéficiaires (**Figure 3**).

- *Les compagnies d'assurance :*

Elles sont régies par le code des assurances et se concentrent principalement sur la couverture individuelle mais intègrent de plus en plus le périmètre des contrats collectifs. Elles peuvent être des sociétés anonymes à but lucratif contrôlées par leurs actionnaires ou alors des sociétés d'assurances mutuelles contrôlées par leurs adhérents et devant constituer néanmoins des fonds propres suffisants pour respecter les contraintes de solvabilité.

- *Les institutions de prévoyance (IP) :*

Régies par le Code de la Sécurité Sociale, elles se concentrent principalement sur les contrats collectifs. Elles sont créées et gérées par les partenaires sociaux (employeurs, employés, entreprises, et adhérents) et sont donc des institutions de droit privé à but non lucratif. Elles disposent d'un Conseil d'administration paritaire, il est constitué à parts égales de représentants des salariés et de représentants des entreprises. C'est ce conseil qui doit définir et mettre en œuvre les garanties dans l'intérêt exclusif des salariés dans l'entreprise. Comme il n'y a aucun actionnariat à rémunérer, les résultats servent donc à améliorer le niveau des garanties, la qualité des services et la sécurité des engagements. Ces institutions peuvent être professionnelles, interprofessionnelles ou d'entreprise. Du fait de leur spécialisation en collectif, elles couvrent surtout des bénéficiaires d'âge actif ou jeunes. Les contrats individuels commencent à se développer depuis la loi Évin du 31 décembre 1989, qui a permis à tous les acteurs de proposer à la fois des contrats individuels et collectifs.

Quel que soit l'organisme d'assurance, il est soumis à des obligations réglementaires en matière de provisionnement et de sécurité financière, contrôlé par l'Autorité de Contrôle Prudentiel (ACPR).

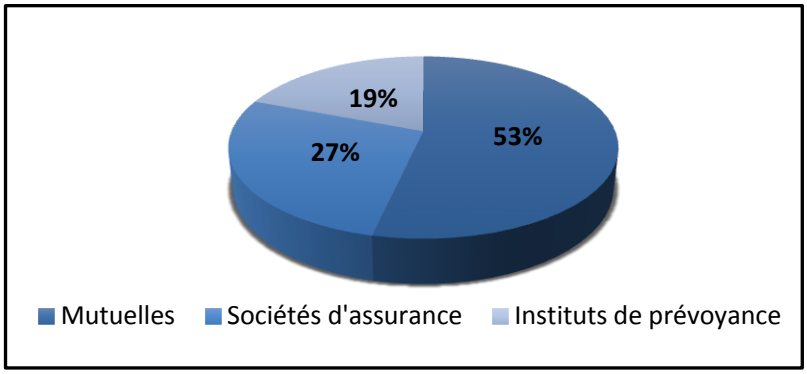

Figure 4 : Répartition du chiffre d'affaire 2014 par les trois organismes complémentaires

c) La répartition des dépenses de santé

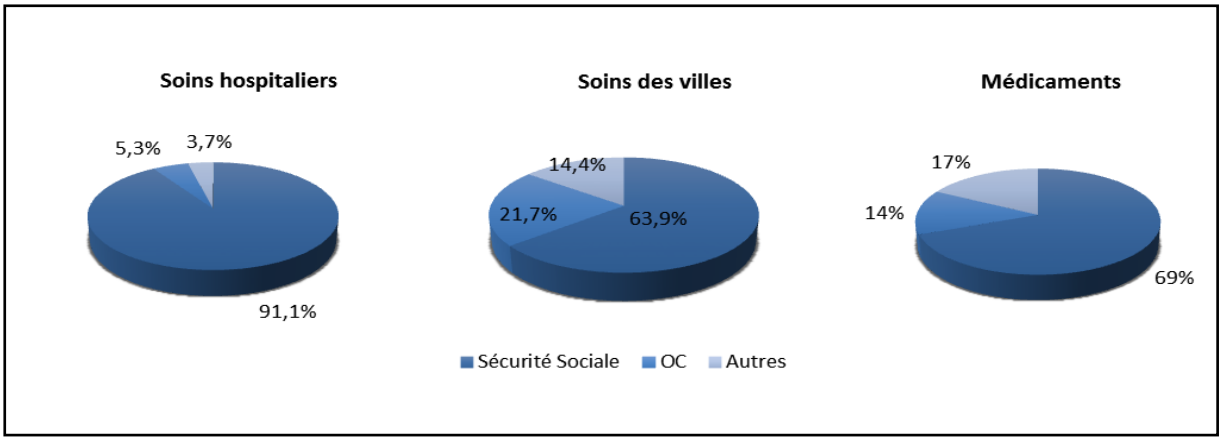

Figure 5 : Répartition des principaux postes de santé de 2014 entre les différents organismes

La Sécurité Sociale

En France, une part de la richesse nationale est consacrée à la production de services de santé. La part de la dépense de santé dans le produit intérieur brut (PIB) est un indicateur qui permet de mettre en perspective les dépenses de santé et leur dynamique avec les ressources du pays. En 2014, la dépense courante de santé s'est élevée à 256,9 milliards d'euros, soit 12% du PIB. Ce montant place la France parmi les pays de l'OCDE qui consacrent le plus de richesse aux dépenses de santé. La Sécurité Sociale a pris en charge en 2014 : 91% des dépenses hospitalières, 64% des soins de ville et 62% des dépenses de biens médicaux. Il existe 3 agrégats de mesure d'évolution des dépenses de santé : la consommation de soins et de biens médicaux (CSBM), la dépense courante de santé (DCS) qui prend en compte la CSBM et ajoute d'autres éléments, et la dépense courante de santé au sens international (DCSi) qui est utilisée pour comparer les dépenses de santé entre les pays. En France, la consommation des soins et biens médicaux (CSBM) représente les trois quarts de la dépense courante de santé avec une dépense de 190,6 milliards d'euros en 2014. Parmi les postes composant la consommation de soins et biens médicaux (CSBM), les dépenses hospitalières représentent la part la plus importante avec 88,6 milliards d'euros en 2014, suivies des soins de ville avec 50 milliards d'euros et des dépenses de biens médicaux avec 13,8 milliards d'euros. Les dépenses de la Sécurité Sociale continuent d'augmenter chaque année entrainant un déficit pour la Sécurité Sociale. C'est pourquoi, la loi de financement de la Sécurité Sociale vise à maitriser les dépenses sociales et de santé. Elle détermine les conditions nécessaires à l'équilibre financier de la Sécurité Sociale et fixe les objectifs de dépenses en fonction des prévisions de recettes. Ainsi, dans le cadre du projet de loi de financement de la Sécurité Sociale (PLFSS), l'évolution des dépenses d'Assurance maladie est maitrisée en fixant un objectif national (ONDAM).

L'ONDAM

Le champ de l'objectif national des dépenses d'Assurance maladie (ONDAM) est celui des remboursements de la Sécurité Sociale qui constituent les trois quarts des dépenses de santé. La loi de financement de la Sécurité Sociale pour 2016 a été exécutée à 185,2 milliards d'euros comme prévu dans l'objectif de dépenses. La progression de 1,8% des dépenses dans le champ de l'ONDAM en 2016 est la plus faible depuis sa création. L'objectif pour 2017 serait de 190,7 milliards d'euros soit une progression de 2,1% par rapport aux dépenses estimées pour 2016.

Les organismes complémentaires

Selon les comptes nationaux de la santé, 29 milliards d'euros ont été versés en 2014 au titre de l'assurance complémentaire. Les organismes complémentaires ont financé une part toujours croissante de la consommation de soins et de biens médicaux. En 2014, elle est de 13,5%. Ce financement est en hausse en raison de la prise en charge des dépassements d'honoraires des médecins et de la prise en charge du forfait hospitalier. En revanche, en ce qui concerne le remboursement des médicaments, la participation des organismes complémentaires est en baisse du fait de l'instauration de franchise de 50 centimes par boîte de médicaments. En 2014, les mutuelles restent l'acteur principal du marché de la complémentaire santé puisqu'elles représentent 53% des prestations versées par les organismes complémentaires (**Figure 4**). L'Assurance maladie complémentaire participe donc à la réduction du reste à charge des ménages. En France en 2014, le reste à charge des ménages s'élevait à 8,5% de la CSBM.

Les prestations versées par les organismes complémentaires en 2016 sont restées stables par rapport à 2015 avec cependant une baisse de 6% au niveau des contrats à adhésion individuelle, compensée par une hausse de 15% au niveau des contrats collectifs.

2. Les mécanismes de remboursements

a) Le fonctionnement du remboursement par la Sécurité Sociale

Les dépenses de soins peuvent être remboursées à toute personne assurée du fait de sa propre activité professionnelle et à ses ayants droits.
Les prestations de la branche maladie du régime général de la Sécurité Sociale sont de deux natures : des prestations en espèces et des prestations en nature.
Les prestations en espèces des régimes obligatoires ont pour objet de compenser une partie des baisses de revenus subis par les assurés du fait de leur état de santé, temporaire ou durable. Hors accident du travail, il y a trois grandes catégories de prestations :

- Les indemnités journalières (IJ), versées en cas d'arrêt de travail (AT) pour maladie ou hospitalisation.
- Assurance maternité, destinée à compenser la perte de salaire pendant le congé maternité.
- Assurance invalidité qui indemnise la perte de capacité de travail sous forme d'un versement d'une pension.

Les prestations en nature sont destinées au remboursement total ou partiel, des dépenses médicales, paramédicales et des frais d'hospitalisation. Dans la suite, le périmètre de l'étude se limite aux prestations en nature et donc à l'étude de la consommation de prestations médicales.
La Sécurité Sociale rembourse une partie des frais réels correspondant au coût global d'un acte, d'un soin, d'une prestation. Et cela pour toute personne résidant sur le territoire français, peu importe sa situation d'emploi, de ressources ou sa profession. L'autre partie des frais est remboursé par la complémentaire santé selon le niveau de garantie de l'assuré. La somme totale du remboursement ne peut pas excéder les frais réels que l'assuré a engagés.

- Pour chaque acte remboursé, il existe une assiette de remboursement, appelé Base de Remboursement de la Sécurité Sociale (BR), et anciennement Tarif de Convention (TC).

- Sur cette Base de Remboursement, la Sécurité Sociale rembourse une certaine part, appelée Taux de Remboursement.

- Depuis le 1er janvier 2005, une **participation forfaitaire d'1 euro** est de plus déduite des montants des remboursements de la Sécurité Sociale pour les consultations ou actes pratiqués par un médecin, les examens radiologiques et les analyses de biologie médicale.

- Le principe est le paiement par le bénéficiaire d'un « **ticket modérateur** » : la part qui reste à la charge du bénéficiaire après le remboursement par le régime d'Assurance maladie obligatoire. Le ticket modérateur est donc la part de la Base de remboursement qui n'est pas prise en charge par la Sécurité Sociale et qui ne comprend pas non plus la participation forfaitaire restant obligatoirement à la charge de l'assuré. Ce dernier peut alors être pris en charge par la complémentaire santé à laquelle il est adhérent.

- Les dépassements d'honoraires sont appliqués par quelques praticiens à honoraires libres (spécialistes) en supplément du tarif de convention.

Les remboursements en frais de santé de la Sécurité Sociale peuvent ainsi être synthétisés par le schéma qui suit :

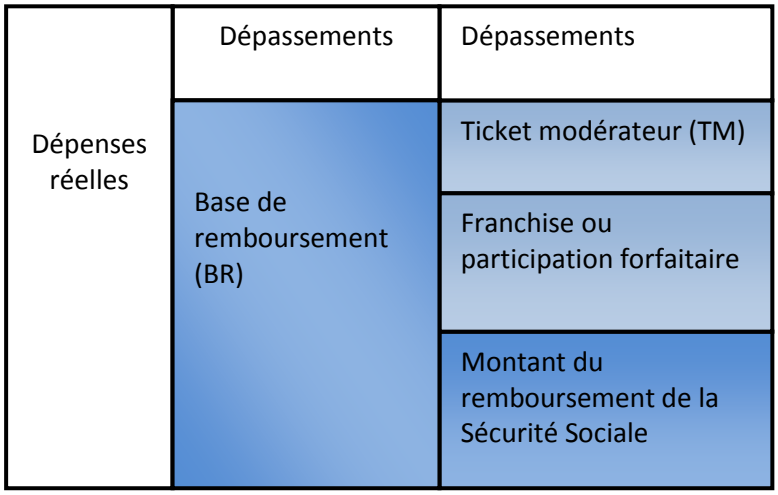

Figure 6 : Mécanisme de remboursement

Le schéma ci-dessus permet de bien saisir le fonctionnement de remboursement, et dans quelle mesure chaque acteur intervient. Le remboursement proposé par la Sécurité Sociale s'échelonnant sur plusieurs taux, n'est que partiel et laisse donc à la charge de l'assuré une part plus ou moins importante de la dépense. Cette base de remboursement dépendra en particulier de plusieurs variables tels que le régime duquel dépend l'assuré (général, Alsace-Moselle ou autres) l'importance de l'acte ou encore l'état de santé de l'assuré (maladies graves, recours à des médicaments irremplaçables). C'est dans cet intervalle qu'interviennent les organismes complémentaires. Leurs objectifs sont de limiter ce reste à charge pour l'assuré et même pour des niveaux de garanties plus élevés, de prendre en charge les dépassements d'honoraires. Il est à noter que le remboursement total versé par tous les organismes ne peut jamais excéder les frais réels engagés par l'assuré, ceci vérifiant le principe que l'assuré ne peut et ne doit pas s'enrichir grâce à sa consommation de soins.

Le taux de remboursement ou de prise en charge par l'Assurance maladie obligatoire varie selon la nature des frais engagés**.** Le tableau ci-dessous dresse les différents remboursements par poste de soins de la Sécurité Sociale :

Postes	REGIME GENERAL	REGIME ALSACE-MOSELLE
Hospitalisation médicale et chirurgicale		
▪ Frais de séjour	80% BR	100% BR
▪ Transport accepté	65% BR	100% BR
Actes médicaux		
▪ Généralistes, spécialistes	70% BR	90% BR
▪ Actes de chirurgie (ADC), actes techniques (ATM)	70% BR	90% BR
▪ Actes d'imagerie médicale, actes d'échographie	70% BR	90% BR
▪ Auxiliaires médicaux, analyses	60% BR	90% BR
Pharmacie		
▪ Médicaments reconnus comme irremplaçables ▪ Médicaments rendus majeurs ou importants ▪ Médicaments à service médical rendus modérés ▪ Médicaments à service médical rendus faibles	100% BR 65% BR 30% BR 15% BR	100% BR 90% BR 80% BR 15% BR
Dentaire		
▪ Consultation chirurgien-dentiste	70% BR	90% BR
▪ Soins dentaires	70% BR	90% BR
▪ Prothèses dentaires	70% BR	90% BR
Prothèses non dentaires		
▪ Prothèses auditives	60% BR	90% BR
▪ Orthopédie & autres prothèses	60% BR	90% BR
Optique		
▪ Monture	60% BR	90% BR
▪ Verres	60% BR	90% BR
▪ Lentilles acceptées	60% BR	90% BR
Cure thermale (acceptée SS)		
▪ Frais de traitement et honoraires	70% BR	90% BR
▪ Frais de voyage et hébergement	65% BR	65% BR

Tableau 1 : Récapitulatif des remboursements de la Sécurité Sociale

b) Le remboursement des soins par l'assurance maladie complémentaire

Avant de présenter les différents postes, il convient de définir les actes médicaux qui sont classifiés selon la classification commune des actes médicaux (CCAM). Les actes médicaux décrits par la CCAM sont les actes professionnels relevant de la compétence des membres des seules professions médicales.

La répartition des dépenses de santé se fait en regroupant les actes médicaux selon des postes, les principaux postes des soins étant les suivants :

- Actes médicaux (qui correspondent aux consultations et visites de médecins et spécialistes)
- Hospitalisation
- Pharmacie
- Optique
- Dentaire
- Autres (maternité, cures thermales…)

La nature des prestations garantie par les régimes complémentaires sont des garanties généralement complémentaires de celles qui sont obtenues dans un régime de base obligatoire. Il existe globalement cinq grands niveaux de remboursement complémentaires, les uns dépendant de celui effectué par la Sécurité Sociale, les autres pouvant être définis par rapport à une limite maximale d'intervention.

Les organismes complémentaires présentent leurs prestations santé sous forme de grille de garantie, indiquant les taux de remboursement ou forfaits qui seront appliqués pour chaque acte médical. Ces taux peuvent être exprimés de plusieurs façons :

- %TM : Prise en charge de tout ou partie du ticket modérateur
- %FR : Prise en charge des frais réels avec ou sans limite
- %RSS : Remboursement à hauteur de X% du montant de remboursement de la Sécurité Sociale
- %BRSS : Remboursement à hauteur d' Y% de la base de remboursement de la Sécurité Sociale
- Forfait : Remboursement à caractère forfaitaire. Il est notamment retenu pour les frais sans intervention de la Sécurité Sociale (maternité et chambre particulière par exemple) ou pour les frais pour lesquels le niveau du tarif de convention est « dérisoire » par rapport à la dépense moyenne réelle (exemple : l'optique).

L'assurance maladie complémentaire s'est largement développée depuis quelques années. La création de la couverture universelle complémentaire (CMU-C), en 2000, puis de l'aide à la complémentaire santé (ACS) en 2005, dispositifs destinés aux ménages aux revenus les plus modestes, ont permis l'élargissement à des populations qui en étaient jusqu'alors exclues.

II. Les évolutions réglementaires récentes

1. L'ANI de 2013

La loi n° 2013-504 du 14 juin 2013 ayant transposé l'accord national interprofessionnel (ANI) du 11 janvier 2013, porte sur la sécurité de l'emploi et notamment sur la généralisation de la complémentaire santé pour les salariés, et l'amélioration de la portabilité des garanties santé et prévoyance pour le demandeur d'emploi. L'ANI de 2013 modifie les devoirs des employeurs en matière de complémentaire santé.
Le dispositif de la généralisation de la complémentaire santé permettra de favoriser l'accès aux soins pour tous les salariés qui bénéficieront d'une couverture santé d'entreprise dont une partie sera pris en charge par l'employeur. Les entreprises devront donc obligatoirement mettre en place une complémentaire santé pour leur salarié au plus tard avant le 1er janvier 2016. Différentes étapes ont été mises en place depuis le 1er juin 2013 jusqu'à la mise en application de la loi le 1er janvier 2016.
Le décret du 8 décembre 2014 fixe des garanties minimales que l'on appelle « le panier de soins ». Cette couverture minimale comprend la prise en charge :

- De l'intégralité du ticket modérateur restant à la charge de l'assuré pour les consultations, actes et prestations remboursables par l'Assurance maladie obligatoire ;
- Du forfait journalier hospitalier sans limitation de durée ;
- Des dépenses de frais dentaires prothétiques et d'orthopédie dentofaciale à hauteur de 125% de la base de remboursement de la Sécurité Sociale ;
- D'un forfait tous les deux ans pour les frais d'optique à hauteur de 100 euros pour les verres simples, 200 euros pour les verres complexes et 150 euros pour les équipements mixtes.

2. Le contrat d'accès aux soins (CAS)

Le contrat d'accès aux soins est entré en vigueur le 1er décembre 2013, issu de la négociation sur l'encadrement des dépassements d'honoraires conclue par la signature d'un avenant n°8 à la convention médicale. L'objectif est de maitriser les dépassements d'honoraires et ainsi permettre aux patients d'être mieux remboursés.
Une distinction est faite entre les professionnels de santé, les médecins de secteur 1 dits les médecins conventionnés et ceux du secteur 2 non conventionnés. Les médecins du secteur 1 respectent les tarifs conventionnés de la Sécurité Sociale tandis que ceux du secteur 2 fixent librement le tarif des consultations. C'est en signant avec l'Assurance maladie un contrat d'accès aux soins que ces derniers font le choix de limiter leurs dépassements d'honoraires. De plus, le montant de la consultation ne doit pas être supérieur à 100% de la base de remboursement de la Sécurité Sociale. En échange de cet engagement, les médecins adhérents (médecins CAS) bénéficient d'avantages sociaux et de la cotisation de toutes les majorations réservées au secteur 1. Le contrat d'accès aux soins, souscrit sur la base du volontariat est effectif pour une durée de 3 ans avec une possibilité de résiliation à la date d'anniversaire du contrat.
L'avantage pour les patients est d'être mieux remboursé en bénéficiant des mêmes bases de remboursements de la Sécurité Sociale appliquées au secteur 1 et également une prise en charge privilégiée pour les dépassements d'honoraires par les complémentaires santé. Par exemple, pour une consultation chez un médecin CAS, la base de remboursement est de 28 euros tandis que chez un médecin non CAS, la base de remboursement est de 23 euros.
Les complémentaires santé participent au contrat d'accès aux soins, elles proposent une prise en charge différente pour les dépassements d'honoraires d'un médecin CAS ou non CAS.
Le CAS a été rebaptisé OPTAM et OPTAM-CO au 1er janvier 2017.

3. Le contrat responsable de 2014

La loi du 13 août 2004 relative à l'Assurance maladie a procédé à une réforme de l'Assurance maladie. Elle a responsabilisé le patient en instaurant des franchises et une contribution forfaitaire de 1 euro. Elle a également créé la notion de contrat responsable afin de responsabiliser les patients, les complémentaires et les praticiens.

La loi de financement de la Sécurité Sociale de 2014 a reconfiguré le contrat solidaire et responsable en posant trois grands principes :

- L'élargissement du champ des prestations couvertes par le contrat responsable ;
- L'introduction de niveaux minimums et maximums de prise en charge ;
- L'introduction de conditions permettant de renforcer la mutualisation des risques couverts.

Le décret du 18 novembre 2014 est venu compléter le dispositif mis en œuvre par la loi du 13 août 2004 en imposant de nouvelles modalités pour le contrat responsable. Il rend obligatoire le respect de certains plafonds de remboursements et définit le nouveau cahier des charges des contrats responsables. Ce

décret a pour objectif d'encadrer les dépenses de santé, en particulier de limiter les dépassements d'honoraires et de baisser les tarifs en optique. Le contrat responsable ouvre droit à un régime social et fiscal avantageux. Les assureurs sont soumis à une taxe dont le taux varie selon le type de contrat. Depuis le 1er janvier 2014, les contrats non responsables sont soumis à la taxe TSCA (Taux spécial sur les contrats d'assurances) passée de 9% à 14%. Les contrats responsables continuent de bénéficier d'un taux dérogatoire de 7%.

Avant la mise en place de ce décret, la notion de contrat responsable se limitait au respect de remboursements sur certaines prestations et à l'interdiction de prise en charge de certaines franchises ou majorations. Désormais, le contrat responsable se caractérise par l'instauration de plafonds de remboursements en vue de limiter les dépassements d'honoraires et les tarifs en optique.

Les dépassements d'honoraires des médecins qui n'ont pas signé le contrat d'accès aux soins (CAS) ne seront ainsi remboursés qu'à hauteur de 200% maximum de la base de remboursement de la Sécurité Sociale. En revanche, la prise en charge sera supérieure si le médecin a conclu un CAS. L'encadrement des garanties vise également l'optique. Le plafond de remboursement est fixé à 470 euros pour une paire de lunettes à verres simples, 750 euros pour des verres complexes et jusqu'à 850 euros pour un équipement complexe à forte correction. La prise en charge des montures reste limitée à 150 euros.

4. La convention médicale de 2016

La nouvelle convention médicale régissant les relations entre les professionnels de santé et l'Assurance maladie a été approuvée par arrêté du 20 octobre 2016 et signée le 25 août 2016 avec trois syndicats représentatifs des médecins libéraux. Elle poursuit l'objectif de la convention médicale de 2011 à faire progresser la santé, développer la prévention et l'accès aux soins pour tous. Plusieurs mesures permettent de répondre à ces enjeux :

- La revalorisation de 2 euros du tarif de la consultation de référence pour les médecins généralistes au 1er mai 2017 pour l'aligner sur celui des spécialistes : la base de remboursement passe ainsi de 23 euros à 25 euros, par conséquent le ticket modérateur passe de 6,90 euros à 7,50 euros. Les généralistes de secteur 2 non signataires du contrat d'accès aux soins (CAS) ne bénéficieront pas de cette revalorisation, dans la mesure où la base de remboursement des spécialistes de secteur 2 non signataires du CAS est de 23 euros.
- La revalorisation de la consultation coordonnée : passage de 28 euros à 30 euros en juillet 2017.
- La création de nouveaux tarifs de consultations pour des prises en charge plus complexes : prévention de MST seront dorénavant valorisées à 46 euros à partir de novembre 2017, les consultations très complexes seront valorisées à 60 euros à partir de novembre 2017.
- La rénovation du dispositif de maîtrise des dépassements d'honoraires, OPTAM et OPTAM-CO (ce dernier réservé aux chirurgiens et gynécologues obstétriciens) pour encourager tous les médecins exerçant en secteur 2 à stabiliser leurs dépassements et à accroître la part des soins facturés aux tarifs opposables.
- Une ROSP (rémunération sur objectif de santé publique) renforcée et élargie.

Ces mesures qui sont majoritairement des mesures tarifaires, vont impacter à la hausse les prestations versées par les complémentaires santé. La première mesure, la revalorisation de 2 euros de la consultation et de la visite à domicile des médecins généralistes au 1er mai 2017 peut être la plus impactante pour les complémentaires santé. L'impact dépend :

- Du poids des consultations et visites de généralistes dans les prestations versées ;
- Du niveau de couverture des contrats et les taux de dépassement pratiqués par les médecins du secteur 2.

5. Le règlement arbitral en dentaire 2017

Suite à l'échec des négociations entre les chirurgiens-dentistes, l'UNCAM (Union nationale des caisses d'Assurance maladie), et l'UNOCAM (Union nationale des organismes d'assurance maladie complémentaire) pour la mise en place d'un avenant n°4 à la convention nationale dentaire, une proposition de règlement arbitral a été proposée par Bertrand Fragonard[2], approuvée par la Ministre des Affaires Sociales et de la Santé par arrêté du 29 mars 2017. Ce règlement arbitral prendra effet à partir du 1er janvier 2018 et les mesures seront échelonnées sur la période 2018-2021.

Face à un taux de renoncement aux soins dentaires pour raisons financières jugé trop important, le principal objectif du règlement est un rééquilibrage de l'activité des chirurgiens-dentistes au profit des soins conservateurs et chirurgicaux, tout en diminuant le reste à charge des ménages. Ce rééquilibrage passe par la revalorisation de la base de remboursement des soins conservateurs et des actes prothétiques les plus courants, ainsi que par le plafonnement des tarifs des actes prothétiques. La revalorisation et le plafonnement sont progressifs, et échelonnés de 2018 à 2021.

Au-delà de ce mécanisme de rééquilibrage, le règlement arbitral prévoit trois autres mesures :

- La revalorisation de certains tarifs maximaux pour les bénéficiaires de la CMU complémentaire. Cette revalorisation sera ensuite, par arrêté, appliquée aux bénéficiaires de l'ACS (Aide au paiement d'une Complémentaire Santé);
- La prise en charge d'un bilan parodontal, suivi le cas échéant, de soins parodontaux, au profit des patients diabétiques ;
- Une meilleure prise en charge des personnes en situation de handicap mental sévère, avec la possibilité de valoriser les techniques de sédation consciente telles que le MEOPA.

Le règlement arbitral impacte les complémentaires santé par plusieurs mesures :

- La revalorisation de la base de remboursement des soins dentaires : les soins conservateurs et de prévention représentent une part faible des honoraires des dentistes mais une majorité de leurs actes. A partir du 1er janvier 2018, les soins conservateurs les plus courants seront revalorisés et les tarifs seront majorés de 40% à 70%. Cette revalorisation sera échelonnée sur 4 ans.
- L'évolution de la base de remboursement des actes prothétiques : la base de remboursement (BR) des actes prothétiques les plus courants sera révisée. Par exemple, une revalorisation pour la couronne dentaire : aujourd'hui à 107,50 euros la BR passera à 112,50 euros en 2018 puis à 120 euros dès 2019.
- Le plafonnement des tarifs des actes prothétiques : ce dispositif consiste en la mise en place d'un plafonnement propre à chaque acte, avec une différenciation entre les couronnes. Le prix limite de facturation instauré correspond à l'honoraire maximal facturable par le chirurgien-dentiste au patient. Le plafonnement sera échelonné de 2018 à 2021.

L'impact des évolutions sur l'ensemble des prestations versées par les complémentaires santé sera propre à chaque segment, et dépendra du poids des prothèses dentaires, et des soins dentaires dans le portefeuille, et du niveau de garanties proposé.

Les dernières informations concernant le règlement arbitral dentaire est un décalage de l'entrée en vigueur au 1er janvier 2019. La nouvelle ministre de la Santé a souhaité une reprise de dialogue avec les

[2] Magistrat à la cour des Comptes

représentants des chirurgiens-dentistes pour une négociation tarifaire avec l'Assurance maladie et le report au 1er janvier 2019 du règlement arbitral. Toutefois, les mesures permettant d'améliorer la couverture des bénéficiaires de la CMU-C et de bénéficier, pour les personnes touchant l'ACS de tarifs plafonnés sur les prothèses, entreront en vigueur au 1er octobre 2017 comme prévu dans le règlement arbitral.

III. Le périmètre AG2R La Mondiale

1. Le groupe AG2R La Mondiale

AG2R La Mondiale est né de l'union d'AG2R et de La Mondiale, deux groupes bénéficiant d'expertises complémentaires, respectivement en retraite complémentaire, prévoyance et santé, en épargne et en retraite supplémentaire.

Créé en 1905, le groupe La Mondiale avait pour but initial de compléter les revenus des salariés de petites entreprises et des travailleurs indépendants. Quant à l'Association Générale de Retraite par Répartition (AGRR), qui deviendra AG2R en 1992, elle a été créée plus tard en 1951, devenant la première caisse de retraite par répartition pour les salariés non cadres. Jusqu'à aujourd'hui, le groupe s'est développé en France et à l'international, devenant ainsi en 2008 la SGAM AG2R LA MONDIALE, une institution mondialement reconnue en assurances de personnes, notamment grâce à La Mondiale Europartner.

Depuis sa création, AG2R La Mondiale est un groupe de protection sociale, paritaire et mutualiste lui permettant ainsi de placer l'intérêt de ses 15 millions d'assurés et ayants droits au cœur de ses préoccupations. AG2R La Mondiale couvre l'ensemble des besoins de protection sociale des personnes et de leur famille, collectifs comme individuels, en santé, prévoyance, épargne et retraite. Il s'adresse à toutes les catégories d'assurés, quel que soient leur âge, leur statut social ou leur secteur professionnel. En tant que société de personnes, le groupe laisse ses assurés prendre les décisions par le biais de leurs représentants membres des organes de gouvernance. La richesse du groupe est consacrée au développement de nouvelles prestations dans l'intérêt de l'assuré ou directement réinvestie dans les fonds propres de l'entreprise assurant ainsi un avenir stable à ses collaborateurs.

Le fonctionnement paritaire d'AG2R La Mondiale se caractérise par une gestion assurée par un Conseil d'Administration paritaire, composé pour moitié de :

- Représentants patronaux, membres adhérents ayant adhéré à un règlement de l'institution ou souscrit un contrat auprès de celle-ci ;
- Représentants salariés, désignés par les unions départementales des confédérations syndicales de salariés : CGT, FO, CFDT, CGC, CFTC.

Le groupe est composé de trois types de structures. Il est donc régi par l'union de trois codes : de la Sécurité Sociale, des assurances et de la mutualité. En effet, AG2R La Mondiale est un groupe de protection sociale composé :

- D'institutions de retraite complémentaire Agirc et Arrco ;
- De mutuelles d'assurance et ;
- De deux instituts de prévoyance.

Le schéma ci-dessous présente les différentes entités du groupe AG2R La Mondiale :

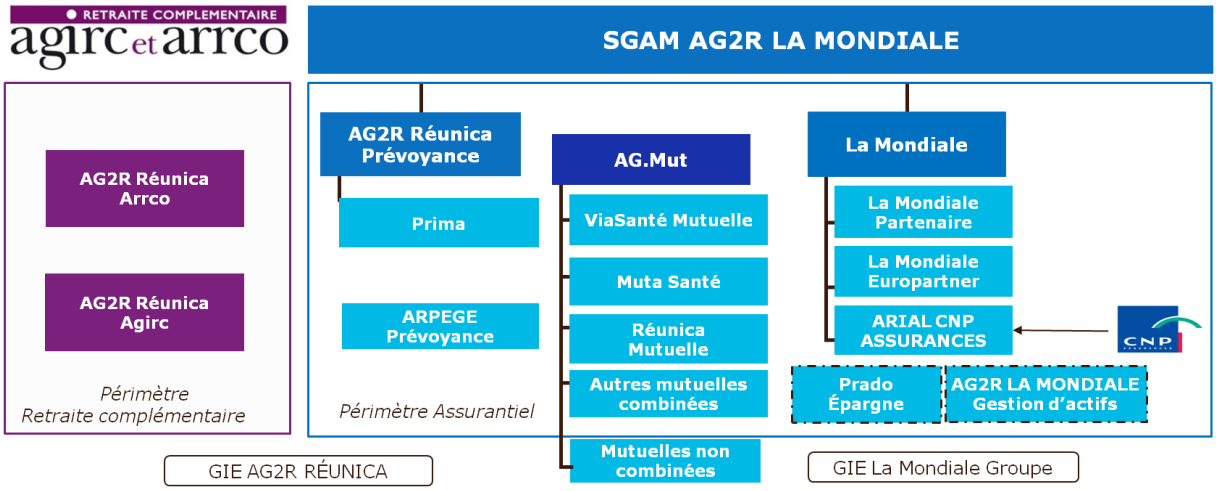

Figure 7 : Les institutions d'AG2R La Mondiale

Avec une collecte brute totale de 28,2 milliards d'euros en 2016, AG2R La Mondiale réalise 18,1 milliards d'euros au titre de la retraite complémentaire obligatoire et 10,1 milliards d'euros au titre des activités d'assurance.

Le portefeuille santé réalise un chiffre d'affaire de 2,2 milliards d'euros et représente ainsi 22% du chiffre d'affaire du groupe en accompagnant 3 millions de bénéficiaires. Les parts respectives des différentes entités sont présentées ci-dessous :

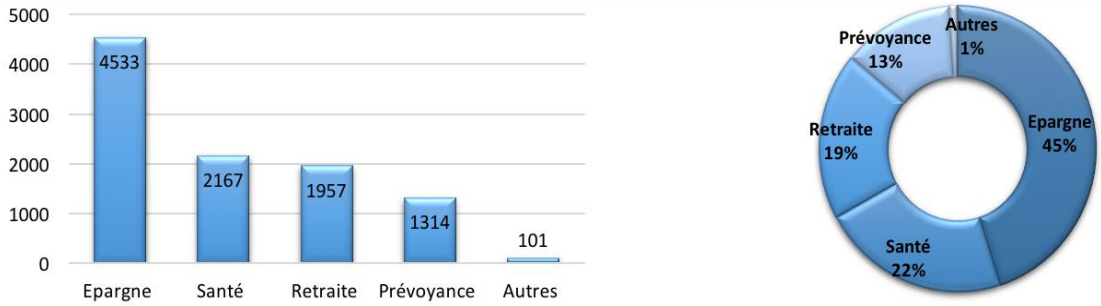

Figure 8 : Ventilation du chiffre d'affaire 2016 par risque

2. Le portefeuille santé d'AG2R La Mondiale

a) L'assurance collective et individuelle

Le portefeuille santé d'AG2R La Mondiale est divisé en deux segments : le segment individuel et le segment collectif. Dans le cadre de l'étude de la dérive, on se focalisera sur le segment individuel. Le segment collectif est divisé en trois parties : le portefeuille standard, le portefeuille sur-mesure et le portefeuille conventionnel. Le portefeuille standard, ou également appelé mode P, concerne les entreprises de moins de 100 salariés à qui l'on propose des gammes prédéfinies (en termes de garanties et de tarifs). Le portefeuille sur-mesure, ou également appelé mode H, concerne les entreprises de plus de 100 salariés qui ont la possibilité de créer librement leurs garanties avec les niveaux de couverture qui les intéressent. Enfin, un contrat est considéré comme relevant d'un suivi conventionnel s'il dépend d'une

Convention Nationale pour laquelle le groupe AG2R La Mondiale a mis en place une offre spécifique, par exemple dans le cadre d'une désignation ou d'une recommandation.
Quant au portefeuille individuel, l'assuré souscrit à titre individuel et peut faire bénéficier ses ayants droits. Le portefeuille est composé de plusieurs gammes, dont chaque gamme comporte plusieurs garanties et offre le choix de plusieurs niveaux de couverture. Le choix de ces garanties dépend du besoin du bénéficiaire, des prestations de remboursement désirées. C'est pourquoi, chaque gamme offre des formules modulables pour être mieux adaptées aux besoins de chacun. L'assurance santé individuelle présente des risques d'anti-sélection et de sélection adverse.

b) Identification des risques

- *Le risque d'anti-sélection*

L'anti-sélection se traduit par le fait de détenir une information cachée dont les assurés détiennent sur leur propre risque, et qui n'est pas accessible, et observable par les assureurs.
Le risque d'anti-sélection se trouve essentiellement en assurance individuelle, et dans une moindre mesure, en assurance collective lorsque l'adhésion est facultative. En effet, en santé collective, le risque d'anti-sélection est moins ressenti dans la mesure où le contrat couvre un ensemble de salariés. Ceux-ci ne sont pas forcément concernés par les mêmes besoins. Cependant, il peut être identifié et analysé dans le cadre des adhésions facultatives (Contrat de base facultatif mis en place par l'entreprise, surcomplémentaire souscrite par le salarié). Cela se traduit par le fait que le salarié souscrit au contrat facultatif car il a conscience de ses futurs besoins. Un salarié n'ayant aucun problème de vue ne souscrira pas à une surcomplémentaire sur l'optique. De même, si l'entreprise exige dans le contrat des garanties spécifiques, on peut prévoir dans une moindre mesure l'intégration d'une anti sélection.
Ainsi, ce risque est dû à une dissymétrie d'information entre l'assureur et les assurés. L'assureur ne pouvant mesurer précisément cette présence d'asymétrie d'information appliquera une prime moyenne sur des types d'assurés en fonction de leurs paramètres de risque induisant en conséquence de l'anti-sélection.

- *Le risque de sélection adverse*

Dans le cadre de l'assurance santé, le risque de sélection adverse ou aléa moral se caractérise par le fait d'adopter sa consommation à ses garanties. L'assuré n'adoptera pas le même comportement en fonction de son niveau de couverture. Il modifiera son comportement face au risque en fonction de sa couverture ainsi que de son niveau de revenu. En effet, lorsque les soins sont remboursés d'une part par la Sécurité Sociale, et d'autre part par les complémentaires santé, le reste à charge des patients peut s'avérer faible pour certains soins. L'assuré peut donc augmenter par exemple sa fréquence de consultations chez le médecin. Il a donc tendance à consommer davantage.
De plus, ce risque est d'autant plus important en fonction du revenu. En effet, l'assuré sera incité à consommer plus souvent car le reste à charge lui est moins contraignant.

IV. Objectif du mémoire

Afin de répondre à ses objectifs de rentabilité et à ses obligations de solvabilité, l'assureur a besoin de prévoir l'évolution de la sinistralité de son portefeuille dans les années futures. Or, les prestations de santé versées chaque année sont impactées par l'évolution naturelle de la consommation de soins ainsi que par les évolutions réglementaires illustrées dans le paragraphe II. De ce fait, l'indicateur central sur lequel s'appuie l'assureur pour décider de faire évoluer les cotisations de ses adhérents (indexation tarifaire) est cette fameuse « dérive » de la consommation de soins. Dans ce contexte, le mémoire a pour objectif de mettre en place un modèle prédictif de la dérive afin d'obtenir des prévisions à court terme.
Plusieurs méthodologies sont mises en œuvre :

- Dans un premier temps, une approche macroéconomique sera présentée.
- Dans un second temps, une approche individuelle sera utilisée visant à intégrer tous les facteurs individuels et éventuellement des facteurs exogènes par le biais de méthodes économétriques.

Définition de la dérive

Il faut savoir tout d'abord que la consommation des soins de santé augmente avec l'âge, et est différente entre les hommes et les femmes.
La dérive des soins de santé est l'évolution du coût du risque santé, c'est à dire l'évolution des prestations payées en moyenne par assuré, toutes choses égales par ailleurs (sans évolution des garanties du contrat, sans évolution de l'âge des assurés, sans évolution de la répartition hommes/ femmes).
Chaque année, à âge constant, la consommation moyenne par personne protégée augmente naturellement du fait :

- Des facteurs épidémiologiques : augmentation de la prévalence des maladies chroniques (diabète, hypertension, cholestérol, …), des cancers, à cause des facteurs environnementaux et comportementaux.
- Des coûts plus importants induits par le progrès technique (nouvelles molécules plus efficaces et plus coûteuses, nouveaux dispositifs médicaux, etc.).
- De l'inflation (dépassements d'honoraires, coût des dispositifs médicaux de type optique, etc.).

Cette croissance doit donc être maitrisée en mettant en place des stratégies de régulation du système de santé. Si aucune action de régulation n'était mise en place, la dérive naturelle serait de l'ordre de 4%. Pour contenir cette dérive naturelle et réduire ainsi le déficit de l'Assurance maladie, le Parlement fixe chaque année l'ONDAM (objectif national des dépenses d'Assurance Maladie), l'enveloppe dont dispose l'Assurance maladie pour financer les prestations. En 2017, l'ONDAM est fixé à +1,75% par rapport à 2016. La LFSS (loi de financement de la Sécurité Sociale) vient ensuite préciser les conditions d'atteinte de cet ONDAM (baisses de prix, amélioration de l'efficience des soins, etc.).
Le gouvernement présente des mesures de régulation dans le PLFSS (Projet de loi de financement de la Sécurité Sociale). Ces mesures peuvent être de trois types :

- Maîtriser l'évolution des prix, par négociation avec les professionnels de santé ou encore par fixation autoritaire (baisse des prix des médicaments par exemple) ;
- Mettre en place des actions de prévention de l'évolution des maladies chroniques ;
- Diminuer les dépenses en diminuant les taux de remboursements (ce qui induit un transfert de charge vers l'assurance complémentaire).

Une des missions de l'actuaire consiste à effectuer une prévision de la dérive naturelle pour les années à venir et à analyser les mesures du PLFSS chaque année.

Partie 2 : Etude macroéconomique

L'assurance complémentaire santé, qui par nature renforce les remboursements de la Sécurité Sociale, a un besoin de connaitre l'évolution des dépenses de santé. En effet, elle doit mesurer cette évolution, la dérive, pour connaitre la progression future de son portefeuille et appliquer les augmentations de tarif appropriées.
Cette partie présentera les résultats d'une première approche macroéconomique mise en œuvre dans le cadre des décisions d'indexation tarifaire 2018. Les limites d'une telle approche seront discutées, afin d'introduire la nécessité d'une approche micro-économique mise en œuvre dans un second temps.

I. Contexte et périmètre de l'étude

1. Contexte de l'étude

Dans un contexte de contrat responsable, il est difficile d'extraire une dérive pure pour l'année 2016. En effet, des importantes modifications ont été apportées aux garanties en 2016 dans le cadre de la mise en conformité au contrat responsable. Il n'est donc pas possible de dégager une tendance 2016 pure. Nous ne pouvons calculer avec certitude l'impact du contrat responsable sur tout le portefeuille servant à calculer la dérive. L'évolution de la consommation entre 2015 et 2016 contient donc l'effet contrat responsable par conséquent, il est difficile d'obtenir une mesure précise de la dérive. C'est pour cette raison qu'il est nécessaire de réaliser une étude afin de calculer une dérive pure. Cette étude a pour but de calculer une tendance sur le portefeuille individuel avant 2016, puis, obtenir une estimation de la dérive à horizon 2018. Les années 2017 et 2018 sont marquées par des mesures réglementaires impactant les dépenses de soins de santé. Cette étude de dérive prendra donc en compte ces effets conjoncturels : mesures de la convention médicale comme la revalorisation de 2 euros du tarif de la consultation des médecins généralistes au 1er mai 2017.

2. Périmètre de l'étude

L'étude est réalisée sur le portefeuille individuel du SI AG2R. La période d'étude des prestations est de trois ans (2014-2015-2016), données arrêtées à fin mars 2017. Les prestations versées par AG2R La Mondiale pour des sinistres survenus en 2016 peuvent être versées en 2017. L'adhérent ayant deux ans pour se faire rembourser. C'est pour cela que les études de dérive peuvent être vues à N+1 ou à N+2, pour avoir des années les plus complètes possibles. Dans notre cas, il s'agira d'une vision à N+1. L'étude est réalisée sur un portefeuille fermé c'est à dire sur la base des bénéficiaires présents du 1er janvier 2014 au 31 décembre 2016.
L'étude de la dérive sur les années N-3 à N-1 permet d'étudier la dérive sur des survenances terminées, tandis qu'une étude sur la dérive sur les années N-2 à N permet d'avoir des premiers éléments sur l'année en cours mais, qui sont moins robustes car on regarde la sinistralité sans recul. C'est pour cette raison que l'étude est effectuée sur les années 2014-2015-2016.

3. Méthodologie

Cette partie s'attache à décrire les étapes d'extraction des données, et la méthodologie de l'étude.

Extraction des données :
Une première étape consiste à extraire des données par le logiciel SAS. La base des prestations détaille l'ensemble des prestations versées par AG2R La Mondiale pour les gammes du portefeuille individuel

pour tous les actes médicaux. L'affectation des postes, et sous-postes pour les différents actes médicaux s'effectue par la récupération des identifiants de garanties (idgar) qui permet d'identifier les actes médicaux qui sont répartis selon des postes et sous-postes (garanties), et ainsi leur associer le montant des prestations versées par AG2R La Mondiale.
La base des bénéficiaires rassemble les données démographiques et sociodémographiques comme l'âge, le sexe, le régime, le département du bénéficiaire. Le portefeuille individuel est composé de plusieurs gammes que l'on répartit en deux types : les gammes d'actifs et les gammes de seniors. Cette distinction s'effectue par rapport à l'âge moyen des bénéficiaires présents dans chaque gamme. Ainsi, les gammes d'actifs rassemblent les gammes dont les bénéficiaires sont âgés de moins de 60 ans tandis que les gammes de seniors rassemblent les bénéficiaires de plus de 60 ans. Cette segmentation est établie car la consommation des soins de santé est très différente selon l'âge (argument qui sera développé par la suite).

Méthodologie de l'étude :
Après avoir extrait les données des prestations et des bénéficiaires, l'étude consiste à réaliser une **étude macroéconomique** de l'évolution de la dérive et pouvoir la projeter à l'horizon 2018.
Dans un premier temps, les montants des prestations sont regroupés par postes, sous-postes et par gammes. De plus, le calcul de l'âge moyen par gamme permet d'établir une segmentation entre les gammes d'actifs et de seniors. Cette segmentation est établie car le comportement des actifs et des seniors en terme de consommation de soins de santé est différente. D'une part, la structure des dépenses est différente (dépenses hospitalières, et de pharmacie prépondérantes chez les seniors). D'autre part, la hausse des dépenses chez les seniors s'explique pour partie par l'importance des soins qui précèdent le décès. En effet, la dernière année de vie concentre une part importante des dépenses de santé. La dépense des seniors est donc élevée sur certains postes du fait de ces consommants.
L'étude est réalisée sur un portefeuille fermé, c'est à dire que ce sont les mêmes bénéficiaires présents sur 3 ans, donc l'âge moyen du portefeuille, par construction, augmente d'un an exactement entre chaque année d'étude. Comme expliqué précédemment, l'étude consiste à obtenir une dérive pure. Pour ce faire, il convient de retirer les effets non liés à l'évolution de la consommation. L'âge constitue un de ces effets, c'est pour cette raison que nous estimons l'effet âge qui sera déflaté de la dérive. L'estimation de la dérive des années 2017 et 2018, s'effectue en calculant une tendance de l'évolution de la consommation des soins de santé de 2015/2014, et ainsi projeter cette tendance à court terme. La tendance n'est pas calculée à partir de l'évolution de la consommation 2016/2015 car celle-ci est biaisée par l'effet contrat responsable. A cette tendance 2015/2014 s'ajoute l'estimation des effets conjoncturels en point de dérive : l'estimation des nouveaux impacts de la convention médicale et du règlement arbitral dentaire.

II. Résultats de l'étude

Dans cette partie, nous présenterons les dépenses des soins de santé pour les années 2014 et 2015 et leur évolution. Cette présentation des dépenses se fera en distinguant les actifs des seniors. Puis, nous déterminerons la tendance de l'évolution de la consommation des soins de santé entre les années 2015/2014. Enfin, nous projetterons cette tendance pour estimer la dérive des années 2017 et 2018.

1. Présentation des dépenses

L'évolution de la consommation des soins de santé est présentée par postes en distinguant les actifs et les seniors. Le regroupement se fait selon 6 postes :

- Les actes médicaux ;
- Autres prestations ;
- Hospitalisation ;
- Pharmacie ;
- Optique ;
- Dentaire.

Le nombre de bénéficiaires présents ces trois années consécutives est de 191 617, soit 55% des bénéficiaires du portefeuille individuel de 2014. Sur ces 191 617 bénéficiaires, on compte 47 644 actifs, dont l'âge moyen en 2014 est de 34 ans et 143 973 seniors dont l'âge moyen en 2014 est de 68 ans.
Le coût moyen par bénéficiaire est le montant total remboursé par l'assureur (on parlera de consommation totale) divisé par le nombre de bénéficiaires présents dans le portefeuille sur la période considérée.

Dans la suite, le terme consommation sera employé pour désigner les prestations versées par AG2R La Mondiale.

$$Coût\ moyen = \frac{consommation}{nombre\ de\ bénéficiaires}$$

L'évolution de la consommation entre l'année N et N-1 peut s'écrire de la façon suivante :

$$Evolution\ \ N/N-1 = \frac{coût\ moyen\ N}{coût\ moyen\ N-1} - 1$$

où le coût moyen est la consommation moyenne par bénéficiaire ou le remboursement assureur moyen par bénéficiaire.
Les tableaux ci-dessous présentent l'évolution de la consommation par poste entre les années 2014 et 2015 :

Pour des raisons de confidentialité, les consommations moyennes ne sont pas affichées.

Actifs	Poids du poste en 2014	Poids du poste en 2015	Poids du poste en 2016	Evolution de la consommation 2015/2014
Actes médicaux	18,1%	17,8%	16,6%	0,1%
Autres prestations	17,2%	17,7%	17,4%	4,3%
Dentaire	14,3%	14,9%	15,2%	6,3%
Hospitalisation	15,6%	14,4%	16,2%	-5,9%
Optique	18,1%	18,9%	18,4%	6,1%
Pharmacie	16,8%	16,3%	16,2%	-1,1%
Total hors hospitalisation	**84%**	**86%**	**84%**	**3,0%**
Total	**100%**	**100%**	**100%**	**1,6%**

Tableau 2 : Evolution de la consommation de soins des gammes d'actifs entre les années 2014 et 2015

En 2015, pour les gammes d'actifs, les postes optique et actes médicaux représentent la part de remboursement la plus importante. D'autre part, l'évolution de la consommation du poste hospitalisation est négative -5,9%, c'est pour cette raison qu'une ligne supplémentaire a été rajoutée pour étudier l'évolution de la consommation sans le poste hospitalisation.

Seniors	Poids du poste en 2014	Poids du poste en 2015	Poids du poste en 2016	Evolution de la consommation 2015/2014
Actes médicaux	10,7%	10,2%	9,2%	-0,3%
Autres prestations	19,6%	19,6%	19,2%	4,7%
Dentaire	6,4%	6,2%	5,6%	2,3%
Hospitalisation	29,1%	32,0%	36,0%	15%
Optique	4,4%	4,5%	5,1%	5,3%
Pharmacie	29,9%	27,5%	25,0%	-3,4%
Total hors hospitalisation	**71%**	**68%**	**64%**	**0,4%**
Total	**100%**	**100%**	**100%**	**4,6%**

Tableau 3 : Evolution de la consommation de soins des gammes de seniors entre les années 2014 et 2015

Pour les gammes de seniors, l'évolution de la consommation pour le poste hospitalisation est très élevée 15% ce qui peut s'expliquer par la nature des bénéficiaires. En effet, les seniors auront tendance à consommer davantage sur le poste hospitalisation en fin de vie. D'autre part, sur le poste dentaire leur consommation est moins élevée comparée à celle des actifs, l'évolution est de 2,3% comparée aux actifs à 6,3%. D'ailleurs, l'évolution globale de la consommation des actifs et des seniors est très disproportionnée: une évolution de 1,6% pour les actifs comparée à une évolution de 4,6% pour les seniors.

Le poids de chaque poste varie de moins de 1% entre les 3 années, on présentera donc dans la suite seulement le poids des postes de la première année, l'année 2014.

2. Calcul de la tendance

Dans un premier temps, nous allons calculer la dérive pure et ensuite nous calculerons la tendance de l'évolution de la consommation.
La dérive pure consiste à déflater l'effet âge à la dérive. L'estimation de l'effet âge est établie grâce à une courbe des âges qui est appliquée par module SCH (soins courants et hospitalisation), et OD (optique et dentaire). C'est pourquoi, le regroupement des consommations se fait en regroupant les postes par modules SCH et OD. Le module SCH est composé des postes : actes médicaux, autres prestations, hospitalisation, et pharmacie. Tandis que le module OD est composé des postes optique et dentaire.
La courbe des âges a par construction un âge de référence. Pour cet âge le coefficient OD ou SCH est égal à 1. C'est donc à partir de cet âge de référence que sont interprétés les autres coefficients pour les différents âges. Par exemple, pour le module SCH concernant une personne de 33 ans le coefficient est de 0,47 c'est à dire qu'une personne de 33 ans consomme 0,47 fois plus qu'une personne de 65 ans (âge de référence). Tandis qu'une personne de 84 ans consomme 1,84 fois plus qu'une personne de 65 ans.

L'effet âge est le taux de croissance entre deux âges. Par exemple, le taux de croissance à 34 ans pour le module SCH est l'évolution du coefficient SCH de la courbe des âges entre 35 et 34 ans. Ainsi sur le périmètre de notre étude, l'effet âge entre 2015 et 2014 est de 2,12% pour les actifs puisque l'âge moyen pour les actifs passe de 34 ans à 35 ans entre les années 2014 et 2015 (le portefeuille étant fermé, l'âge augmente seulement d'un an).

Ci-dessous, l'effet âge des actifs et des seniors :

		Effet âge	
		2015/2014	2016/2015
Actifs	SCH	2,12%	2,14%
	OD	1,89%	1,85%
Seniors	SCH	2,88%	2,98%
	OD	0,53%	0,49%

Tableau 4 : Effet âge des actifs et des seniors

La dérive pure de 2015/2014 est obtenue en calculant l'évolution de la consommation entre la consommation 2015 hors effet âge et la consommation de l'année 2014. La consommation 2015 hors effet âge correspond à la consommation 2015 ôtée de l'effet âge. Le tableau ci-dessous présente la dérive pure pour les actifs et les seniors.

	Actifs	Seniors
	Dérive pure 2015/2014	Dérive pure 2015/2014
SCH	-2,6%	1,8%
OD	4,2%	3,0%
TOTAL	-0,4%	1,9%

Tableau 5 : Dérive pure 2015/2014 pour les gammes d'actifs et seniors

La dérive pure des actifs est négative de -0,4% portée par la dérive négative du module SCH (-2,6%). Les questions à se poser sont: quelle tendance doit-on prolonger pour les actifs ? Ne doit-on pas appliquer une marge de prudence pour le module SCH face à cette tendance trop négative (-2,6%) ? Pour répondre à ces questions, il faut zoomer sur les différentes dérives du module SCH afin de connaître quel poste a une dérive trop faible.

Actifs		Dérive pure 2015/2014
SCH	Actes médicaux	-2,0%
	Autres prestations	2,1%
	Hospitalisation	**-7,9%**
	Pharmacie	-3,2%
TOTAL SCH		-2,6%

Tableau 6 : Dérive pure 2015/2014 du module SCH sur les gammes d'actifs

Nous remarquons qu'il s'agit du poste hospitalisation, sa dérive est trop faible -7,9% pour pouvoir être projetée. De plus, en regardant le baromètre santé de l'année précédente, nous constatons que l'évolution de la consommation pour ce poste n'est pas stable : l'évolution de la consommation 2014/2013 est de 8,0% pour atteindre -4,8% entre 2015/2014. Ces arguments nous amènent donc à modifier la dérive du poste hospitalisation pour le calcul de la tendance, ainsi nous mettons à zéro la

dérive de ce poste. La dérive du module SCH après modification est de -0,8%. Dans la suite, il faudra tenir compte de cette modification en prenant cette marge de prudence dans la tendance à projeter.
Le tableau ci –dessous présente la tendance modifiée pour les actifs :

Actifs	Dérive pure modifiée 2015/2014
SCH	-0,8%
OD	4,2%
TOTAL	0,8%

Tableau 7 : Tendance modifiée 2015/2014 des gammes d'actifs

La tendance à projeter permettant d'estimer les dérives des années postérieures est la dérive 2015/2014 présentée dans le tableau précédent pour les actifs (Tableau 7) et plus haut pour les seniors (Tableau 5).

3. Estimation de la dérive des années 2017 et 2018

a) Estimation de la dérive pour l'année 2017

Pour obtenir une estimation de la dérive 2017/2016 il faut projeter la tendance 2015/2014 en y ajoutant les estimations conjoncturelles de l'année 2017. La convention médicale impacte les prestations versées de l'année 2017 notamment avec la revalorisation de la consultation médicale. Cet impact est donc mesuré sur le portefeuille individuel en détaillant l'impact pour les actifs et les seniors. L'estimation de la consommation 2017 est ainsi obtenue en projetant la tendance 2015/2014 sur la consommation 2016 et en y ajoutant le coût de la convention médicale.

		2017
Impact de la convention médicale sur le module SCH	Actifs	0,60%
	Seniors	0,25%

Tableau 8 : Impact de la convention médicale en 2017 sur le module SCH

Des études ont été établies pour déterminer l'impact de la convention médicale sur l'ensemble des prestations du portefeuille individuel. Pour une meilleure précision, il faut donc calculer cet impact seulement sur le module SCH en distinguant les actifs et les seniors.
Le module SCH est impacté par la convention médicale de +0,6 points de dérive pour les actifs et de +0,25 points de dérive pour les seniors.
L'estimation de la consommation moyenne par bénéficiaire pour l'année 2017 sans correction des effets conjoncturels est obtenue en appliquant la tendance de la consommation 2015/2014 à la consommation 2016 :

$$Consommation_{2017}^{estimée} = Consommation_{2016} * (1 + tendance_{2015/2014})$$

A cette consommation estimée est additionnée l'estimation de l'impact de la convention médicale sur le module SCH. Le module OD n'est pas modifié puisque le règlement arbitral dentaire prévoit des impacts sur les prestations seulement à partir de l'année 2018.

Ainsi, est obtenue une dérive pure 2017/2016 de 1,3% pour les actifs et de 2,2% pour les seniors, représentée dans les tableaux ci-dessous :

Actifs	Tendance projetée	Coût convention médicale 2017	Dérive pure 2017/2016
SCH	-0,8%	0,60%	-0,2%
OD	4,2%		4,2%
TOTAL	**0,8%**	**0,5%**	**1,3%**

Tableau 9 : Estimation de la dérive 2017 des gammes d'actifs

Seniors	Tendance projetée	Coût convention médicale 2017	Dérive pure 2017/2016
SCH	1,8%	0,25%	2,1%
OD	3,0%		3,0%
TOTAL	**1,9%**	**0,2%**	**2,2%**

Tableau 10 : Estimation de la dérive 2017 des gammes de seniors

Ces résultats de dérives seront comparés dans la suite aux hypothèses de dérives d'organismes extérieurs comme la FNMF (Fédération nationale de la Mutualité française) ou le BIPE (Bureau d'information et de prévisions économiques).

b) Estimation de la dérive pour l'année 2018

La méthode d'estimation de la dérive de l'année 2018 est similaire à celle de l'année 2017. La tendance est appliquée sur la consommation estimée 2017 pour obtenir la consommation 2018 estimée. L'année 2018 sera marquée par les impacts de la convention médicale mais également du règlement arbitral dentaire. Il faudra donc prendre en compte ces effets.

Comme pour la convention médicale, l'impact du règlement arbitral dentaire est mesuré sur l'ensemble des prestations du portefeuille individuel. L'impact sur le module OD est calculé en divisant à l'impact global le poids des postes optique et dentaire.

$$Impact\ OD = \frac{impact\ global}{poids\ OD}$$

Le tableau ci-dessous présente les impacts réglementaires de l'année 2018 :

		2018
Impact de la convention médicale sur le module SCH	**Actifs**	1,12%
	Seniors	0,48%
Impact du règlement arbitral dentaire sur le module OD	**Actifs et seniors**	0,75%

Tableau 11 : Impact des effets conjoncturels pour l'année 2018

Ainsi, l'impact du règlement arbitral dentaire est de 0,75 points de dérive en 2018 pour les actifs et les seniors. L'impact de la convention médicale sur le module SCH pour l'année 2018 est calculé de la même façon que pour l'année 2017.L'impact obtenu est de 1,12 points de dérive pour les actifs et de 0,48 points de dérive pour les seniors.
La dérive pure 2018/2017 est ainsi obtenue en ajoutant ces impacts à la consommation 2018 estimée.
Les résultats d'estimation de dérive 2018/2017 sont présentés dans les tableaux ci-dessous :

Actifs	Tendance	Coûts conventions	Dérive pure 2018/2017
SCH	-0,8%	1,12%	0,3%
OD	4,2%	0,75%	5,0%
TOTAL	**0,8%**	**1,1%**	**1,9%**

Tableau 12 : Estimation de la dérive 2018 pour les gammes d'actifs

Seniors	Tendance	Coûts conventions	Dérive pure 2018/2017
SCH	1,8%	0,48%	2,3%
OD	3,0%	0,75%	3,7%
TOTAL	**1,9%**	**0,5%**	**2,5%**

Tableau 13 : Estimation de la dérive 2018 pour les gammes de seniors

Les coûts des conventions sont plus importants pour les actifs : 1,1% en 2018 comparé à 0,5% pour les seniors. Cependant l'estimation de la dérive est plus élevée chez les seniors 2,5% que pour les actifs 1,9%.

Au vu des résultats obtenus, on pourrait se demander s'il faudrait ajouter une marge de prudence à l'estimation des taux de dérives. Ainsi, la comparaison des dérives avec les organismes extérieurs peut être bénéfique.
Les hypothèses de dérive du BIPE et de la FNMF sont construites à partir des données de la CNAMTS (Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés), globalisées sur les portefeuilles collectifs et individuels. Les hypothèses de dérive de la FNMF pour l'année 2016 est de 2,1% et pour l'année 2018 de 2,6%. Pour l'année 2018, le 2,6% se décompose en 2,3% de dérive pour l'ONDAM et de 0,3% de dérive pour la FMF .Tandis que les hypothèses de dérives communiquées par le BIPE sont de 1,8% pour l'année 2016 et de 2,3% en 2017. On compare l'estimation de la dérive 2017 du BIPE à celle obtenue par l'étude menée.
Il faut donc détailler les hypothèses de dérives 2017 par poste du BIPE à la structure du portefeuille de l'étude. Dans un premier temps, les poids des postes médicaux de l'année 2016 du portefeuille sont calculés. Ensuite, sont appliquées les estimations du BIPE en points de dérive pour les différents postes afin d'obtenir une estimation de dérive adaptée à la structure du portefeuille de l'étude. Finalement, nous obtenons au global une dérive de 2,2% pour les actifs et de 2,1% pour les seniors. L'écart obtenu entre les dérives du BIPE appliquées à la structure de notre portefeuille et la dérive estimée de cette étude est minime. En effet, par exemple, pour les seniors la dérive obtenue en 2017 est de 2,2% tandis que celle estimée par le BIPE est de 2,1%. Par conséquent, on peut donc conclure que les dérives obtenues par application de tendance donnent une bonne estimation des dérives à horizon 2 ans.

Le tableau suivant est un récapitulatif des résultats obtenus :

	2015	2016	2017	2018
Gammes d'actifs	0,8%	Non mesurée en raison de modifications de garanties	1,3%	1,9%
Gammes de seniors	1,9%		2,2%	2,5%

Tableau 14 : Récapitulatif des dérives des années 2015 à 2018

III. Limite d'une approche macroéconomique

L'approche de l'étude précédente réalisée sur la dérive est une approche macroéconomique dans la mesure où les données sont agrégées. L'approche macroéconomique permet d'appréhender des effets globaux. Elle s'attache alors à déterminer les conditions qui réalisent des niveaux équilibrés entres les bénéficiaires d'une part, et les prestations d'autre part. Cette méthode d'analyse repose sur un certain nombre d'hypothèses : au niveau macroéconomique, il s'agit de regrouper les individus selon leur consommation dans les différents postes de santé, la consommation individuelle n'est donc pas prise en compte. Pour un acte médical donné, il s'agit d'agréger le volume des prestations ainsi que le nombre de personnes couvertes.

D'autre part, l'agrégation du nombre de bénéficiaires couverts dans l'étude précédente est scindée en deux niveaux : les gammes d'actifs et de seniors. Ainsi, on ne tient pas compte des différents âges. En effet, la dérive pure est estimée en déflatant un effet âge, cependant cet effet âge est calculé à partir d'un seul âge, âge moyen du portefeuille.

Dans l'étude précédente, le poste hospitalisation met en évidence la limite d'une approche macroéconomique. En effet, comme pour les actifs et seniors, la dérive du poste hospitalisation doit être étudiée séparément car elle représente une évolution différente par rapport aux autres postes.

De plus, en se concentrant sur les gammes de seniors, on remarque que leur consommation est très élevée, il se peut que cette croissance soit poussée par la consommation des bénéficiaires en dernière année de vie. Parallèlement, les bénéficiaires ont tendance à consommer davantage leur première année d'adhésion. Cependant l'approche macroéconomique ne permet pas de mettre en évidence ces effets.

En outre, la variabilité de la dérive entre les zones géographiques n'est pas prise en compte.

C'est pourquoi une approche individuelle est nécessaire pour capter tous les effets liés à la dérive, et ainsi mesurer une dérive précise.

L'approche individuelle s'intéresse à la structure de la consommation médicale en considérant les facteurs individuels comme l'âge, le sexe, le niveau de couverture. Ces facteurs varient selon le périmètre étudié. En effet, s'il s'agit du périmètre collectif, l'ajout de certaines variables comme la catégorie socioprofessionnelle, le type de régime doit être pris en compte.

La dépense des soins de santé varie selon l'âge et diffère selon le sexe. En effet, le niveau des dépenses pour les actes médicaux pour les femmes s'élève vers les âges de 20 ans (dépenses de gynécologie). De plus, les dépenses hospitalières augmentent significativement à l'âge de 60 ans. La variable de l'âge est donc à prendre en compte dans la dérive. Or, cette dernière n'est pas mise en évidence dans une étude macroéconomique. De même, pour le sexe, la structure des dépenses marque une différence entre les hommes et les femmes. A structure d'âge identique, les dépenses de soins totales ne marquent pas de différence mais selon les différents postes, il existe une distinction.

Partie 3 : Description et construction d'une base de données

Pour modéliser la dérive des soins de santé sur le portefeuille individuel, il est nécessaire de construire une base de données regroupant les consommations des soins de santé par postes et sous-postes (garanties) associées à chaque bénéficiaire. Cette partie a donc pour but d'expliquer comment les bases de données, qui vont être utilisées lors de la modélisation, ont été créées et de faire une étude descriptive.
Dans un premier temps, le périmètre à étudier doit être délimité pour pouvoir construire une base de données et choisir les différentes variables à prendre en compte. Puis, dans un second temps, une étude descriptive des variables est établie pour une meilleure compréhension lors de l'étape de modélisation.

I. Construction d'une base de données

Cette première partie sera consacrée au data management, c'est à dire le traitement de données permettant la production de données valides destinées à l'analyse statistique. La constitution de la base de données est la première étape de l'analyse. L'infocentre d'AG2R La Mondiale est constitué de nombreuses tables sous SAS contenant l'historique des informations sur les contrats, les bénéficiaires, les prestations versées, comprenant donc de nombreuses variables. Ces différentes bases vont permettre d'extraire des données et créer des variables d'analyse. Notre but est d'obtenir une table de données de consommation pour chaque bénéficiaire et pour des gammes du portefeuille individuel.

1. Périmètre de l'étude

Le périmètre de l'étude est un portefeuille fermé sur deux ans : années N-N+1. Un portefeuille fermé signifie que seulement les bénéficiaires présents deux années consécutives sont pris en compte, c'est à dire des bénéficiaires présents du 1er janvier de l'année N au 31 décembre de l'année N+1. Les années étudiées sont les années 2013-2014.
L'étude sera restreinte à deux gammes les plus récentes et encore commercialisées du portefeuille individuel : les gammes Santé Actif et Santé Senior. Le choix de ces gammes repose sur deux arguments : premièrement, elles représentent une part importante du portefeuille individuel et deuxièmement en combinant ces deux gammes l'ensemble des âges est rassemblé. En effet, ces gammes regroupent un ensemble d'individus ayant différents âges et non une restriction sur une moyenne d'âge précise. Ainsi, est capté l'ensemble des âges de 0 à plus de 100 ans.

- **Présentation du produit étudié :**

Le portefeuille individuel d'AG2R La Mondiale est composé d'une vingtaine de gammes. Ces différentes gammes recouvrent des individus de moyennes d'âges différentes. Ainsi, en combinant les gammes Santé Actif et Santé Senior, différentes moyennes d'âges sont prises en compte. En 2014, l'âge moyen de la gamme Santé Actif est de 37,8 ans et de 67,9 ans pour la gamme Santé Senior. A noter que l'adhésion à la gamme Santé Senior est une adhésion dès 55 ans et sans limite d'âge. Ces deux gammes offrent une couverture sur un ensemble de soins médicaux remboursés par la Sécurité Sociale mais également des soins non remboursés par la Sécurité Sociale. Le remboursement peut être exprimé en pourcentage de la BR ou en forfait, par exemple pour les frais hospitaliers un forfait de 18€/jour maximum est attribué.

Les produits Santé Actif et Santé Senior sont déclinés en plusieurs formules, c'est à dire que les bénéficiaires ont le choix entre plusieurs formules ou options. Pour chaque formule, un tableau de garantie est défini par l'assureur.
Ci-dessous un exemple de tableau de garantie pour un niveau de couverture de base (entrée de gamme) pour la gamme Santé Actif (avant la mise en conformité au contrat responsable) :

Hospitalisation médicale et chirurgicale	**Secteur conventionné**
Frais de séjour	100% BR
Actes de chirurgie, d'anesthésie, autres honoraires	100% BR
Hospitalisation à domicile	100% BR
Forfait hospitalier	18€/jour maxi
Chambre particulière	20€/jour
Frais d'accompagnement	Non pris en charge
Transport prescrit accepté par le RO	100%BR
Maternité	
Frais de séjour	100% BR
Autres honoraires	100% BR
Chambre particulière	20€/jour
Actes médicaux	
Généralistes	100% BR
Spécialistes	100% BR
Actes de chirurgie, actes techniques	100% BR
Actes d'imagerie médicale et d'échographie	100% BR
Analyses	100% BR
Auxiliaire médicaux	100% BR
Pharmacie	100% TFR*
Dentaire	
Soins dentaires	100% BR
Prothèses dentaires remboursées RO	100% BR
Orthodontie acceptée ou refusée RO	Non pris en charge
Parodontologie/Implantologie	Non pris en charge
Prothèses dentaires non remboursées RO	Non pris en charge
Orthopédie, Autres prothèses acceptées RO	
Prothèses, petits et grands appareillages	100% BR
Optique	
Monture/ Verre/ Lentilles acceptées, lentilles refusées y compris jetables	Année 1 : 100€ +RRO* Année 2 : 150€ +RRO Année 3 : 200€ +RRO
Chirurgie réfractive des yeux	100€/œil par an
Actes de prévention	
Détartrage annuel complet et sous-gingival, effectué en 2 séances maximum	100% BR
Vaccin diphtérie, tétanos et poliomyélite	100% BR
Cure thermale acceptée RO	
Traitement et honoraires	Non pris en charge
Voyage et hébergement	Non pris en charge
Bien-être	50€/an

* TFR : Tarif forfaitaire de responsabilité, RRO : Remboursement du régime obligatoire

Tableau 15 : Niveau de couverture de base pour la gamme Santé Actif

Le terme employé pour désigner le niveau de garantie est : « le niveau de couverture ».
Les différentes formules proposées par ces produits se différencient de part leur niveau de couverture, et de part leurs tarifs. Ces deux gammes offrent plusieurs niveaux de couverture selon les demandes des bénéficiaires. Ces niveaux de couverture sont modulables, c'est à dire qu'un bénéficiaire peut choisir une couverture de base pour les soins médicaux et avoir un niveau de couverture élevé pour les soins optiques et dentaires. En effet, un bénéficiaire peut souscrire à un niveau de base et améliorer ses remboursements en choisissant un renfort sur les postes hospitalisation et actes médicaux, il peut par exemple améliorer ses remboursements de spécialistes ou limiter ses frais d'hospitalisation.
Il existe également une garantie « Bien être » pour bénéficier d'un forfait en médecines douces, c'est-à-dire pour des consultations chez l'ostéopathe, le diététicien. Sur cette garantie « Bien- être », il est également possible d'avoir un renfort afin d'obtenir un forfait plus élevé.
Un niveau de couverture de base est proposé, il s'agit d'un remboursement de prestations à hauteur du remboursement du Régime Obligatoire. Il est possible d'obtenir des niveaux de couvertures plus élevés en choisissant des niveaux plus élevés sur les postes actes médicaux et optique-dentaire. C'est-à-dire en choisissant un niveau de garantie plus élevé sur le poste actes médicaux, on obtient également un niveau de couverture plus élevé sur le poste hospitalisation, transport et maternité.
Le niveau de couverture sur le poste hospitalisation est toujours de 100% du TFR, cependant le niveau de couverture pour la chambre particulière et les frais d'accompagnement varie selon le niveau de couverture choisi pour le poste actes médicaux. De plus, il est possible d'augmenter son niveau de couverture simultanément sur les postes optique et dentaire, c'est-à-dire en optant pour une élévation de couverture sur le poste optique, le niveau de couverture du poste dentaire est également réévalué à la hausse et inversement.
Ainsi les différents niveaux de couvertures sont présentés pour les postes actes médicaux, optiques et dentaires.

Les niveaux de couverture sont identifiables à partir du libellé de garantie. Le tableau ci-dessous présente quelques libellés d'option pour la gamme Santé Actif :

Libellé du niveau de couverture	Poste actes médicaux		Poste optique		Poste dentaire	
	Niveau	Remboursement	Niveau	Remboursement	Niveau	Remboursement
Modulo	1	100%BR	1	Année 1 : 100€ +RRO Année 2 : 150€ +RRO Année 3 : 200€ +RRO	1	100%BR
Modulo + ODA3	1	100%BR	4	Année 1 : 400€ +RRO Année 2 : 450€ +RRO Année 3 : 500€ +RRO	4	350%BR
Modulo + HAM1 + ODA2	2	150%BR	3	Année 1 : 500€ +RRO Année 2 : 550€ +RRO Année 3 : 600€ +RRO	3	250%BR

Tableau 16 : Exemple de libellé d'option pour la gamme Santé Actif

Le périmètre étant défini nous pouvons définir quelles variables prendre en compte.

2. Présentation des variables

Initialement, nous disposons de deux principales bases de données sur l'infocentre d'AG2R La Mondiale :

- Une base de données « bénéficiaires » regroupant les informations pour chaque bénéficiaire couvert par AG2R La Mondiale.
- Une base de données « prestations » détaillant l'historique de consommation de tous les assurés. Chaque ligne de la table des prestations représente un remboursement versé au bénéficiaire et contient les informations sur la date de survenance, la nature du soin et le montant remboursé par AG2R La Mondiale.

Nous avons eu besoin d'utiliser d'autres bases de données afin de récupérer toutes les informations nécessaires à notre étude mais celles-ci ne seront pas présentées.

a) Table des bénéficiaires

L'extraction des bénéficiaires dans le système d'information d'AG2R La Mondiale se fait en récupérant les numéros de contrat et les bénéficiaires associés. A chaque numéro de contrat individuel lui est associé un ou plusieurs bénéficiaires.

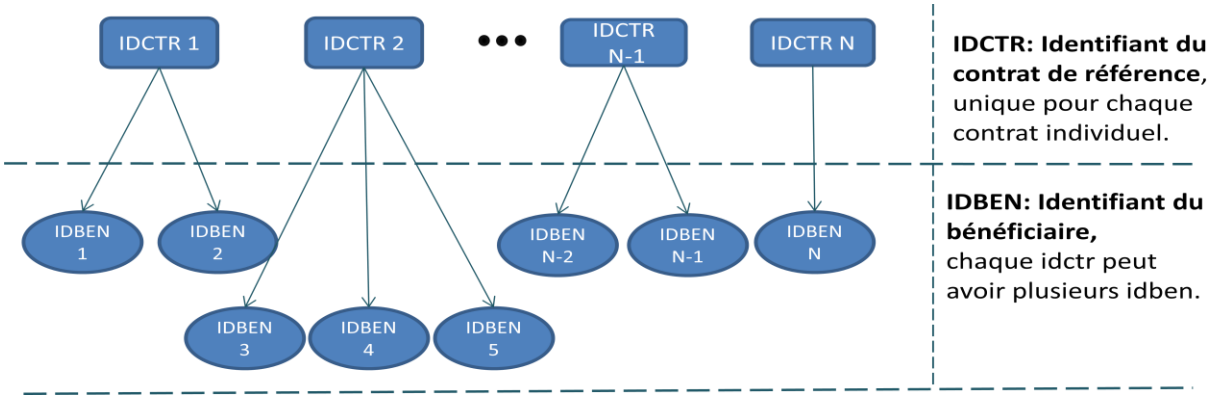

Figure 9 : Schéma des identifiants des bénéficiaires

La base de données des « bénéficiaires » sur l'infocentre nous permet de récupérer des informations sur les bénéficiaires. Celle-ci nous permet de construire une base de données des bénéficiaires propre au périmètre de l'étude. Les principales variables sont les suivantes :

- Le sexe du bénéficiaire ;
- Les catégories du type de bénéficiaires se distinguent selon trois types :
  AD : Adhérent, CJ : Conjoint et EN : enfant.
- La date de naissance du bénéficiaire ;
- La date d'adhésion du bénéficiaire. Cette date correspond à l'adhésion à un produit et non au contrat. Cela se traduit par le fait qu'un même individu peut avoir deux dates d'adhésion différentes s'il a changé de produit.
- La date de fin d'adhésion du bénéficiaire ;
- L'année et le mois de survenance du sinistre. Pour le risque santé, il s'agit de la date de demande de remboursement des soins de santé ;
- La variable durée indiquant pour chaque année combien de temps le bénéficiaire a été couvert par le produit associé (c'est à dire le temps de présence) ;
- Le nom de la gamme souscrit par le bénéficiaire, dont les gammes retenues sont les gammes Santé Actif et Santé Senior.

Dans un premier temps, nous avons sélectionné les bénéficiaires présents les années N et N+1 en supprimant les bénéficiaires dont le temps de présence est inférieur à zéro. De plus, nous avons calculé l'âge des bénéficiaires à partir de la date de naissance où l'on supprime les individus dont les âges sont inférieurs à zéro et supérieurs à 120 ans.

b) Table des prestations

L'étape suivante consiste à récupérer l'ensemble des prestations associées aux bénéficiaires. L'étude se base du point de vue de l'assureur, c'est à dire les prestations versées par AG2R La Mondiale. Les variables récupérées de la base de données des prestations sont celles présentées ci-dessous :

- L'identifiant du bénéficiaire ;
- Le numéro de contrat auquel il est rattaché ;
- Le produit souscrit ;
- Le poste et la garantie (appelé également sous-poste) auquel le bénéficiaire a consommé (la définition des postes est constituée à partir de l'identifiant de garantie) ;
- Le nombre d'actes consommés par poste ;
- La date de soin ;
- Le montant des frais réels en centimes d'euros ;
- Le montant des remboursements versé par AG2R La Mondiale en centimes d'euros.

Le traitement effectué sur cette base a consisté à récupérer les actes de consommation pour définir différents postes de soins présentés dans le tableau ci-dessous :

Postes	Garantie (Sous-poste)
Actes médicaux	Consultations et visites, actes de spécialités et médecines douces
Hospitalisation	Frais de séjour, honoraires, forfait hospitalier, chambre particulière et frais accompagnant
Pharmacie	Pharmacie, vaccins
Dentaire	Soins dentaires, prothèses dentaires, implants dentaires, orthodontie, parodontologie, inlay et onlay
Optique	Verres et monture, lentilles, chirurgie de la myopie
Autres prestations	Auxiliaires médicaux, analyses, actes imagerie, prothèses non dentaires, transport, cure thermale, maternité, frais obsèques

Tableau 17 : Définition des postes de soins

Le périmètre de l'étude est limité aux actes consommés entre le 1er janvier de l'année N et le 31 décembre de l'année N+1. De plus, un retraitement est effectué sur les montants remis en euros et non en centimes d'euros.

Les variables de la table des prestations sont de plusieurs types, les variables concernant les données du sinistre comme le montant des consommations et des variables permettant de relier les données de prestations à la base de données des « bénéficiaires » pour ainsi les fusionner, et obtenir une base de données spécifique au périmètre de l'étude.

c) Ajout de variables et création de variables d'intérêt

- **Ajout de variables :**

Les bases de données présentées précédemment ne permettent pas d'obtenir toutes les informations nécessaires afin de segmenter les bénéficiaires. Ainsi sont ajoutées plusieurs variables associées à chaque bénéficiaire :

- Le régime de couverture de l'assuré : on distingue deux régimes :
    - le Régime général et ;
    - le Régime d'Alsace-Moselle.

  Une distinction entre ces deux régimes est établie car leur niveau de remboursement est différent.
- Le département et la région : le département de l'assuré a pu être récupéré en fonction de la résidence de l'assuré. La formation des régions à partir des départements sera expliquée plus en détails par la suite.
- Le niveau de couverture : le remboursement versé par AG2R La Mondiale varie selon le niveau de couverture choisi par l'assuré. Ainsi l'évolution de consommation pour un bénéficiaire ayant un niveau de couverture élevé sur un poste sera différente pour un bénéficiaire ayant un niveau de couverture faible.

➢ **Création de variables d'intérêt :**

Afin de mesurer l'impact de certaines variables sur la dérive des soins de santé, il est nécessaire de créer des variables d'intérêt. Ainsi, sont créées deux variables d'intérêt :

- Une variable d'intérêt portant sur les affaires nouvelles ;
- Une variable d'intérêt concernant les décès.

La variable « affaire nouvelle » est créée pour permettre de déceler les individus sur le périmètre N-N+1 ayant souscrits à un produit l'année N (c'est à dire que ce sont des nouveaux arrivants). Ces individus peuvent avoir un comportement différent sur leur consommation de soins de santé. En effet, un assuré venant de souscrire aura tendance à consommer davantage la première année que la seconde. C'est donc avec cet indicateur que pourra être mesuré cet effet.

La variable est nommée « anN » pour affaire nouvelle de l'année N. Elle prend les modalités suivantes :

- valeur de 1 s'il s'agit d'une première année de souscription à un produit ;
- 0 dans le cas contraire.

De même, un bénéficiaire aura tendance à consommer davantage sa dernière année de vie, il faut donc ajouter une variable qui permet de distinguer les individus qui décèderont l'année N+2 et mesurer ce phénomène. Ainsi, nous récupérons les dates de décès associées à chaque bénéficiaire. Cependant les dates de décès ne sont disponibles que pour le bénéficiaire ouvrant le contrat c'est à dire l'adhérent. Or un contrat est constitué de plusieurs bénéficiaires : l'ouvrant droit, le conjoint et les enfants. L'impact de la dernière année de vie pourra donc être mesuré que sur certaines personnes et non sur l'ensemble des individus d'un contrat. **Nous pouvons donc obtenir un biais du fait que les dates de décès ne sont pas obtenues pour tous les bénéficiaires**. Il faudra donc en tenir compte dans la suite. La variable est nommée « décès-N+2 » qui prend la valeur 1 si l'assuré est décédé l'année N+2 et 0 dans le cas contraire.

Les bénéficiaires présentant les anomalies suivantes sont retirés de la base de données :

- Le département est inconnu ou mal renseigné ;
- Les individus ne relevant pas du Régime Général ou de l'Alsace-Moselle.

Après suppression des anomalies, le nombre de bénéficiaires sur le périmètre 2013-2014 est de 33 080.

d) Regroupement des niveaux de couverture

Les niveaux de couverture sont identifiables à partir du libellé de garantie. Seulement, il existe un nombre important de libellés, il faut donc les assembler pour un même niveau de couverture. Ainsi, un regroupement est établi en distinguant jusqu'à 4 à 5 niveaux de couvertures sur les différents postes.

Pour le poste pharmacie, le remboursement est toujours de 100% du TFR (Tarif de responsabilité). Cependant, le fait d'avoir un niveau de couverture élevé sur le poste actes médicaux, entraine les bénéficiaires à consulter plus fréquemment des généralistes et donc consommer davantage sur le poste pharmacie. Certains actes peuvent donc présenter des corrélations.
Pour le poste hospitalisation, la garantie « frais de séjour » est toujours remboursée à 100% de la BR ainsi que la garantie « Forfait Hospitalier » d'un forfait de 18€/jour maximum. Pour les autres garanties du poste hospitalisation, le niveau de couverture varie en fonction du choix des niveaux de couverture sur le poste « actes médicaux ». Ainsi, le remboursement de la garantie « actes de chirurgie, d'anesthésie, autres honoraires » est du même niveau de remboursement que pour les garanties du poste actes médicaux.

Pour le poste actes médicaux, le niveau de couverture est composé de 5 niveaux :

- Niveau 1 : Remboursement de 100% de la BR ;
- Niveau 2 : Remboursement de 150% de la BR ;
- Niveau 3 : Remboursement de 200% de la BR ;
- Niveau 4 : Remboursement de 300% de la BR ;
- Niveau 5 : Remboursement de 400% de la BR.

Pour le poste optique, le niveau de couverture est réparti en 5 niveaux, exprimé en forfait :

- Niveau 0 : Non pris en charge ;
- Niveau 1 :

Pour la gamme Santé Senior:
- Année 1 : 100€ + RRO ;
- Année 2 : 125€ + RRO ;
- Année 3 : 150€ + RRO ;

Pour la gamme Santé Actif :
Année 1 : 100€ + RRO
Année 2 : 150€ + RRO
Année 3 : 200€ + RRO

- Niveau 2 :

Pour la gamme Santé Senior:
- Année 1 : 150€ + RRO ;
- Année 2 : 200€ + RRO ;
- Année 3 : 250€ + RRO ;

Pour la gamme Santé Actif :
Année 1 : 200€ + RRO
Année 2 : 250€ + RRO
Année 3 : 300€ + RRO

- Niveau 3 :

Pour la gamme Santé Senior:
- Année 1 : 250€ + RRO ;
- Année 2 : 300€ + RRO ;
- Année 3 : 350€ + RRO ;

Pour la gamme Santé Actif :
Année 1 : 300€ + RRO
Année 2 : 350€ + RRO
Année 3 : 400€ + RRO

- Niveau 4 :

Pour la gamme Santé Senior:
- Année 1 : 350€ + RRO ;
- Année 2 : 400€ + RRO ;
- Année 3 : 450€ + RRO ;

Pour la gamme Santé Actif :
Année 1 : 400€ + RRO
Année 2 : 450€ + RRO
Année 3 : 500€ + RRO

- Niveau 5 :

Pour la gamme Santé Senior:
- Année 1 : 400€ + RRO ;
- Année 2 : 450€ + RRO ;
- Année 3 : 500€ + RRO ;

Pour la gamme Santé Actif :
Année 1 : 500€ + RRO
Année 2 : 550€ + RRO
Année 3 : 600€ + RRO

Sur le poste dentaire, le remboursement des soins dentaires est toujours de 100% de la BR sauf pour le niveau 4 de la gamme Santé Actif qui est de 200% de la BR. Les différents niveaux de couverture s'appliquent donc pour les garanties « Inlay/Onlay « et « les prothèses dentaires ».
Pour le poste dentaire, le niveau de couverture est réparti en 5 niveaux :

- Niveau 1 : Niveau de base : remboursement de 100% de la BR ;
- Niveau 2 : Remboursement de 150% de la BR pour la gamme Santé Senior et 200% de la BR pour la gamme Santé Actif ;
- Niveau 3 : Remboursement de 200% de la BR pour la gamme Santé Senior de 250% de la BR pour la gamme Santé Actif ;
- Niveau 4 : Remboursement de 300% de la BR pour la gamme Santé Senior et de 350% de la BR pour la gamme Santé Actif ;
- Niveau 5 : Remboursement de 400% de la BR pour la gamme Santé Senior et de 450% de la BR pour la gamme Santé Actif.

e) Création des régions

L'information de la zone géographique des bénéficiaires est le département. L'étude de la zone géographique sera restreinte à la métropole. Cette donnée qualitative doit être regroupée. En effet, le portefeuille présente des faibles volumes dans certaines zones, les bénéficiaires ne sont pas uniformément répartis sur la métropole. Il est donc nécessaire d'effectuer un premier regroupement des départements par région. Pour ce faire, on se base sur les 22 régions de la métropole.

Figure 10 : Carte de France avec les différents départements

Les régions sont ainsi créées en regroupant plusieurs départements comme par exemple la région Ile-de-France est constituée de 8 départements : 75, 77, 78, 91, 92, 93, 94 et 95. La région Picardie regroupe les départements : 02, 60 et 80.

Ainsi nous obtenons la carte des régions de France avec 22 régions en France métropolitaine.

Figure 11 : Carte de France en regroupant les départements en 22 régions

f) Ajout des données en open data

- **Les données ajoutées :**

Des données externes peuvent permettre d'obtenir une meilleure mesure de la dérive. En effet, les variables présentées précédemment (données présentes dans l'infocentre) sont seulement des données spécifiques aux bénéficiaires et prestations. L'ajout de données en open data enrichit donc la base de données.
Les données en open data nécessitent des retraitements pour les associer aux données dans la base créée.
Les données ajoutées sont les suivantes :

- données météorologiques ;
- données épidémiologiques.

La météo peut avoir un impact sur la consommation des soins de santé. En effet, les agents causatifs de maladie sont extrêmement sensibles à la température. De plus, les changements climatiques en France peuvent avoir un impact sur des pathologies associées à certains aléas climatiques, et donc une consommation de soins de santé. Ces pathologies sont par exemple les états de stress post aléas exceptionnels : inondations. De même, l'augmentation d'ensoleillement estival incite la population à rester plus longtemps à l'extérieur et, par ce fait, celle-ci est directement exposée aux rayonnements solaires, ce qui peut provoquer des pathologies comme les cancers cutanés. Les vagues de chaleur sont également des phénomènes météorologiques pouvant provoquer une augmentation de la consommation de soins de santé. De plus, les vagues de chaleur conduisent à des niveaux élevés de pollution, ces conditions contribuent à l'augmentation de pathologie respiratoire et donc de consommation de soins. De même, en période de faibles températures, un changement dans la consommation de soins de santé peut être observé.
Ces modifications climatiques peuvent donc avoir des conséquences sur la dérive.
De plus, les épidémies peuvent avoir un impact sur la consommation de certains postes. En effet, une épidémie de grippe une année donnée aura comme conséquence une augmentation de consommation sur les garanties « Consultations et visites », « Pharmacie » ou sur le poste « Hospitalisation ».

- **Le retraitement des données :**

Les données météorologiques sont récupérées du site *Météo France.* Ces données nécessitent un retraitement. En effet, les informations récoltées sont des informations mensuelles pour une année mais uniquement pour certains départements de France. De plus, la zone géographique retenue dans la base de données créée est la région. Les données météorologiques sont donc assemblées par région. Dans un premier temps, la création de région est établie en faisant la moyenne des données météorologiques des départements pour chaque région. Dans un second temps, pour les régions dont les informations météorologiques ne sont pas renseignées une moyenne des données des plus proches régions est attribuée. Enfin, l'information temporelle dans la base de données créée est l'année donc les données mensuelles sont transformées en annuelles.

Les données épidémiologiques proviennent de la source suivante : *Réseau Sentinelles.* Les données récupérées sont des données hebdomadaires, il faut donc les transformer en données mensuelles et annuelles. De plus, comme pour les données météorologiques, les informations ne sont pas renseignées pour certaines régions.
Après retraitement de ces données, celles-ci sont ajoutées à la base de données en les fusionnant par la variable commune qui est la région.

II. Analyse descriptive des données

L'assureur procède à une segmentation de ses assurés par rapport à des variables jugées par la théorie et l'historique pertinents pour mesurer l'évolution de la consommation de ses bénéficiaires. Ces variables sont aussi appelées facteurs discriminants. Ces facteurs font intervenir les caractéristiques propres du bénéficiaire. Les prestations sont quant à elles réparties selon différents postes. Une étude statistique permet de définir les critères de sélection dans la modélisation, le regroupement de certaines modalités de variables et d'écarter si nécessaire certaines variables dont le volume de données est insuffisant.
Cette analyse descriptive est établie sur la base de données globale. Cependant la modélisation ne sera pas faite sur l'ensemble de la base de données. En effet, dans une étude de recherche et d'une modélisation « tête par tête » on se concentrera sur un certain poste de soins pour mieux observer les résultats.

1. Répartition hommes/femmes

En santé, la variable du sexe est retenue car la consommation de certains postes de soins médicaux comme par exemple les consultations, la pharmacie ou les analyses varie selon le sexe. En effet, les femmes consomment de manière générale plus de soins de santé que les hommes. Par exemple, une femme vers l'âge de 21 ans aura une consommation plus importante sur le poste « actes médicaux » pour la garantie : « Consultations des spécialistes » (exemple : les consultations gynécologiques) qu'un homme du même âge.

Le portefeuille de notre étude est équilibré au niveau de la séparation hommes/femmes.

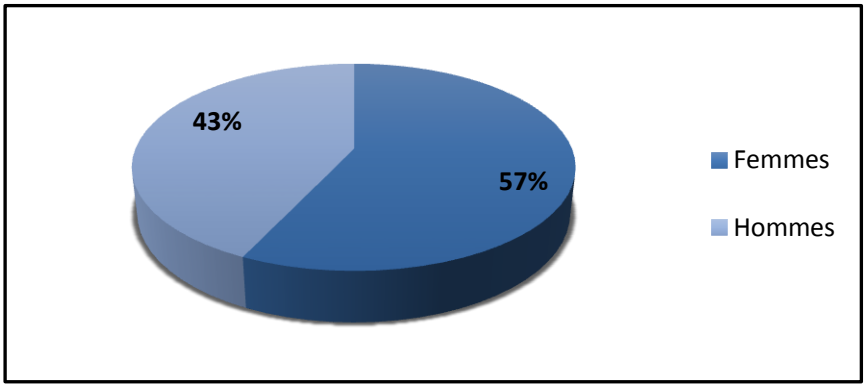

Figure 12 : Répartition des bénéficiaires selon le sexe

On distingue également trois types de bénéficiaires dans le rang familial : l'adhérent, le conjoint et l'enfant. On peut faire abstraction de cette variable dans la modélisation, car c'est l'âge et le sexe qui caractérisent le fait qu'un enfant a une consommation différente que l'adhérent.

2. Répartition par âge

L'âge est un facteur discriminant en santé, c'est le facteur le plus discriminant car plus une personne vieillit, plus sa santé se dégrade et plus elle est susceptible de consommer en soins médicaux. Ce facteur est d'ailleurs encore plus important qu'il ne l'était auparavant avec pour principales raisons l'augmentation de l'espérance de vie et l'amélioration des techniques et services médicaux. D'ailleurs, l'augmentation de l'espérance de vie accroit la durée des séjours d'hospitalisation.

L'âge est une variable continue, la création de classes d'âges (étape de discrétisation) permettra dans la suite d'assembler des individus ayant les mêmes caractéristiques en terme de consommation de soins de santé. De plus, la formation de classes d'âge permet une meilleure visibilité et fait disparaître les valeurs extrêmes. En effet, les valeurs extrêmes sont affectées à la première et dernière classe, par exemple des individus âgés de 100 ans seront affectés à la tranche d'âge des 75 ans et plus.
La discrétisation n'a pas toujours recours à une méthode statistique, parfois à des critères métiers, c'est sur ce critère que seront créées les classes d'âge. En effet, la création des classes d'âges est effectuée au vu des observations des consommations de soins de santé et confirmée par la courbe des âges utilisée pour tarifier les produits du portefeuille santé d'AG2R La Mondiale. Ainsi seront retenues 7 tranches d'âges :

- 1ère classe d'âge : les bénéficiaires de moins de 16 ans ;
- 2ème classe d'âge : les bénéficiaires entre 17 et 24 ans ;
- 3ème classe d'âge : les bénéficiaires entre 25 ans et 35 ans ;
- 4ème classe d'âge : les bénéficiaires entre 36 ans et 44 ans ;
- 5ème classe d'âge : les bénéficiaires entre 45 ans et 58 ans ;
- 6ème classe d'âge : les bénéficiaires entre 59 ans et 74 ans ;
- 7ème classe d'âge : les bénéficiaires de plus de 75 ans.

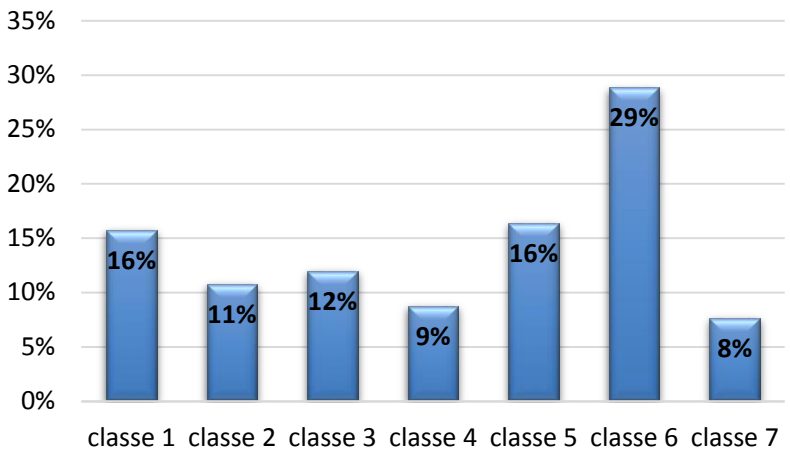

Figure 13 : Répartition des bénéficiaires selon les classes d'âge

La figure précédente montre que les bénéficiaires entre 59 ans et 74 ans sont les plus nombreux. De plus, cette figure met en avant la répartition uniforme entre les différentes classes d'âge.
L'âge a une importance sur les consommations médicales, on pourrait dire que l'effet âge est prédominant sur l'évolution de la consommation des soins.

3. Répartition par niveau de couverture

Une personne bénéficiant d'un niveau de couverture élevé aura tendance à plus consommer qu'une personne ayant un niveau de couverture de base. Ainsi, sur certaines garanties le montant remboursé par la complémentaire santé sera élevé pour les bénéficiaires ayant un haut niveau de couverture. La répartition du niveau de couverture est présentée pour un poste de soins en particulier : les actes médicaux.

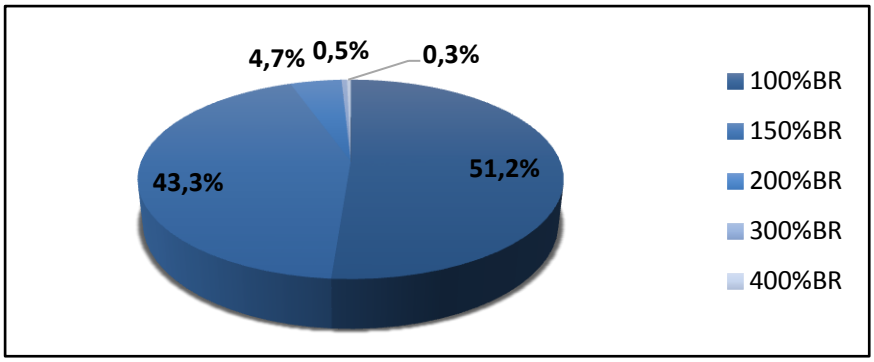

Figure 14 : Répartition des niveaux de couverture pour le poste actes médicaux

Le niveau de couverture pour le poste actes médicaux est composé de 5 niveaux de couverture. L'effectif étant trop faible pour les niveaux 4 et 5, un regroupement de ces niveaux est donc établi pour ces remboursements à hauteur de 300% et 400% de la BR.

4. Répartition par zone géographique

La zone géographique peut avoir une conséquence sur la consommation médicale.
On constate une densité d'offres de soins différente entre les régions. Cela peut s'expliquer par une différence de mode de vie, ainsi, il y a plus de soins techniques dans les grandes villes.
Par conséquent, l'évolution de la consommation sera distincte en comparant la région d'Ile de France et la Bretagne. Il faut noter également que l'évolution peut être différente entre les régions car la complémentaire santé d'AG2R La Mondiale commercialise plus ses produits dans certaines régions.

Régions	Proportion des assurés par région	Proportion de la population française par région [3]
Alsace	2,7%	2,9%
Aquitaine	5,6%	5,2%
Auvergne	1,3%	2,1%
Basse-Normandie	2,2%	2,3%
Bourgogne	3,1%	2,6%
Bretagne	5,7%	5,1%
Centre	4,6%	4,0%
Champagne-Ardenne	3,6%	2,1%
Corse	**0,4%**	**0,5%**
Franche-Comté	1,1%	1,8%
Haute-Normandie	3,2%	2,9%
Ile-de-France	**16,7%**	18,8%
Languedoc-Roussillon	2,3%	4,3%
Limousin	1,5%	1,2%
Lorraine	3,5%	3,7%
Midi-Pyrénées	2,8%	4,6%
Nord-Pas-de-Calais	3,9%	6,4%
Pays-De-La-Loire	5,7%	5,7%
Picardie	6,2%	3,0%
Poitou-Charentes	4,1%	2,8%
Provence-Alpes-Côte-D'azur	**13,2%**	**7,8%**

Tableau 18 : Proportion des bénéficiaires par région

Les différences de proportion de bénéficiaires par région s'expliquent par le fait que la population française n'est pas répartie uniformément sur le territoire.
Dans la base de données étudiée, les régions ayant une proportion de bénéficiaires les plus importantes sont l'Ile de France et la région PACA (Provence-Alpes-Côte D'Azur). Tandis que la Corse est la région la moins représentée.
D'autre part, en comparant les proportions de bénéficiaires de notre jeu de données à la proportion de la population française, on remarque que la région PACA représente un plus grand pourcentage de bénéficiaires sur le portefeuille étudié (13,2% bénéficiaires contre 7,8% proportion de la population française).

5. Répartition par régime

Les bénéficiaires sont répartis selon deux régimes :

- Le régime général et ;
- Le régime d'Alsace Moselle.

[3] Source : Insee, Recensement de la population française 2013.

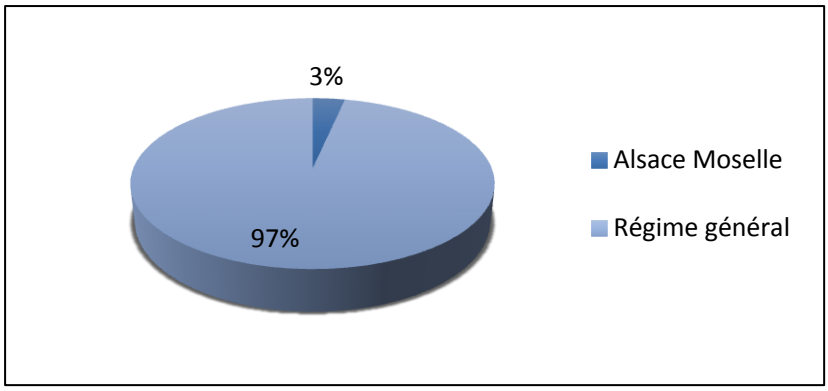

Figure 15 : Répartition des bénéficiaires par régime

Un très faible pourcentage des bénéficiaires sont couverts par le régime d'Alsace Moselle. La suite de l'étude sera donc restreinte au régime général. Le nombre de bénéficiaires étudié est donc de 31 953 en régime général.

6. Statistiques des prestations totales

Les prestations médicales sont réparties en six postes médicaux :

- Actes médicaux ;
- Autres prestations ;
- Hospitalisation ;
- Pharmacie ;
- Optique ;
- Dentaire.

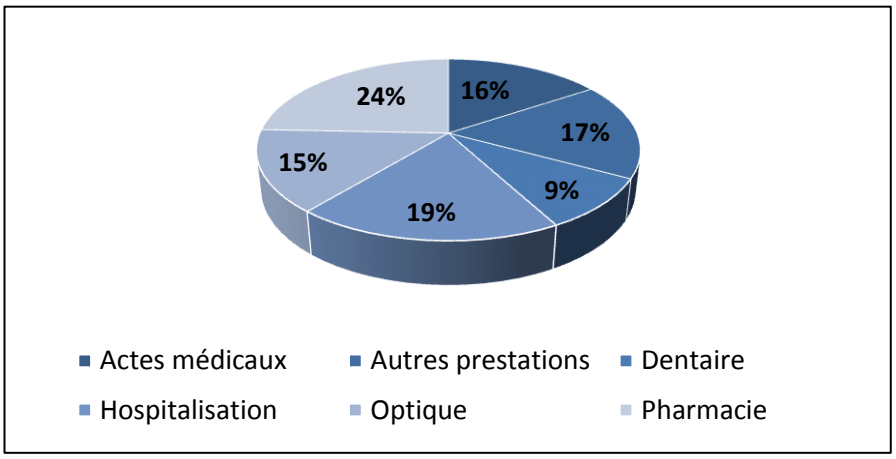

Figure 16 : Poids des dépenses médicales sur le portefeuille étudié

Le poids de ces postes est différent. Les postes majeurs sont les postes : pharmacie, hospitalisation et actes médicaux.
En effet, le poids de ces postes est différent du fait d'un besoin qui varie selon les âges.

- Les enfants ont des besoins en hospitalisation, actes médicaux et pharmacie. Mais, ils n'ont pas besoin de soins en optique et dentaire. En grandissant, leur besoin dans ces deux postes peut augmenter, cependant le coût de ces dépenses pour les organismes complémentaires n'est pas important. En effet, les soins dentaires représentent un faible coût pour les organismes complémentaires car il s'agit essentiellement de soins avec peu de dépassements d'honoraires.

Quant au poste optique, les corrections pour ces âges sont faibles et donc représentent un faible coût.

- Les adolescents ont les mêmes besoins, auxquels il faut ajouter les soins en orthodontie, soins généralement peu remboursés par la Sécurité Sociale. De même, pour l'optique, des corrections plus importantes peuvent apparaître à l'adolescence et donc un coût plus important pour les organismes complémentaires.
- Les femmes vers l'âge de 20-30 ans auront un besoin plus important dans les postes actes médicaux et hospitalisation. En effet, la visite chez des spécialistes aura tendance à augmenter les coûts des organismes complémentaires car ces soins présentent d'importants dépassements.
- A partir de 40 ans, les besoins des soins de santé sont essentiellement de l'optique et du dentaire. En effet, à ces âges, par exemple en dentaire, de nouveaux besoins de soins apparaissent qui sont plus coûteux comme les prothèses dentaires.
- Enfin pour les personnes âgées, les besoins des soins de santé reposent sur les postes pharmacie, hospitalisation et actes médicaux. Les besoins en optique et dentaire sont très faibles.

A travers cette description des besoins de santé le long de la vie d'un bénéficiaire, nous comprenons donc que les postes les plus importants sont la pharmacie, l'hospitalisation et les actes médicaux.

Comme vu précédemment, la démographie influe sur le niveau de consommation de soins. En effet :

- La consommation de soins augmente avec l'âge : une population plus âgée consommera plus qu'une population qui lui est semblable en tout point, mais plus jeune ;
- Les femmes ont tendance à plus consommer que les hommes ;
- Les habitants de régions urbanisées dans laquelle la densité de médecins est plus élevée auront également tendance à plus consommer, ou à consommer des soins plus onéreux.

D'une manière générale il est préférable d'avoir peu de facteurs, mais fiables, plutôt qu'un nombre important dont la connaissance est approximative.
Après avoir établie une description des variables constituant la base de données servant à la modélisation, l'étape suivante est donc la modélisation proprement dite.

Partie 4 : Modélisation statistique économétrique

Nous cherchons à modéliser la dérive des soins de santé. Deux approches de modélisation peuvent être utilisées :

- Modéliser la consommation des soins de santé et ensuite obtenir la dérive ;
- Modéliser l'évolution de la consommation des soins de santé qui est la dérive.

Ces deux approches utilisent une modélisation par un modèle linéaire généralisé (appelé GLM dans la suite).

Dans un premier temps, la théorie d'une modélisation GLM sera exposée. Dans un second temps, les retraitements et les résultats obtenus seront présentés. Enfin, sera abordée une discussion critique sur les limites d'une modélisation GLM.

I. Théorie du modèle linéaire généralisé

Les modèles linéaires généralisés, GLM, ont été introduits par John Nelder et Robert Wedderburn en 1972.
Dans cette partie, seront présentés les fondements théoriques du modèle linéaire généralisé. Avant de développer le modèle linéaire généralisé, quelques rappels sur le modèle linéaire seront introduits.

1. Rappel sur le modèle linéaire

Dans un modèle en régression linéaire, l'objectif de la modélisation est de prédire et d'expliquer par une relation de type linéaire, une variable dite expliquée (dépendante ou à expliquer) en fonction d'un ensemble de variables dites variables explicatives. Ce modèle peut s'écrire sous la forme suivante :

$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

Avec :

- Y la variable que l'on souhaite expliquer, appelée également variable à expliquer ou variable réponse ;
- X_j, la $j^{ème}$ variable explicative ;
- p, le nombre de variables explicatives ;
- β_j , les paramètres du modèle à estimer, ce sont les coefficients dont la valeur détermine l'impact de chaque variable explicative sur la variable Y ;
- ε, l'erreur du modèle. Il s'agit de l'écart entre la valeur réelle observée de Y et sa valeur théorique obtenue par le modèle. On pose les hypothèses suivantes sur la variable ε :

  $$\mathrm{E}(\varepsilon) = 0 \text{ et } \mathrm{Var}(\varepsilon) = \sigma^2$$

  où σ^2 est un paramètre inconnu, à estimer. On parle de modèle linéaire gaussien si nous ajoutons une hypothèse de normalité des résidus ε, c'est-à-dire que ε est distribué selon une loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 inconnu.
- Les hypothèses posées sur ε impliquent les caractéristiques suivantes de Y :

$$\mathrm{E(Y)} = \sum_{\mathrm{j=1}}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{X_j} \text{ et } \mathrm{Var(Y)} = \sigma^2$$

En moyenne Y s'écrit comme une combinaison linéaire des X_j, c'est pour cela que le modèle s'appelle modèle linéaire.

En conclusion, les paramètres à estimer dans le modèle sont β et σ^2.
L'estimation de ε est $\hat{\varepsilon} = Y - X\hat{\beta}$, où $\hat{\beta}$ est l'estimateur des moindres carrés ordinaires donné par l'expression de $\hat{\beta} = (X'X)^{-1}X'Y$.
L'estimation des paramètres est basée sur les n observations des individus et donc pour la i[ème] observation le modèle s'écrit :

$$y_i = \sum_{j=1}^{p} \beta_j \, x_i^j + \varepsilon_i$$

Avec :

- y_i, la variable explicative pour l'individu i $(i = 1 \text{ à } n)$;
- p variables explicatives $x^1, x^2, \dots, x^p$ prenant les valeurs $x_i^1, x_i^2, \dots, x_i^p$ pour l'individu i;
- ε_i sont identiquement distribués selon la loi $N(0, \sigma^2)$ avec σ^2 constante.

En notation matricielle le modèle s'écrit (comme indiqué plus haut) :

$$Y = \underbrace{X\beta}_{effet\ fixe} + \underbrace{\varepsilon}_{effet\ aléatoire}$$

Avec :

- Y, le vecteur colonne des n observations de la variable expliquée ;
- X, la matrice des observations des p vecteurs X_i de dimension (n, p);
- β, le vecteur colonne des p coefficients de régression ;
- ε , le vecteur des erreurs.

2. Le modèle linéaire généralisé

Le modèle linéaire gaussien impose le respect d'hypothèses, par exemple une condition sur la normalité de la variable à expliquer Y. Dans certaines circonstances, cette contrainte n'est pas vérifiée (des tests de normalité peuvent être effectués pour vérifier si la variable Y satisfait la condition de normalité). On a donc recours au modèle linéaire généralisé qui est une extension du modèle linéaire permettant de modéliser les observations par une loi mieux adaptée.
L'objectif des modèles linéaires généralisés est donc de généraliser le modèle gaussien à un ensemble de lois.

Les modèles GLM sont caractérisés par trois composantes :

- La distribution de la variable à expliquer Y appartient à la famille exponentielle naturelle ;
- Le prédicteur linéaire ;
- La fonction de lien (notée g, fonction inversible) qui décrit la relation entre l'espérance de la variable réponse et les variables explicatives.

Le modèle GLM s'écrit donc :

- $Y \sim Loi_{exp}$
- $\mu = E(Y)$
- $\eta = g(\mu)\ avec\ \eta = X\beta$

Loi de la famille exponentielle :

Pour appliquer ces modèles, la loi de probabilité de la variable Y doit appartenir à la famille exponentielle naturelle qui contient entre autres des lois aussi usuelles que la loi normale, la loi de Bernoulli, la loi binomiale, la loi de Poisson, la loi Gamma. Ces lois ont une forme de densité commune sous forme exponentielle.

Une variable aléatoire Y appartient à la famille exponentielle si sa densité peut s'écrire sous la forme suivante :

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

Avec :

- θ, le paramètre naturel de la famille exponentielle ;
- ϕ, le paramètre de dispersion, un paramètre de nuisance ;
- c (.) une fonction continue et dérivable ;
- b (.) une fonction continue et trois fois dérivable : b'(.) est inversible c'est à dire que $(b')^{-1}$ existe.

En général, on a $a(\phi)$ au lieu de ϕ où $a(.)$ est une fonction continue et dérivable.

Dans un modèle GLM nous avons donc les propriétés suivantes pour Y :

- $E(Y) = \mu = b'(\theta)$
- $V(Y) = b''(\theta)\phi$

Le prédicteur linéaire :

Le prédicteur linéaire est défini par :

$$\eta = X\beta$$

où β est un vecteur de paramètres inconnus de taille p et X la matrice des variables explicatives de dimension $n \times p$.

Lien canonique :

Chacune des lois de la famille exponentielle possède une fonction spécifique, dite fonction de lien canonique. La fonction de lien est la fonction qui permet de lier les variables explicatives à la prédiction. Le lien canonique est telle que :

$$\eta = X\beta \text{ prédicteur linéaire, donc telle que } g(\mu) = \theta$$
$$g(\mu) = \theta, or\ \mu = b'(\theta)\ donc\ \ g(.) = b'(.)^{-1}$$

A noter qu'il existe d'autres fonctions de lien non canonique utilisées en pratique. Mais, si dans la modélisation aucune raison de choisir une fonction de lien spécifique ne s'impose, le choix par défaut consiste à choisir la fonction de lien naturel.

Le modèle linéaire généralisé part du même principe que le modèle linéaire gaussien, cependant il existe des différences. Dans un modèle GLM, on ne modélise pas directement la variable à expliquer, c'est une fonction de cette variable (appelée fonction de lien canonique) qui est modélisée. La variable à expliquer Y et son espérance ne se traduisent plus nécessairement par une transformation linéaire mais c'est $g(E(Y))$ qui est désormais exprimée par une relation linéaire ($\eta = X\beta\ prédicteur\ linéaire$) en fonction des variables explicatives.

Le choix de la distribution pour le modèle généralisé est déterminé par la nature des données, le type de la variable réponse Y. Pour une variable réponse binaire, la distribution peut être une loi de Bernoulli ou une loi binomiale, une loi de Poisson pour des comptages. Ce tableau dresse les principales lois appartenant à la famille exponentielle et leurs liens canoniques associés :

Lois	Fonction de lien canonique	Nom du lien
Bernoulli/Binomiale	$g(\mu) = logit$ $= log\left(\frac{\mu}{1-\mu}\right)$	Logit
Poisson	$g(\mu) = log(\mu)$	Log
Gamma	$g(\mu) = -\frac{1}{\mu}$	Réciproque
Gaussienne	$g(\mu) = \mu$	Identité

Tableau 19 : Les GLM usuels

En assurance, la fonction de lien la plus souvent utilisée est la fonction log. Les coefficients obtenus lors de la modélisation sont alors multiplicatifs et permettent une meilleure lisibilité. En effet, ces coefficients peuvent alors être interprétés facilement comme des taux de majoration ou de minoration par rapport à un individu de référence.

Les modèles linéaires généralisés permettent donc d'étendre la méthode de régression linéaire à un ensemble de lois plus larges et correspondant plus à la sinistralité réelle couverte par l'assureur. De plus, contrairement aux régressions linéaires, il est désormais possible de traiter des variables catégorielles.

a) Lois usuelles de type exponentiel

Les lois usuelles appartenant à la famille exponentielle sont :

- La loi Normale ;
- La loi Gamma ;
- La loi Exponentielle ;
- La loi de Poisson ;
- La loi Binomiale ;
- La loi de Tweedie.

<u>Loi Normale :</u>

Utiliser une loi normale dans une modélisation GLM correspond en fait à une régression linéaire. Soit Y une variable aléatoire suivant une loi normale d'espérance μ et de variance σ^2, sa densité s'écrit donc :

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

La loi normale $\mathcal{N}(\mu, \sigma^2)$ appartient donc à la famille exponentielle avec les paramètres suivants:

- $\theta = \sigma^2$;
- $\phi = \sigma^2$;
- $b(\theta) = \frac{\sigma^2}{2}$;
- $c(y, \theta) = \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$

La densité sous forme exponentielle s'écrit donc :

$$f(y) = \exp\left(\frac{y\mu - \frac{\sigma^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right)$$

Loi log-normale :

La loi log-normale est une des lois la plus utilisée en assurance pour la modélisation des coûts. C'est une loi dérivée de la loi normale, avec la propriété suivante :

Y suit une loi log-normale si et seulement si $log(Y)$ suit une loi normale.

La densité d'une loi log-normale de paramètre μ et σ^2 est :

$$f(y) = \frac{1}{y * \sqrt{2\pi\sigma^2}} * \exp\left(\frac{-(\log(y) - \mu)^2}{2\sigma^2}\right)$$

Cette loi n'appartient pas à la famille de lois exponentielles, il n'est donc pas possible de lui appliquer un GLM. Cependant, en posant $X = \log(Y)$, la nouvelle variable X ainsi créée suit alors une loi normale d'espérance μ et de variance σ^2.

Les notions fondamentales d'un GLM et les différentes composantes qui interagissent dans ce modèle ont été présentées plus haut. Il faut à présent présenter les méthodes d'estimations des variables explicatives.

b) Estimations des paramètres

Les estimations des paramètres se fait par maximum de vraisemblance ou plutôt en maximisant la log-vraisemblance. En effet, la famille de distribution des GLM ont une structure exponentielle, il est donc plus aisé de maximiser la log-vraisemblance.

Soit Y la variable à expliquer dont les observations sont rangées dans le vecteur Y. Considérons p variables explicatives dont les observations sont rangées dans la matrice design X, $X = (X_1, \dots , X_P)$, β un vecteur de p paramètres et g la fonction de lien (fonction de lien non canonique).

Pour n observations supposées indépendantes, la log-vraisemblance s'écrit :

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \ln f(y_i\,; \theta_i\,;\, \phi) = \sum_{i=1}^{n} \mathcal{L}_i$$

où

$$\mathcal{L}_i = \frac{y_i\theta_i \, - b(\theta_i)}{\phi} + c(y_i, \phi)$$

Ce procédé implique de définir trois notations déjà vues précédemment :

- μ_i : l'espérance conditionnelle de Y par rapport à la i^{ème} observation ;
- θ_i : la valeur du paramètre θ pour la i^{ème} observation ;
- η_i : le prédicteur linéaire de la i^{ème} observation.

Calculons

$$\frac{\partial \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \mathcal{L}_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

Comme :

$\frac{\partial \mathcal{L}_i}{\partial \theta_i} = \frac{y_i - \mu_i}{\phi}$

$\frac{\partial \theta_i}{\partial \mu_i} = \left[\frac{\partial \mu_i}{\partial \theta_i}\right]^{-1} = \left[b^{"}(\theta_i)\right]^{-1} = \frac{Var(y_i)}{\phi}$ car $Var(y_i) = b^{"}(\theta_i)\phi$

$\frac{\partial \mu_i}{\partial \eta_i}$ dépend de la fonction de lien η_i

$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

Les équations de vraisemblance sont donc pour $j = 1, \dots, p$:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

Dans le cas où le lien canonique est utilisé, plusieurs simplifications interviennent (se référer à l'Annexe A : Compléments sur les GLM) ainsi les équations de vraisemblance s'écrivent pour $j = 1, \dots, p$:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi} x_{ij} = 0$$

Ce sont des équations non linéaires en β dont la résolution requiert des méthodes itératives telles que l'algorithme de Newton-Raphson et IRLS (Iterative Reweighted Least squares). Pour plus de détails sur ces algorithmes se référer à l'Annexe A : Compléments sur les GLM.
Deux questions peuvent se poser en étudiant plusieurs modèles. Comment sélectionner le « bon » modèle et comment le valider ?

c) Sélection du modèle

Dans un bon nombre d'études statistiques, les variables potentiellement explicatives ne sont pas forcément des variables pertinentes, c'est à dire pertinentes pour expliquer la variable Y.
Dans un GLM, deux composantes permettent d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations. Il s'agit de la déviance et de la statistique ou test de Pearson. Dans cette partie sera présentée seulement la déviance.

Déviance :
Le modèle estimé est comparé au modèle dit saturé, c'est à dire le modèle possédant autant de paramètres que d'observations. Notons que le modèle saturé est le modèle le plus complexe, complet. Donc, si un modèle plus simple a une vraisemblance proche de la vraisemblance du modèle saturé, on le préférera.

La déviance D vaut :

$$D = 2(\mathcal{L}_{sat} - \mathcal{L}),$$

où $\mathcal{L}$ et $\mathcal{L}_{sat}$ sont respectivement les log-vraisemblance dans le modèle et dans le modèle saturé.

Il apparaît clairement que plus la déviance est grande moins le modèle est bon car le modèle est éloigné de la réalité et donc non adapté. Mais, à partir de quel niveau peut-on juger que le modèle est adéquat ? Il est possible d'établir un seuil de significativité à l'aide de loi de Khi 2 afin de répondre à ce problème. Asymptotiquement, D suit une loi de χ^2 (khi 2) à $n - p$ degrés de liberté, ce qui permet de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.
Se comparer au modèle saturé n'est pas toujours utile, il est parfois judicieux de comparer deux modèles emboités $\mathcal{M}_1$ et $\mathcal{M}_2$ respectivement avec p_1 et p_2 variables explicatives ($p_1 > p_2$) donc $\mathcal{M}_2 \subset \mathcal{M}_1$. La différence de déviance est un test permettant de choisir entre les deux modèles.

d) Robustesse du modèle

Les modèles possibles ne sont pas forcément emboités, l'utilisation du test de la déviance a donc ses limites. En effet, pour modéliser la dérive, différents GLM peuvent être réalisés avec des paramétrages différents (utilisation d'autres variables explicatives, changement de fonction de lien). D'autres critères permettent de comparer des modèles qui ne sont pas emboités les uns dans les autres. Plus la vraisemblance (ou son log) d'un modèle est grand, meilleur est le modèle. Mais la valeur élevée de la vraisemblance (ou son log) n'est pas suffisante pour faire un choix. L'inconvénient est que la vraisemblance (ou son log) d'un modèle augmente avec la complexité p du modèle, c'est à dire avec l'ajout de variables explicatives. Sélectionner le modèle qui maximise la vraisemblance (ou son log) revient à choisir le modèle saturé. Mais si l'on sélectionne le modèle qui a le plus de variables explicatives, certes celui-ci est plus précis, mais l'inconvénient est qu'il est moins robuste. Plusieurs critères existent pour parer à ce problème.
Deux critères seront exposés, mais il en existe bien d'autres. Les critères de sélection introduits sont : le critère AIC et le critère BIC. Ces critères maximisent la vraisemblance tout en pénalisant les grands modèles.
Le critère AIC (critère d'information d'Akaike) basé sur l'information de Kullback-Leibler est une mesure de la qualité d'un modèle. Il contient un facteur pénalisant en fonction du nombre de paramètres.

L'expression du critère AIC est définie pour un modèle à p paramètres par :

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

Le critère BIC (critère de Schwarz) qui se place dans un contexte bayésien de sélection de modèle, avec une approche basée sur le facteur de Bayes, est défini pour un modèle à p paramètres et n observations par :

$$BIC = -2\mathcal{L} + p * log(n)$$

On choisira le modèle qui possède le plus petit critère. Notons que certains logiciels utilisent les critères –AIC et –BIC; pour ceux-là, on choisira le modèle qui a le critère le plus grand.

e) Validation du modèle

Une fois le modèle choisi, il faut le valider d'un point de vue global par des tests d'adéquation par exemple. L'ajustement peut se faire variable par variable. De plus, il faut vérifier la cohérence d'un modèle en procédant à l'analyse des résidus. Plusieurs types de résidus peuvent être définis. L'examen des résidus est primordial pour valider un modèle. Cette analyse peut permettre d'une part de vérifier la qualité d'ajustement, et d'autre part de mettre en évidence des valeurs aberrantes.

- ***Résidus bruts***

Les résidus bruts sont définis pour tout $i = 1, \dots, n$ par :

$$\hat{\xi}_i = Y_i - \hat{Y}_i$$

La définition la plus naturelle, consiste à quantifier l'écart entre l'observation Y_i et sa prédiction $\hat{Y}_i$ par le modèle. Ils permettent de quantifier l'ajustement du modèle observation par observation. Mais, ils sont difficiles à comparer car ils n'ont pas toujours la même variance. Il est donc difficile de les comparer à un comportement type attendu. C'est pourquoi sont introduits les *résidus de Pearson*.

- ***Résidus de Pearson***

Les résidus de Pearson sont les résidus bruts normalisés par une variance estimée définis par :

$$r_{p_i} = \frac{Y_i - \hat{Y}_i}{\sqrt{V(Y_i)}}$$

où $V(Y_i)$ désigne la variance théorique de Y_i .
Cependant pour les *résidus de Pearson*, leur variance dépend de l'observation i.

- ***Résidus de Pearson standardisés***

Les résidus de Pearson standardisés sont obtenus en normalisant les résidus de Pearson par l'effet levier :

$$r_{s_i} = \frac{Y_i - \hat{Y}_i}{\sqrt{(1 - h_{ii})V(Y_i)}}$$

où h_{ii}, désigne le levier, c'est-à-dire le terme diagonal de la matrice $H = X(X'X)^{-1}X'$ dans le cas où la matrice de design X est de rang plein.

Une autre approche consiste à définir les *résidus de déviance*.

- ***Résidus de déviance***

Les résidus de déviance mesurent à quel point la log-vraisemblance pour l'observation i est loin de la log-vraisemblance pour cette même observation dans le cas du modèle saturé. Ils sont définis par :

$$r_{d_i} = signe(Y_i - \hat{Y}_i)\sqrt{2(\mathcal{L}_{sat} - \mathcal{L})}$$

Pour rendre ces résidus comparables entre eux, il faut les corriger pour prendre en compte l'influence de chaque observation, ainsi les *résidus de déviance standardisés* en tiennent compte.

- ***Résidus de déviance standardisés***

Les résidus de déviance standardisés sont définis par :

$$r_{d_i} = signe(Y_i - \hat{Y}_i)\sqrt{\frac{2(\mathcal{L}_{sat} - \mathcal{L})}{1 - h_{ii}}}$$

Intuitivement, une observation ayant un résidu de déviance élevé est une observation ayant une grande influence sur l'estimation des paramètres du modèle et doit donc être examinée.
Les aspects théoriques d'un modèle GLM reposent sur des hypothèses fortes. Il est donc nécessaire de vérifier la véracité de ces hypothèses. Les analyses graphiques jouent un grand rôle dans la validation des hypothèses notamment l'analyse des résidus.

Analyse de l'hypothèse de normalité :

Le QQ-plot est un outil graphique permettant de visualiser l'adéquation d'une série numérique à une distribution théorique de référence. En reportant sur l'axe des ordonnées les fractiles correspondant à la distribution observée $\hat{\xi}$ et sur l'axe des abscisses ceux correspondant à la distribution théorique $\xi \sim \mathcal{N}(0,1)$, sont obtenus des points alignés sur la première bissectrice. L'hypothèse de normalité des résidus est alors vérifiée.

Analyse de l'homoscédasticité:

Afin de vérifier l'homoscédasticité, un graphique prédiction linéaire en fonction des résidus est réalisé.

II. Résultats

Comme expliqué plus haut, dans l'introduction de la partie 4, deux approches sont testées pour modéliser la dérive. Une première approche consiste à modéliser la consommation des soins de santé et ensuite obtenir la dérive. Une seconde approche consiste à modéliser le taux d'évolution. De plus, dans un esprit de recherche, la modélisation de la dérive est établie sur un poste de soins de santé en particulier. En fait, il s'agit de sélectionner la garantie ayant le plus grand poids au sein d'un même poste et de modéliser sa dérive. Ainsi pour le poste « actes médicaux », la garantie « Consultations et visites » est sélectionnée. **La modélisation de la dérive sera restreinte au Régime Général.**
On se concentre sur la garantie « Consultations et visites ». Une analyse descriptive de celle-ci est présentée pour permettre une meilleure connaissance des variables.

Analyse descriptive de la garantie « Consultations et visites » :

Paramètres		Poids
Sexe	**Femme**	57%
	Homme	43%
Classe d'âge	**1**	15,2%
	2	10,3%
	3	12,3%
	4	8,7%
	5	16,1%
	6	29,3%
	7	8,1%
Gamme	**Santé Actif**	57%
	Santé Senior	43%
Niveau de couverture	**Niveau 1**	52%
	Niveau 2	43%
	Niveau 3	4%
	Niveau 4	1%

Tableau 20 : Statistique descriptive de la garantie « Consultations et visites »

Sur la garantie « Consultations et visites », 31 953 bénéficiaires sont présents sur 22 régions en Métropole. Cependant, la zone géographique nécessite un regroupement en un nombre moins important de régions, car les bénéficiaires ne sont pas uniformément répartis. C'est pourquoi, nous regroupons les régions en 13 régions, ainsi la fréquence pour chaque région est supérieure à 4%, excepté pour la Corse où le poids est inférieur à 1%.

Régions	Poids
Alsace/ Champagne-Ardenne/ Lorraine	7,0%
Aquitaine/ Limousin/ Poitou-Charentes	11,6%
Auvergne/ Rhône-Alpes	8,0%
Bourgogne/Franche-Comté	4,3%
Bretagne	5,9%
Centre-Val-De-Loire	4,8%
Corse	0,4%
Haute-Normandie/ Basse-Normandie	5,6%
Ile-de-France	17,3%
Languedoc-Roussillon/Midi-Pyrénées	5,2%
Nord-Pas-de-Calais/Picardie	10,5%
Pays-De-La-Loire	5,9%
Provence-Alpes-Côte-D'azur	13,5%

Tableau 21 : Répartition de la variable région pour la garantie « Consultations et visites »

Pour les variables d'intérêt créées, la modalité 0 de la variable affaire nouvelle est plus représentée que celle de la modalité 1. C'est-à-dire, que la majorité des bénéficiaires ne sont pas de nouveaux adhérents, ils représentent environ 30% des bénéficiaires étudiés. En ce qui concerne la variable d'intérêt décelant les individus qui décèderont l'année N+2, celle-ci est surreprésentée par les individus en vie. Seulement 1% de la population décéderont l'année N+2.

Paramètres		Poids
Affaire nouvelle	**0**	69%
	1	31%
Décès N+2	**0**	99%
	1	1%

Tableau 22 : Statistique descriptive des variables d'intérêt pour la garantie « Consultations et visites »

1. Modélisation de la consommation

Dans un premier temps, seulement les variables techniques sont introduites dans la modélisation c'est-à-dire que les variables ajoutées en open data ne sont pas introduites**. La variable réponse est la variable montant de la consommation.** Les variables explicatives sont les variables introduites dans le paragraphe précédent « Analyse descriptive de la garantie Consultations et visites ».

Dans le cadre d'une modélisation GLM, une des hypothèses forte de celle-ci est de considérer que la variable réponse suit une loi de la famille exponentielle. Ainsi, les deux modèles les plus classiques permettant de modéliser les coûts individuels sont :

- le modèle Gamma ;
- le modèle log-normal, ou plutôt un modèle Gaussien sur le logarithme des coûts.

a) Choix de la distribution

Nous essayons de percevoir la courbe de consommation pour visualiser la tendance de consommation du portefeuille et de déterminer une loi usuelle la décrivant au mieux.
Lorsqu'on modélise une consommation, le plus souvent la distribution de celle-ci est une log-normale. La variable réponse est donc transformée par le logarithme. Cependant, dans notre étude, la variable réponse ne suit pas une loi log-normale. En effet, nous modélisons la consommation de soins de santé en prenant en compte la non-consommation. Il existe donc un pic de montants à zéro qui correspondent aux bénéficiaires n'ayant pas consommé l'une des deux années ou les deux années.

Méthode de Box-Cox :

On peut donc utiliser la méthode de Box-Cox qui permet de transformer les données vers la normalité. Cependant, avant de définir quelle transformation appliquer par la méthode de Box-Cox, il faut ajouter arbitrairement une constante à la variable réponse c'est à dire :

$$Y' = Y + constante,$$

car celle-ci prend des valeurs nulles, et pour effectuer une transformation de Box-Cox, il est nécessaire d'obtenir une série à termes positifs.
Après cette modification de la variable réponse, il faut appliquer une transformation par la méthode de Box-Cox. L'objectif est d'obtenir une distribution normale des données après transformation, par exemple une transformation logarithmique. La méthode Box-Cox permet de dilater les différences au niveau des queues de distributions à gauche et réduire les différences sur la queue de distribution à droite. La méthode Box-Cox cherchera la meilleure transformation possible, pas nécessairement une transformation logarithmique. Suite à cette transformation, la variable réponse peut être une loi normale.
La transformation de Box-Cox est définie ainsi :
Si x est une série de données, opérer une transformation de Box-Cox conduit à modéliser la série x':

$$x' = f(x,\lambda) = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} \; si \; \lambda \neq 0 \\ \log(x) \; si \; \lambda = 0 \end{cases}$$

Plusieurs valeurs de λ ont été testées, dont la valeur λ=0. Cependant, après les différentes transformations la variable réponse ne suit toujours pas une loi normale. On se penche donc vers une autre approche.

Distribution Tweedie :

Une autre méthode consiste à trouver une distribution prenant en compte ces montants à zéro. On a donc recours à la distribution « tweedie ». La distribution Tweedie apparaît comme le modèle théorique adéquat pour tenir compte du grand pourcentage de zéros dans la variable réponse. La famille de distribution Tweedie est un cas particulier de la famille exponentielle naturelle. Leur fonction de variance est bien spécifique, elle lie la variance et l'espérance selon la relation suivante :

$$V(Y) = am^{p}$$

Avec :

- $p \in]-\infty, 0] \cup [1, +\infty[$, paramètre qui contrôle la variance de la distribution ;
- $a > 0$, paramètre de dispersion ;
- m la moyenne de Y, $(m = E(Y))$.

Pour :

- $p = 0$, on retrouve une distribution normale ;
- $p = 1$, une distribution de Poisson ;
- $1 < p < 2$, loi composé Poisson-Gamma, car la distribution est continue pour des valeurs supérieures à zéro et une masse positive pour 0.
- $p = 2$, la distribution obtenue est une Gamma ;
- $p = 3$, la distribution est une inverse Gaussienne.

Les distributions Normale, Poisson, Gamma, et Inverse Gaussienne sont donc des cas particuliers de la loi Tweedie.
L'intérêt de cette loi, lorsque $1 < p < 2$, est de permettre de gérer un nombre important de valeurs nulles.

b) Limites de cette modélisation

Cependant, la base de données utilisée comporte des individus présents deux années, c'est à dire des données répétées. On a donc un individu par ligne qui est répété car les informations sont associées pour les deux années. Les informations de chaque bénéficiaire ne changent pas entre les années, seulement l'âge évolue d'un an par construction du portefeuille fermé et également la variable à expliquer qui est le montant de consommation. La limite à cette modélisation, c'est-à-dire en considérant simultanément les deux années est d'obtenir une dépendance entre les individus, et donc une corrélation des variables. Cette dépendance pourrait être écrasée par l'ensemble des individus, c'est à dire par la masse d'individus. On ne retiendra donc pas cette modélisation du fait de cette corrélation.
On aurait pu penser à l'utilisation de modèle mixte, c'est-à-dire pour des données répétées au fil du temps. En effet, la consommation de soins de santé est observée pour des individus au cours du temps. On aurait donc pu croire à l'utilisation d'un modèle mixte en modélisant le montant de consommation en fonction des variables explicatives dont l'année.
Cependant, en utilisant cette modélisation, par exemple l'effet âge serait calculé pour les deux années séparément, c'est-à-dire qu'on obtiendrait un effet âge pour l'année 2013, et un effet âge pour l'année 2014. On n'obtient donc pas un seul effet âge au global. Or, l'objectif est d'obtenir une dérive qui contient tous les effets y compris l'effet âge au global.

c) Pistes de modélisation

Modélisation des consommations en deux temps :
La dérive doit donc être modélisée en modélisant la consommation des deux années séparément. C'est-à-dire, modéliser les consommations en deux temps. Dans un premier temps, il faut modéliser la consommation de l'année 2013, puis dans un second temps celle de l'année 2014 pour ainsi obtenir une dérive estimée 2014/2013.

Pour se faire, il faut percevoir les courbes de consommation des années 2013 et 2014 pour ainsi choisir une distribution servant dans les GLM.
Les figures ci-dessous représentent la distribution de la variable réponse « montant » des années 2013 à gauche et 2014 à droite :

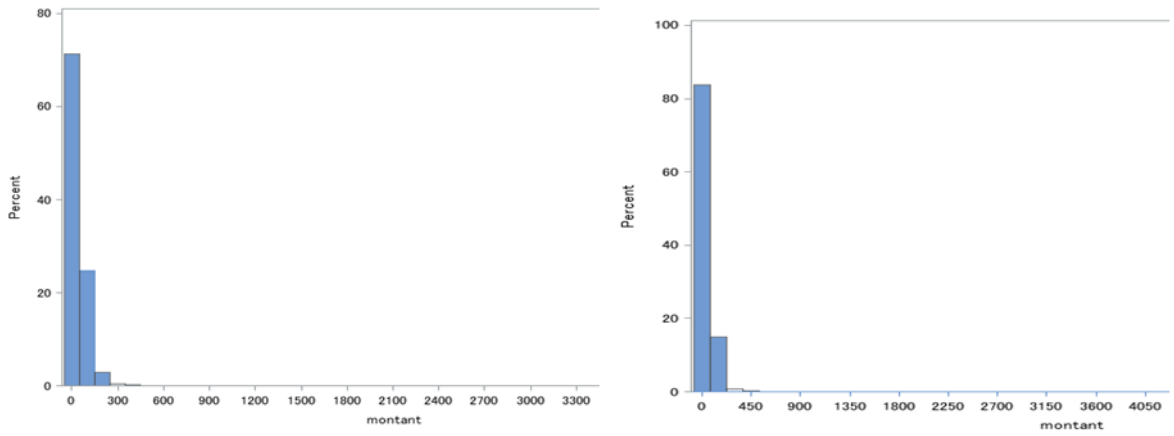

Figure 17 : Distribution de la variable réponse (montant de consommation) pour les années 2013 à gauche et 2014 à droite

Ces figures mettent en avant le pic de zéros et donc le choix de la distribution Tweedie. Pour estimer la dérive une possibilité serait donc d'estimer dans un premier temps le montant de consommation pour l'année N, en fixant des paramètres de référence (par exemple un individu de sexe féminin, habitant en Ile de France etc.). Dans un second temps, de façon similaire, le montant de consommation de l'année $N+1$ est modélisé en fixant les mêmes paramètres de références de l'année N.
Ainsi, nous obtenons un montant estimé l'année N pour un individu i exprimé par une fonction de la façon suivante :

$$Y_i^N = f(\, intercept^N + \sum_{j=1}^{p} \beta_j^N \left(X_i^j\right)^N \,)$$

Avec :

- $i = 1$ à n, où n est le nombre d'individus de la base de données considérée ;
- Y_i^N, le montant de consommation de l'année N du $i^{\text{ème}}$ individu ;
- $l'intercept^N$, le coefficient de l'année N correspondant à l'individu de référence pour lequel les modalités des variables explicatives ont été définies ;
- p variables explicatives $(X^1)^N, (X^2)^N, \dots, (X^p)^N$ prenant les valeurs $X_i^1, X_i^2, \dots, X_i^p$ pour l'individu i l'année N.
-

De même, pour l'année $N+1$ le montant de consommation s'exprime de la façon suivante :

$$Y_i^{N+1} = f(\, intercept^{N+1} + \sum_{j=1}^{p} \beta_j^{N+1} \left(X_i^j\right)^{N+1} \,)$$

Avec :

- $i = 1$ à n, où n est le nombre d'individus de la base de données considérée ;
- Y_i^{N+1}, le montant de consommation de l'année $N+1$ du $i^{\text{ème}}$ individu ;
- $l'intercept^{N+1}$, le coefficient de l'année $N+1$ correspondant à l'individu de référence pour lequel les modalités des variables explicatives ont été définies ;
- p variables explicatives $(X^1)^{N+1}, (X^2)^{N+1}, \dots, (X^p)^{N+1}$ prenant les valeurs $X_i^1, X_i^2, \dots, X_i^p$ pour l'individu i l'année $N+1$.

Les valeurs des variables $\left(X_i^j\right)^{N+1}$ ne prennent pas les même valeurs que $\left(X_i^j\right)^N$ pour toutes les modalités. En effet, les valeurs de certaines variables ont pu évoluer. Le portefeuille construit est un portefeuille fermé donc par construction de celui-ci, il s'agit des mêmes bénéficiaires présents sur deux

ans (sans changements de caractéristiques propres à l'individu). Cependant, l'âge des bénéficiaires évolue entre les deux années, les modalités de la variable « classe d'âge » seront donc différentes.

Ainsi la dérive $N+1/N$ est la différence des coefficients obtenus de la modélisation pour l'année $N+1$ et N. Cependant, cette méthode est juste uniquement dans le cas où il y a très peu de variabilité entre les coefficients obtenus de la modélisation GLM de l'année N, et les coefficients obtenus du GLM de l'année $N+1$. C'est-à-dire, que les coefficients n'évoluent pas. C'est pour cette raison que cette méthode ne sera pas retenue, car pour certaines variables explicatives il pourrait y avoir une différence entre les coefficients de l'année N et $N+1$.
Ci-dessous, un tableau mettant en avant des différences obtenues pour des coefficients du GLM entre les années N et $N+1$ de la variable classe d'âge :

Paramètres		Coefficients de l'année N	Coefficients de l'année $N+1$
Classe d'âge	1	-0.1966	-0.3013
	5	-0.1327	-0.1671

Tableau 23 : Différence des coefficients obtenus des GLM pour les années N et N+1

Dans la même idée, une autre piste de modélisation est de tester les modèles dit « *zero-inflated*». Ces modèles sont utilisés lorsqu'il y a un pic de zéros dans un modèle de comptage, typiquement un modèle de Poisson ou binomial négatif. Pour appliquer ce modèle à notre base, il faut donc transformer la variable réponse (le montant de consommation), variable continue en variable discrète, et dans ce cas, utiliser les modèles ZIP pour « **Z**ero **I**nflated **P**oisson », cela signifie que la variable réponse suit une loi de Poisson en présence d'une masse de zéros.

Modélisation de l'évolution absolue :
Une autre approche consiste donc à ne pas avoir de données répétées et donc modéliser l'évolution absolue. C'est-à-dire construire une base de données avec une ligne par individu en lui associant l'évolution absolue du montant de consommation. Ainsi, l'information de la consommation des deux années est contenue sur une seule ligne. La variable à expliquer est donc la variable Y' qui s'écrit de la façon suivante :

$$Y' = montant_{N+1} - montant_N$$

A chaque bénéficiaire sont ajoutées les variables techniques, c'est à dire le sexe, le régime, la gamme, la région, les variables d'intérêt, ainsi que l'âge moyen de chaque bénéficiaire entre les deux années.
En observant la distribution de cette variable Y', on remarque que ce n'est pas une distribution usuelle paramétrique. On obtient également un pic de zéros au milieu de la distribution. Elle pourrait s'apparenter à une loi normale.
Ces résultats nous poussent donc à nous orienter vers une autre approche de modélisation en créant des profils d'individus qui regroupent les assurés ayant les mêmes caractéristiques. C'est-à-dire, de ne pas considérer une approche purement individuelle.

2. Modélisation du taux d'évolution

Dans le paragraphe précédent, une des pistes de modélisation est la modélisation du taux d'évolution absolu. C'est pourquoi, dans le même esprit de modélisation, le taux d'évolution relatif est considéré sur des profils d'individus.

Pour modéliser le taux d'évolution, il faut introduire la variable taux d'évolution dans la base de données 2013-2014 définie par :

$$Taux\ évolution\ N+1/N = \frac{Consommation\ N+1}{Consommation\ N} - 1$$

Ce taux d'évolution est exprimé en pourcentage.
La première étape avant de modéliser le $\boldsymbol{taux\ d'évolution\ N+1/N}$ **(variable réponse)** est de créer des profils d'individus.

a) Retraitements des données

La création de profils d'individus se fait en croisant plusieurs variables afin d'obtenir des profils ayant des caractéristiques communes et d'associer le nombre de bénéficiaires par profil. Les variables utilisées sont les variables définies dans le paragraphe introductif de la partie II. Résultats :

- Le sexe ;
- La classe d'âge ;
- La gamme ;
- Le niveau de couverture ;
- Les régions ;
- La variable affaire nouvelle ;
- La variable décès N+2.

La création de profils va permettre de grouper plusieurs individus dans un même profil et de calculer une consommation moyenne par profil. Le calcul d'une consommation moyenne limite donc le nombre de montants à zéro.
Dans un premier temps, le croisement de plusieurs individus est effectué avec toutes les variables internes définies plus haut. Cependant, la variable décelant les décès n'est pas représentative des bénéficiaires comme expliqué dans la partie création de variables. Elle peut représenter un biais, c'est pour cela que le croisement des individus est effectué en ne considérant pas cette caractéristique.
Plusieurs types de profils d'individus ont été créés en croisant à chaque fois différentes variables introduites plus haut, et ensuite en modélisant la dérive. Ci-dessous un exemple de croisement d'individus pour la création de profils d'individus:

Profil	Sexe	Classe d'âge	Gamme	Niveau de couverture	Affaire nouvelle	Nombre d'individus
1	F	1	Santé Actif	Niveau 1	**0**	762
2	F	1	Santé Actif	Niveau 1	**1**	335
3	F	1	Santé Actif	**Niveau 2**	0	766
4	F	1	Santé Actif	Niveau 2	1	368

Tableau 24 : Exemple de création de profil d'individus

Ce tableau montre comment sont créés les profils d'individus. On regroupe dans un premier temps, les individus dont les modalités des variables sont identiques. Dans un second temps, est calculé le nombre d'individus par profil, enfin le montant moyen par profil.
Par exemple, pour les deux premières lignes, les modalités des variables de croisement sont les mêmes sauf pour la variable affaire nouvelle (prenant la valeur de **0** pour le premier profil et **1** pour le deuxième profil) créant ainsi deux profils d'individus différents. De même, pour le troisième profil la modalité de la

variable niveau de couverture (niveau 2) est différente du profil 1 (niveau 1), créant ainsi un nouveau profil.
Cet exemple de création de profils a utilisé les caractéristiques des individus suivantes : sexe, classe d'âge, gamme, niveau de couverture et affaire nouvelle.

Ainsi, la meilleure estimation de la dérive a été obtenue en considérant les variables suivantes :

- Le sexe ;
- La classe d'âge ;
- La gamme ;
- Le niveau de couverture ;
- La variable affaire nouvelle.

On présentera donc les résultats de la modélisation avec les variables précédentes.
Le nombre d'individus au sein d'un même profil varie de 1 individu à 1 723 individus. Le tableau ci-dessous met en avant le nombre d'individus pouvant composer un profil :

Niveau	Nombre d'individus par profil
La valeur maximale	1 723
La médiane	28
La valeur minimale	1

Tableau 25 : Etendue de la variable "nombre d'individus »

La médiane nous indique que 50% des profils d'individus sont composés de plus de 28 individus et 50% de moins de 28 individus.

Pour calculer un taux d'évolution, un montant à zéro peut poser problème c'est pourquoi sont supprimés les profils dont les montants sont à zéros. On comptabilise 3 profils dont la consommation est nulle. Ces profils regroupent 4 individus du jeu de données. Ils représentent donc moins de 1% des profils d'individus.
Après ce retraitement des montants à zéro, l'étape suivante est d'écarter dans la modélisation les taux d'évolution qui sont trop élevés ou trop faibles. C'est-à-dire, d'écarter les bénéficiaires qui sont en marge. Nous traitons des taux d'évolution par individu donc l'étendue du taux d'évolution est entre -91% et 666%, dont 50% des bénéficiaires de la base ont un taux d'évolution de l'ordre de -4,5%. Nous allons donc considérer dans la modélisation les taux d'évolution entre -30% et 30%, en écartant ainsi les individus possédant un taux d'évolution en marge. En choisissant de considérer l'étendue de ce taux d'évolution, on écarte moins de 2% de bénéficiaires. On peut donc effectuer cette restriction sans perte d'informations.

b) Choix de la distribution

Au vu de l'allure de la courbe des taux d'évolution c'est une loi Normale qui est sélectionnée.

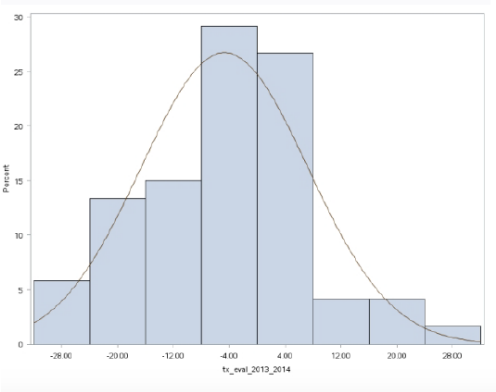

Figure 18 : Allure de la distribution de la variable réponse *taux d'évolution*

On commence donc par s'assurer de la validité de l'hypothèse de normalité de la variable réponse. Certes, le nombre d'observations peut être jugé suffisamment grand pour effectuer l'approximation normale, mais il peut s'avérer que cela n'est pas toujours vérifié. Des tests empiriques comme les coefficients d'asymétrie et d'aplatissement sont donc utilisés. La loi normale est caractérisée par un coefficient d'asymétrie (Skewness) et d'aplatissement (Kurtosis) nul. (Se référer à l'Annexe B : Tests empiriques de normalité - paramètres de forme pour plus de détails sur ces coefficients).
Il paraît donc naturel de calculer ces indicateurs. Si ces indicateurs sont suffisamment proches de la valeur 0, l'hypothèse de compatibilité avec la loi normale ne peut être rejetée. Les coefficients obtenus sont les suivants :

Coefficients	Valeurs
Skewness	0,2091416
Kurtosis	0,26880427

Tableau 26 : Skewness et Kurtosis de la variable taux d'évolution

Nous constatons quant aux résultats des tests, que l'adéquation à la loi normale paraît plausible.

On remarque qu'en incluant les points extrêmes des taux d'évolution et en ne se concentrant pas que sur les taux d'évolution entre -30% et 30%, ces mêmes indicateurs prennent des valeurs sensiblement différentes. Par exemple, si nous considérons des taux d'évolution entre -100% et 60% le Skewness a pour valeur -0,5617408 et le Kurtosis 1,86102797, confirmant qu'un individu s'écartant significativement de la majorité de la population peut fausser les résultats.

De plus, l'utilisation d'un outil graphique le QQ-plot (quantile-quantile plot) permet également de comparer la pertinence de l'ajustement d'une distribution à un modèle théorique. Si les données sont compatibles avec la loi normale, une droite ajuste au mieux le nuage de points. Ci-dessous, le QQ-plot obtenu :

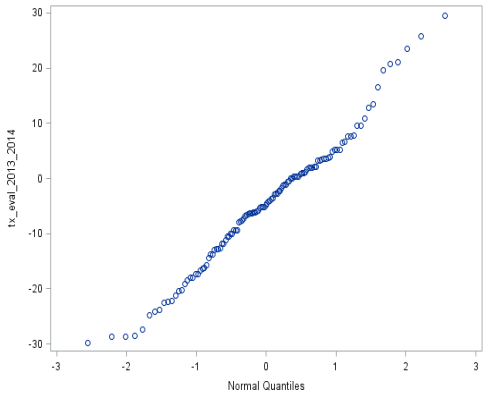

Figure 19 : QQ-plot de la variable réponse

D'autres indicateurs peuvent être utilisés pour apprécier rapidement l'écart à la loi normale. Par exemple, la loi étant symétrique, l'écart entre la médiane et la moyenne ne devrait pas être élevé. Dans notre jeu de données, la médiane est égale à -4,67311 et la moyenne à -4,69276. Mais ce dispositif permet seulement d'apprécier la symétrie de la distribution.

Il existe également des tests statistiques pour vérifier l'adéquation à la loi normale. A tout test est associé un risque α $(souvent\ \alpha = 5\%)$ dit de première espèce. Il s'agit de la probabilité de rejeter l'hypothèse de normalité alors qu'elle est vraie. Pour le test de Kolmogorov-Smirnov, (se référer à l'Annexe C : Les différents tests de normalité , pour plus d'explications sur ce test) par exemple, une p-valeur > 0,05 permet d'accepter l'appartenance à loi. Dans notre jeu de données la p-valeur est supérieure à 0,150: ainsi on ne peut rejeter l'hypothèse selon laquelle la variable réponse suit une loi normale. Les résultats obtenus par ces tests statistiques sont influencés par la taille de l'échantillon. De ce fait, les approches empiriques et graphiques gardent toute leur importance.

Ainsi, l'hypothèse de normalité de la variable réponse est retenue.

c) Les coefficients obtenus

Dans cette partie nous présentons les résultats obtenus en précisant l'individu de référence. Dans le GLM avec l'utilisation de la loi normale, réalisé sur le taux d'évolution 2013-2014 l'individu de référence utilisé possède les caractéristiques suivantes :

- De sexe féminin ;
- Appartenant à la gamme Santé Actif ;
- Couvert par un niveau de base : niveau 1 ;
- Ayant entre 45 ans et 58 ans : classe d'âge 5 ;
- N'étant pas un nouvel adhérent (AN=0).

Pour rappel la modélisation est effectuée pour des bénéficiaires du Régime général.
Toutes les variables présentées dans cette modélisation sont significatives. Les variables considérées comme significatives sont les variables dont la p-valeur est inférieure à 5%. Les coefficients obtenus sont les suivants :

Paramètres		Coefficients
Intercept		-2,227
Sexe	**Femme**	**0**
	Homme	-1,9501
Gamme	**Santé Actif**	**0**
	Santé Senior	2,2696
Niveau de couverture	**Niveau 1**	**0**
	Niveau 2	-3,7134
	Niveau 3	-3,7946
	Niveau 4	-2,0041
Classe d'âge	**1**	-5,6709
	2	-0,0226
	3	2,9507
	4	2,953
	5	**0**
	6	2,9468
	7	-1,7199
Affaire nouvelle	**Non**	**0**
	Oui	2,9468

Tableau 27 : Les coefficients obtenus du GLM

D'après les résultats obtenus, nous interprétons les résultats ainsi : sur la garantie « Consultations et visites » :

- Les hommes ont une évolution de consommation moins élevée que les femmes ;
- Les bénéficiaires de la gamme Santé Senior consomment plus que ceux de la gamme Santé Actif ;
- L'évolution de consommation pour des niveaux de couverture supérieurs à 100% de la BR est moins élevée que l'évolution de consommation du niveau de couverture de base ;
- Les bénéficiaires des classes d'âges 3, 4 et 6 ont une évolution de consommation plus élevée qu'un bénéficiaire entre 45 et 58 ans.
- L'évolution de consommation est plus élevée pour les nouveaux adhérents.

A partir de ces résultats, le taux d'évolution estimé est déterminé par :

$$taux\ évolution_{\ estimé} = intercept + \sum_{j=1}^{p} \beta_j X_j$$

Avec :

- l'*intercept* : le coefficient correspondant à l'individu de référence pour lequel les modalités des variables qualitatives et quantitatives ont été définies plus haut ;
- β_j : correspond au coefficient obtenu pour chaque paramètre j associé à la variable X_j ;
- X_j : correspond à la valeur de la $j^{\text{ème}}$ variable explicative.

Avec les résultats du tableau précédent nous obtenons les coefficients suivants :

$$taux\ évolution_{\ estimé} = -2{,}227 - 1{,}9501 * Cosex_H + \cdots + 2{,}9468 * Affaire\ nouvelle_{Oui}$$

d) Validation de la modélisation

Dans un premier temps, le contrôle des variables a été effectué en s'assurant que les variables sont significatives. Dans un second temps, il faut s'assurer des hypothèses de normalités des résidus.
Au niveau des tests statistiques, les coefficients retenus lors de la modélisation sont significatifs et nous permettent d'obtenir une estimation du taux d'évolution. Cependant, ces tests ne sont pas suffisants. Afin de valider les résultats obtenus, il faut les tester sur une nouvelle base de données et comparer la dérive

observée et estimée. En effet, lors de la modélisation, les variables ont été sélectionnées en deux temps : dans un premier temps, la significativité des variables a été observée à l'aide de la p-valeur et dans un second temps, la sélection des variables s'est effectuée grâce au critère AIC.
Mais également, on observe si l'ajout ou la suppression de certaines variables améliore l'estimation de la dérive.
Les résultats de la modélisation précédente (de la base de données 2013-2014) sont donc testés sur un nouveau jeu de données, c'est-à-dire que les coefficients obtenus lors de la modélisation GLM sont appliqués à une nouvelle base de données. Ainsi, une base de données est créée. Cette base de données est constituée des bénéficiaires présents entre le 1er janvier de l'année 2014 et le 31 décembre de l'année 2015. La création de cette base de données est effectuée de la même manière que la base de données présentée dans la Partie 3 : Description et construction d'une base de données.
Une analyse descriptive des variables de la garantie « Consultations et visites » de la base de données 2014-2015 est présentée ci-dessous :

Paramètres		Poids
Sexe	**Femme**	56%
	Homme	44%
Classe d'âge	**1**	15%
	2	9%
	3	12%
	4	8%
	5	16%
	6	31%
	7	9%
Gamme	**Santé Actif**	55%
	Santé Senior	45%
Niveau de couverture	**Niveau 1**	54%
	Niveau 2	41%
	Niveau 3	4%
	Niveau 4	1%

Tableau 28 : Statistique descriptive des variables techniques de la base de données 2014-2015

Le nombre de bénéficiaires de cette base de données en gardant la restriction au Régime général est de 48 563.
La séparation du sexe sur ce jeu de données est équilibrée. Il en est de même pour la répartition des gammes entre Santé Actif et Santé Senior.

Comme pour la base créée à partir des bénéficiaires présents entre le 1er janvier de l'année 2013 et le 31 décembre de l'année 2014, le nombre de personnes décédant l'année N+2 est également de 1%. D'autre part, la majorité des bénéficiaires ont souscrit il y a plus d'un an. On remarque que moins de 20% sont des nouveaux adhérents comme le montre le tableau ci-dessous :

Paramètres		Poids
Affaire nouvelle	0	82%
	1	18%
Décès N+2	0	99%
	1	1%

Tableau 29 : Poids des variables d'intérêt de la base 2014-2015

Le tableau ci-dessous complète l'analyse des variables précédentes : la région la plus représentée est l'Ile de France avec 18% des bénéficiaires suivi de la région de PACA.

Régions	Poids
Alsace/ Champagne-Ardenne/ Lorraine	6,6%
Aquitaine/ Limousin/ Poitou-Charentes	12,0%
Auvergne/ Rhône-Alpes	8,3%
Bourgogne/Franche-Comté	3,9%
Bretagne	5,9%
Centre-Val-De-Loire	4,6%
Corse	0,4%
Haute-Normandie/ Basse-Normandie	4,9%
Ile-de-France	18,0%
Languedoc-Roussillon/Midi-Pyrénées	5,4%
Nord-Pas-de-Calais/Picardie	8,4%
Pays-De-La-Loire	6,2%
Provence-Alpes-Côte-D'azur	15,4%

Tableau 30 : Répartition des régions de la base 2014-2015

Après la création d'un jeu de données regroupant les bénéficiaires présents du 1er janvier de l'année 2014 au 31 décembre de l'année 2015, l'étape suivante est la création de profils d'individus. Celle-ci est établie de façon similaire au jeu de données servant à la modélisation. On compte 175 profils d'individus différents regroupant de 1 à 3 413 individus par profil. Parmi ces 175 profils, 7 profils ont un montant à zéro, il faut donc les supprimer pour pouvoir calculer la variable taux d'évolution. L'étendue de la variable taux d'évolution est de -90% à plus de 1000%, dont 50% des bénéficiaires ont un taux d'évolution de l'ordre de 2%. Puisque nous observons des taux d'évolution individu par individu, ceux-ci peuvent donc avoir des valeurs élevées tant en négatives que positives. Ainsi, une personne consultant un médecin généraliste trois fois la première année et deux fois plus l'année suivante, son montant total de consommation de soins sera considérablement plus élevé cette même année. L'évolution de consommation de cette dernière sera donc importante.
Enfin, l'application des coefficients obtenus lors de la modélisation aux profils d'individus de 2014/2015 créés, permet d'estimer un taux d'évolution 2014/2015 pour chaque profil d'individu.

Le montant 2015 estimé par profil est obtenu par la relation suivante :

$$montant_{estimé}^{2015} = montant^{2014} * (1 + taux\ d'évolution_{estimé}^{2014/2013})$$

Où, le $taux\ d'évolution_{estimé}^{2014/2013}$ est le taux d'évolution obtenu lors de la modélisation GLM.

Pour calculer la dérive 2015/2014 estimée, il faut tout d'abord calculer le coût total des dépenses estimées de 2015 et le coût de 2014 pour chaque profil d'individu. Chaque profil d'individu est constitué

de plusieurs individus, ainsi pour obtenir le coût total des dépenses estimées il faut multiplier le montant estimé par le nombre d'individu :

$$co\hat{u}t\ total\ {}^{2015}_{estimé} = montant^{2015}_{estimé} * freq$$

Où, la variable $freq$ représente le nombre d'individus par profil.
Pour des raisons de confidentialité, les coûts totaux ne seront pas affichés.

La dérive 2015/2014 observée sur le jeu de données est obtenue par la formule suivante :

$$dérive\ \ 2015/2014_{observée} = \frac{co\hat{u}ts\ totaux^{2015}}{co\hat{u}ts\ totaux^{2014}} - 1$$

Où, les $co\hat{u}ts\ totaux^{2015}$ et $co\hat{u}ts\ totaux^{2014}$ représentent respectivement la somme des coûts pour l'année 2015 et la somme des coûts de l'année 2014.
La valeur obtenue pour la $dérive\ \ 2015/2014_{observée}$ est de -4,7%.

Tandis que la dérive 2015/2014 estimée est obtenue par la formule suivante :

$$dérive\ \ 2015/2014_{estimée} = \frac{co\hat{u}ts\ totaux\ {}^{2015}_{estimé}}{co\hat{u}ts\ totaux^{2014}} - 1$$

Où, les $co\hat{u}ts\ totaux\ {}^{2015}_{estimé}$ et $co\hat{u}ts\ totaux^{2014}$ représentent respectivement la somme des coûts estimé pour l'année 2015 et la somme des coûts de l'année 2014.
Sa valeur est de -3,1%. La dérive 2015/2014 est donc surestimée de 1,6 point.

Cependant, on retrouve une dérive négative pour la dérive estimée de même que la dérive observée.
Les dérives estimées et observées sont de même signe. L'estimation de la dérive donne donc le sens d'évolution des dépenses pour un poste de soins de santé. Quant à l'écart entre l'estimation et l'observation, il peut être obtenu par des facteurs non maitrisables, c'est-à-dire des variables qui peuvent être inobservables ou difficilement modélisables. En effet, des facteurs exogènes sont difficilement quantifiables. Il est donc difficile d'obtenir une mesure précise de la dérive. Cette modélisation tente d'obtenir une mesure de la dérive à court terme par le biais de certains facteurs quantifiables (comme le sexe, l'âge). Toutefois, l'estimation précise de la dérive est difficilement modélisable, il faut donc tenir compte d'une marge de prudence.

3. Limites des modèles GLM

Dans les paragraphes précédents, les méthodes de régression linéaire présentent des limites importantes :

- Dans le cas de modèle linéaire généralisé, il est fréquent que les hypothèses d'homogénéité de la variance des résidus ou de la normalité ne soient pas vérifiées, ce qui entraîne l'utilisation de modèles erronés.
- Le modèle GLM est dit paramétrique, en effet, il nécessite de préciser une loi pour la variable réponse qui est souvent une contrainte difficile à remplir. L'approche GLM impose de faire une hypothèse sur la forme de la loi conditionnelle de la variable expliquée Y en fonction des variables explicatives.

Cette hypothèse peut s'avérer fausse, on prend donc un risque pour le modèle. Il est alors possible de chercher à modéliser directement la forme de l'espérance conditionnelle, mais sans faire l'hypothèse de loi complète de la variable expliquée. Une solution est donc d'utiliser les modèles GAM (Modèle Additif Généralisé) qui relâchent l'hypothèse de linéarité entre la variable à expliquer et les variables explicatives. Il est de même possible d'utiliser des modèles de régression non paramétriques. Une telle démarche permettrait de faire diminuer le risque de modèle en prenant des hypothèses moins restrictives.

Conclusion

Dans le cadre du pilotage du risque santé, il est important pour les organismes complémentaires d'assurance santé de prévoir la sinistralité de leur portefeuille, la « dérive » des prestations des soins de santé. Ces dernières années, les dépenses de santé ont beaucoup évolué. On constate une accélération des dépenses relatives aux soins de santé et une augmentation de la part des remboursements dans ces dépenses par les organismes complémentaires. Les principaux facteurs d'évolution des dépenses de santé sont l'évolution des récentes réglementations et les facteurs environnementaux.

Dans le cadre de l'indexation tarifaire, afin de connaitre la progression future du portefeuille, l'assureur a besoin de mesurer la dérive. Dans ce mémoire, nous nous sommes intéressés à l'estimation de la dérive des soins de santé à court terme.

L'objectif de ce mémoire a donc été de mesurer à court terme la dérive des soins de santé, évolution de la consommation de soins, sur le segment individuel et de regarder l'apport d'une modélisation de l'évolution individuelle.

Initialement, a été effectué une estimation de la dérive pour les années 2017 et 2018 par une approche macroéconomique. Dans cette approche, l'estimation de la dérive a été établie par la construction d'une tendance. En effet, en raison d'importantes modifications apportées aux garanties en 2016 dans le cadre de la mise en conformité au contrat responsable, il n'est pas possible de dégager une tendance 2016 pure. C'est pourquoi, une tendance sur les années antérieures a été construite.

On parle d'approche macroéconomique dans le sens où l'évolution des dépenses est agrégée. Pour un acte médical donné, le volume des prestations et le nombre de bénéficiaires couverts sont agrégés. Néanmoins, une approche macroéconomique permet uniquement d'appréhender des effets globaux. L'apport d'une approche individuelle peut être donc nécessaire pour capter les effets liés à la dérive.

Pour mettre en place la modélisation, la première étape a consisté en la création d'une base de données. Cette étape, a été longue à réaliser. En effet, le choix du périmètre étudié et la récolte de données fiables ont été nécessaires à la modélisation.
Une fois la base de données créée, la modélisation de la dérive a été réalisée en plusieurs étapes. La première étape a consisté en la modélisation de la consommation de santé en deux temps, c'est-à-dire la modélisation de la consommation de l'année N puis ensuite celle de l'année N+1 par un modèle linéaire généralisé, pour ensuite obtenir la dérive de l'année N+1/N. Cependant, la distribution de la consommation est difficile à modéliser. En effet, la prise en compte des non consommants fait apparaître un pic de zéros dans la distribution. De plus, l'inconvénient de cette méthode pour l'estimation de la dérive est le fait de devoir avoir très peu de variabilité entre les coefficients obtenus lors des deux modélisations par un modèle linéaire généralisé. Or ce n'est pas forcément le cas en pratique.
Après avoir essayé d'estimer la dérive par le biais de la modélisation de la consommation de soins de santé, la deuxième étape a été de s'orienter directement vers la modélisation de l'évolution de la consommation. Cependant, lors de la précédente étape de modélisation un nombre important de montants nuls a été constaté. De ce fait, pour diminuer le nombre de montants nuls, la création de profil d'individus a été nécessaire pour ainsi regrouper plusieurs individus et associer à chaque profil un coût moyen. Un modèle linéaire a été conservé pour modéliser l'évolution des consommations. En pratique, les résultats de cette modélisation ne sont pas tout à fait satisfaisants dans le sens où ils ne donnent pas

une mesure précise de la dérive. Cependant, ce modèle a permis de donner l'ordre de grandeur et le sens de l'évolution pour un poste de soins. De plus, il a permis de quantifier les variables influentes sur la dérive.
Les variables ajoutées en open data dans la base de données ont été écartées, car la modélisation en intégrant les variables techniques (sexe, âge…) a été difficile à mener. C'est pourquoi l'ajout des variables en open data n'a pas été pris en compte afin de maitriser avant tout l'intégration des variables techniques.

Pour conclure, l'estimation d'une dérive par une approche individuelle en prenant en compte plusieurs facteurs peut être difficile à obtenir. En effet, des facteurs exogènes non observables viennent influencer le comportement des dépenses de chaque individu. Cependant cette modélisation peut mettre en avant l'influence des variables sur la dérive.

Dans une approche de recherche, ce mémoire a été réalisé sur un portefeuille individuel, il serait intéressant de modéliser la dérive en considérant un portefeuille collectif qui donnerait une vision globale de l'estimation de la dérive. De plus, l'approche d'un portefeuille collectif est différente en terme de structure faisant intervenir d'autres variables.
D'autre part, on pourrait s'interroger sur l'apport d'une approche d'apprentissage statistique ne nécessitant pas d'hypothèse de lois sur la variable réponse.

Glossaire

Pour bien interpréter les différents indicateurs présents au long du rapport, il est nécessaire de définir le vocabulaire utilisé.

ACS : Aide au paiement d'une complémentaire santé

ANI : Accord national interprofessionnel

BIPE : Bureau d'information et de prévisions économiques

BR : Base de remboursement de la Sécurité Sociale

CAS : Contrat d'accès aux soins

CNAF : Caisse Nationale d'Allocations Familiale

CNAM : Caisse Nationale d'Assurance Maladie

CNAV : Caisse Nationale d'Assurance Vieillesse

CSBM : Consommation des soins et biens médicaux

FNMF : Fédération Nationale de la Mutualité Française

GLM : Modèle linéaire généralisé

OD : Optique et dentaire

ONDAM : Objectif National des dépenses d'assurance maladie

PACA : La région de Provence-Alpes-Côte-D'azur

PLFSS : Loi de financement de la Sécurité Sociale

RO : Régime obligatoire de base, il peut s'agir soit du régime général, soit du régime des travailleurs non-salariés ou soit du régime local (Alsace-Moselle).

RRO : Remboursement du Régime obligatoire

SCH : Soins courants et hospitalisation

TFR : Tarif de Responsabilité

TM : Ticket modérateur, il représente la part de dépense qui reste à la charge de l'assuré, une fois déduit le remboursement de la Sécurité Sociale.

TSCA : Taux spécial sur les contrats d'assurance

Table des figures

Table des tableaux

Bibliographie

Sites internet :

- Site de l'Assurance Maladie : www.ameli.fr;
- Site de la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques : www.dress.sante.gouv.fr ;
- Site de l'Institut de Recherche et Documentation en Economie de la Santé : www.irdes.fr;
- Site de l'Institut National de la Statistique et des Etudes Economiques : www.insee.fr ;
- Site de l'Argus de l'Assurance : www.argusdelassurance.com ;
- Site du groupe AG2R La Mondiale : www.ag2rlamondiale.fr ;
- Support pour SAS : support.sas.com ;
- www.ressources-actuarielles.net ;
- Site de ressources mathématiques de l'Université de Toulouse: www.wikistat.fr;
- Les données météorologiques : www.meteofrance.fr ;
- Les données épidémiologiques : www.sentiweb.fr ;

Etudes :

- BIPE, *Observatoire complémentaire santé*, Présentation du 29 juin 2017
- BIPE, *Observatoire complémentaire santé*, Support de présentation du 30 mars 2017
- FNMF, *Etude « Coût du risque de l'assurance maladie complémentaire 2017-2018 »,* Support de présentation
- DRESS, *La complémentaire santé, Acteurs, bénéficiaires, garanties,* édition 2016
- *Les comptes de la Sécurité Sociale, Résultats 2015, Prévisions 2016 et 2017*, rapport septembre 2016

Mémoire d'actuariat :

- MERDIGNAC M., Modélisation de la consommation en santé en fonction de variables exogènes
- BOSSANT V., Frais de santé des organismes complémentaires, Analyse sur plusieurs portefeuilles
- GOURLIER S., Analyse de la rentabilité d'un produit en santé individuelle

Livres et note de cours :

- MEZZIANI K., [2014], *Introduction aux modèles linéaires généralisés,* Cours, Université Paris-Dauphine
- TUFFERY S., [2012], *Data Mining et statistique décisionnelle, L'intelligence des données, 4e édition, Technip.*
- CHARPENTIER A., [2013], *Modèles linéaires généralisés*, Transparents de cours
- RAKOTOMALALA R., [2011], *Test de normalité, Techniques empiriques et tests statistiques,* Cours, Université Lumière Lyon 2
- COLLETAZ G., [2017], *Statistique non paramétrique,* Cours, Université d'Orléans

Annexes

Annexe A : Compléments sur les GLM

➢ **Equation de vraisemblance dans le cas de lien canonique :**

Les équations de la vraisemblance sont dans le cas général donné pour $j = 1, \dots, p$:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

Dans le cas où le lien canonique est utilisé, plusieurs simplifications interviennent :

- $\eta_i = \theta_i = X_i{}'\beta$,
- $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = v''(\theta_i) = \frac{Var(y_i)}{\phi}$

Ainsi,

$$\frac{\partial \mathcal{L}_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{Var(y_i)} \frac{Var(y_i)}{\phi} x_{ij} = \frac{(y_i - \mu_i)}{\phi} x_{ij}$$

Donc, les équations de vraisemblance pour un lien canonique s'écrivent pour $j = 1, \dots, p$:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi} x_{ij} = 0$$

➢ **Algorithme IRLS /de Newton-Raphson :**

Les équations de vraisemblance sont en générales transcendantes, c'est à dire qu'il n'est en général pas possible de donner une expression analytique de l'estimateur de maximum de vraisemblance (EMV) en les résolvant. Une solution est d'utiliser des procédures itératives d'optimisation dont la démarche est la suivante :

Algorithme :

Départ : β^0

$k \leftarrow 1$

Répéter :

$\beta^{k+1} \leftarrow \beta^k + A^k \nabla \mathcal{L}_n(\beta^k)$

$k \leftarrow k + 1$

Jusqu'à $\beta^{k+1} \approx \beta^k$ et/ou $\mathcal{L}_n(\beta^{k+1}) \approx \mathcal{L}_n(\beta^k)$

L'expression de A^k pour les deux algorithmes :

- L'algorithme IRLS (Iterative Reweighted Least squares) avec :

$$A^k = -\left[\mathrm{E}_\beta(\nabla^2 \mathcal{L}(\beta^k))\right]^{-1}$$

- L'algorithme de Newton-Raphson (RS) est une méthode de gradient (recommandé quand le critère est concave) dont la direction est donnée par $\nabla \mathcal{L}$ et le pas par :

$$A^k = -\left[\nabla^2 \mathcal{L}(\beta^k)\right]^{-1}$$

Annexe B : Tests empiriques de normalité - paramètres de forme

- **Le coefficient d'asymétrie (Skewness) :**

Le coefficient d'asymétrie (proposé par Fisher) d'une variable aléatoire X est défini par :

$$\gamma_1 = \frac{\mu_3}{\sigma^3},$$

avec σ l'écart type et $\mu_3 = \mathrm{E}(X - \mathrm{E}(\mathrm{X}))^3$.

Le coefficient d'asymétrie donne une mesure du degré d'asymétrie de la distribution. Lorsque $\gamma_1 = 0$, la distribution est symétrique, sinon l'asymétrie penchera vers la gauche ou vers la droite suivant que γ_1 soit négatif ou positif. Si $\gamma_1 < 0$, on dit qu'il y a une asymétrie à gauche et que la moyenne est plus petite que la médiane.

L'expression du coefficient d'asymétrie sous SAS s'obtient en remplaçant :

- La variance σ^2 par son estimateur sans biais $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x)}^2$
- μ_3 par son estimateur sans biais

Ainsi, γ_1 est estimé par l'expression suivante :

$$\frac{n^2}{(n-1)(n-2)}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Dans le cas d'une distribution symétrique, ce coefficient est nul.

- **Le coefficient d'aplatissement (Kurtosis) :**

Le coefficient d'aplatissement (proposé par Fisher) d'une variable aléatoire X est défini par :

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3,$$

avec σ l'écart type et $\mu_4 = \mathrm{E}(X - \mathrm{E}(\mathrm{X}))^4$.

C'est une quantité mesurant l'épaisseur des queues de la distribution. Si $\gamma_2 = 0$, la distribution a des queues gaussiennes. Si $\gamma_2 > 0$, les queues sont plus épaisses que la loi normale. Pour $\gamma_2 < 0$, les queues sont plus légères que la loi normale.

L'expression du coefficient d'asymétrie sous SAS s'obtient en remplaçant :

- La variance σ^2 par son estimateur sans biais s^2
- μ_4 par son estimateur sans biais

Ainsi, γ_2 est estimé par l'expression suivante :

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(\frac{x_{i-\bar{x}}}{s})^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Dans le cas d'une distribution normale, ce coefficient est nul.

Annexe C : Les différents tests de normalité

- **Test de Kolmogorov-Smirnov (KS):**

Principe :

Le test de Kolmogorov-Smirnov est un test non paramétrique qui consiste à mesurer l'écart en valeur absolue entre la fonction de répartition (fonction de densité cumulée) de la variable testée et la fonction de répartition d'une variable gaussienne (ou plus généralement, de toute variable continue dont on veut comparer la distribution à celle de la variable observée). En d'autres termes, ce test consiste à comparer deux fonctions de répartition.

Ce test permet de déterminer si deux échantillons suivent une même loi statistique et de déterminer si un échantillon suit une loi statistique connue.

Contexte :

On considère ainsi une variable aléatoire X possédant une fonction de répartition que l'on veut comparer à une fonction de répartition théorique continue. On souhaite tester les hypothèses suivantes :

$$H_0: "\ X\ suit\ la\ loi\ \mathcal{L}\ "\ \ \text{contre}\ \ H_1: "X\ \ ne\ suit\ pas\ la\ loi\ \mathcal{L}"$$

Si $(X_1, \dots, X_n)$ est un n-échantillon de X, la fonction de répartition empirique associée à cet échantillon est :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i \leq x)$$

où $1(x_i \leq x)$ est la fonction indicatrice définie par :

$$1(x_i \leq x) = \begin{cases} 1\ si\ \ x_i \leq x \\ 0 \quad\ sinon \end{cases}$$

$F_n(x)$ est la fonction de répartition empirique associée aux données (proportion des observations dont la valeur est inférieure ou égale à x) et $F(x)$ est la fonction de répartition associée à la loi $\mathcal{L}$.

L'idée du test de Kolmogorov-Smirnov est que plus $F_n(x)$ diffère de $F(x)$ plus le rejet de H_0 est significatif.

Le test :

Le test consiste à calculer l'écart suivant :

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

qui sera l'écart de décision, ou fonction discriminante du test.

Dans un premier temps on peut montrer aisément que la distribution de la statistique D_n, ne dépend pas de la fonction supposée $F()$. En effet,

$$\begin{aligned} |F_n(x) - F(x)| &= \left|\frac{1}{n}\sum_{i=1}^{n} 1(x_i \leq x) - F(x)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n} 1[F(x_i) \leq F(x)] - F(x)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n} 1[y_i \leq y] - y\right| \end{aligned}$$

où dans la dernière égalité $y_i = F(x_i)$ est la réalisation d'une uniforme sur $[0,1]$.

Ainsi, $|F_n(x) - F(x)| = |F_{uni}(y) - y|$, où $F_{uni}(y)$ est la fonction de répartition empirique d'une uniforme sur $[0,1]$. En conséquence,

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

ne dépend pas de $F()$.

D'autre part, le théorème de Kolmogorov stipule :

Théorème de Kolmogorov : *Pour un ensemble de n variables aléatoires $i.i.d$ de fonction de répartition continue F, on a $P\left(\sqrt{n}\ D_n \leq x\right) \underset{n\to\infty}{\longrightarrow} K(x)$, où $K(x)$ est la fonction de répartition de Kolmogorov définie par :*

$$K(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2}$$

Pour des faibles valeurs de n il existe des tables donnant les valeurs critiques aux seuils de risques usuels. Pour les tailles d'échantillon suffisamment grandes on peut utiliser les propriétés asymptotiques et donc calculer $K(x)$.

- **<u>Les tests de Cramer Von Mises et d'Anderson-Darling :</u>**

Le test de Cramer Von Mises repose sur le même principe que le test de Kolmogorov-Smirnov, la statistique de test calcule la somme des carrés des écarts en valeur absolue entre les deux fonctions de répartition. Il est souvent plus puissant que le test de Kolmogorov-Smirnov.
Le test d'Anderson-Darling est variante du test de Kolmogorov-Smirnov, qui donne plus d'importance aux queues de distribution.

Ces deux tests débutent avec la même logique que le test KS, à savoir examiner la distance entre la fonction de répartition théorique supposée sous H_0 et la fonction de répartition empirique construite sur l'échantillon $F_n(x)$. Ils diffèrent sur deux points :

- Leur distance fait intervenir l'écart quadratique, $(F_n(x) - F(x))^2$, et non plus l'écart absolu $|F_n(x) - F(x)|$,
- Alors que dans KS on regarde seulement la distance maximale entre les deux fonctions, ils vont considérer l'ensemble des observations.

Leur expression générale est donc :

$$Q = n\int_{-\infty}^{+\infty}\left(F_n(x) - F(x)\right)^2\psi(x)dF(x),$$

où $\psi(x)$ est une fonction de pondération qui va caractériser l'un ou l'autre test.

Une conséquence majeure de ces variations est que la distribution des nouvelles statistiques, et donc leurs valeurs critiques vont dépendre de le fonction $F(x)$ retenue par l'hypothèse nulle.

1. *Le test de Cramer-Von-Mises*
   La fonction de pondération est $\psi(x) = 1$ et la statistique de test est :
   $$W^2 = \sum_{i=1}^{n}(\frac{2i-1}{2n} - F(x_{(i)}))^2 + \frac{1}{12n}$$
   où $x_{(i)}$ est la $i^{\text{ème}}$ plus petite valeur de l'échantillon.
   Une grande valeur de la statistique W^2 est un signe défavorable à H_0 : on va rejeter cette hypothèse lorsque W^2 est supérieur à sa valeur critique.
2. *Le test d'Anderson-Darling*

La fonction de pondération est $\psi(x) = [F(x)(1-F(x))]^{-1}$ et la statistique de test est

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-n)\log F\left(x_{(i)}\right) + (2n+1-2i)\log(1-F\left(x_{(i)}\right))\right]$$

Dans les tests de Cramer-Von-Miles un même poids est donné à toutes les observations quel que soit leur rang. Dans le test d'Anderson-Darling, la fonction de pondération utilisée donne plus de poids aux observations situées dans les queues de distribution. La statistique A^2 peut donc être intéressante à regarder lorsqu'on veut précisément prêter attention aux écarts entre les deux fonctions pour les valeurs situées dans les queues de distribution.